Parallel Coordinate Plots of Mixed-Type Data

Il Youp Kwak\(^1\), Myung-Hoe Huh\(^2\)

Abstract

Parallel coordinate plot of Inselberg (1985) is useful for visualizing dozens of variables, but so far the plot’s applicability is limited to the variables of numerical type. The aim of this study is to extend the parallel coordinate plot so that it can accommodate both numerical and categorical variables. We combine Hayashi’s (1950, 1952) quantification method of categorical variables and Hurley’s (2004) endlink algorithm of ordering variables for the parallel coordinate plot. In line with our former study (Kwak and Huh, 2008), we develop Andrews’ type modification of conventional straight-lines parallel coordinate plot to visualize the mixed-type data.

Keywords: Statistical graphics; parallel coordinate plot; Andrews’ plot; mixed-type data; endlink algorithm; Hayashi’s quantification methods.

1. Background and Aim

For the dataset with \( p (\geq 3) \) numerical variables, we use more often parallel coordinate plot (PCP) of Inselberg (1985) because of its compactness. Compared with \( p \times p \) scatterplot matrix, conventional PCP contains only \( p - 1 \) diagrams. In PCP, we may rearrange the order of variables to ease the data exploration, using Hurley’s (2004) endlink algorithm which joins the nearest endpoints of ordered clusters. For instance, suppose that the inter-distances among five objects (variables) \( X_1 \) to \( X_5 \) are specified as follows:

\[
D = \begin{pmatrix}
0.0 & 0.3 & 0.1 & 0.7 & 0.4 \\
0.3 & 0.0 & 0.6 & 1.0 & 0.9 \\
0.1 & 0.6 & 0.0 & 0.8 & 0.5 \\
0.7 & 1.0 & 0.8 & 0.0 & 0.2 \\
0.4 & 0.9 & 0.5 & 0.2 & 0.0
\end{pmatrix}
\]

Then, \( X_1 \) and \( X_3 \) are joined firstly, since \( d_{13} \) is the smallest. Secondly, \( X_4 \) and \( X_5 \) are joined, since \( d_{45} \) is the second smallest. Thirdly, \( X_1 \) and \( X_2 \) are joined to yield the ordered sequence of \( X_2-X_1-X_3 \). Fourthly, \( X_1 \) and \( X_5 \) are selected for the next join, but the pair is not acceptable since \( X_1 \) is not an endpoint of the sequences. Instead, \( X_3 \) and \( X_5 \) are joined. Hence \( X_2-X_1-X_3-X_5-X_4 \) forms the completed list.

\(^1\) Graduate Student in Master’s Course, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: iykwak@korea.ac.kr

\(^2\) Professor, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea. Correspondence: stat420@korea.ac.kr
Applicability of the endlink algorithm to PCP, however, is limited to the dataset or subset data consisting of numerical variables. The aim of this study is to extend the parallel coordinate plot so that it can accommodate both numerical and categorical variables.

Simply consider the dataset of \( n \) observations containing two numerical and two categorical variables, \( X_1, X_2, V_1, V_2 \). We assume that two numerical variables \( X_1 \) and \( X_2 \) are given in standardized forms, \( x_1 \) and \( x_2 \), with mean 0 and standard deviation (SD) 1. For the variable \( V_1 \) with \( k_1 \) categories, we assume it is represented in dummy coding matrix \( Z_1 \) with \( k_1 \) columns, one column for each category. Similarly, we represent the variable \( V_2 \) with \( k_2 \) categories by the dummy coding matrix \( Z_2 \) with \( k_2 \) columns.

There are three types in the pair of two variables: numerical-numerical, numerical-categorical (or categorical-numerical) and categorical-categorical. For numerical-numerical pair, the association of two variables is measured by Pearson’s product moment correlation and the data points on the parallel axis are linked as in conventional PCP. For two other types of pairs, we apply Hayashi’s (1950, 1952) quantification methods to quantify categories and measure the association between variables as follows.

For numerical-categorical pair, say \( X_1 \) and \( V_1 \) with \( k_1 \) categories, Hayashi’s quantification can be formulated as

\[
\max_{a_1} \frac{x_1 Z_1 a_1}{n - 1} \quad \text{subject to } a_1^t Z_1^t Z_1 a_1 / (n - 1) = 1 \text{ and } 1_n^t Z_1 a_1 = 0,
\]

where \( x_1 \) is \( n \times 1 \), \( Z_1 \) is \( n \times k_1 \), \( a_1 \) is the \( k_1 \times 1 \) vector of quantified values for \( k_1 \) categories in \( V_1 \) and \( 1_n \) is the \( n \times 1 \) vector of elements all equal to 1. The second restriction in (1.1) requires that the \( n \times 1 \) quantified vector \( Z_1 a_1 \) of \( V_1 \) should have mean 0 and the first restriction together with the second restriction requires that \( Z_1 a_1 \) should have SD 1. By Lagrangian multiplier method, one can easily show that

\[
a_1 = \frac{D_1^{-1} Z_1^t x_1}{\left( x_1^t Z_1 D_1^{-1} Z_1^t x_1 / n - 1 \right)^{\frac{1}{2}}},
\]

where \( D_1 = Z_1^t Z_1 \) is the \( k_1 \times k_1 \) diagonal matrix, with diagonal elements equal to the observed frequencies of respective categories in \( V_1 \). The optimized value of (1.1) is equal to Pearson’s correlation between \( X_1 \) and \( Z_1 a_1 \), the quantified variable of \( V_1 \). This method, known as Hayashi’s Quantification Method I or II in Japan (Huh, 1999), can be considered as regression analysis on dummy variables or a special case of canonical correlation analysis.

For categorical-categorical pair, say \( V_1 \) with \( k_1 \) categories and \( V_2 \) with \( k_2 \) categories, Hayashi’s quantification can be formulated as

\[
\max_{a_1, a_2} \frac{a_1^t Z_1^t Z_2 a_2}{n - 1} \quad \text{subject to } a_1^t Z_1^t Z_1 a_1 / (n - 1) = 1, \quad 1_n^t Z_1 a_1 = 0,
\text{ and } a_2^t Z_2^t Z_2 a_2 / (n - 1) = 1, \quad 1_n^t Z_2 a_2 = 0,
\]
where $Z_1$ is $n \times k_1$, $Z_2$ is $n \times k_2$, $a_1$ and $a_2$, respectively, are the $k_1 \times 1$ and $k_2 \times 1$ vectors of quantification values for $k_1$ categories in $V_1$ and for $k_2$ categories in $V_2$. It is well known that $a_1$ and $a_2$ can be obtained via the singular value decomposition of

$$G = D_1^{-1} Z_1' Z_2 D_2^{-1},$$

where $D_1 = Z_1' Z_1$ and $D_2 = Z_2' Z_2$. More specifically, the solution vectors $a_1$ and $a_2$ of (2.1) are given by

$$a_1 = \left( \frac{D_1}{n-1} \right)^{-\frac{1}{2}} u_1 \quad \text{and} \quad a_2 = \left( \frac{D_2}{n-1} \right)^{-\frac{1}{2}} u_2,$$

where $u_1$ and $u_2$ are left and right singular vectors of $k_1 \times k_2$ matrix $G$ corresponding to the largest singular value except the trivial root. The optimized value of (2.1) is equal to Pearson’s correlation between $Z_1 a_1$ and $Z_2 a_2$, the quantified variables of $V_1$ and $V_2$, respectively. This method, known as Hayashi’s Quantification Method III (Huh, 1999), is closely connected to correspondence analysis (Huh, 1989). There appeared several papers on correspondence analysis in Korean journals (Yang and Huh, 1999; Choi and Huh, 1999; Choi et al., 2005).

In that way, we may determine the correlation between any types of variable. In the next section, we will propose a PCP for mixed-type data via Hurley’s endlink algorithm, sequentially applying Hayashi’s quantification methods to categorical variables.

2. Modification of Endlink Algorithm

Parallel coordinate plot appears differently depending on the order of variables. For the purpose, we want to use Hurley’s (2004) endlink algorithm which connects the nearest endpoints of ordered sequence of variables. The problem is that the distances between pairs of variables are not readily available in the case of mixed-type data.

We propose a modified version of Hurley’s (2004) endlink algorithm to determine the order variables of numerical and/or categorical type in the parallel coordinate plot. Suppose that there are variables of numerical and/or categorical type.

Step 1: We make a $p \times p$ correlation matrix $R$ among variables of numerical and/or categorical type. For the pair of variables of which at least one variable is not numerical, we use Hayashi’s quantification methods to acquire the correlation coefficient. From $R = \{r_{ij}\}$, we derive the distance matrix $D = \{d_{ij}\}$ by

$$d_{ij} = 2(1 - r_{ij}), \quad \text{for} \quad i, j = 1, \ldots, p.$$

Step 2: Join the closest ends of chained variables. If all variables are chained to form a single cluster, then stop.

Step 3: If any variable of newly joined pair is categorical, replace its categorical codes by the quantified values related to the counter variable and change the variable type from categorical to numerical. Return to Step 1.
We will illustrate our algorithm by two scenarios for the simulated dataset in which two variables \((X_1 \text{ and } X_2)\) are numerical and two variables \((V_1 \text{ and } V_2)\) are categorical.

Scenario 1.

Cycle 1: \(V_1\) and \(V_2\) are quantified related to \(X_1\) and \(X_2\), all separately. Also, \(V_1\) and \(V_2\) are mutually quantified. The pair of \(X_1\) and \(V_1\) is selected. \(V_1\) is replaced by quantified values \(\tilde{V}_1\) related to \(X_1\) and the variable type is changed to numerical. We have a chain of \(\tilde{V}_1 - X_1\).

Cycle 2: \(V_2\) is quantified related to \(X_1, X_2\) and \(\tilde{V}_1\), all separately. The pair of \(X_1\) and \(X_2\) is selected. Thus we have a chain of \(\tilde{V}_1 - X_1 - X_2\).

Cycle 3: \(V_2\) is quantified related to \(X_2\) and \(\tilde{V}_1\), all separately. The pair of \(X_2\) and \(V_2\) is selected. Categorical \(V_2\) is quantified with respect to \(X_2\), turned into numerical \(\tilde{V}_2\). Thus we have a chain of \(\tilde{V}_1 - X_1 - X_2 - \tilde{V}_2\).

Scenario 2.

Cycle 1: \(V_1\) and \(V_2\) are quantified related to \(X_1\) and \(X_2\), all separately. Also, \(V_1\) and \(V_2\) are mutually quantified. The pair of \(V_1\) and \(V_2\) is selected. \(V_1\) and \(V_2\) are replaced by quantified values \(\tilde{V}_1\) and \(\tilde{V}_2\) and the variable type is changed to numerical. We have a chain of \(\tilde{V}_1 - \tilde{V}_2\).

Cycle 2: The pair of \(X_1\) and \(\tilde{V}_1\) is selected. Thus we have a chain of \(X_1 - \tilde{V}_1 - \tilde{V}_2\).

Cycle 3: The pair of \(X_2\) and \(X_1\) is selected. Thus we have a chain of \(X_2 - X_1 - \tilde{V}_1 - \tilde{V}_2\).

In Scenario 1, we simulated 100\((=n)\) observations of \((X_1, X_2, X_3, X_4)\) from a multivariate normal distribution with the zero means and the covariance matrix

\[
\Sigma_1 = \begin{pmatrix}
1.0 & 0.6 & 0.8 & 0.0 \\
0.6 & 1.0 & 0.2 & 0.4 \\
0.8 & 0.2 & 1.0 & 0.1 \\
0.0 & 0.4 & 0.1 & 1.0
\end{pmatrix}.
\]

Then, \((X_3, X_4)\) are discretized into categorical variables \((V_1, V_2)\) via

\[
V_1 = \begin{cases}
1, & \text{if } X_3 \leq -1.5, \\
2, & \text{if } -1.5 < X_3 \leq -0.5, \\
3, & \text{if } -0.5 < X_3 \leq 0.5, \\
4, & \text{if } 0.5 < X_3 \leq 1.5, \\
5, & \text{if } X_3 > 1.5,
\end{cases}
\quad V_2 = \begin{cases}
1, & \text{if } X_4 \leq -1, \\
2, & \text{if } -1 < X_4 \leq 1, \\
5, & \text{if } X_4 > 1.
\end{cases}
\tag{2.1}
\]

Running our algorithm, \(V_1\) and \(V_2\) are quantified to

\[
\tilde{V}_1 = \begin{pmatrix}
-2.28 \\
-1.15 \\
0.06 \\
0.93 \\
1.64
\end{pmatrix}, \quad \tilde{V}_2 = \begin{pmatrix}
-1.88 \\
-0.07 \\
1.57
\end{pmatrix}
\]
and we have the ordered cluster $\wedge_1 - X_1 - X_2 - \wedge_2$. Figure 2.1 shows the PCP (left) and Andrews’ type PCP (right). Andrews’ type PCP or APCP is the Andrews’ plot (Andrews, 1972) for the orthogonal-transformed dataset, so that the variables appear in the designated order (Kwak and Huh, 2008). In Scenario 2, we simulated 100(=n) observations of $(X_1, X_2, X_3, X_4)$ from a multivariate normal distribution with the zero means and the covariance matrix

$$
\Sigma_2 = \begin{pmatrix}
1.0 & 0.4 & 0.6 & 0.2 \\
0.4 & 1.0 & 0.0 & 0.1 \\
0.6 & 0.0 & 1.0 & 0.8 \\
0.2 & 0.1 & 0.8 & 1.0
\end{pmatrix}.
$$

Then, $(X_3, X_4)$ are discretized into categorical variables $(V_1, V_2)$ according to (2.1).
Running our algorithm, $V_1$ and $V_2$ are quantified to

$$\tilde{V}_1 = \begin{pmatrix} -3.02 \\ -0.98 \\ -0.01 \\ 0.88 \\ 1.70 \end{pmatrix}, \quad \tilde{V}_2 = \begin{pmatrix} -2.00 \\ 0.08 \\ 1.55 \end{pmatrix},$$

and we have the ordered cluster $X_2 - X_1 - \tilde{V}_1 - \tilde{V}_2$. Figure 2.2 shows the PCP (left) and Andrews’ type PCP (right).

3. Cars93 Data

Cars93 data, available at R’s **MASS** library, consists of 93 records on automobile models. Among 27 characteristics available for each automobile, we included 20 variables for analysis: (Hereafter, categorical variables are underlined) Type, AirBag, DriverTrain, Cylinders, EngineSize, Man.trans.avail, Fuel.tank.capacity, Passengers, Length, Wheelbase, Width, Weight, Origin, MPG.city, MPG.highway, Horsepower, RPM, Rev.per.mile, Turn.circle and Price. We omitted one record which has non-numerical value on Cylinders.

Figure 3.1 shows APCP of Cars93 data set. The individual curves are colored by Price (light color for low price and dark color for high price). In the plot, we may find the categorical variable Type, quantified to −1.21 for “Small”, −0.93 for “Sporty”, −0.21 for “Compact”, 0.55 for “Midsize”, 1.37 for “Van” and 1.50 for “Large”, are located between two numerical variables, Fuel.tank.capacity and Wheelbase. Average Fuel.tank.capacity by Type are $-1.22, -0.30, -0.17, 0.56, 1.32, 0.75$, while average Wheelbase by Type are $-1.10, -0.84, -0.19, 0.50, 1.24, 1.36$ (in standardized unit).
Figure 4.1: APCP’s for German Credit Data: Good credit cases as reference (Upper), Contrasted to bad credit cases as supplementary observations (Lower).

We clearly see that \texttt{MPG.highway}, \texttt{MPG.city}, and \texttt{Rev.per.mile} form one cluster of variables with inter-correlations 0.94, 0.70 and \texttt{Price}, \texttt{Horsepower}, \texttt{Cylinders}, \texttt{EngineSize}, \texttt{Width}, \texttt{Weight}, \texttt{Fuel.tank.capacity}, \texttt{Type}, \texttt{Wheelbase}, \texttt{Length}, and \texttt{Turn.circle} form another group with inter-correlations 0.78, 0.79, 0.69, 0.87, 0.88, 0.90, 0.82, 0.90, 0.82, 0.74.

4. German Credit Data

German Credit data, available at \texttt{http://mlearn.ics.uci.edu/MLSummary.html}, contains financial and socio-demographic information on 1000\(=n\) individuals. Number of measured variables are 20\(=p\) except the classification code for credit outcome (good/bad). Among the variables, seven variables are numerical and the remaining thirteen variables are categorical.
Figure 4.2: Mean curve and its 95% confidence limits of bad credit cases: Derived from the lower plot of Figure 4.1

The upper APCP of Figure 4.1 shows the good credit cases as reference. In contrast, the lower APCP shows bad credit cases as supplementary observations. Overall features of two plots are not lucid, so we draw the mean curve and plot its 95% confidence limits of bad credit cases separately in Figure 4.2.

In Figure 4.2, we can see that bad credit individuals tend to gather at

1) negative values of $X_6$ (savings: Category 1 = -0.89, 4 = 0.63, 2 = 1.01, 3 = 1.21, 5 = 1.23) and $X_1$ (checking: 1 = -1.86, 3 = -0.76, 2 = 0.28, 4 = 0.72),

2) positive values of $X_{20}$ (foreign worker: 2 = -4.49, 1 = 0.22), $X_2$ (duration) and $X_5$ (amount),

3) positive values of $X_{19}$ (telephone: 1 = -0.84, 2 = 1.18), $X_{17}$ (job: 2 = -0.41, 3 = -0.34, 4 = 1.33, 1 = 5.48) and $X_7$ (employment: 3 = -0.41, 4 = -0.22, 2 = -0.04, 5 = -0.15, 1 = 4.08),

4) positive values of $X_{15}$ (housing: 2 = -0.35, 1 = -0.16, 3 = 3.14),

5) negative values of $X_{13}$ (age),

6) negative values of $X_{16}$ (number of credits) and $X_3$ (history: 1 = -0.91, 2 = -0.90, 0 = 0.67, 3 = 0.67, 4 = 1.21).

Thus bad credit individuals can be typified by

1) savings ($X_6$) less than 100 and checking ($X_1$) < 0,

2) foreign worker ($X_{20}$), large duration ($X_2$) and larger amount ($X_5$),

3) telephone owner ($X_{19}$), manager/self-employed/qualified employee/officer ($X_{17}$) and unemployed ($X_7$),
4) free house (X15),

5) young (X13),

6) small number of credits (X16) and all credits paid back duly/existing credits paid duly till now (X3).

In that way, we see the difference between two groups of individuals with additional information on the clustered list of variables carrying the disparity.

5. Concluding Remark

This study is aimed to represent the mixed type data on PCP. Combining and modifying Hayashi’s quantification method of categorical variables and Hurley’s endlink algorithm for ordering variables, we made a PCP and its variation for mixed type data. Usefulness of proposed graphs are demonstrated via two real datasets, Cars93 and German Credit data.

References


[Received May 2008, Accepted June 2008]