Discriminant analysis algorithm to classify Hickson’s compact groups of galaxies

H. I. Abdel-Rahman* and Sabry A. Mohammed*©

*Astronomy Department, National Research Institute of Astronomy and Geophysics, Helwan, Cairo, Egypt; ©Department of Mathematics, Faculty of Science and Arts, Shaqra University, Shaqra, Saudi Arabia; ©Department of Physics, College of Science and Humanities, Hawtat sudair, Majmaah University, Hawtat sudair, Saudi Arabia

ABSTRACT
In the present paper, we introduce an algorithm to classify Hickson’s compact groups of galaxies. The algorithm uses the discrimination analysis technique to determine the best set of indicators that can be used to construct the models in which the classification validity of the Hickson’s compact groups could be predicted, and thus distinguish between groups that have classification problems. We have concluded that the earlier classifications of the 92 Hickson’s compact groups of galaxies are correct.

1. Introduction
Groups of galaxies are systems that have a small number of members (galaxies) in the small region and satisfy certain criterion. They contain approximately half of the galaxies within a magnitude limited redshift survey, Ramella et al. (2002), Ramella et al. (1997), Giuricin et al. (2000), Tucker et al. (2000), Merchán and Zandivarez (2002), Holmberg (1950) and Humason et al. (1956).

There are many different catalogues of groups of galaxies depending on many different criteria such as position, magnitude, mean separation distance between members in the group, and mean surface brightness (Rose 1977; Hickson 1982; Hickson et al. 1992, Hickson 1993; Allam and Tuker 2000; Lee et al. 2004; de Carvalho et al. 2005; Wang et al. 2008).

Compact groups of galaxies (CGGs) are systems contain a few numbers of galaxies (<10) that are close to each other. Hickson (1982) has visually inspected all the red Palomar sky survey reprints covering about 67% of the sky. He used the following criteria

\[ n \geq 4 \quad \text{with} \quad m \geq m_B + 3R_N \geq 3R_G \quad \text{and} \quad \mu_G \leq 26 \]

where \( n \) is the number of members, \( m_B \) is the estimated magnitude of the brightest group member, \( R_G \) is the radius of the smallest circle containing the group members, \( R_N \) is the distance from the centre of this circle to the nearest non-member satisfying the same magnitude condition and \( \mu_G \) is the mean surface brightness contained by the circle.

In this article, using discriminant analysis we try to check whether the classification of Hickson groups current poses true or that there is an error in the classification and affiliation of some galaxies to other groups.

The organisation of the paper is as follows. Section 2 is devoted to the method of calculations. In Section 3 computational algorithm is outlined. The results reached are described in Section 4, while the conclusion is outlined in Section 5.

2. The method
We develop a linear model that classifies Hickson groups through the identification of the following:

1. The galaxies that belong to Hickson’s groups is true, for the purpose of determining the boundary point between the discriminatory signs for Hickson groups that suffer from problems of the classification.
2. The most important variables in the composition of the discriminant function.
3. The ability of the proposed model of discrimination between Hickson’s groups which suffer or not from problems in the classification.

The main assumption of the research includes the ability of linear model proposed, consisting of a set of variables for the problems of the classification of the groups which reached using the discriminant analysis to distinguish between groups.

Suppose that, there are \((K)\) groups and each of the group includes \((n_i)\) from observations so that \(i = 1, 2, \ldots, k\).
The observation vector \((y)\) of each group could be converted by discriminant analysis using:
\[
z_{ij} = a y_{ij} \tag{1}\]
The finding of a vector \((a)\), that maximises the differences between vector averages \(z_{ij}\), will go through the resolution of the following equation:
\[
(E^{-1}H - \lambda I)a = 0 \tag{2}
\]
where
\[
H = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})'
\]
\[
= \sum_{i=1}^{k} \frac{1}{n_i} y_i y_i' - \frac{1}{N} y y'
\]
and
\[
E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'
\]
\[
= \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} y_{ij}' - \sum_{i=1}^{k} \frac{1}{n_i} y_i y_i'.
\]
The solution of formula (2) lead to the extraction of Eigenvalues \((\lambda_i)\) and the corresponding Eigenvectors \((a_i)\) for the matrix \((E^{-1}H)\), Rencher (2002).

Therefore, the largest Eigenvalues \((\lambda_i)\) represent the maximum value of:\-
\[
\lambda = \frac{a'Ha}{a'Ea} \tag{3}
\]
where \(a\) is Eigenvector and \(a'\) is its transpose.

Thus, the first discriminant function that maximises the difference between the averages of groups are
\[
z_1 = a_1' y \tag{4}
\]
Let that, there are \((S)\) Eigenvectors, then the discriminant functions that maximise the difference between the averages of groups could be written as
\[
z_1 = a_1' y, z_2 = a_2' y, \ldots, z_s = a_s' y
\]
And using the Eigenvalues of the matrix \(E^{-1}H\) to find the relative importance to the discriminant analysis as
\[
\lambda_i = \frac{SSH(z)}{SSE(z)} \tag{5}
\]
Selection of independent variables which comprise the model by choosing the variables that have the highest value \((F)\) and the minimum value of Wilks Lambda (Rencher 2002). The rate of \((F)\) the contribution of independent variables in discrimination between groups, after taking into account the changes introduced by the other of the discriminating variables. Wilks Lambda measures the degree of divergence between the groups.

The Standardisation discriminant coefficients \((b)\) could be obtained by solving from the following equation (Hintze 2007)
\[
y^* = b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \tag{6}
\]
Where:
- \(y^*\): The Standardisation Discriminant value
- \(x_n\): The Standardisation Discriminant variable
- \(b_n\): The Standardisation Discriminant coefficient
- \(n\): The numbers of The Standardisation Discriminant variables. It equals (the number of groups – 1).

Un-standardisation Discriminant coefficients are the b values in the following equation
\[
y = b_1 s_1 + b_2 s_2 + b_3 s_3 + \ldots + b_n s_n + C \tag{7}
\]
Where:
- \(C\): constant.
- \(s_i\): Un-standardisation Discriminant Variables.
- \(b_i\): Un-standardisation Discriminant coefficients.
- \(y\): The Un-standardisation Discriminant value.

We do the accuracy test of the discriminant function as follows (Walfgang 2003):

1. (1) test the validity of forecasting
   This is done by finding the value of the discriminating coefficient of the Equation (4) by multiplying the unstandardized discriminant coefficient for each ratio or variable in real terms, and sum the products of each ratio within the discriminating equation in addition to the sum or subtract a fixed number of them. Comparing the discriminating value of the group with the actual values of the group, we classify the individual within this group or another.

2. (2) The ability to the discriminant function of discrimination between groups
   To test the ability of the discriminant function of discrimination between groups be based on statistical indicators as follows:
   a) The Eigenvalues
   We use the Eigenvalues, to know the ability of discrimination between groups where the high value of the Eigenvalues is an indication of the ability of the discrimination between groups. In addition, rewrite formula (3) as follows:
   \[
   \lambda_i = \frac{SSH(z)}{SSE(z)} \tag{8}
   \]
   where
   - \(SSE = \sum_{ij} (z_{ij} - \bar{z}_i)^2\)
   - \(SSH = n \sum_{i=1}^{k} (\bar{z}_i - \bar{z})^2\)
   b) Canonical correlation
   The canonical correlation measures the goodness of the discriminant function, where the high value of the correlation coefficient indicates the quality of goodness of the discriminant function and its squares equal to the determination coefficient.
   c) Wilkes Lambda Test
We use Wilkes Lambda is to test the ability of the discrimination between groups as follows

\[ V_m = -\left[ N - 1 - \frac{1}{2}(p + k) \right] \ln \Lambda_m \]  \hspace{1cm} (9)

where

\[ \Lambda_m = \prod_{i=m}^{1} \frac{1}{1 + \lambda_i} \],

where \( \lambda \) has the chi-square distribution. If the calculated value is less than the tabulated value, i.e. the discriminant function has the ability to discriminate between groups.

(d) F test

We used the F test to investigate the statistical significance of the ability of the discriminant function to separate the groups as

\[ F = \frac{1 - \frac{\Lambda_m}{\Lambda_m}}{\frac{df_2}{df_1}} \]  \hspace{1cm} (10)

Where \( df_1 \) and \( df_2 \) are the degrees of freedom. If the value of the calculated F is larger than the value of the tabulated F, this means that discriminant function is able to discriminate between groups.

3. Computational algorithm

- Computational sequence
  - Input parameters: distance, apparent magnitude, velocity, right ascension, declination and colour index.
  - (1) Using the discriminant analysis to convert the observation vector \( (y) \) for each group of formula (1).
  - (2) Compute the matrix \( E^{-1}H \).
  - (3) Solve the Equation (2) to get the vector \( (a) \) and the Eigenvalues.
  - (4) Compute the largest Eigenvalues from Equation (3).
  - (5) Compute the standardisation discriminant coefficients from Equation (6).
  - (6) Compute the unstandardization discriminant coefficients from Equation (7).
  - (7) Test the validity (accuracy) of the discriminant function by the following:
    - 7–1 find the discriminant coefficients of Equation (4) to classify the individuals in its different groups.
    - 7–2 test the ability of the discriminant function between groups of some indicators such as Eigenvalues from Equation (3) and Canonical correlation (spare of the correlation coefficient).
    - 7–3 Compute Wilkes Lambda from Equation (9).
- The algorithm is completed

4. Results and discussions

4.1. Data

We used the data taken from the lists of HCGs (Hickson 1982, 1993) and Hickson et al. (1992) which gives the group individuals’ data as equatorial coordinates, the heliocentric radial velocity of the galaxy, estimated standard error of the radial velocity, the type of the galaxy (Hubble Type), length of the semi-major and semi-minor axes and the asymptotic magnitude corrected for the internal and external extinction.

Tables 1 and 2 are the frequency distribution of the number of groups of galaxies and the correlation matrix between variables respectively.

4.2. Selection of the independent variables

We will select the independent variables that are used to find the suggested model that discriminate between the individuals of Hickson’s galaxy groups according to the highest and lowest values of \( F \) and Wilkes Lambda respectively as listed in the Table 3:

We noted from the Table 3 that, all variables are very important in the discrimination between groups because the values of the calculated \( F \) calculated (statistic) amounting more than the minimum of the crisis to enter the variable in the analysis. In addition, the values of the Wilks’ Lambda are very small and are equal to zero.

4.3. Determination of the discriminant standardised coefficients

Table 4 illustrates of the measure the actual contribution of independent variables for the dependent variable has been the composition of the discriminant function.

We noted from Table 4 that, the independent variables (right ascension), which amounted to the standard value (1.006) was a major contributor in proving the first discriminant function. The second discriminant function (distance) the major contributors to the composition of that function with the standard value is (2.218), and the largest contributor to the third discriminant function is the (declination) with Standard value is (1.001). colour index and velocity are the major contributions in 4\textsuperscript{th} and 5\textsuperscript{th}
Table 2. Parameters correlations matrix.

|                        | R_A_Degree | Dec._Degree | V        | Distance | m         | Colour Index |
|------------------------|------------|-------------|----------|----------|-----------|-------------|
| Correlation            | 1.000      | 0.588       | 0.088    | 0.086    | 0.016     | 0.059       |
| Dec._Degree            | 0.588      | 1.000       | −0.019   | 1.000    | 0.929     | 1.000       |
| V                      | 0.088      | −0.019      | 1.000    | 0.929    | 1.000     | 1.000       |
| Distance               | 0.086      | −0.026      | 0.892    | 1.000    | 0.075     | 0.192       |
| m                      | 0.016      | 0.033       | 0.127    | 0.075    | 1.000     | 0.473       |
| Colour Index           | 0.059      | 0.060       | 0.208    | 0.192    | 0.473     | 1.000       |

Table 3. Wilks' Lambda.

| Step | No. of Variables | Lambda | df1 | df2 | df3 | Statistic | df1 | df2 | Sig. |
|------|------------------|--------|-----|-----|-----|-----------|-----|-----|------|
| 1    | 1                | 0.000  | 1   | 91  | 291 | 63365634.73  | 91  | 291 | .000 |
| 2    | 2                | 0.000  | 2   | 91  | 291 | 14246079.46  | 182 | 580 | .000 |
| 3    | 3                | 0.000  | 3   | 91  | 291 | 18342467.82  | 334 | 1024 | .000 |
| 4    | 4                | 0.000  | 4   | 91  | 291 | 18342467.82  | 334 | 1024 | .000 |
| 5    | 5                | 0.000  | 5   | 91  | 291 | 18342467.82  | 334 | 1024 | .000 |
| 6    | 6                | 0.000  | 6   | 91  | 291 | 18342467.82  | 334 | 1024 | .000 |

Table 4. Standardised canonical discriminant function coefficients.

| Function          | 1      | 2      | 3      | 4      | 5      | 6      |
|-------------------|--------|--------|--------|--------|--------|--------|
| R_A_Degree        | 1.006  | 0.042  | −0.013 | −0.003 | 0.000  | 0.000  |
| Dec._Degree       | −0.048 | 0.558  | 1.001  | 0.004  | −0.003 | −0.001 |
| V                 | 0.032  | −1.969 | −0.116 | 1.166  | 1.029  | 0.054  |
| Distance          | −1.228 | 2.218  | −0.088 | 0.004  | −0.023 | −0.002 |
| m                 | 0.011  | 0.118  | −0.010 | −0.509 | −1.733 | 1.000  |
| Colour Index      | −0.042 | −0.077 | −0.064 | 1.151  | −0.010 | −0.018 |

Table 5. Canonical discriminant function coefficients.

| Function          | 1      | 2      | 3      | 4      | 5      | 6      |
|-------------------|--------|--------|--------|--------|--------|--------|
| R_A_Degree        | 36.701 | 1.525  | −4.60  | −0.94  | 0.007  | 0.003  |
| Dec._Degree       | −47.3  | 0.25   | 9.920  | 0.042  | −0.025 | −0.006 |
| V                 | 0.000  | −0.006 | 0.000  | −0.001 | 0.003  | 0.000  |
| Distance          | −17.77 | 30.692 | −1.220 | −0.052 | −0.317 | −0.030 |
| m                 | 0.31   | −0.12  | 0.31   | 0.520  | 0.177  | 1.022  |
| Colour Index      | 0.076  | 0.139  | 0.011  | 0.2087 | −0.019 | −0.032 |
| (Constant)        | 691.5693 | −3332.23 | 73.32 | 12.340 | 364.0 | −14.594 |

4.4. Determination of the discriminant unstandardized coefficients

To get the discriminant equations, we use the unstandardization discriminant coefficients as shown in Table 5.

In Table 5, we extracted the classification signs for Hickson groups, as well as a forecast of the situation with regard to the classification by the proposed model. We have the same classification of Hickson, then the Hickson classification is true.

4.5. The ability of the discriminant function to distinguish between groups

To discover the function of discrimination between the groups that have the ability to rely on Eigenvalues and the ratio of interpreter variance and canonical correlation coefficient in order to discriminate the significant statistical functions for Hickson’s groups, we construct Table 6.

From Table 6, we note that the ratio of the interpreter variance for the first discriminant function had reached 80.6%, in addition, 19.1 and 0.3 for the second and third discriminant function and the correlation coefficient are 1. This means the function is high goodness of fit, and the determinant coefficient equals 1. This means that the proposed model has succeeded by 100% in discriminating groups and this is illustrated in Table 7.

We note from Table 7, that all the discriminant functions are significant under the significant level (0.05), because of (p-value = Sig.) equal to (0.000) less than the significance level (0.05).

4.6. Accuracy of the model

To test the accuracy of the model we must make sure that the value of Kaba is greater than or equal to 0.7,
Table 7. Wilks’ Lambda.

| Test of Function(s) | Wilks’ Lambda | Chi-square | df | Sig. |
|---------------------|---------------|------------|----|------|
| 1 through 6         | 0.00          | 17709.739  | 546 | .000 |
| 2 through 6         | 0.00          | 11509.942  | 450 | .000 |
| 3 through 6         | 0.00          | 6390.269   | 356 | .000 |
| 4 through 6         | 0.00          | 2668.612   | 264 | .000 |
| 5 through 6         | 0.029         | 1176.887   | 174 | .000 |
| 6                   | 0.577         | 183.362    | 86  | .000 |

Table 8. Symmetric measures.

| Measure of Agreement | Kappa | Asymp. Std. Error | Approx. T | Approx. Sig. |
|----------------------|-------|-------------------|-----------|--------------|
| N of Valid Cases     | 383   | 1.00              | 181.888   | .00          |

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Table 8 shows that the value of the Kaba is 1.00, which refers to the highest forecast accuracy. Then the model is excellent to discriminate between the Hickson groups.

From the above, we found that all groups have the correct classification and coincide with the Hickson classification.

5. Conclusion

In this paper, we introduced a procedure to predict the discrimination between Hickson’s compact groups of galaxies. Discrimination analysis is used to guess the discrimination between the groups that suffer from problems in the classification of some primeval galaxies and we get the following:

(1) The study pointed out that the right ascension has a major contribution in the first discriminant function.
(2) For the second discriminant function, distance has a substantial contribution.
(3) For the third discriminant function, declination has a substantial contribution.
(4) For the fourth discriminant function, the colour index has a substantial contribution.
(5) For the fifth discriminant function, the velocity has a substantial contribution.
(6) For the sixth discriminant function, the apparent magnitude has a substantial contribution.
(7) All groups have a correct classification and coincide with the Hickson classification.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Allam S, Tucker D. 2000. Compact groups of galaxies in the Las Campanas redshift survey. AN. 321:101.
de Carvalho RR, Gonçalves TS, Iovino A, Kohl-Moreira JL, Gal RR, Djoygovero SG. 2005. A catalog of distant compact groups using the digitized second palomar observatory sky survey. AJ. 130:425.
Giuricin G, Marinoni C, Ceriani L, Pisani A. 2000. Nearby optical galaxies: selection of the sample and identification of groups. ApJ. 543:178G.
Hickson P. 1982. Systematic properties of compact groups of galaxies. ApJ. 255:382.
Hickson P. 1993. Atlas of Compact Groups of Galaxies (Special Issue). Astrophys Lett Commun. 29:1.
Hickson P, Mendes de Oliveira C, Huchra JP, Palumbo G. 1992. Dynamical properties of compact groups of galaxies. ApJ. 399:353H.
Hintze J. 2007, “NCSS statistical system” Kaysville Utah.
Holmberg E. 1950. A photometric study of nearby galaxies. Medd Lunds Obs Ser. 2(128):5.
Humason ML, Mayall. NU, Sandage AR. 1956. Redshifts and magnitudes of extragalactic nebulae. AJ. 61:97.
Lee BC, Allam S, Tucker DL, Annis J, Johnston D, Scranton R, Acoy E, Bahcall NA, 14 coauthors. 2004. A Catalog of Compact Groups of Galaxies in the SDSS Commissioning Data. AJ. 127:1811L.
Leech NL, Barrett KC, Morgan GA. 2005. SPSS for intermediate statistics: Use and interpretation. Mahwah (NJ): Lawrence Erlbaum.
Merchán M, Zandivarez A. 2002. Galaxy groups in the 2dF galaxy redshift survey: the catalogue. MNRAS. 335:216M.
Ramella M, Armando. P, Geller MJ. 1997. Groups of galaxies in the Northern CfA redshift survey. AJ. 113:483R.
Ramella M, Geller MJ, Pisani A, Da Costa LN. 2002. The UZC-SSRS2 group catalog. AJ. 123:2976R.
Rencher AC. 2002. Methods of multivariate analysis. 2nd. Brigham Young University, John Wiley & Sons, Inc.
Rose JA. 1977. A survey of compact groups of galaxies. ApJ. 211:311.
Tucker DL, Oemler A Jr., Hashimoto Y, Shectman SA, Kirshner RP, Lin H, Landy SD, Schechter PL, Allam SS. 2000. Loose groups of galaxies in the Las Campanas redshift survey. ApJS. 130:237T.
Wolfgang H. 2003. Applied multivariate statistical analysis. Berlin and Louvain-la- Neuve.
Wang Y, Yang X, Mo HJ, van Den Bosch FC, Weinmann SM, Chy Y. 2008. The clustering of SDSS galaxy groups: mass and color dependence. ApJ. 687:919.