Research Report

KSTS/RR-06/001
February 13, 2006

High dimensional data visualisation: Textile plot

by

Natsuhiko Kumasaka
Ritei Shibata

Natsuhiko Kumasaka
Keio University
Ritei Shibata
Keio University

Department of Mathematics
Faculty of Science and Technology
Keio University

3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522 Japan

KSTS/RR-06/001
February 13, 2006

High Dimensional Data Visualisation: Textile Plot

Natsuhiko Kumasaka, Ritei Shibata

Fundamental Science and Technology, Keio University, Yokohama, Japan
Department of Mathematics, Keio University, Yokohama, Japan

Abstract

Textile plot is a parallel coordinate plot on which the whole location and scales of the axes are simultaneously chosen so that all connecting lines, each of which signifies an observation, are aligned as horizontally as possible. Textile plot can visualise not only numerical data but also ordered or unordered categorical data or the mix of those. Textile plot makes easier for user to understand the given data because of the design principle. Various attributes are displayed together with the parallel coordinate plot and the missing values are dealt with properly. Several theorems necessary for the computation of Textile plot are given and two outstanding features Knot and Neat Weft are characterised by several simple conditions. The design principle of the Textile plot is also discussed into detail.

1 Introduction

Parallel coordinate plot has been frequently used for exploring high dimensional data. It is proposed by Inselberg [5] as a tool for visualising a high dimensional geometry on a two-dimensional display. Wegman [10] has developed it into a tool for visualising multivariate data. The basic idea of the parallel coordinate plot is to place all axes in parallel on a two dimensional display, and the coordinates of each observation are connected by segments from left to right to signify an observational point in the high dimensional space. The parallel coordinate plot is a good way of visualising very high-dimensional data, but it is also true that it becomes difficult to understand any mechanism behind the data as the number of the intersections of connecting lines increases.

Email address: `kumasaka@stat.math.keio.ac.jp` (Natsuhiko Kumasaka).
URL: `http://www.stat.math.keio.ac.jp/DandDIV/` (Natsuhiko Kumasaka).

Textile plot is a solution to such a problem. The location and scales of the axes are chosen so that all connected lines are aligned as horizontally as possible. In other words, given data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, they are always transformed into p coordinate vectors,

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

to make a Textile plot with a common coordinate system through axes. The parameters α_j's and β_j's are chosen so that the sum of squares $\sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{\xi}\|^2$ is minimised to a vector $\boldsymbol{\xi}$ which is also determined so as the sum of squares to be minimised. Textile plot is named after a fabric into which warp and weft yarns are woven. Our criterion reflects the fact that any weft yarns run as horizontally as possible to make a neat textile. This is a good contrast to the parallel coordinate plot, on which the location and scale of each axis are chosen so that all coordinate points on the axis fill up the axis, that is, the given data vector $\boldsymbol{x}_j$ is always transformed into

$$\boldsymbol{y}_j = \frac{\boldsymbol{x}_j - \min(\boldsymbol{x}_j)\mathbf{1}}{\max(\boldsymbol{x}_j) - \min(\boldsymbol{x}_j)}, \quad j = 1, \ldots, p,$$

before display. Note that the transformation is determined axis by axis and no global criterion is employed in the parallel coordinate plot.

Textile plot does not only make easier for people to understand relationship between adjacent axes, but also make possible to grasp a global relationship among data vectors. Furthermore, any ordered or unordered categorical data can be displayed on a plot together with numerical data in which missing values can exist. This is due to our general criterion to choose the location and scales of the axes. For example, the positional vector for the levels of a categorical data vector is automatically determined by the criterion as far as the data vector is encoded by a set of contrasts. The result does not depend on the choice of the contrast.

There are several related works which have been done in the field of *Homogeneity analysis* [3]. In homogeneity analysis, any categorical data vectors are quantified so as to make the distance from the *object scores* as small as possible. Usually a 2D plot is produced to visualise the first two quantified vectors but sometimes a parallel coordinate plot is employed to visualise all quantified vectors, which is called *Optimised Parallel Coordinate Plot*([6]). In fact, the plot produces the same picture as that produced by Textile in a specific case where all data vectors are categorical and no missing value exists. The objective of the Textile plot is, however, different from that of homogeneity analysis. Textile plot is a tool for visualising and browsing high dimensional data as it is. Symbols for points displayed are carefully chosen so that any necessary and sufficient information is provided in a neat way. Furthermore, the order of axis is also carefully chosen so as to give a clear image of the given data to the user.

In the next section we will establish several theorems for the computation to produce a Textile plot.

2 Choice of Location and Scales

2.1 Numerical Data

We will use the following notations. The norm $\|\boldsymbol{x}\|_{\boldsymbol{v}}^2 = \sum_{i=1}^n v_i x_i^2$ is a weighted norm of a vector $\boldsymbol{x}$ with a non-negative weight vector $\boldsymbol{v}$, and $\boldsymbol{x} \cdot \boldsymbol{v}$ and $\boldsymbol{x}/\boldsymbol{v}$ are element-wise product and division of the vectors $\boldsymbol{x}$ and $\boldsymbol{v}$.

We first consider the case when all n-dimensional data vectors are numerical. Let us denote them by a matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = (x_{ij};\ 1 \le i \le n,\ 1 \le j \le p)$. As is described in the previous section, in the Textile plot the given data vectors $\{\boldsymbol{x}_j,\ j = 1, \ldots, p\}$ are always transformed into the coordinate vectors,

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

for the display, where $\mathbf{1}$ is the vector of all ones. The location $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ and the scale $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ parameter vectors are simultaneously chosen so as the sum of squares

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

to be minimised. Here $\boldsymbol{w}_j$ is the jth column of a weight matrix $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p) = (w_{ij};\ 1 \le i \le n,\ 1 \le j \le p)$ of zero or one to indicate missing values. That is, $w_{ij} = 0$ if the value of x_{ij} is missing or invalid, otherwise 1. The use of such a weighted norm implies that all missing values will be neglected for the choice of location and scales of the axes on the Textile plot.

The vector $\boldsymbol{\xi}$ is not predetermined but the choice $\boldsymbol{\xi} = \boldsymbol{m} = \sum_{j=1}^p \boldsymbol{w}_j \cdot \boldsymbol{y}_j / \boldsymbol{w}$ gives a solution to minimise $S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ for any $\boldsymbol{y}_j$'s, since

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

where $\boldsymbol{w} = \sum_{j=1}^p \boldsymbol{w}_j$. We will call the vector $\boldsymbol{m}$ the *mean vector* for $\boldsymbol{y}_j$'s, the vector of the mean positions of each observations or records.

We need here a constraint for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to avoid trivial solutions like $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0}$. A natural constraint would be that the total dispersion of the points on the textile plot, $\sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2_{\boldsymbol{w}_j}$ remains a constant, for example, the effective number of the points $N = \sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}$, where $\bar{y}_{\cdot j} = \boldsymbol{w}_j^T\boldsymbol{y}_j/\mathbf{1}^T\boldsymbol{w}_j$ is the mean of the coordinates on the j th axis.

From the decomposition,

$$
\begin{aligned}
S^2(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{m}) &= \sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2_{\boldsymbol{w}_j} \\
&= \sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2_{\boldsymbol{w}_j} + \sum_{j=1}^{p} \|\bar{y}_{\cdot j}\mathbf{1}\|^2_{\boldsymbol{w}_j} - \|\boldsymbol{m}\|^2_{\boldsymbol{w}} \\
&= N + \sum_{j=1}^{p} \|\bar{y}_{\cdot j}\mathbf{1}\|^2_{\boldsymbol{w}_j} - \|\boldsymbol{m}\|^2_{\boldsymbol{w}}, \qquad (3)
\end{aligned}
$$

we see that it is enough to find $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which minimise

$$
f(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{j=1}^{p} \|\bar{y}_{\cdot j}\mathbf{1}\|^2_{\boldsymbol{w}_j} - \|\boldsymbol{m}\|^2_{\boldsymbol{w}}
$$

under the constraint that

$$
\sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2_{\boldsymbol{w}_j} = N. \qquad (4)
$$

Such an optimal choice of location and scales of the axes of the Textile plot always exists since $f(\boldsymbol{\alpha},\boldsymbol{\beta})$ is bounded above $-N$.

The function $f(\boldsymbol{\alpha},\boldsymbol{\beta})$ is rewritten as

$$
f(\boldsymbol{\alpha},\boldsymbol{\beta}) = \boldsymbol{\alpha}^T\mathbf{A}_{11}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\mathbf{A}_{12}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{A}_{22}\boldsymbol{\beta} \qquad (5)
$$

by using the matrix notations,

$$
\begin{aligned}
\mathbf{A}_{11} &= -\Big(\boldsymbol{w}_j^T(\boldsymbol{w}_k/\boldsymbol{w});\ j,k=1,\dots,p\Big) + \mathrm{diag}(\mathbf{1}^T\boldsymbol{w}_j;\ j=1,\dots,p), \\
\mathbf{A}_{12} &= \Big(\boldsymbol{w}_j^T(\boldsymbol{w}_k \cdot \boldsymbol{x}_k/\boldsymbol{w});\ j,k=1,\dots,p\Big) - \mathrm{diag}(\boldsymbol{w}_j^T\boldsymbol{x}_j;\ j=1,\dots,p),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{A}_{22} = &-\Big((\boldsymbol{w}_j \cdot \boldsymbol{x}_j)^T(\boldsymbol{w}_k \cdot \boldsymbol{x}_k/\boldsymbol{w});\ j,k=1,\dots,p\Big) \\
&+\mathrm{diag}\Big((\boldsymbol{w}_j^T\boldsymbol{x}_j)^2/\mathbf{1}^T\boldsymbol{w}_j;\ j=1,\dots,p\Big).
\end{aligned}
$$

Also the constraint (4) is rewritten as

$$
\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta} = N, \qquad (6)
$$

where $\mathbf{B} = \mathrm{diag}(\|\boldsymbol{x}_j - \bar{x}_{\cdot j}\mathbf{1}\|^2_{\boldsymbol{w}_j};\ j = 1, \ldots, p)$ and $\bar{x}_{\cdot j} = \boldsymbol{w}_j^T \boldsymbol{x}_j / \mathbf{1}^T \boldsymbol{w}_j$ is the mean of the data vector $\boldsymbol{x}_j$.

Since the constraint (6) is related only to the parameter vector $\boldsymbol{\beta}$, the solution $\hat{\boldsymbol{\alpha}}$ should always satisfy

$$\mathbf{A}_{11}\hat{\boldsymbol{\alpha}} = \mathbf{A}_{12}\hat{\boldsymbol{\beta}}$$

for a solution $\hat{\boldsymbol{\beta}}$. To write explicitly $\hat{\boldsymbol{\alpha}}$, we need a generalised inverse of $\mathbf{A}_{11}$ since $\mathbf{A}_{11}$ is singular. In fact, by using the Moor-Penrose inverse $\mathbf{A}_{11}^+$ [7] of $\mathbf{A}_{11}$ it is written as

$$\hat{\boldsymbol{\alpha}} = \mathbf{A}_{11}^+\mathbf{A}_{12}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{A}_{11}^+\mathbf{A}_{11})\boldsymbol{z} \tag{7}$$

where $\boldsymbol{z}$ is arbitrary p-dimensional vector. Noting the fact that

$$(\mathbf{A}_{12}\boldsymbol{\beta})^T(\mathbf{I} - \mathbf{A}_{11}^+\mathbf{A}_{11})\boldsymbol{z} = (\mathbf{A}_{11}\boldsymbol{\alpha})^T(\mathbf{I} - \mathbf{A}_{11}^+\mathbf{A}_{11})\boldsymbol{z} = 0$$

which follows from the definition of the generalised inverse $\mathbf{A}_{11}^+$, we have

$$\begin{aligned} f(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}) &= \boldsymbol{\beta}^T(-\mathbf{A}_{12}^T\mathbf{A}_{11}^+\mathbf{A}_{12} + \mathbf{A}_{22})\boldsymbol{\beta} - 2\boldsymbol{z}^T(\mathbf{I} - \mathbf{A}_{11}^+\mathbf{A}_{11})^T\mathbf{A}_{12}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^T(-\mathbf{A}_{12}^T\mathbf{A}_{11}^+\mathbf{A}_{12} + \mathbf{A}_{22})\boldsymbol{\beta}. \end{aligned} \tag{8}$$

It is now clear that the solution $\hat{\boldsymbol{\beta}}$ is an eigenvector of $\mathbf{A} = \mathbf{A}_{12}^T\mathbf{A}_{11}^+\mathbf{A}_{12} - \mathbf{A}_{22}$ with respect to $\mathbf{B}$ for the largest eigenvalue.

We need the following assumptions for Theorem 1.

Assumption 1 *There is no record of all missing values.*

Assumption 2 *No $\boldsymbol{x}_j$ is a vector of a unique value, for $j = 1, \ldots, p$.*

The Assumption 1 assures that all elements of $\boldsymbol{w}$ are positive, and the Assumption 2 assumes that no trivial data vector is included in the given data set.

Theorem 1 *For the given numerical data vectors $\boldsymbol{x}_j$, $j = 1, \ldots, p$, which satisfy* Assumptions 1 *and* 2*, an optimal choice of location and scale vectors is given by $\hat{\boldsymbol{\alpha}} = \mathbf{A}_{11}^+\mathbf{A}_{12}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{A}_{11}^+\mathbf{A}_{11})\boldsymbol{z}$ for an arbitrary p-dimensional vector $\boldsymbol{z}$, and $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}}^T\mathbf{B}\hat{\boldsymbol{\beta}} = N$, which is an eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$ for the largest eigenvalue.*

Optimal choice of location and scales is not necessarily unique. However, if $\mathrm{rank}(\mathbf{A}_{11}) = p - 1$, then the choice of $\boldsymbol{\alpha}$ becomes essentially unique and $\hat{\boldsymbol{\alpha}} = c\mathbf{1} + \mathbf{A}_{11}^+\mathbf{A}_{12}\hat{\boldsymbol{\beta}}$ for an arbitrary global constant c. This is because $\mathrm{Ker}(\mathbf{A}_{11}) = \mathrm{span}\{\mathbf{1}\}$ if $\mathrm{rank}(\mathbf{A}_{11}) = p - 1$. Here note that $\mathbf{1} \in \mathrm{Ker}(\mathbf{A}_{11})$ always holds true. The uniqueness of $\hat{\boldsymbol{\beta}}$ is of course equivalent to that of the eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$ for the largest eigenvalue.

Theorem 1 becomes simpler if no missing value exists.

Corollary 1 *If there is no missing value in* $\mathbf{X}$*, an optimal choice of the location and scales is given by*

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

and

$$\hat{\beta}_j = \frac{1}{\|\boldsymbol{x}_j - \bar{x}_{\cdot j}\mathbf{1}\|}\gamma_j,\ j = 1, \ldots, p,$$

where α_0 *is an arbitrary constant,* $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ *is an eigenvector of the sample correlation matrix of* $\boldsymbol{x}_j$*,* $j = 1, \ldots, p$ *for the largest eigenvalue and* $\|\boldsymbol{\gamma}\|^2 = N = np$*.*

PROOF. It is easily seen that

$$\mathbf{A}_{11}^{+} = \frac{1}{n^2}\mathbf{A}_{11} = \frac{1}{n}\left(\mathbf{I} - \frac{1}{p}\mathbf{1}\mathbf{1}^T\right),$$

and

$$\mathbf{A}_{12} = -n\left(\mathrm{diag}(\bar{\boldsymbol{x}}) - \frac{1}{p}\mathbf{1}\bar{\boldsymbol{x}}^T\right),$$

where $\bar{\boldsymbol{x}} = (\bar{x}_{\cdot 1}, \ldots, \bar{x}_{\cdot p})^T$. Then we have

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= \mathbf{A}_{11}^{+}\mathbf{A}_{12}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{A}_{11}^{+}\mathbf{A}_{11})\boldsymbol{z} \\
&= \frac{1}{n}\mathbf{A}_{12}\hat{\boldsymbol{\beta}} + \frac{1}{p}\mathbf{1}\mathbf{1}^T\boldsymbol{z} \\
&= \frac{1}{p}\left(\frac{\mathbf{1}^T\mathbf{X}\hat{\boldsymbol{\beta}}}{n} + \mathbf{1}^T\boldsymbol{z}\right)\mathbf{1} - \bar{\boldsymbol{x}} \cdot \hat{\boldsymbol{\beta}} \\
&= \alpha_0\mathbf{1} - \bar{\boldsymbol{x}} \cdot \hat{\boldsymbol{\beta}}.
\end{aligned}$$

The sample correlation matrix of $\boldsymbol{x}_j$, $j = 1, \ldots, p$ can be written as

$$\mathbf{R} = p\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}},$$

since $p\mathbf{A}/n$ is the sample covariance matrix of $\boldsymbol{x}_j$, $j = 1, \ldots, p$. Therefore we see that $\hat{\boldsymbol{\beta}} = \mathbf{B}^{-\frac{1}{2}}\boldsymbol{\gamma}$. □

2.2 Numerical and Categorical Data

Even if some of data vectors are categorical, a similar theorem holds true to Theorem 1 as far as such categorical data vectors are encoded by a set of contrasts.

By introducing a $q \times (q-1)$ contrast matrix $\mathbf{C}$ whose column vectors are linearly independent of $\mathbf{1}$ for an n-dimensional categorical data vector $\boldsymbol{x}$ with q levels, we can encode it into an $n \times (q-1)$ data matrix $\mathbf{X}$. It is then transformed into a coordinate vector

$$\boldsymbol{y} = \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta}$$

for the display of the Textile plot. The location parameter is now not a single value but a $\boldsymbol{\beta}$ is a $(q-1)$-dimensional vector for a categorical data vector $\boldsymbol{x}$.

Example 1 *The categorical data vector* $\boldsymbol{x} = (A, B, C)^T$ *is transformed into the coordinate vector,*

$$\boldsymbol{y} = \alpha \mathbf{1} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha + \beta_1 \\ \alpha + \beta_2 \end{pmatrix},$$

if a Treatment contrast $\mathbf{C}^T = (\mathbf{0}, \mathbf{I})$ *is used.*

To cover cases where both numerical and categorical data vectors exist, we will use a matrix notation $\mathbf{X}_j$ for numerical data vector $\boldsymbol{x}_j$ where $q_j = 2$ for the consistency of notations. Then, given data matrices $\{\mathbf{X}_j,\ j = 1 \ldots, p\}$ they are combined into a $n \times Q$ data matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, where $Q = \sum_{j=1}^{p}(q_j - 1)$. The index set

$$\mathscr{I}_j = \left\{ \sum_{i=1}^{j-1}(q_i - 1) + 1, \ldots, \sum_{i=1}^{j}(q_i - 1) \right\}$$

signifies the columns in $\mathbf{X}$ corresponding to the data vector $\boldsymbol{x}_j$. The whole index set is then

$$\mathscr{I} = \bigcup_{j=1}^{p} \mathscr{I}_j = \{1, \ldots, Q\}.$$

By using notations $\boldsymbol{v}(\mathscr{K})$ or $\mathbf{M}(\mathscr{K}, \mathscr{L})$ for the sub-vector or the sub-matrix specified by index set $\mathscr{K}$ and $\mathscr{L}$ [4], we can write the transformation to coordinate vector in a general manner,

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

where $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_p)$ and $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_Q)$ are scale and location parameter vectors, respectively.

The matrix notations used in the previous section have to be re-defined except $\mathbf{A}_{11}$ given both numerical and categorical data vectors. The $\mathbf{A}_{12}$ is a $p \times Q$ matrix such that

$$\mathbf{A}_{12}(j, \mathscr{I}_k) = \begin{cases} \boldsymbol{w}_j^T(\boldsymbol{w}_k \cdot \mathbf{X}_k / \boldsymbol{w}) & j \neq k, \\ \boldsymbol{w}_j^T(\boldsymbol{w}_k \cdot \mathbf{X}_k / \boldsymbol{w}) - \boldsymbol{w}_j^T \mathbf{X}_j & j = k, \end{cases}$$

for $j, k = 1, \dots, p$. The $\mathbf{A}_{22}$ and $\mathbf{B}$ are $Q \times Q$ matrices as

$$\mathbf{A}_{22}(\mathscr{I}_j, \mathscr{I}_k) = \begin{cases} -(\boldsymbol{w}_j \cdot \mathbf{X}_j)^T(\boldsymbol{w}_k \cdot \mathbf{X}_k/\boldsymbol{w}) & j \neq k, \\ -(\boldsymbol{w}_j \cdot \mathbf{X}_j)^T(\boldsymbol{w}_k \cdot \mathbf{X}_k/\boldsymbol{w}) & \\ \qquad +\mathbf{X}_j^T\boldsymbol{w}_j\boldsymbol{w}_j^T\mathbf{X}_j/(\mathbf{1}^T\boldsymbol{w}_j) & j = k, \end{cases}$$

and

$$\mathbf{B}(\mathscr{I}_j, \mathscr{I}_k) = \begin{cases} \mathbf{O} & j \neq k, \\ \mathbf{X}_j^T(\boldsymbol{w}_j \cdot \mathbf{X}_j) & \\ \qquad -\mathbf{X}_j^T\boldsymbol{w}_j\boldsymbol{w}_j^T\mathbf{X}_j/(\mathbf{1}^T\boldsymbol{w}_j) & j = k, \end{cases}$$

for $j, k = 1, \dots, p$.

We have used the notations $\cdot$ and $/$ in a slightly extended way to accommodate matrix in place of vector, that is, $\boldsymbol{v} \cdot \mathbf{Z} = (\boldsymbol{v} \cdot \boldsymbol{z}_1, \dots, \boldsymbol{v} \cdot \boldsymbol{z}_r)$ and $\mathbf{Z}/\boldsymbol{v} = (\boldsymbol{z}_1/\boldsymbol{v}, \dots, \boldsymbol{z}_r/\boldsymbol{v})$ for an n-dimensional vector $\boldsymbol{v}$ and an $n \times r$ matrix $\mathbf{Z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_r)$.

To show the Textile plot does not depend on the choice of the contrast, we need the following lemma.

Lemma 1 *Given two $q \times (q-1)$ contrast matrices $\mathbf{G}$ and $\mathbf{H}$, there always exist a $(q-1)$-dimensional vector $\boldsymbol{c}$ and a non-singular $(q-1) \times (q-1)$ matrix $\mathbf{D}$ such that*

$$\mathbf{H} = \mathbf{1}\boldsymbol{c}^T + \mathbf{G}\mathbf{D}.$$

PROOF. Since $\mathrm{rank}(\mathbf{G}) = q - 1$ and all column vectors of $\mathbf{G}$ are linearly independent of $\mathbf{1}$, there exist $c_i \in \mathbb{R}$ and $\boldsymbol{d}_i \in \mathbb{R}^{q-1}$ such that

$$\boldsymbol{h}_i = c_i\mathbf{1} + \mathbf{G}\boldsymbol{d}_i$$

for each column vector $\boldsymbol{h}_i$ of $\mathbf{H}$, $i = 1, \dots, q-1$. Therefore, we have the desired result by taking $\boldsymbol{c} = (c_1, \dots, c_{q-1})$ and $\mathbf{D} = (\boldsymbol{d}_1, \dots, \boldsymbol{d}_{q-1})$.The non-singularity of $\mathbf{D}$ is clear if we note that,

$$(\mathbf{1}, \mathbf{H}) = (\mathbf{1}, \mathbf{G}) \begin{pmatrix} 1 & \boldsymbol{c}^T \\ \mathbf{0} & \mathbf{D} \end{pmatrix}.$$

□

Theorem 2 *Given numerical or categorical data vectors $\boldsymbol{x}_j$, $j = 1, \dots, p$, which satisfy* Assumptions 1 *and* 2*, an optimal choice of location and scale vectors are given by $\hat{\boldsymbol{\alpha}} = \mathbf{A}_{11}^{+}\mathbf{A}_{12}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{A}_{11}^{+}\mathbf{A}_{11})\boldsymbol{z}$ for an arbitrary p-dimensional* *vector* $\boldsymbol{z}$ *and* $\hat{\boldsymbol{\beta}}$ *such that* $\hat{\boldsymbol{\beta}}^T\mathbf{B}\hat{\boldsymbol{\beta}} = N$*, which is an eigenvector of* $\mathbf{A}$ *with respect to* $\mathbf{B}$ *for the largest eigenvalue. The result is independent of the choice of the contrast.*

PROOF. The proof of the former part of the theorem is exactly the same as that of Theorem 1, so that it is enough to show that the result is independent of the choice of the contrast. Assume that two different contrasts $\mathbf{G}$ and $\mathbf{H}$ are used for a categorical data vector. Then, two different encoded matrices $\mathbf{X_G}$ and $\mathbf{X_H}$ are obtained, the coordinate vector $\boldsymbol{y}$ is differently represented as

$$\boldsymbol{y} = \alpha\mathbf{1} + \mathbf{X_G}\boldsymbol{\beta},$$

or

$$\boldsymbol{y} = \alpha\mathbf{1} + \mathbf{X_H}\boldsymbol{\beta}.$$

However, we see that Lemma 1implies that there exist a $(q-1)$-dimensional vector $\boldsymbol{c}$ and a $(q-1)\times(q-1)$ nonsingular matrix $\mathbf{D}$ such that

$$\mathbf{X_H} = \mathbf{1}\boldsymbol{c}^T + \mathbf{X_G}\mathbf{D}.$$

Therefore, the linear spaces of $\boldsymbol{y}$ are the same although the representations are different, so that the same optimal choice of scale and location parameter vectors can be obtained. $\square$

Theorem 2 becomes simpler if no missing value exists. The matrix $\mathbf{A}$ becomes

$$\mathbf{A} = \frac{1}{p}\left(\mathbf{X}^T\mathbf{X} - \frac{1}{n}\mathbf{X}^T\mathbf{1}\mathbf{1}^T\mathbf{X}\right), \tag{9}$$

and the matrix $\mathbf{B}$ becomes

$$\mathbf{B}(\mathscr{I}_j, \mathscr{I}_k) = \begin{cases} \mathbf{O} & j \neq k, \\ \mathbf{X}_j^T\mathbf{X}_j - \mathbf{X}_j^T\mathbf{1}\mathbf{1}^T\mathbf{X}_j/n & j = k, \end{cases} \tag{10}$$

for $j, k = 1, \ldots, p$.

Corollary 2 *If there is no missing value in* $\mathbf{X}$*, an optimal choice of the locations is given by*

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

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* $\hat{\boldsymbol{\beta}}$ *such that* $\hat{\boldsymbol{\beta}}^T\mathbf{B}\hat{\boldsymbol{\beta}} = N$ *as an eigenvector of* $\mathbf{A}$ *in (9) with respect to* $\mathbf{B}$ *in (10) for the largest eigenvalue.*

2.3 General Result

We have not taken into consideration the case when some of categorical data vectors are ordered. For such ordered categorical data vector, we have to retain the order of the levels for the display.

Example 2 *Given an ordered categorical data vector* $\boldsymbol{x} = (Small, Medium, Large)^T$*, if it is encoded by a contrast,*

$$\mathbf{C} = \begin{pmatrix} 0 & \cdots & 0 \\ 1 & \ddots & \vdots \\ \vdots & \ddots & 0 \\ 1 & \cdots & 1 \end{pmatrix} \tag{11}$$

then the coordinate vector is parametrised as

$$\boldsymbol{y} = \alpha\mathbf{1} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha + \beta_1 \\ \alpha + \beta_1 + \beta_2 \end{pmatrix}.$$

To retain the order of the levels as Small < Medium < Large, we need an additional constraint

$$\beta_1 \geq 0, \;\; \beta_2 \geq 0.$$

or

$$\beta_1 \leq 0, \;\; \beta_2 \leq 0,$$

since the direction, upward or downward, of each axis is arbitrary in the Textile plot.

As is seen from this example, we need an additional inequality constraint if an ordered categorical data vector were included in the given data vectors. Although the resulting coordinate vector remains unchanged even by the choice of the constraint, the inequality constraint becomes more complicated. Therefore, we hereafter use the constraint as in (11) for ordered categorical data vector and assume that the levels are in an increasing order.

To simplify the problem, we assume that the first r data vectors $\boldsymbol{x}_k$, $k = 1, \ldots, r$ are ordered categorical data vectors and $\boldsymbol{x}_k$, $k = r+1, \ldots, p$ are other types of data vectors. Then the problem is to minimise

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \mathbf{A}_{11} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \mathbf{A}_{12} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{A}_{22} \boldsymbol{\beta} \tag{12}$$

under the equality constraint

$$\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} = N \tag{13}$$

and the inequality constraint

$$\boldsymbol{\beta}(\mathscr{I}_k) \geq \mathbf{0} \quad \text{or} \quad \boldsymbol{\beta}(\mathscr{I}_k) \leq \mathbf{0}, \quad k = 1, \ldots, r. \tag{14}$$

Here, the matrices $\mathbf{A}_{11}, \mathbf{A}_{12}$ and $\mathbf{A}_{22}$ are the same matrices as before and the notations $\geq$ or $\leq$ are used as an element-wise inequality for two vectors. That is, $\boldsymbol{u} \geq \boldsymbol{v}$ for $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^k$ according to $u_i \geq v_i$, for all $i = 1, \ldots, k$.

Since $\hat{\boldsymbol{\alpha}}$ is still given the same as in (7), it is enough to consider the minimisation problem of the $\boldsymbol{\beta}$ by introducing the following Lagrangian function,

$$L(\boldsymbol{\beta}; \lambda, \boldsymbol{\mu}) = -\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} + \lambda(\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} - N) + \sum_{k=1}^{r} \boldsymbol{\mu}(\mathscr{I}_k)^T \boldsymbol{\beta}(\mathscr{I}_k),$$

where λ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_R)^T$ are Lagrange multipliers and $\mathbf{A}$ and $\mathbf{B}$ are the same as in the previous section. The following proposition is an application of the well known result, for example, see Proposition 1.29 in [1].

Proposition 1 *The solution $\hat{\boldsymbol{\beta}}$ is characterised by the following condition. There exist $\hat{\lambda}$ and $\hat{\boldsymbol{\mu}}$ such that*

$$\nabla_{\boldsymbol{\beta}} L(\hat{\boldsymbol{\beta}}; \hat{\lambda}, \hat{\boldsymbol{\mu}}) = \mathbf{0}, \tag{15}$$

and either

$$\hat{\boldsymbol{\mu}}(\mathscr{I}_k) \geq \mathbf{0}, \qquad \hat{\boldsymbol{\mu}}(\mathscr{I}_k) \cdot \hat{\boldsymbol{\beta}}(\mathscr{I}_k) = \mathbf{0}, \qquad \hat{\boldsymbol{\beta}}(\mathscr{I}_k) \leq \mathbf{0} \tag{16}$$

or

$$\hat{\boldsymbol{\mu}}(\mathscr{I}_k) \leq \mathbf{0}, \qquad \hat{\boldsymbol{\mu}}(\mathscr{I}_k) \cdot \hat{\boldsymbol{\beta}}(\mathscr{I}_k) = \mathbf{0}, \qquad \hat{\boldsymbol{\beta}}(\mathscr{I}_k) \geq \mathbf{0} \tag{17}$$

is satisfied for $k = 1, \ldots, r$.

The condition (16) or (17) in the proposition implies that any element of $\hat{\boldsymbol{\mu}}(\mathscr{I}_k) = -\nabla_{\boldsymbol{\beta}(\mathscr{I}_k)} g(\hat{\boldsymbol{\beta}}, \hat{\lambda})$ should be zero as far the corresponding element of the solution $\hat{\boldsymbol{\beta}}$ is not on the boundary of the inequality constraint in (14), where

$$g(\boldsymbol{\beta}; \lambda) = -\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} + \lambda(\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} - N).$$

We have a practical algorithm in the following theorem from this observation.

Theorem 3 *Given ordered categorical data vectors $\boldsymbol{x}_j$, $j = 1, \ldots, r$ and other data vectors $\boldsymbol{x}_j$, $j = r + 1, \ldots, p$, which satisfy Assumptions 1 and 2, an* *optimal choice of the locations is given by* $\hat{\boldsymbol{\alpha}} = \mathbf{A}_{11}^{+}\mathbf{A}_{12}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{A}_{11}^{+}\mathbf{A}_{11})\boldsymbol{z}$ *for an arbitrary p-dimensional vector* $\boldsymbol{z}$*. That of the scales* $\hat{\boldsymbol{\beta}}$ *can be obtained by finding all index sets* $\mathscr{I}_0 \subseteq \bigcup_{k=1}^{r} \mathscr{I}_k$ *such that*

(1) $\hat{\boldsymbol{\beta}}(\mathscr{I}_0) = \mathbf{0}$*, and* $\hat{\boldsymbol{\beta}}(\mathscr{I}_0^c)$ *such that* $\hat{\boldsymbol{\beta}}(\mathscr{I}_0^c)\mathbf{B}(\mathscr{I}_0^c, \mathscr{I}_0^c)\hat{\boldsymbol{\beta}}(\mathscr{I}_0^c) = N$ *is an eigenvector of* $\mathbf{A}(\mathscr{I}_0^c, \mathscr{I}_0^c)$ *with respect to* $\mathbf{B}(\mathscr{I}_0^c, \mathscr{I}_0^c)$ *for the largest eigenvalue* $\hat{\lambda}$*, where* $\mathscr{I}_0^c = \mathscr{I} \setminus \mathscr{I}_0$*,*
(2) either $\hat{\boldsymbol{\beta}}(\mathscr{I}_k \cap \mathscr{I}_0^c) > \mathbf{0}$ *or* $\hat{\boldsymbol{\beta}}(\mathscr{I}_k \cap \mathscr{I}_0^c) < \mathbf{0}$ *is satisfied for* $k = 1, \ldots, r$*,*

select an $\mathscr{I}_0$ *for which the* $\hat{\lambda}$ *is the largest, and normalise it so that* $\hat{\boldsymbol{\beta}}\mathbf{B}\hat{\boldsymbol{\beta}} = N$*.*

PROOF. From Proposition 1, we see that it is enough to compare the solutions which satisfy conditions (1) and (2) for all possible $\mathscr{I}_0$'s. Note that

$$L(\hat{\boldsymbol{\beta}}; \hat{\lambda}, \hat{\boldsymbol{\mu}}) = g(\hat{\boldsymbol{\beta}}, \hat{\lambda}) = -\hat{\lambda}N$$

for any such solutions, then the desired result follows.

3 Design of Textile Plot

We have developed several theorems to find an optimal choice of location and scales of each axes. A simple Textile plot would be a plot of the points according to the coordinate vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ on the parallel axes with a coordinate system. However, the aim of Textile plot is not only to visualise data but also to assist human being to understand various aspect of the given data. Therefore it is necessary to display only the points but also some other helpful information on the Textile plot. We will describe various aspects of the design of Textile plot in the following sections.

"A graphical method is successful only if the decoding process of human being from the given graphics is effective" (Cleveland [2]).

3.1 Point Display on a Warp Yarn

In this section, we will discuss the way of point display on a warp yarn. It is clear that the display of enough attributes of each data vector is indispensable. One of such attributes is so called *data type*. Besides a well known distinction between quantitative or categorical, a further classification is needed to be informative, particularly in case of high dimensional data. We first classify a data vector into numerical or non-numerical and further classify numerical data into *continuous* or *discrete*, and classify non-numerical data into *ordered*

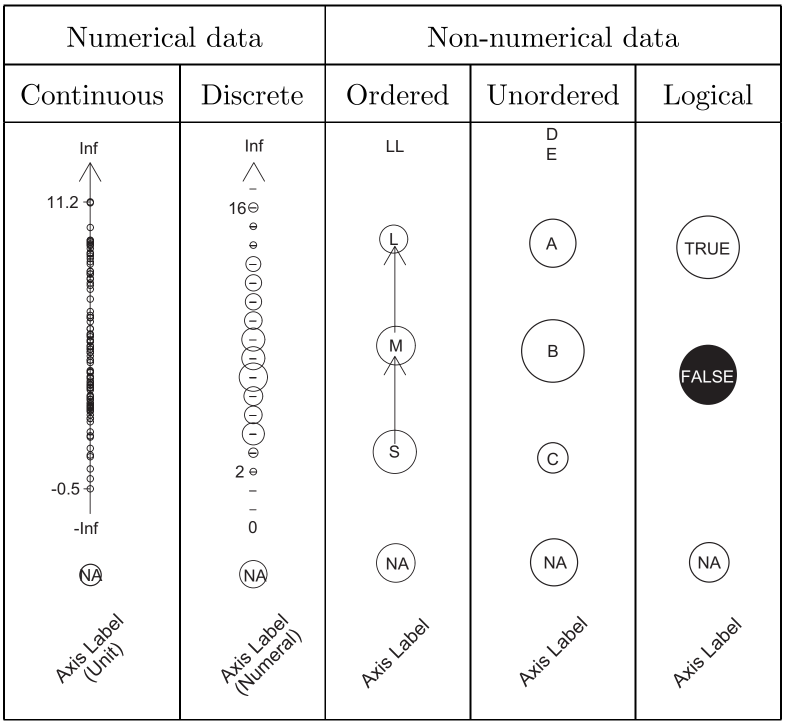

Fig. 1. Design of Point Display on a Warp Yarn

category, *unordered category* or *logical*. Although any further classification is possible, we restrict our attention into such five data types in order to avoid superfluity.

Figure 1 illustrates the way of point display on a warp yarn for each data type. In case of numerical data, indication of possible values is quite useful for understanding the data. The possible values are indicated by a continuous line if it is continuous and by ticks otherwise. Also, the possible maximum and minimum are shown by figures at the both end of the continuous line or of a series of the ticks. Then, it makes possible for user to understand well the background of the data but also makes easier to distinguish two data types, continuous or discrete. An arrow head placed on either end of the warp yarn indicates a direction of coordinates. It is upward on the jth warp yarn if $\beta_j \geq 0$, and downward otherwise. Circles are placed on each warp yarn instead of points. We can understand the coordinate from the centre and the number of duplication from the area. This is similar idea to that for categorical data in parallel coordinate plot [11]. The minimum and maximum of the given data vector are shown by labels with tick marks on the left hand side of these coordinates.

In case of non-numerical data, no definite coordinate exists a priori but we have now the coordinate vector by a transformation as is described in the previous section. The levels are indicated by circles on a warp yarn. Each level can be identified by the level name in the circle. We can see relative

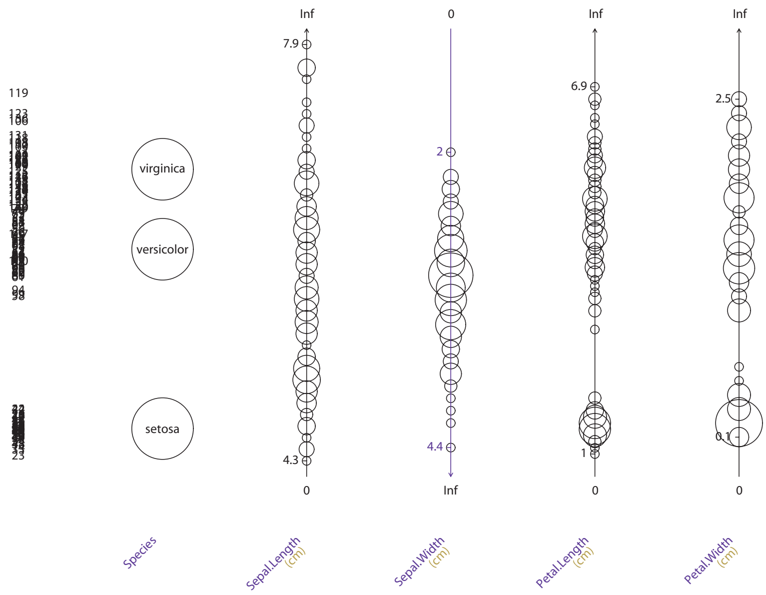

Fig. 2. Display of Warp Yarns for Iris Data

frequencies from the area of the circle. This design is consistent with that for numerical data. Zero frequency levels are indicated on the top of the display by the name without circle. This is similar to the display of possible values in case of discrete data. If it is ordered category, the levels are connected by a sequence of arrows to indicate a natural order of the levels. If it is a logical data, the circle for FALSE is daubed with black in order to stand out logical data from categorical one.

In any case, existence of missing values is indicated by a circle with "NA", which is separately placed on the bottom of each warp display. The area of the circle is again proportional to the number of missing values. As the label for each warp yarn, we can make use of the name of data vector if it is available. It is hard to understand the meaning without such labels. The unit or numeral attribute is indispensable in case of continuous or discrete data, so that it is indicated together with the label.

3.2 Display of Warp Yarns

According to the design of point display, all coordinate vectors $\boldsymbol{y}_j$, $j = 1, \ldots, p$ are displayed as the warp yarns in a Textile plot. However, there is an extra warp yarn so called *ID warp yarn* on the leftmost of the Textile plot. This is for identification of each record or observation, and the ID labels are placed according to the coordinate vector

$$\boldsymbol{y}_0 = \frac{1}{\lambda}(\boldsymbol{m} - \bar{m}\mathbf{1}) + \bar{m}\mathbf{1},$$

where $\bar{m} = \mathbf{1}^T \boldsymbol{m}/n$, because the coordinate vector of a categorical data vector with all different values is equal to the vector $\boldsymbol{m}$ except a multiple constant. It is shown in the following proposition.

Proposition 2 *The coordinate vector $\boldsymbol{y}_j$ of a categorical data vector with all different values is equal to the vector $\boldsymbol{m}$ except for a constant multiplication.*

PROOF. Assume that the first r data vectors are ordered categorical. Let us define a function

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \lambda) &= f(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda \left(\sum_{k=1}^{p} \|\boldsymbol{y}_k - \bar{y}_{\cdot k}\mathbf{1}\|^2_{\boldsymbol{w}_k} - N \right) + \sum_{k=1}^{r} \boldsymbol{\mu}(\mathscr{I}_k)^T \boldsymbol{\beta}(\mathscr{I}_k) \\ &= \sum_{k=1}^{p} \|\bar{y}_{\cdot k}\mathbf{1}\|^2_{\boldsymbol{w}_k} - \|\boldsymbol{m}\|^2_{\boldsymbol{w}} + \lambda \left(\sum_{k=1}^{p} \|\boldsymbol{y}_k - \bar{y}_{\cdot k}\mathbf{1}\|^2_{\boldsymbol{w}_k} - N \right) \\ &\quad + \sum_{k=1}^{r} \boldsymbol{\mu}(\mathscr{I}_k)^T \boldsymbol{\beta}(\mathscr{I}_k) \end{aligned}$$

with Lagrange multiplier λ and $\boldsymbol{\mu}$. If $\boldsymbol{x}_j$ is the unordered categorical data vector with n levels, we see from $\partial L/\partial \alpha_j = 0$ that

$$\mathbf{1}^T \boldsymbol{y}_j = \mathbf{1}^T \boldsymbol{m}$$

and from $\partial L/\partial \boldsymbol{\beta}(\mathscr{I}_j) = \mathbf{0}$ that

$$\lambda \mathbf{X}_j^T \boldsymbol{y}_j = \mathbf{X}_j^T \boldsymbol{m}$$

for a $n \times (n-1)$ encoded matrix $\mathbf{X}_j$ such that

$$\mathbf{1}^T \mathbf{X}_j = \mathbf{0}, \quad \mathbf{X}_j^T \mathbf{X}_j = \mathbf{I}.$$

Note here that

$$\mathbf{X}_j \mathbf{X}_j^T = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T,$$

we have

$$\lambda(\boldsymbol{y}_j - \bar{m}\mathbf{1}) = \boldsymbol{m} - \bar{m}\mathbf{1}. \tag{18}$$

Therefore the resulting coordinate vector $\boldsymbol{y}_j$ should satisfy (18) for $\lambda = \hat{\lambda}$ which attains the minimum. $\square$

Therefore, it is not necessary to include an ID data vector for the computation of coordinate vector even if it exists, but the ID data vector is used for labelling the points on the ID warp yarn. The sequence number of observations or records will be used as the labels if no ID data vector is given.

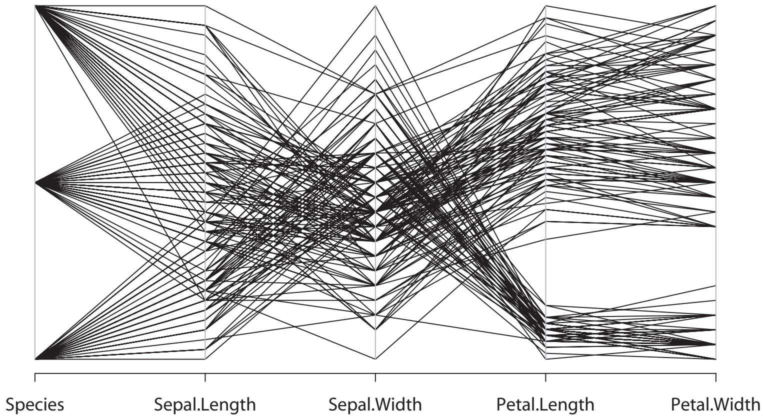

Fig. 3. Parallel Coordinate Plot of Iris Data

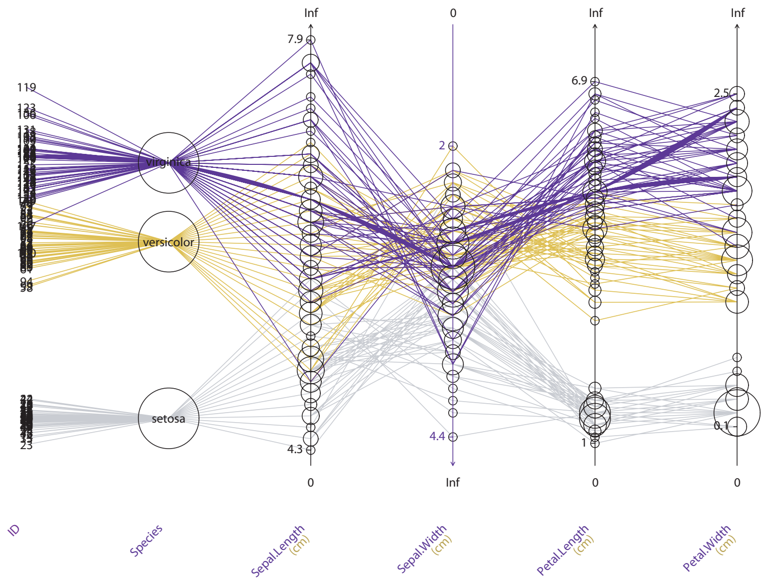

Fig. 4. Simple Minded Textile Plot of Iris Data

Figure 2 is an example of warp yarns for the coordinate vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_5$ of famous Iris data. The user can at once understand there are one unordered categorical data vector and four continuous measurement in this data set. The different size of circles on continuous warp yarns indicate that there are many duplicated values. This is simple because the precision is up to 10^{-1}. It is also interesting to note that the direction of the axis for Sepal width is upside down.

3.3 Weft Interlace

As same as in parallel coordinate plot, the coordinates on warp yarns are connected in order to identify data points. Figure 3 is a parallel coordinate plot of the Iris data and Figure 4 is a simple minded Textile plot. The effect of

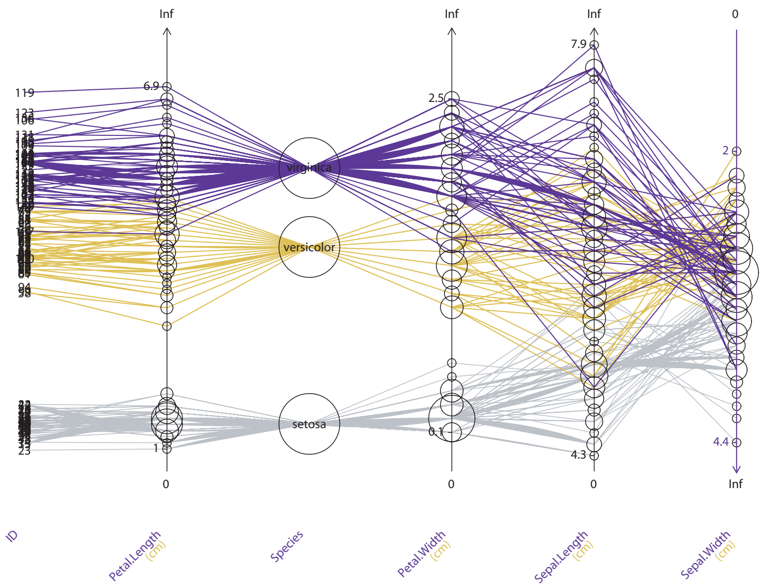

Fig. 5. Textile Plot of Iris Data

choosing proper location and scale for each data vector is significant in Species and Sepal width. Properly placed labels of Species suggests that Iris setosa is isolated from other species in the Textile plot. The upside down direction of the axis for Sepal width makes the connection look simpler. However, the order of warp yarns is carefully chosen in Textile plot. The Textile plot becomes more persuasive if an appropriate order is selected.

There would be various criteria to make an order of the warp yarns. One of such criteria is based on the distance $\|\boldsymbol{y}_j - \boldsymbol{m}\|$, $j = 1, \ldots, p$ from the mean vector $\boldsymbol{m}$. The identification of data points becomes easier by this criterion. The coordinate vector for left hand side yarn of a yarn is closer to $\boldsymbol{m}$. However, the ID warp yarn is always placed on the leftmost.

The coordinates of a data points are connected by segments but there are some cases when no coordinate exists on a warp yarn due to the existence of missing values. The weft yarn is then disconnected on the warp yarn. It is possible to make a more detailed design of the segments, for example, coloured, varying width, varying thickness and so forth. However, we leave such a design problem for further investigation.

Figure 5 is now a Textile plot of the Iris data. The weft yarns are identified by colours to the three species. The difference among species is clear in the plot and the most important data vectors Petal width and length are placed on the left and the right of the categorical vector Species, and those data vectors are placed nearest to the ID warp yarn. Such an observation is well known in discriminant analysis but easily seen from Textile plot without any complicated calculation. However, advantage of data visualisation by Textile plot is not only on a classification of data but also whole understanding of the

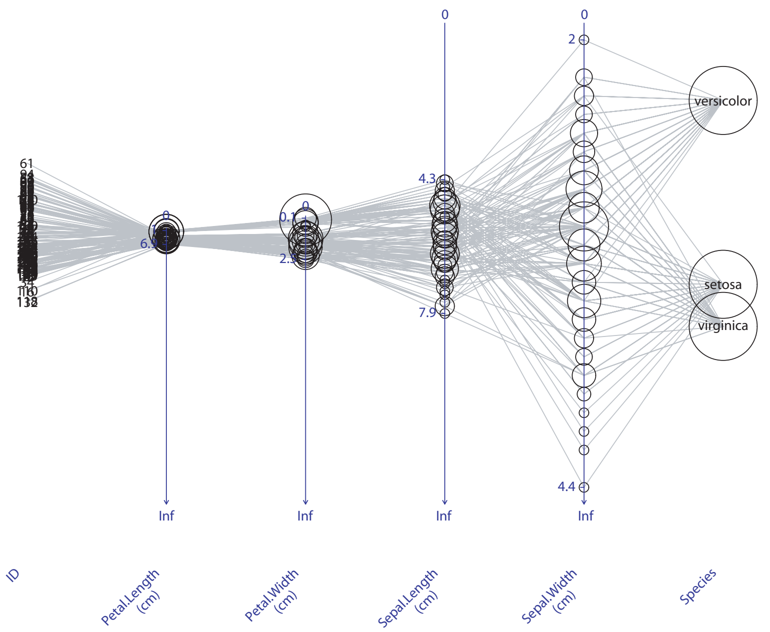

Fig. 6. The Second Piece of Textile Plot in a Suit

given data. For example, we can see marginal distribution of each data vector, correlations and so forth.

3.4 Several Textile Plots

There are two cases where several Textile plots in a suit. The first case is when the maximum eigenvalue in Theorem 3 has multiplicity. Then we will find several orthogonal eigenvectors for the maximum eigenvalue and produce several Textile plots for every eigenvectors.

Another case is when we want to see the results of other types of horizontal alignment. Then we can produce different Textile plots for a given data by choosing different eigenvectors in Theorem 3. This is equivalent to choose different type of location and scales. In other words, it is to choose different mean vector $\boldsymbol{m}$ orthogonal to that for the largest eigenvalue.

Figure 6 is an example of such a Textile plot for Iris data, where the eigenvector for the second largest eigenvalue has been chosen. The left most warp yarn indicates again the ID warp yarn with sequence numbers. Now, we can see the role of the data vector Sepal width which was not so clear in Figure 5.

4 Properties of Textile Plot

Two outstanding features can be found in Textile plot. One of them is an existence of *Knot* on a warp yarn. Another is *Neat weft* which is required for a pretty textile. In this section, we will give a condition for a knot produced on a warp yarn or a condition for neat weft produced between two warp yarns.

To simplify the problem, let us assume that there is no missing value and no ordered categorical data vector in the given data. Under these assumptions, we can assume without loss of generality that the data matrices $\mathbf{X}_j$, $j = 1, \ldots, p$ are normalised so that

$$\mathbf{1}^T \mathbf{X}_j = \mathbf{0} \quad \text{and} \quad \mathbf{X}_j^T \mathbf{X}_j = \mathbf{I}, \quad j = 1, \ldots, p. \tag{19}$$

Note that the textile plot is invariant under change of location and scales of the original data vectors or the choice of contrasts mentioned in Section 3. By using such normalised data matrices, we define a $n \times q$ data matrix

$$\mathbf{X}_{-j} = (\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_p),$$

where $q = Q - (q_j - 1)$.

We also assume that $\alpha_0 = 0$ in Corollary 2 since the choice of α_0 does not change the appearance of Textile plot. This implies that the mean vector $\boldsymbol{m}$ is orthogonal to $\mathbf{1}$.

4.1 Knot on a Warp Yarn

The knot is a point on a warp yarn, which all weft yarns go through. Mathematically saying, a knot is produced on the jth warp yarn when the scale parameter selected is zero, that is, $\hat{\boldsymbol{\beta}}(\mathscr{I}_j) = \mathbf{0}$. Before giving a necessary and sufficient condition for $\hat{\boldsymbol{\beta}}(\mathscr{I}_j) = \mathbf{0}$ in Theorem 4, we need the following lemma for the proof of the theorem. Consider a $Q \times Q$ symmetric matrix $\mathbf{C}$, and partition $\mathbf{C}$ into

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix}, \tag{20}$$

where $\mathbf{C}_{22}$ is a $q \times q$ sub matrix for a $q < Q$. We also denote the eigenvalues of $\mathbf{C}_{22}$ as $\lambda_1, \ldots, \lambda_q$ in a descending order, and the corresponding eigenvectors as $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_q$. Then we have the lemma.

Lemma 2 *Assume that the largest eigne value λ_1 of $\mathbf{C}_{22}$ has no multiplicity. Let $\hat{\boldsymbol{\gamma}}$ be the $\boldsymbol{\gamma}$ which maximises $\boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma}$ under the constraint $\|\boldsymbol{\gamma}\| = 1$. A* *necessary and sufficient condition for the first $Q-q$ elements of $\hat{\boldsymbol{\gamma}}$ to be 0 is given by*

$$\mathbf{C}_{12}\boldsymbol{p}_1 = \mathbf{0} \tag{21}$$

and

$$\mathbf{C}_{12}(\lambda_1\mathbf{I} - \mathbf{C}_{22})^{+}\mathbf{C}_{12}^T < \lambda_1\mathbf{I} - \mathbf{C}_{11} \tag{22}$$

in a sense of positive definiteness.

PROOF. Let us partition the vector $\boldsymbol{\gamma}$ to

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix}$$

in parallel with the partition of $\mathbf{C}$. If $\hat{\boldsymbol{\gamma}}$ is partitioned in a similar way, the $\hat{\boldsymbol{\gamma}}_1$ is the vector of the first $Q-q$ elements of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}}_1 = \mathbf{0}$ is equivalent to

$$\boldsymbol{\gamma}^T\mathbf{C}\boldsymbol{\gamma} < \hat{\boldsymbol{\gamma}}^T\mathbf{C}\hat{\boldsymbol{\gamma}} = \lambda_1 \tag{23}$$

for any $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\gamma}\| = 1$ other than $\hat{\boldsymbol{\gamma}}$.

The condition (23) can be rewritten as

$$f(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) < \lambda_1 \tag{24}$$

for any $0 < \varepsilon \leq 1$ and any $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ such that $\|\boldsymbol{\gamma}_1\|^2 = \varepsilon$ and $\|\boldsymbol{\gamma}_2\|^2 = 1-\varepsilon$, where

$$f(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = \boldsymbol{\gamma}_1^T\mathbf{C}_{11}\boldsymbol{\gamma}_1 + 2\boldsymbol{\gamma}_1^T\mathbf{C}_{12}\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2^T\mathbf{C}_{22}\boldsymbol{\gamma}_2.$$

For a fixed $\boldsymbol{\gamma}_1$ and ε, the maximum of $f(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ with respect to $\boldsymbol{\gamma}_2$ under the constraint $\|\boldsymbol{\gamma}_2\|^2 = 1-\varepsilon$ is attained by $\boldsymbol{\gamma}_2^*$ such that

$$(\lambda\mathbf{I} - \mathbf{C}_{22})\boldsymbol{\gamma}_2^* = \mathbf{C}_{12}^T\boldsymbol{\gamma}_1, \tag{25}$$

where λ is a Lagrange multiplier. By using the Moor-Penrose inverse of $\lambda\mathbf{I} - \mathbf{C}_{22}$, a solution of (25) is given by

$$\boldsymbol{\gamma}_2^* = (\lambda\mathbf{I} - \mathbf{C}_{22})^{+}\mathbf{C}_{12}^T\boldsymbol{\gamma}_1. \tag{26}$$

The Lagrange multiplier λ is chosen so that $\|\boldsymbol{\gamma}_2\|^2 = 1-\varepsilon$. We will show that we can always find a $\lambda > \lambda_2$ for any $0 < \varepsilon \leq 1$. We see that

$$(\lambda\mathbf{I} - \mathbf{C}_{22})^{+} = \begin{cases} \displaystyle\sum_{i=2}^{q} \frac{\boldsymbol{p}_i\boldsymbol{p}_i^T}{\lambda_1 - \lambda_i} & \lambda = \lambda_1, \\ \displaystyle\sum_{i=1}^{q} \frac{\boldsymbol{p}_i\boldsymbol{p}_i^T}{\lambda - \lambda_i} & \lambda \neq \lambda_1, \end{cases}$$

and $\boldsymbol{p}_1 = \hat{\boldsymbol{\gamma}}_2$ since $\hat{\boldsymbol{\gamma}}_2$ is the eigenvector of $\mathbf{C}_{22}$ for the largest eigenvalue λ_1, and $\mathbf{C}_{12}\hat{\boldsymbol{\gamma}}_2 = \lambda_1 \mathbf{C}_{12}\boldsymbol{p}_1 = \mathbf{0}$. Then

$$\|(\lambda \mathbf{I} - \mathbf{C}_{22})^{+}\mathbf{C}_{12}^T\boldsymbol{\gamma}_1\|^2 = \sum_{i=2}^{q} \frac{\boldsymbol{\gamma}_1^T\mathbf{C}_{12}\boldsymbol{p}_i\boldsymbol{p}_i^T\mathbf{C}_{12}^T\boldsymbol{\gamma}_1}{(\lambda-\lambda_i)^2}$$

for any $\lambda > \lambda_2$. Here we have already shown that $\hat{\boldsymbol{\gamma}}_1 = \mathbf{0}$ implies (21). Therefore it is enough to show (23) is equivalent to (22).

By normalising $\boldsymbol{\gamma}_1$ as $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_1/\|\boldsymbol{\gamma}_1\|$, we can rewrite $\|\boldsymbol{\gamma}_2^*\|^2 = 1-\varepsilon$ as

$$\frac{1}{\varepsilon} = 1 + \sum_{i=2}^{q} \frac{\tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{12}\boldsymbol{p}_i\boldsymbol{p}_i^T\mathbf{C}_{12}^T\tilde{\boldsymbol{\gamma}}_1}{(\lambda-\lambda_i)^2} \tag{27}$$

for $\lambda > \lambda_2$. The right hand side of (27) is now independent of ε and a monotone decreasing function of λ from ∞ to 1 as λ moves from λ_2 to ∞. This leads to that we can find a λ for any given $0 < \varepsilon \le 1$. Here, we employed a convention that $\lambda = \infty$, that is, $\boldsymbol{\gamma}_2^* = \mathbf{0}$ if $\varepsilon = 1$. Now,

$$\begin{aligned} f(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2^*) = {} & \boldsymbol{\gamma}_1^T\mathbf{C}_{11}\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_1^T\mathbf{C}_{12}(\lambda\mathbf{I}-\mathbf{C}_{22})^{+} \\ & \times(2\lambda\mathbf{I}-\mathbf{C}_{22})(\lambda\mathbf{I}-\mathbf{C}_{22})^{+}\mathbf{C}_{12}^T\boldsymbol{\gamma}_1 < \lambda_1 \end{aligned}$$

is equivalent to

$$\tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{11}\tilde{\boldsymbol{\gamma}}_1 < \frac{\lambda_1}{\varepsilon} - \tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{12}(\lambda\mathbf{I}-\mathbf{C}_{22})^{+}(2\lambda\mathbf{I}-\mathbf{C}_{22})(\lambda\mathbf{I}-\mathbf{C}_{22})^{+}\mathbf{C}_{12}^T\tilde{\boldsymbol{\gamma}}_1.$$

Substituting $1/\varepsilon$ by the right hand side of (27), we further rewrite the inequality above as

$$\tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{11}\tilde{\boldsymbol{\gamma}}_1 < \sum_{i=2}^{q} \frac{(\lambda_1+\lambda_i-2\lambda)\tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{12}\boldsymbol{p}_i\boldsymbol{p}_i^T\mathbf{C}_{12}^T\tilde{\boldsymbol{\gamma}}_1}{(\lambda-\lambda_i)^2} + \lambda_1. \tag{28}$$

We see now that (28) is equivalent to (23) for any $\tilde{\boldsymbol{\gamma}}_1$ with $\|\tilde{\boldsymbol{\gamma}}_1\| = 1$ and $\lambda > \lambda_2$. Let us evaluate the lower bound of the right hand side of the inequality (28). The minimum of the right hand side of (28) for $\lambda > \lambda_2$ is attained at $\lambda = \lambda_1$ since the gradient with respect to λ is

$$2(\lambda-\lambda_1)\sum_{i=2}^{q} \frac{\tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{12}\boldsymbol{p}_i\boldsymbol{p}_i^T\mathbf{C}_{12}^T\tilde{\boldsymbol{\gamma}}_1}{(\lambda-\lambda_i)^3}.$$

Therefore (23) is equivalent to the condition that

$$\tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{11}\tilde{\boldsymbol{\gamma}}_1 < -\sum_{i=2}^{q} \frac{\tilde{\boldsymbol{\gamma}}_1^T\mathbf{C}_{12}\boldsymbol{p}_i\boldsymbol{p}_i^T\mathbf{C}_{12}^T\tilde{\boldsymbol{\gamma}}_1}{(\lambda_1-\lambda_i)} + \lambda_1 \tag{29}$$

for any $\tilde{\gamma}_1$ with $\|\tilde{\gamma}_1\| = 1$. Note that the inequality (29) is equivalent to

$$\mathbf{C}_{12} \sum_{i=2}^{q} \frac{\boldsymbol{p}_i \boldsymbol{p}_i^T}{\lambda_1 - \lambda_i} \mathbf{C}_{12}^T < \mathbf{C}_{11} - \lambda_1 \mathbf{I}.$$

Then, it is clear that this is equivalent to (22) if we remember the definition of the Moor-Penrose inverse $(\lambda_1 \mathbf{I} - \mathbf{C}_{22})^+$. □

We shall use the singular value decomposition $\mathbf{UDV}^T$ of $\mathbf{X}_{-j}$, where the diagonal elements of $\mathbf{D} = \mathrm{diag}(d_j;\ j = 1, \ldots, q)$ are singular values arranged in the order that $d_1 > d_2 \geq \cdots \geq d_q \geq 0$, and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q)$ and $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)$ are column orthogonal matrices.

Assumption 3 *The multiplicity of the largest singular value d_1 of $\mathbf{X}_{-j}$ is 1.*

Assumption 3 implies that $d_1 > 1$, since

$$\mathrm{tr}(\mathbf{X}_{-j}^T \mathbf{X}_{-j}) = \sum_{i=1}^{q} d_i^2 = q.$$

The following theorem gives us necessary and sufficient condition for a knot produced on the jth warp yarn under Assumption 3.

Theorem 4 *Assume that there is no missing value in $\mathbf{X}$ and no ordered categorical data in the given data. Under* Assumption 3, *a necessary and sufficient condition for a knot produced on the jth warp yarn is that*

$$\mathbf{X}_j^T \boldsymbol{u}_1 = \mathbf{0} \tag{30}$$

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

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

PROOF. Note that $\hat{\boldsymbol{\beta}}_j = \mathbf{0}$ is equivalent to the fact that the first $k_j - 1$ elements are 0 of the eigenvector of

$$\mathbf{C} = \begin{pmatrix} \mathbf{X}_j^T \mathbf{X}_j & \mathbf{X}_j^T \mathbf{X}_{-j} \\ \mathbf{X}_{-j}^T \mathbf{X}_j & \mathbf{X}_{-j}^T \mathbf{X}_{-j} \end{pmatrix}$$

for the largest eigenvalue. It follows from Theorem 2 by putting $\mathbf{A} = \mathbf{C}/p$ and $\mathbf{B} = \mathbf{I}$. Then, the necessary and sufficient condition is given by (21) and (22) in Lemma 2. In (21),

$$\begin{aligned}\mathbf{C}_{12}\boldsymbol{v}_1 = \mathbf{X}_j^T\mathbf{X}_{-j}\boldsymbol{v}_1 &= \mathbf{X}_j^T\mathbf{U}\mathbf{D}\mathbf{V}^T\boldsymbol{v}_1\\ &= d_1\mathbf{X}_j^T\boldsymbol{u}_1\end{aligned}$$

which is equivalent to (30). In (22), the left hand side of the inequality is rewritten as

$$\begin{aligned}\mathbf{C}_{12}(\lambda_1\mathbf{I}-\mathbf{C}_{22})^+\mathbf{C}_{12}^T &= \mathbf{X}_j^T\mathbf{X}_{-j}(d_1^2\mathbf{I}-\mathbf{X}_{-j}^T\mathbf{X}_{-j})\mathbf{X}_{-j}^T\mathbf{X}_j\\ &= \mathbf{X}_j^T\mathbf{U}\mathbf{D}\mathbf{V}^T(d_1^2\mathbf{I}-\mathbf{V}\mathbf{D}^2\mathbf{V}^T)\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{X}_j\\ &= \mathbf{X}_j^T\mathbf{U}\mathbf{D}(d_1^2\mathbf{I}-\mathbf{D}^2)^+\mathbf{D}\mathbf{U}^T\mathbf{X}_j\\ &= \mathbf{X}_j^T\mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T\mathbf{X}_j,\end{aligned}$$

and the right hand side of the inequality is rewritten as

$$\lambda_1\mathbf{I}-\mathbf{C}_{11} = d_1^2\mathbf{I}-\mathbf{X}_j^T\mathbf{X}_j = (d_1^2-1)\mathbf{I}.$$

We have now the desired result. □

Note that $\boldsymbol{u}_1$ is proportional to the mean vector $\boldsymbol{m}_{-j}$ to which all coordinate vectors on the textile plot of $\mathbf{X}_{-j}$ are aligned. Therefore, the condition (30) says that any column vector of $\mathbf{X}_j$ is orthogonal to $\boldsymbol{m}_{-j}$. However, such an orthogonality is not enough for a knot produced. The projected size of $\mathbf{X}_j$ on the range space of $\mathbf{X}_{-j}$ should be small enough relative to the size of $\mathbf{X}_{-j}$. The following corollary gives us a simplified sufficient condition for a knot produced on a warp yarn.

Corollary 3 *Under the same assumption as in* Theorem 4*, a sufficient condition for a knot produced on the* j*th warp is that*

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

and all eigenvalues of $\mathbf{X}_j^T\mathbf{U}\mathbf{U}^T\mathbf{X}_j$ *are less than* $(d_1^2-d_2^2)(d_1^2-1)/d_2^2$.

PROOF. It is enough to note that

$$\boldsymbol{z}^T(\mathbf{X}_j^T\mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T\mathbf{X}_j)\boldsymbol{z} \le \frac{d_2^2}{d_1^2-d_2^2}\boldsymbol{z}^T(\mathbf{X}_j^T\mathbf{U}\mathbf{U}^T\mathbf{X}_j)\boldsymbol{z}$$

for any (k_j-1)-dimensional vector $\boldsymbol{z}$. □

The sufficient condition in Corollary 3 becomes simpler if the original data vector $\boldsymbol{x}_j$ for the jth warp is numerical. Then $\mathbf{X}_j$ in Corollary 3 is a vector and $\mathbf{X}_j^T\mathbf{U}\mathbf{U}^T\mathbf{X}_j$ is a scalar value. Therefore, it is easy to check $\mathbf{X}_j^T\boldsymbol{u}_1 = 0$ and $\mathbf{X}_j^T\mathbf{U}\mathbf{U}^T\mathbf{X}_j < (d_1^2-d_2^2)(d_1^2-1)/d_2^2$.

Example 3 *If $\mathbf{X}_j^T \boldsymbol{u}_1 = \mathbf{0}$, a sufficient condition for a knot produced on the jth warp is that $d_1^2 > d_2^2 + 1$.*

The condition $d_1^2 > d_2^2 + 1$ for the largest and the second largest singular values d_1 and d_2 requires that column vectors of $\mathbf{X}_{-j}$ are not so much distinct each other.

If we assume a stronger assumption that $\mathbf{X}_j^T \mathbf{X}_{-j} = \mathbf{O}$, then $\mathbf{X}_j^T \mathbf{U} \boldsymbol{\Delta} \mathbf{U}^T \mathbf{X}_j = \mathbf{O}$. Therefore we have the following example from Corollary 4.

Example 4 *If $\mathbf{X}_j^T \mathbf{X}_{-j} = \mathbf{O}$, a knot is always produced on the jth warp.*

4.2 Neat Weft

Another outstanding feature of a textile plot is *neat weft*, which means that all weft yarns are horizontally aligned between two warp yarns. Neat weft between the jth and the $j+1$th warp yarns is when the coordinate vectors $\boldsymbol{y}_j$ and $\boldsymbol{y}_{j+1}$ are identical.

Lemma 3 *Assume that there is no missing value in the given data and no ordered categorical data is involved. A necessary and sufficient condition for $\boldsymbol{y}_j = \boldsymbol{y}_k$ is given by*

$$\mathrm{Proj}_{\mathbf{X}_j}(\boldsymbol{m}) = \mathrm{Proj}_{\mathbf{X}_k}(\boldsymbol{m}), \tag{31}$$

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

PROOF. Note that

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

where $\hat{\alpha}_j = 0$ provided that $\alpha_0 = 0$ and $\mathbf{1}^T \mathbf{X}_j = \mathbf{0}$. Therefore $\boldsymbol{y}_j = \boldsymbol{y}_k$ is equivalent to

$$\mathbf{X}_j \boldsymbol{\beta}(\mathscr{I}_j) = \mathbf{X}_k \boldsymbol{\beta}(\mathscr{I}_k).$$

If we remember that $\lambda \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}$ for the maximum eigenvalue λ of $\mathbf{X}^T \mathbf{X}$, the necessary and sufficient condition is

$$(\mathbf{X}_j \mathbf{X}_j^T) \mathbf{X} \boldsymbol{\beta} = (\mathbf{X}_k \mathbf{X}_k^T) \mathbf{X} \boldsymbol{\beta}.$$

It is enough to note that $\boldsymbol{m} = \mathbf{X} \hat{\boldsymbol{\beta}} / p$. $\square$

Theorem 5 *Assume that there is no knot on an adjacent pair of warps. If the data vectors for the pair of warps are both numerical, then a necessary and sufficient condition for a neat weft between the pair of warps is that the* *data vectors are identical except difference of locations and scales. If the data vectors are unordered categorical, then a sufficient condition is that the data vectors are identical except difference of the labels of levels.*

PROOF. The former part of the theorem follows from Lemma 3, if we note the normalisation assumption given in (19). The latter part follows from Lemma 1. From the condition, there exists an nonsingular matrix $\mathbf{D}$ such that

$$\mathbf{X}_j = \mathbf{X}_{j+1}\mathbf{D},$$

since $\mathbf{X}_j$ and $\mathbf{X}_{j+1}$ are both normalised as in (19). The matrix $\mathbf{D}$ is then an orthogonal matrix, since

$$\mathbf{I} = \mathbf{X}_j^T\mathbf{X}_j = \mathbf{D}^T\mathbf{X}_{j+1}^T\mathbf{X}_{j+1}\mathbf{D} = \mathbf{D}^T\mathbf{D}.$$

Therefor we have

$$\mathbf{X}_j\mathbf{X}_j^T = \mathbf{X}_{j+1}\mathbf{D}\mathbf{D}^T\mathbf{X}_{j+1}^T = \mathbf{X}_{j+1}\mathbf{X}_{j+1}^T$$

and the desired result follows from Lemma 3. □

5 Textile Plot and Optimised Parallel Coordinate Plot

The optimised parallel coordinate plot is proposed by Michailidis and Leeuw [6] in the context of homogeneity analysis. In homogeneity analysis, the main concern is to find a proper quantification of given categorical data vectors $\boldsymbol{x}_j$, $j = 1, \ldots, p$. The quantified vectors in case of no missing values are defined as $\boldsymbol{y}_j = \mathbf{Z}_j\boldsymbol{\varphi}_j$, $j = 1, \ldots, p$ so as to minimise

$$\sigma_2(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p, \boldsymbol{\xi}) = \frac{1}{p}\sum_{j=1}^{p} \|\boldsymbol{\xi} - \boldsymbol{y}_j\|^2$$

under the constraint

$$\mathrm{Var}(\boldsymbol{\xi}) = \frac{1}{n}\|\boldsymbol{\xi} - \bar{\xi}\mathbf{1}\|^2 = 1,$$

where $\bar{\xi} = \sum_{i=1}^{n} \xi_i/n$. Here $\mathbf{Z}_j$, $j = 1, \ldots, p$ are indicator matrices for the data vector $\boldsymbol{x}_j$, $j = 1, \ldots, p$. The ith row of $\mathbf{Z}_j$ is an indicator vector for the ith element of $\boldsymbol{x}_j$, in which the element corresponding to the level of the ith element of $\boldsymbol{x}_j$ is 1 and 0 otherwise. The optimised parallel coordinate plot is a parallel coordinate plot of $\boldsymbol{y}_j$, $j = 1, \ldots, p$ where all axes share the same coordinate system.

We can see similarities between textile plot and optimised parallel coordinate plot. In fact, the criterion $\sigma_2(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p, \boldsymbol{\xi})$ is equivalent to $S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ in (1).

This is because a space spanned by a quantified vector $\boldsymbol{y}_j$ is equivalent to that for the coordinate vector $\boldsymbol{y}_j$ in Textile plot, since the range space of $\mathbf{Z}_j$ is equal to that of the matrix $(\mathbf{1}, \mathbf{X}_j)$, where $\mathbf{X}_j$ is an encoded matrix with any contrasts matrix. Therefore the quantified vectors are considered to be the coordinate vectors $\boldsymbol{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}(\mathscr{I}_j)$, $j = 1, \ldots, p$ so as to minimise $S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{m})$ in (3) under the constraint $\mathrm{Var}(\boldsymbol{m}) = 1$.

It is clear that the problems yield the same solution $\hat{\boldsymbol{\alpha}}$ given in Corollary 2, since the constraints are both independent of $\boldsymbol{\alpha}$. By introducing the means $\bar{y}_{\cdot j} = \sum_{i=1}^n y_{ij}/n$ and $\bar{m} = \sum_{i=1}^n m_i/n$, the sum of squares S^2 can be written as

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

where $\mathbf{A}$ in (9) and $\mathbf{B}$ in (10). Then the solution $\hat{\boldsymbol{\beta}}$ under the constraint $\mathrm{Var}(\boldsymbol{m}) = 1$ is given by the $\boldsymbol{\beta}$ which minimises $S^2(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta})$ under the constraint $\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} = N$, since $\mathrm{Var}(\boldsymbol{m}) = \boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta}/N$.

On the other hand, the optimal choice of scales in Textile plot is given by $\boldsymbol{\beta}$ which minimises $S^2(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta})$ under the constraint $\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} = N$. The following proposition tells us that the solutions are the same.

Proposition 3 *The eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$ for the largest eigenvalue gives us the solution for the minimisation problem of $g(\boldsymbol{\beta}) = -\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta}$ irrespective of constraint $\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} = 1$ or $\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} = 1$ except a constant multiplication.*

PROOF. It is enough to note that, the solution $\hat{\boldsymbol{\beta}}$ for the minimization of $\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta}/\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta}$ is equivalent to that of $-\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta}/\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta}$ except for a constant multiplication. The solution $\hat{\boldsymbol{\beta}}$ immediately follows from the eigen vector of $\mathbf{A}$ with respect to $\mathbf{B}$ for the largest eigenvalue. $\square$

From Proposition 3, we see that both optimised parallel coordinate plot and Textile plot yield the same picture, provided that given data vectors are all categorical and no missing value exists. This is rather a coincidence because the problems are in a dual relation as is seen in the proof. Proposition 3 suggests another view of textile plot. The locations and scales on each axes of textile plot are chosen so that the coordinate vectors are close to the mean vector $\boldsymbol{m}$ normalised as $\|\boldsymbol{m} - \bar{m}\mathbf{1}\| = 1$. However, it is worthy of noting that the aims are different. The quantification itself is the aim of homogeneity analysis, but data visualisation and data browsing is the aim of textile plot.

6 Concluding Remarks

We have proposed a new high dimensional data visualisation technique called Textile plot with a hope that this will be a key to understand phenomena hidden behind the data. Textile plot is a generalisation of the Parallel coordinate plot and accepts any kind of data, although the Parallel coordinate plot accepts only numerics.

Two outstanding features of Textile plot is knot and neat weft. A knot indicates that the data vector is isolated from other data vectors, and a neat weft indicates existence of a linear relation between two measurements or equivalence of two categorical data.

It is quite important to develop an efficient algorithm which can be applied for very high dimensional data or many ordered categorical data vectors. Also, introduction of dynamic display or interactive display mentioned in [8] or [9] is necessary to improve user interface. We leave such problems for further investigation.

References

[1] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods* (Academic Press, Inc., 1982).

[2] W.S. Cleveland, *The Elements of Graphing Data* (Bell Telephone Laboratories, Inc., Murray Hill, NJ, 1985).

[3] A. Gifi, *Nonlinear Multivariate Analysis* (John Wiley & Sons Ltd, 1990).

[4] R. Horn and C. Johnson, *Matrix Analysis* (Cambridge University Press, 1985).

[5] A. Inselberg, The plane with parallel coordinates, *The Visual Computer* **1** (1985) 69–91.

[6] G. Michailidis and J. de Leeuw, Data visualization through graph drawing. *Computational Statistics* **16** (2001) 435–450.

[7] C.R. Rao and S.K. Mitra, *Generalized Inverse of Matrices and its Applications* (John Wiley & Sons, Inc., 1971).

[8] M. Theus, Interactive data visualization using mondrian. *Journal of Statistical Software* **7** (2002).

[9] A. Unwin, C. Volinsky and S. Winkler, Parallel coordinates for exploratory modeling analysis, *Computational Statistics & Data Analysis* **43** (2003) 553–564.

[10] E. Wegman, Hyperdimensional data analysis using parallel coordinates. *Journal of The American Statistical Association* **85** (1990) 664–675.

[11] G.J. Wills, Selection: 524,288 ways to say this is interesting, *Proceedings of the 1996 IEEE Symposium on Information Visualization* (IEEE Computer Society Washington, DC, USA, 1996) 54–60.

Department of Mathematics
Faculty of Science and Technology
Keio University

Research Report

2005

[05/001] T. Kawazoe, *On the Littlewood-Paley g function and the Lusin area function on real rank 1 semisimple Lie groups*, KSTS/RR-05/001, March 28, 2005

[05/002] A. Sako, T. Suzuki, *Partition functions of supersymmetric gauge theories in noncommutative* $\mathbb{R}^{2D}$ *and their unified perspective,* KSTS/RR-05/002, April 25, 2005

[05/003] K. Aoyama, K. Shimizu, *A generalization of the inverse trinomial,* KSTS/RR-05/003, June 23, 2005

[05/004] T. Tanaka, *Algebraic independence of modified reciprocal sums of products of Fibonacci numbers,* KSTS/RR-05/004, July 27, 2005

[05/005] Y. Aoki, T. Kato, *New method of ground reflection coefficient reconstruction and shape recovery by another SAR signal modeling,* KSTS/RR-05/005, October 12, 2005

[05/006] A. Sako, T. Suzuki, *Dimensional reduction of Seiberg-Witten monopole equations, N=2 noncommutative supersymmetric field theories and Young diagrams,* KSTS/RR-05/006, November 28, 2005

2006

[06/001] N. Kumasaka, R. Shibata, *High dimensional data visualisation: Textile plot*, KSTS/RR-06/001, February 13, 2006