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. Thanks to Hurley's (2004) endlink algorithm, we may re-arrange the order of variables in PCP to ease the data exploration. PCP's applicability, 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.

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.

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, Dept. of Statistics, Korea University. Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea. Correspondence: stat420@korea.ac.kr

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

$$\max_{a_1}\ x_1^t Z_1 a_1 \,/\, (n-1) \tag{1.1}$$
$$\text{subject to } a_1^t Z_1^t Z_1 a_1 \,/\, (n-1) = 1 \ \text{ and } \ 1_n^t Z_1 a_1 = 0,$$

where 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

$$a_1 = D_1^{-1} Z_1^t x_1 \,/\, (x_1^t Z_1 D_1^{-1} Z_1^t x_1 \,/\, (n-1))^{1/2},$$

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 is known as Hayashi's Quantification Method I or II (Huh, 1999).

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

$$\max_{a_1, a_2}\ a_1^t Z_1^t Z_2 a_2 \,/\, (n-1)$$
$$\text{subject to } \ a_1^t Z_1^t Z_1 a_1 \,/\, (n-1) = 1 \ , 1_n^t Z_1 a_1 = 0, \tag{1.2}$$
$$\text{and } \ a_2^t Z_2^t Z_2 a_2 \,/\, (n-1) = 1 \ , 1_n^t Z_2 a_2 = 0,$$

where 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

$$G = D_1^{-1/2} Z_1^t Z_2 D_2^{-1/2}.$$

More specifically, the solution vectors a_1 and a_2 of (1.2) are given by

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

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 (1.2) 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 is known as Hayashi's Quantification Method III (Huh, 1999).

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 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 joins the nearest endpoints of ordered clusters. 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

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

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 a scenario for the simulated dataset in which two variables (X_1 and X_2) are numerical and two variables (V_1 and V_2) are categorical.

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 $\tilde{V}_1 - X_1 - X_2 - \tilde{V}_2$.

In the above scenario, we simulated for $100(=n)$ observations of (X_1, X_2, X_3, X_4) from a multivariate normal distribution with the zero means and the covariance matrix

$$\Sigma = \begin{pmatrix} 1.0 & 0.6 & 0.9 & 0.0 \\ 0.6 & 1.0 & 0.2 & 0.4 \\ 0.9 & 0.2 & 1.0 & 0.1 \\ 0.0 & 0.4 & 0.1 & 1.0 \end{pmatrix} .$$

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

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

Running our algorithm, V_1 and V_2 are quantified to

$$\tilde{V_1} = \begin{pmatrix} -2.15 \\ -1.04 \\ 0.04 \\ 0.99 \\ 1.79 \end{pmatrix}, \quad \tilde{V_2} = \begin{pmatrix} -1.59 \\ -0.16 \\ 2.28 \end{pmatrix},$$

and we have the ordered cluster $\tilde{V_1} - X1 - X2 - \tilde{V_2}$. 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 appears in the designated order (Kwak and Huh, 2008).

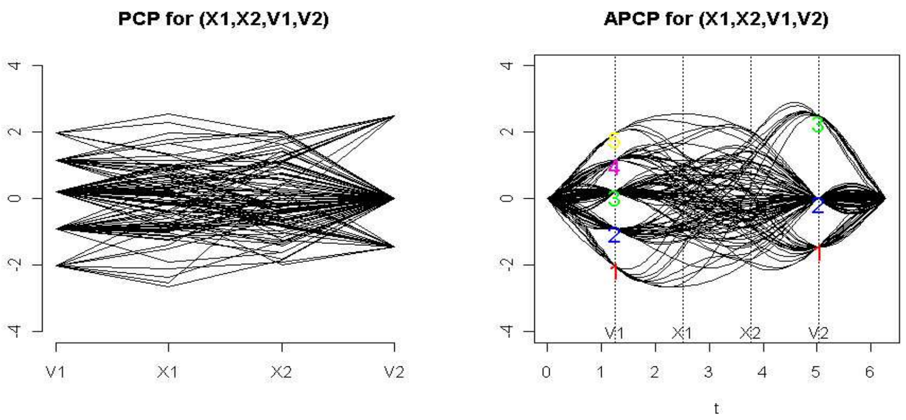

Figure 2.1: PCP[left] and APCP[right] for the simulated dataset of four variables.

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) <u>`Type`</u>, <u>`AirBags`</u>, <u>`DriverTrain`</u>, `Cylinders`, `EngineSize`, <u>`Man.trans.avail`</u>, `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).

We clearly see that `MPG.highway`, `MPG.city`, and `Rev.per.mile` form one cluster of variables with inter-correlations 0.94, 0.70 and `Price`, `Horsepower`, `Cylinders`, `EngineSize`, `Width`, `Weight`, `Fuel.tank.capacity`, `Type`, `Wheelbase`, `Length`, and `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.

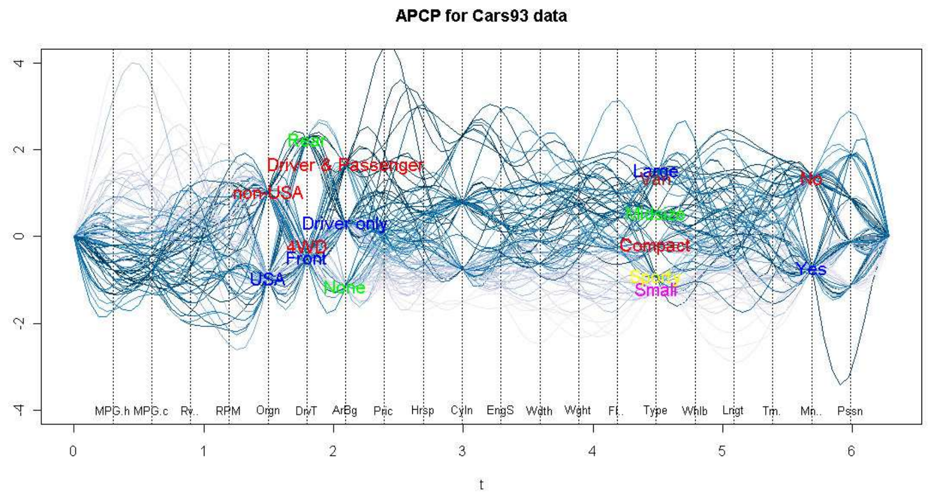

Figure 3.1: APCP for Cars93 Data.

4. German Credit Data

German Credit data, available at *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. 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

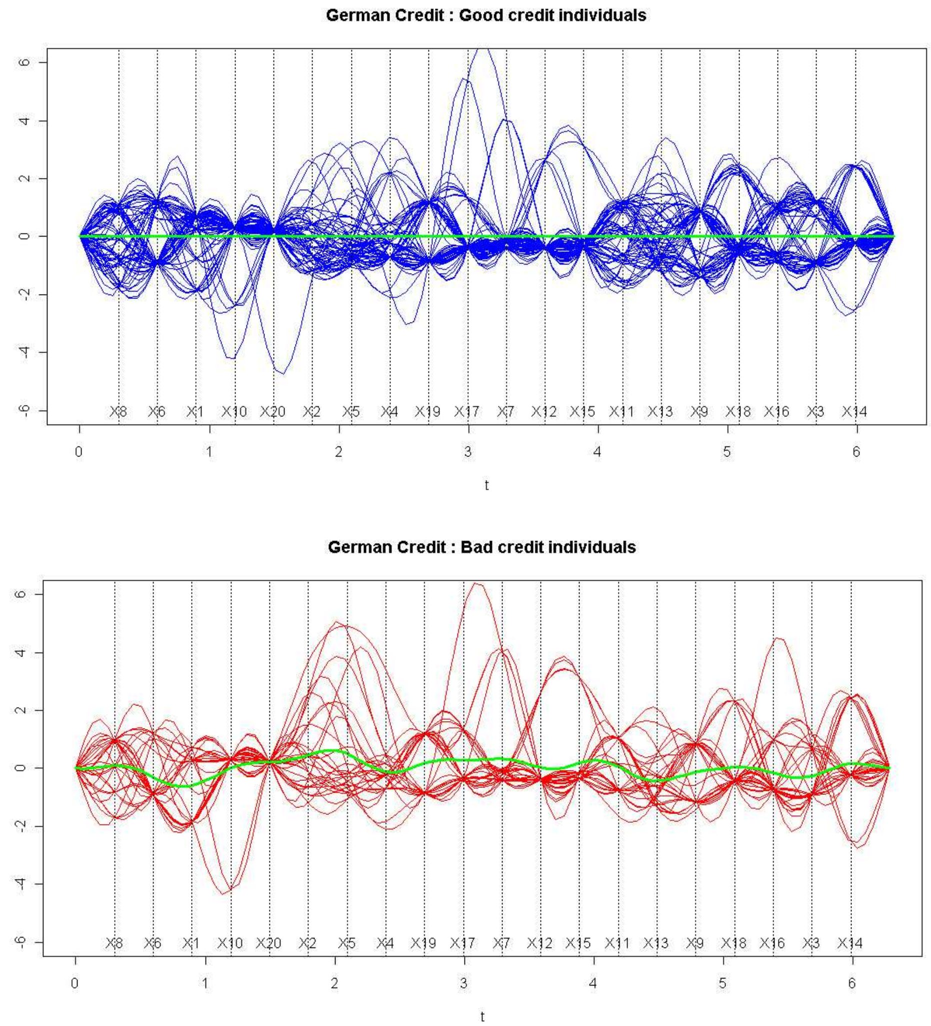

Figure 4.1: APCP's for German Credit Data: Good credit cases as reference [Upper], Contrasted to bad credit cases as supplementary observations [Lower].

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 $X6$ (savings: Category 1=-0.89, 4=0.63, 2=1.01, 3=1.21, 5=1.23) and $X1$ (checking: 1=-1.86, 3=-0.76, 2=0.28, 4=0.72),
2) positive values of $X20$ (foreign worker: 2=-4.49, 1=0.22), $X2$ (duration) and $X5$ (amount),
3) positive values of $X19$ (telephone: 1=-0.84, 2=1.18), $X17$ (job: 2=-0.41, 3=-0.34,

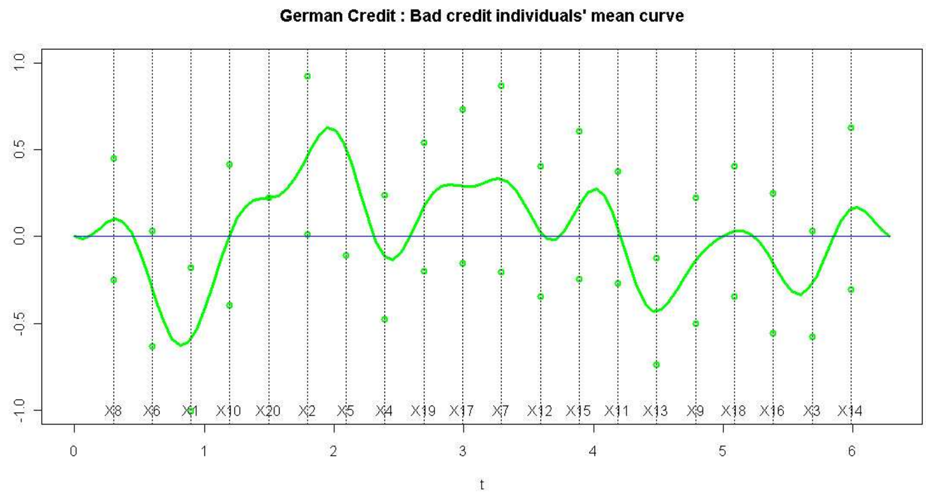

Figure 4.2: Mean curve and its 95% confidence limits of bad credit cases: Derived from the lower plot of Figure 4.1

4=1.33, 1=5.48) and $X7$ (employment: 3=-0.41, 4=-0.22, 2=-0.04, 5=-0.15, 1=4.08),

4) positive values of $X15$ (housing: 2=-0.35, 1=-0.16, 3=3.14),
5) negative values of $X13$ (age),
6) negative values of $X16$ (number of credits) and $X3$ (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 ($X6$) less than 100 and checking ($X1$) < 0,
2) foreign worker ($X20$), large duration and larger amount,
3) telephone owner ($X19$), manager/self-employed/qualified employee/officer ($X17$), and unemployed ($X7$),
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 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

Andrews, D. F. (1972). Plots of high-dimensional data, *Biometrics,* **85**, 125–136.

Hayashi, C. (1950). On the quantification of qualitative data from the mathematico-statistical point of view, *Annals of Institute of Statistical Mathematics,* **2(1)**, 35–47.

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view, *Annals of Institute of Statistical Mathematics,* **3(2)**, 69–98.

Huh, M.H. (1999). *Quantification Methods for Multivariate Data,* Freedom Academy, Seoul (written in Korean).

Hurley, C.B. (2004). Clustering visualizations of multidimensional data, *Journal of Computational and Graphical Statistics,* **13**, 788–806.

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

Kwak, I.Y. and Huh, M.H. (2008). Andrews' plot for extended uses, *Communications of The Korean Statistical Society,* **15**, 87–94.