Principal Component Analysis and Quasar Identification Techniques

Angelica Rivera

March 15, 2016

Abstract

Principal Component Analysis (PCA) is one of the most common and useful data analysis techniques to perform on a set of observations with variables that may be correlated with one another. PCA can extract the most important relationships in a data set by projecting the data into an orthogonal space where the weighted eigenvectors describe the amount of variance in the data set. These eigenvectors are obtained by the singular value decomposition of the original data set, and are composed of linear coefficients which will project the original observables into the new orthogonal space. The linear combinations resulting from this multiplication are called factor scores. The most strongly correlated observables will have factor scores that are largest in magnitude.

Although there are several ways to execute PCA, this paper will focus on the PCA of a correlation matrix in order to extract emission−line ratios most relevant to the classification of radio−loud vs. radio−quiet quasars (Boroson and Green 1992). I will be using a subset of quasar data from the aforementioned paper in order to confirm their results and to give a clearer illustration of the methods of PCA.

1 Difficulties in Quasar Identification

Historically there have been a number of difficulties in accurate quasar (quasi−stellar object) identification. These objects can be extremely old (they range from redshifts $z < 0.1$ to $z > 3$) and contain supermassive black holes that are actively (at their own present times) consuming large amounts of matter. As a result, quasars emit relativistic jets of energy perpendicular to the disk, and appear as point sources in the sky (when the observer is able to view the jet either head−on or at an angle). Quasars may also

be obscured by an accretion disk composed of dust/infalling matter. Their photometric outputs may therefore vary over time. As such, even radio observations may be found to be faulty in determining whether the output of a quasar is truly radio-loud or radio-quiet. A more precise measure of radio output was therefore required to confirm this classification. Boroson and Green (1992) recorded a series of emission line observations consisting of 87 quasars (all of which were at redshift < 0.5) across a variety of wavelengths, in the hopes of determining ratios of line-emission strengths or equivalent widths that would be reliable indicators of the strength of a quasar's radio emission. All observations were taken at Kitt Peak National Observatory, using the 2.1 meter telescope and the Gold Spectrograph, with a TI $800x800$ CCD camera. Two 300 g*mm^{-1} gratings were used to take into account the various redshifts of the sources. One grating was blazed at 4000 Angstroms and the other at 6750 Angstroms. I selected a subset of 25 quasars (including 9 radio-loud, 12 radio-quiet, and 4 flat spectrum quasars) on which to perform a PCA analysis. A complete list of these quasars is included in Table 1, and the corresponding observations are listed in Tables 2 and 3. Targets chosen had observations of all potential observables (as several of those used by Boroson and Green were missing an α_{ox} measure).

2 Observables and Weighting

PCA analysis is always enacted on a matrix C with a observations, and b variables. If a correlation PCA is desired, each of the columns of C must be normalized such that each of the b columns have averages equal to 0. To do this, one must subtract the average of column b from each of the elements of the column (Abdi & Williams). Additionally, if the units of the observed variables are not uniform, the columns have to be normalized such that each of the variables is divided by it's norm (the square root of the sum of all of the squared elements in the column). Multiplying $C * C^T$ would then give a matrix of correlation coefficients (Tables 4 and 5). Note that the diagonal elements of the correlation matrix will consist solely of 1's, as the correlation coefficient of a variable and itself is equal to 1 (Palmer).

---

3 PCA Analysis and SVD

In order to obtain the eigenvectors for the analysis, it is first necessary to perform singular value decomposition (SVD) on the matrix C. SVD is

a technique commonly used to identify eigenvalues and eigenvectors for a matrix that is not square. In this case, our matrix C would be decomposed into:

$$C = U * S * V^T \tag{1}$$

where S^2 gives the eigenvalues of the matrix C, U is an a x r matrix of left singular vectors (where r is the rank of matrix C) and V^T is a b x r matrix of right singular vectors. (In other words, the columns of U are the eigenvectors of $C * C^T$, and the columns of V are the eigenvectors of $A^T * A$. The eigenvectors that describe the projections of the original variables onto their principal components are those that make up the columns of V. In terms of determining which column/eigenvectors are most relevant in determining correlations between variances, it is first necessary to find the total variance of the data table, where the variance equals the summed squares of each column. PCA calculates principal components (or eigenvectors) which have the property that the first eigenvector is that which describes the largest possible variance, the second being orthonormal to the first and describing the next greatest possible variance, etc. This occurs as the eigenvalues are equal to the summed squares of the factor components that they correspond to (Abdi & Williams). The contribution of a component is therefore the value of that component squared over it's corresponding eigenvalue. The factor scores are obtained by either taking U∗S from the SVD, or from multiplying C by V.

4 Results

As expected, the first five eigenvectors of the analysis describe the greatest amount of variance (%37.4, %21.3, %10.82, %10.28, and %7.8, respectively). The projections of each eigenvector are listed in Table 6. Since the first two eigenvectors hold the most variance, they most clearly reflect the strongest relationships among the original set of observables. Boroson and Green determined that the first two eigenvectors are guided by a strong anticorrelation in FeII and OIII (Eigenvector 1) and the inverse correlation between HeII and optical luminosity, Mv (Eigenvector 2). Indeed today these two eigenvectors are most commonly known as Eigenvector 1 and Eigenvector 2 in the current literature (Richards et. al 2011). My results agree with this determination, as I obtained correlation coefficients of +0.648 and +0.573 for these pairs, respectively. However, the difference between the first pair of projections was much greater than that found by Boroson and Green, and

the second pair had a difference that was much less (*i.e.* -0.14 and -0.669 for FeII and MO[III] and 0.478 and 0.358 for HeII and Mv, respectively. I believe that these differences arise from the use of a smaller subset of data. I attempted to represent a range of quasar radio types, selecting 9 that were radio−loud, 12 that were radio−quiet, and 4 that were flat. Figures 1, 2, and 3 show that the quasars selected represented what would appear to be an adequate range over these values. However, Boroson and Green used a selection of quasars much greater than my own, and which contained many more radio-quiet quasars, whereas in my subset the ratio of radio-loud to radio quiet quasars was about comparable. For the most part, this "selection effect" appears to have affected only the calculation of the RFeII projection.

5 Conclusion

Although Boroson and Green were not the first to use PCA to identify quasars (i.e. the "Baldwin Effect", discovered in 1977, has been used to describe the anticorrelation between the luminosity and equivalent width of CIV), their work was key in determining emission strengths from a variety of quasar types at low redshift. In recent years, a series of other researchers have added observations (including x-ray spectral index) into Boroson and Green's eigenvector matrix to obtain classifications at higher redshifts (Richards et al.). Much progress has been made towards the determination of the strength of a quasar's energy output through various types of emission; now we may turn our attention towards other related questions, such as determining the mass accretion rates of the quasars themselves.

References

[1] Abdi H, Williams, LJ. "Principal Component Analysis", *Wires Computational Statistics* 2 (2010): 433-459. Web. 8 March 2016.

[2] Baldwin, J.A., 1977, *ApJ*, 214, 679

[3] Boroson, T.A., and Green, R.F., 1992, *ApJ*, 80, 109

[4] Palmer, Michael. "A Glossary of Ordination-related terms". okstate.edu. okstate, n.d. Web. 3 March 2016.

[5] Press, William H., et al., *Numerical Recipes in C, The Art of Scientific Computing*, New York: Cambridge University Press, 1988. Web.

[6] Richards, G.T. et al, 2011, *ApJ*, 141, 141

Table 1: Quasar Observations

PG QSO	Redshift	Date Observed	Exposure Time	Radio Classification
0007+106	0.089	Sep18 1990	1200	Flat
1226+023	0.158	Feb16 1990	900	Flat
1302−102	0.286	Apr22 1990	2000	Flat
2209+184	0.70	Sep18 1990	600	Flat
0003+158	0.450	Oct10 1990	3600	Steep
1004+130	0.240	Apr21 1990	3600	Steep
1100+772	0.313	Feb19 1990	1951	Steep
1048−090	0.344	Apr22 1990	2620	Steep
1211+143	0.085	Feb15 1990	750	Steep
1425+267	0.366	Apr23 1991	3600	Steep
2251+113	0.323	Oct11 1990	3600	Steep
1545+210	0.266	Sep19 1990	2400	Steep
1704+608	0.371	Sep20 1990	2423	Steep
0003+199	0.025	Sep08 1990	500	Quiet
0049+171	0.064	Sep18 1990	1800	Quiet
0844+349	0.064	Feb15 1990	1300	Quiet
0934+013	0.05	Apr21 1990	3600	Quiet
1534+580	0.03	Feb16 1990	2400	Quiet
1519+226	0.137	Feb20 1990	3600	Quiet
1435−067	0.129	Feb17 1990	1800	Quiet
1352+183	0.158	Feb20 1990	3000	Quiet
2233+134	0.325	Oct10 1990	3600	Quiet
2214+139	0.067	Sep18 1990	500	Quiet
1552+085	0.119	Feb17 1990	2400	Quiet
1613+658	0.129	Apr23 1990	3600	Quiet

Table 2: Emission-Line Strengths and Properties

PG QSO	Mv	LogR	α_{ox}	EW Hβ	R5007	R4686	RFeII	Peak5007
0007+106	-23.85	2.29	1.06	101	0.42	0.02	0.35	3.07
1226+023	-27.15	3.06	1.32	113	0.04	0.03	0.57	0.33
1302-102	-26.6	2.27	1.49	28	0.33	0	0.6	1.36
2209+184	-23.14	2.15	1.35	115	0.13	0	0.44	1.67
0003+158	-26.92	2.24	1.39	91	0.28	0.16	0	2.7
1004+130	-25.97	2.36	1.92	43	0.15	0	0.23	1.6
1100+772	-25.86	2.51	1.36	90	0.46	0.05	0.21	3.99
1048-090	-25.83	2.58	1.41	81	0.34	0.08	0.09	4.45
1211+143	-24.6	1.39	1.22	84	0.14	0.16	0.52	0.55
1425+267	-26.18	1.73	1.68	93	0.38	0.02	0.11	4.31
2251+113	-26.24	2.56	1.8	82	0.23	0.03	0.32	1.69
1545+210	-25.63	2.62	1.28	96	0.34	0.02	0	3.66
1704+608	-26.38	2.81	1.6	28	0.94	0	0	6.5
0003+199	-22.14	-0.57	1.25	95	0.23	0.28	0.62	0.8
0049+171	-21.81	-0.49	1.24	136	0.72	0.03	0	3.99
0844+349	-23.31	-1.52	1.53	76	0.1	0.14	0.89	0.55
0934+013	-21.43	-0.42	1.29	92	0.55	0.32	0.48	1.89
1534+580	-21.44	-0.15	1.27	97	0.81	0.4	0.27	5.31
1519+226	-23.76	-0.05	1.48	105	0.03	0.06	1.01	0.16
1435-067	-24.1	-1.15	1.4	142	0.09	0.05	0.45	0.59
1352+183	-24.13	-0.96	1.41	133	0.07	0.06	0.46	0.58
2233+134	-25.18	-0.55	1.64	67	0.17	0.05	0.89	0.77
2214+139	-23.39	-1.3	1.83	107	0.08	0.03	0.32	0.87
1552+085	-23.72	-0.35	1.69	46	0.06	0.05	1.02	0.22
1613+658	-24.22	0	1.47	110	0.18	0.02	0.38	1.99

Where Mv is the absolute visual magnitude, LogR the ratio of radio to optical flux density, and α_{ox} the X-ray to optical spectral index.

Table 3: Emission-Line Strengths and Properties (cont.)

PG QSO	Hβ FWHM*	Hβ shift	Hβ shape	Hβ asymm	$MO[III]^{+}$
0007+106	5100	0.18	1.05	-0.046	-27.91
1226+023	3520	0.038	1.142	0.044	-28.85
1302-102	3400	0.027	1.021	-0.024	-29.02
2209+184	6500	-0.07	1.192	0.051	-26.1
0003+158	4760	-0.46	1.143	-0.163	-30.44
1004+130	6300	0.169	1.355	0.065	-28
1100+772	6160	0.063	1.107	-0.097	-29.89
1048-090	5620	0.069	1.218	-0.224	-29.44
1211+143	1860	0.012	1.151	-0.003	-27.26
1425+267	9410	0.052	1.204	-0.052	-30.06
2251+113	4160	-0.135	1.216	-0.083	-29.45
1545+210	7030	0.086	1.184	-0.095	-29.42
1704+608	6560	0.042	1.28	-0.288	-29.95
0003+199	1640	-0.043	1.198	0.068	-25.49
0049+171	5250	0.021	1.058	-0.047	-26.78
0844+349	2420	0.068	1.099	0.059	-25.53
0934+013	1320	-0.067	1.205	-0.084	-25.7
1534+580	5340	-0.032	1.024	0.044	-26.18
1519+226	2220	0.041	1.104	0.095	-25.15
1435-067	3180	-0.028	1.126	0.029	-26.84
1352+183	3600	0.023	1.072	-0.021	-26.62
2233+134	1740	-0.015	1.181	0.071	-27.85
2214+139	4550	0.119	1.248	0.164	-25.76
1552+085	1430	-0.01	1.203	0.069	-24.88
1613+658	8450	-0.056	1.155	-0.207	-27.47

$*-$full width at half maximum.
+-Forbidden transition.

Table 4: Correlation Matrix for Observed Properties

Property	Mv	$LogR$	α_{ox}	EW Hβ	R5007	R4686	RFeII	Peak5007
Mv	+1.000	-0.726	-0.347	+0.406	+0.149	+0.573	+0.255	-0.110
$LogR$	-0.726	+1.000	-0.042	-0.357	+0.215	-0.364	-0.490	+0.430
α_{ox}	-0.347	-0.042	+1.000	-0.458	-0.268	-0.385	+0.110	-0.160
EW Hβ	+0.406	-0.357	-0.458	+1.000	-0.188	+0.093	-0.121	-0.151
R5007	+0.149	+0.215	-0.268	-0.188	+1.000	+0.282	-0.588	+0.879
R4686	+0.573	-0.364	-0.385	+0.093	+0.282	+1.000	+0.098	+0.061
RFeII	+0.255	-0.490	+0.110	-0.121	-0.588	+0.098	+1.000	-0.771
Peak5007	-0.110	+0.430	-0.160	-0.151	+0.879	+0.061	-0.771	+1.000
Hβ FWHM	-0.316	+0.470	+0.111	+0.071	+0.341	-0.391	-0.730	+0.665
Hβ shift	+0.080	-0.006	+0.083	-0.106	+0.005	-0.324	+0.094	+0.045
Hβ shape	-0.260	+0.215	+0.621	-0.367	-0.117	-0.206	-0.174	+0.024
Hβ asymm	+0.383	-0.493	+0.173	+0.169	-0.535	+0.113	+0.621	-0.664
MOIII	+0.839	-0.791	-0.076	+0.265	-0.301	+0.370	+0.648	-0.540

Table 5: Correlation Matrix for Observed Properties (cont.)

Property	Hβ FWHM	Hβ shift	Hβ shape	Hβ asymm	M[OIII]
Mv	-0.316	+0.080	-0.260	+0.383	+0.839
$LogR$	+0.470	-0.006	+0.215	-0.493	-0.791
α_{ox}	+0.111	+0.083	+0.621	+0.173	-0.076
EW Hβ	+0.071	-0.106	-0.367	+0.169	+0.265
R5007	+0.341	+0.005	-0.117	-0.535	-0.301
R4686	-0.391	-0.324	-0.206	+0.113	+0.370
RFeII	-0.730	+0.094	-0.174	+0.621	+0.648
Peak5007	+0.665	+0.045	+0.024	-0.664	-0.540
Hβ FWHM	+1.000	+0.124	+0.174	-0.485	-0.556
Hβ shift	+0.124	+1.000	+0.073	+0.252	+0.152
Hβ shape	+0.174	+0.073	+1.000	-0.090	-0.137
Hβ asymm	-0.485	+0.252	-0.090	+1.000	+0.662
MOIII	-0.556	+0.152	-0.137	+0.662	+1.000

Table 6: PCA Eigenvectors

Property	1st Eigenvector	2nd Eigenvector	3rd Eigenvector	4th Eigenvector	5th Eigenvector
Eigenvector variance	% 37.4	% 21.3	% 10.82	% 10.28	% 7.8
Mv	-0.292	+0.358	+0.031	-0.133	+0.246
LogR	+0.362	-0.129	-0.456	+0.273	+0.383
α_{ox}	-0.301	+0.193	-0.356	+0.412	-0.366
EW Hβ	+0.152	-0.091	-0.022	+0.420	-0.078
R5007	-0.122	+0.258	-0.308	-0.422	+0.174
R4686	+0.044	+0.478	-0.425	+0.126	-0.179
RFeII	-0.140	+0.168	+0.247	-0.005	+0.037
Peak 5007	+0.285	+0.392	-0.129	-0.262	-0.123
Hβ FWHM	-0.189	-0.227	+0.029	-0.040	-0.404
Hβ shift	+0.142	-0.140	-0.115	-0.485	-0.237
Hβ shape	-0.160	-0.508	-0.537	-0.237	-0.010
Hβ asymm	+0.155	-0.005	-0.008	+0.042	+0.498
MOIII	-0.669	-0.063	-0.090	+0.065	+0.325

Figure 1:

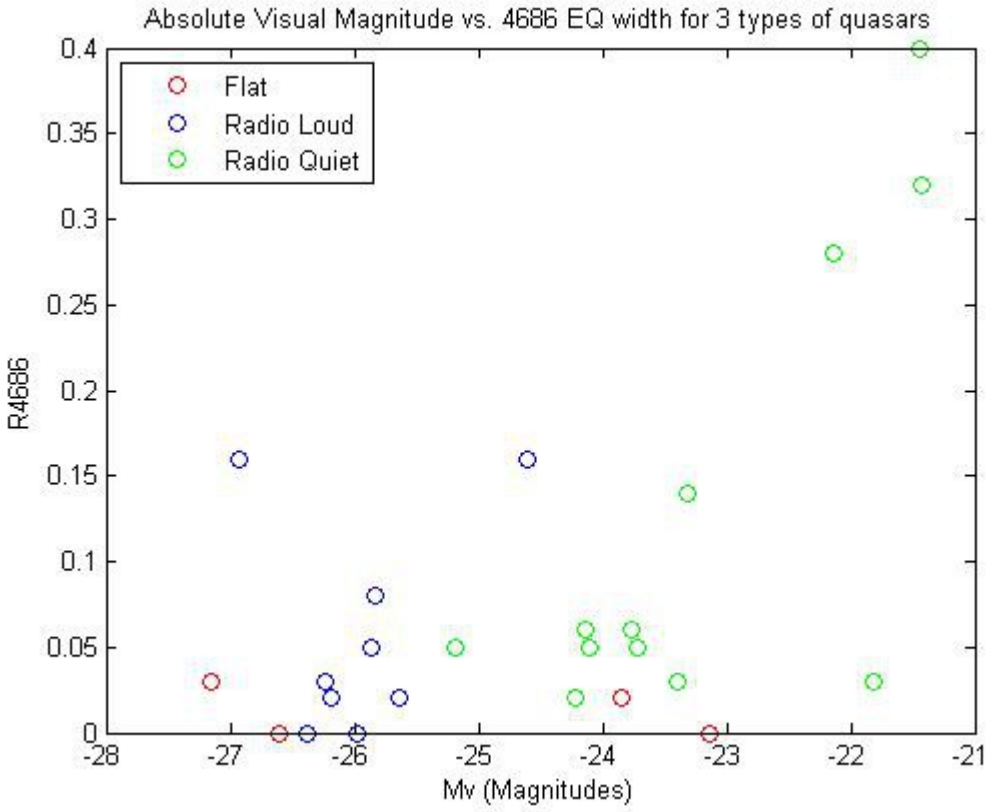

Absolute Visual Magnitude vs. 4686 EQ width for 3 types of quasars
Flat
Radio Loud
Radio Quiet
R4686
Mv (Magnitudes)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-28
-27
-26
-25
-24
-23
-22
-21

Figure 2:

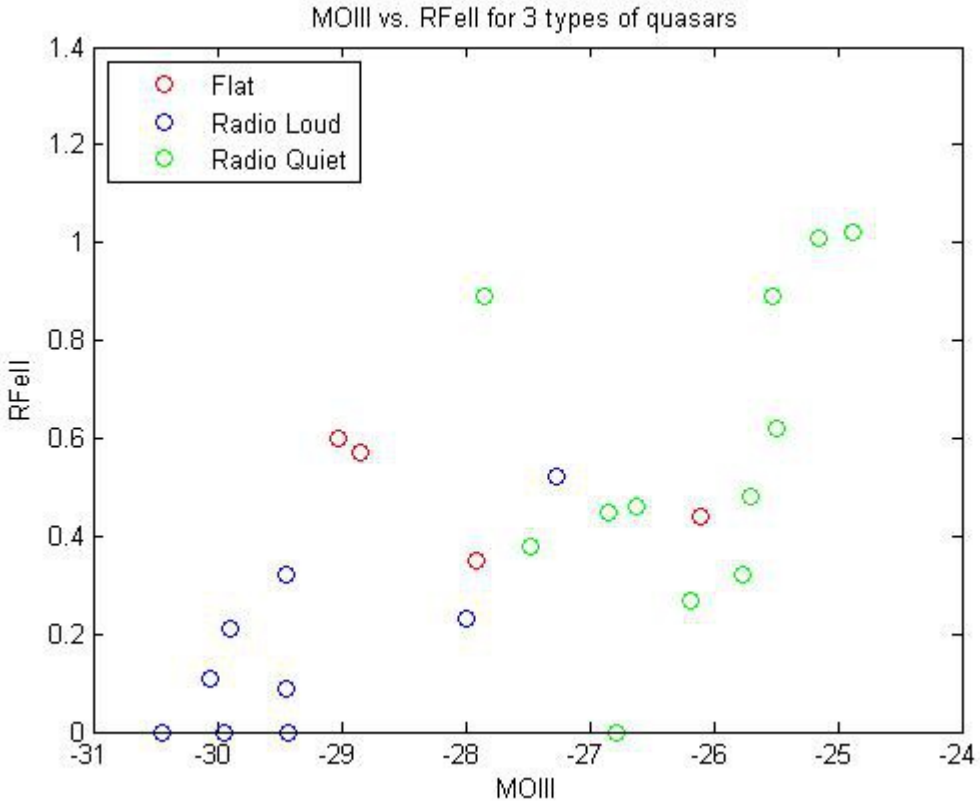
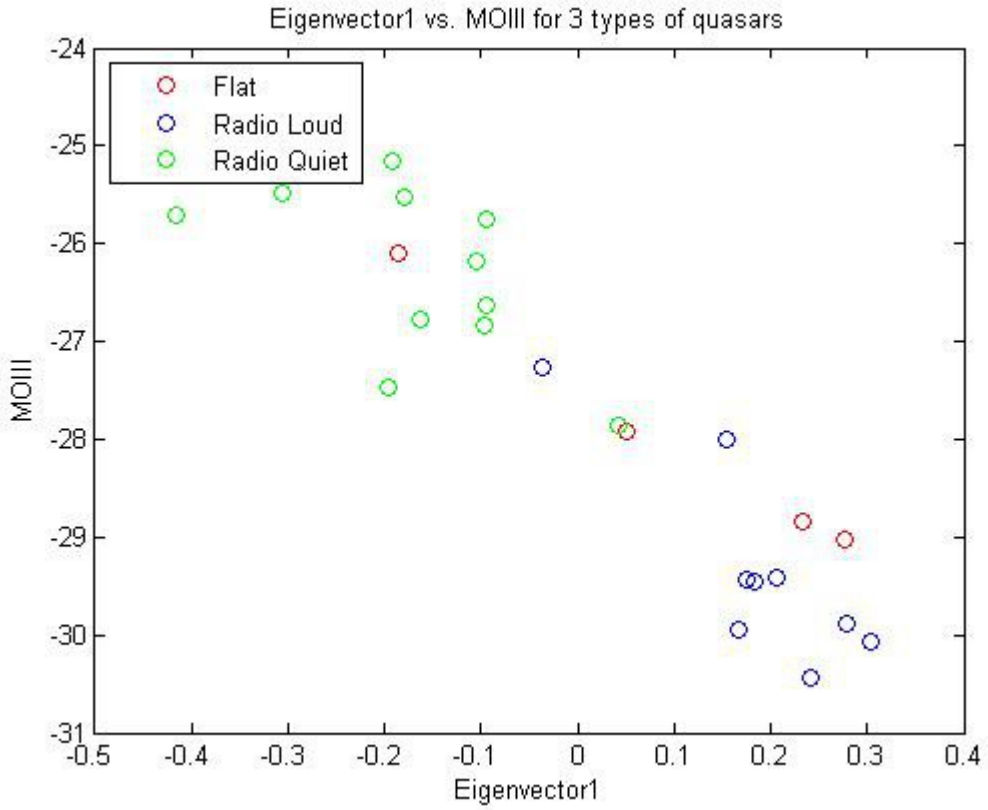

MOIII vs. RFeII for 3 types of quasars
Flat
Radio Loud
Radio Quiet
RFeII
MOIII
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-31
-30
-29
-28
-27
-26
-25
-24

Figure 3:


Eigenvector1 vs. MOIII for 3 types of quasars
Flat
Radio Loud
Radio Quiet
MOIII
Eigenvector1
-24
-25
-26
-27
-28
-29
-30
-31
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4