THE ASTROPHYSICAL JOURNAL, 492:98–109, 1998 January 1

AUTOMATIC REDSHIFT DETERMINATION BY USE OF PRINCIPAL COMPONENT ANALYSIS. I. FUNDAMENTALS

KARL GLAZEBROOK, ALISON R. OFFER, AND KATHRYN DEELEY
Anglo-Australian Observatory
Received 1997 June 30; accepted 1997 August 13

ABSTRACT

With the advent of very large redshift surveys of tens to hundreds of thousands of galaxies, reliable techniques for automatically determining galaxy redshifts are becoming increasingly important. The most common technique currently in common use is the cross-correlation of a galactic spectrum with a set of templates. This series of papers presents a new method based on principal component analysis. The method generalizes the cross-correlation approach by replacing the individual templates with a simultaneous linear combination of orthogonal templates. This effectively eliminates the mismatch between templates and data and provides for the possibility of better error estimates. In this paper, the first of a series, the basic mathematics is presented along with a simple demonstration of the application.

Subject headings: galaxies: distances and redshifts — methods: observational — surveys

1. INTRODUCTION

The development of fiber-based spectrographs capable of observing hundreds of objects simultaneously has led to the advent of many large redshift surveys intended to further our understanding of the large-scale structure, clustering, and evolution of galaxies (for a review, see Strauss 1996). Examples include the just-completed Las Campañas Redshift Survey of 26,000 galaxies (Shectman et al. 1996) and the Two Degree Field (2dF) Redshift Survey (Taylor et al. 1998; Maddox et al. 1998), which started in 1997 and will measure the redshifts of over 250,000 galaxies over the next several years.

Because of the sheer size of these surveys, it is becoming very important to develop methods of reliably and quantifiably measuring the redshifts of the galaxy spectra without manual intervention. For example, in the 2dF survey, a method with a 95% success rate would still leave 12,500 spectra to be inspected manually, a very large task. Ideally any automatic redshift calculation should also give an accurate error estimate and confidence rating for each redshift to indicate which 12,500 galaxies of the 250,000 need further, possibly manual, attention.

At the current time the most successful and widespread method of automatic redshift measurement is cross-correlation analysis (Tonry & Davis 1979). In this method, the galaxy spectrum is cross-correlated with a series of template spectra corresponding to a sequence of standard galaxy or stellar types. The size of the largest peak in the cross-correlation function is an indication of the quality of the match between the galaxy and the template spectrum. The position and width of the peak give the redshift and an "error" on the redshift. If the galactic and template spectra agreed exactly, then a sharp correlation peak would be found, but in practice it is unlikely that the galactic spectrum will exactly match any of the template spectra. Depending on the size of the mismatch, the redshift may or may not be correct—the "error" is merely a measure of the accuracy of the location of the peak and not an indication of its true worth. Tonry & Davis (1979) presented a formulation for the error on peak location, which was improved upon by Heavens (1993).

A series of templates consisting of different types of galactic spectra, individually tested, is not necessarily the optimal template set to use. It would be preferable to generalize the concept of cross-correlation to use a simultaneous linear combination of templates with expansion coefficients that depend on the redshift. With a suitable choice of template spectra, the mismatch between the data and a linear combination of a small number of template spectra could be reduced to an arbitrarily small amount. Any residual would be due only to the random component of the observational noise.

In this paper, the first of a series of papers, a method is presented for achieving this. The method, which we will call "PCAZ," is based upon the use of principal component analysis to make the general linear problem amenable to efficient computation. The fundamental mathematics is presented in § 2, and a simple demonstration based upon some sample 2dF galaxy spectra is shown in § 3. Subsequent papers will present in more detail the robust error-analysis methods and software for implementing the PCAZ algorithm.

2. THE MATHEMATICS BEHIND PCAZ

2.1. *Standard Cross-Correlation Revisited*

Consider a galaxy spectrum G_λ (with normally distributed errors, variance σ_λ^2) requiring a redshift z and a single template spectrum T_λ. If both the galactic spectrum and the template spectrum are binned on the same logarithmic wavelength grid, the likelihood that the galaxy and the template are the same, except for redshift and normalization, can be written

$$-\log \mathcal{L} \propto \chi^2 = \sum_\lambda \frac{1}{\sigma_\lambda^2} [G_\lambda - a(z) T_{\lambda+\Delta}]^2 , \qquad (1)$$

where the sum is over discrete wavelength bins ($\lambda = 1, 2, 3, \ldots$), and Δ is the linear shift along the logarithmic grid owing to the redshift, $\Delta \propto \log(1+z)$. The redshift-dependent coefficient of the template is $a(z)$. At any particular redshift z, we can find the value of $a(z)$ that maximizes the likelihood (i.e., gives the best match between the galaxy and the template) by setting $\partial\chi^2/\partial a = 0$. This gives

$$a(z) = \frac{\sum_\lambda (1/\sigma_\lambda^2) G_\lambda T_{\lambda+\Delta}}{\sum_\lambda (1/\sigma_\lambda^2) T_{\lambda+\Delta}^2} . \qquad (2)$$

It can be seen that $a(z)$ in equation (2) is simply proportional to the cross-correlation function of the galaxy spectrum with the template spectrum for the case where the variance is ignored. Substituting this value of $a(z)$ into equation (1) and simplifying gives

$$\chi^2 = \sum_\lambda \frac{1}{\sigma_\lambda^2} G_\lambda^2 - a(z)^2 \left(\sum_\lambda \frac{1}{\sigma_\lambda^2} T_{\lambda+\Delta}^2 \right)^{-1} . \quad (3)$$

The minimum of χ^2 as a function of redshift occurs when $a(z)$ is maximized, so that finding the peak of the cross-correlation function (or CCF) is exactly equivalent to finding the maximum likelihood redshift at which the single template best matches the data (again, including the variance introduces complications; see § 2.4).

This maximum likelihood basis for cross-correlation is fundamental to the linear generalization but has not been remarked upon in the astronomical literature. The approach that has historically been used to assign a confidence or quality value to the redshift has been based upon the height of the CCF peak above the CCF "noise" (see, e.g., Heavens 1993). However much of this "noise" is caused by a systematic mismatch between the template and the data and is not observational noise, and thus the assumption that peaks are uncorrelated is invalid. With the formulation given above and realistic errors, it would be possible to assign a true likelihood value and hence confidence intervals to a peak if the template and the galaxy were identical and differed only because of the observational noise and the redshift.

2.2. *Linear Generalization*

The standard cross-correlation method of Tonry & Davis (1979) tests the candidate galaxy spectrum against a range of template spectra individually. The linear generalization presented here essentially assumes that a galaxy spectrum can be expanded as a linear sum of template spectra. This in principle allows the systematic mismatch between galaxy and template to be arbitrarily reduced and hence a realistic likelihood to be assigned to an output redshift.

Initially, for simplicity, we will consider how one solves for the values of the coefficients at zero redshift. We assume that a galaxy spectrum is represented by an n-dimensional vector $\boldsymbol{G}$, where n is the number of wavelength bins. The m template spectra are represented by the rows of an $m \times n$ matrix $\boldsymbol{T}$. The galaxy spectrum is then fitted by a linear combination of templates with coefficients a_j,

$$G_\lambda \simeq \sum_j a_j T_{j\lambda} . \quad (4)$$

The coefficients, a_j, may be found by following the same maximum likelihood recipe used above and minimizing χ^2, where now

$$\chi^2 = \sum_\lambda \frac{w_\lambda^2}{\sigma_\lambda^2} \left(G_\lambda - \sum_j a_j T_{j\lambda} \right)^2 . \quad (5)$$

We have now introduced w_λ as representing a general wavelength-dependent weighting function (which might be used, for example, to emphasize particular spectral features). Setting $\partial\chi^2/\partial a_i = 0$ leads to the matrix equation

$$\boldsymbol{Ca} = \boldsymbol{TG}' , \quad (6)$$

where the elements of the vector $\boldsymbol{G}'$ are $G'_\lambda = w_\lambda^2 G_\lambda/\sigma_\lambda^2$ and the elements of the $m \times m$ correlation matrix, $\boldsymbol{C}$, are given by

$$C_{ij} = \frac{w_\lambda^2}{\sigma_\lambda^2} T_{i\lambda} T_{j\lambda} . \quad (7)$$

Direct inversion of the $\boldsymbol{C}$ matrix to obtain the a_j coefficients is clearly impractical. Not only would it be numerically intensive to do this at many trial redshifts, but the presence of very small eigenvalues (see § 2.2 below) would lead to large numerical instabilities. However, if the template spectra (the rows of the matrix $\boldsymbol{T}$) are replaced by a basis set of orthogonal vectors, the transformed correlation matrix will be diagonal, and the problem simplifies. Principal component analysis (PCA) is the tool used to select the orthogonal vectors.

2.3. *Principal Component Analysis*

Principal component analysis is a technique frequently used for data compression and classification (Kendall & Stuart 1966; Murtagh & Heck 1987). In particular, direct PCA of spectral data similar to that used here has been used for classification of galactic spectra (Mittaz, Penston, & Snijders 1990; Connolly et al. 1995; Folkes, Lahav, & Maddox 1996; Sodré & Cuevas 1997) and for classification of QSO spectra (Francis et al. 1992).

In essence, PCA finds the "best" representation of a set of data by a set of orthogonal vectors, or principal components, which can be combined linearly to reconstitute the data. The components are ordered in terms of significance in a least-squares sense, and data compression is achieved by retaining only the most significant principal components.

PCA can be formulated in two different but equivalent ways, both of which have been used for spectral classification. Consider a set of m template spectra sampled at n discrete wavelengths. The elements of the matrix $\boldsymbol{T}$ can be pictured as a series of row vectors, each of which is a point representing a spectrum in n-dimensional wavelength space. Alternatively, the data can be thought of as column vectors with each point in m-dimensional template space being the set of fluxes in an individual wavelength bin. A PCA in the template space diagonalizes the elements of the $m \times m$ correlation matrix,

$$C_{ij} = \sum_\lambda T_{i\lambda} T_{j\lambda} , \quad (8)$$

where, for now, we ignore weights and variance factors for clarity in the discussion. A PCA in wavelength space diagonalizes the elements of the $n \times n$ correlation matrix,

$$D_{\lambda_1\lambda_2} = \sum_i T_{i\lambda_1} T_{i\lambda_2} . \quad (9)$$

The two approaches are equivalent, and in principle they will lead to the same eigenvalues and principal components that are related by a simple transformation (Murtagh & Heck 1987). In many ways the wavelength space is more intuitive for spectral classification, and Mittaz et al. (1990), Francis et al. (1992), Folkes et al. (1996), and Sodré & Cuevas (1997) have used a PCA in wavelength space for this. Connolly et al. (1995), who used only a small number of spectra, chose to work with the reduced dimensionality of template space. In order to show the link between the cross-correlation method and PCA clearly and because we have fewer template spectra than wavelength bins, we have

chosen to follow Connolly et al. and perform the diagonalization in template space.

In practice this means taking the set of templates and constructing from them a set of orthogonal "eigentemplates." The matrix $\boldsymbol{C}$ is diagonalized to yield a set of eigenvalues,

$$\boldsymbol{C} = \boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^T \text{ where } \boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_1 & 0 & 0 & \dots \\ 0 & \Lambda_2 & 0 & \dots \\ 0 & 0 & \Lambda_3 & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}, \quad (10)$$

and an associated matrix of m-dimensional eigenvectors, $\boldsymbol{R}$, which are the principal components in template space. The diagonalization is accomplished by standard numerical techniques such as singular-value decomposition (Mittaz et al. 1990). The matrix $\boldsymbol{R}$ defines a transformation between the template spectra and a set of n-dimensional orthogonal eigentemplates, the principal components in wavelength space $E_{i\lambda}$,

$$E_{i\lambda} = \sum_k R_{ki}\, T_{k\lambda} \, . \quad (11)$$

This is essentially the Karhumen-Loève transform (Murtagh & Heck 1987). The resulting eigentemplates satisfy the orthogonality property

$$\sum_\lambda E_{i\lambda} E_{j\lambda} = \Lambda_i \delta_{ij} \, , \quad (12)$$

where δ_{ij} is the Kronecker δ. The eigenvalues Λ_i represent the contribution of each eigentemplate to the set of templates in a least-squares sense. If the principal components are arranged in order of decreasing eigenvalue, it can be shown (Kendall & Stuart 1966) that the first principal component in either space is the line along which the cloud of points is the most elongated (has the greatest variance). Equivalently, the first principal component is the line for which the sum of the squared perpendicular distances of the points from the line is a minimum. Similarly, if the points are projected onto a hyperplane orthogonal to the first principal component, the second principal component is the line in that hyperplane along which the projected distribution is most elongated. Representing the data in terms of just the first principal component would be equivalent to approximating the cloud of points by a line and characterizing each point in terms of its projected distance along the line. Representing the data in terms of the first two principal components is equivalent to projecting the cloud of points onto a plane.

The spectra within the template set can be represented to any given accuracy by a linear combination of eigentemplates,

$$G_\lambda \simeq \sum_{j=1}^{p} b_j E_{j\lambda} \, , \quad (13)$$

where p is the number of eigentemplates retained. Since the eigentemplates are orthogonal, the corresponding expansion coefficients are given by

$$b_j = \sum_\lambda \frac{G_\lambda E_{j\lambda}}{\Lambda_j} \, , \quad (14)$$

where Λ_j is the jth eigenvalue derived above.

In practice, only a subset of the principal components represent real correlations and anticorrelations between the spectra within the template set. The remaining principal components may contain a large fraction of uncorrelated noise, in which case they can be discarded. Folkes et al. (1996) show that the number of significant principal components, p, depends on the quality of the template data set. Reconstruction of the template spectra from the first p principal components effectively filters out much of the noise. In the case where the input template set consists of a few very high S/N spectra, it may be desirable to retain all the eigentemplates—in this case the PCA analysis can be viewed as a shortcut for speeding up the solving of equation (6) for a large number of redshifts.

To apply PCA to redshift determination, it is necessary to assume that the template set is sufficiently general that any galactic spectrum not included in the original template set can also be represented to the required accuracy by a summation over the first p principal components. Essentially, we are assuming that the correlations within the template set reflect a global correlation across all galaxies in the survey. Allowance for abnormal objects such as stars, active galaxies, and quasars can be made by including example spectra of these in the template set or by discriminating against bad matches (see § 3.2).

2.4. *Relation to Cross-Correlation and Redshift Determination*

The discussion of PCA above is general and up to this point follows the spirit of the spectral classification of Mittaz et al. (1990), Francis et al. (1992), Connolly et al. (1995), and Folkes et al. (1996). The extra step is to include the redshift z as an additional variable. Weighting can be retained but must be tied to the rest frame of the templates, and the variance must be assumed to be independent of wavelength (but see below). The coefficients of the eigentemplates then become

$$b_j(z) = \sum_\lambda \frac{w^2_{\lambda+\Delta} G_\lambda E_{j(\lambda+\Delta)}}{\Lambda_j} \, , \quad (15)$$

where each $b_j(z)$ is the cross-correlation function of the galaxy spectrum with the jth eigentemplate weighted by the corresponding eigenvalue. If g_k is the discrete Fourier transform of the galaxy spectra, G_λ, and e_{jk} is the discrete fourier transform of $w^2_\lambda E_{j\lambda}$, then the coefficients $b_j(z)$ are given by the inverse Fourier transform of the product of g_k and e_{jk},

$$b_j(z) = \frac{1}{N\Lambda_j} \sum_{k=0}^{N-1} g_k e_{jk} \exp\left(\frac{2\pi i k \Delta}{N}\right). \quad (16)$$

The orthogonality gives the simple relation for the joint likelihood,

$$-\log \mathscr{L} \propto \chi^2 = \frac{1}{\sigma^2}\left[\sum_\lambda w^2_{\lambda+\Delta} G^2_\lambda - \sum_j \Lambda_j b_j^2(z)\right], \quad (17)$$

where the variance is σ^2. The minimum of equation (17) gives the maximum likelihood redshift, z, through

$$z_{\Delta_{\min}} = 10^{\Delta_{\min}\delta(\log\lambda)} - 1 \, , \quad (18)$$

where $\Delta_{\min}$ is the shift that gives the minimum value of χ^2. Note that the single cross-correlation function in equation (3) has been replaced in equation (17) by a weighted sum of the squares of the individual cross-correlation functions. This is a natural result given that the eigentemplates are orthogonal.

To preserve the orthogonality of templates with redshift, it is not possible to weight by a wavelength-dependent variance because the errors will be tied to the observed frame. For a strongly wavelength-dependent variance, the method still gives the optimal fit in a least-squares sense by minimizing the function

$$f = \sum_{\lambda} \left[G_{\lambda} - \sum_{i} b_i(z) E_{i,\lambda+\Delta} \right]^2 . \quad (19)$$

One can then reintroduce σ_{λ} for a final pass and calculate a true likelihood for the final redshift, though this may not be the absolute maximum likelihood because it is not used in the calculation of the values of b_j. In practice, this will mean that a few more objects may fail to have their redshifts determined within specified likelihood bounds.

It should be noted that the logarithmic wavelength scale used for redshift determination gives the correct weighting of spectral features. Since $\Delta\lambda \ll \lambda$, then $\log(\lambda + \Delta\lambda) - \log(\lambda) \approx \Delta\lambda \log(e)/\lambda$, so the fractional wavelength range per bin is constant across the spectrum. The important features in the eigenfunctions are the spectral lines that should be equally weighted. For a classical grating, $\Delta\lambda \propto \lambda$ for unresolved lines, and for Doppler broadening, the same holds true. Thus logarithmic binning gives correct equal weighting of features. Of course most real systems give close to a linear wavelength scale, so the spectra must be resampled to logarithmic bins, which will introduce covariance between neighboring pixels. However this will be very small as the wavelength scale only changes very slowly across the spectrum.

The PCAZ method has numerous advantages over previous methods in the literature:

1. Because it is just a set of cross-correlations, the standard fast Fourier transform (FFT) method can be used to compute the values of $b_j(z)$ efficiently. The simultaneous combination of m eigentemplates takes the same computer time as doing m templates separately.

2. Existing cross-correlation codes can be used with little modification. They need only to be provided with only orthogonalized eigentemplates instead of the normal templates as inputs and to have some provision made for combining the cross-correlation functions in quadrature afterward.

3. Emission-line galaxies are easily handled by PCAZ. The standard cross-correlation method gives relatively poor results for these because emission-line ratios vary much more than absorption lines and hence can no longer be accounted for by a small number of standard galaxy spectra. However, with the extra freedom given by a linear combination of eigentemplates, variable line ratios can be accommodated. This freedom means the method is robust against other wavelength-dependent variations such as very approximate or absent flux calibration of the input spectra.

4. High S/N eigenspectra can be created from a large set of noisy data as well as from a small set of high S/N spectra because each eigenspectrum represents an average of that mode over the data. This would be especially suitable for a deep redshift survey where many of the weak ultraviolet absorption features would be missing from local templates. A few hundred high-redshift galaxies could have their redshifts measured manually. Eigenspectra constructed from these could be used to measure the rest automatically.

5. The ability to calculate a likelihood means a true confidence could be assigned to a redshift; thus future science analyses of survey statistics such as the power spectrum of galaxy clustering $P(k)$ could include a realistic probability distribution of redshift errors rather than neglecting them. This is especially important for the next generation of very large surveys.

6. The maximum likelihood reconstruction from the coefficients $b_j E_{j\lambda}$ is a noise-filtered version of the data, which is useful for other analyses.

7. The coefficients b_j have independent errors and could be used as the basis for classification schemes for faint spectra, either by themselves or as input into other systems such as artificial neural net algorithms (see, e.g., Folkes et al. 1996).

8. The provision of weights allows templates to be defined only in regions of interest, for example, around strong lines. This would be particularly suitable for very faint, low-S/N data where one might wish to search for weak emission lines appearing above the noise. With weights, the rest of the noisy, possibly undetected, continuum can be excluded from the χ^2.

2.5. *Practicalities*

There are a number of important practicalities involved in using the PCA formalism to determine redshifts. The first is the issue of mean subtraction. It is usual in PCA to subtract the mean of the distribution from each point; in the case of spectral classification, the mean spectrum is subtracted from each of the template spectra prior to orthogonalization. This is equivalent to moving the origin of the PCA coordinate system to the center of the distribution of points. However, strictly, a redshifted mean spectrum should also be subtracted from the candidate spectrum whose redshift is not yet known. Because of this, the mean spectrum was not subtracted prior to orthogonalization.

A second important point is continuum subtraction. Spectral classification schemes have avoided continuum subtraction in order to retain as much spectral information as possible (Mittaz et al. 1990; Connolly et al. 1995; Folkes et al. 1996). However, continuum subtraction is more important for redshift determination. Continuum subtraction reduces the smoothly varying background to zero and essentially has the same effect as filtering out the long-period Fourier components of the spectra. Without continuum subtraction, the cross-correlation functions show a broad peak representing the cross-correlation of the two apodized continua, with a small spectral cross-correlation peak superimposed.

A final practicality is the normalization of the template spectra. Francis et al. (1992), Folkes et al. (1996), and Sodré & Cuevas (1997) normalize to unit flux,

$$\sum_{\lambda} T_{\lambda} = 1 . \quad (20)$$

The alternative is to normalize to unit scalar product (Connolly et al. 1995),

$$\sum_{\lambda} T_{\lambda}^2 = 1 . \quad (21)$$

With continuum subtraction, the resulting spectra oscillate about zero so normalization to unit scalar product was used.

3. EXAMPLES

In this section, the method is illustrated using a set of sample sky-subtracted spectra. The method was developed

TABLE 1

GALACTIC SPECTRA USED TO CONSTRUCT THE FIRST SET OF EIGENSPECTRA[a]

Galaxy	Morphology
NGC 3379.....	E0
NGC 4889.....	E4
NGC 5248.....	Sbc
NGC 2276.....	Sc
NGC 4485.....	Sm/Im

[a] From Kennicutt 1992.

using test spectra from a variety of sources; we choose to illustrate its effectiveness here using some early data recently taken from the 2dF galaxy survey for which the algorithm is being developed. The 2dF survey is more comprehensively described elsewhere; see Taylor (1994) and Taylor et al. (1998) for a description of the 2dF instrument and Maddox et al. (1998) for an introduction to the galaxy survey. The data described here consist of two test fields, SGP 463 and NGP 359, taken during 2dF commissioning

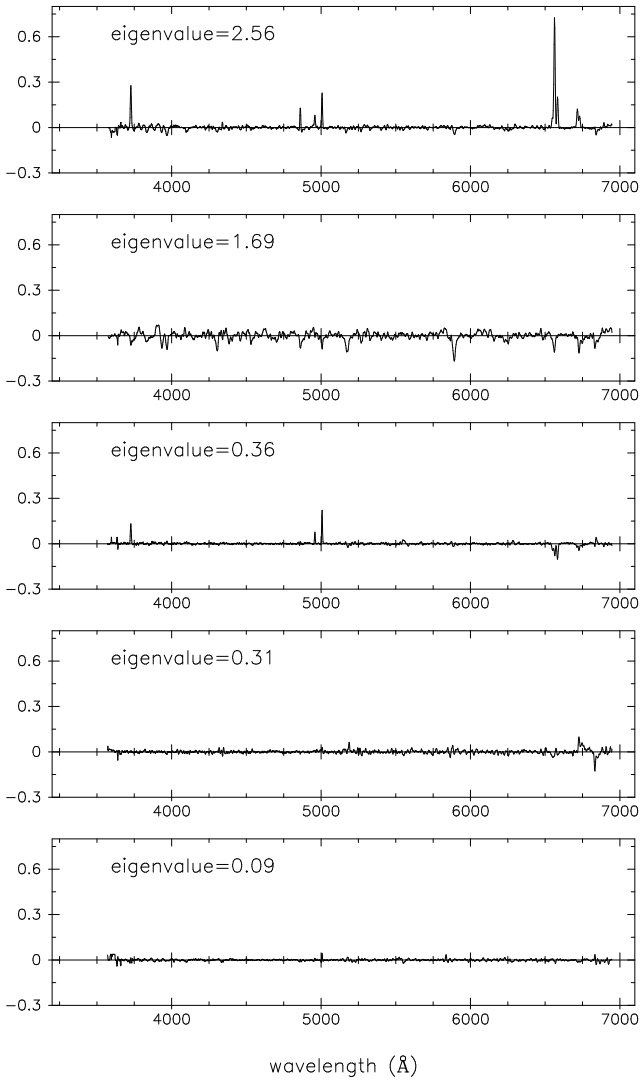

FIG. 1.—Eigenfunctions obtained using the Kennicutt (1992) spectra listed in Table 1. The vertical scale is the flux per unit wavelength in normalized units.

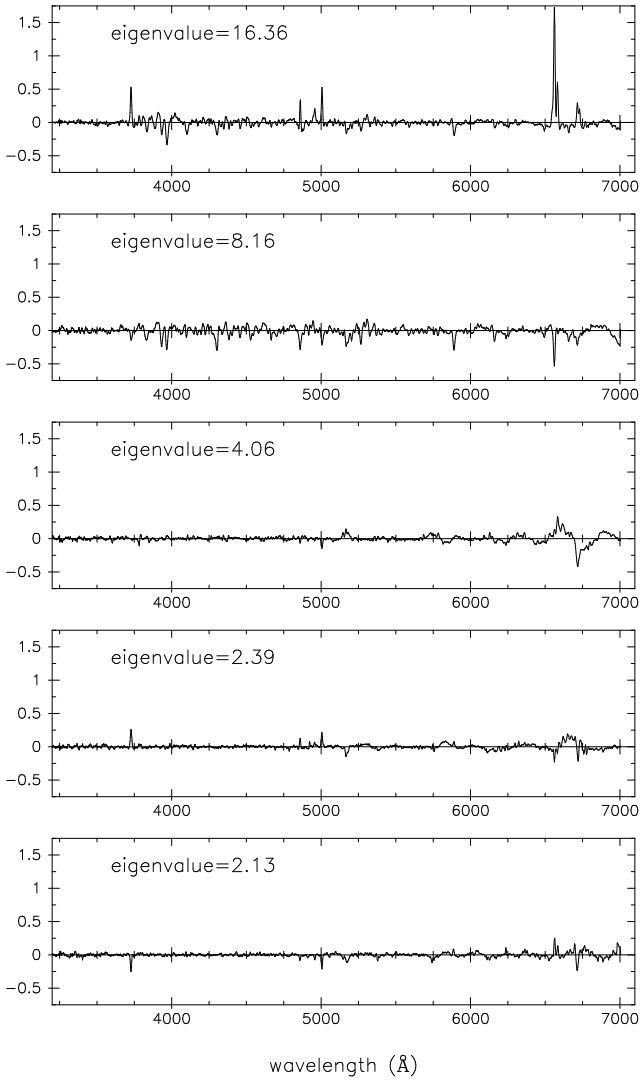

FIG. 2.—First five eigenfunctions obtained using a sample of 91 2dF spectra.

for the survey in 1997 January–April. The galaxies are selected from the Automatic Plate Measuring Facility (APM) survey (Maddox et al. 1990) with $b_j < 19.7$.

The 2dF spectra spanned a wavelength range of 3810–8227 Å with a 2 pixel resolution of around 8.4 Å (FWHM). Two fields were considered, one as template spectra and one as candidate spectra whose redshift was to be determined. Redshifts had been previously assigned to both data sets by visual inspection (M. Colless and K. Glazebrook 1997,

TABLE 2

REDSHIFTS AND EXPANSION COEFFICIENTS FOR SPECTRA (A) TO (E)[a]

Spectrum	PCAZ Redshift	Visual Inspection Redshift	$b_1(z)$	$b_2(z)$	$b_3(z)$
a	0.0674	0.0676	1.193	0.025	0.029
b	0.1411	0.1412	0.075	0.833	0.001
c	0.2379	0.2384	0.069	0.752	0.003
d	0.1809	0.1809	1.104	0.176	−0.068
e	0.0600	...	0.683	0.139	0.003

[a] From Fig. 3.

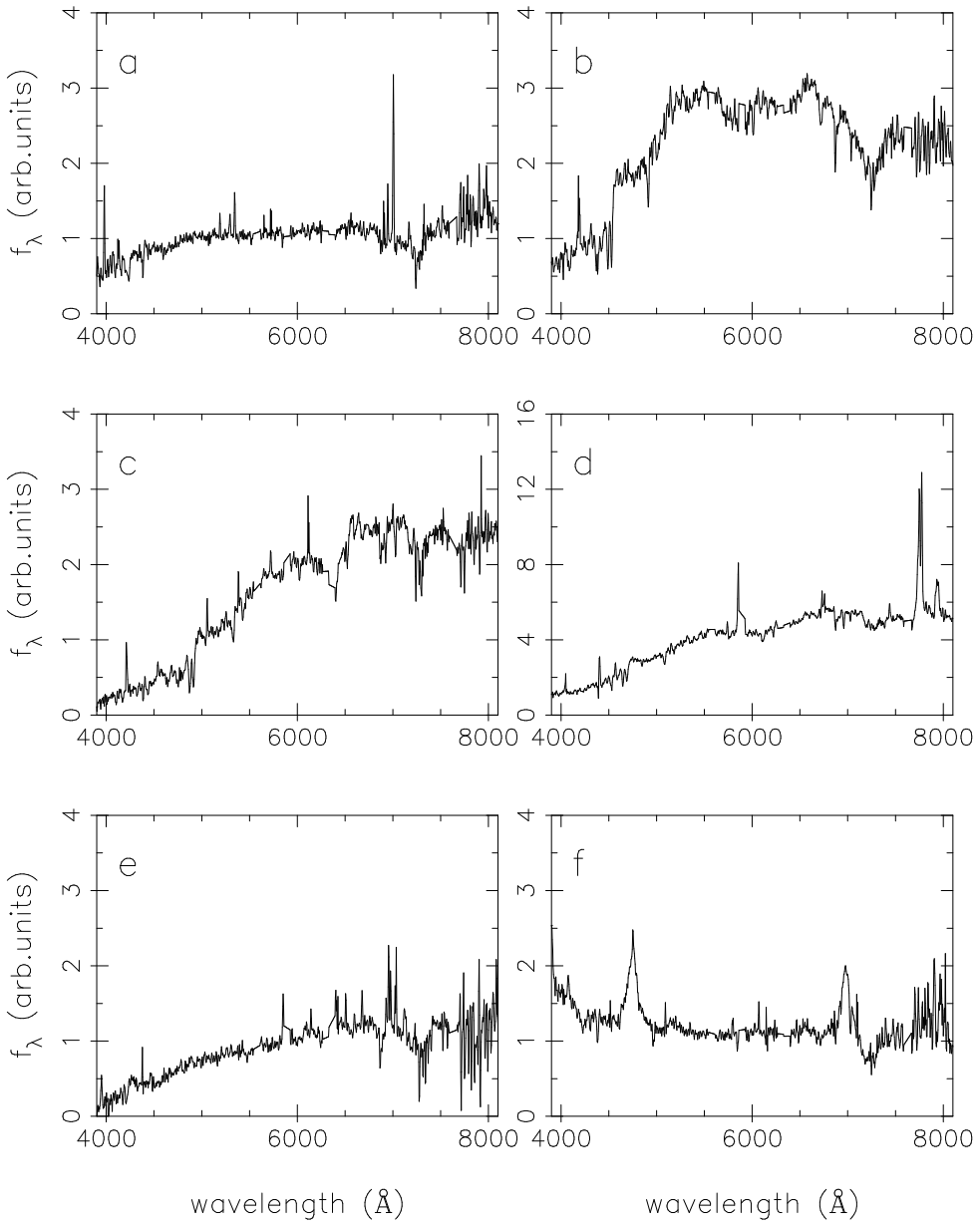

FIG. 3.—The six 2dF spectra discussed in the text. The spectra have been corrected for the sky residuals and divided by the instrument response function.

private communication). This gives a typical accuracy of $\Delta z \simeq 0.0005$, set by the spectral resolution. The template field contained a total of 91 galaxies for which redshifts had been assigned, and the candidate field contained 104 galaxies with known redshifts. The typical S/N of the continuum was 10–30 at 5500 Å, which should be typical for the survey spectra. These are quite high S/N, and we expect a variety of methods to work well; in the analysis below we add artificial noise to degrade the spectra to test the robustness of the method.

3.1. *Eigenspectra*

Two sets of eigenspectra were constructed. The first used five high-S/N template spectra taken from an atlas of integrated spectra of local galaxies (Kennicutt 1992). The five spectra chosen are listed in Table 1. They cover a wavelength range of 3600–7050 Å. The spectra were rebinned on a log wavelength grid with a grid spacing of $\delta \log_{10}(\lambda/\text{Å}) = 1.7 \times 10^{-4}$. The second set of spectra were derived from the 2dF data itself. The NGP 359 field was used. The 91 spectra with well-determined redshifts were corrected for redshift and used as the template set. A wavelength grid of 3100 to 7007 Å was used for these with a grid spacing of $\delta \log_{10}(\lambda/\text{Å}) = 1.8 \times 10^{-4}$.

The template spectra were continuum subtracted and normalized prior to orthogonalization. For simplicity, a constant variance and unit weights were assumed at all wavelengths. The 2dF spectra were fluxed using an approximate mean 2dF response curve derived from photometric standards. Continuum subtraction was done at each point by subtracting the local median calculated over a window 100 bins wide centered at that point. The continuum-subtracted spectra were normalized so that the sum of the squares of fluxes in the continuum-subtracted spectrum was unity. With this normalization and these unit weights, the first term on the right-hand side of equation (17) equals one, leading to a particularly simple expression for χ^2. The resulting normalized spectra were orthogonalized using a

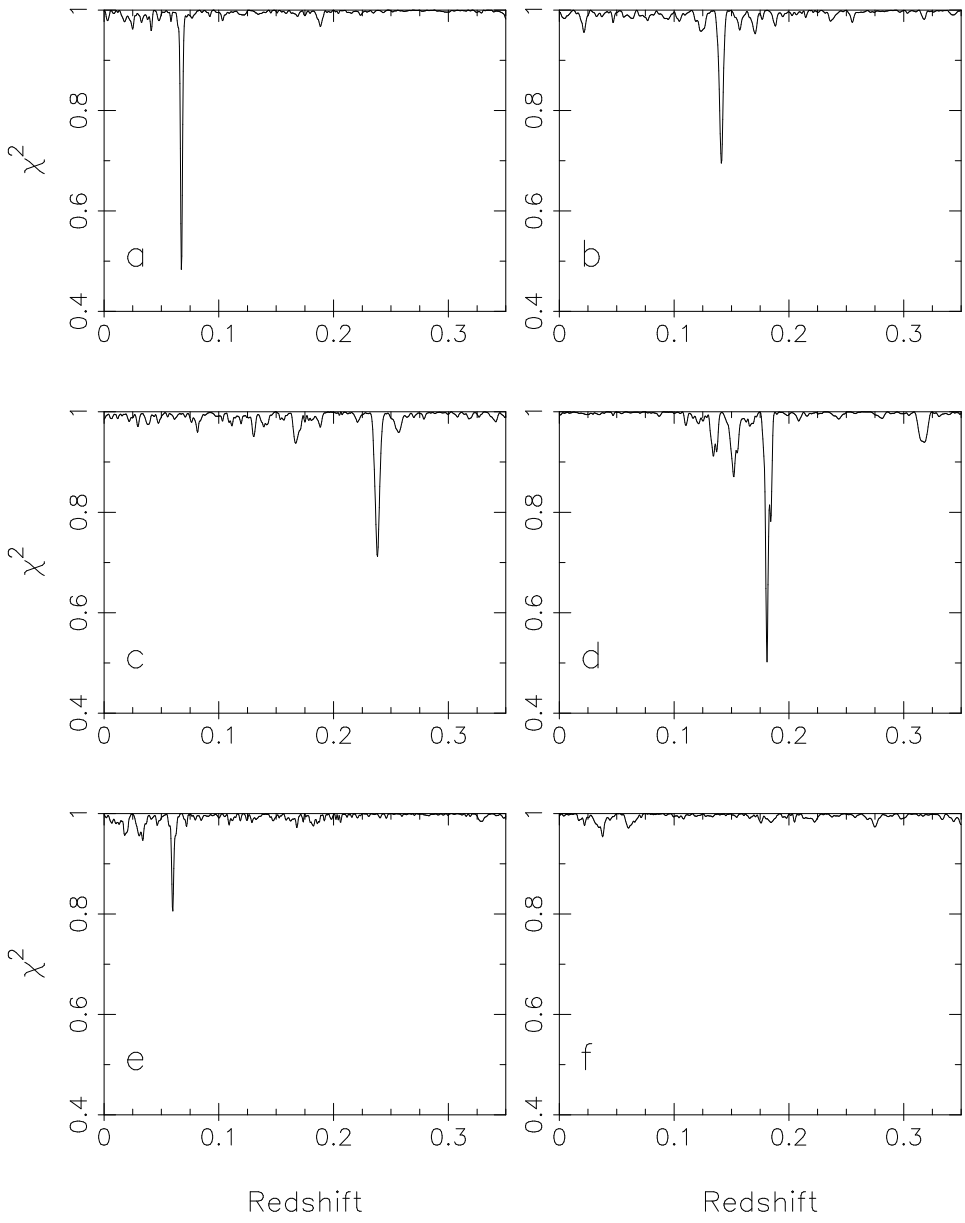

FIG. 4.—χ^2 functions (multiplied by a constant variance) corresponding to the spectra discussed in the text

standard singular-value decomposition routine. A selection of the resulting eigenspectra are shown in Figures 1 and 2.

The five eigenfunctions derived from the Kennicutt (1992) spectra are shown in Figure 1. Figure 2 shows the 2dF eigenspectra with the five highest eigenvalues. Orthogonalization of 91 spectra leads to 91 eigenspectra, but, as discussed in § 2, many of these represent noise. Five 2dF eigentemplates and three Kennicutt eigentemplates were retained for the redshift determination. It can be seen from Figure 1 that the first two eigenfunctions, derived from the different data sets, are very similar. These two account for $> 80\%$ of the variation in the input data. The higher order eigenfunctions come out differently for the different data sets, which is to be expected given the effect of noise on the exact location of the principal components.

3.2. *Redshift Determination*

Redshifts were calculated using both the Kennicutt eigenspectra shown in Figure 1 and the 2dF eigenspectra shown in Figure 2. As most spectral information is contained in the eigenspectra with the highest eigenvalues, only the first three Kennicutt eigenspectra and the first five 2dF eigenspectra were retained.

The 2dF spectra showed a number of residual sky features in the regions of strong atmospheric emission and absorption lines. Where these are the strongest features in the spectrum, there is a danger that the correlation between the strong peaks in the eigenspectra (particularly the strong Hα line) and the sky residuals will be greater than the correlation between the templates and the much weaker galaxy spectra. As a preprocessing step before orthogonalization, sky residuals were removed in 60 Å bands around 5577, 5892, 6300, 6363, and 7610 Å. The missing spectral bands were interpolated using a least-squares fit to the spectrum on either side, and the spectra were rebinned onto the same wavelength grid as the eigenspectra.

The rebinned spectra were continuum subtracted and normalized in the same way as the template spectra. The

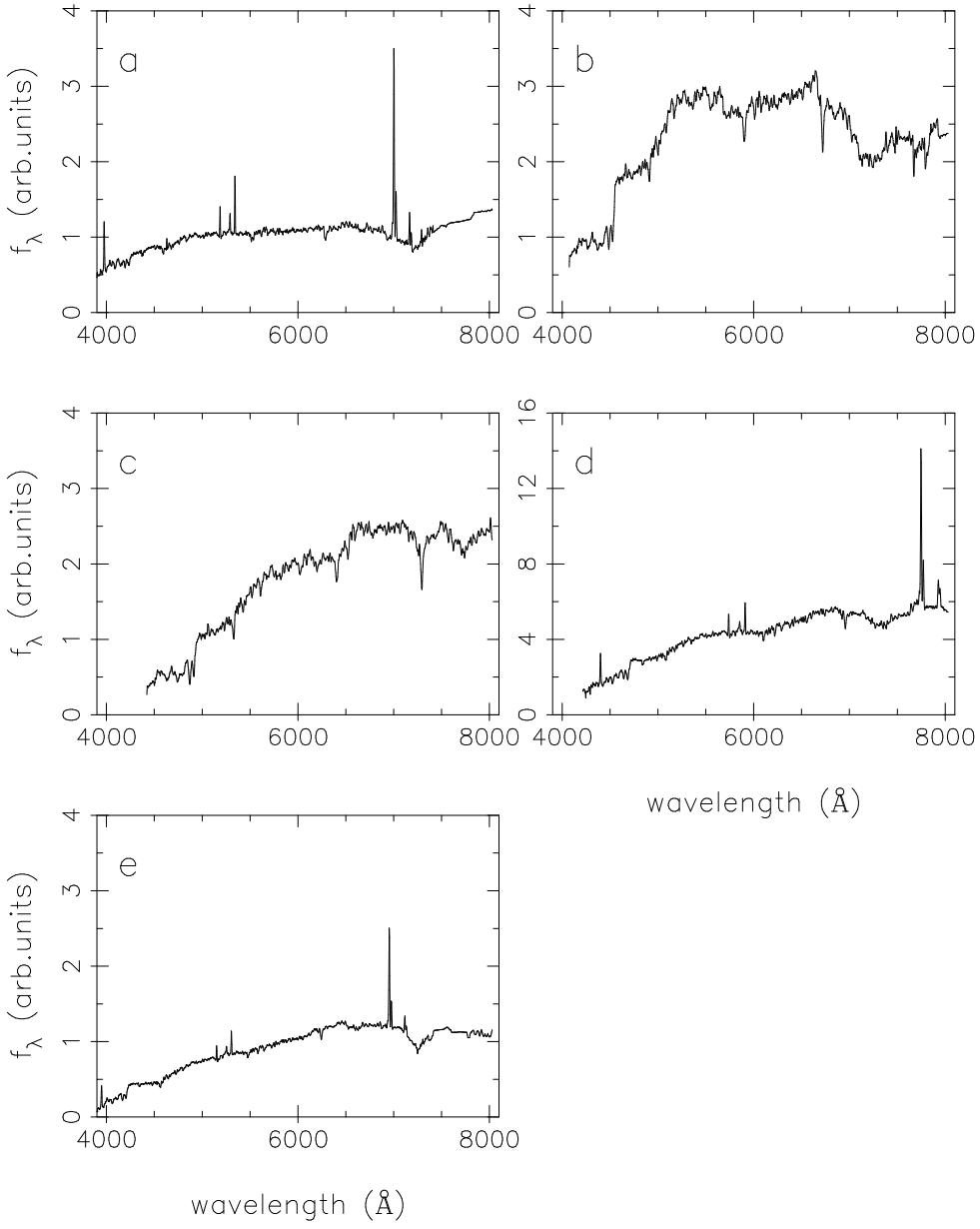

FIG. 5.—Noise-filtered reconstructions of the five spectra (*a*)–(*e*). The spectra are reconstructed from the first three Kennicutt (1992) eigenfunctions.

expansion coefficients, $b_j(z)$, can be quickly and efficiently found using fast Fourier transforms. The FFT algorithms are most efficient if the total length of the series, N, is equal to a power of 2. In addition, because the FFT treats the series as a periodic function of period N, N must be greater than the sum of the length of the galactic and template spectra to avoid errors in the cross-correlation calculation. To this end, both template and galactic spectra were zero-padded to the power of 2 greater than the sum of their lengths.

To illustrate the procedure, Figure 3 shows a selection of six of the input spectra. They have been corrected for the sky residuals but not yet continuum subtracted or normalized. The results are discussed using the Kennicutt (1992) eigenspectra. Figure 4 shows the corresponding χ^2 functions obtained. Calculated and manual redshifts are given in Table 2 along with the associated expansion coefficients.

Spectra (*a*)–(*d*) in Figure 3 are typical of the majority of the spectra studied. They give calculated redshifts that agree well with the manual redshifts. The corresponding χ^2 functions show clear minima giving unambiguous determinations of the redshifts. Spectrum (*e*) is noisier and had no manual redshift assigned to it previously; however, the χ^2 function gives a clear, albeit weaker, peak at a redshift of 0.06. Spectrum (*f*) is the spectrum of a quasar included as a deliberate outlier. As expected, the method clearly fails to find a redshift for this spectrum since there are no quasar spectra in the template set. There will always be a minimum value of χ^2, but it is clear from inspection of the corresponding χ^2 function that the associated redshift estimate is unreliable.

The failure of the method to find a redshift for the quasar illustrates the importance of including all spectral types of interest in the template set. The method will fail to find a

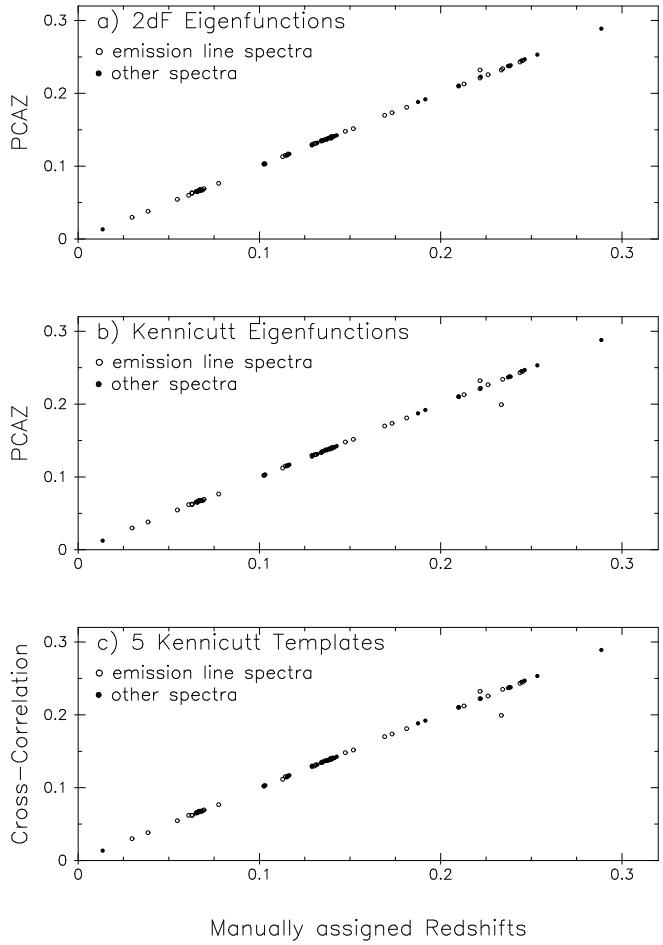

FIG. 6.—Comparison with manual redshifts in the SGP 463 2dF field for the three automated methods discussed in the text: (*a*) PCAZ redshifts determined using the eigenfunctions derived from the NGP 359 2dF field, (*b*) PCAZ redshifts determined using the eigenfunctions derived from the Kennicutt (1992) templates, and (*c*) simple cross-correlation with the Kennicutt templates, picking the best peak.

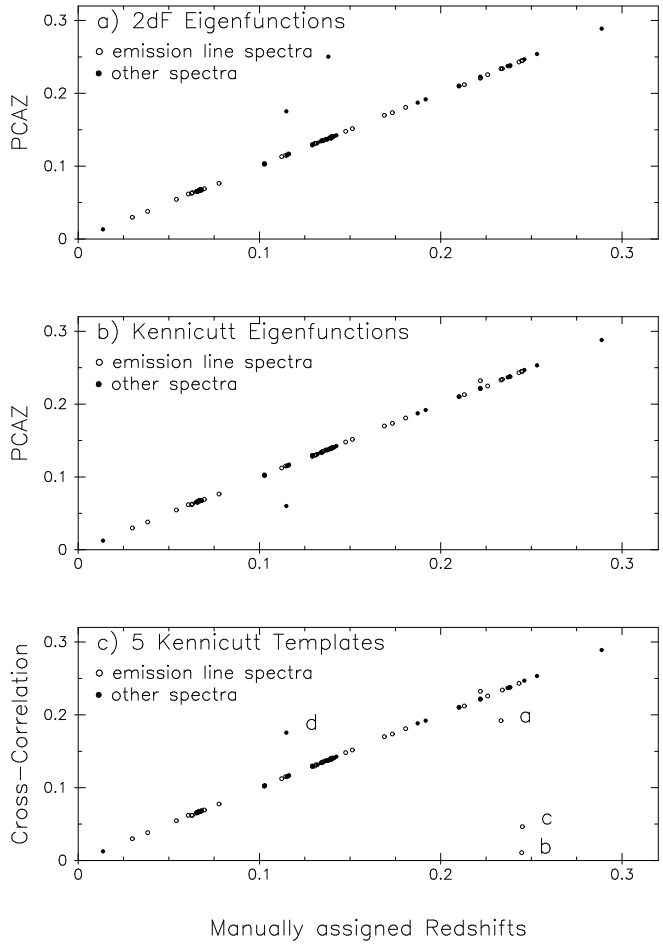

FIG. 7.—As in Fig. 6, this time with the continuum S/N degraded to the range 3–10 for both the test data (NGP 351) and in the 2dF field (SPG 463) used to construct the eigenfunctions in (*a*).

redshift for galaxies whose spectra differ fundamentally from the template set (e.g., in instrumental resolution). However, in principle the method will work for all spectral types, including emission-line galaxies, provided that the relevant spectral types lies within the *m*-dimensional space spanned by the eigenspectra. The power of the PCA method lies in its ability to reduce the dimensionality of linear problems from many templates to a few eigenspectra with no loss of accuracy and in its resulting ability to filter out the noise from noisy templates.

A side effect of this method is the ability to reconstruct "filtered" versions of the spectra from the eigentemplates. With only a few eigentemplates, the relative strength of the emission and absorption lines may not have fully converged, but a comparison of the reconstructed and original spectra helps to clarify how the method works. Figure 5 shows the reconstructed spectra corresponding to the first five original spectra shown in Figure 3. No reconstruction is given for the quasar spectrum.

Of course PCA will fail to reduce the problem space in nonlinear cases; a practical example might be if one has a sample of AGNs with very broad lines covering a large continuous range in velocity. However, for redshift work, most galaxy spectra are unresolved or only marginally resolved, in which case the variation can be accommodated in the eigenspectra.

The ultimate test of the method comes with a larger scale comparison of the manually determined redshifts with the calculated redshifts. Figure 6 shows the comparison between the manually determined redshifts and the two sets of PCAZ redshifts calculated with the two sets of eigenspectra. It is clear that the agreement between the PCAZ redshifts and the manually determined redshifts are very good for this field, with a success rate greater than 98%. The 2dF eigenfunctions performed the best, giving only one mismatch at a redshift of 0.23. Clearly we need somewhat more than 104 spectra to determine the error rate at this high level of success—something like 1000–2000 spectra are needed. We will look at this in more detail in the next paper in this series.

Poor sky subtraction remains a possible source of error in the automatic redshift determination. The PCAZ method took less than 2 minutes of computer time to calculate the 104 redshifts. The measured scatter of the points on the line is $\Delta z \leq 0.0005$, which is what we expected from the instrumental resolution.

With the PCAZ code, it is trivial to turn off the steps of orthogonalization and quadratic combination of cross-correlation functions—this enables us to reproduce the results of simple CCF analysis with the same template set.

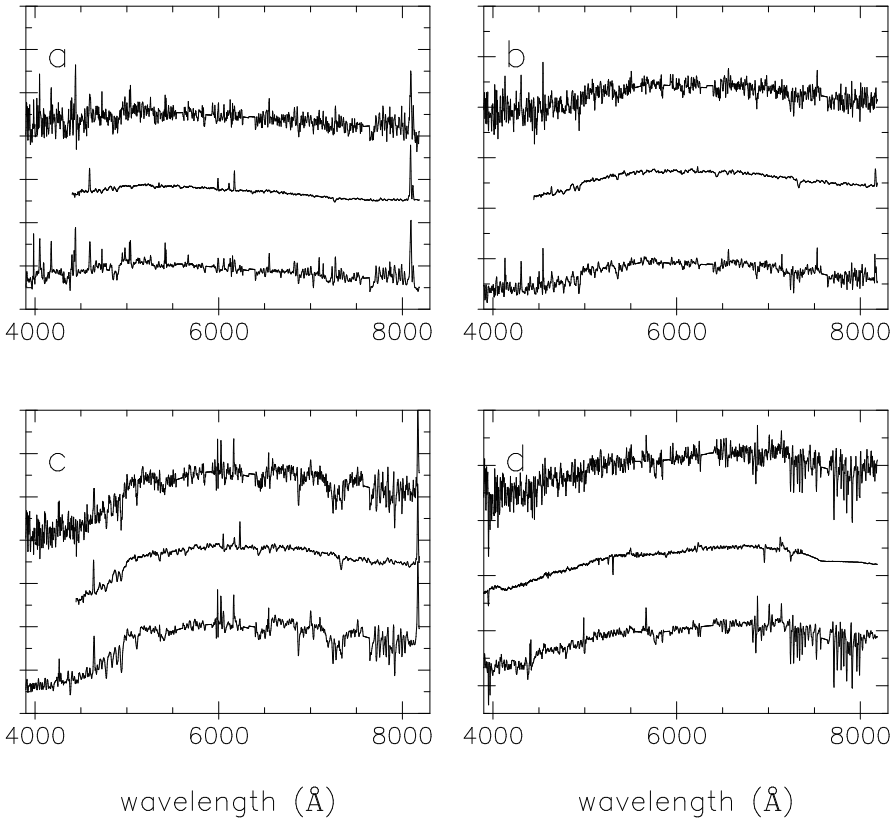

FIG. 8.—Noise-degraded spectra and reconstructions corresponding to points a–d in Fig. 7. The top curve in each panel is the noise-degraded input spectrum; the second curve is the PCA reconstruction from the first three Kennicutt (1992) eigenfunctions. The lowest curve is the original 2dF spectrum from which the input spectrum was derived.

This is also shown in Figure 6, where the Kennicutt (1992) template with the highest CCF peak gives the CCF redshift. It can be seen that for these high S/N spectra, the results are similar whether or not the templates are diagonalized. This simply reflects the excellent quality of the 2dF spectra with highly significant features for the algorithms to select. We anticipated that the PCAZ method would perform better than simple CCF for lower S/N spectra (much of the initial testing was done with such spectra before we had access to 2dF data). To demonstrate this, we add artificial Gaussian noise to the 2dF data, both data and templates, and decrease the continuum S/N by a factor of 3, so the galaxies typically have S/N = 3–10, and repeat our analyses. Rerunning the PCA analysis gives virtually identical 2dF eigenfunctions; the redshift results are shown in Figure 7. It is evident that PCAZ still performs at the 98% level while the CCF method has dropped to a 93% success rate. These results were obtained with minimal manual intervention and illustrate that PCAZ is more robust in the low-S/N regime.

Figure 8 shows four of the spectra for which the nonorthogonal cross-correlation method fails, labeled (*a*)–(*d*) in Figure 7. The lower curve in each panel shows the original 2dF spectrum. The top curve shows the spectrum plus added noise. The central curve is the PCA reconstruction of the noisy spectrum. Figure 9 shows the corresponding cross-correlation functions and χ^2 functions for the four spectra. The PCAZ results were calculated using the first three Kennicutt eigenfunctions derived from the same five Kennicutt spectra used for the nonorthogonal cross-correlation method. For the spectra (*a*)–(*c*), the PCAZ method correctly locates the redshift of the noisy spectra. The corresponding cross-correlation functions clearly show a peak at the same redshift, but the noise peaks are as large. PCAZ simultaneously uses many templates, effectively averaging over the CCF noise.

The fourth spectrum shows a case where both methods fail. The correct redshift is 0.115, but the presence of a sharp noise spikes, especially at 3950 Å, has introduced spurious correlations.

As surveys progress to thousands and tens of thousands of galaxies, we expect this relative advantage to increase: the derived eigenfunctions will include more subtle natural variations in the range of galaxy spectral features and will average over larger numbers of galaxies.

4. CONCLUSIONS

A new method of automatic redshift determination has been developed and shown to be capable of reproducing manually determined redshifts with a minimal amount of manual intervention. The method is a superior generalization of cross-correlation and has the potential to provide a sounder mathematical basis for confidence in the final redshifts. The expansion coefficients generated can be used to reconstruct noise-filtered versions of the spectra and have the potential to be used for a basic classification of the spectra. The method proves more robust in the low S/N regime than independent cross-correlation and has greater potential for very high success rates in upcoming very large redshift surveys.

This concludes the introduction and illustration of the mathematical principles behind PCAZ. In the next paper in

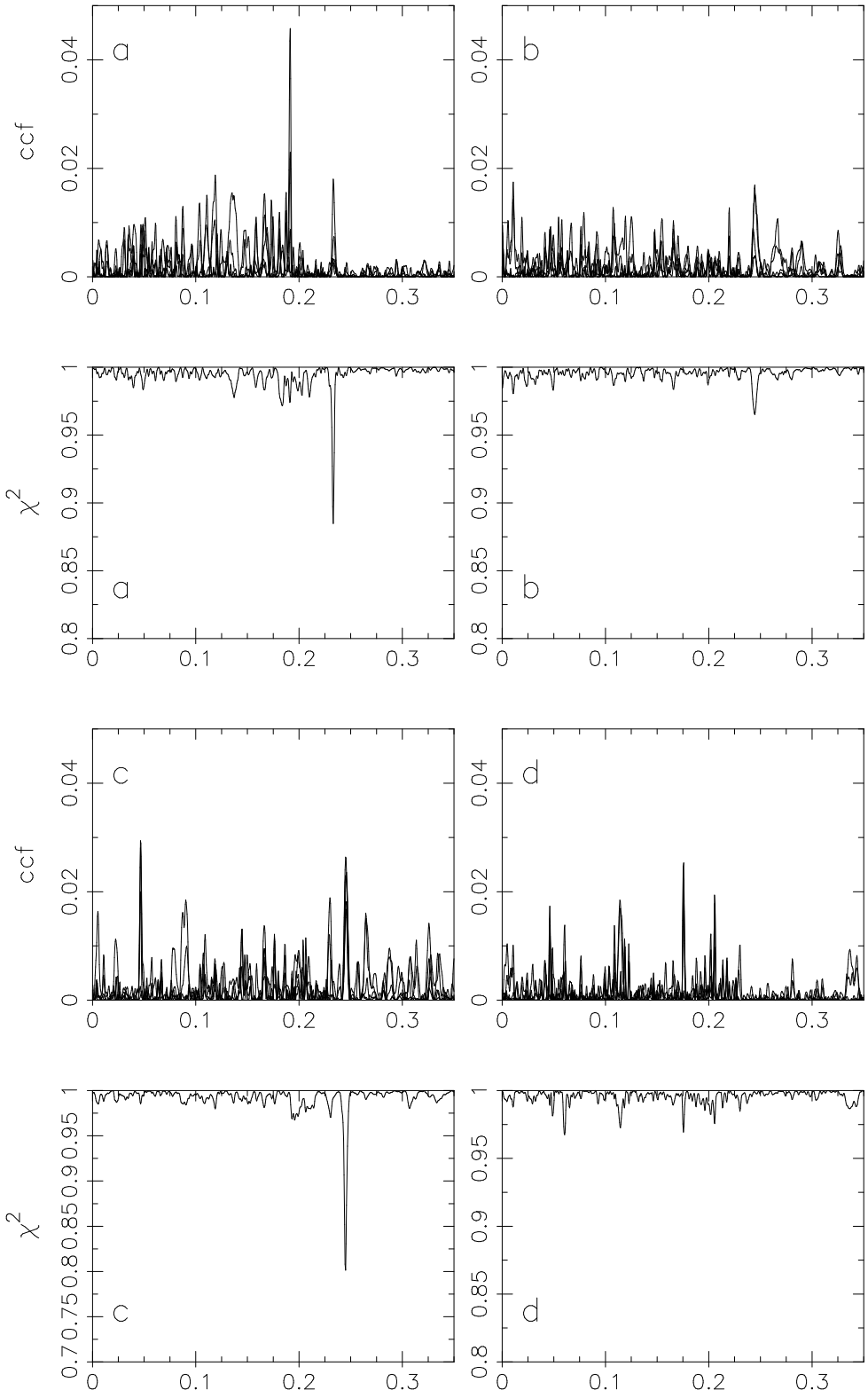

FIG. 9.—Cross-correlation functions from a simple nonorthogonalized cross-correlation approach and χ^2 from PCAZ for the spectra in Fig. 8. The five individual cross-correlation functions for each spectrum are plotted on the same graph.

this series, we will look in more detail at the reliability and the robustness of the method with much larger data sets, and we will consider in detail the treatment of the data with realistic errors and the robustness with S/N and compute typical probability distributions for redshift errors from PCAZ. We will also examine, via simulations, how this affects the measurement from very large redshift surveys of derived bulk galaxy properties such as $P(K)$ and the galaxy luminosity function.

The authors would like to thank Ofer Lahav and Joss Bland-Hawthorn for useful discussions and Matthew

Colless for providing one-half of the manual redshifts. We are grateful to the 2dF Galaxy Survey team for allowing us to use their early data for this paper and to the 2dF commissioning team at the Anglo-Australian Telescope for all their hard work. K. G. B. acknowledges the Anglo-Australian Observatory for their generous support of research for staff astronomers. K. D. was supported by an Anglo-Australian Observatory Summer Studentship. Computing facilities for this research were provided by the Anglo-Australian Observatory.

REFERENCES

Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. 1995, ApJ, 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, 376
Heavens, A. F. 1993, MNRAS, 263, 735
Kendall, M. G., & Stuart, A. 1966, The Advanced Theory of Statistics, Vol. 3 (London: Griffin)
Kennicutt, R. C. 1992, ApJ, 388, 310
Maddox, S. J., Sutherland, W. J., Efstathiou, G., & Loveday, J. 1990, MNRAS, 243, 692
Maddox, S. J., et al. 1998, in preparation
Mittaz, J. P. D., Penston, M. V., & Snijders, M. A. J. 1990, MNRAS, 242, 370
Murtagh, F., & Heck, A. 1987, Multivariate Analysis (Dordrecht: Reidel)
Shectman, S. A., Landy, S. D., Oemler, A., Tucker, D. L., Lin, H., Kirshner, R. P., & Schechter, P. L. 1996, ApJ, 470, 172
Sodré, L., & Cuevas, H. 1997, MNRAS, 287, 137
Strauss, M. A. 1996, in Formation of Structure in the Universe, ed. A. Dekel & J. P. Ostriker (Cambridge: Cambridge Univ. Press)
Taylor, K. 1994, in Wide Field Spectroscopy and the Distant Universe, The 35th Herstmonceux Conference, ed. S. J. Maddox & A. Aragón-Salamanca (Singapore: World Scientific), 15
Taylor, K., et al. 1998, in preparation
Tonry, J., & Davis, M. 1979, AJ, 84, 1511