REGRESSION, DISCRIMINANT ANALYSIS, AND CANONICAL CORRELATION ANALYSIS WITH HOMALS

JAN DE LEEUW

ABSTRACT. It is shown that the `homals` package in R can be used for multiple regression, multi-group discriminant analysis, and canonical correlation analysis. The `homals` solutions are only different from the more conventional ones in the way the dimensions are scaled by the eigenvalues.

1. MORALS

Suppose we have $m+1$ variables, with the first m being *predictors* (or *independent variables*), and the last one the *outcome* (or *dependent variable*). In `homals` [De Leeuw and Mair, 2009] we use `ndim=1,sets=list(1:m,m+1),rank=1` which means the loss function looks like

$$\sigma(x,a,q) = (x - a_{m+1}q_{m+1})'(x - a_{m+1}q_{m+1}) + (x - \sum_{j=1}^{m} a_j q_j)'(x - \sum_{j=1}^{m} a_j q_j)$$

with q_j the quantified or transformed variables. This must be minimized over a, x, q under the conditions that $u'x = u'q_j = 0$ and $x'x = q_j'q_j = 1$, and of course that $q_j \in \mathscr{K}_j$, the appropriate set of admissible transformations.

Write

$$Q = \begin{bmatrix} q_1 & \cdots & q_m \end{bmatrix}$$

and $b = (a_1, \cdots, a_m)$. Also write $s = a_{m+1}$ and $y = q_{m+1}$. Then

$$\sigma(x,a,q) = (x - sy)'(x - sy) + (x - Qb)'(x - Qb).$$

It folllows that

$$s = x'y,$$
$$b = (Q'Q)^{-1}Q'x,$$

Date: Sunday 26th April, 2009 — 1h 8min — Typeset in TIMES ROMAN.

as well as

$$x = \frac{sy + Qb}{\|sy + Qb\|}.$$

Also x is the normalized eigenvector corresponding with the largest eigenvalue of $K = yy' + P$, where $P = Q(Q'Q)^{-1}Q'$. But the non-zero eigenvalues of K are the squares of the non-zero singular values of

$$\begin{bmatrix} y & \mid & Q(Q'Q)^{-\frac{1}{2}} \end{bmatrix}$$

and these are the same as the non-zero eigenvalues of

$$H = \begin{bmatrix} 1 & y'Q(Q'Q)^{-\frac{1}{2}} \\ (Q'Q)^{-\frac{1}{2}}Q'y & I \end{bmatrix}$$

Define the usual regression quantities $\beta = (Q'Q)^{-1}Q'y$ and $\rho^2 = y'Q(Q'Q)^{-1}Q'y$. The eigenvalues of H are $1+\rho$, $1-\rho$, and 1 with multiplicity $m-1$. An eigenvector corresponding with the dominant eigenvalue is

$$\begin{bmatrix} \rho \\ (Q'Q)^{-\frac{1}{2}}Q'y \end{bmatrix}.$$

It follows that an eigenvector corresponding with the dominant eigenvalue of K is $(Q(Q'Q)^{-1}Q' + \rho I)y$, and

$$x = \frac{1}{\rho\sqrt{2(1+\rho)}}(Q(Q'Q)^{-1}Q' + \rho I)y.$$

Thus

$$b = \frac{1}{\rho}\sqrt{\frac{1+\rho}{2}}\beta,$$
$$s = \sqrt{\frac{1+\rho}{2}}.$$

The vector of regression coefficient β is thus proportional to b, and the two are identical if and only if $\rho = 1$. The minimum loss function value is $1-\rho$. Thus, ultimately, we find tranformations q_j of the variables in such a way that the multiple correlation is maximized.

2. CRIMINALS

Again we have $m+1$ variables, with the first m being *predictors* and the last one the *outcome*. But now the outcome is a categorical variable with k categories. In `homals` we use `ndim=p,sets=list(1:m,m+1),rank=c(rep(1, m),p)` where $p<k$. The loss function is

$$\sigma(X,A,Q,Y)=\mathbf{tr}\ (X-GY)'(X-GY)+\mathbf{tr}\ (X-QA)'(X-QA),$$

where G is the indicator matrix of the outcome, and where we require $u'X=u'Q=0$ and $X'X=\mathbf{diag}(Q'Q)=I$. Now we must have at the minimum

$$\begin{aligned} Y&=(G'G)^{-1}G'X,\\ A&=(Q'Q)^{-1}Q'X. \end{aligned}$$

Thus X are the normalized eigenvectors corresponding with the p largest eigenvalues of $K=G(G'G)^{-1}G'+Q(Q'Q)^{-1}Q'$. And X also are the normalized left singular vectors of

$$\left[G(G'G)^{-\frac12}\quad|\quad Q(Q'Q)^{-\frac12}\right].$$

We can find the right singular vectors as the eigenvectors of

$$H=\begin{bmatrix} I & (G'G)^{-\frac12}G'Q(Q'Q)^{-\frac12}\\ (Q'Q)^{-\frac12}Q'G(G'G)^{-\frac12} & I\end{bmatrix}.$$

Now let $U\Psi V'$ be the singular value decomposition of $(G'G)^{-\frac12}G'Q(Q'Q)^{-\frac12}$. Then $\begin{bmatrix}U\\ V\end{bmatrix}$ are the eigenvectors of H corresponding with the largest eigenvalues $I+\Psi$. Take the eigenvectors $\begin{bmatrix}U_p\\ V_p\end{bmatrix}$ corresponding with the p largest singular values Ψ_p. The corresponding left singular vectors are $\tilde{X}=G(G'G)^{-\frac12}U_p+Q(Q'Q)^{-\frac12}V_p$. Because $\tilde{X}'\tilde{X}=2(I+\Psi_p)$ we find

$$X=2^{-\frac12}(G(G'G)^{-\frac12}U_p+Q(Q'Q)^{-\frac12}V_p)(I+\Psi_p)^{-\frac12}.$$

Thus

$$\begin{aligned} Y&=2^{-\frac12}(G'G)^{-\frac12}U_p(I+\Psi_p)^{\frac12},\\ A&=2^{-\frac12}(Q'Q)^{-\frac12}V_p(I+\Psi_p)^{\frac12}, \end{aligned}$$

and

$$X=(GY+QA)(I+\Psi_p)^{-1}.$$

Also note that $Y'G'GY = A'Q'QA = \frac{1}{2}(I+\Psi_p)$, while $Y'G'QA = \frac{1}{2}\Psi_p(I+\Psi_p)$. The minimum value of the loss function is $p - \mathbf{tr}\ \Psi_p$.

Now let us compare these computations with the usual canonical discriminant analysis. There we compute the projector $P = G(G'G)^{-1}G'$ and the between-groups dispersion matrix $B = Q'PQ$ and we solve the generalized eigenvalue problem $BZ = TZ\Lambda$, where $T = Q'Q$ is the total dispersion. The problem is normalized by setting $Z'TZ = I$. Thus, using the p largest eigenvalues, $Q'G(G'G)^{-1}G'QZ_p = Q'QZ_p\Lambda_p$. This immediately gives $\Lambda_p = \Psi_p^2$. Also $(Q'Q)^{\frac{1}{2}}Z_p = V_p$ or $Z_p = \sqrt{2}A(I+\Psi_p)^{-\frac{1}{2}}$. For the group means $M_p = (G'G)^{-1}G'QZ_p$ we find $M_p = \sqrt{2}Y(I+\Psi_p)^{-\frac{1}{2}}$. Thus both Z_p and M_p are simple rescalings of A and Y. `homals` find the transformations of the variables that maximizes the sum of the p largest singular values of $(G'G)^{-\frac{1}{2}}G'Q(Q'Q)^{-\frac{1}{2}}$.

3. CANALS

Canonical correlation analysis with `homals` has $m1+m2$ variables, and we use `ndim=p,sets=list(1:m1,m1+(1:m2)),rank=c(rep(1,m1+m2)`. The loss is

$$\sigma(X,A,Q) = \mathbf{tr}\ (X - Q_1A_1)'(X - Q_1A_1) + \mathbf{tr}\ (X - Q_2A_2)'(X - Q_2A_2).$$

Since our analysis of discriminant analysis in `homals` never actually used the fact that G was an indicator, the results are exactly the same as in the previous section (with the obvious substitutions).

In classical canonical correlation analysis the function $\mathbf{tr}\ R'Q_1'Q_2S$ is maximized over $R'Q_1'Q_1R = I$ and $S'Q_2'Q_2S = I$. This means solving

$$\begin{aligned} Q_1'Q_2S &= Q_1'Q_1R\Phi, \\ Q_2'Q_1R &= Q_2'Q_2S\Phi. \end{aligned}$$

From `homals`, as before,

$$\begin{aligned} A_1 &= 2^{-\frac{1}{2}}(Q_1'Q_1)^{-\frac{1}{2}}U_p(I+\Psi_p)^{\frac{1}{2}}, \\ A_2 &= 2^{-\frac{1}{2}}(Q_2'Q_2)^{-\frac{1}{2}}V_p(I+\Psi_p)^{\frac{1}{2}}. \end{aligned}$$

In canonical analysis $\Phi = \Psi$ and

$$R = (Q_1'Q_1)^{-\frac{1}{2}}U_p = \sqrt{2}A_1(I+\Psi_p)^{-\frac{1}{2}},$$
$$S = (Q_2'Q_2)^{-\frac{1}{2}}V_p = \sqrt{2}A_2(I+\Psi_p)^{-\frac{1}{2}}.$$

Again we see the same type of rescaling of the canonical weights.

Note that `homals` does *not* find the transformations that maximize the sum of the *squared* canonical correlations, which is the target function in the original CANALS approach [Young et al., 1976; Van Der Burg and De Leeuw, 1983]. Maximizing the square of the canonical correlations means maximizing a different *aspect* of the correlation matrix [De Leeuw, 1988, 1990].

References

J. De Leeuw. Multivariate Analysis with Linearizable Regressions. *Psychometrika*, 53:437–454, 1988.

J. De Leeuw. Multivariate Analysis with Optimal Scaling. In S. Das Gupta and J. Sethuraman, editors, *Progress in Multivariate Analysis*, Calcutta, India, 1990. Indian Statistical Institute.

J. De Leeuw and P. Mair. Homogeneity Analysis in `R`: the Package `homals`. *Journal of Statistical Software*, (in press), 2009.

E. Van Der Burg and J. De Leeuw. Non-linear Canonical Correlation. *British Journal of Mathematical and Statistical Psychology*, 36:54–80, 1983.

F. W. Young, J. De Leeuw, and Y. Takane. Regression with Qualitative and Quantitative Data: and Aletrnating Least Squares Approach with Optimal Scaling Features. *Psychometrika*, 41:505–529, 1976.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: `deleeuw@stat.ucla.edu`

URL, Jan de Leeuw: `http://gifi.stat.ucla.edu`