Automated Data Analysis in Astronomy, **69-82**

Principal Component Analysis and its Application to Stellar Spectra

Harinder P. Singh

Department of Physics, Sri Venkateswara College, University of Delhi, New Delhi - 110 021, India.
email: hps@ttdsvc.ernet.in

Coryn A.L. Bailer-Jones

Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany.

Ranjan Gupta

IUCAA, P.O.Box 4, Ganeshkhind, Pune - 411 007, India.

Abstract. Principal Component Analysis (PCA), also known as the Karhunen-Loéve or Hotelling transform, provides an elegant method for achieving reduction in dimensionality of a data set. The reduced data set can then be used as input to another analysis, e.g., involving artificial neural nets. Sometimes, the principal components are an end in themselves and may be subjected to interpretation. PCA identifies linear combinations of raw parameters accounting for maximum variance in a data set. A compression of the data is obtained by ignoring those components that represent the least variance in the data. In this paper, we describe the transform which reduces a general $n \times m$ data array into an $n \times p$ array (where $p << m$) and review some applications of the principal component analysis with special emphasis on classification of stellar spectra.

Key words: principal component analysis, numerical methods, data analysis, classification, spectra

1. Introduction

Principal Component Analysis has been used in several areas of astronomy, including stellar spectral classification (Deeming 1964, Whitney 1983), galaxy spectral classification

(Connolly et al. 1995, Folkes, Lahav & Maddox 1996, Ronen, Aragón-Salamanca & Lahav 1999), quasar spectral classification (Francis et al. 1992) and as a data compression tool for further analysis by neural nets in stellar and galaxy spectral classification (Storrie-Lombardi et al. 1994, Lahav et al. 1996, Singh, Gulati & Gupta 1998, Bailer-Jones 1996, Bailer-Jones, Irwin & von Hippel 1998a).

Deeming (1964) used the PCA technique for the formulation of classification parameters from the spectrophotometric measurements on stellar spectra. A specific case of the late-type giants was used for testing and classification parameters were reported for 84 G and K giants. The technique was found to be not only effective in giving the formulation of the parameters but also in indicating the number of parameters necessary to describe the sample of stars.

Whitney (1983) used PCA for spectral classification of a set of 53 A and F stars. His data set consisted of a 53×47 matrix with 47 photoelectric measurements for each star over the wavelength range 3500$\mathring{A}$ to 4000$\mathring{A}$. He applied PCA to this data set and then performed a regression on the 3 most significant components, achieving an average classification error of 1.6 spectral subtype.

2. Derivation of the principal components

It is desirable to reduce the dimensionality of an $n \times m$ data array to obtain an $n \times p$ array where $p << m$. We can use the principal component analysis to achieve this.

The principal component analysis transforms the original set of m variables by way of an orthogonal transformation to a new set of uncorrelated variables or principal components. The technique amounts to a straight forward rotation from the original axes to the new ones and the principal components are derived in decreasing order of importance. A successful derivation means that the few p components account for most of the variation in the original data (Chatfield & Collins 1980, Murtagh & Heck 1987).

Let f_{ij} be the flux in the j^{th} flux bin for the i^{th} star for a total of m flux values and n stars. Defining the elements of the $n \times m$ matrix, $\mathbf{X}$, by x_{ij}, we have

$$x_{ij} = \frac{f_{ij} - \overline{f}_j}{s_j \sqrt{n}}, \tag{1}$$

with

$$\overline{f}_j = \frac{1}{n} \sum_{i=1}^{n} f_{ij}, \tag{1a}$$

and

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (f_{ij} - \overline{f}_j)^2, \tag{1b}$$

where $\overline{f}_j$ is the mean flux in the j^{th} bin and s_j is its standard deviation. The PCs, $\mathbf{e}$, are defined as those vectors in this space for which the data projection onto them is maximised. Mathematically, the PCs are those vectors which maximise

$$(\mathbf{X}.\mathbf{e})^T(\mathbf{X}.\mathbf{e}) - \lambda(\mathbf{e}^T.\mathbf{e}), \tag{2}$$

where λ is a Lagrange multiplier introduced to ensure that the PCs are normalised. This reduces to the eigenvalue equation

$$\mathbf{C}\,\mathbf{e_1} = \lambda_1\,\mathbf{e_1}. \tag{3}$$

where C is the correlation matrix with elements

$$C_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik} = \frac{1}{n}\sum_{i=1}^{n}(f_{ij} - \overline{f}_j)(f_{ik} - \overline{f}_k)/s_j\,s_k. \tag{4}$$

The maximum variance is given by the largest eigenvector $\mathbf{e_1}$ associated with the largest eigenvalue λ_1. The next (second) axis is to be orthogonal to the first and another solution of equation (3) gives the second largest eigenvalue λ_2 and the corresponding eigenvector or principal component $\mathbf{e_2}$.

3. Application of PCA to Stellar Spectra: Two Case Studies

We apply the PCA to two sets of stellar spectra, both representing a wide range of normal spectral types (O–M) and luminosity classes. However, they differ in their source, number and resolution, and hence show some interesting differences.

3.1. First case study

In this first case, the data consist of 213 spectra (55 of O to M type stars from Silva & Cornell (1992) and 158 spectra from Jacoby, Hunter & Christian (1984) libraries. The 213 spectra were preprocessed to yield 659 flux values for each spectrum at a resolution of 11Å(Singh et al. 1998). Table 1 gives the eigenvalues of this 213×659 data array in decreasing order of importance as explained by the percentage of variance corresponding to each eigenvalue.

For the correlation matrix C_{jk}, the diagonal terms are all unity. Thus the sum of the diagonal terms or the sum of the variances of the standardized variables is equal to m (659). Therefore, the sum of the eigenvalues is also equal to m. Hence, the proportion of the total variation accounted for by the jth component is λ_j/m. For example, corresponding to second principal component we have $\lambda_2 = 194.52$ and hence the second component accounts for λ_2/m or $\sim$ 29% of the variance.

The first twenty principal components account for 99.9 % of the variance of the difference spectra about the mean spectrum. The first twenty components thus give a

data compression of 33 : 1. Even the first two components account for 95.8 % of the variance and if used will give an impressive data compression of 330 : 1. An enormous reduction in the dimensionality of the data array has been achieved.

Table 1. Percentage of variance for the first 20 eigenvectors for the data in the first case study

Config. No.	Eigenvalue	Variance (as %)	Cumul. Percentage
1.	437.0134	66.3146	66.3146
2.	194.5184	29.5172	95.8318
3.	18.3041	2.7776	98.6094
4.	2.3279	0.3532	98.9626
5.	2.1442	0.3254	99.2880
6.	1.4618	0.2218	99.5098
7.	0.8470	0.1285	99.6383
8.	0.4169	0.0633	99.7016
9.	0.3661	0.0556	99.7571
10.	0.2081	0.0316	99.7887
11.	0.1678	0.0255	99.8142
12.	0.1356	0.0206	99.8347
13.	0.0922	0.0140	99.8487
14.	0.0878	0.0133	99.8621
15.	0.0707	0.0107	99.8728
16.	0.0630	0.0096	99.8823
17.	0.0521	0.0079	99.8903
18.	0.0506	0.0077	99.8979
19.	0.0452	0.0069	99.9048
20.	0.0371	0.0056	99.9104

In every application, one needs to make a decision as to how many principal components to retain in order to have an effectively compressed dataset. A Scree graph, which is a plot of λ_j versus j, helps to identify between the "large" and the "small" eigenvalues. However, one should not routinely ignore the smaller components (components corresponding to the smaller eigenvalues) as the last few components may carry information that is useful in some applications. Fig. 1 shows a Scree graph for first ˘20˘ eigenvalues as evaluated in Table 1.

Assuming that the first p principal components are sufficient to retain the behaviour of the original m variables, we now have a $(n \times p)$ matrix E_p of eigenvectors. We also find the projection vector $\mathbf{A}$ onto the p principal components by using

$$\mathbf{A} = \mathbf{x}\,\mathbf{E_p}, \tag{5}$$

where $\mathbf{x}$ is a vector of fluxes defined in equation (1) and can be represented by

$$\mathbf{x} = \mathbf{A}\,\mathbf{E_p^{-1}}. \tag{6}$$

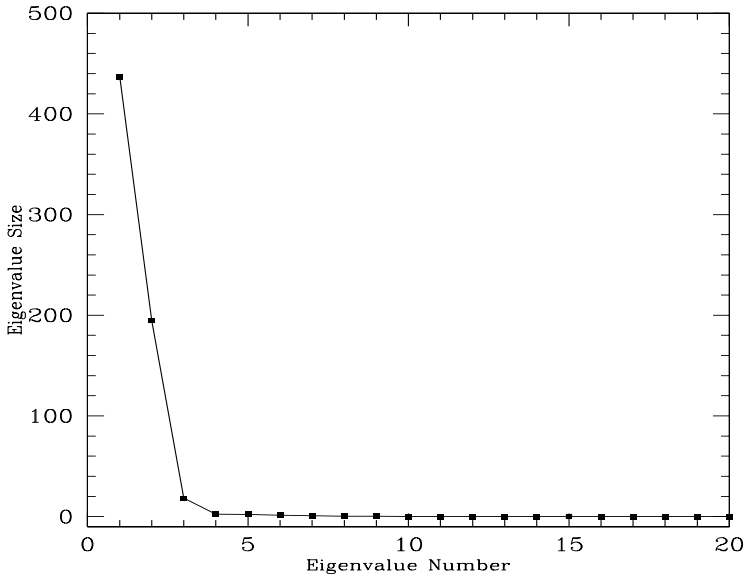

Figure 1. Scree graph for eigenvalues in Table 1 for the 213×659 array (first case study)

We obtain the final spectra $\mathbf{x_{rec}}$ by multiplying $\mathbf{x}$ with $s_j \sqrt{n}$ and adding the mean spectrum. It may be noted that in equation (6), $\mathbf{A}$ is a $(n \times p)$ matrix and E_p^{-1} is a $(p \times m)$ matrix and hence the reconstructed spectra is the original $(n \times m)$ matrix.

In Fig. 2 we show the reconstructed spectra for stars of four different spectral types out of the 158 test spectra by using the first twenty, ten, five, two and one principal components. For each star, the reconstructed spectrum has been offset on an arbitrary scale to avoid overlapping. While the early type stars have more spectral features around 4000 Å, the K type star has large number of spectral features spread in the whole wavelength range. As may be noticed, the first component does not seem to allow us to recover the original spectra. This may be due to non-linearity in the data as the PCA has the inherent weakness of assuming linearity of data. It could also be due to the effect of noise on the deduction of principal components. It has been suggested (Lahav et al. 1996) that the criterion of maximum variance may underestimate the importance of minor principal components, components that contribute significantly less to the cumulative variance.

Another way to test the replicabilty of different principal components is by computing the residue (Δ), the difference between the reconstructed and the original flux values for each wavelength. We define $\Delta\, PC1$ as the residue when only the first principal component is used, $\Delta\, PC2$ when the first two principal components are used for reconstruction, and so on. In Fig. 3 we plot residues taken in pairs. Fig. 3a shows the behaviour of $\Delta\, PC1$ against $\Delta\, PC2$ for the 213 objects. The data points form a line which is tilted towards the x-axis, which denotes $\Delta\, PC1$. This means that the residue from the reconstruction using the first two components is smaller. Fig. 3b shows the plot of $\Delta\, PC1$ against $\Delta\, PC20$. The 213 points arrange themselves in a line parallel to the x-axis as the residue

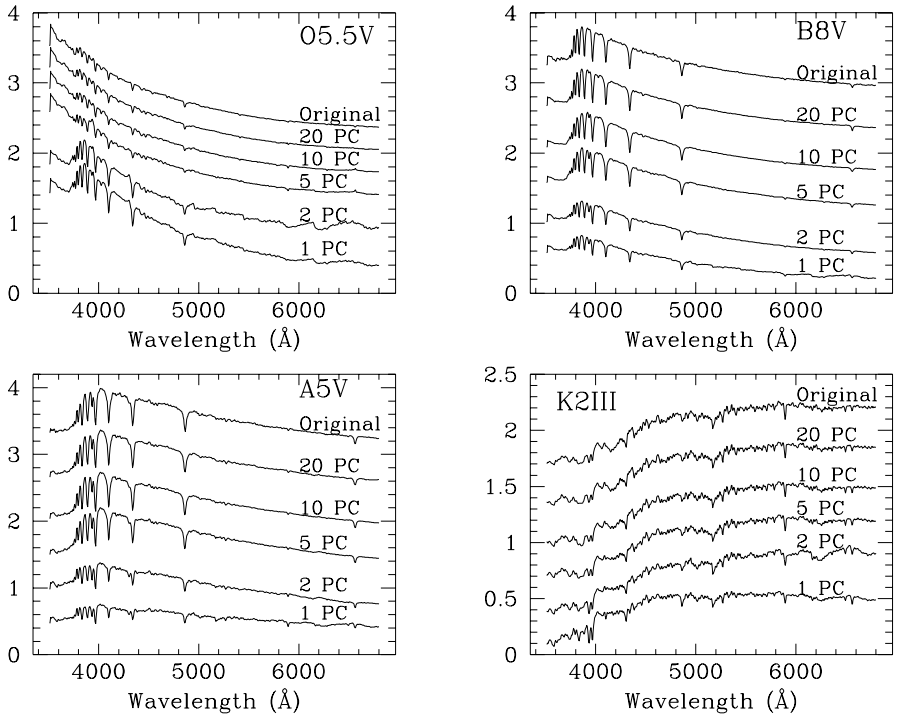

Figure 2. Some reconstructed spectra using 1, 2, 5, 10, and 20 principal components

$\Delta\, PC20$ is close to zero, showing the excellent replicabilty of the first twenty principal components. Figs. 3b and 3c show plots of residue values between PC5 and PC20 respectively. The data points now spread along both the axes showing the approach of $\Delta\, PC5$ and $\Delta\, PC10$ towards $\Delta\, PC20$.

In general, the reconstruction quality, R, varies with the number of PCs used in the reconstruction. This is defined as

$$R = 100\% \frac{\sum_{j=1}^{j=p} \lambda_j}{\sum_{j=1}^{j=m} \lambda_j} \tag{7}$$

The projection of the 213 (n) row-points on the p principal components forms a $(n \times p)$ array. In Fig. 4, the upper and the lower plots show the projections on the first and the twentieth principal components against spectral type respectively (for sp. type coding see G94a). The upper figure shows that the different spectral types occupy distinct regions except for the late type stars. As may be seen from Singh et al. (1998), the ANN using the projections on to the first principal components is unable to satisfactorily classify the late type stars. The lower figure shows that the projections on to the twentieth component are less significant.

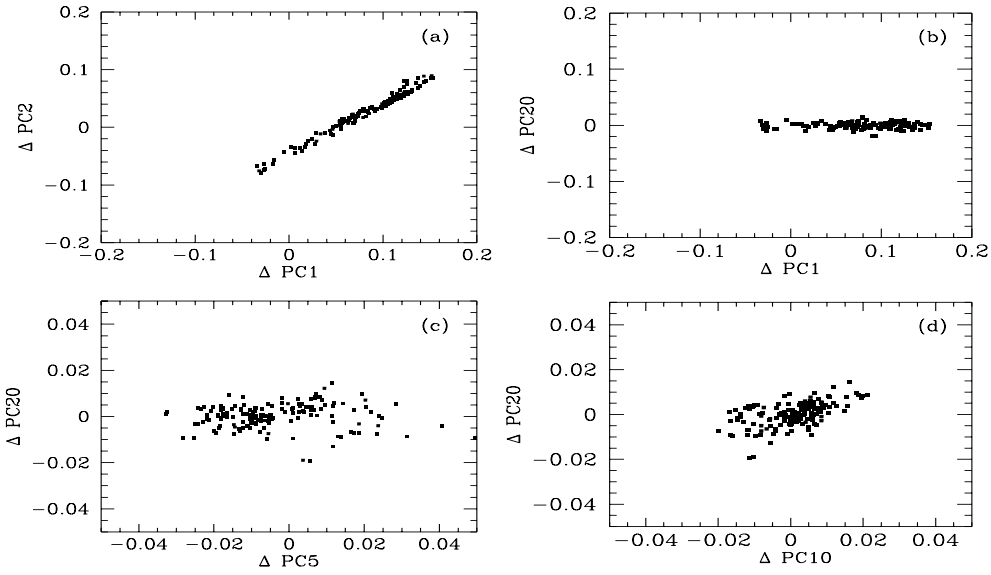

Figure 3. Residuals of reconstructed spectra for the 213 stars plotted against each other

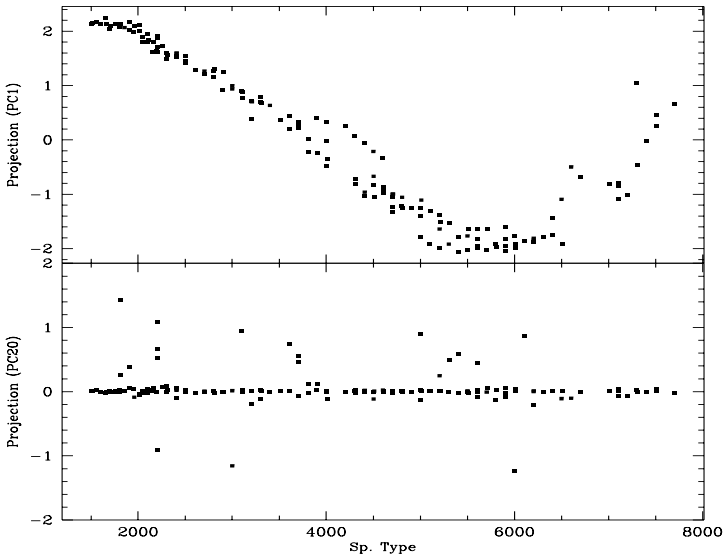

Figure 4. The projection of row-vector on the first principal component (top) and the 20th principal component (bottom)

3.2. Second case study

The data in this case study consist of 5144 stellar spectra taken from scans of objective prism plates from the Michigan spectral survey (Bailer-Jones 1996, Bailer-Jones et al. 1998a,b). Each spectrum has 820 flux elements covering the wavelength range is 3802–5186Å, with the dispersion ranging from 1.1Å/pixel at the blue end to 2.8Å/pixel at the red end. The continuum has not been removed from the spectrum, but the spectra have been normalised to have the same total flux (area under the spectrum). The spectra consist of O–M stars with luminosity classes between III and V.

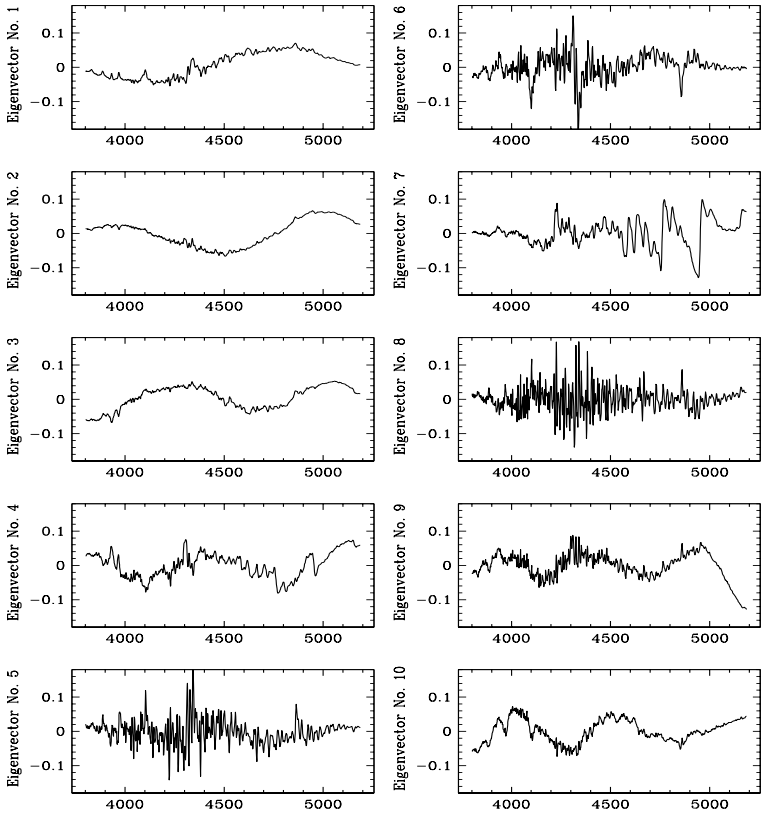

Figure 5. The first ten principal components from the second case study. The principal components are normalized eigenvectors plotted against wavelength (in Angstroms)

Fig. 5 shows the first ten PCs. Both the stellar continuum and individual spectral lines are distributed across many eigenvectors. For example, the Ca II H&K lines at 3934Å

and 3969Å are distinct in the first four eigenvectors as well as the average spectrum. (Note that the sign of the eigenvectors is arbitrary, as the coefficients can be negative.) That the features do not separate into different components is not surprising, as from the physics of line formation we know that a spectrum is not a linear combination of spectral features. However, some features are nonetheless better represented in certain components. For example, the TiO bands, which extend redward from about 4500Å, and which are characteristic of M stars are more strongly represented in the 7^{th} PC.

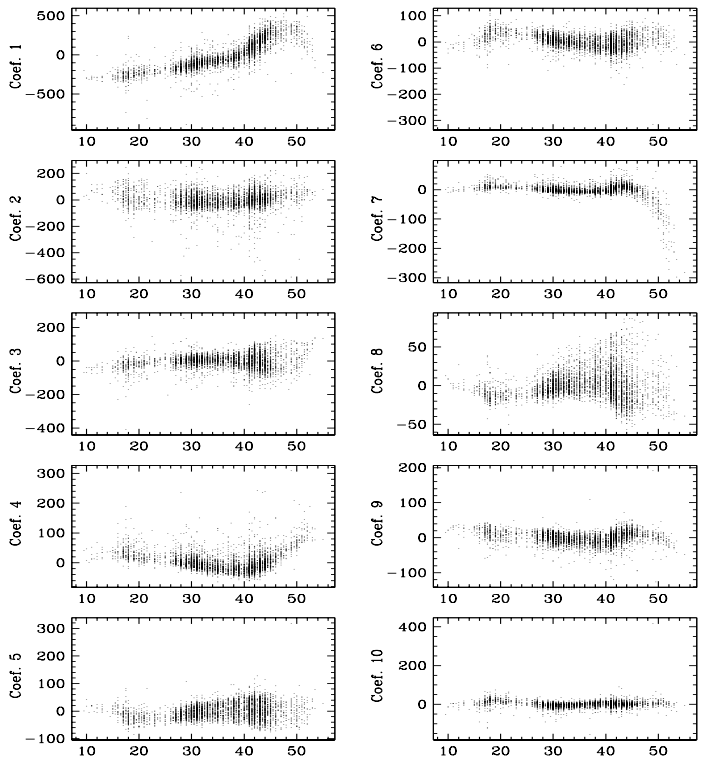

Figure 6. The first ten principal component coefficients from the second case study plotted against spectral type (SpT), where an O3 star has SpT=1 and a M9 SpT=57.

Fig. 6 shows the correlation between the the first ten PC coefficients for all 5144 stars and their spectral type. No single component correlates strongly and linearly with spectral type, which is not surprising as spectral type has a complex dependence on spectral features, whereas the PCs are just linearly related to the original spectra. Some components show some correlation over a limited range of spectral type (for example, the 7^{th} component correlates well for M stars), but such correlations are often quite noisy.

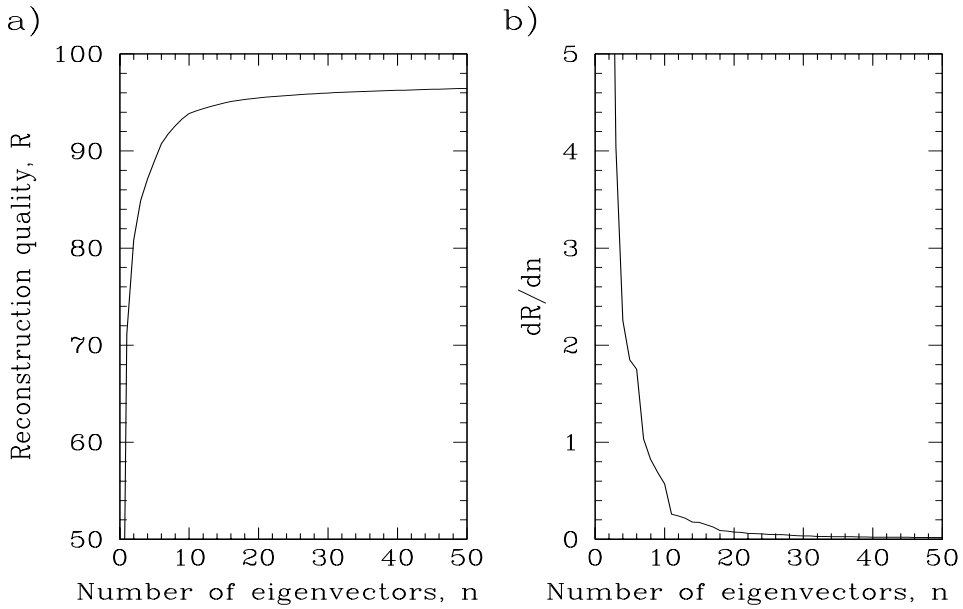

Figure 7. Quality of spectral reconstruction, R, for the data in the second case study As $\frac{dR}{dn} \approx 0$ at $n \sim 25$, the first 25 components can be used for an optimal reconstruction, with the remaining eigenvectors dominated by noise

This implies that nonlinear classifiers could achieve better results than linear ones on these data.

Fig. 7 shows how the reconstruction quality, R, varies with the number of PCs used in the reconstruction. We see that once we have used the first 25 components, the derivative of R is more or less zero, so R improves only very slowly as more components are added. Thus with the first 25 components we can compress the data by a factor of 33 and still explain 95.8% of the variance. Comparing this with Table 1, we see that the data set from the first case study attain this value of R with only the first 3 components, that is with a compression factor of 220, i.e. the present data set is less compressible. This implies that there is more variance in the data in the present case study than in the first one. Although both data sets cover a similar range of spectral types, the second one has many more spectra, 5144 as opposed to 213, and are at higher resolution. Another contribution is noise. The turnover point in Fig. 7 corresponds to where noise begins to dominate the reconstruction. That this turnover occurs once a smaller fraction of the components are used indicates that this has less noise, i.e. higher SNR, than the second data set. This we can also see when we consider the values of R: Whereas the second data set achieves an R of 95.8% once 3% of the PCs are used, the first data set achieves R=99.9% for the same fraction of PCs used

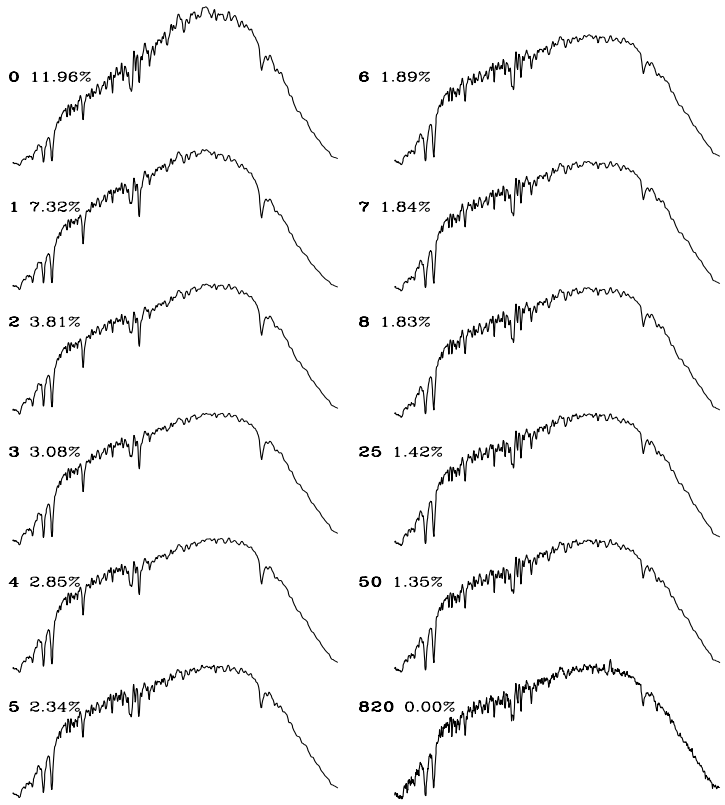

Figure 8. Reconstruction of a G3 V star using successive numbers of principal components (indicated in bold). The empirical reconstruction error, E, is shown as a percentage. Zero PCs corresponds to the mean spectrum, and 820 the original spectrum

Fig. 8 shows the appearance of a G star reconstructed with increasing numbers of PCs. A rapid improvement in the overall appearance is seen after adding the first one or two PCs, after which only more subtle changes to the strengths of certain lines are be made. As the reconstruction quality, R, is a statistical measure over the whole data set, it is not that useful in a specific case. Thus we define the reconstruction error as

$$E = \frac{100\%}{S} \sum_{j=1}^{j=m} |y_j - x_j| \tag{8}$$

where x_j and y_j are the fluxes in the original and reconstructed spectra respectively, and S is the total area under the spectra (which serves to normalise the error). A histogram

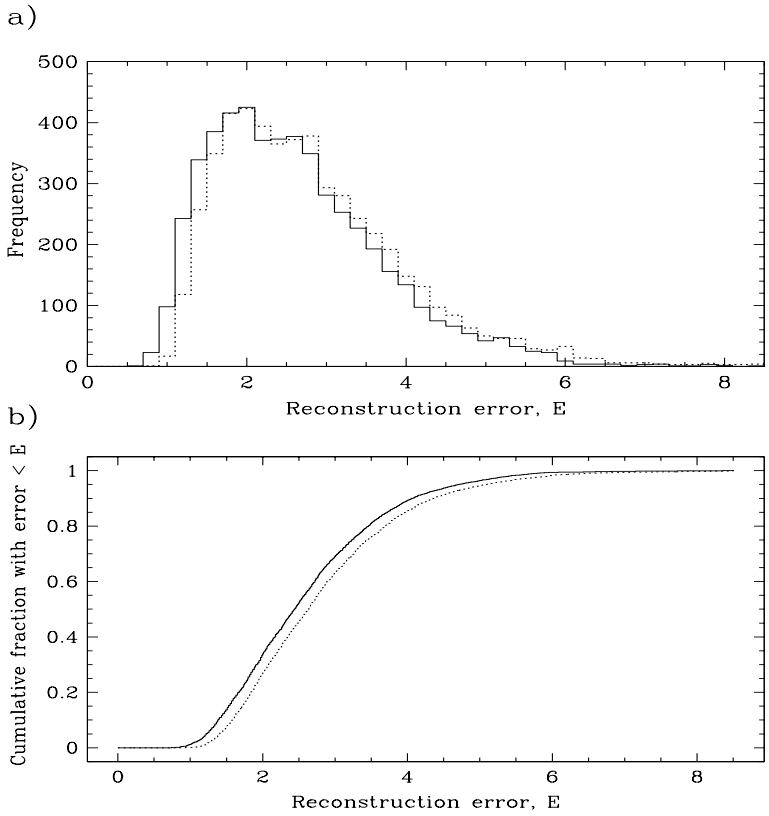

Figure 9. Frequency distribution of the empirical reconstruction error, E. The solid line shows the errors for a 50 component reconstruction and the dashes line for a 25 component reconstruction. (a) histogram of the reconstruction errors. (b) the cumulative distribution of (a)

of these errors for all 5144 spectra is shown in Fig. 9 using both 25 or 50 components in the reconstruction. Such an error is useful as a measure of how the overall appearance of the spectrum has been reconstructed, but cannot be taken as a measure of how well the star could be classified. Classification may rely on the strength of weak lines (e.g. certain iron lines in the case of metallicity determination), yet these will have little influence on E. Thus the quality of data compression and reconstruction may have very little to do with classification. A simple example of this is shown in Fig. 10, where an F star with a strange emission feature is reconstructed using the first 25 PCs. The feature is completely absent in the reconstruction. This is because none of the 5144 stars used to create the PCs had such a feature, so it is not explicitly represented in the PCs. Of course, once enough PCs are used – all if necessary – any feature can be reconstructed, but if the feature was not common to the original data it will require a lot of PCs. The

more common a feature is to the stars, the more strongly it will be represented in the significant PCs. Thus a PC reconstruction will lead to the loss of rare features (see Bailer-Jones 1996).

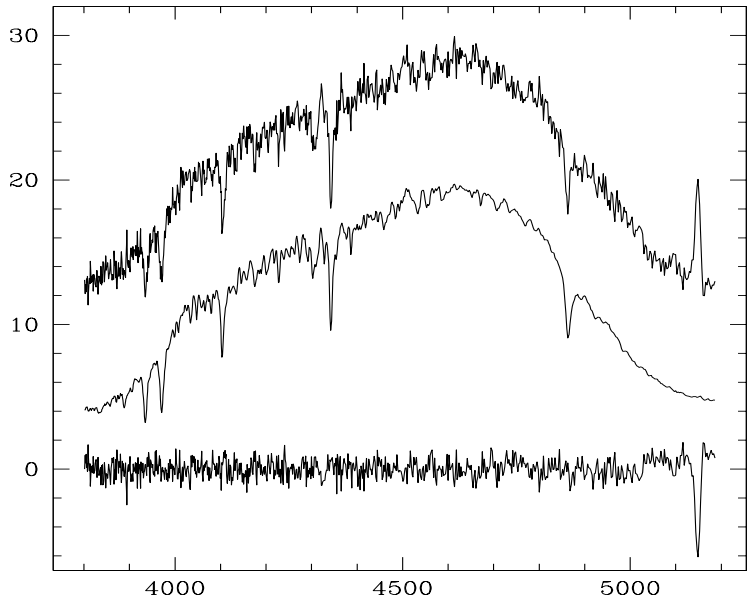

Figure 10. Reconstruction of an F5 V star with a bogus emission spike. The top spectrum is the original, the middle the reconstruction using the first 25 PCs, and the bottom the residual spectrum (reconstructed minus original). The horizontal scale is in Å

In certain contexts, this filtering of rare features can be an advantage. The emission feature in Fig. 10 is actually a piece of dust present on the photographic plate during scanning, so having it filtered in this case is a useful feature of PCA. In the same way, PCA can be used to detected spectra which deviate considerably from the set used to create the PCs. In this case study, the PCs represent stellar spectra. Thus, if we attempt to reconstruct a quasar, or galaxy, or background sky with these PCs, then the reconstruction error will be very high, typically much higher than the 6% which contains almost all of the stellar spectra in Fig. 9 (Bailer-Jones et al. 1998a). Thus by projecting all observed spectra onto the PCs and assigning a threshold, we can automate the rejection of contaminating spectra. (The example in Fig. 10 has E=5.3%, so is a borderline case for rejection.)

4. Conclusions

We demonstrated, by way of our two case studies, that the PCA can be used as an effective tool to compress the data. In both the test cases, we were able to compress the

spectra by a factor of 30 while retaining over 96and 99showed how PCA can be used to filter out bogus spectral features or identify unusual spectra. We also find that PCA is unable to reconstruct weak or rare features. We also tested the performance of ANNs using projections from the PCA as input and find that similar accuracy in classification can be achieved with first 10 or 20 PCs instead of using the complete data set.

References

Bailer-Jones, C. A. L., 1996, PhD thesis, Univ. Cambridge
Bailer-Jones, C. A. L., Irwin, M., von Hippel, T., 1998a, MNRAS, 298, 361.
Bailer-Jones, C. A. L., Irwin, M., von Hippel, T., 1998b, MNRAS, 298, 1061.
Chatfield C., Collins A. J., 1980, Introduction to Multivariate Analysis. Chapman & Hall, London.
Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L.,
Calzetti D., 1995, AJ, 110, 1071.
Folkes S. R., Lahav O., Maddox, S. J., 1996, MNRAS, 283, 651.
Francis P. J., Hewett P. C., Foltz C. B., Chaffee F. H., 1992, ApJ, 398, 476.
Gulati R. K., Gupta R., Gothoskar P., Khobragade S., 1994a, ApJ, 426, 340 (G94a).
Gulati R. K., Gupta R., Gothoskar P., Khobragade S., 1994b, Vistas in Astronomy, 38, 293 (G94b).
Gulati R. K., Gupta R., Gothoskar P., Khobragade S., 1996, Bull. Astron. Soc. India, 24, 21 (G96).
Jacoby G. H., Hunter D. A., Christian C. A., 1984, ApJS, 56, 257 (JHC).
Lahav O., Naim A., Sodré L., Jr., Storrie-Lombardi M. C., 1996, MNRAS, 283, 207.
Murtagh F., Heck A., 1987, Multivariate Data Analysis, Reidel, Dordrecht.
Odewahn S. C., Stockwell E. B., Pennington R. L., Humphreys R. M., Zumach W. A. 1992, AJ, 103, 318.
Silva D. R., Cornell M. E., ApJS, 81, 865 (SC).
Singh, H. P., Gulati, R. K., Gupta, R., 1998, MNRAS, 295, 312.
Sodré L. Jr., Cuevas H., 1994, Vistas in Astronomy, 38, 331.
Storrie-Lombardi M. C., Irwin M J., von Hippel T., Storrie-Lombardi L. J., 1994, Vistas in Astronomy, 38, 331.
von Hippel T., Storrie-Lombardi L. J., Storrie-Lombardi M. C., Irwin M.,1994, MNRAS, 269, 97.
Whitney, C. A., 1983, A&A, 51, 443.