LOGISTIC BIPLOTS FOR CATEGORICAL DATA

José Luis Vicente Villardón
villardon@usal.es
Universidad de Salamanca (Spain)

SUMMARY

Classical Biplot methods allow for the simultaneous representation of individuals and continuous variables in a given data matrix. When variables are binary, categorical or ordinal, a classical linear biplot representation is not suitable. We propose a linear biplot representation based on logistic response models. The coordinates of individuals and variables are computed to have logistic responses along the biplot dimensions. The method is related to logistic regression in the same way that Classical Biplot Analysis (CBA) is related to linear regression. Thus we refer to the method as Logistic Biplot (LB). In the same way as Linear Biplots are related to Principal Components Analysis, Logistic Biplots are related to Latent Trait Analysis or Item Response Theory. The geometry of those kinds of biplots is studied and their usefulness in Data Mining is illustrated using data on SNPs (Single Nucleotide Polymorphisms) from the HAPMAP project.

Key Words: Logistic Biplot; Categorical Data; Regression biplots.

1. Logistic Biplot for Binary Data

Let $\mathbf{X}_{IxJ}$ be a data matrix in which the rows correspond to I individuals and the columns to J binary characters. Let $p_{ij} = E(x_{ij})$ the expected probability that the character j be present at individual i, and x_{ij} the observed probability, either 0 or 1, resulting in a binary data matrix. The *S-dimensional* logistic biplot in the logit scale is formulated as

$$\log it(p_{ij}) = \log(\frac{p_{ij}}{1-p_{ij}}) = b_{j0} + \sum_{s=1}^{S} b_{js}a_{is} = b_{j0} + \mathbf{a}_i'\mathbf{b}_j \qquad \left[p_{ij} = \frac{e^{b_{j0}+\sum_k b_{jk}a_{ik}}}{1+e^{b_{j0}+\sum_k b_{jk}a_{ik}}} \right]$$

where a_{is} and b_{js} ($i=1, ...,I; j=1, ...,J; s=1, ..., S$) are the model parameters used as row and column markers respectively. The model is a generalized (bi)linear model having the logit as a link function. In matrix form, $\log it(\mathbf{P}) = \mathbf{1}_I \mathbf{b}_0' + \mathbf{AB}'$, where $\mathbf{P}$ is the matrix of expected probabilities, $\mathbf{1}_I$ is a vector of ones; $\mathbf{b}_0$ is the vector containing the constants, $\mathbf{A}$ and $\mathbf{B}$ are the matrices containing the markers for the rows and columns of $\mathbf{X}$. The constants b_{j0} have been added because it is not possible to center the data matrix in the same way as in linear biplots. The constant is the displacement of the gravity centre in the same way as it is the first ordination axis in Correspondence Analysis. The model is a latent trait model for binary data, being the row coordinates the scores of the individuals on the latent trait. Although the biplot in the logit scale may be useful, it would be more interpretable in a probablility scale.

The points predicting different probabilities are on parallel straight lines on the biplot; this means that predictions on the logistic biplot are made in the same way as on the linear biplots, i. e., projecting a row marker $\mathbf{a}_i=(a_{i1}, a_{i2})$ onto a column marker $\mathbf{b}_j=(b_{j1}, b_{j2})$. (See VICENTE-VILLARDON, et al. , 2006).

2. LOGISTIC BIPLOT FOR CATEGORICAL DATA

Let $\mathbf{X}_{IxJ}$ be a data matrix containing the values of J categorical variables -each with K_j ($j=1, ..., J$) categories- for I individuals, and let $\mathbf{G}_{IxL}$ be the corresponding indicator matrix with $L = \sum_j (K_j - 1)$ columns. The last category of each variable will be used as a baseline. Let $p_{i(jk)}$ the expected probability that the category k of variable j be

present at individual *i*. In the he multinomial logistic latent trait model we assume that the log-odds of each response (relative to the last category) follows a linear model

$$\log(\frac{p_{i(jk)}}{p_{i(jK_j)}}) = b_{j0} + \sum_{s=1}^{S} b_{(jk)s} a_{is} = b_{j0} + \mathbf{a}_i' \mathbf{b}_{(jk)}$$

where a_{is} and $b_{(jk)s}$ ($i=1, \ldots, I$; $j=1, \ldots, J$; $k=1, \ldots, K_j-1$ $s=1, \ldots, S$) are the model parameters. In matrix form, $\mathbf{O} = \mathbf{1}_I \mathbf{b}_0' + \mathbf{A}\mathbf{B}'$, where $\mathbf{O}_{IxL}$ is the matrix containing the expected odds, defines a biplot for the odds. Although the biplot for the odds may be useful, it would be more interpretable in terms of predicted probabilities and categories. The points predicting different probabilities are no longer on parallel straight lines (see the figure 1 with the response surfaces); this means that predictions on the logistic biplot are not made in the same way as in the linear biplots, the surfaces define now prediction regions for each category as shown in the graph.

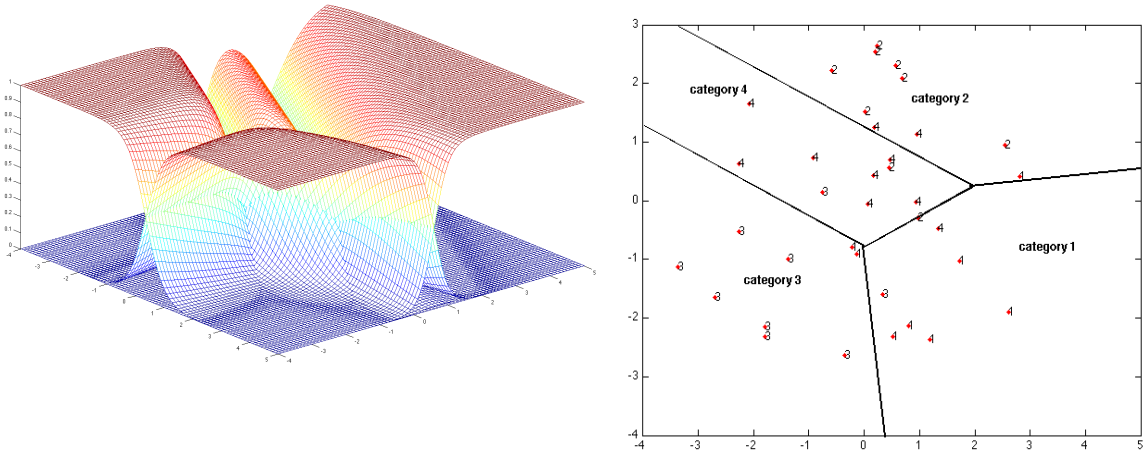
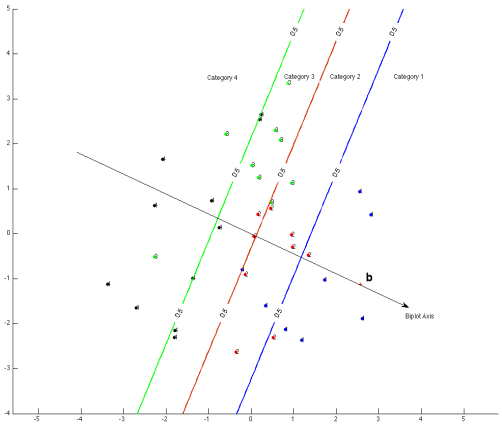


Figure 1: Response surfaces and Prediction Regions for a Nominal Logistic Biplot.

Prediction regions are closely related to Voronoi diagrams in Computational Geometry but changing distances by probabilities.

3. LOGISTIC BIPLOT FOR ORDINAL DATA

Let $\mathbf{X}_{IxJ}$ be a data matrix containing the values of J ordinal variables -each with K_j ($j=1, \ldots, J$) categories- for I individuals, and let $\mathbf{G}_{IxL}$ be the cumulative indicator matrix with $L = \sum_j (K_j - 1)$ columns. The last category of each variable will be used as a baseline. Let $p_{i(j \le k)} = P(x_{ij} \le k)$. The ordinal logistic latent trait model for the cumulative probabilities is



$$p_{i(j \le k)} = \frac{e^{b_{jk} + \sum_s b_{js} a_{is}}}{1 + e^{b_{jk} + \sum_s b_{js} a_{is}}}$$

$$\log it(p_{i(j \le k)}) = \log\left[\frac{p_{i(j \le k)}}{p_{i(j \le k)}}\right] = b_{jk} + \sum_s b_{js} a_{is}$$

The equations define a biplot in the logit scale that shares the geometry of the binary case for each category. Observe that each category have a different constant but the same slopes, that means that the prediction direction is common to all categories and just the

prediction markers are different. The parameters *b* define the direction of the projection; the representation subspace can be divided into prediction regions, for each category, delimited by parallel straight lines. (see figure).

4. PARAMETER ESTIMATION

- Alternated generalized regressions and interpolations. (Maximum Likelihood), (GOWER & HAND, 1986; GABRIEL, 1998; VICENTE-VILLARDON, 2006)
- Marginal Maximum Likelihood (As in Item Response Theory, BAKER, 1992).
- Iterative Majorization. (VERBOON et al., 1994).
- Heuristic approach for big data matrices: External Logistic Biplots (Logistic fits on the Principal Coordinates), (DEMEY et al., 2008)

5. APPLICATIONS

- HAPMAP Data. (DEMEY et al., 2008)
- Sugar cane germoplasm. (ZAMBRANO, et al., 2007)
- Innovation profiles in Portugal. (VICENTE, et al., 2010)
- Irregular working force in Spain. (PATINO et al., 2010)

REFERENCES

BAKER, F.B. (1992): *Item Response Theory. Parameter Estimation Techniques*. Marcel Dekker. New York.

GABRIEL, K. R. (1998). Generalised bilinear regresión. *Biometrika*, 85: 689 – 700.

GOWER, J. C. & HAND, D. (1986): *Biplots*. Chapman & Hall. London.

DEMEY, J., VICENTE-VILLARDON, J. L., GALINDO, M.P. & ZAMBRANO, A. (2008) Identifying Molecular Markers Associated With Classification Of Genotypes Using External Logistic Biplots**.** *Bioinformatics*, 24(24):2832-2838.

PATINO MC., VICENTE P., VICENTE-VILLARDÓN JL. Y GALINDO P. (2010) Identifying Inmigrant Women patterns by external Logistic Biplots. *Sociological Methodology* (En revisión).

VERBOON, P. & HEISER, W. J. (1994). Resistant Lower Rank Approximation of Matrices by Iterative Majorization. *Computational Statistics & Data Analysis.* 18: 457-467.

VICENTE, M.P.; NORONHA, T. AND NIJKAMP, P. (2010). Institutional capacity to dynamically innovate: an application to the portuguese case. *Technological Forecasting & Social Change*. (En prensa)

VICENTE-VILLARDON, J. L., GALINDO M. P. & BLAZQUEZ, A. (2006). Logistic Biplots. In "*Múltiple Correspondence Análisis And Related Methods*". Grenacre, M & Blasius, J. Eds. Chapman and Hall. Boca Ratón.

ZAMBRANO, A. Y., MARTÍNEZ, G., GUTIÉRREZ, E., MANZANILLA E., VICENTE-VILLARDÓN, J. L. & DEMEY, J. (2007). Marcador RAPD asociado a la resistencia a *Fusarium Oxysporum* en *Musa*. *Interciencia*. Vol 32, nº 11: 775-779. (2007).