Journal of the Korean
Statistical Society
Vol. 24, No. 2, 1995

Resistant h-Plot for a Sample Variance-Covariance Matrix [†]

Yong-Seok Choi[1]

ABSTRACT

The h-plot is a graphical technique for displaying the structure of one population's variance-covariance matrix. This follows the mathematical algorithm of the principal component biplot based on the singular value decomposition. But it is known that the singular value decomposition is not resistant, i.e., it is very sensitive to small changes in the input data. In this article, since the mathematical algorithm of the h-plot is equivalent to that of principal component biplot, using the algorithm for the resistant principal component biplot of Choi and Huh (1994), we derive the resistant h-plot.

KEYWORDS: h-plot, Principal component biplot, Resistant, Singular value decomposition.

1. INTRODUCTION

The central idea of many multivariate analyses is dimension reduction. The dimension reduction is easily given by the singular value decomposition which

[†] This paper was supported(in part) by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1993.

[1] Department of Statistics, Pusan National University, Pusan, 609-735, Korea.
E-mail: yschoi@hyowon.pusan.ac.kr

is one of the most useful methods in the areas of matrix computation. Traditionally, the eigensystem is used for dimension reduction in many multivariate analyses. As a practical matter, however, there are reasons for preferring the use of the singular value decomposition (Belsley, Kuh and Welsch, 1980, p. 99). Some multivariate analyses (principal component biplot, correspondence analysis and principal factor analysis) are similar to principal component analysis which is a representative method for dimension reduction based on it.

Specially, since the h-plot follows the mathematical algorithm of the principal component biplot of Gabriel (1971), this is also based on the singular value decomposition. But it is known that the singular value decomposition of the data matrix is not resistant, i.e., it is very sensitive to small changes in the input data (Choi and Huh, 1994). As some multivariate analyses (principal component biplot, correspondence analysis and principal factor analysis), the h-plot based on it is influenced by outliers in data matrix and does not give stable graphical techniques.

In Section 2, we briefly provide the classical h-plot based on the singular value decomposition. In Section 3, we provide the resistant version of h-plot using the algorithm for the resistant singular value decomposition in Choi and Huh (1994, Theorem). We call this the resistant h-plot. Finally, in Section 4, we give two numerical illustrations with discussions.

2. THE CLASSICAL h-PLOT: BASED ON THE SINGULAR VALUE DECOMPOSITION

Consider the $n \times p$ variables-centered data matrix $\widetilde{\boldsymbol{X}}$ such as $\widetilde{\boldsymbol{X}} = (x_{ij} - \bar{x}_{.j})$ (with $\bar{x}_{.j} = \sum_i x_{ij}/n, \quad i = 1, \dots, n; j = 1, \dots, p$). In biplot and principal component biplot, this kind of centering is usually adopted (Bradu and Gabrel, 1978; Jolliffe, 1986, p. 86). Note that a p-variate sample variance-covariance matrix $\boldsymbol{S}$ is given by

$$\boldsymbol{S} = \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}/n, \tag{2.1}$$

The singular value decomposition of matrix $\widetilde{\boldsymbol{X}}$ with rank r can be written

$$\widetilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{D}_\lambda\boldsymbol{V}', \tag{2.2}$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$ are $n \times r$, $p \times r$ matrices with orthogonal columns $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$, $k = 1, \ldots, r$, respectively and $\boldsymbol{D}_\lambda = diag(\lambda_1, \ldots, \lambda_r)$ with singular values $\lambda_1 \geq \cdots \geq \lambda_r$. From (2.1) and (2.2), we have

$$n\boldsymbol{S}\boldsymbol{v}_k = \lambda_k^2\boldsymbol{v}_k, k = 1, \ldots, r. \tag{2.3}$$

This is the usual form of the traditional eigensystem for principal component analysis except the factor n. So, the singular value decomposition of (2.2) and the traditional eigensystem of (2.3) are main approaches for principal component biplot (Gabriel, 1971; Jolliffe, 1986, pp. 75-77; Seber, 1984, pp. 207-208).

Now consider the construction of the classical h-plot. In Corsten and Gabriel (1976), they provided the construction of 2-dimensional h-plot for a matrix $\boldsymbol{S}$ based on the eigensystem (2.3). However, we note that in operating directly on the $n \times p$ matrix $\widetilde{\boldsymbol{X}}$, the singular value decomposition avoids the additional computation burden of forming $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}$ in (2.1) (Belsley, Kuh and Welsch, 1980, p. 99).

Therefore we provide the construction of 2-dimensional h-plot based on the singular value decomposition of (2.2) as:

Step 1: We obtain the largest two singular values, λ_1 and λ_2, and the corresponding the right singular vectors, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

Step 2: We compute the $p \times 2$ matrix

$$\boldsymbol{H} = n^{-1/2}(\lambda_1\boldsymbol{v}_1, \lambda_2\boldsymbol{v}_2). \tag{2.4}$$

Step 3: we have the 2-dimensional h-plot with coordinates providing by rows of $\boldsymbol{H}$.

And as a measure of the quality for this 2-dimensional h-plot, Gabriel (1971) provides

$$\rho_2^{(4)} = (\lambda_1^4 + \lambda_2^4) / \sum_{k=1}^{r} \lambda_k^4,$$

where λ_i, $i = 1, 2$ and r are noted in (2.2). He calls this the goodness of fit.

3. THE RESISTANT h-PLOT: BASED ON THE RESISTANT SINGULAR VALUE DECOMPOSITION

3.1 Construction of the Resistant h-Plot

We know that in previous section, the matrix $\boldsymbol{H}$ of (2.4) can be simply obtained from the singular value decomposition (2.2).

However, it is well known that the sample mean as location estimator is not resistant, i.e., it is very sensitive to small changes in the input data matrix. The sample mean which is not resistant influences the variable-centered matrix $\widetilde{\boldsymbol{X}}$. Also this influences the sample variance-covariance matrix $\boldsymbol{S}$ in (2.1).

So as in principal component biplot, both approaches (traditional eigensystem (2.3) and singular value decomposition (2.2)) for h-plot are not resistant. Thus this h-plot is influenced by outlying observations and then is not resistant.

Now Choi and Huh (1994, Theorem) provided the resistant singular value decomposition of an $n \times p$ data matrix $\widetilde{\boldsymbol{X}}^*$ of rank r centered at a robust location estimate. And we used the median scale estimator. Calculation of the resistant singular value decomposition can be done using the iterative procedure with Andrew's $\psi(\cdot)$ function given by

$$\psi(t) = \begin{cases} c\sin(t/c), & \text{for } 0 \le t < c\pi, \\ 0, & \text{for } t \ge c\pi. \end{cases}$$

Actually, $(c\pi)^2$ is 95 percentile point of χ^2 distribution with $p - s$ degrees of freedom. As discussed in Section 2, for the 2-dimensional h-plot, we must take $s = 2$.

We note that the resistant singular value decomposition can be written by

$$\widetilde{\boldsymbol{X}}^* = \boldsymbol{U}\boldsymbol{D}_{\lambda^*}\boldsymbol{V}', \tag{3.1}$$

where $\boldsymbol{U}$ is an $n \times r$ matrix such that $\boldsymbol{U}'\boldsymbol{D}_w\boldsymbol{U} = \boldsymbol{I}_r$, $\boldsymbol{V}$ is a $p \times r$ matrix of eigenvectors $\widetilde{\boldsymbol{X}}^{*\prime}\boldsymbol{D}_w\widetilde{\boldsymbol{X}}^*$ such that $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_r$, and $\boldsymbol{D}_{\lambda^*} = \text{diag}(\lambda_1{}^*, \ldots, \lambda_r{}^*)$ with λ_k^{*2} is the k-th eigenvalue of $\widetilde{\boldsymbol{X}}^{*\prime}\boldsymbol{D}_w\widetilde{\boldsymbol{X}}^*$.

In fact, $\boldsymbol{D}_w = \text{diag}(w_1, \ldots, w_n)$ is an $n \times n$ diagonal matrix with the diagonal elements $w_i = \psi(\|\tilde{\boldsymbol{x}}_i{}^* - \hat{\boldsymbol{x}}_i{}^*\|/\hat{\sigma})/(\|\tilde{\boldsymbol{x}}_i{}^* - \hat{\boldsymbol{x}}_i{}^*\|/\hat{\sigma})$, $i = 1, \ldots, n$. The calculation of w_i can be done using the iterative procedure of Choi and Huh (1994). Here $\tilde{\boldsymbol{x}}_i{}^*$ denotes the i-th row of $\widetilde{\boldsymbol{X}}^*$ and can be viewed as n points in a p-dimensional space $\mathcal{R}^p$. Let $\hat{\boldsymbol{x}}_i{}^*$ in a subspace of dimension $s(1 \leq s \leq p)$ of $\mathcal{R}^p$ be the nearest point of an arbitrary point $\tilde{\boldsymbol{x}}_i{}^*$ in $\mathcal{R}^p$. And we use the median scale estimator $\hat{\sigma} = [\text{med}_i(\|\tilde{\boldsymbol{x}}_i{}^* - \hat{\boldsymbol{x}}_i{}^*\|^2)/\chi^2_{.50(p-s)}]^{1/2}$, $i = 1, \ldots, n$ where $\chi^2_{.50(p-s)}$ is 50 percentile point of χ^2 distribution with $p - s$ degrees of freedom. So we note that in $\boldsymbol{D}_w$, the diagonal elements having zero or nearly zero denote the notable observations in data.

And the resistant singular value decomposition (3.1) gives the weighted variance-covariance matrix

$$\boldsymbol{S}^* = \widetilde{\boldsymbol{X}}^{*\prime}\boldsymbol{D}_w\widetilde{\boldsymbol{X}}^*/n^*, \tag{3.2}$$

where $n^* = \sum_{i=1}^n w_i = \boldsymbol{1}_n'\boldsymbol{D}_w\boldsymbol{1}_n$.

From the (3.2), we obtain the form of the resistant eigensystem

$$n^*\boldsymbol{S}^*\boldsymbol{v}_k = \lambda_k^{*^2}\boldsymbol{v}_k\ , k = 1, \ldots, r. \tag{3.3}$$

Of course, we can provide a construction of a resistant h-plot for $\boldsymbol{S}^*$ based on the resistant eigensystem (3.3). However as a strict analogy with construction of the classical h-plot in Section 2, we can simply provide a construction of a resistant h-plot based on the resistant singular value decomposition (3.1). This proceeds as follows:

Step 1: We obtain the largest two resistant singular values, λ_1^* and λ_2^*, and the corresponding the right resistant singular vectors, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

Step 2: We compute the $p \times 2$ matrix

$$\boldsymbol{H}^* = n^{*-1/2}(\lambda_1^* \boldsymbol{v}_1, \lambda_2^* \boldsymbol{v}_2). \tag{3.4}$$

Step 3: We have an optimal 2-dimensional plot with the coordinates providing by rows of (3.4).

Then this 2-dimensional plot is a resistant version for the classical h-plot. From now, we call this the resistant h-plot for a sample variance-covariance $\boldsymbol{S}$. And with respect to the previous construction of the classical h-plot, we note that a resistant h-plot for a sample variance-covariance matrix $\boldsymbol{S}$ is the h-plot for a weighted sample variance-covariance $\boldsymbol{S}^*$.

3.2 Geometric Interpretations of the Resistant h-Plot

In Subsection 3.1, the matrix $\boldsymbol{H}^*$ of (3.4) can be represented as $\boldsymbol{H}^* = (\boldsymbol{h}_1^*, \ldots, \boldsymbol{h}_p^*)'$ where $\boldsymbol{h}_j^{*'} = n^{*-1/2}(\lambda_1^* v_{j1}, \lambda_2^* v_{j2}), j = 1, \ldots, p$. Thus the rows $\boldsymbol{h}_j^{*'}(j = 1, \ldots, p)$ provide the coordinate for the resistant h-plot and the variable j is represented by an arrow from the origion to its vertex at $(n^{*-1/2}\lambda_1^* v_{j1}, n^{*-1/2}\lambda_2^* v_{j2})$.

From Choi and Huh (1994, Lemma), the resistant approximations of covariances, variances and correlations are given by

$$s_{jk}^* \simeq \boldsymbol{h}_j^{*'}\boldsymbol{h}_k^* \tag{3.5}$$

$$s_j^{*^2} \simeq \|\boldsymbol{h}_j^*\|^2 \tag{3.6}$$

$$r_{jk} \simeq \cos(\theta), \tag{3.7}$$

where $\simeq$ denotes "resistant approximation".

Therefore we have the geometric interpretations of these equations as follows:

(3.5) is the (j, k)-th element of $\boldsymbol{S}^*$.

(3.6) is the squared lenght of arrow $\boldsymbol{h}_j^*$ and resistantly approximates the variance of the j-th variable ($j = 1, \ldots, p$).

(3.7) is the correlation denoting by the cosine of the angle θ between two rows $\boldsymbol{h}_j^*$ and $\boldsymbol{h}_k^*, j \neq k$.

3.3 Goodness of Resistant Fit

We know $\rho_2^{(4)}$ as a measure of an 2-dimensional h-plot. Its algebraic and mathematical calculations are based on the lower rank least squares approximation of Householder and Young (1938).

Now we need a measure of the quality for an optimal 2-dimensional resistant h-plot. Choi and Huh (1994) provided a lower rank resistant approximation. Therefore with using $\boldsymbol{S}^*$ and $\boldsymbol{H}^*\boldsymbol{H}^{*\prime}$ instead of $\widetilde{\boldsymbol{X}}^*$ and $\widetilde{\boldsymbol{X}}^*_{(2)}$, their discussions enoughly lead to the goodness of resistant approximation measures for the resistant h-plot,

$$\begin{aligned}
\rho_2^{*(4)} &= 1 - \| \boldsymbol{S}^* - \boldsymbol{H}^*\boldsymbol{H}^{*\prime} \|^2 / \| \boldsymbol{S}^* \|^2, \\
&= 1 - \sum_{k=3}^{r} (\lambda_k^{*^2})^2 / \sum_{k=1}^{r} (\lambda_k^{*^2})^2, \\
&= \sum_{k=1}^{2} \lambda_k^{*^4} / \sum_{k=1}^{r} \lambda_k^{*^4}.
\end{aligned}$$

We call this a goodness of resistant fit. Note that $\lambda_k^{*^2}$, $\boldsymbol{H}^*$ and $\boldsymbol{S}^*$ are noted in Subsection 3.1.

4. NUMERICAL ILLUSTRATIONS

Example 1. The census-tract data (Johnson and Wichern, 1992, Table 8.2. p. 392) provids 14 tract infomations on 5 socioeconomic variables for the Madison, Wisconsin area. Choi and Huh (1994) applied the resistant principal component analysis to this data.

The 2-dimensional h-plot is given in Fig. 1 with the goodness of fit 99.60%. We note that since the variables MSY(median school years) and MVH(median value home) have a similar pattern, their correlation is high and, naturally, their angle must be small. Also TOP(total population), TOE(total employment) and HSE(health services employment) have the same characteristic variables and so their angles must be small. But these interpretations in Fig. 1 are not clear.

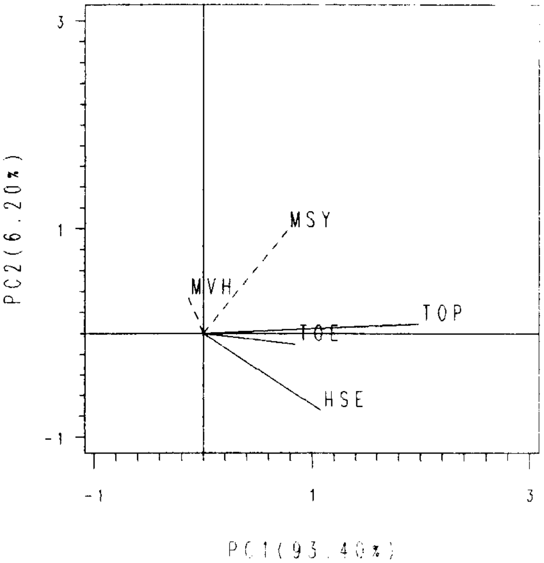


Figure 1. Classical h-Plot

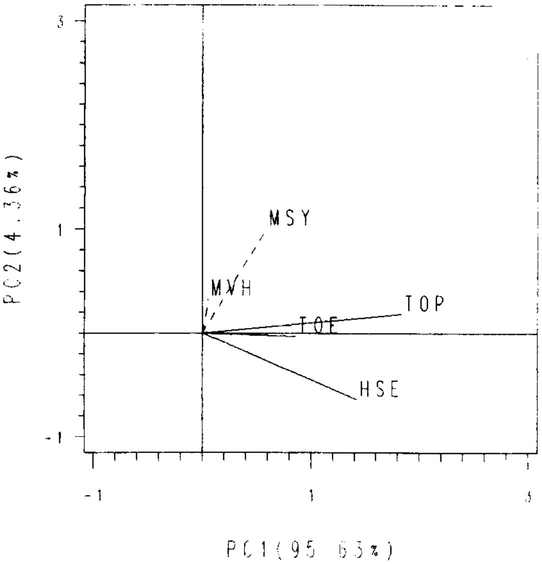


Figure 2. Resistant h-Plot

for the Census-Tract Data.

Now we will obtain the resistant h-plot. As noted in Subsection 3.1, we use the Andrew's $\psi(\cdot)$ function with c=0.89 where $(c\pi)^2$ is 95 percentile point of χ^2 distribution with 3 degrees of freedom. And we use 1.54 for the median scale estimate. Then the final weights used in computing resistant singular value decomposition (3.1) are in the diagonal matrix

$$\boldsymbol{D_w} = \mathrm{diag}(0.00, 0.00, 0.90, 0.43, 0.76, 0.81, 1.00,$$

$$0.00, 0.91, 0.88, 0.77, 0.00, 0.00, 0.00).$$

In $\boldsymbol{D_w}$, the elements having 0.00 denote the notable observations in data. So in our census-tract data, we know that observations 1, 2, 8, 12, 13 and 14 are notable.

The 2-dimensional resistant h-plot is shown in Fig. 2 with the goodness of fit 99.99%. By reducing the influence of the notable observations, Fig. 2 gives somewhat lucid intepretations, i.e., the angles between the same characteristic variables of Fig. 2 are smaller than those of Fig. 1.

Example 2. The open-closed book data (Mardia, Kent and Bibby, 1979, Table 1.2.1) consists of five variables for eighty-eight observations. From this data, as a matter of convenience we make an artificial data with the 4-th, 8-th, ..., 88-th observations of original data.

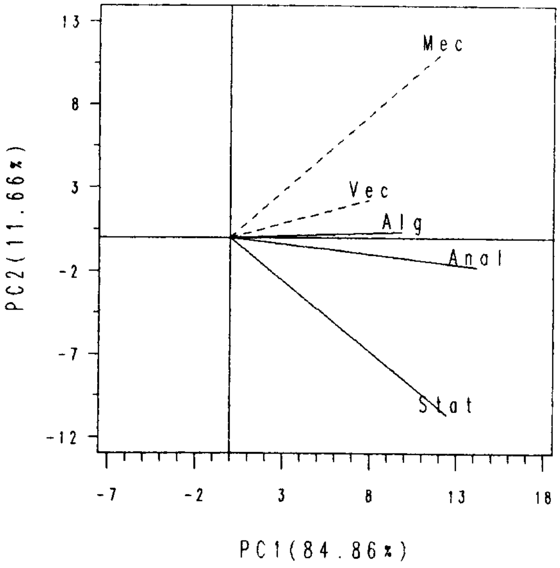
The 2-dimensional h-plot is given in Fig. 3 with the goodness of fit 96.52%. Fig. 3 shows that though the variables Vec(Vectors) and Alg(Algebra) are different type of examinations each other, their angle is amall and so they have higher correlation. Also the variables Mec(Mechanics) and Vec(Vectors) must have a small angle because their types of examinations are the same. And we note that the variables Alg(Algebra), Ana(Analysis) and Stat(Statistics) must have the same pattern. But Fig. 3 doesn't give these clear interpretations.

Now consider the resistant h-plot with Andrew's $\psi(\cdot)$ function as defined in Example 1. And also we use the median scale estimator 1.54. Then we have final weights used in computing resistant singular value decomposition

$$\begin{aligned} \boldsymbol{D_w} = \mathrm{diag}(&0.79, 0.85, 0.89, 1.00, 0.80, 0.58, 0.47,\\ &0.51, 0.03, 0.34, 0.78, 0.47, 0.60, 0.00,\\ &0.40, 0.11, 0.40, 0.57, 0.49, 0.00, 0.76, 0.81). \end{aligned}$$

In $\boldsymbol{D_w}$, we note that the 9-th, 14-th and 20-th diagonal elements have 0.03, 0.00 and 0.00 respectively. So their numbers of original data are 36, 56 and 80 respectively.

Thus the resistant h-plot with the goodness of fit 99.58% is given in Fig. 4. It gives the precise display of variables by reducing the influence of notable observations. In particular, we note that their angles between the variables of same pattern of Fig. 4 are smaller than those of Fig. 3.



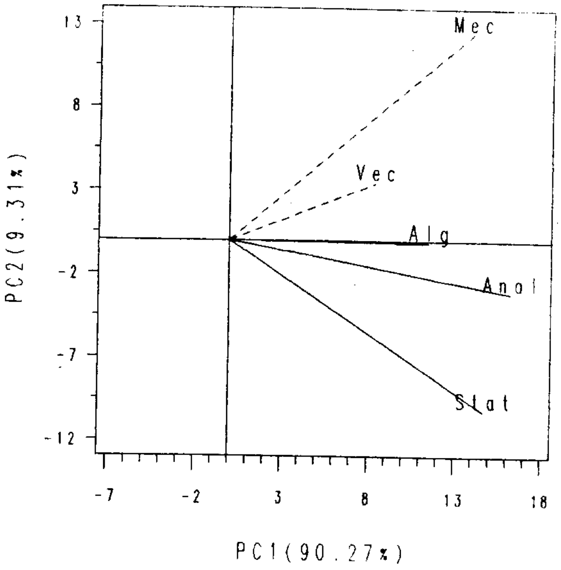


Figure 3. Classical *h*-Plot **Figure 4.** Resistant *h*-Plot
for the Open-Closed Book Data.

REFERENCES

(**1**) Belsley, D., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

(**2**) Bradu, D. and Gabriel, K.R. (1978). The Biplot as a Diagnostic Tool for Models of Two-Way Tables. *Technometrics*, **20**, 47–68.

(**3**) Choi, Y.S. and Huh, M.H. (1994). Resistant Singular Value Decomposition and its Applications. *Unpublished.*

(4) Corsten, L.C.A. and Gabriel, K.R. (1976). Graphical Exploration in Comparing Variance Matrices. *Biometrics*, **32**, 851–863.

(5) Gabriel, K.R. (1971). The Biplot Graphics Display of Matrices with Applications to Principal Component Analysis. *Biometrika*, **58**, 453–467.

(6) Householder, A.S. and Young, G. (1938). Matrix Approximation and Latent Roots. *American Mathematical Monthly*, **45**, 165–171.

(7) Jolliffe, I.T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.

(8) Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* Prentice-Hall, New Jersey.

(9) Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press.

(10) Seber, G.A. (1984). *Multivariate Observations.* Wiley, New York.