

연속형과 범주형의 혼합자료에 대한 h-plot

Author and Position

❑ **Junsu Park:** Master's course, Department of Statistics, Pusan National University

❑ **Yong-seok Choi:** Professor, Department of Statistics, Pusan National University

Text

1. Introduction

❖ Gabriel(1971, 1981)에 의해서 주로 개발된 행렬도(biplot)는 이원표 자료행렬(two-way data matrix)의 행(rows)과 열(columns)을 하나의 그래프에 동시에 나타내어 이들의 관계를 알아보는 다변량 그래픽 분석기법이다.

❖ 대표적인 행렬도로 주성분행렬도(principal component biplot)가 있으며, 국내에서는 최용석(1992)이 처음으로 주성분행렬도의 이론과 응용을 보였다. 특히, Choi(1995)는 자료행렬에서 열만의 관계를 나타내는 Corsten과 Gabriel(1976)의 h-plot의 저항성(resistant) 버전을 제시하였는데, 이 h-plot도 행렬도의 한 부류로 생각할 수 있다.

❖ 주성분행렬도나 h-plot은 연속형 자료행렬이 주어졌을 때에만 얻을 수 있다. 그래서 Gower(1992)는 자료행렬에 연속형 변수와 더불어 범주형 변수가 포함되어 있는 경우, 일반화행렬도(generalized biplots)를 사용할 것을 제안하였다. 이는 혼합자료(mixed data)가 주어졌을 때, 계량형 다차원척도법(metric multidimensional scaling)의 알고리즘을 적용하여 개체들의 정보를 저차원 공간상에 표현한 후에 연속형과 범주형 변수들의 정보를 추가적으로 가상점으로 나타내어 얻을 수 있다.

❖ 일반화행렬도는 연속형 변수와 범주형 변수들 간의 척도(scale) 차이가 크기 때문에 하나의 행렬도 상에 한꺼번에 표현할 수 없다. 또한, 연속형 변수들 간의 상관관계를 행렬도 상에 표현할 수가 없고, 자료의 개체 수가 많아질수록 알고리즘 계산시간이 오래 걸린다는 단점이 있다.

❖ 본 연구에서는 연속형과 범주형의 혼합자료가 주어졌을 때, h-plot의 알고리즘을 적용하여 일반화행렬도의 단점을 보완하고자 한다. 여기서 연속형 자료행렬은 상관행렬과의 관계를 이용하여 척도를 조정하고자 하며, 범주형 자료행렬은 다중대응분석(multiple correspondence analysis)의 다중표시행렬(multiple indicator matrix)로 재표현하여 척도를 조정하고자 한다.

2. h-plot and Multiple Indicator Matrix

2.1. h-plot

❖ 일반적으로 크기가 $n \times p$ 인 연속형 변수들의 자료행렬을 다음과 같이 정의한다.

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, i = 1, \dots, n; j = 1, \dots, p$$

❖ 열을 구성하는 변수 간에 측정단위가 다른 경우에는 다음과 같이 표준화된 새로운 자료행렬로 정의한다.

$$\tilde{\mathbf{X}} = \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right), \text{ 여기서 } \bar{x}_j = \sum_{i=1}^n x_{ij} / n, s_j^2 = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 / (n-1)$$

❖ 계수가 r 인 자료행렬 $\tilde{\mathbf{X}}$ 의 비정칙치분해(singular value decompositon)는 다음과 같다.

$$\tilde{\mathbf{X}} = \mathbf{U} \mathbf{D}_\lambda \mathbf{V}^t$$

여기서 행렬 $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_r)$ 는 $\mathbf{U}^t \mathbf{U} = \mathbf{I}_r$ 인 크기가 $n \times r$ 의 열 직교행렬이고, 행렬 $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_r)$ 는 $\mathbf{V}^t \mathbf{V} = \mathbf{I}_r$ 인 크기가 $p \times r$ 의 열 직교행렬이고. 또한 $\mathbf{D}_\lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$ 는 크기가 $r \times r$ 인 대각행렬로 $\lambda_1 \geq \dots \geq \lambda_r > 0$ 관계를 갖는 비정칙치(singular value)를 대각원소로 한다.

❖ 비정칙치분해는 다음과 같이 인자분해(factorization)된다.

$$\tilde{\mathbf{X}} = \mathbf{U}(\mathbf{V} \mathbf{D}_\lambda)^t = \mathbf{G} \mathbf{H}^t$$

전통적으로 $\mathbf{G} = \mathbf{U}$ 와 $\mathbf{H} = \mathbf{V} \mathbf{D}_\lambda$ 로 두기 때문에 주성분행렬도를 $\mathbf{G} \mathbf{H}^t$ 행렬도라 부르기도 한다. 여기서 크기가 $n \times p$ 인 는 행(개체)에 대한 정보를 담고 있으며, 크기가 $p \times r$ 인 $\mathbf{H}$ 는 열(변수)에 대한 정보를 담고 있는 열 좌표행렬이다.

❖ 자료행렬 $\mathbf{X}$ 에 대한 상관행렬을 $\mathbf{R}$ 이라고 할 때, 표준화 된 자료행렬 $\tilde{\mathbf{X}}$ 와 상관행렬 $\mathbf{R}$ 사이에는 대수적으로 다음의 관계가 성립한다.

$$\tilde{\mathbf{X}}^t \tilde{\mathbf{X}} = \mathbf{H} \mathbf{H}^t = (n-1) \mathbf{R}$$

❖ 여기서 행렬 $\mathbf{H}$ 를 다음과 같이 새로 정의하자.

$$\hat{\mathbf{H}} = \frac{1}{\sqrt{n-1}} \mathbf{H} = (\hat{\mathbf{h}}_1, \dots, \hat{\mathbf{h}}_p)^t$$

❖ 그러면 크기가 $r \times 1$ 인 벡터 $\hat{\mathbf{h}}_j, j = 1, \dots, p$ 는 행렬 $\hat{\mathbf{H}}$ 의 j 번째 행 벡터이다. Corsten과 Gabriel(1976)은 $\hat{\mathbf{h}}_j$ 들을 사용하여 변수들에 대한 정보를 나타내는 행렬도를 ‘h-plot’이라 정의하였다.

❖ h-plot의 기하적 성질은 다음과 같다.

$$\hat{\mathbf{h}}_j^t \hat{\mathbf{h}}_m \doteq r_{jm} \\ \cos(\theta_{jm}) \doteq r_{jm}$$

2.2. Multiple Indicator Matrix

❖ 일반적으로 크기가 $n \times p$ 인 범주형 변수들의 자료행렬을 다음과 같이 정의한다.

$$\mathbf{O} = (o_{ik}), i = 1, \dots, n; k = 1, \dots, p$$

이때 p 개의 범주형 변수들 중 k 번째 변수가 c_k 개의 범주를 가진다고 하면, n 개의 개체들에 대하여 각각의 범주별로 0과 1로만 나타낸 크기가 $n \times c_k$ 인 분할계획행렬(partition design matrix) $\mathbf{Z}_k$ 로 표현 가능하다.

❖ p 개의 범주형 변수들에 대한 다중표시행렬(multiple indicator matrix)은 다음과 같은 $n \times c$ 인 행렬로 정의될 수 있다.

$$\mathbf{Z} = \left[\begin{array}{c|c|c|c} \mathbf{Z}_1 & \mathbf{Z}_2 & \cdots & \mathbf{Z}_p \end{array} \right]$$

여기서 $c = \sum_{k=1}^p c_k$ 로 개의 범주형 변수들의 범주들을 모두 더한 값이다.

❖ 다중표시행렬 $\mathbf{Z}$ 의 각 열들의 합을 $z_{.h} = \sum_{i=1}^n z_{ih}$ 이라고 했을 때, 이를 대각원소로 하는 크기가 $c \times c$ 인 대각행렬 $\mathbf{D}_z$ 는 다음과 같이 정의된다.

$$\mathbf{D}_z = \text{diag}(z_{.1}, z_{.2}, \dots, z_{.c})$$

3. h-plot of Mixed Data

❖ 자료행렬 $\mathbf{X}$ 의 크기가 $n \times p_1$ 이고, 자료행렬 $\mathbf{O}$ 의 크기가 $n \times p_2$ 일 때, 두 자료행렬을 열 결합(column combine)시킨 크기가 $n \times (p_1 + p_2)$ 인 자료행렬을 다음과 같이 정의하자.

$$\left[\begin{array}{c|c} \mathbf{X} & \mathbf{O} \end{array} \right]$$

❖ 척도를 조정한 크기가 $n \times p_c$ 인 새로운 혼합자료행렬은 다음과 같다.

$$\mathbf{Y} = \left[\begin{array}{c|c} (n-1)^{-1/2} \tilde{\mathbf{X}} & \mathbf{Z} \mathbf{D}_z^{-1} \end{array} \right]$$

여기서 $p_c = p_1 + c$ 이고, c 는 다중표시행렬 $\mathbf{Z}$ 의 열의 수이다.

❖ 연속형 변수들의 자료행렬과 범주형 변수들의 자료행렬의 관계는 다음의 $\mathbf{Y}^t \mathbf{Y}$ 를 통하여 알 수 있다.

$$\mathbf{Y}^t \mathbf{Y} = \left[\begin{array}{cc} (n-1)^{-1} \tilde{\mathbf{X}}^t \tilde{\mathbf{X}} & (n-1)^{-1/2} \tilde{\mathbf{X}}^t \mathbf{Z} \mathbf{D}_z^{-1} \\ (n-1)^{-1/2} \mathbf{D}_z^{-1} \mathbf{Z}^t \tilde{\mathbf{X}} & \mathbf{D}_z^{-1} \mathbf{Z}^t \mathbf{Z} \mathbf{D}_z^{-1} \end{array} \right]$$

여기서 연속형 변수들 간의 관계를 나타낸 부분인 $(n-1)^{-1} \tilde{\mathbf{X}}^t \tilde{\mathbf{X}}$ 는 상관행렬 $\mathbf{R}$ 과 같고, $(n-1)^{-1/2} \tilde{\mathbf{X}}^t \mathbf{Z} \mathbf{D}_z^{-1}$ 는 연속형 변수와 범주형 변수들의 관계를 나타낸 부분으로서 표준화 연속형 변수들의 범주 별 평균이 되고, $\mathbf{D}_z^{-1} \mathbf{Z}^t \mathbf{Z} \mathbf{D}_z^{-1}$ 는 범주형 변수들 간의 관계를 나타낸 부분으로서 버트행렬 $\mathbf{Z}^t \mathbf{Z}$ 의 척도를 재조정된 부분이 된다.

❖ 혼합자료행렬 $\mathbf{Y}$ 에 대한 2차원 h-plot 알고리즘은 다음과 같다.

[1단계] $\mathbf{Y}$ 에 대해 비정칙치분해를 한다.

$$\mathbf{Y} = \mathbf{U}^* \mathbf{D}_\lambda^* \mathbf{V}^{*t}$$

여기서 행렬 $\mathbf{U}^* = (\mathbf{u}_1^*, \dots, \mathbf{u}_r^*)$ 는 $\mathbf{U}^{*t} \mathbf{U}^* = \mathbf{I}_{p_c}$ 인 크기가 $n \times p_c$ 의 열 직교행렬이고, 행렬 $\mathbf{V}^* = (\mathbf{v}_1^*, \dots, \mathbf{v}_r^*)$ 는 $\mathbf{V}^{*t} \mathbf{V}^* = \mathbf{I}_{p_c}$ 인 크기가 $p_c \times p_c$ 의 열 직교행렬이다. 또한, $\mathbf{D}_\lambda^* = \text{diag}(\lambda_1^*, \dots, \lambda_{p_c}^*)$ 는 크기가 $p_c \times p_c$ 인 대각행렬로 $\lambda_1^* \geq \dots \geq \lambda_{p_c}^* > 0$ 관계를 갖는 비정칙치를 대각원소로 한다.

[2단계] 2차원 h-plot의 열 좌표행렬을 계산한다.

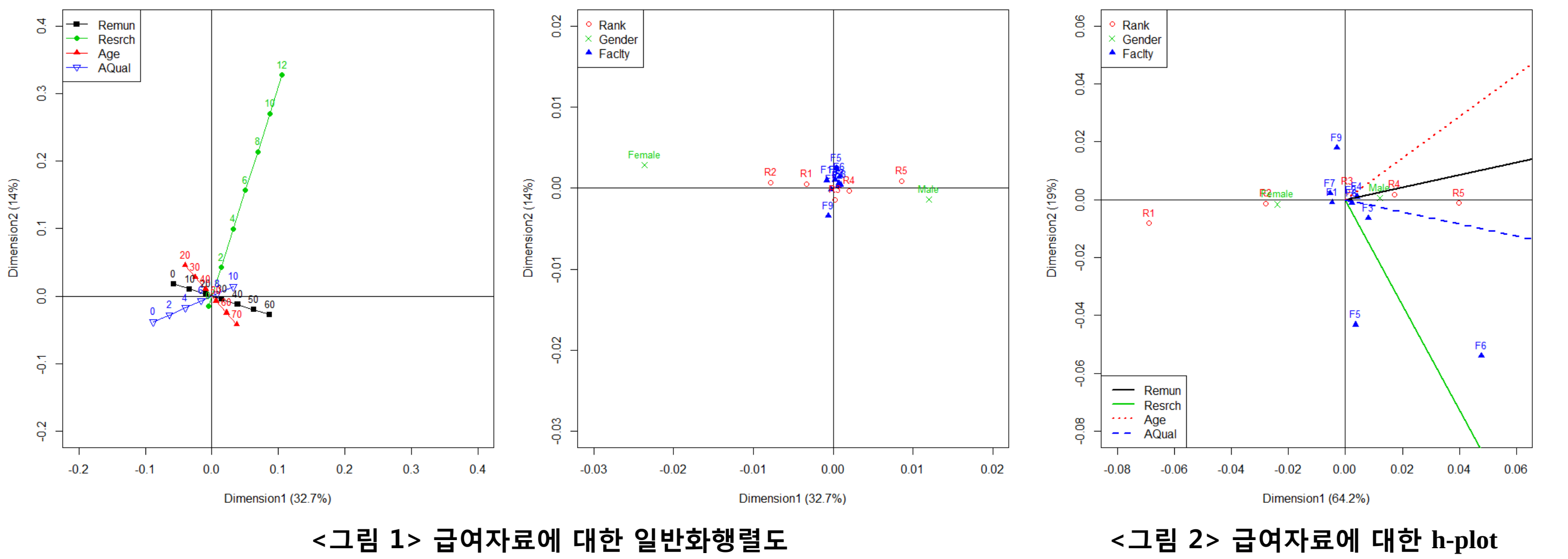
$$\hat{\mathbf{H}}_{(2)}^* = \frac{1}{\sqrt{n-1}} (\mathbf{v}_1^{*t} \lambda_1^*, \mathbf{v}_2^{*t} \lambda_2^*)$$

[3단계] 2차원 h-plot에 대한 근사적합도(goodness-of-fit of the approximation)를 계산한다.

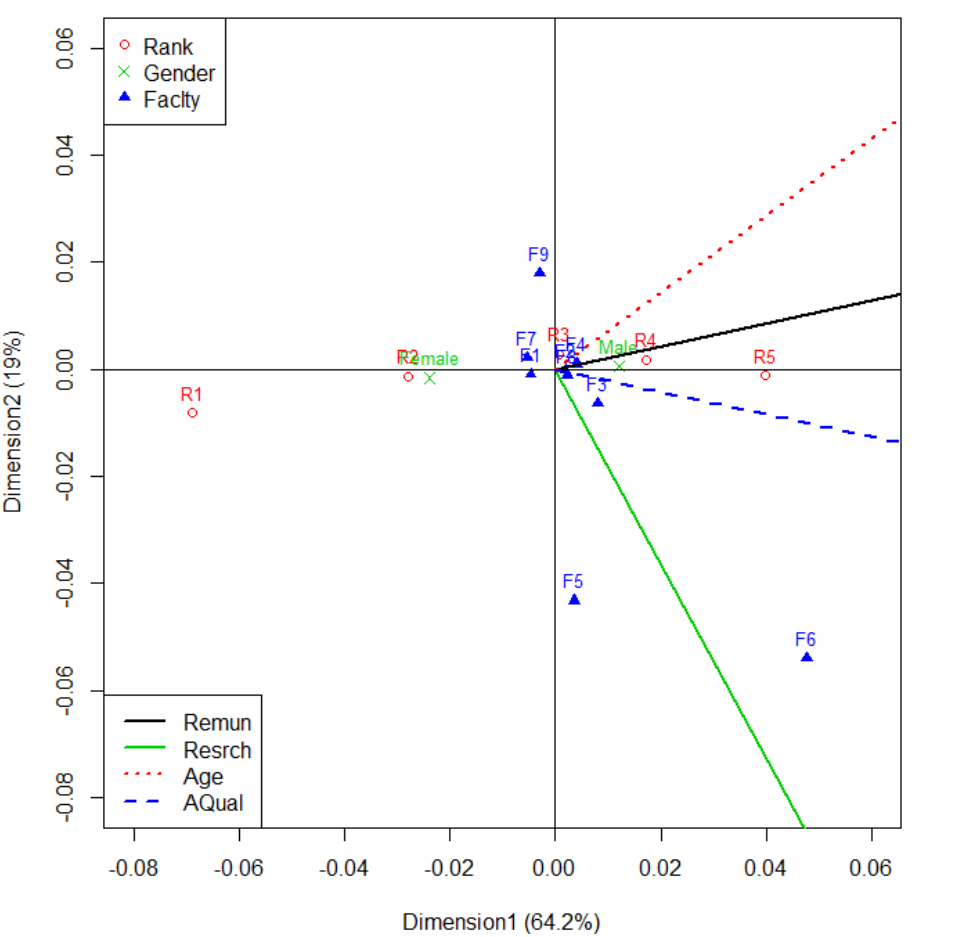
$$\begin{aligned} fit &= \left(1 - \frac{\| \mathbf{R}^* - \hat{\mathbf{H}}_{(2)}^* \hat{\mathbf{H}}_{(2)}^{*t} \|^2}{\| \mathbf{R}^* \|^2} \right) \times 100\% \\ &= \left(1 - \frac{\sum_{q=3}^{p_c} (\lambda_q^{*2})^2}{\sum_{q=1}^{p_c} (\lambda_q^{*2})^2} \right) \times 100\% \\ &= \left(\frac{\lambda_1^{*4} + \lambda_2^{*4}}{\sum_{q=1}^{p_c} \lambda_q^{*4}} \right) \times 100\% \end{aligned}$$

4. Illustrations

❖ 급여 자료는 Stellenbosch University의 2002년 교직원 728명에 대해 급여(Remun), 연구결과(Resrch), 나이(Age), 학문적 자격(AQual)의 4가지 연속형 변수와 학문적 위치(Rank; R1~R5), 성별(Gender; Male, Female), 학부(Facly; F1~F9)의 3가지 범주형 변수를 측정한 자료이다. 급여 자료에 대한 일반화행렬도와 h-plot은 각각 <그림 1>과 <그림 2>이며, 이에 대한 근사적합도는 <표 1>에 나타내었다.



<그림 1> 급여자료에 대한 일반화행렬도

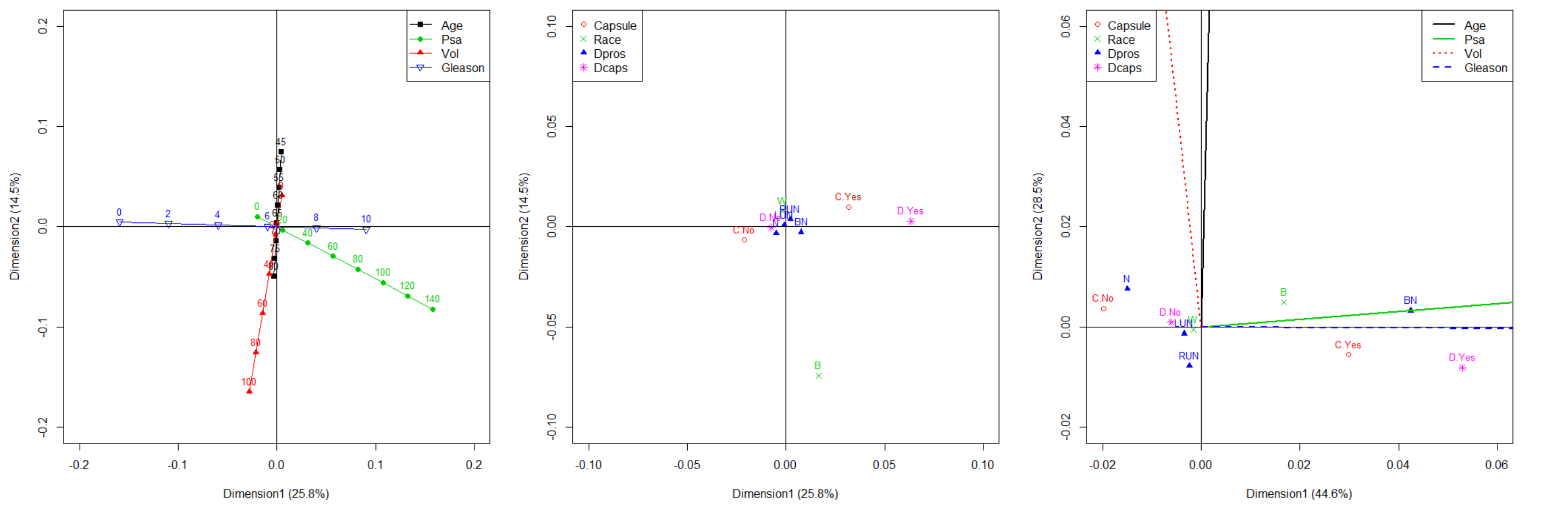


<그림 2> 급여자료에 대한 h-plot

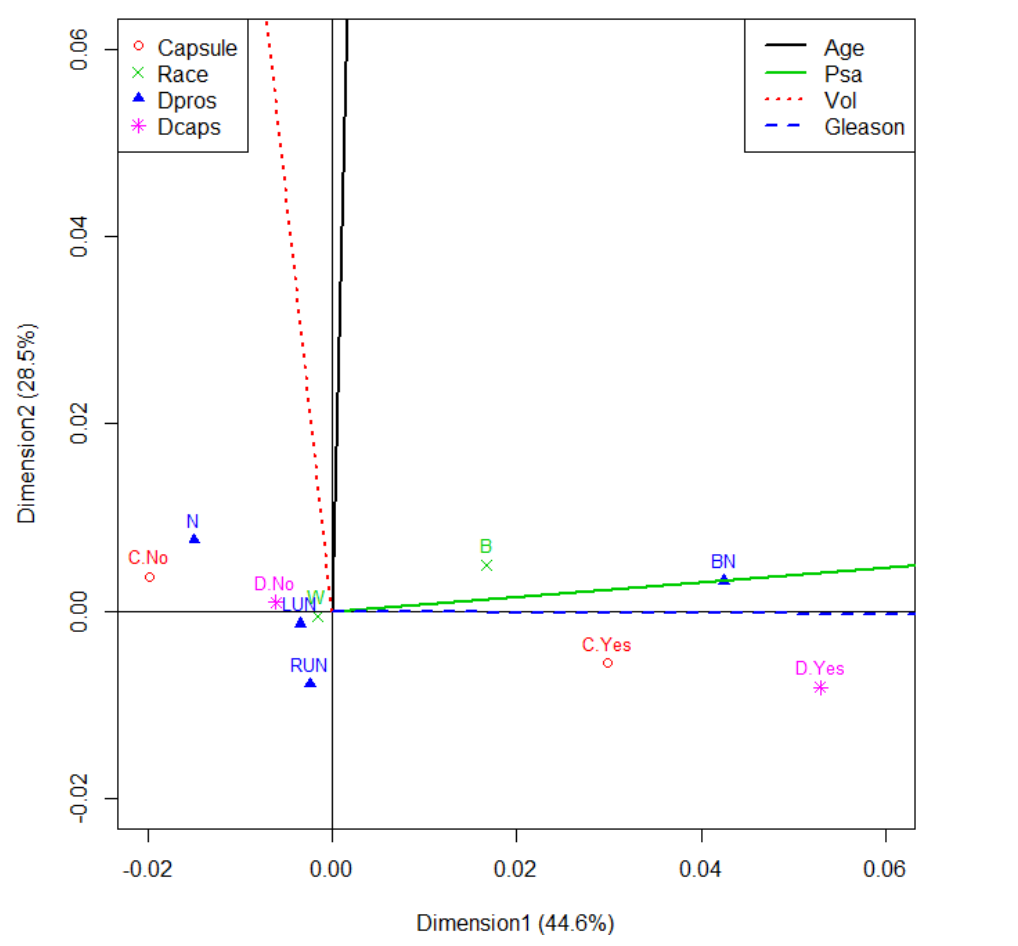
<표 1> 급여자료에 대한 일반화행렬도와 h-plot의 근사적합도

	축	근사적합도(%)	누적 근사적합도(%)
일반화행렬도	1	32.67	46.65
	2	13.98	
h-plot	1	64.20	83.21
	2	19.01	

❖ 전립선 암자료는 미국 오하이오 주립대학교 암센터에서 조사된 것으로 대상자 380명 중 결측값이 있는 4명을 제외한 총 대상자가 376명인 자료이다. 기초 검사에서 나이(Age), 전립선 항원값(Psa), 종양 크기(Vol), 글레존 점수 총계(Gleason)의 4가지 연속형 변수와 전립선암 전이 여부(Capsule; C.Yes, C.No), 인종(Race; W, B), 직장 검사 결과(Dpro; N, LUN, RUN, BN), 직장과의 연관 여부(Dcaps; D.Yes, D.No)의 4가지 범주형 변수를 측정했다. 전립선 암자료에 대한 일반화행렬도와 h-plot은 각각 <그림 3>과 <그림 4>이며, 이에 대한 근사적합도는 <표 2>에 나타내었다.



<그림 3> 전립선 암자료에 대한 일반화행렬도



<그림 4> 전립선 암자료에 대한 h-plot

<표 2> 전립선 암자료에 대한 일반화행렬도와 h-plot의 근사적합도

	축	근사적합도(%)	누적 근사적합도(%)
일반화행렬도	1	25.82	40.31
	2	14.49	
h-plot	1	44.62	73.16
	2	28.54	