High-dimensional Data Visualisation: the Textile Plot

Natsuhiko Kumasaka, Ritei Shibata

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

Abstract

The textile plot is a parallel coordinate plot in which the ordering, locations and scales of the axes are simultaneously chosen so that the connecting lines, each of which represents a case, are aligned as horizontally as possible. Plots of this type can accommodate numerical data as well as ordered or unordered categorical data, or a mixture of these different data types. Knots and parallel wefts are features of the textile plot which greatly aid the interpretation of the data. Several practical examples are presented which illustrate the potential usefulness of the textile plot as an aid to the interpretation of multivariate data.

1 Introduction

Parallel coordinate plots have become a routine device with which to explore high dimensional data. This type of plot was originally proposed by Inselberg [10] as a tool for visualising high dimensional geometries using a two-dimensional display. Wegman [20] developed it as a tool for visualising high dimensional data. The basic idea of the parallel coordinate plot is to place axes, representing each observed variable or attribute, in parallel in a two dimensional display. For a given data point observed in a high dimensional space, its associated coordinates on adjacent axes are then connected by straight lines. Thus, each case is represented in the display by a trajectory made up of a series of connected straight lines. The parallel coordinate plot is one possible way of visualising high-dimensional data.

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

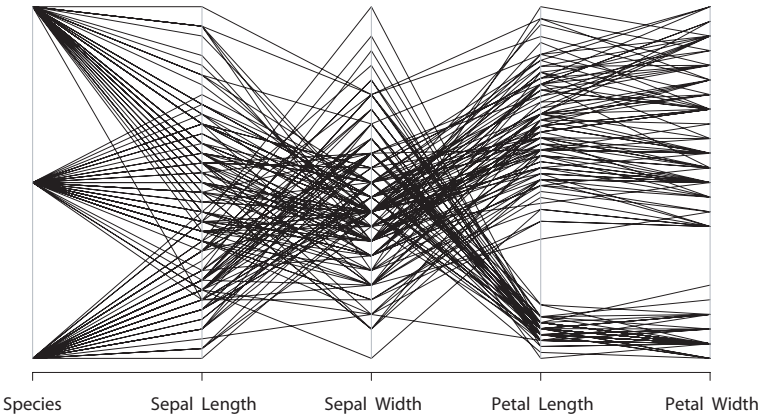

Fig. 1. Parallel coordinate plot for the iris data.

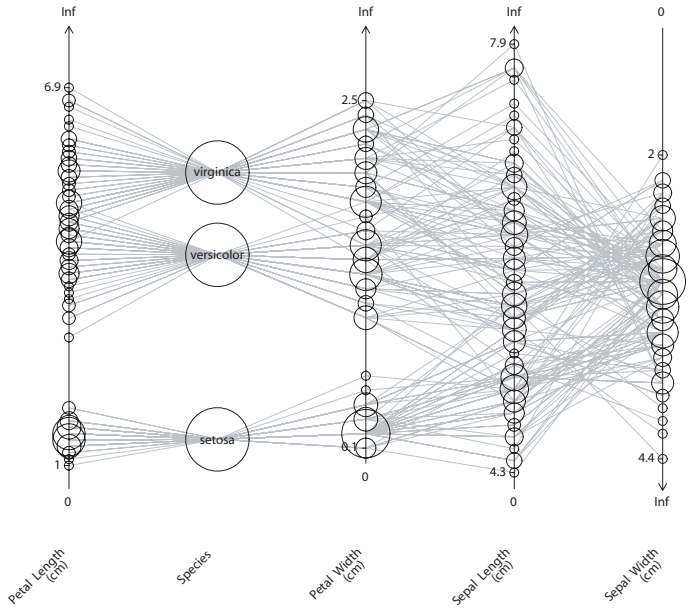

Fig. 2. Textile plot for the iris data.

Figure 1 shows an example of a parallel coordinate plot for Fisher's famous iris data [1]. Although this data set involves only relatively few dimensions, we use it as an initial example because of its simplicity and familiarity to statisticians. We have simply assigned the numbers 1 to 3 to the different species arranged in alphabetical order, namely Setosa, Versicolor and Virginica, since no definitive specification of how this should be done is given in the definition of the parallel coordinate plot. This lack of specification within the definition of the parallel coordinate plot tends to lead to a less than optimal display of categorical data. Moreover, we consider that displaying just the name of each attribute as the only written information makes the plot far too terse in the sense that it does not include important information that can be helpful to the user in their interpretation of the data.

The textile plot has been designed with problems such as these in mind. And

we would contend that the textile plot in Figure 2 provides an improved graphical representation of the data when compared with the parallel coordinate plot in Figure 1. This is because the ordering (from left to right) of the axes, and their (vertical) locations and their scales, are simultaneously chosen so that the connecting lines are aligned as horizontally as possible. As a consequence of introducing an objective criterion for determining the ordering of the axes and their locations and scales, appropriate numerical values can be assigned to the different categories associated with categorical data. For example, in Figure 2, Versicolor is positioned closer to Virginica than to Setosa. Another consequence is that the *direction* of the scale used to order "high" and "low" values on an axis can vary between axes. Thus, for example, in Figure 2, the direction of the scale used on the axis for *Sepal Width* runs opposite to that of the others.

Additional information is also depicted within the textile plot. As commented previously, the coordinates associated with the data points are indicated on the axes. However, where there is a point that is repeated, a circle is included within the plot with an area that is proportional to the number of replicates associated with the point in question. For the iris data, almost all of the data values are replicated due to the fact that a precision of just one decimal place was used when recording the values of the four continuous variables measured. Overall, the textile plot is a better representation of the iris data than the corresponding parallel coordinate plot as it provides a clearer and more comprehensive representation of the data. For example, the well known fact that *Petal Length* and *Petal Width* are important indicators of *Species* is readily seen from the textile plot, whereas this fact is not so easily established from a consideration of the parallel coordinate plot.

The name "textile plot" was derived by analogy to the process of fabric production in which warp and weft yarns are woven. A fabric is considered to be a "good" one if its weft yarns run as horizontally as possible. Because of the use of the horizontalisation criterion, the textile plot not only makes it easier for the user to understand the relationships that might exist between adjacent axes (i.e. variables or attributes), but it also allows one to identify potential linear relationships or orthogonalities that might exist between data vectors. Of course, such features are heavily dependent upon a careful choice of ordering of the axes. Furthermore, ordered and unordered categorical data can be displayed on the plot as well as numerical data with missing values.

Because of its construction, the textile plot is related to the optimised parallel coordinate plot proposed by Michailidis and de Leeuw [12] developed in the field of homogeneity analysis [7] and used for displaying categorical data. We will discuss the relationship between these two plots in greater detail in Section 7. Displaying categorical data using a parallel coordinate plot is also discussed in Rosario et al. [16].

2 Determination of locations and scales

In this section, we consider the location and scale transformation of the data where the latter would have been preprocessed by applying a suitable non-linear transformation prior to the location and scale transformation.

First we will explain the criterion we use for determining the locations and scales of the axes. Let $\boldsymbol{x}_j$ denote the vector of n observations on attribute j ($j = 1, \ldots, p$). Then each row of the data matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ gives us a p-dimensional observation. If the data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are all numeric, then they are simply transformed into p coordinate vectors

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

where $\mathbf{1}$ is a vector of ones, which results in a textile plot with a common coordinate system. The vector $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})^T$ gives us the coordinates of the n observations on the jth axis. The degree to which each connecting line on the textile plot is horizontal can be measured by the sum of squared deviations from a horizontal line at level ξ_i, that is

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

for the ith line connecting the points at the levels $y_{i1}, \ldots, y_{ip}$. Then our criterion would be to choose α_j and β_j, $j = 1, \ldots, p$, so that

$$\sum_{i=1}^{n} \sum_{j=1}^{p} (y_{ij} - \xi_i)^2 = \sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{\xi}\|^2$$

is minimised. The vector $\boldsymbol{\xi}$ also has to be chosen to minimise the sum of squares since the levels ξ_i, $i = 1, \ldots, n$ are unknown a priori.

This approach is in contrast with the parallel coordinate plot where

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

since the locations and scales are chosen axis by axis so that the coordinate points fill up the range of each axis.

In the textile plot, a categorical data vector $\boldsymbol{x}_j$ is first encoded into a data matrix $\mathbf{X}_j$ by an appropriate set of contrasts [3] and then transformed into a numerical coordinate vector,

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}_j.$$

The location parameter α_j and the scale parameter vector $\boldsymbol{\beta}_j$ are chosen simultaneously, using the same criterion as before. The coordinates for the three categories of *Species* on the second axis in Figure 2 are determined in this way. It is worthy of note that the resulting coordinate vector $\boldsymbol{y}_j$ is independent of the choice of the set of contrasts.

We hereafter assume the following for simplicity of presentation.

Assumption 1 *None of the data vectors nor the cases contains just missing values.*

Assumption 2 *No data vector consists of just a single value.*

Assumption 3 *The number of variables is larger than or equal to the number of cases.*

These assumptions do not cause any practical problems because we can delete any data vectors or cases which violate Assumption 1 beforehand, and set $\beta_j = 0$ or $\boldsymbol{\beta}_j = \mathbf{0}$ for any data vectors that violate Assumption 2. Such a modification does not affect the choice of locations and scales for any other data vectors.

We first consider the case where all data vectors are numeric and generalise the results to other cases in subsequent subsections.

2.1 Numerical Data

In the textile plot, the sum of squared deviations is not properly defined if there are missing values in the data. To reflect the existence of a missing value, we introduce the weight vectors $\boldsymbol{w}_j$, $j = 1, \ldots, p$ whose elements of zero or one are used to indicate missing values in $\boldsymbol{x}_j$, $j = 1, \ldots, p$. That is, the ith element w_{ij} of $\boldsymbol{w}_j$ is 0 if the corresponding element x_{ij} of $\boldsymbol{x}_j$ is missing; otherwise w_{ij} is 1. Using the notation $\|\boldsymbol{x}\|^2_{\boldsymbol{v}} = \sum_{i=1}^n v_i x_i^2$ for the norm with a weighting vector $\boldsymbol{v}$, we can formally define the sum of squares

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

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ is the vector of location parameters and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of scale parameters. Then, the use of such a weighted norm implies that missing values do not contribute to the sum of squared deviations, but the missing information itself is retained for display on the textile plot.

By using the notation $\boldsymbol{x} \cdot \boldsymbol{v}$ and $\boldsymbol{x}/\boldsymbol{v}$ to denote an element-wise product and the division of the vectors $\boldsymbol{x}$ and $\boldsymbol{v}$, it is readily seen that the solution $\hat{\boldsymbol{\xi}} =$

$\boldsymbol{m} = \sum_{j=1}^{p} \boldsymbol{w}_j \cdot \boldsymbol{y}_j / \boldsymbol{w}$ for $\boldsymbol{\xi}$ minimises $S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ since

$$S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2_{\boldsymbol{w}_j} + \sum_{j=1}^{p} \|\boldsymbol{m} - \boldsymbol{\xi}\|^2_{\boldsymbol{w}_j}, \tag{3}$$

where $\boldsymbol{w} = \sum_{j=1}^{p} \boldsymbol{w}_j$. Throughout the paper we will refer to $\boldsymbol{m}$ as the *mean vector* since it is a vector of the mean positions for the connecting lines within the textile plot. We would point out to the reader that $\boldsymbol{m}$ is not the vector of the means of each coordinate vector.

We need a constraint on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ so as 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 constant. For example, that it equals the effective number of the points $N = \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij}$. Here $\bar{y}_{\cdot j} = \boldsymbol{w}_j^T \boldsymbol{y}_j / \mathbf{1}^T \boldsymbol{w}_j$ is the mean of the coordinate vector $\boldsymbol{y}_j$. This constraint is equivalent to $\sum_{ij}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = N$ when there are no missing values, and $\bar{y}_{\cdot\cdot} = \sum_{ij} y_{ij}/N$.

The decomposition

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

indicates that all that is required is 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

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

A solution to this constrained minimisation problem always exists since $f(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is bounded below by $-N$.

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

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

where

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

and

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

Constraint (4) can also be rewritten as

$$\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta} = N \tag{6}$$

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

Then, a solution $\hat{\boldsymbol{\alpha}}$ is a solution of the equation

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

provided that $\hat{\boldsymbol{\beta}}$ is a solution, since constraint (6) is only effective for the parameter vector $\boldsymbol{\beta}$. An explicit expression for $\hat{\boldsymbol{\alpha}}$ is

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

where $\mathbf{A}_{11}^+$ is the Moore-Penrose inverse [15] of $\mathbf{A}_{11}$ and $\boldsymbol{z}$ is an arbitrary p-dimensional vector.

The value of the function $f(\boldsymbol{\alpha},\boldsymbol{\beta})$ at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ becomes

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

since

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

Therefore, 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 $\mathbf{B}$ associated with the largest eigenvalue.

Proposition 1 *For given numerical data vectors $\boldsymbol{x}_j$, $j = 1,\ldots,p$, which satisfy* Assumptions 1 *and* 2*, a solution which minimises $S^2(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{m})$ under the constraint* (4) *is given by $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\alpha}} = \mathbf{A}_{11}^+\mathbf{A}_{12}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{A}_{11}^+\mathbf{A}_{11})\boldsymbol{z}$ for an arbitrary p-dimensional vector $\boldsymbol{z}$ and $\hat{\boldsymbol{\beta}}$ is that eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$ associated with the largest eigenvalue such that $\hat{\boldsymbol{\beta}}^T\mathbf{B}\hat{\boldsymbol{\beta}} = N$.*

Note that the solution referred to above is not necessarily unique. However, if $\mathrm{rank}(\mathbf{A}_{11}) = p - 1$, then the choice of $\boldsymbol{\alpha}$ is essentially unique and can be written as $\hat{\boldsymbol{\alpha}} = c\mathbf{1} + \mathbf{A}_{11}^+\mathbf{A}_{12}\hat{\boldsymbol{\beta}}$ for an arbitrary global constant c. This is because $\{\boldsymbol{z};\ \mathbf{A}_{11}\boldsymbol{z} = \mathbf{0}\} = \mathrm{span}\{\mathbf{1}\}$ if $\mathrm{rank}(\mathbf{A}_{11}) = p - 1$. Note here that $\mathbf{A}_{11}\mathbf{1} = \mathbf{0}$

always holds true. The choice of $\hat{\boldsymbol{\beta}}$ is unique as far as the eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$ associated with the largest eigenvalue is unique.

Proposition 1 becomes simpler if there are no missing values.

Corollary 1 *If there are no missing values in the data, then a solution 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 and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ is the eigenvector of the sample correlation matrix of $\boldsymbol{x}_j$ associated with the largest eigenvalue, satisfying $\|\boldsymbol{\gamma}\|^2 = N = np$.

The proof is given in Appendix A.

2.2 Numerical and Categorical Data

If $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is a categorical data vector with q categories, then the element of the coordinate vector $\boldsymbol{y}$ takes only q different values on an axis. By denoting such values as $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$, the coordinate vector can be written as

$$\boldsymbol{y} = \mathbf{Z}\boldsymbol{\gamma}, \tag{9}$$

where the (i, k)th element z_{ik} of an $n \times q$ indicator matrix $\mathbf{Z}$ is 1 if x_i is equal to the kth category; otherwise z_{ij} is 0. If an $n \times (q-1)$ matrix

$$\mathbf{X} = \mathbf{Z}\mathbf{C}$$

is defined by a $q \times (q-1)$ contrast matrix $\mathbf{C}$ such that $\text{rank}(\mathbf{C}) = q-1$ and the columns are all linearly independent of $\mathbf{1}$, it is easily seen that $\text{Range}(\mathbf{Z}) = \text{Range}\{\mathbf{Z}(\mathbf{1}, \mathbf{C})\} = \text{Range}\{(\mathbf{1}, \mathbf{X})\}$. Therefore (9) can be rewritten as

$$\boldsymbol{y} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta} \tag{10}$$

by replacing $\boldsymbol{\gamma}$ by the parameters α and $\boldsymbol{\beta}$. The discussion above implies that, for the case of a categorical data vector, it is enough to encode $\boldsymbol{x}$ to $\mathbf{X}$ through $\mathbf{Z}$ and then apply the same minimisation criterion as used for a numerical data vector. It is clear that the resulting coordinates $\hat{\boldsymbol{\gamma}}$ of the q categories are independent of the choice of the contrast matrix $\mathbf{C}$.

Example 1 *The data vector* Species $\boldsymbol{x} = (\text{Setosa}, \ldots, \text{Setosa}, \text{Versicolor}, \ldots, \text{Versicolor}, \text{Virginica}, \ldots, \text{Virginica})^T$ *in the iris data is categorical. The coor-*

dinate vector is represented as $\boldsymbol{y} = \alpha\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}$ *with*

$$\mathbf{X}_1 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix},$$

when the treatment contrast $\mathbf{C} = (\mathbf{0}, \mathbf{I})^T$ *is used, where* $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$*. Thus the coordinate vector is parametrised as* $\boldsymbol{y} = (\alpha, \ldots, \alpha, \alpha + \beta_1, \ldots, \alpha + \beta_1, \alpha + \beta_2, \ldots, \alpha + \beta_2)^T$*.*

To cover cases where both numerical and categorical data vectors exist, we consistently use the matrix notation $\mathbf{X}_j$ in place of the numerical data vector $\boldsymbol{x}_j$ by letting $q_j = 2$. Such matrices $\{\mathbf{X}_j,\ j = 1 \ldots, p\}$ are combined into an $n \times Q$ data matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ where $Q = \sum_{j=1}^{p}(q_j - 1)$. Then by using the notation $\boldsymbol{v}(\mathscr{K})$ or $\mathbf{M}(\mathscr{K}, \mathscr{L})$ for the sub-vector or the sub-matrix specified by index sets $\mathscr{K}$ and $\mathscr{L}$ [8], we can generally write the coordinate vector as

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

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. Here

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

is an index set corresponding to the sub-matrix $\mathbf{X}_j$ of $\mathbf{X}$, such that

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

We now have the following proposition. Here the matrix $\mathbf{A}_{11}$ is the same as before but the matrices $\mathbf{A}_{12}$, $\mathbf{A}_{22}$ and $\mathbf{B}$ are defined in a slightly extended way. Their explicit definitions can be found in Appendix B.

Proposition 2 *For the given numerical or categorical data vectors* $\boldsymbol{x}_j$*,* $j = 1, \ldots, p$*, which satisfy* Assumptions 1 *and* 2*, a solution which minimises* $S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{m})$ *under the constraint* (4) *is given by* $\hat{\boldsymbol{\alpha}}$ *and* $\hat{\boldsymbol{\beta}}$*, where* $\hat{\boldsymbol{\alpha}} = \mathbf{A}_{11}^{+}\mathbf{A}_{12}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{A}_{11}^{+}\mathbf{A}_{11})\boldsymbol{z}$ *for an arbitrary p-dimensional vector* $\boldsymbol{z}$*, and* $\hat{\boldsymbol{\beta}}$ *is the eigenvector of* $\mathbf{A}$ *with respect to* $\mathbf{B}$ *associated with the largest eigenvalue such that* $\hat{\boldsymbol{\beta}}^T\mathbf{B}\hat{\boldsymbol{\beta}} = N$*.*

Proposition 2 becomes simpler if no missing values exist. The matrix $\mathbf{A}$ becomes

$$\mathbf{A} = \frac{1}{p}\left(\mathbf{X}^T\mathbf{X} - \frac{1}{n}\mathbf{X}^T\mathbf{1}\mathbf{1}^T\mathbf{X}\right), \tag{11}$$

and the matrix $\mathbf{B}$ becomes

$$\mathbf{B}(\mathscr{I}_j, \mathscr{I}_k) = \begin{cases} \mathbf{O} & j \neq k, \\ \mathbf{X}_j^T \mathbf{X}_j - \mathbf{X}_j^T \mathbf{1}\mathbf{1}^T \mathbf{X}_j / n & j = k, \end{cases} \tag{12}$$

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

Corollary 2 *If there are no missing values in the data, then a solution 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 for the scales is given by* $\hat{\boldsymbol{\beta}}$ *which is the eigenvector of* $\mathbf{A}$ *in* (11) *with respect to* $\mathbf{B}$ *in* (12) *associated with the largest eigenvalue such that* $\hat{\boldsymbol{\beta}}^T \mathbf{B} \hat{\boldsymbol{\beta}} = N$*.*

Example 2 *For the iris data, the data matrix is*

$$\mathbf{X} = (\mathbf{X}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5)$$

where $\mathbf{X}_1$ *is the same* 150×2 *matrix as in* Example 1 *for* Species *and* $\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ *and* $\boldsymbol{x}_5$ *are the numerical data vectors for* Sepal Length*,* Sepal Width*,* Petal Length *and* Petal Width*, respectively. Using* Corollary 2*, we find that*

$$\hat{\boldsymbol{\beta}} = (50.57710, 73.10587, 32.61262, -34.70152, 17.55146, 39.71478)^T$$

and

$$\hat{\boldsymbol{\alpha}} = (-41.22766, -190.56643, 106.09412, -65.95838, -47.63126)^T$$

provided that $\alpha_0 = 0$*. Then the coordinate vectors are written as, for example,*

$$\boldsymbol{y}_1 = (-41.22766)\mathbf{1} + \mathbf{X}_1 \begin{pmatrix} 50.57710 \\ 73.10587 \end{pmatrix}, \tag{13}$$

and $\boldsymbol{y}_2 = (-190.56643)\mathbf{1} + (32.61262)\boldsymbol{x}_2$*. Equation* (13) *implies that the categories* Setosa*,* Versicolor *and* Virginica *are located at* $\hat{\alpha}_1 = -41.22766$*,* $\hat{\alpha}_1 + \hat{\beta}_1 = 9.349441$ *and* $\hat{\alpha}_1 + \hat{\beta}_2 = 31.878215$*, respectively, on the axis for* Species*.*

2.3 General Result

Now, we have to consider the case in which some of the data vectors are ordered categorical. Clearly, the order of the categories within an ordered categorical data vector has to be retained on the corresponding axis of the textile plot, otherwise the plot will mislead the user.

A natural choice of a contrast matrix $\mathbf{C}$ for an ordered categorical data vector with q categories is

$$c_{ij} = \begin{cases} 1, & i > j \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

for $i = 1, \ldots, q$ and $j = 1, \ldots, q-1$, as is illustrated in the following example.

Example 3 *Consider an ordered categorical data vector* $\boldsymbol{x} = (\text{Small}, \text{Medium}, \text{Large}, \text{Medium})^T$. *Then the coordinate vector can be written as*

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

where $\beta_1, \beta_2 \leq 0$ *or* $\beta_1, \beta_2 \geq 0$ *to retain the order of the categories on the corresponding axis of the textile plot.*

As noted before, the choice of contrast does not affect the result even when an ordered categorical data vector is present in the data. However it is not as advantageous to use a contrast other than $\mathbf{C}$ in (14), since the constraint on the scale parameters would not be as simple as in Example 3. Hereafter we assume that the contrast matrix for ordered categorical data vector is always $\mathbf{C}$ as given in (14).

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

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \mathbf{A}_{11} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \mathbf{A}_{12} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{A}_{22} \boldsymbol{\beta} \tag{15}$$

under the equality constraint

$$\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} = N \tag{16}$$

together with the inequality constraints

$$\boldsymbol{\beta}(\mathscr{I}_k) \geq \mathbf{0} \quad \text{or} \quad \boldsymbol{\beta}(\mathscr{I}_k) \leq \mathbf{0}, \quad k = 1, \ldots, r. \tag{17}$$

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

By noting that $\hat{\boldsymbol{\alpha}}$ is still given as in (7), we obtain the following theorem by

applying a well-known constrained minimisation result (see e.g. Proposition 1.29 in Bertsekas [4]).

Theorem 1 *If the given data vectors satisfy* Assumptions 1 *and* 2*, then a solution which minimises* $S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{m})$ *under the constraints* (16) *and* (17) *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 for the scales,* $\hat{\boldsymbol{\beta}}$*, can be obtained by selecting an index set* $\mathscr{I}_0 \subseteq \bigcup_{k=1}^{r} \mathscr{I}_k$ *such that*

(1) $\hat{\boldsymbol{\beta}}(\mathscr{I}_0) = \mathbf{0}$*, and* $\hat{\boldsymbol{\beta}}(\mathscr{I}_0^c)$ *is an eigenvector of* $\mathbf{A}(\mathscr{I}_0^c, \mathscr{I}_0^c)$ *with respect to* $\mathbf{B}(\mathscr{I}_0^c, \mathscr{I}_0^c)$ *associated with the largest eigenvalue* $\hat{\lambda}$*, such that* $\hat{\boldsymbol{\beta}}(\mathscr{I}_0^c)\mathbf{B}(\mathscr{I}_0^c, \mathscr{I}_0^c)\hat{\boldsymbol{\beta}}(\mathscr{I}_0^c) = N$*, 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* $k = 1, \ldots, r$*,*

for which the $\hat{\lambda}$ *is the largest.*

A straight-forward algorithm to find the solution $\hat{\boldsymbol{\beta}}$ is the following.

(1) Find all possible index sets $\{\mathscr{I}_0\}$ for which conditions (1) and (2) are satisfied.
(2) Find an $\mathscr{I}_0^*$ in $\{\mathscr{I}_0\}$ for which $\hat{\lambda}$ is the largest.
(3) Then $\hat{\boldsymbol{\beta}}(\mathscr{I}_0^*)$ is the solution.

Further sophistication of the algorithm is possible in various ways but we leave that for future investigation.

3 Further Details of the Textile Plot

In the previous section we developed several proposals for determining an optimal choice of locations and scales. We are now in a position to plot the points of $\boldsymbol{y}_j$ on a parallel axis $j = 1, \ldots, p$ using a common coordinate system. However, as Cleveland [5] states "A graphical method is successful only if the decoding process from the given graphic by the viewer is effective". Thus, our aim in designing the textile plot was not only to graphically represent the data points themselves but also to assist the user in their interpretation of the data. With this aim in mind, it would appear reasonable to display any other information that might be helpful to the user in the textile plot together with the data.

Here we introduce two technical terms convenient for describing the design principle of the textile plot. Since a textile is a fabric produced by weaving warps together with wefts, we call the display of each data vector together with any necessary information a *warp*, and the connected straight lines defining the trajectory of a case a *weft*.

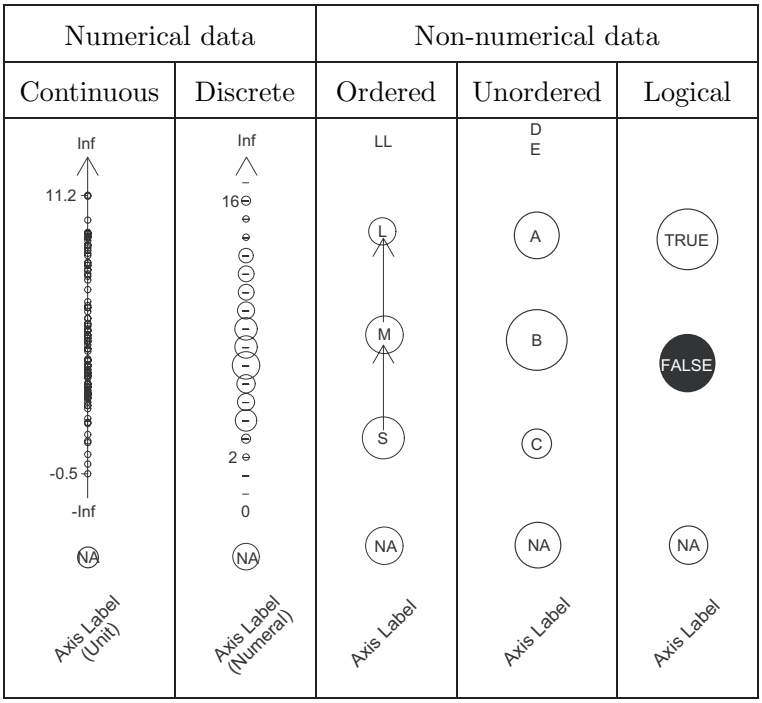

Fig. 3. Different forms of warp.

3.1 Warps

A warp in the textile plot is an integrative display of the information associated with a data vector. Clearly, *data type* is one of the important attributes of any data vector. The distinction between merely being quantitative or qualitative is generally not enough for an informative display, particularly in the case of high dimensional data. We first classify a data vector as being numerical or non-numerical and further classify the former as being *continuous* or *discrete*. The latter might be further classified as being *ordered categorical*, *unordered categorical* or *logical*. For simplicity, in what follows we restrict our attention to these five main data types.

Figure 3 illustrates how points are displayed on a warp for each data type. In the case of numerical data, the indication of the possible values aids in the understanding of the data. The possible values are indicated by a continuous vertical line if the data are continuous, and by tick-marks otherwise. The maximum and minimum possible values are identified at the ends of each axis. This enables the user to understand the background to the data beyond the distinction of merely being continuous or discrete. An arrowhead placed on either end of the warp indicates the direction of "low" to "high" for each coordinate axis. It points upwards on the jth warp if $\beta_j \geq 0$, and downwards otherwise. If there is a point that is repeated, a circle centred at the point coordinates is placed on each warp with an area proportional to the number of repeated values at that point. A similar idea to this was introduced by Parabox

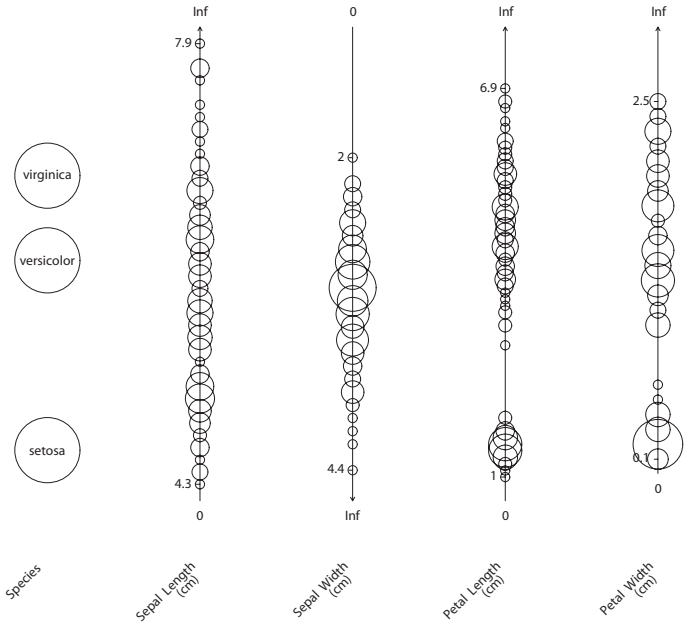

Fig. 4. Display of the warps for the iris data

[21] for a parallel coordinate plot designed to accommodate categorical data. An alternative would be to place a histogram on each warp as in [14]. However, the plot soon becomes illegible as the dimensionality of the data increases due to the fact that so many histograms have to be positioned on the plot. Also, class intervals have to be subjectively chosen for the histograms when the data are continuous. We chose to use circles because of their simplicity and because they do not involve any of the subjectivity inherent in the use of histograms. The minimum and maximum of a data vector are also indicated by the figures which appear to the left of each axis.

In the case of non-numerical data, each category can be identified by its category name placed on the coordinate. Also, relative frequencies are indicated by the area of a circle. Clearly, this design is consistent with that for numerical data. Zero frequency categories are indicated at the top of the display by their category names without circles. This is similar to the display of possible values in the case of discrete data. If the vector is ordered categorical then the categories are connected by a sequence of arrows to indicate their natural order. If the vector is logical, the circle for FALSE is filled to distinguish logical from categorical.

In all cases, missing values are indicated using circles, the areas of which are proportional to the number of missing values, placed at the bottom of each warp. Each warp is tagged with a label and its units (in the case of numerical data). The display design described here is clearly only one of many other possible choices. For example, points for continuous data could be displayed using a histogram, for instance. However, we decided to use circles for the points so as to maintain consistency over different data types. Figure 4 displays

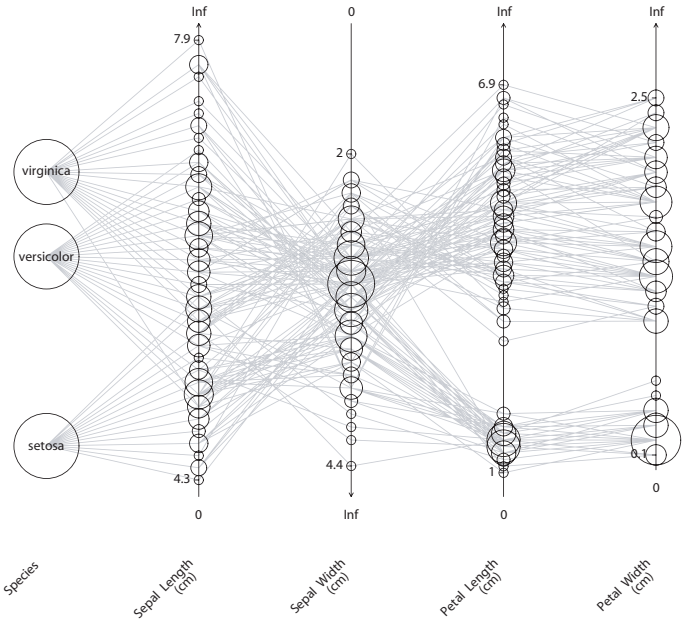

Fig. 5. Wefts overlaid on the display of the warps for the iris data.

the five warps for the iris data.

Care needs to be taken if some of the vectors are identifying (ID) vectors, since these are categorical data vectors for which the values are all distinct. Any ID vector can be excluded in the computation of the location and scales since the coordinate vector of ID vectors is always equal to $\boldsymbol{m}$ except for a constant shift and multiplication, as is proved in Appendix B. The display of the ID warp is optional. If it is required, $\boldsymbol{m}$ is used as the coordinate vector for the ID warp. We can see how horizontalised the wefts are from the ideal coordinates.

3.2 Wefts

A weft on the textile plot is traced out by the linked line segments for each case, although the segments will be disconnected if there are any missing values. Figure 5 is a textile plot of the iris data where all the wefts are overlaid on Figure 4. The display of wefts is simpler than that of warps, since each weft corresponds to just one individual case. Various attributes of a weft, such as its width, line type, colour etc, can be introduced to distinguish certain cases from others, but this is probably better done through interaction with the user. We leave such design enhancements and the construction of a user friendly environment with which to produce textile plots to further investigation.

3.3 Order of Warps

Different orders for the warps displayed in a textile plot give the user different impressions of the data. Certainly, it is rather rare that a natural order of the warps might previously be known. In all other scenarios, it would appear reasonable to order warps using some objective criterion. In the context of the parallel coordinate plot, Ankerst et al. [2] proposed a method that maximises the sum of similarity measures between two adjacent axes on a parallel coordinate plot. A more general discussion of this problem and its potential solution can be found in Yang et al. [22]. They proposed two dimension-ordering techniques; similarity-oriented dimension ordering based on the similarity measures in Ankerst et al. and importance-oriented dimension ordering based on the result of principal component analysis. These two approaches are also implemented in the DAVIS [6] software.

However, the situation is somewhat different in the textile plot. We have already introduced a criterion for choosing locations and scales, and the order of the warps can also be determined using the same criterion. Here we propose two different methods for determining the order of the warps. One is based on the distance to the mean vector $\boldsymbol{m}$, and the other is based on the absolute deviations between two adjacent warps. The former is closely related to the importance-oriented dimension ordering and is good for the classification of wefts. The latter is related to the similarity-oriented dimension ordering and is good for the classification of warps.

3.3.1 Distance to the Mean Vector (Classification of Wefts)

The distances $\|\boldsymbol{y}_j - \boldsymbol{m}\|_{\boldsymbol{w}_j}/\|\boldsymbol{w}_j\|$, $j = 1, \ldots, p$, can be used to determine the order of the warps, since the locations and scales are chosen to minimise $\sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2_{\boldsymbol{w}_j}$. The normalisation by $\|\boldsymbol{w}_j\|$ reflects the effective number of observations. If the warps are arranged from left to right according to ascending distance, the leftmost warps are then considered to be the most important warps for the classification of wefts. This is because the mean vector $\boldsymbol{m}$ essentially gives us a set of ideal coordinates for each case. The warps in Figure 2 are ordered by this criterion and show that the wefts passing through warps *Petal Length*, *Species* and *Petal Width* are roughly classified into three groups. This suggests that the criterion might be useful for classification of the wefts or, equivalently, the cases.

3.3.2 Distance Between Warps (Classification of Warps)

A natural choice of distance between two adjacent warps on the textile plot is the mean absolute deviation $\sum_{i=1}^{n} |y_{ij} - y_{ik}|/n$ where the jth and the kth

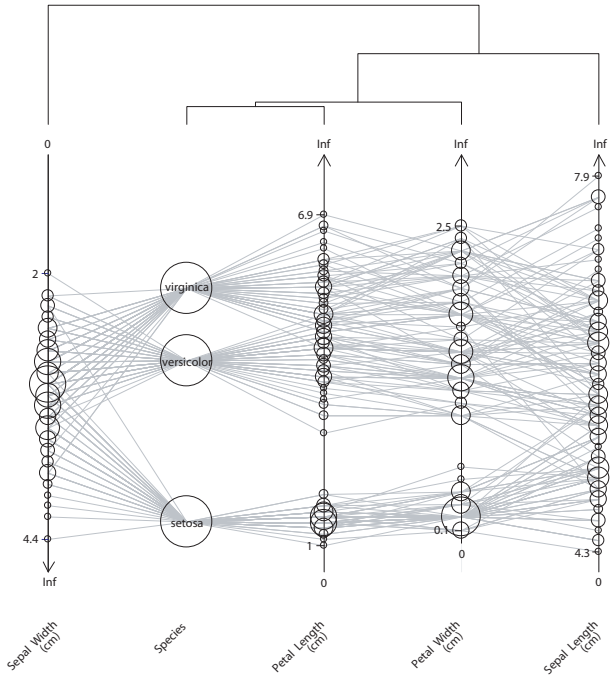

Fig. 6. Textile plot for the iris data with the locations of the warps ordered using a clustering algorithm.

warps are adjacent. Allowing for missing values, this distance becomes

$$\frac{1}{\boldsymbol{w}_j^T \boldsymbol{w}_k} \sum_{i=1}^{n} w_{ij} w_{ik} |y_{ij} - y_{ik}|.$$

One way of ordering the warps is to apply a clustering algorithm based on the above distances. For example, the ordered single end-linkage clustering algorithm proposed by Hurley [9] can be employed, although it was originally proposed for the rearrangement of the axes on a parallel coordinate plot. This algorithm provides an order for the warps together with a dendrogram. Figure 6 shows a textile plot of the iris data in which the ordering of the warps was determined using this particular clustering algorithm. It can be seen that the most similar warps are *Species* and *Petal Length*, followed by *Petal Width*. The distance between two adjacent warps on the dendrogram is indicated by the height of a merge of two clusters of warps.

4 Significant Features of the Textile Plot

Two important features which are sometimes found on a textile plot are a unique *knot* on a warp and completely *parallel wefts* between two adjacent warps, as illustrated in Figure 7. A knot is a point on a warp where all the wefts intersect, indicating that the warp is unrelated to the others. They arise

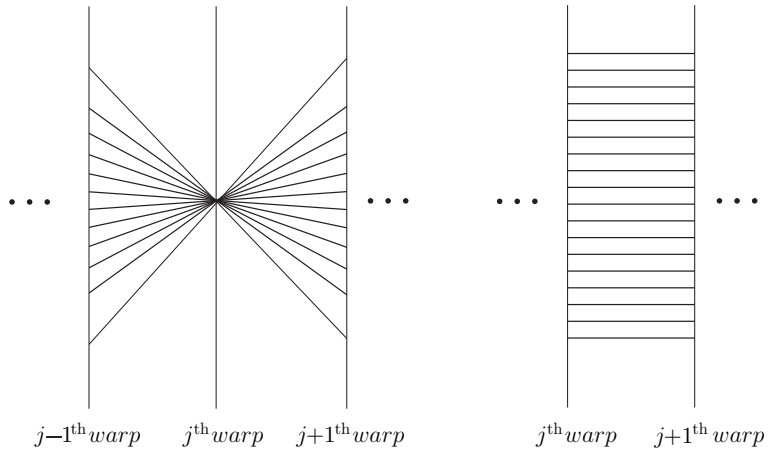

Fig. 7. Stylised representations of a unique knot on the jth warp (left) and completely parallel wefts for the jth and $(j+1)$st warps (right).

as a result of choosing the locations and scales so that all wefts are aligned as horizontally as possible, and are a feature specific to the textile plot. To clarify the conditions under which they are produced, we will prove that they occur when a data vector is essentially orthogonal to the other data vectors in the textile plot. Parallel wefts occur when all the wefts in two adjacent warps are horizontally aligned. It is intuitively clear that parallel wefts imply the linear dependence of two numerical data vectors, but the converse is not so clear. Later in this section we will discuss the conditions for parallel wefts, including the case where categorical data vectors are involved.

Unique knots and complete parallel wefts, therefore, indicate two extremes; a form of independence on the one hand and perfect linear dependence on the other. In practice, we can omit such warps to simplify the textile plot, as we illustrate in Section 5.2.

To simplify things, we will assume that there are no missing values and no ordered categorical data vectors in the given data set. Under Assumption 2 in Section 2.1, we can further assume that the data matrices $\mathbf{X}_j$, $j = 1, \ldots, p$ are normalised so that, without loss of generality,

$$\mathbf{1}^T \mathbf{X}_j = \mathbf{0} \quad \text{and} \quad \mathbf{X}_j^T \mathbf{X}_j = \mathbf{I}, \quad j = 1, \ldots, p. \tag{18}$$

Note that the textile plot is invariant under location and scale shifts of the original data vector or of the choice of contrasts. We also assume that $\alpha_0 = 0$ in Corollary 2 since the choice of α_0 does not change the appearance of the textile plot. This assumption implies that the mean vector $\boldsymbol{m}$ is always orthogonal to the vector $\mathbf{1}$.

A unique knot on the jth warp is produced when the selected scale parameter is zero, that is, $\hat{\boldsymbol{\beta}}(\mathscr{I}_j) = \mathbf{0}$. Define

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

which is now an $n \times q$ matrix with $q = Q - (q_j - 1)$. As we will now show, the singular value decomposition $\mathbf{UDV}^T$ of $\mathbf{X}_{-j}$ plays an important role in the occurrence of a unique knot. Here the diagonal elements of $\mathbf{D} = \mathrm{diag}(d_j;\ j = 1, \ldots, q)$ are singular values arranged in the order $d_1 \geq d_2 \geq \cdots \geq d_q \geq 0$, and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q)$ and $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)$ are column-orthogonal matrices.

Theorem 2 *Assume that there are no missing values in* $\mathbf{X}$ *and no ordered categorical data vectors in the data. Under the assumption that the multiplicity of the largest singular value* d_1 *of* $\mathbf{X}_{-j}$ *is* 1*, a necessary and sufficient condition for a unique knot to occur on the* j*th warp is that*

$$\mathbf{X}_j^T \boldsymbol{u}_1 = \mathbf{0} \tag{19}$$

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*

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

The proof of Theorem 2 is given in Appendix C.

Note that $\boldsymbol{u}_1$ is proportional to the mean vector $\boldsymbol{m}_{-j}$ for the data matrix $\mathbf{X}_{-j}$, around which all coordinate vectors on the textile plot of $\mathbf{X}_{-j}$ are aligned. Therefore, condition (19) specifies that any column vector of $\mathbf{X}_j$ is orthogonal to $\boldsymbol{m}_{-j}$, which is the mean vector with the jth element omitted. However, as the theorem tells us, orthogonality is not enough to produce a unique knot. The projected size of $\mathbf{X}_j$ on the range space of $\mathbf{X}_{-j}$ has to be small enough relative to the size of $\mathbf{X}_{-j}$.

Note that

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

holds true for any $(q_j - 1)$-dimensional vector $\boldsymbol{z}$. The following corollary gives us a simplified sufficient condition for the occurrence of a unique knot on a warp.

Corollary 3 *Under the same assumption as in* Theorem 2*, a sufficient condition for the occurrence of a unique knot 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* $(d_1^2 - d_2^2)(d_1^2 - 1)/d_2^2$.

The sufficient condition given 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 $\mathbf{X}_j^T\mathbf{U}\mathbf{U}^T\mathbf{X}_j$ has a scalar value. Therefore, it is easy to check if $\mathbf{X}_j^T\boldsymbol{u}_1 = 0$ and $\mathbf{X}_j^T\mathbf{U}\mathbf{U}^T\mathbf{X}_j < (d_1^2 - d_2^2)(d_1^2 - 1)/d_2^2$. Even if the jth data vector is not numerical, the following example gives us a simple sufficient condition, since all the eigenvalues of $\mathbf{X}_j^T\mathbf{U}\mathbf{U}^T\mathbf{X}_j$ are less than or equal to 1.

If $\mathbf{X}_j^T\boldsymbol{u}_1 = \mathbf{0}$, a sufficient condition for a unique knot to occur on the jth warp is that $d_1^2 > d_2^2 + 1$. As is shown in Appendix D, this condition is equivalent to

$$p - 2 - \frac{1}{n}S^2(\hat{\boldsymbol{\alpha}}_{-j}, \hat{\boldsymbol{\beta}}_{-j}, \boldsymbol{m}_{-j}) > d_2^2 \geq d_3^2 \geq \cdots \geq d_q^2,$$

where $\hat{\boldsymbol{\alpha}}_{-j}$ and $\hat{\boldsymbol{\beta}}_{-j}$ are the solutions for the location and scale parameter vectors, respectively, for $\mathbf{X}_{-j}$. This condition is satisfied when all of the wefts in the textile plot of $\mathbf{X}_{-j}$ are well-aligned.

If we make the stronger assumption that $\mathbf{X}_j^T\mathbf{X}_{-j} = \mathbf{O}$, then $\mathbf{X}_j^T\mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T\mathbf{X}_j = \mathbf{O}$ so that $\mathbf{X}_j^T\boldsymbol{u}_1 = \mathbf{0}$. The following gives us a simpler condition for a unique knot. If $\mathbf{X}_j^T\mathbf{X}_{-j} = \mathbf{O}$, a unique knot is always produced on the jth warp.

As is seen from the proof of Lemma 1 or of Theorem 2 given in Appendix C, all the wefts will intersect near a point if $\mathbf{X}_j^T\mathbf{X}_{-j}$ is close to $\mathbf{O}$ because of the continuity of the eigenvalue problem.

4.2 Completely Parallel Wefts

Completely parallel wefts between the jth and the $(j+1)$st warps occur when the coordinate vectors $\boldsymbol{y}_j$ and $\boldsymbol{y}_{j+1}$ are identical. To see this, it is enough to consider the case when no unique knot occurs on either of the two warps. In the case of two numerical data vectors, a necessary and sufficient condition for completely parallel wefts is that the data vectors are identical except for differences in locations and scales. Here, the necessary part is trivial, but the sufficiency requires proving. To do so, one can consider, without loss of generality, the case where $\mathbf{X}_j = \pm\mathbf{X}_k$. Then $\boldsymbol{y}_j = \boldsymbol{y}_k$ follows from the fact that $\boldsymbol{\beta}(\mathscr{I}_j) = \pm\boldsymbol{\beta}(\mathscr{I}_k)$, since $\hat{\boldsymbol{\beta}}$ is the solution of $\lambda\mathbf{B}\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\beta}$ with $\mathbf{B} = \mathbf{I}$ and $\mathbf{A} = (\mathbf{X}^T\mathbf{X})/p$ in this case.

When the two data vectors are both categorical, it is hard to derive a necessary and sufficient condition. However, a sufficient condition is that there is a one-to-one association between the categories of the two vectors. That is, the

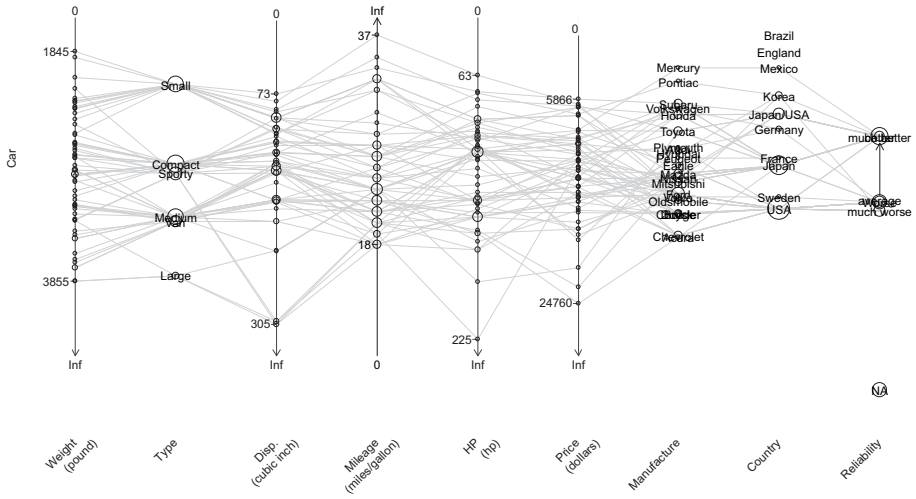

Fig. 8. Textile plot for the automobile data.

two data vectors are identical except for the difference in the labels for the categories.

All wefts between two warps will be reasonably well aligned horizontally if the projection of the mean vector $\boldsymbol{m}$ on the range space of $\mathbf{X}_j$ is close to that of $\mathbf{X}_{j+1}$, again because of the continuity of the eigenvalues.

5 Practical Examples

In this section we present two examples of the use of the textile plot. The first considers the automobile data set used as example data in S-Plus [17]. Within this data set there are three types of data vector: numerical, unordered and ordered categorical. The second example considers the body measurement data from [11] for 318 Japanese people. In this data set, 54 variables (49 body measurements and 5 other attributes) were recorded for each subject.

5.1 Automobile Data

Figure 8 presents a textile plot of the automobile data in which the warps have been ordered by their distances to the mean vector as described in Section 3.3.1. In this data set there are five numerical data vectors (*Weight*, *Mileage*, *Displacement*, *Horse Power* and *Price*), three unordered categorical data vectors (*Country*, *Manufacturer* and *Type*) and an ordered categorical data vector (*Reliability*). Each data vector contains measurements made on 60 cars. Note that the scale used on the *Mileage* warp in Figure 8 runs in the opposite direction to the scales of the other numerical variables. As can

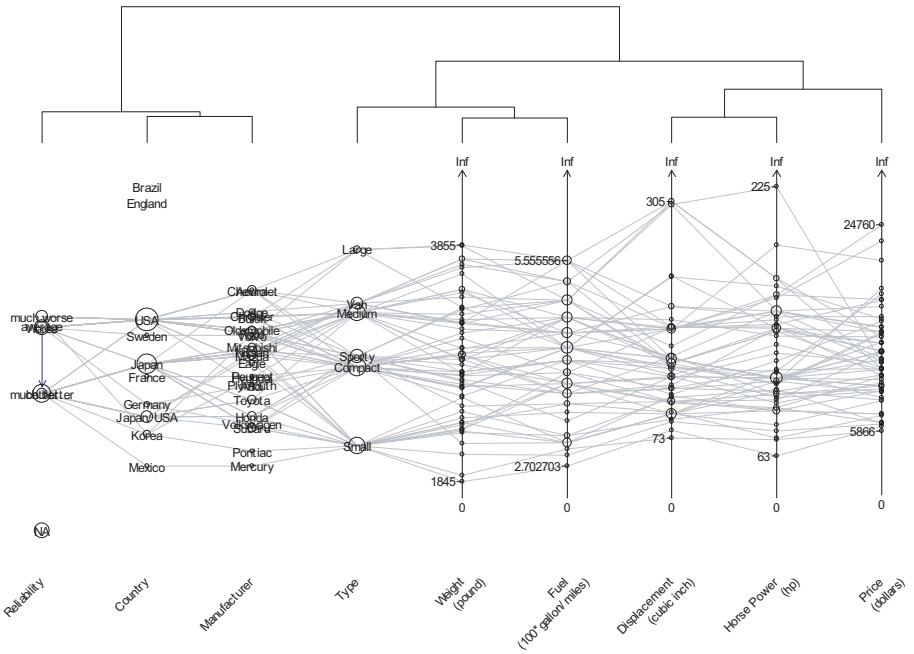

Fig. 9. Textile plot of the automobile data.

be seen, the wefts passing through this warp confirm the nonlinear pattern. Since gallons per mile is as sensible a measure of gas consumption as miles per gallon, it would appear reasonable to attempt to straighten the relationship by applying an inverse transformation to *Mileage*.

Figure 9 shows the equivalent textile plot obtained after transforming the variable *Mileage* into the variable *Fuel* using the inverse transformation referred to above. As can be seen from the dendrogram, the nine warps are classified into two groups.

The first group consists of the warps *Reliability*, *Country* and *Manufacturer*. A significant feature of this group is that the five categories of *Reliability* are clustered into two. One is formed from the categories *Much Worse*, *Worse* and *Average* and the other from the categories *Better* and *Much Better*. This implies that, at least for these data, only two categories are actually required in order to describe the reliability of a car. One can also see clearly how the reliability of a car is related with the *Country* in which it was manufactured and the *Manufacturer*. The plot appears to suggest that cars manufactured in Korea and Mexico are more reliable. Nevertheless, one has to be somewhat cautious with this interpretation because the number of cars manufactured in these two countries is far smaller than the numbers of cars manufactured in other countries. Note that the number of observations for each country is indicated by the area of each circle on the *Country* warp.

The second group is formed from the warps *Type*, *Weight*, *Fuel*, *Displacement*, *Horse Power* and *Price*. It can also be seen that the six warps are further

classified into two subgroups; one comprising of *Type*, *Weight* and *Fuel*, and the other made up from *Displacement*, *Horse Power* and *Price*. The warps in the first of these two subgroups are related to the size of a car and those in the latter subgroup are related to the size of engine and the price of a car.

5.2 Body Measurement Data

As mentioned previously, our second data set is comprised of 49 different body measurements and 5 other attributes collected for 318 Japanese people. The names of all 54 data vectors are listed in Table 1. What each of the 49 body measurements, other than *Body Mass*, represents is identified in Figure 10, where the numbers refer to those used to identify the data vectors in Table 1.

Figure 11 is a textile plot of the body measurement data where warps are ordered by their distances to the mean vector as described in Section 3.3.1. The representations of the three right most warps (*School*, *Occupation* and *Race*) indicate that these variables take only one value each. Moving left, the next two warps (*Bicristal Breadth* and *Toe I Angle*) come close to having a unique knot. It is therefore advisable to delete such warps from the textile plot since they are unable to discriminate between subjects. Clearly, then, textile plots with warps ordered by distance to the mean vector are useful for identifying warps that are redundant.

Figure 12 is a textile plot for the remaining 49 warps once the five most extreme warps to the right of the previous textile plot where removed from the analysis. In this plot the warps were ordered using the clustering algorithm described in Section 3.3.2. It can be seen from the dendrogram towards the top of the textile plot that the warps are classified into three main groups: the first formed by the first 11 warps on the left of the plot, the second by the next 17 warps, and the third by the last 21 warps. Figures 13, 14 and 15 are textile plots for each of these three groups of warps. These figures provide clearer representations of the relationships that exist between the warps within each group.

In Figure 13 the most extreme four warps on the left (20, 39, 40 and 47) are measurements of skinfold thickness at four points on the body. These variables are identified using asterisks in Figure 10. The remaining seven warps are measurements related to *Gender*. It is known that structural differences between males and females occur mainly in the shoulders and hands, but the plot also shows that these differences also appear in *Bicondylar Humerus*, *Medial Malleolus Height* and *Lateral Malleolus Height*. It is interesting to note that the scales for the four most extreme warps on the left run in the opposite direction to those for the other warps. This is simply because the measurements

Table 1
Variables associated with the body measurement data

No.	Data Vector	No.	Data Vector
1	Gender	28	Hand Breadth
2	Age	29	Hand Length From Crease
3	School	30	Hand Length From Stylion
4	Occupation	31	Hand Thickness
5	Race	32	Heel Breadth
6	Body Mass	33	Hip Circumference
7	Stature	34	Instep Length
8	Iliac Spine Height Standing	35	Lateral Epicondyle Height
9	Shoulder (Biacromial) Breadth	36	Lateral Malleolus Height
10	Head Length	37	Maximum Body Height
11	Head Breadth	38	Medial Malleolus Height
12	Chest Circumference	39	Subscapular Skinfold Thickness
13	Waist Circumference	40	Suprailiac Skinfold Thickness
14	Calf Circumference	41	Suprasternal Height
15	Ball Angle	42	Symphyseal Height
16	Ball Breadth	43	Thigh Circumference
17	Bicondylar Femur	44	Toe I Angle
18	Bicondylar Humerus	45	Toe V Angle
19	Bicristal Breadth	46	Total Head Height
20	Calf Skinfold Thickness	47	Triceps Skinfold Thickness
21	Cristal Height	48	Trochanterion Height
22	Fibular Instep Length	49	Upper Arm Circumference
23	Foot Breadth	50	Upper Arm Circumference flexed
24	Foot Circumference	51	Upper Arm Length
25	Foot Length	52	Upper Limb Length
26	Forearm Circumference	53	Waist Breadth
27	Forearm Length	54	Waist Height

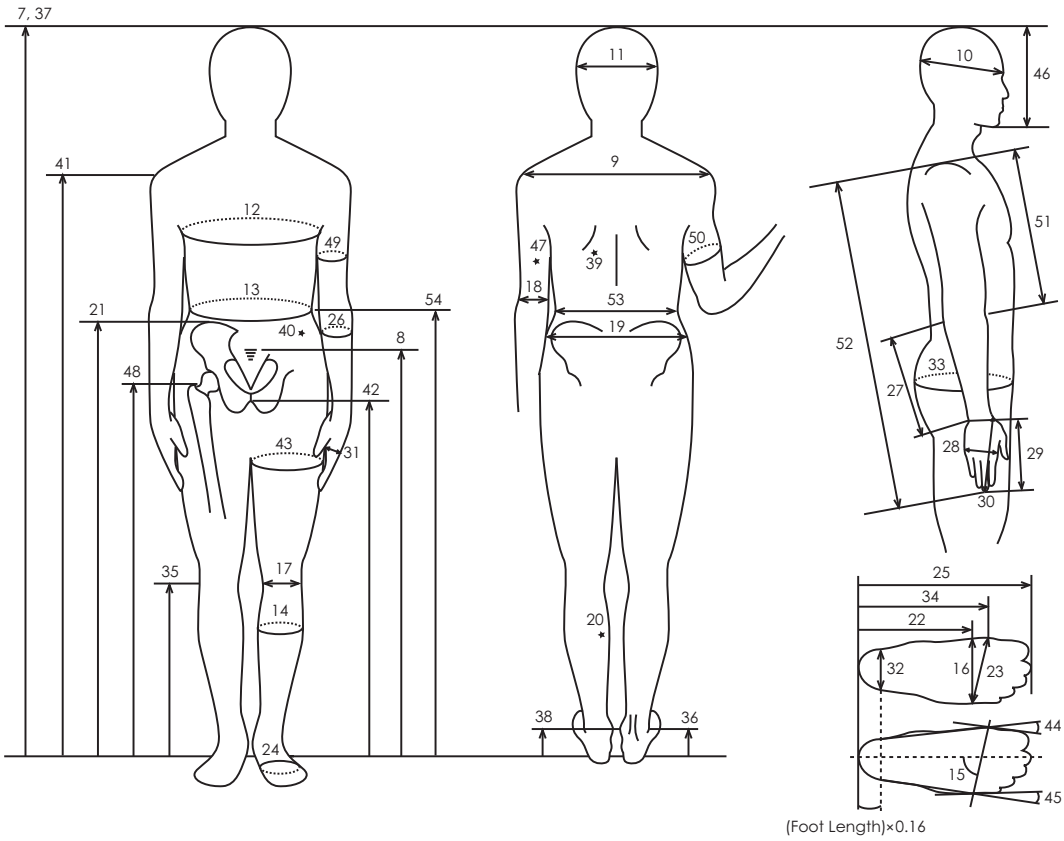

Fig. 10. Identification of the characteristics of the body measured.

for skinfold thickness tend to be higher for females.

In Figure 14, the three most extreme warps on the left are arm measurements and the next nine warps are related to height and leg length. The remaining five warps are related to foot or hand length. As a whole, the wefts in this textile plot are almost parallel because the measurements considered are strongly related to human height.

In Figure 15, the subgroup on the left consists of warps related to circumference measurements and weight. The subgroup on the right consists of warps related to foot and head measurements, and age. In this data set, the ages of examinees are clustered into two groups; one young (around 20 years old) and the other old (around 70), which is clearly evident from an inspection of the warp for *Age*. As expected, foot and head measurements reflect the different age cohorts of the people considered. Note that *Ball Angle* manifests something close to a unique knot, which indicates that it is orthogonal to the other measurements and is unable to discriminate reliably between subjects.

The above observations are preliminary ones made after an initial exploratory analysis of the data based on a consideration of textile plots alone. Clearly, further investigation would be required to probe the issues raised in more

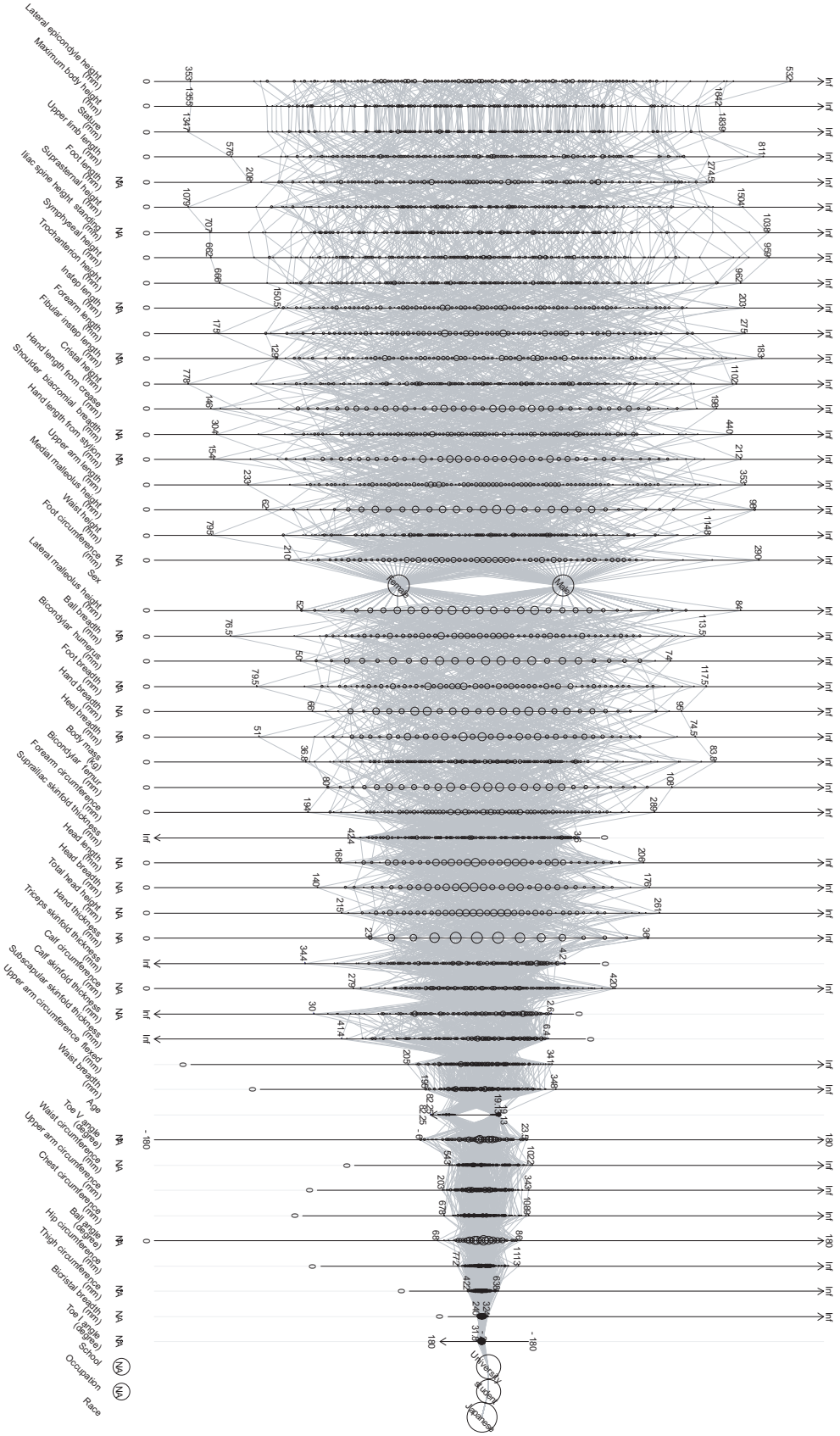

Fig. 11. Textile plot for the full body measurement data set.

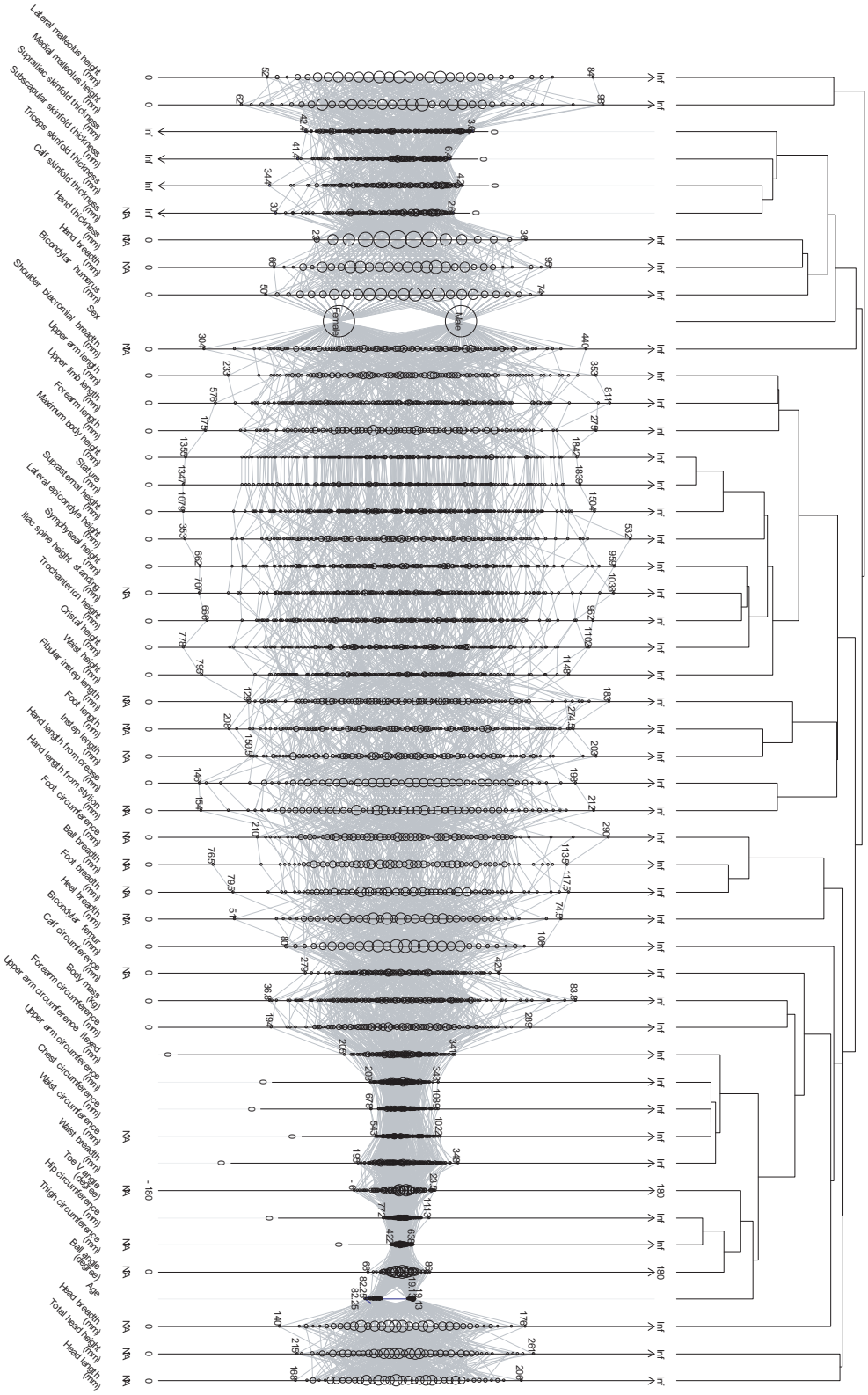

Fig. 12. Textile plot of the reduced body measurement data set where the 49 warps are ordered using the ordered single end-linkage clustering algorithm.

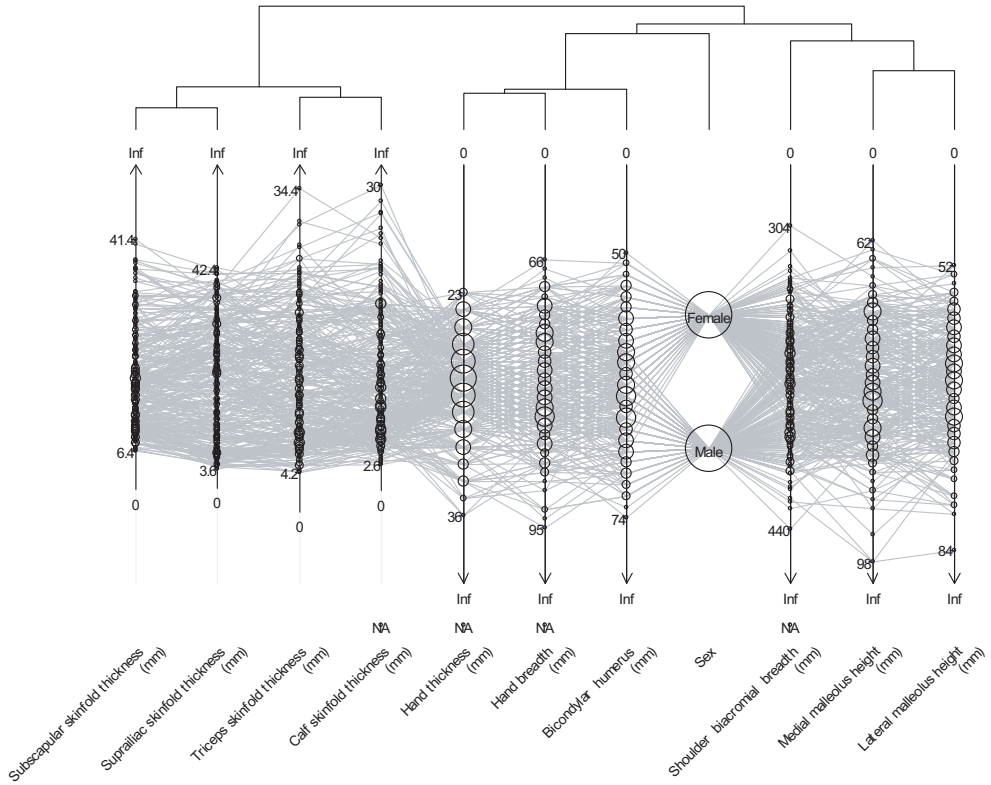

Fig. 13. Textile plot for the first group of 11 warps identified in the body measurement data set.

detail. Thus, the great value of the textile plot is that it can provide the user with informative graphical representations of high dimensional data which will often suggest potential avenues for subsequent further exploratory, or even confirmatory, data analysis.

6 Computational Aspects

6.1 Scalability

The most time consuming part of the computation of the textile plot is to find the eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$, as in Proposition 1, and calculate the coordinate vectors $\boldsymbol{y}_j$, $j = 1, \ldots, p$. Figure 16 provides a graphical summary of the computer time required to obtain the coordinate vectors as a function of the number of data vectors, p, and cases, n. This diagram corresponds to the situation in which all the data vectors are numeric. The program used to perform the calculations was written in C and incorporated an algorithm for solving the generalised eigenvalue problem available from *Lapack* [13]. The machine used was a PC with a Xeon 3.2 GHz dual-core processor with 2GB of

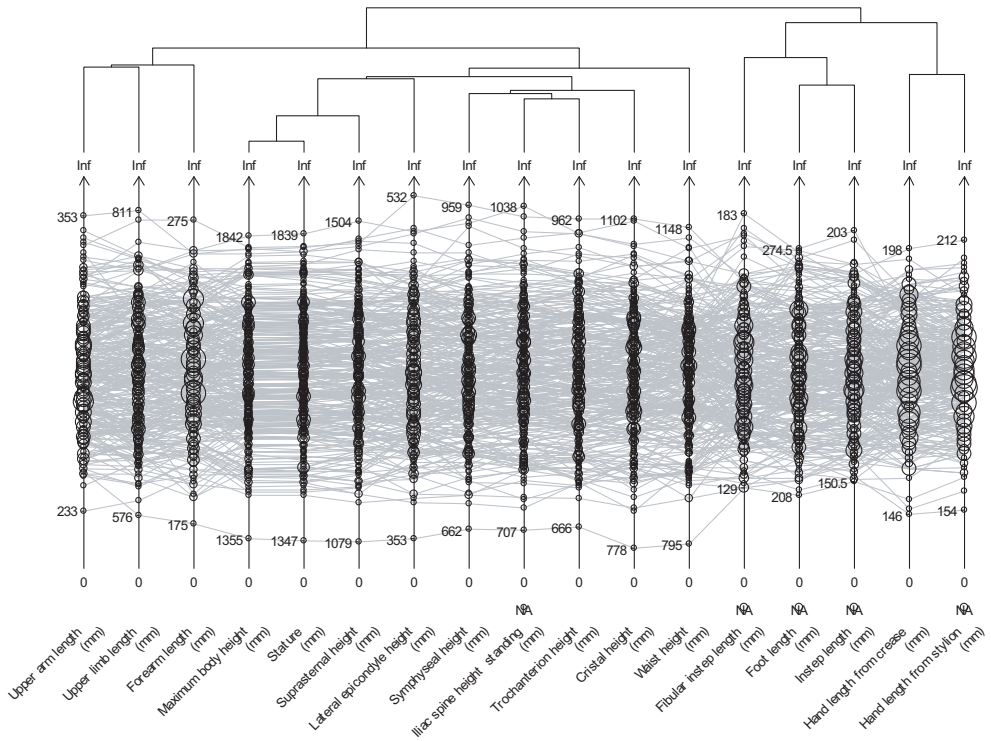

Fig. 14. Textile plot for the second group of 17 warps identified in the body measurement data set.

memory. From this diagram it can be seen that the computational burden is very light indeed for p-values up to order 100 and n-values up to order 1000. Clearly, for larger values of p and n the time required to perform the calculations can become considerably larger. However, note that for data sets with up to 100 dimensions and 10000 cases, the computations can be performed in under four seconds of computer time. Since there is no obvious way of decreasing the computational burden associated with computing the coordinate vectors for the textile plot, we hope that Moore's law continues to hold.

6.2 Inequality constraint

We are currently searching for an improved algorithm with which to find a solution when the data include ordered categorical data vectors. As described in Theorem 1, optimisation with inequality as well as equality constraints is necessary in order to find all those index sets, the $\mathscr{I}_0$'s, from which to select an optimal $\mathscr{I}_0$. The current implementation is to search all of the $\mathscr{I}_0$'s, but this is computationally expensive and the burden of computation increases with the number of ordered categorical data vectors.

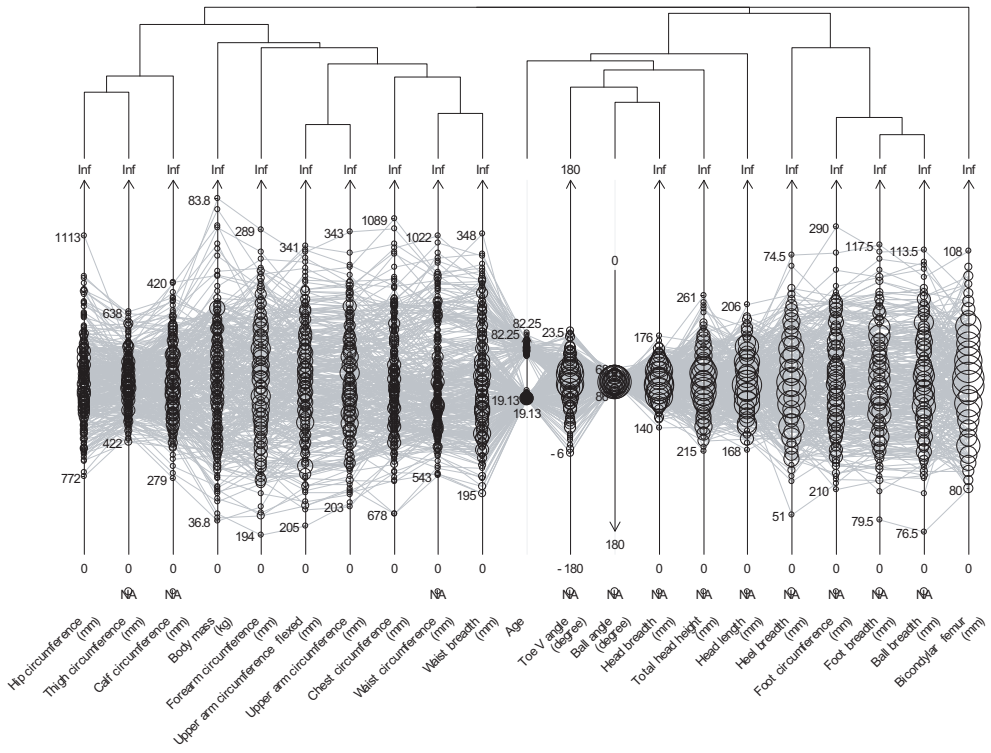

Fig. 15. Textile plot for the third group of 21 warps identified in the body measurement data set.

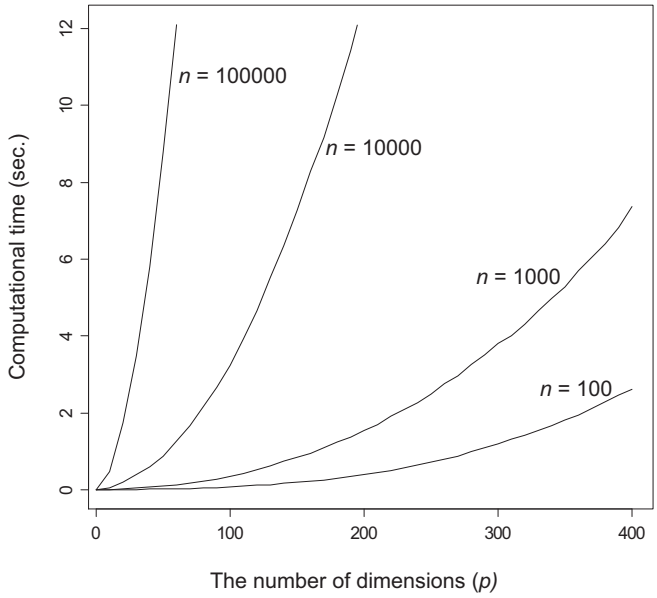

Fig. 16. Time required to compute the coordinate vectors as a function of p and n.

7 The Textile Plot and the Optimised Parallel Coordinate Plot

The optimised parallel coordinate plot was proposed by Michailidis and de Leeuw [12] in the context of homogeneity analysis where the main objective

is to find quantified vectors $\boldsymbol{y}_j$, $j = 1, \ldots, p$, for given categorical data vectors $\boldsymbol{x}_j$, $j = 1, \ldots, p$. The optimised parallel coordinate plot is a parallel coordinate plot of the resulting quantified vectors in which all the axes share a common coordinate system.

Let $\boldsymbol{x}_j$, $j = 1, \ldots, p$, be unordered categorical data vectors with no missing values and $\mathbf{Z}_j$, $j = 1, \ldots, p$, be the indicator matrices for $\boldsymbol{x}_j$ defined as in Section 2.2. The quantified vectors $\boldsymbol{y}_j = \mathbf{Z}_j\boldsymbol{\gamma}_j$, $j = 1, \ldots, p$, are defined so as to minimise

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

under the constraint

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

where $\bar{\xi} = \sum_{i=1}^{n} \xi_i/n$. In the same way as in Section 2.1, the minimisation of $\sigma_2(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p, \boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ yields the solution $\boldsymbol{\xi} = \boldsymbol{m}$. Therefore, the problem is to minimise

$$\begin{aligned} p\sigma_2(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p, \boldsymbol{m}) &= \sum_{j=1}^{p} \|\boldsymbol{m} - \boldsymbol{y}_j\|^2 \\ &= \sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2 - p\|\boldsymbol{m} - \bar{m}\mathbf{1}\|^2 + \sum_{j=1}^{p} \|\bar{y}_{\cdot j}\mathbf{1} - \bar{m}\mathbf{1}\|^2, \end{aligned} \tag{20}$$

under the constraint $\|\boldsymbol{m} - \bar{m}\mathbf{1}\|^2 = n$, where $\bar{m} = \sum_{i=1}^{n} m_i/n$. Recalling the relation

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

from (10), we see that the last term on the right hand side of (20) depends only on the location parameter vector $\boldsymbol{\alpha}$ and it vanishes when $\boldsymbol{\alpha}$ is taken to be $\hat{\boldsymbol{\alpha}}$ as given in Corollary 2. Thus the problem is to find $\boldsymbol{\beta}$ so as to minimise $\sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2 = \boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta}$ under the constraint $p\|\boldsymbol{m} - \bar{m}\mathbf{1}\|^2 = \boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta} = N$. In contrast, for the textile plot the problem is to find $\boldsymbol{\beta}$ so as to maximise $\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}$ under the constraint $\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta} = N$. However, it is clear that the two problems above yield the same solution $\hat{\boldsymbol{\beta}}$ since the eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$ associated with the largest eigenvalue is equal to the eigenvector of $\mathbf{B}$ with respect to $\mathbf{A}$ associated with the smallest eigenvalue.

Thus we see that the optimised parallel coordinate plot and the textile plot yield the same picture in the restricted case in which all the data vectors are categorical with no missing values. However, the aim of homogeneity analysis is the quantification of categorical data vectors, whereas the motivation for the textile plot is as an aid to the visualisation and exploration of the data.

8 Concluding Remarks

We have proposed a new data visualisation technique, which we have named the textile plot, with the hope that it will be adopted as a fundamental tool for exploring for relationships within high dimensional data sets.

The textile plot, whose name was derived by analogy to the production of a fabric in which warp and weft yarns are interwoven, is a generalisation of the parallel coordinate plot. Data vectors of any type (numerical, unordered or ordered categorical) can be displayed on warps in a concise way so as to provide valuable graphical and numerical summary of the data. The wefts, which trace out the trajectory of each case, are aligned as horizontally as possible so as to accentuate the differences between cases. Two important features of the textile plot, unique knots and completely parallel wefts, are also characterised by simple conditions.

It is important to develop an efficient algorithm for very high dimensional data sets or data sets containing a large number of ordered categorical data vectors. The introduction of dynamic or interactive displays such as those mentioned in [6], [18] or [19] would also be important improvements to the user interface. Such developments are left for further investigation.

Acknowledgements

The authors thank Dr. Peter Thomson at Statistics Research Associates Limited and Dr Arthur Pewsey at the University of Extremadura for their helpful suggestions. This research was partly supported by the 21st Century COE Program at Keio University: Integrative Mathematical Sciences.

Appendix A. Proof of Corollary 1

Note 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

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

holds ture for any constant α_0.

On the other hand,

$$\hat{\boldsymbol{\beta}} = \mathbf{B}^{-\frac{1}{2}}\boldsymbol{\gamma} \tag{21}$$

holds true, where $\boldsymbol{\gamma}$ is the eigenvector of the sample correlation matrix of the $\boldsymbol{x}_j$'s associated with the largest eigenvalue and $\|\boldsymbol{\gamma}\|^2 = N = np$. Note that

$$\mathbf{A}\hat{\boldsymbol{\beta}} = \lambda_{\max}\mathbf{B}\hat{\boldsymbol{\beta}}$$

is equivalent to

$$(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\boldsymbol{\gamma} = \lambda_{\max}\boldsymbol{\gamma}$$

and $p\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}$ is the sample correlation matrix. Therefore (21) can be written as

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

Appendix B. The matrices in Proposition 2

The matrix $\mathbf{A}_{12}$ is a $p \times Q$ matrix with

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

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

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

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 $j, k = 1, \ldots, p$. Here the notation $\cdot$ and / is used in a slightly extended way to accommodate matrices as well as vectors; that is, $\boldsymbol{v} \cdot \mathbf{Z} = (\boldsymbol{v} \cdot \boldsymbol{z}_1, \ldots, \boldsymbol{v} \cdot \boldsymbol{z}_r)$ and $\mathbf{Z}/\boldsymbol{v} = (\boldsymbol{z}_1/\boldsymbol{v}, \ldots, \boldsymbol{z}_r/\boldsymbol{v})$ for an n-dimensional vector $\boldsymbol{v}$ and an $n \times r$ dimensional matrix $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_r)$.

Appendix C. The coordinate vector of a categorical data vector with all distinct values

We can encode a categorical data vector $\boldsymbol{x}_j$ containing values which are all distinct into an $n \times (n-1)$ matrix $\mathbf{X}_j$ by choosing a proper encoding matrix such that $\mathbf{1}^T \mathbf{X}_j = \mathbf{0}$. Then we have

$$\begin{aligned}
\mathbf{A}(\mathscr{I}_j, \mathscr{I})\hat{\boldsymbol{\beta}} &= \mathbf{A}_{12}(\{1, \ldots, p\}, \mathscr{I}_j)^T \mathbf{A}_{11}^{+} \mathbf{A}_{12} \hat{\boldsymbol{\beta}} - \mathbf{A}_{22}(\mathscr{I}_j, \mathscr{I})\hat{\boldsymbol{\beta}} \\
&= \mathbf{A}_{12}(\{1, \ldots, p\}, \mathscr{I}_j)^T \hat{\boldsymbol{\alpha}} - \mathbf{A}_{22}(\mathscr{I}_j, \mathscr{I})\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}_j^T(\boldsymbol{w}_1/\boldsymbol{w}, \ldots, \boldsymbol{w}_p/\boldsymbol{w})\hat{\boldsymbol{\alpha}} + \mathbf{X}_j^T(\boldsymbol{w}_1 \cdot \mathbf{X}_1/\boldsymbol{w}, \ldots, \boldsymbol{w}_p \cdot \mathbf{X}_p/\boldsymbol{w})\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}_j^T \sum_{k=1}^{p} \boldsymbol{w}_k \cdot \{\hat{\alpha}_k \mathbf{1} + \mathbf{X}_k \hat{\boldsymbol{\beta}}(\mathscr{I}_k)\}/\boldsymbol{w} \\
&= \mathbf{X}_j^T \sum_{k=1}^{p} \boldsymbol{w}_k \cdot \boldsymbol{y}_k/\boldsymbol{w} \\
&= \mathbf{X}_j^T \boldsymbol{m}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{B}(\mathscr{I}_j, \mathscr{I})\hat{\boldsymbol{\beta}} &= \mathbf{B}(\mathscr{I}_j, \mathscr{I}_j)\hat{\boldsymbol{\beta}}(\mathscr{I}_j) \\
&= \mathbf{X}_j^T \mathbf{X}_j \hat{\boldsymbol{\beta}}(\mathscr{I}_j) \\
&= \mathbf{X}_j^T \boldsymbol{y}_j,
\end{aligned}$$

because $\mathbf{B}$ is a block diagonal matrix. Therefore $\mathbf{A}\hat{\boldsymbol{\beta}} = \hat{\lambda}\mathbf{B}\hat{\boldsymbol{\beta}}$ implies that

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

where $\hat{\lambda}$ is the matrix eigenvalue given in Proposition 1. Note that

$$
\begin{aligned}
\mathbf{1}^T \boldsymbol{m} &= \mathbf{1}^T \sum_{k=1}^{p} \boldsymbol{w}_k \cdot \boldsymbol{y}_k / \boldsymbol{w} \\
&= \mathbf{1}^T \sum_{k=1}^{p} \boldsymbol{w}_k \cdot \{\hat{\alpha}_k \mathbf{1} + \mathbf{X}_k \hat{\boldsymbol{\beta}}(\mathscr{I}_k)\} / \boldsymbol{w} \\
&= \mathbf{1}^T (\boldsymbol{w}_1 / \boldsymbol{w}, \ldots, \boldsymbol{w}_p / \boldsymbol{w}) \hat{\boldsymbol{\alpha}} + \mathbf{1}^T (\boldsymbol{w}_1 \cdot \mathbf{X}_1 / \boldsymbol{w}, \ldots, \boldsymbol{w}_p \cdot \mathbf{X}_p / \boldsymbol{w}) \hat{\boldsymbol{\beta}} \\
&= n\hat{\alpha}_j - \mathbf{A}_{11}(j, \{1, \ldots, p\}) \hat{\boldsymbol{\alpha}} + \mathbf{A}_{12}(j, \mathscr{I}) \hat{\boldsymbol{\beta}} \\
&= n\hat{\alpha}_j = \mathbf{1}^T \boldsymbol{y}_j,
\end{aligned}
$$

and hence we obtain the desired result

$$
\boldsymbol{y}_j = \frac{1}{\hat{\lambda}} (\boldsymbol{m} - \bar{y}_{\cdot j} \mathbf{1}) + \bar{y}_{\cdot j} \mathbf{1}.
$$

Appendix D. Proof of Theorem 2

Before giving the proof of Theorem 2, we need the following lemma.

Consider a $Q \times Q$ symmetric matrix $\mathbf{C}$, partitioned as

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

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

Lemma 1 *Assume that the largest eigenvalue λ_1 of $\mathbf{C}_{22}$ has no multiplicity. Let $\hat{\boldsymbol{\gamma}}$ be that $\boldsymbol{\gamma}$ 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 that*

$$
\mathbf{C}_{12} \boldsymbol{p}_1 = \mathbf{0} \tag{23}
$$

and

$$
\mathbf{C}_{12} (\lambda_1 \mathbf{I} - \mathbf{C}_{22})^{+} \mathbf{C}_{12}^T < \lambda_1 \mathbf{I} - \mathbf{C}_{11} \tag{24}
$$

holds true in the sense of positive definiteness.

PROOF. We first partition the vector $\boldsymbol{\gamma}$ as

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

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

$$\boldsymbol{\gamma}^T \mathbf{C} \boldsymbol{\gamma} < \hat{\boldsymbol{\gamma}}^T \mathbf{C} \hat{\boldsymbol{\gamma}} = \lambda_1 \tag{25}$$

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

Condition (25) can be rewritten as

$$f(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) < \lambda_1 \tag{26}$$

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 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 that $\boldsymbol{\gamma}_2^*$ for which

$$(\lambda \mathbf{I} - \mathbf{C}_{22}) \boldsymbol{\gamma}_2^* = \mathbf{C}_{12}^T \boldsymbol{\gamma}_1, \tag{27}$$

where λ is a Lagrange multiplier. By using the Moore-Penrose inverse of $\lambda \mathbf{I} - \mathbf{C}_{22}$, a solution to (27) is given by

$$\boldsymbol{\gamma}_2^* = (\lambda \mathbf{I} - \mathbf{C}_{22})^+ \mathbf{C}_{12}^T \boldsymbol{\gamma}_1. \tag{28}$$

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 $\mathbf{C}_{22}$ associated with 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 $\lambda > \lambda_2$. Since we have already shown that $\hat{\boldsymbol{\gamma}}_1 = \mathbf{0}$ implies (23), it is sufficient to show that (25) is equivalent to (24).

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

$$\frac{1}{\varepsilon} = 1 + \sum_{i=2}^{q} \frac{\tilde{\boldsymbol{\gamma}}_1^T \mathbf{C}_{12} \boldsymbol{p}_i \boldsymbol{p}_i^T \mathbf{C}_{12}^T \tilde{\boldsymbol{\gamma}}_1}{(\lambda - \lambda_i)^2} \tag{29}$$

for $\lambda > \lambda_2$. The right hand side of (29) is now independent of ε and a monotone decreasing function of λ, ranging from ∞ to 1 for $\lambda_2 \leq \lambda < \infty$. Thus we can find a λ for any given $0 < \varepsilon \leq 1$. Here, we have employed the convention that $\lambda = \infty$, that is, $\boldsymbol{\gamma}_2^* = \mathbf{0}$, if $\varepsilon = 1$. Now,

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

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 (29), we can rewrite the inequality above as

$$\tilde{\boldsymbol{\gamma}}_1^T \mathbf{C}_{11} \tilde{\boldsymbol{\gamma}}_1 < \sum_{i=2}^{q} \frac{(\lambda_1 + \lambda_i - 2\lambda) \tilde{\boldsymbol{\gamma}}_1^T \mathbf{C}_{12} \boldsymbol{p}_i \boldsymbol{p}_i^T \mathbf{C}_{12}^T \tilde{\boldsymbol{\gamma}}_1}{(\lambda - \lambda_i)^2} + \lambda_1. \tag{30}$$

We now see that (30) is equivalent to (25) for any $\tilde{\boldsymbol{\gamma}}_1$ with $\|\tilde{\boldsymbol{\gamma}}_1\| = 1$ and $\lambda > \lambda_2$. Let us evaluate the lower bound for the right hand side of the inequality (30). The minimum of the right hand side of (30) 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 (25) is equivalent to the condition that

$$\tilde{\boldsymbol{\gamma}}_1^T \mathbf{C}_{11} \tilde{\boldsymbol{\gamma}}_1 < -\sum_{i=2}^{q} \frac{\tilde{\boldsymbol{\gamma}}_1^T \mathbf{C}_{12} \boldsymbol{p}_i \boldsymbol{p}_i^T \mathbf{C}_{12}^T \tilde{\boldsymbol{\gamma}}_1}{(\lambda_1 - \lambda_i)} + \lambda_1 \tag{31}$$

for any $\tilde{\boldsymbol{\gamma}}_1$ with $\|\tilde{\boldsymbol{\gamma}}_1\| = 1$. Note that the inequality (31) 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 (24) if we remember the definition of the Moore-Penrose inverse $(\lambda_1 \mathbf{I} - \mathbf{C}_{22})^+$. □

Using the above result, we have the following proof of Theorem 2.

PROOF.

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

$$\mathbf{C} = \begin{pmatrix} \mathbf{X}_j^T\mathbf{X}_j & \mathbf{X}_j^T\mathbf{X}_{-j} \\ \mathbf{X}_{-j}^T\mathbf{X}_j & \mathbf{X}_{-j}^T\mathbf{X}_{-j} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{X}_j^T\mathbf{U}\mathbf{D}\mathbf{V}^T \\ \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{X}_j & \mathbf{V}\mathbf{D}^2\mathbf{V}^T \end{pmatrix}$$

are 0 for the largest eigenvalue, since $\mathbf{A} = \mathbf{C}/p$ and $\mathbf{B} = \mathbf{I}$ as in (11) and (12). Then, by applying Lemma 1, we have the necessary and sufficient condition,

$$\mathbf{X}_j^T\mathbf{U}\mathbf{D}\mathbf{V}^T\boldsymbol{v}_1 = \mathbf{0} \quad \text{and} \quad \mathbf{X}_j^T\mathbf{X}_{-j}(d_1^2\mathbf{I} - \mathbf{X}_{-j}^T\mathbf{X}_{-j})\mathbf{X}_{-j}^T\mathbf{X}_j < (d_1^2 - 1)\mathbf{I}.$$

The result follows on noting that

$$\mathbf{X}_j^T\mathbf{U}\mathbf{D}\mathbf{V}^T\boldsymbol{v}_1 = d_1\mathbf{X}_j^T\boldsymbol{u}_1$$

and

$$\begin{aligned} \mathbf{X}_j^T\mathbf{X}_{-j}(d_1^2\mathbf{I} - \mathbf{X}_{-j}^T\mathbf{X}_{-j})\mathbf{X}_{-j}^T\mathbf{X}_j &= \mathbf{X}_j^T\mathbf{U}\mathbf{D}\mathbf{V}^T(d_1^2\mathbf{I} - \mathbf{V}\mathbf{D}^2\mathbf{V}^T)\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{X}_j \\ &= \mathbf{X}_j^T\mathbf{U}\mathbf{D}(d_1^2\mathbf{I} - \mathbf{D}^2)^+\mathbf{D}\mathbf{U}^T\mathbf{X}_j \\ &= \mathbf{X}_j^T\mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T\mathbf{X}_j. \qquad \square \end{aligned}$$

Appendix E. Unique knot condition

Using the matrices $\mathbf{A}$ and $\mathbf{B}$ as in Section 2.2, we can write the sum of squared deviations as

$$S^2(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \boldsymbol{m}) = -\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta} + \boldsymbol{\beta}\mathbf{B}\boldsymbol{\beta}.$$

Therefore, the minimum sum of squared deviations becomes

$$\begin{aligned} S^2(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \boldsymbol{m}) &= -\hat{\lambda}\hat{\boldsymbol{\beta}}^T\mathbf{B}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}\mathbf{B}\hat{\boldsymbol{\beta}} \\ &= N(1 - \hat{\lambda}), \end{aligned} \tag{32}$$

where $\hat{\lambda}$ is the largest eigenvalue of $\mathbf{A}$ with respect to $\mathbf{B}$. Provided there are no missing values and every sub-matrix of $\mathbf{X}_{-j} = (\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_p)$ is normalised as in (18), we have $\mathbf{A} = \mathbf{X}_{-j}^T\mathbf{X}_{-j}/p$ and $\mathbf{B} = \mathbf{I}$. Therefore $\hat{\lambda}$ can be written as $\hat{\lambda} = d_1^2/(p-1)$ where d_1 is the largest singular value of $\mathbf{X}_{-j}$. Then we can rewrite (32) for the data matrix $\mathbf{X}_{-j}$ into

$$S^2(\hat{\boldsymbol{\alpha}}_{-j}, \hat{\boldsymbol{\beta}}_{-j}, \boldsymbol{m}_{-j}) = n(p-1)\left(1 - \frac{d_1^2}{p-1}\right).$$

We now have

$$d_1^2 = p - 1 - \frac{1}{n} S^2(\hat{\boldsymbol{\alpha}}_{-j}, \hat{\boldsymbol{\beta}}_{-j}, \boldsymbol{m}_{-j}).$$

From the condition $d_1^2 > d_2^2 + 1$, we obtain

$$p - 2 - \frac{1}{n} S^2(\hat{\boldsymbol{\alpha}}_{-j}, \hat{\boldsymbol{\beta}}_{-j}, \boldsymbol{m}_{-j}) > d_2^2.$$

References

[1] E. Anderson, The Irises of the Gaspe Peninsula. *Bulletin of the American Iris Society* **59** (1935) 2–5.

[2] M. Ankerst, S. Berchtold and D. A. Keim, Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data. *Proceedings of the IEEE Symposium on Information Visualization, InfoVis '98,* (1998) 52–60.

[3] J. M. Chambers and T. J. Hastie, *Statistical Models in S* (Wadsworth and Brooks/Cole, Pacific Grove CA, 1992).

[4] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods* (Academic Press, New York 1982).

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

[6] DAVIS Home Page, `http://stat.skku.ac.kr/myhuh/DAVIS.html`/ (2007).

[7] A. Gifi, *Nonlinear Multivariate Analysis* (Wiley, Chichester 1990).

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

[9] C. B. Hurley, Clustering Visualizations of Multidimensional Data, *Journal of Computational and Graphical Statistics* **13** (2004) 788–806.

[10] A. Inselberg, The Plane with Parallel Coordinates, *The Visual Computer* **1** (1985) 69–91.

[11] M. Kouchi, M. Mochimaru, H. Iwasawa and S. Mitani, Anthropometric Database for Japanese Population 1997-98, Japanese Industrial Standards Center in Agency of Industrial Science and Technology, Ministry of International Trade and Industry (2000).

[12] G. Michailidis and J. de Leeuw, Data Visualization Through Graph Drawing. *Computational Statistics* **16** (2001) 435–450.

[13] LAPACK Home Page, `http://www.netlib.org/lapack/` (2006).

[14] K. B. Pratt and G. Tschapek, Visualizing Concept Drift, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003) 735–740.

[15] C.R. Rao and S.K. Mitra, *Generalized Inverse of Matrices and its Applications* (Wiley, New York, 1971).

[16] G. E. Rosario, E. A. Rundensteiner, D. C. Brown and M. O. Ward, Mapping Nominal Values to Numbers for Effective Visualization, *Proceeding of Information Visualization* (2003) 113-120.

[17] S-Plus, Insightful Corporation Home Page, `http://www.insightful.com/` (2006).

[18] M. Theus, Interactive Data Visualization Using Mondrian. *Journal of Statistical Software* **7** (2002).

[19] A. Unwin, C. Volinsky and S. Winkler, Parallel Coordinates for Exploratory Modeling Analysis, *Computational Statistics & Data Analysis* **43** (2003) 553–564.

[20] E. Wegman, Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association* **85** (1990) 664–675.

[21] 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.

[22] Yang Jing, Peng Wei, M. O. Ward and E. A. Rundensteiner, Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets, *Proceedings of Information Visualization* (2003) 105-112.