A Source Classification Algorithm for Astronomical X-ray Imagery of Stellar Clusters

by

Susan M. Hojnacki

B.S. Electrical Engineering, Syracuse University

M.S. Computer Engineering, Rochester Institute of Technology

M.S. Imaging Science, Rochester Institute of Technology

A dissertation submitted in fulfillment of the
requirements for the degree of Doctor of Philosophy
at the Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

May 2005

Signature of the Author ______________________________

Accepted by ______________________________

Coordinator, Ph.D. Degree Program Date

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE

ROCHESTER INSTITUTE OF TECHNOLOGY

ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

Ph.D. DEGREE DISSERTATION

The Ph.D. Degree Dissertation of Susan M. Hojnacki
has been examined and approved by the
dissertation committee as satisfactory for the
dissertation required for the
Ph.D. degree in Imaging Science

Joel H. Kastner, Ph.D., Dissertation Advisor

Steven M. LaLonde, Ph.D.

Michael W. Richmond, Ph.D.

Carl Salvaggio, Ph.D.

Date

DISSERTATION RELEASE PERMISSION

ROCHESTER INSTITUTE OF TECHNOLOGY

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE

Title of Dissertation:

A Source Classification Algorithm for Astronomical X-ray Imagery of Stellar Clusters

I, Susan M. Hojnacki, hereby grant permission to Wallace Memorial Library of R.I.T. to reproduce my dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Signature ____

Date

A Source Classification Algorithm for Astronomical X-ray Imagery of Stellar Clusters

by

Susan M. Hojnacki

Submitted to the Chester F. Carlson Center for Imaging Science
in partial fulfillment of the requirements for the Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

The Chandra X-ray Observatory (*Chandra*) is producing images with outstanding spatial resolution using low-noise, fast-readout CCDs. Among many other things, X-ray images and spectra help astronomers study star formation and galactic evolution. Currently, X-ray astronomers classify one X-ray source at a time by visual inspection and use of model-fitting software. This approach is useful for studying the physics of bright individual sources but is time consuming for analyzing large images of rich fields of X-ray sources, such as stellar clusters. Objective and efficient techniques from the fields of multivariate statistics, pattern recognition, and hyperspectral image processing, are needed to analyze the growing *Chandra* image archive.

An image processing algorithm has been developed that orders the given X-ray sources based on hard versus soft X-ray emission and then groups the ordered X-ray sources into clusters based on their spectral attributes. The algorithm was applied to imaging spectroscopy of the Orion Nebula Cluster (ONC) population of more than 1000 X-ray emitting stars. As an initial test of the algorithm, images of the ONC from the *Chandra* archive were analyzed. The final spectral classification algorithm was applied to a sample of sources selected from among the more than 1600 X-ray sources detected in the *Chandra* Orion Ultradeep Project. Clustering results have been compared with known optical and infrared properties of the population of the ONC to assess the algorithm's ability to identify groups of sources that share common attributes.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AAS	American Astronomical Society
ACIS	Advanced CCD Imaging Spectrometer
ACIS-I	ACIS-Imaging
ANN	artificial neural network
APED	Astrophysical Plasma Emissivity Database
ASAS	All Sky Automated Survey
ASCA	Advanced Satellite for Cosmology and Astrophysics
AXAF	Advanced X-ray Astrophysics Facility
BI	backside-illuminated
CCD	charge-coupled device
CIAO	*Chandra* Interactive Analysis of Observations
COUP	*Chandra* Orion Ultradeep Project
CXO	Chandra X-ray Observatory
DEC	declination
FI	frontside-illuminated
FOV	field of view
FWHM	full-width half-maximum
HETG	High Energy Transmission Grating
HRC	High Resolution Camera
HRMA	High Resolution Mirror Assembly
IDL	Interactive Data Language
IPC	Imaging Proportional Counter
IR	infrared
ISIS	Interactive Spectral Interpretation System
LETG	Low Energy Transmission Grating
NIR	near infrared
NCC	normalized correlation coefficient
ObsIds	Observation Ids
ONC	Orion Nebula Cluster
PCA	principal component analysis
PMS	pre-main-sequence

PSF	point spread function
QE	quantum efficiency
RA	right ascension
ROSAT	Roentgen Satellite
SAS	Statistical Analysis Software
SIM	Science Instrument Module
XMM-Newton	X-ray Multi-Mirror Mission-Newton
XRB	X-ray background

Acknowledgements

This research was funded in part by grants from the Eastman Kodak Company and the Smithsonian Astrophysical Observatory.

The following people provided input and technical advice and I would like to thank each one of them: Eric Feigelson, Konstantin Getman, Giusi Micela, Norbert Schulz, David Huenemoerder, and Vinay Kashyap.

I would like to thank the members of my thesis committee for providing me with invaluable input during the course of my research. Dr. LaLonde taught me to question all the results, to continually ask “why”, and to go beyond the numerical answer to find its meaning. Dr. Richmond provided me with endless thought-provoking suggestions, ideas, and motivation. Dr. Salvaggio provided the imaging science and remote sensing point of view, balancing out the astronomy aspects of my research.

I’d like to thank all my friends who stood by me throughout the past 8+ years and the crazy 80+ hours per week of work and school. I’m thankful for their support and for dragging me out on bicycle rides to give my brain a break.

I am extremely grateful to my parents for teaching me perseverance and determination; for my Father’s unquestioning support and patience during my long pursuit of this degree; and for my Mother’s understanding when I missed family get-togethers and holidays. I owe my Mother several Mother’s Days, with interest.

Finally, I’d like to thank my advisor, Dr. Joel H. Kastner. He listened to all my tales of woe and always got me back on track. He never micromanaged my research and was a constant source of energy and enthusiasm. One must never underestimate the importance of having a good advisor.

Chapter 1

Introduction

A large fraction of the Chandra X-ray Observatory[1] (*Chandra*) observing time has been devoted to the study of young star clusters and, consequently, large datasets exist from these observations of rich stellar fields. X-ray images help astronomers study new star formation and galactic evolution. However, the physical processes responsible for X-ray emission from recently formed stars are not fully understood and are presently hotly debated within the X-ray astronomy community[2, 3, 4]. The growth of the *Chandra* archive of X-ray observations of young clusters has fueled this vigorous debate concerning the characterization of X-ray emission from young stars [5, 6, 7].

A typical *Chandra* charge-coupled device (CCD) observation of a young stellar cluster results in detection of X-ray emissions from tens to hundreds of very young stars. An example of this is shown in *Chandra*'s dramatic deep ~80 ks image of the Orion Nebula Cluster (ONC, Figure 1.1). *Chandra* has resolved more than 1000 X-ray emitting sources in this single image of the ONC, including X-ray sources associated with externally illuminated structures that are presumably planet-forming circumstellar disks[8,9].

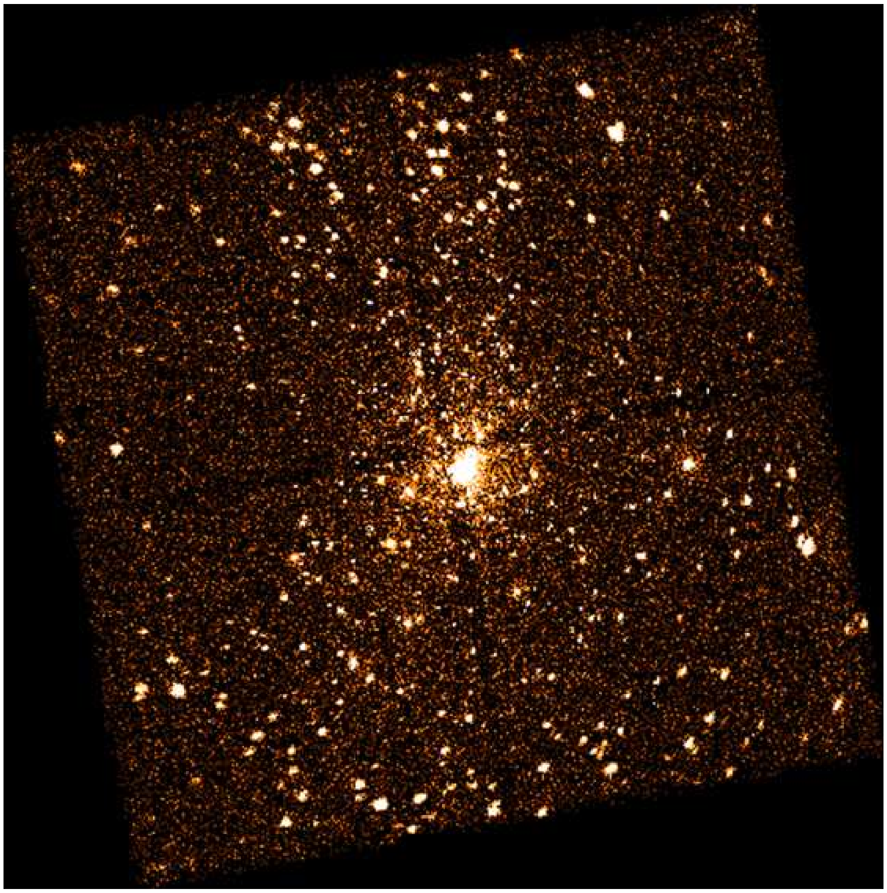


Figure 1.1: Chandra X-ray Observatory image of the ONC[8].

In addition, a new set of problems have been uncovered by X-ray images of young stellar clusters[5,8,9,10]. Among the challenges and puzzles are:

- Only very weak trends have been found when attempting to correlate model parameters derived from spectral fitting of individual sources (e.g., X-ray luminosity and temperature; X-ray absorbing column and visual extinction)

- There is no apparent relationship between the intensity of X-ray emission and the presence of circumstellar disks. For example, Preibisch et al.[11] have found weak anti-correlation between X-ray luminosity and indicators of accretion rate.
- Some X-ray sources show distinct spectral features that can be attributed to emission from specific ions; most do not
- A very wide range of temporal behavior has been detected, from long-term flaring to episodic, short X-ray bursts[12]
- Approximately 17.6% of the ~1616 detected X-ray sources in and around the ONC have no visible or infrared (IR) counterparts[68]

These puzzling observations are being studied by analyzing data from the *Chandra* Orion Ultradeep Project (COUP), an ~838 ks exposure of the ONC obtained over a nearly continuous period of ~10 days in January of 2003[12] (Figure 1.2).

Classification of X-ray sources is traditionally accomplished by visual inspection of individual X-ray source spectra and subsequent fitting of each source spectrum to various models, either manually, or by use of model-fitting software programs. One X-ray source is analyzed at a time using this approach and classification success is measured visually. This approach is useful for studying the physics of bright, individual sources. However, this can be a time consuming approach for analyzing large datasets created from observations of rich stellar fields.

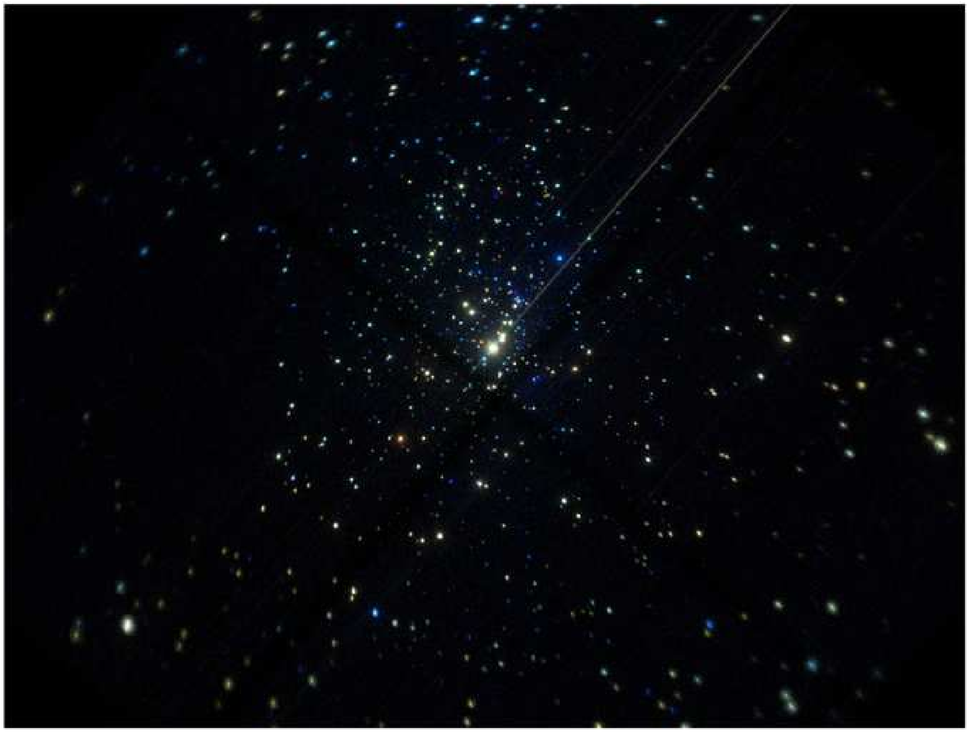


Figure 1.2: *Chandra* image of the ONC from the COUP observation.

The wealth of multidimensional data currently being produced by the X-ray CCD detector arrays onboard *Chandra* represents a far-reaching problem pervasive to many current astronomical missions. That is, the data archives of current missions have surpassed their predecessors, both in terms of number of sources detected and the information content available for each source. Given the detection of a very large number of X-ray sources, each of which is potentially well-resolved spectrally, spatially, and temporally, how does one best extract and analyze the available information? Is it possible to group detected sources into distinct categories or classes in an unbiased manner in order to better guide subsequent spectral analyses of individual sources? These questions suggest use of objective model-independent methods for spectral clustering of X-ray sources: methods that can take advantage of the vast collection of high-spatial resolution CCD spectral data now being acquired by *Chandra*.

My research involved exploring solutions to this problem using multivariate statistical and pattern recognition techniques. Use of techniques from these fields is not new to astronomical data analysis (see Chapter 4), but are previously untested in the context of X-ray spectral data from *Chandra*. The goal of my research was to develop an X-ray source clustering algorithm with the following capabilities:

- Find natural groupings of X-ray sources in stellar clusters
- Process large datasets created from observations of rich stellar fields
- Perform without a priori information concerning the nature of the sources
- Use an approach that is objective and model-independent
- Consist of as few manual steps as possible

Sources within the same group may be sufficiently similar to be treated identically for the purpose of further astronomical analysis, where this would be impossible for the whole heterogeneous star field.

The expected scientific significance of this approach includes the potential to:

- Determine relationships between X-ray and visible spectral classes and parameters
- Uncover classes of sources that do not fit any existing models
- Identify extreme outliers of interest among all the sources in a stellar field

- Identify groups of sources that have no visible or IR counterparts or that are poorly characterized in other wavelength regimes
- Identify groups of contaminating and interloping sources so that researchers can eliminate them from subsequent statistical studies
- Increase productivity of X-ray archival research due to the ability of the resulting algorithm to process and categorize larger quantities of data than could be done manually

Chapter 2 contains a brief background on X-ray astronomy. In Chapter 3, I provide a description of the relevant subsystems of *Chandra* and its imaging capabilities. Chapter 4 contains a review of applications of multivariate statistical and pattern recognition techniques to current and past astronomical problems. Challenges specific to X-ray data are also provided in Chapter 4. Chapter 5 contains a description of the mathematical techniques used in my research. In Chapter 6, I define the multivariate variables used as input into the algorithm. A proof of concept is presented in Chapter 7. The X-ray source classification algorithm is then detailed in Chapter 8. The analysis of results is presented in Chapter 9. Finally, a summary is presented in Chapter 10.

Chapter 2

X-ray Astronomy

2.1 History

X-ray astronomy dates back to 1949 when it was discovered that the Sun emits X-rays[13] . Since that time, many interesting sources of X-ray emission have been discovered in the universe. In the early 70's, NASA's Uhuru[14] astronomy satellite discovered a number of X-ray binary stars, in which an ordinary star orbits a super dense neutron star that emits X-rays as it pulls matter from the ordinary star. In the late 70's and early 80's, NASA's Einstein Observatory discovered that cataclysmic variable stars in our own galaxy emit X-rays when they are in outburst. The Einstein Observatory also collected the first X-ray images of pulsars and supernova remnants. The imaging ability of the Einstein Observatory changed the way X-ray astronomers conduct their research, with the detection of thousands of discrete sources of X-ray emission. This trend toward high-resolution X-ray imaging spectroscopy accelerated in the mid 90's with the advent of

Roentgen Satellite[15] (ROSAT). ROSAT, a joint project of the United States, Great Britain, and Germany, was used to expand the number of known X-ray sources to over 60,000. The availability of ROSAT proportional counter data led to the widespread use of X-ray hardness ratios (the Hertzsprung-Russell diagrams of X-ray astronomy) for source classification[16] .

The Advanced Satellite for Cosmology and Astrophysics[17] (ASCA), the follow-on to ROSAT, featured improved spectral resolution, albeit with inferior spatial resolution. ASCA's demonstration of the application of CCDs in X-ray astronomy paved the way for *Chandra* and the X-ray Multi-Mirror Mission-Newton[18] (XMM-Newton). *Chandra*, one of NASA's Great Observatories, was launched in 1999. Within months, an X-ray source at the center of our galaxy that is believed to be a supermassive black hole was discovered from the X-rays emitted from superheated matter nearing its event horizon.

2.2 X-ray Properties

The wavelength range for the X-ray portion of the electromagnetic spectrum is from about 0.01 nm to about 10 nm, which corresponds to a range of 0.1 Å to 100 Å, (10 Å = 1 nm = 10^{-9} m). The wavelength of an X-ray photon is less than a millionth of a centimeter: about a thousand times shorter than a visible-light photon. Extremely hot gases and charged particles moving at nearly the speed of light emit X-rays. Material that is at a very high temperature (millions of degrees Kelvin) emits X-rays. Temperatures this high can occur in extremely dense objects, in large magnetic fields, or from explosive forces.

The energies of X-ray photons are typically measured in electron volts and range from 0.1 keV to 10 keV. Higher energy X-rays are referred to as “hard” X-rays while lower energy X-rays are referred to as “soft” X-rays. The boundary between the two types is not well defined, but is generally placed around 2 keV [19]. The highest energy X-rays can penetrate more deeply into a substance than soft X-rays, and therefore, require a denser detector containing material that is more massive.

X-ray photons emitted by a constant source or a source that is at least constant for some time interval will form an independent Poisson process for each energy interval. The counts in a given time interval will then be a Poisson-distributed random variable[20] .

2.3 X-rays from Young Stars

A star spends most of its life in what is known as the “main-sequence phase” in which it produces power by nuclear fusion of hydrogen into helium. Young stars are called pre-main-sequence (PMS) stars if they have not yet begun to burn hydrogen. These very young stars are constantly changing in X-ray brightness, sometimes within half a day. Star birth occurs within dense, molecule-rich and dust-rich cores of interstellar gas clouds. As the star-generating part of the core collapses, it flattens so as to conserve angular momentum. The central region of the collapsing cloud will form a star, while the flattened structure surrounding this protostar can eventually form planets orbiting the star. This flattened structure is called a protoplanetary disk and can be quite thick. The cloud core can be optically opaque, such that visible and even infrared (IR) light cannot escape the star's immediate vicinity, particularly if the star is viewed through its own disk almost edge-on. However X-ray photons are somewhat more penetrating than even IR photons, especially at energies greater than 2 keV [9]. A large number of PMS stars in the ONC have only been detectable in X-rays thus far. Therefore, X-ray astronomy may be used to penetrate these star-forming regions to detect stars in very early stages of formation that are inaccessible to optical and IR observations.

Young stars, with or without surrounding, planet forming disks, emit X-rays at rates thousands of times higher than middle-aged stars such as the Sun. These X-rays often are emitted during flares that are thought to arise from the release of energy stored in highly tangled magnetic fields near the surface of the star, similar to magnetic flares from the Sun. However, young stars release much more frequent and violent flares, reaching temperatures of ~100 x 10^8 Kelvin[10]. It is possible that some of this energy release is derived from magnetic reconnection events resulting from interactions between a young star and its circumstellar, protoplanetary disk[21]. Newborn stars at the center of nebulae emit extremely strong bursts of X-rays. One particular rich sample of PMS stars can be observed in a relatively compact region within the Great Nebula in Orion. This cluster is called the Orion Nebula Cluster (ONC)[8].

2.4 Orion Nebula Cluster

At a distance of about 450 parsecs[a], the ONC is the richest stellar nursery in the solar neighborhood. Within the ONC radius of less than ~3 parsecs is an association of young stars (< 1 Myr), most of them X-ray sources. At the core of the ONC is a very young, closely packed group of stars and protostars that are only a few hundred thousand years old. Many of these stars emit extremely strong bursts of hard X-rays. A *Chandra* Advanced CCD Imaging Spectrometer – Imaging (ACIS-I, see Chapter 3) image of the ONC is shown in Figure 1.1. The detected sources range from a few photon counts to several thousand photon counts. Some of the detected X-ray sources are very faint, resulting in approximately only 6 detected photons[22]. Figure 2.1 shows the Hubble Space Telescope image of the Trapezium region of the ONC. Contours from *Chandra* X-ray data of the same region have been overlaid on the optical image. As can be seen in this image, some X-ray sources have no visible counterparts.

2.4.1 X-ray Background

The X-ray background (XRB) was detected during a rocket flight whose scientific purpose was to study X-ray emission from the Moon, but instead found the first extra-solar X-ray source (Sco X-1) and the XRB[23]. Instrumental effects can also contribute to the perceived background radiation.

[a] 1 parsec = 3.26 light years

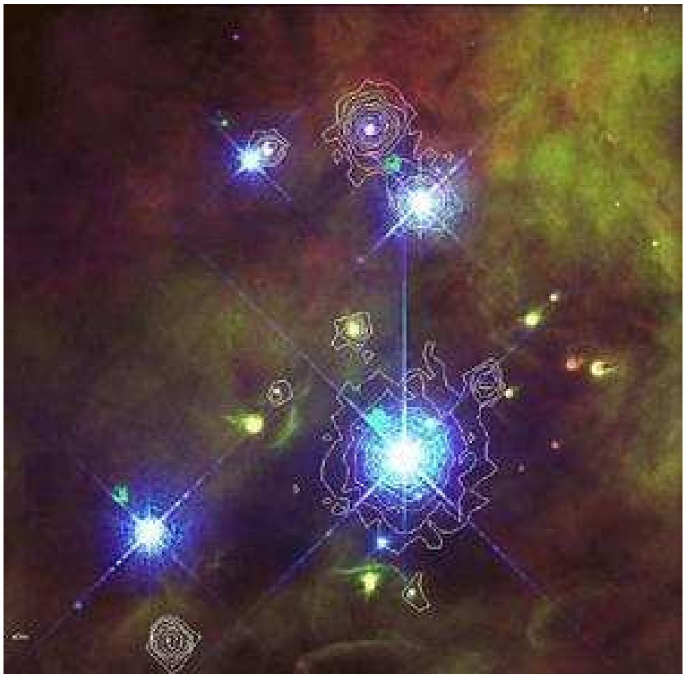


Figure 2.1: Hubble Space Telescope image of the Trapezium region of the ONC[9]. The contour lines from the Chandra X-ray Observatory are overlaid on the visible image.

Chapter 3

Chandra X-ray Observatory

3.1 Background

X-rays are absorbed by the Earth's atmosphere. Therefore, a space-based telescope is needed to image X-ray emitting space-based objects. *Chandra* was carried up on the Space Shuttle Columbia during a night launch on July 23, 1999 from the Kennedy Space Center in Florida. The observatory reached its final orbit location on August 24, 1999, after a series of five burns of the Integral Propulsion System. *Chandra's* orbit is elliptical with a perigee of 250 miles and an apogee of 45,014 miles: more than one-third of the way to the moon (see Figure 3.1). The period is 24 hours and 38 minutes and the Earth's radiation belts are crossed on every orbit. At perigee, *Chandra* travels at approximately 22,000 miles per hour.

3.2 Hardware

A schematic of the observatory is shown in Figure 3.2. The hardware relevant to my research includes the High Resolution Mirror Assembly (HRMA; Figure 3.3) and the Advanced CCD Imaging Spectrometer (ACIS; Figure 3.4).

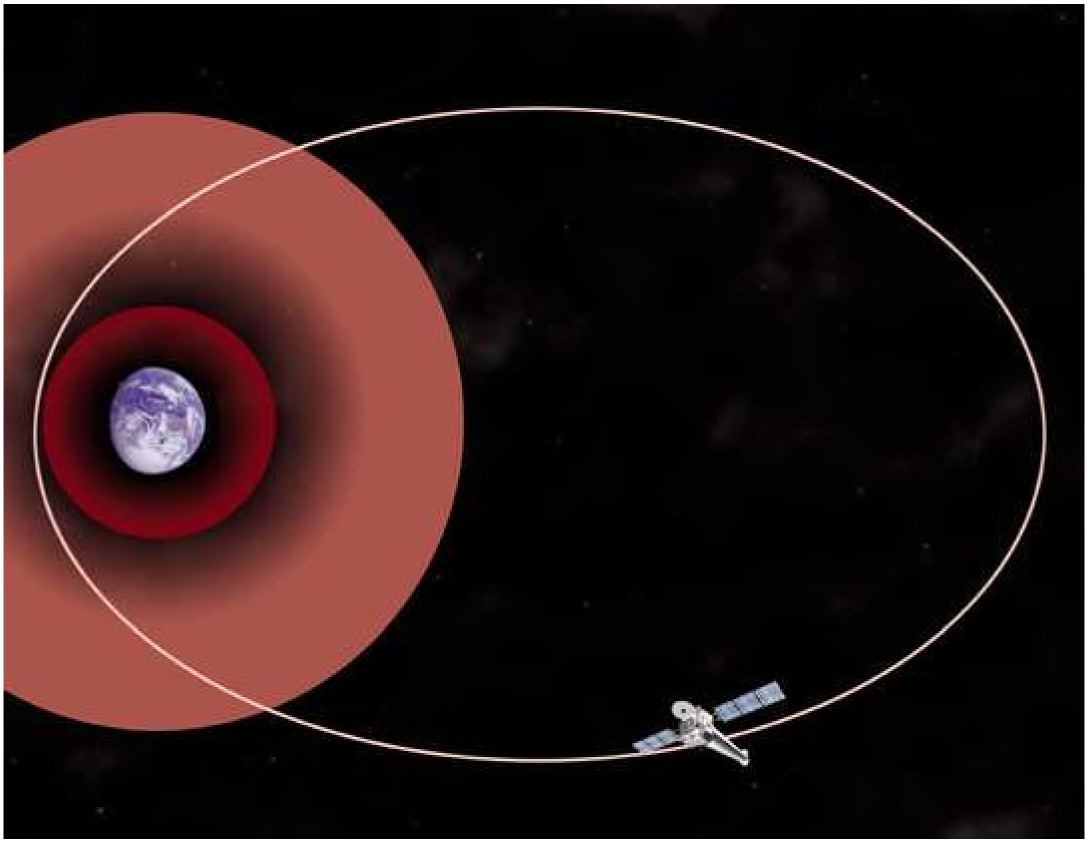


Figure 3.1 The orbit of *Chandra* shown from above. The pink bands encircling the Earth represent the radiation belts (Illustration: *Chandra* X-ray Center/M. Weiss).

3.2.1 HRMA

X-ray telescopes use grazing incidence optics so photons are not absorbed by the optics. *Chandra*'s X-ray mirrors are capable of resolving sources that are of the order of an arcsecond apart. The HRMA consists of two sets of four concentric nested mirrors: one set of paraboloid-shaped mirrors and one set of hyperboloid-shaped mirrors (see Figure 3.3). This configuration increases the photon collection area while deflecting the paths of the photons towards the focal surface.

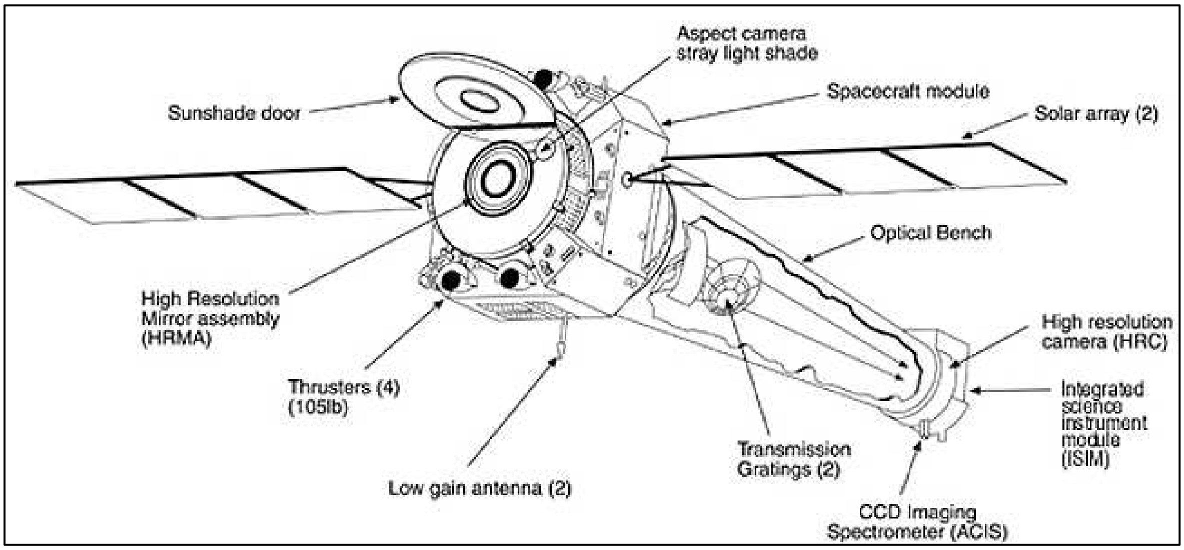

Aspect camera stray light shade
Spacecraft module
Solar array (2)
Sunshade door
Optical Bench
High Resolution Mirror assembly (HRMA)
High resolution camera (HRC)
Thrusters (4) (105lb)
Integrated science instrument module (ISIM)
Transmission Gratings (2)
Low gain antenna (2)
CCD Imaging Spectrometer (ACIS)

Figure 3.2: Schematic of the Chandra X-ray Observatory (Illustration: *Chandra* Proposers' Observatory Guide).

3.2.2 ACIS

X-ray CCDs are essentially similar in design to visible light CCDs. However, in visible light imaging systems, ensembles of photons arrive within a given observing interval at each pixel of the CCD. In contrast, X-ray CCDs are operated in a manner such that, ideally, photons can be counted one at a time. Another key difference involves the number of electrons that are liberated by one photon. Whereas a visible light photon will liberate one electron, an X-ray photon can liberate many electrons within the silicone of the CCD because the number of electrons that are liberated depends on the energy of the photon. Photon energies can be determined if the X-rays are detected individually.

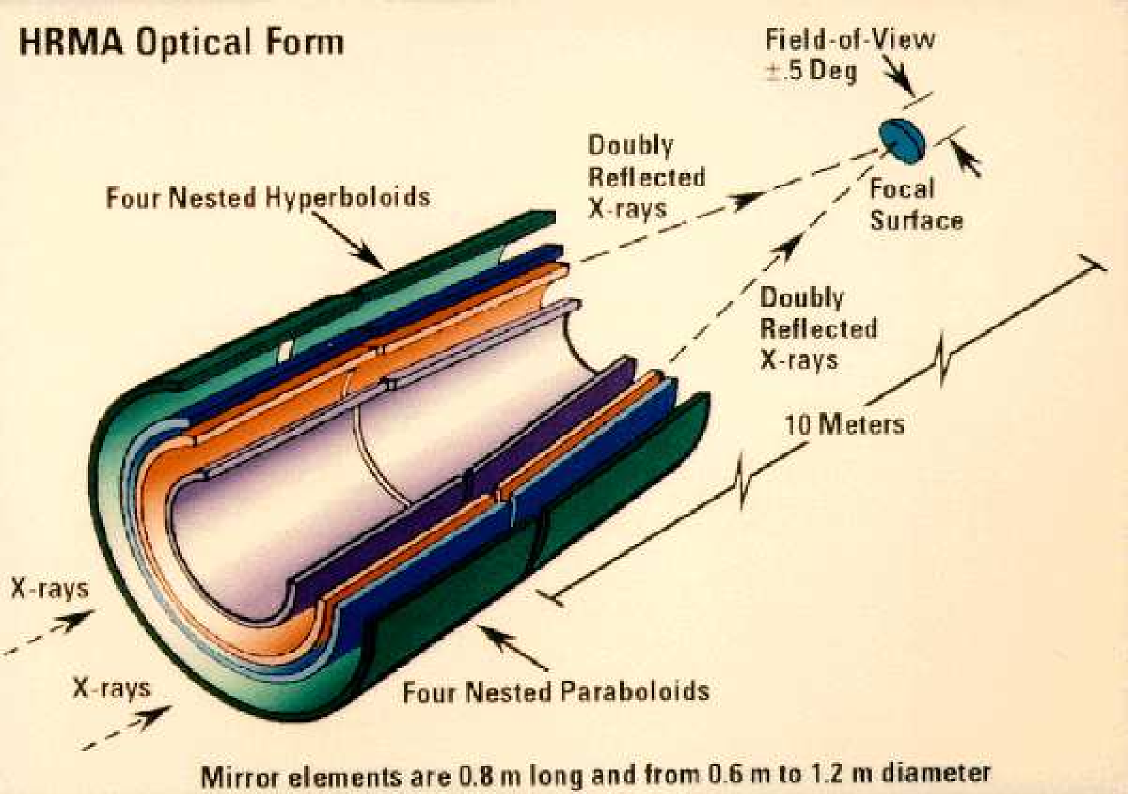


Figure 3.3: High Resolution Mirror Assembly configuration (Illustration: Hughes Danbury Optical Systems).

The field of view (FOV) is the total amount of sky that can be imaged in one frame. The ACIS has an angular resolution of 0.49 arcseconds with an FOV of 16 arcminutes by 16 arcminutes. The ACIS consists of 10 planar CCDs, each with 1024 by 1024 pixels (Figure 3.5) with a pixel size of 24 µm. Four of the CCDs are arranged in a 2x2 array (ACIS-I) and are used for imaging. The remaining six are arranged in a 1x6 array (ACIS-S) and are used either for imaging or as a detector for the transmission grating spectrometers aboard *Chandra*. ACIS-I was used for the archival observations used in my research. If ACIS-I is selected in “imaging” mode, chips I0-I3 plus chips S2 and S3 are used[24].

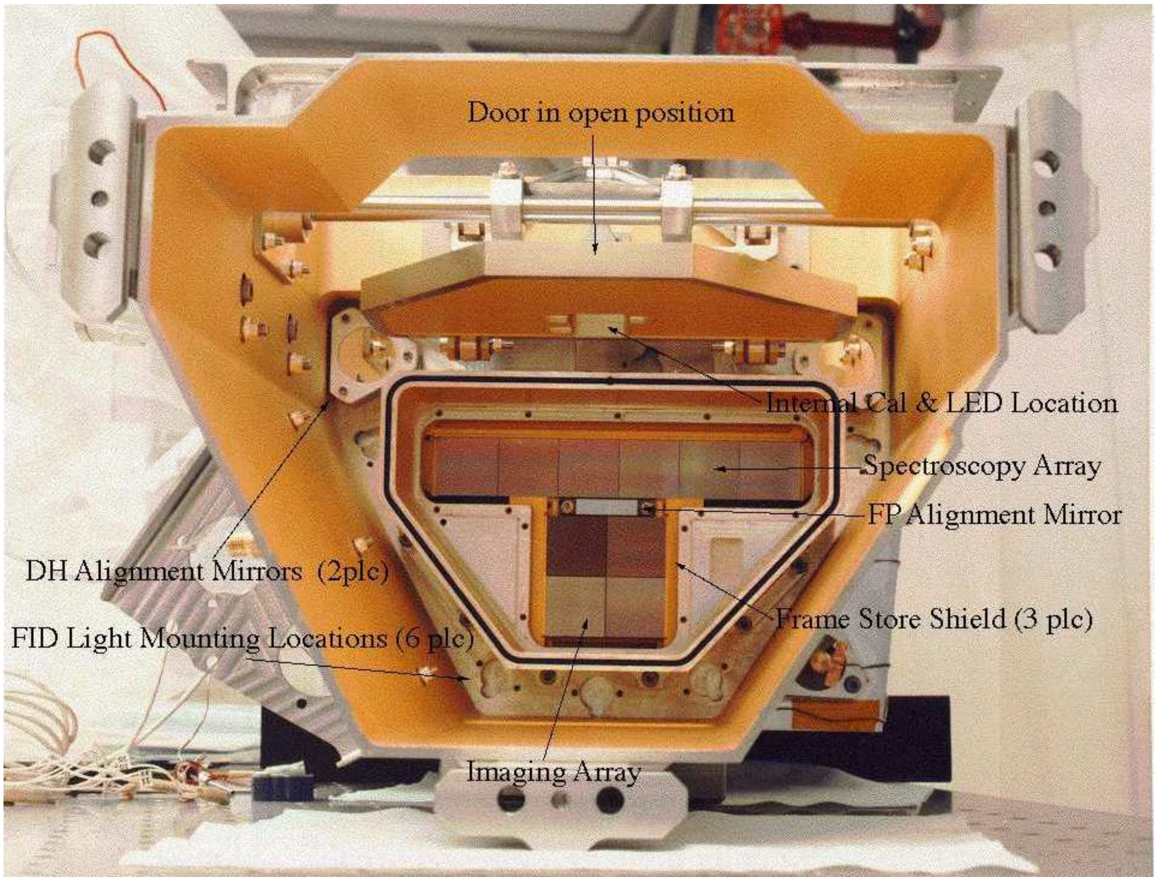

Door in open position
Internal Cal & LED Location
Spectroscopy Array
FP Alignment Mirror
DH Alignment Mirrors (2plc)
Frame Store Shield (3 plc)
FID Light Mounting Locations (6 plc)
Imaging Array

Figure 3.4: Photo of the Advanced CCD Imaging Spectrometer

See Table 3.1 for a summary of ACIS characteristics. Two characteristics of CCDs are quantum efficiency and charge transfer efficiency. Quantum efficiency is the percentage of incident photons that actually produces detectable charge in the depletion region. See Figure 3.7 for the quantum efficiency curve for the ACIS-I chips. Charge transfer efficiency (CTE) is the fraction of charge that is successfully transferred from pixel to pixel during one CCD transfer cycle.

$$\text{CTI} = 1 - \text{CTE}$$

where CTI is the charge transfer inefficiency.

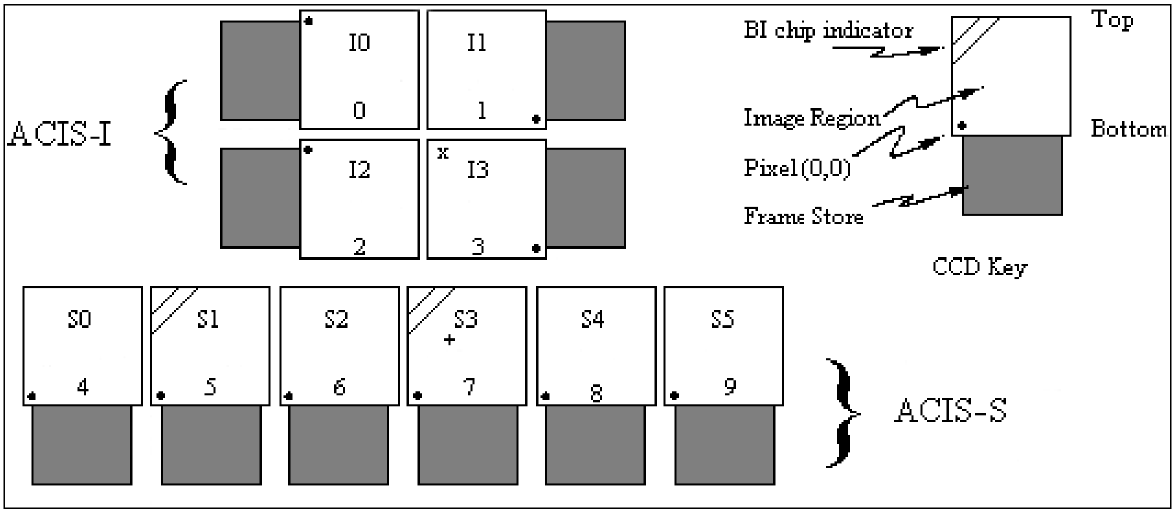


Figure 3.5: A schematic of the ACIS flight focal plane showing the 4 chips used for imaging (ACIS-I) and the 6 chips used for spectroscopy (ACIS-S).

Table 3.1: ACIS Characteristics

CHARACTERISTIC	VALUE
CCD format	1024 by 1024 pixels
Pixel size	24 microns
Array size	ACIS-I : 16.9 by 16.9 arcmin ACIS-S: 8.3 by 50.6 arcmin
On-axis effective area	110 cm2 @ 0.5 keV (FI)
Quantum Efficiency	> 80% between 3.0 and 5.0 keV
frontside illumination	> 30% between 0.8 and 8.0 keV
Quantum Efficiency	> 80% between 0.8 and 6.5 keV
backside illumination	> 30% between 0.3 and 8.0 keV
Charge Transfer Inefficiency (parallel)	FI: ~2x10^{-4} BI: ~2x10^{-5}
Charge Transfer Inefficiency (serial)	BI (S3): ~7x10^{-5} BI (S1): ~1.5x10^{-4} FI: < 2x10^{-5}
System noise	< ~2 electrons (rms) per pixel
Nominal frame time	3.2 sec (full frame)
Event threshold	FI: 38 ADU (~140 eV) BI: 20 ADU (~70 eV)

All but two of the chips on the ACIS are frontside-illuminated (FI). The FI chip gate structures are facing the incident X-ray beam. However, the backs of chips S1 and S3 have had treatments applied to remove insensitive, undepleted, bulk silicon material, thereby leaving the photo-sensitive depletion region exposed. These two chips have their backs facing the HRMA and are called backside-illuminated (BI). They were designed to improve the quantum efficiency at low energies.

Before launch, the ACIS FI CCDs approached the theoretical limit for energy resolution for almost all energies[1]. After launch, it was discovered that there was some degradation in the quality of the FI CCDs, exhibited by the energy resolution as a function of row number with the largest degradation in the farthest row from the frame store region. It is believed that the damage was caused by low energy protons that reached the focal plane during radiation belt passages[1]. As a result, the operating procedure was changed to move the ACIS out of the focal plane during radiation belt passages. Therefore, the resulting energy resolution for the FI CCDs is a function of row number due to the increase in CTI from radiation damage. An ACIS CTI correction has been developed and is now applied as part of the standard processing[25]. The full-width half-maximum (FWHM) of the FI detectors increases with increasing energy (see Figure 3.6). The energy resolution for the two BI CCDs is the same as their pre-launch values.

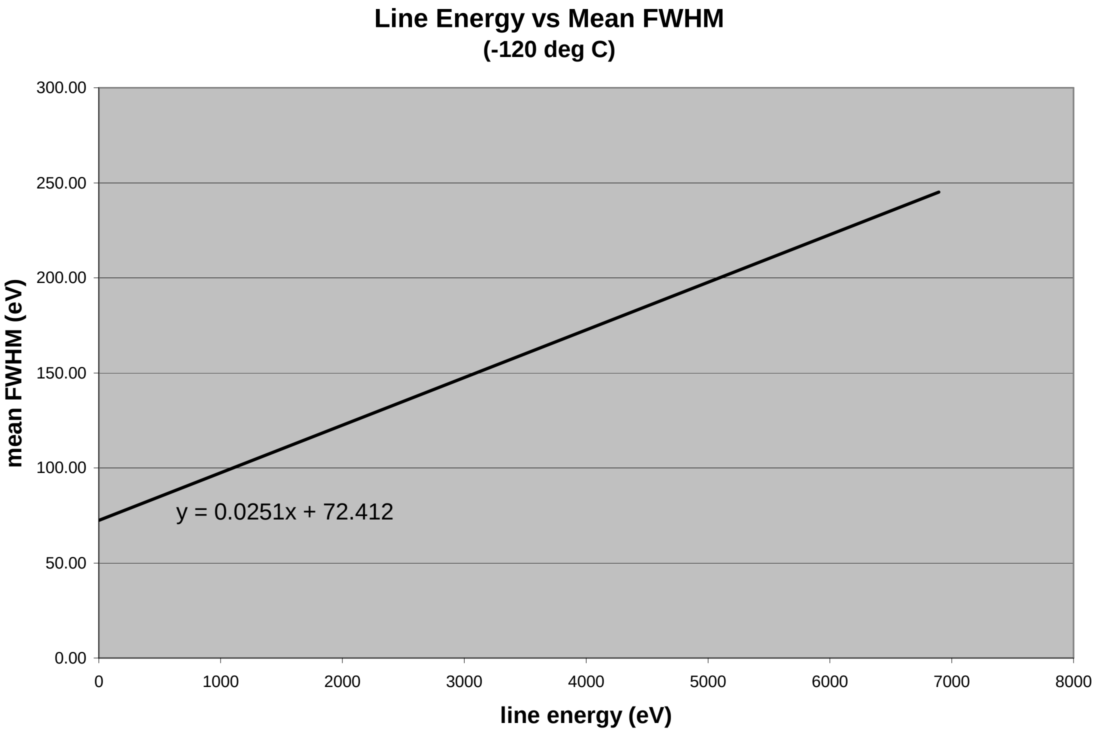


Figure 3.6: Plot showing how the FWHM of the FI CCDs increases with increasing energy. This data is after CTI correction.

There are several sources of noise in a CCD imaging system. One source is photon counting noise (also called shot noise). Photon noise includes random fluctuations in the photon stream of the source due to the quantum nature of light. The rate at which photons are received has a Poisson distribution. Other sources of noise are read noise, due to CCD readout electronics, and thermal noise generated by dark current. The total noise for ACIS is shown in Table 3.1.

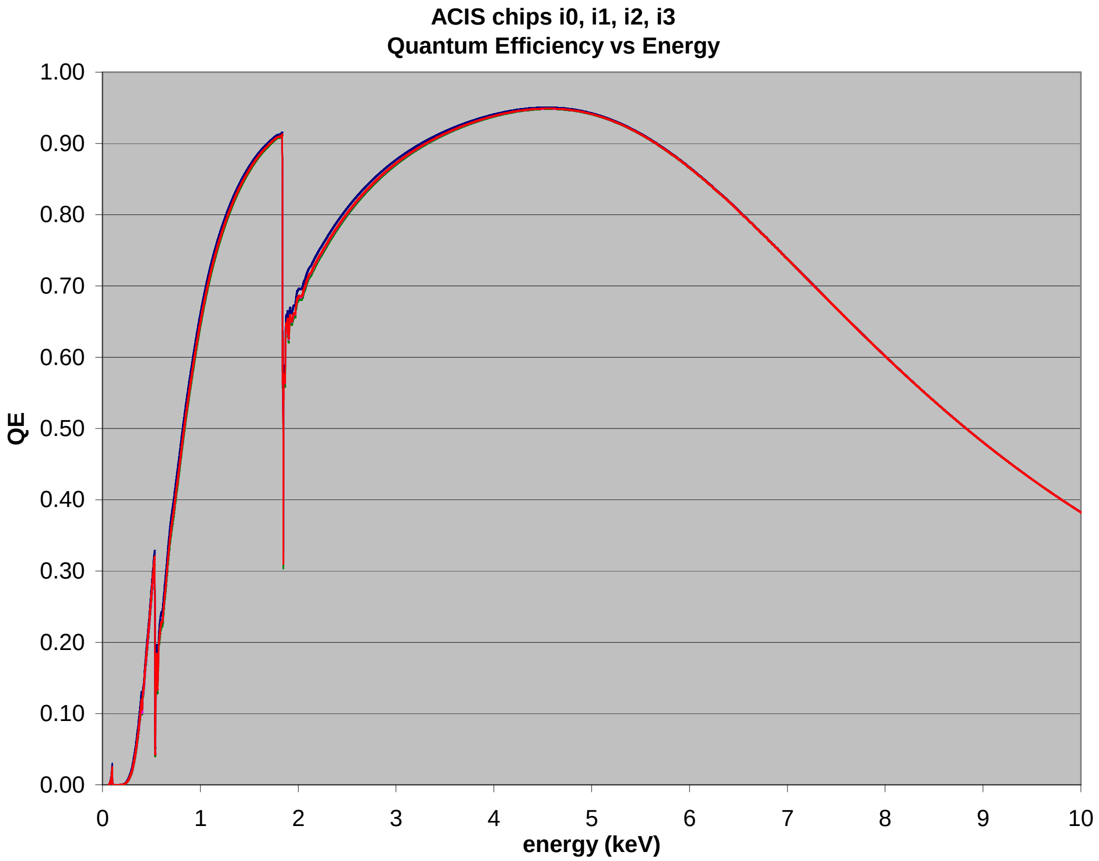


Figure 3.7: Quantum efficiency curves for the four front-illuminated ACIS-I chips showing the absorption features (07/2000 version of the data[b]).

The ACIS operates in X-ray photon counting mode. The energy of a photon with frequency ν is given by

$$E = h\,\nu$$

where h is a constant from quantum theory known as Planck’s constant. The X-ray photon arrival time follows a Poisson distribution. X-ray photons arriving at the ACIS are called events or counts. Software onboard *Chandra* records each event's two-dimensional spatial location, energy, and arrival time. Each event is assigned values for x and y in “sky” coordinates. These coordinates can be converted to a position in right ascension (RA) and declination (DEC). Since

[b] From *Chandra* X-ray Center Calibration Website:
http://cxc.harvard.edu/cal/Acis/Cal_prods/qe/08_11_04/qe.html

the CCD is dithered around on the sky during an observation, there is a complex, although typically very well-determined, time-dependent relationship between CCD pixel x and y, sky x and y, and RA and DEC. Therefore, the energy and arrival time, as well as the position of each photon, are known. Thus, in principle, the data can be represented by a four-way table of counts[26]. Due to instrumental constraints, each of these quantities is binned or rounded, creating a discrete variable.

For ACIS, if an X-ray source is bright, there is a non-negligible probability that two or more photons could land in the same pixel before readout of the ACIS frame. The detector will not be able to discern that there were multiple events and the individual photon energies will be unknown. This is called photon pileup[27]. The nominal frame exposure time is 3.2 seconds (full frame). The amount of time it takes to transfer data to the frame store is approximately 41 ms. The count rate at which a source is flagged as possibly exhibiting pileup for the COUP observation is approximately 0.003 counts/sec/pixel[12].

From the four-way table of counts data, a spectrum and an X-ray light curve can be constructed for each detected source (Figure 3.8). This data provides the potential for astrophysical insight into individual X-ray sources, and, in the case of a rich stellar cluster such as the ONC, to establish the *global* X-ray spectral and temporal properties of various classes of objects (e.g., low-mass versus high-mass pre-main-sequence stars; accreting versus non-accreting stars; cluster members versus contaminating foreground and background X-ray sources).

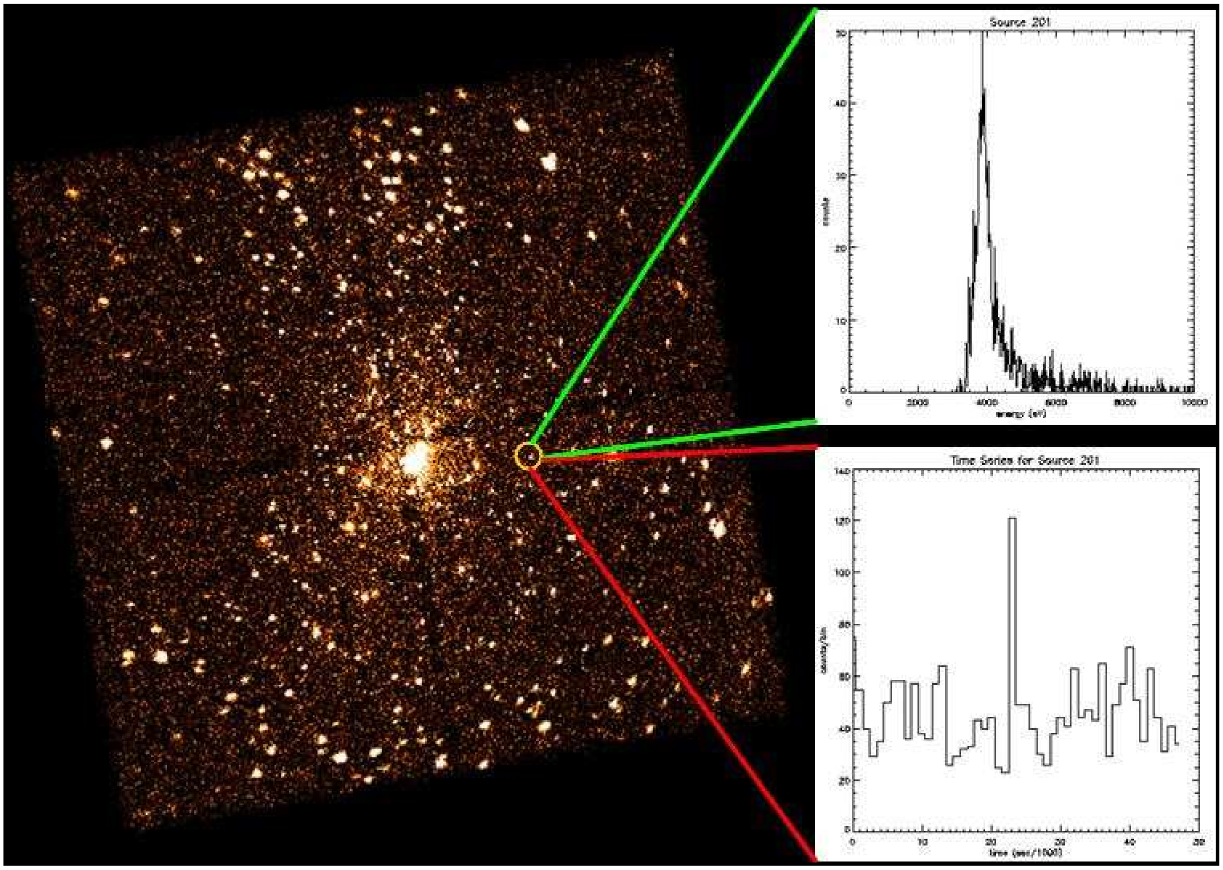


Figure 3.8: Extraction of energy spectrum (top) and light curve (bottom) for a detected X-ray source (Image from Ref. 8).

3.2.3 Heisenberg Uncertainty Principle

It is interesting to look at the Heisenberg Uncertainty Principle as it relates to *Chandra*. A form of the quantum mechanical principle due to Heisenberg states that it is not possible to determine the energy and time of a particle at a specific time. The simultaneous measurement of energy and time for a moving particle entails a limitation on precision (standard deviation) of each measurement. Moreover, the more precise the measurement of energy, the more imprecise the measurement of the time, and vice versa[28]. For example, at a precise time t, the energy of the particle is not determinable to a precision greater than $h/4\pi$.

$$\Delta E \, \Delta t \geq h / 4\pi$$

where,

ΔE is the uncertainty in the energy measurement

Δt is the uncertainty in the time measurement when the energy is measured

h Planck's constant , 6.6262×10^{-34} J s

For *Chandra*, Δt is equal to 3.2 seconds. This requires that the energy resolution of *Chandra* be greater than or equal to 1.02×10^{-16} eV. *Chandra*'s energy resolution well exceeds this number and indeed, current technology does not even approach this number.

3.3 Ground Data Processing

Level 0 processing takes raw *Chandra* telemetry, splits it into products that correspond to the different spacecraft components and then divides the data along observation boundaries. Level 1 processing takes Level 0 output and applies instrument-dependent corrections, including aspect determination (pointing position of *Chandra* versus time), science observation event processing, and calibration[29].

Chapter 4

Astronomical Applications of Data Mining

4.1 Background

Pattern recognition emphasizes feature selection and classification techniques[30]. It is defined as the grouping of objects into distinct classes by examining significant attributes of the objects[31]. The set of these attributes of the objects is called a feature vector. The feature vector method is dependent on finding features that are invariant to the expected changes in the features between the pattern classes and the amount of discriminating information contained in the features chosen[31]. Classification then takes place using a statistical method such as a similarity measure, a distance measure, or a probability function, as in the maximum likelihood method and Bayesian methods. There are two types of classification methods: supervised and unsupervised. In supervised classification or learning, part of the classifier design involves training the classifier using samples for which the class membership is known. The algorithm tries to group the samples of the training set into classes that match their predefined labels. The accuracy of the classifier design is tested on a separate set of sequestered samples. When an acceptable level of accuracy is achieved, the internal state of the classifier is saved. The algorithm is then used to classify new objects of unknown class. An example of a supervised classification method is the neural network. In unsupervised classification, or cluster analysis, the classifier forms "natural" groupings of the input samples[32]. Cluster analysis is a multivariate statistical technique that compares and groups objects based on a set of variables representing characteristics of the objects to be grouped, not on an estimation of those variables themselves. This makes the researcher's definition of the set of variables critical to the success of the clustering[33]. Supervised methods typically outperform unsupervised methods, however they are incapable of discovering new classes of objects and accounting for extreme outliers of possible interest[34].

Combinations of classification techniques, as opposed to a single classification technique, may show better clustering results[35]. Bazell and Aha[36] found that combining the results of an ensemble of classifiers gave better classification results than using an individual classifier.

A literature review was performed to ascertain the types and extent of astronomical research performed using techniques from the fields of multivariate statistics and pattern recognition. Since the objective of my research was to develop a model independent method to classify X-ray sources, independent of a priori knowledge concerning the nature of the sources, methods that analyze one source at a time and attempt to fit X-ray spectra to a model are not included in this review of existing techniques.

A broad search was performed first, to ascertain existing knowledge and breadth of techniques in the field of astronomy in general. Also, this search was kept broad in part to examine:

- Preprocessing required for astronomical data
- Types of attributes that have been selected to classify astronomical objects

- Classification accuracy of various methods for astronomical data

The results of this broad search are presented in section 4.2.

Next, the search was narrowed to focus on research specific to X-ray astronomy. An overview of the relevant research is in section 4.3.

4.2 Application to Astronomy

Statistical clustering and pattern recognition techniques have been used in a variety of areas of astronomy. What follows is not an exhaustive list, but a sampling of the techniques and methods used for various astronomical applications.

Until the early 1980's, galaxy shapes were classified by visual examination[37]. Recently, pattern recognition has been used to automatically classify galaxies into spiral, elliptical, and irregular classes. Burda and Feitzinger[38] used data from the atlas of HII regions in spiral galaxies[39] as input for their classification technique. Preprocessing involved centering the images and normalizing all objects in size and inclination. A relaxed form of the opening and closing morphological operations was used to filter the grayscale density distribution structure of each galaxy to be classified. Five classification parameters, including galaxy inclination and size of the bulge, were extracted from the filtered density distributions. These parameters are dependent upon galaxy morphological type. The mathematical form of the spiral was used for pattern matching. The authors were able to correctly classify 21 out of 24 objects. However, they concluded that this was a poor method of classification for the given data set, because the majority of the galaxies in the input data set have very few HII regions. Another technique[40] used data created by digitally scanning over 50 pictures from The Hubble Atlas of Galaxies[41]. A statistical spatial thresholding method for initial segmentation of the image was applied. The median filter was used to remove salt and pepper noise. A smoothing process was then performed on the boundaries between the segmented regions. In the smoothing process, the input gray-level image and the segmented image were modeled as realizations of Markov Random Fields. The posterior distribution was calculated using Bayes rule. The maximum of the posterior distribution was considered the final segmentation. The following parameters were measured from the final segmented image: a scale-invariant measure of compactness of the closed shape, the distance between the boundary of the segmented region and a fitted elliptical model, and curvature values calculated on each point on the boundary. Using these parameters, spiral and elliptical galaxies were successfully classified. Bazell and Aha[36] tested a Naive Bayes classifier, a backpropagation neural network, and a decision-tree induction algorithm on a sample of 800 galaxies. They started with 22 features of the galaxies, including area, radius of the bulge, peak brightness, and entropy. After examining the correlation matrix of the features, 8 features were eliminated due to significant correlation with other features. The neural network was a fully connected network consisting of 14 input nodes, 10 hidden nodes, and 2 to 6 output nodes corresponding to the number of output classes. An interesting part of their experiment involved the use of an ensemble of classifiers. An ensemble of classifiers is created by using bootstrap replicates of the training set. The predictions of the classifiers in the ensemble are then combined to determine a final class prediction. Bazell and Aha determined that an ensemble approach, as opposed to an individual approach, greatly improved the results for the decision-tree and neural network methods when classifying galaxy morphology. Overall, they concluded that their technique decreased classification error, with improvement as the number of output classes is decreased.

Pattern recognition and neural networks have been used in astrophysical studies of the Sun to predict solar flares[42]. A combination of datasets was used, all of which were acquired at a single site and under the same observing conditions. The datasets included full-disk white light images with high precision of position determination, full-disk Ha images, full-disk magnetograms, full-disk Doppler velocity fields, and full-disk filtergrams. They included a pre-processing step to remove effects caused by non-uniform illumination, and to remove the center-to-limb variation from the solar full-disk images. Another example of replacing human classification with computer-based classification is shown in a study performed using both a supervised and an unsupervised method to classify the neutral hydrogen distribution in 21 cm spectral line images[43]. The supervised method involved cross-correlation of the observed HI distribution with a template that represented the projected supershell model. A noise-corrected estimator of the normalized correlation coefficient was used to measure the quality of the match. The unsupervised method used a dissimilarity measure based on the brightness temperature distribution of the feature. After calculating the dissimilarity for all pairs of features, clustering of the dissimilarity matrix was performed.

Computerized classification techniques have also been used to classify variable stars. Eyer and Blake[44] developed a classification method for periodic variable stars. First, a Fourier decomposition of the light curves was found. Four light curve parameters were then chosen: period, amplitude, skewness, and an amplitude ratio. The parameters were fed into a Bayesian classifier called AutoClass[45]. They applied this algorithm to a subsample of 458 stars from the All Sky Automated Survey (ASAS). They obtained a classification error rate of about 5% for their sample.

Wozniak et al.[34] developed several supervised and unsupervised methods to automatically classify 1781 variable stars. Their input data set consisted of light curves from 5.6% of the total Robotic Optical Transient Search Experiment sky coverage. The variable stars were manually divided into nine classes. Some of the light curve features used include period, amplitude, ratios formed from the amplitudes of the first three Fourier components, and the sign of the largest deviation from the mean. The authors emphasized that the asymmetry of the magnitude distribution must be represented in the feature set chosen. The supervised method, Support Vector Machines, outperformed the unsupervised methods of K-means and AutoClass. The best classification accuracy rate achieved was 90% for the supervised method and 75% for the unsupervised method. However, the authors point out some advantages of using unsupervised methods. The classes with the highest confusion were the Mira variable stars and the long period variable stars. The classification was rerun after reducing the number of classes from nine to four and better results were obtained.

Buccheri et al.[46] presented a self-adaptive clustering method to detect microstructures in the light curves of gamma-ray pulsars. They claim that their method works for low counting statistics in the high-energy range, as well as high counting statistics in the low energy range. The method is based on the single linkage clustering algorithm. The input into the algorithm consists of the residual phases corresponding to the arrival times of the selected gamma-ray photons after sorting in ascending order. The specific dataset they used contains the Crab and Vela pulsars. The dataset was collected by a European Space Agency satellite. The authors obtained very good results without using any a priori information or binning.

Spectra of stars have been classified with methods developed by Heck et al.[47], Bailer-Jones[48], and Vieira and Ponz[49]. Heck et al. argued that the best strategy is to apply multiple methods to the same data set and then compare the results. They used three cluster analysis methods (K-means clustering, single linkage clustering, and modified complete linkage clustering) on stellar data from the Hauck and Lindemann photometric catalogue[50]. Principal component analysis was used with the Euclidean clustering method. Input to each classifier consisted of numerical values of photometric indices from 2849 stars. Overall, they obtained good agreement between the three clustering methods although the misclassified stars were not the same for each method. Due to their results, the authors recommended that either the spectral type or the photometric indices of 249 of the stars in the catalogue should be re-determined.

Bailer-Jones[48] used an artificial neural network (ANN) to automate MK spectral classification. The input data set was taken from the Michigan Spectral Survey[50] and included over 5000 spectra in the wavelength range of 3800 Å to 5200 Å. The ANN was trained on synthetic spectra and then applied to observed spectra to determine the spectral classification, effective temperature, and other physical parameters of the stars. Principal component analysis was used to reduce the dimensionality of the stellar spectra. The reproducibility of neural network classifications was shown with high accuracy for the dwarf and giant classes.

Vieira and Ponz[49] explored two automated classification methods: an ANN and a Self-Organized Map. Their input set consisted of low-dispersion spectra of normal stars with spectral types ranging from O3 to G5. All spectra were corrected for interstellar extinction prior to classification. Sixty-four stars were used for training. Very low error rates were achieved by both methods.

4.3 Application to Astronomical X-ray Data

Automated pattern recognition and classification methods have been successfully implemented for classification of X-ray spectra in certain contexts. Yin et al.[51] applied pattern recognition techniques to spectra obtained by an X-ray spectrometer developed for the Mars rover. The X-ray fluorescence pulse-height spectrum was represented by an n-dimensional vector, where n is the number of channels. The authors used a normalized correlation coefficient (NCC) based on the angle between two n-dimensional vectors: one vector representing the spectra of the sample and the other representing the spectra from a chemical composition table. The value of the NCC is close to one for two spectra with similar structures. All the spectra were attenuated to reduce the magnitude of overly prominent components. They demonstrated that applying their techniques to the raw spectra provided the same discrimination among samples collected by the Mars rover as knowledge of the sample's actual chemical composition. An interesting test that the authors' performed involved re-running the experiment with fewer counts per sample. They tried decreasing the number of counts per sample by two orders of magnitude (from 1,200,000 to 12,000) and still obtained a very high rate of accuracy (97%).

Pattern recognition has been used on active regions of the sun to forecast solar flares[52]. Solar flares were separated into two classes, hazardous and non-hazardous, using radiation in the X-ray range of the active regions of the Sun. Maximum intensity of the X-ray burst and time of the flare's decline were used as parameters for the Topol and Sigma algorithms. A classification accuracy of over 80% was obtained.

Finally, pioneering work by Collura et al.[53] successfully demonstrated a model-independent method to group X-ray sources detected with the Einstein Observatory Imaging Proportional Counter (IPC). Einstein was operational from 1978 thru 1981. The IPC provided full focal plane coverage but only moderate spatial and spectral resolution. The IPC had an FOV of 75 arcmin by 75 arcmin with a spatial resolution of ~1 arcmin, compared to *Chandra*’s ACIS FOV of 16 arcmin by 16 arcmin and a spatial resolution of less than 1 arcsec. The IPC covered an energy range of 0.4 keV to 4 keV, whereas the ACIS energy range is from to 0.2 keV to 10 keV.

Much like the X-ray source clustering method described in Chapter 8 which I developed independently, their technique uses multivariate statistical techniques, including principal component analysis and hierarchical clustering. The authors limited their X-ray data to sources whose X-ray spectra contained more than 50 net counts and those that could be identified with high Galactic latitude entries in one of four catalogs. As a result, their input data did not contain any young stars or A stars. Their results showed that the IPC had sufficient spectral resolution to distinguish between stellar sources and extragalactic sources. In comparison, my research involves the much higher spatial and spectral resolution data currently being produced by *Chandra*.

4.4 X-ray Data Challenges

Observations of some very weak X-ray sources may yield only a few counts per detector element. The photons detected generate an image in which the faint X-ray object appears as a cluster of events embedded in the cosmic background. Since low count X-ray data is not typically normally distributed, classical multivariate methods that require multivariate normal data cannot be used for the analysis of low count X-ray sources. Also, traditional multivariate techniques often assume that the relationships between variables are linear. However, astronomical variables may have nonlinear relationships, such as logarithmic, exponential, or power law[54]. Non-normal data may be made more "normal looking" by performing a transformation of the data, such as a logarithmic or square-root transformation. Normal-theory analyses are then carried out on the transformed data. It has been theoretically shown that count data can often be made more normal by taking the square root of the counts[55]. Therefore, if techniques that assume normality of the data are to be used on non-normal data, a transformation of the data to near normality is often indicated.

Chapter 5

Relevant Mathematical Techniques

Multivariate statistical methods provide a simultaneous analysis of relationships among a set of *p* random variables. These variables consist of measurements taken across a sample of *n* observations, such as people or objects. Multivariate techniques can be used for exploratory analysis to search the relationships among the variables for patterns that are not attributable to chance.

Cluster analysis is a multivariate statistical technique that compares and groups the *n* observations based on the set of *p* variables. Cluster analysis works best when the objects to be grouped have distinct measurable characteristics that are reflected directly in the *p* variables. The *p* variables must be relevant to the classification desired. This makes the definition of the set of variables critical to the success of the clustering[33].

Many clustering algorithms exist and no specific algorithm is generally considered to be the "best". Different algorithms may produce different results for the same set of input data[56]. In addition, the results obtained by most clustering algorithms are sensitive to outliers, because sources of error or variation are not formally considered[57].

Clusters can only be based on the variables that are given in the data. The clusters obtained may be rather sensitive to the particular choice of variables that is made. A different choice of variables, apparently equally reasonable, may result in different clusters.

Three multivariate techniques were used in my algorithm. The first technique, Principal Component Analysis (PCA), is described in section 5.1. Two clustering methods were used: agglomerative hierarchical clustering, described in section 5.2, and a non-hierarchical technique called K-means, described in section 5.3. The clustering algorithms were used to find groups of X-ray sources with similar spectra and to separate out X-ray sources with unusual spectra. In the context of my research, the *n* observations are the detected X-ray sources. The *p* input variables correspond to X-ray spectral bandpasses, which are described in detail in Chapter 6.

5.1 Principal Component Analysis

PCA is a classical multivariate statistical technique that originated in 1901 when Pearson developed the method as a means of fitting planes by orthogonal least squares[58]. It may be used to[58,59,60,61]:

- Transform a number of correlated input variables into uncorrelated ones
- Find linear combinations that result in relatively large variability
- Reduce the size of the dataset for subsequent analyses
- Identify groups of variables that vary together and possibly uncover hidden relationships in the data

Standardizing the variables entails subtracting the mean of the variable (computed across all observations) from the variable, then dividing the resulting value by the standard deviation of the variable (again, computed across all observations). Input variables should be standardized if they are measured on widely differing scales or if the units of measurement are not commensurate. Standardization will minimize differences between existing groups, because if groups are separated well by variable p_i, then the variance of p_i will be large, however, that is desired. The equivalent of standardization can be accomplished by using the correlation matrix as opposed to the covariance matrix in PCA.

PCA can be described algebraically through the data's covariance or correlation matrices, or geometrically via clouds of data points in k-dimensional space[62]. Geometrically speaking, if two or more variables are correlated, the cloud of data points will be most elongated along some direction in this k-dimensional space. PCA removes the correlation between the input variables by rotating the data axes so that the cloud of data points is most elongated along a new axis: the axis of maximum variance of the data[63]. The method of minimization of the sums of squares of the deviations is used to determine the new axis of maximum variance and accomplish this rotation. This occurs subject to the constraint that the new axes are orthogonal. The resulting axis of maximum variance represents the first principal component. This process is repeated to define each subsequent component, in order of decreasing variance. The principal components are then the new random variables specified by the axes of each rigid rotation of the original system of coordinates, and correspond to the successive directions of maximum variance of the cloud of data points. The principal components give the positions of the objects in the new system of coordinates.

PCA generates *p* eigenvalues and eigenvectors from the covariance or correlation matrix. The eigenvalues are the variance explained by each of the principal components. The eigenvectors are linear combinations of the original input variables. They determine the directions of maximum variability and can be interpreted as measuring the importance of the corresponding variable to each principal component. PCA depends solely on the covariance or correlation matrix, not on multivariate normal data[64]. Typically, researchers attempt to assign application specific significance and meaning to the principal components resulting from PCA, but the components are not always interpretable[65].

Although *p* components are required to reproduce the total variability within the dataset, a relatively smaller number of principal components, *k*, may adequately represent most of the original variance. PCA may then be used for data reduction by retaining only those *k* principal components, resulting in a simplified description of the dataset.

PCA has some disadvantages in the context of astronomical problems. First, it can only uncover linear relationships between the input variables. Astronomical variables may have nonlinear relationships, in which case the variables will appear uncorrelated. Second, since PCA is scale dependent, it is sensitive to outliers[57].

PCA was used in the proof of concept algorithm (see Chapter 7) and the final X-ray source classification algorithm (see Chapter 8).

5.2 Agglomerative Hierarchical Clustering

The objective of the agglomerative hierarchical clustering algorithm is to uncover natural groupings of the *n* observations. This method does not assume multivariate normality of the data.

It begins with each of the observations (i.e., sources) as its “own statistical cluster” and the statistical distance (or statistical similarity) between each individual observation and all other individual observations is calculated. In the first step, the closest two (i.e., most similar) observations are joined. In the next step, either a third observation joins the two that were joined in the first step or two other observations are joined together. Close groups (i.e., similar groups) are successively merged in this hierarchical or “nested” fashion, based on the statistical distance (or similarity) measure between each pair of clusters. Cluster merging continues until there is only one large cluster containing all the sources. At this point, the pattern of how the distance (or similarity) values change from step to step is manually examined to find a large jump in the metric value between amalgamations. This identifies the number of clusters in the final partition, if the grouping seems logical for the dataset at hand.

Selection of the final partition can also be accomplished visually by use of a 2-D tree diagram called a dendrogram, which shows the cluster mergers at each step (see Figure 5.1). The distance values for each of the intermediate clustering steps are examined for large gaps to determine the final number of clusters. The dendrogram is then “cut” at the desired distance (or similarity) level to specify the final grouping of observations. Domain knowledge is typically used when determining the final number of clusters. This final partition is the grouping of observations which will, ideally, identify groups whose members share common characteristics.

There are many different metrics that can be used for the statistical measure. For example, measures of distance (dissimilarity) such as Euclidean, Minkowski, Canberra, and Czekanowski, or measures of similarity such as correlation coefficients can be used[66].

One disadvantage of hierarchical clustering is that the selection of the final number of classes (i.e., the location at which to cut the dendrogram) is somewhat heuristic. There is no mathematical basis for choosing a final distance (similarity) level. A second disadvantage of this clustering method is that it cannot transfer an observation (i.e., a source) from one cluster to another if it was grouped incorrectly in an earlier step[57].

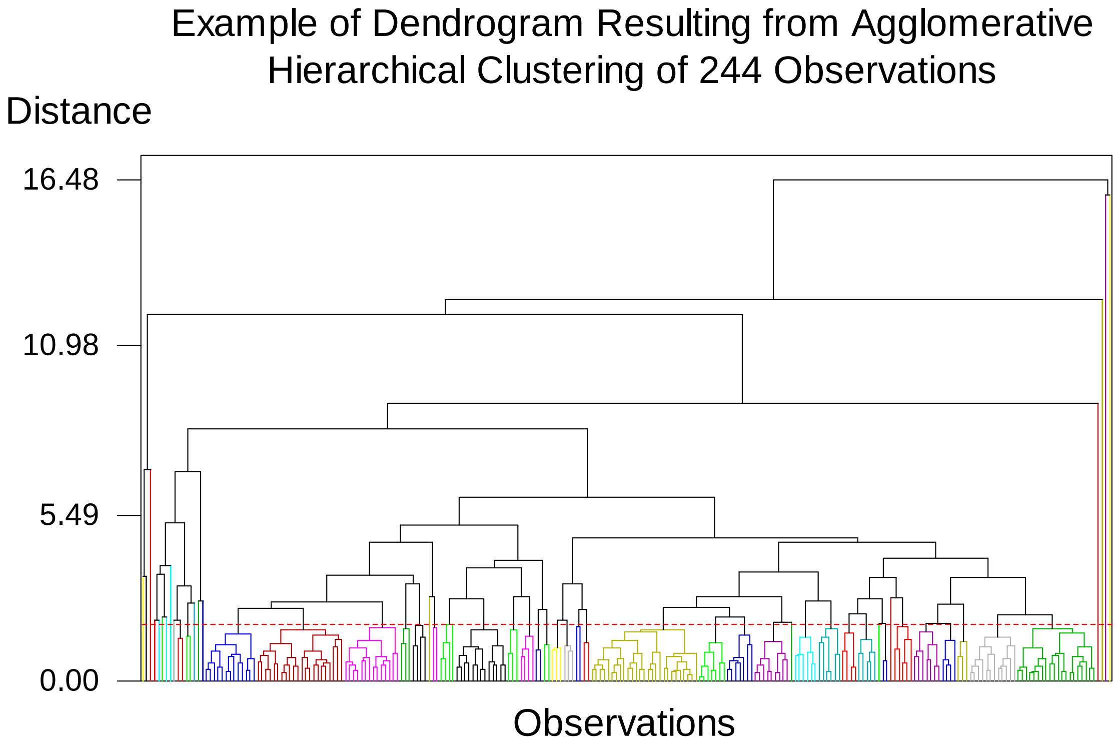


Figure 5.1: Example of a dendrogram. The dashed horizontal red line shows where the dendrogram has been cut at a distance level of approximately 2 units.

Agglomerative hierarchical clustering was used in the proof of concept algorithm (see Chapter 7) and the final source classification algorithm (see Chapter 8).

5.3 K-Means Clustering

K-means is an iterative, non-hierarchical clustering method that groups observations into a collection of K clusters. It begins by partitioning the sources into K clusters, where K is an input to the algorithm and, therefore, must be identified in advance of running K-means. For my algorithm, I used the agglomerative hierarchical clustering algorithm to obtain a value of K to feed into the K-means algorithm. The clusters obtained by the hierarchical clustering algorithm were used to seed the K-means algorithm with an initial set of clusters.

The centroid of a cluster is the center of that cluster. It is represented by a vector containing one number for each variable, where each number is the mean of that variable for the observations in that cluster. First, the centroids (means) for each of the K clusters are then calculated. Next, each observation is examined and reassigned to the cluster with the nearest centroid where necessary, based on the distance measure (see Figure 5.2)**.** Then the centroids are recalculated for each cluster receiving a reassigned observation and also for any clusters losing observations. This is repeated until either no more reassignments take place or a specified number of iterations have been completed. At this point, each cluster contains statistically similar sources, based on the multivariate features passed to the algorithm.

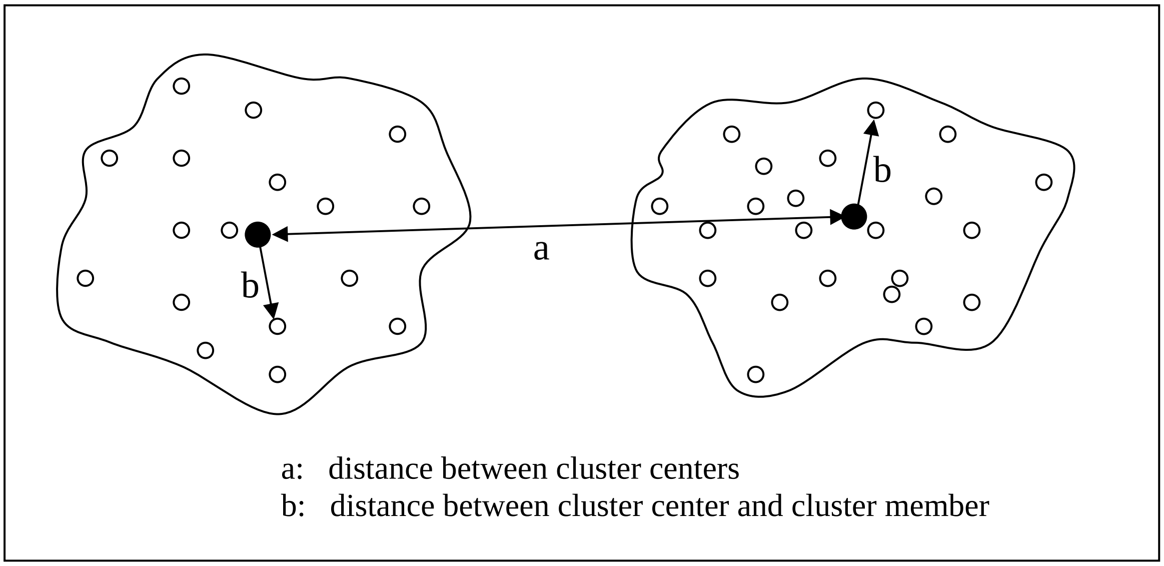

b
a
b
a: distance between cluster centers
b: distance between cluster center and cluster member

Figure 5.2: 2-D schematic showing between-cluster distance and within-cluster distance. The clusters may exist in greater than 2-dimensional space.

The final assignment of observations to clusters is, to some extent, dependent on the initial clusters passed to the algorithm. Most major changes in cluster assignments happen during the first reallocation step[67].

K-means clustering was used in the proof of concept algorithm (see Chapter 7) and the final X-ray source classification algorithm (see Chapter 8).

Chapter 6

Input Variable Selection

6.1 Background

Input variables were chosen that could be used to distinguish the X-ray sources, keeping in mind that a priori information about the type or nature of the X-ray sources could not be used. The projected spatial location (x and y) of each point source is known. However, the distance to the source (z) is not well-determined. Two sources that are close in x and y may be far apart in z, and won't necessarily have the same intrinsic nature. For example, for Orion, analysis indicates that ~10% of COUP sources (~159) are "background" (extragalactic) point sources[68]. Consequently, it can be difficult to draw conclusions about source similarity based solely on spatial proximity or density.

Therefore, the variables chosen had to be based on the raw photon count data. Temporal data was not used at this time.

6.2 X-ray Emission Lines

Emission lines are narrow features in the spectral distribution that are caused when electrons make a transition from one allowed energy state to the next, each one emitting energy in the form of a photon in the process. The photon carries exactly the amount of energy set free by dropping to a lower allowed energy state. Emission lines are typically modeled with a Gaussian distribution, Lorentzian distribution, or delta function[26]. X-ray spectra display emission lines if the spectra are of sufficiently high resolution and/or the emission is strong.

Spectral analysis of individual X-ray sources is performed to determine temperatures of the sources and elemental abundances in the source. Elements with strong enough lines to be observed with current technology are oxygen (O), neon (Ne), magnesium (Mg), silicon (Si), sulphur (S), argon (Ar), calcium (Ca), iron (Fe), and nickel (Ni) [9,69,71]. Mg and Ne are in the energy range dominated by the Fe L-shell lines. With current X-ray spectrometer resolutions, these elements are difficult to measure independently of the Fe L-shell lines. The Fe K-shell lines have larger transition probability widths and are in a more isolated part of the spectrum, so Fe K-shell lines are observationally easier and more reliable to measure[69] .

The X-ray spectrum from 0.4 keV to 8.2 keV was divided into non-overlapping bands. The locations of the bands had to be chosen to emphasize regions of the X-ray spectrum that are important in distinguishing young stars. For this reason, I looked at emission lines from measurements of the spectra of young stars. Some important emission line features are:

- Ne line at ~0.92 and ~1.02 keV
- Si line at 1.7 keV
- Fe Kα emission line at ~6.4 keV; an integral part of AGN phenomenology [70]
- Fe XXV (24 times ionized Fe) at ~6.7 keV; iron atom that has lost 24 of its 26 electrons

Figure 6.1 shows an example of the regions of the X-ray spectrum of a young star called TW Hya[3]. This particular spectrum is weak in iron.

A variety of techniques was used to divide the X-ray spectrum from 0.4 keV to 8.2 keV into bands. I tried three different techniques for selecting the bands. The width of the bands and the number of bands were varied for each technique. Each method and its resulting bands are described in the following sections. The resulting X-ray spectral bands became the input variables for the classification algorithm. The algorithm was run on the number of counts in each spectral band for the source dataset.

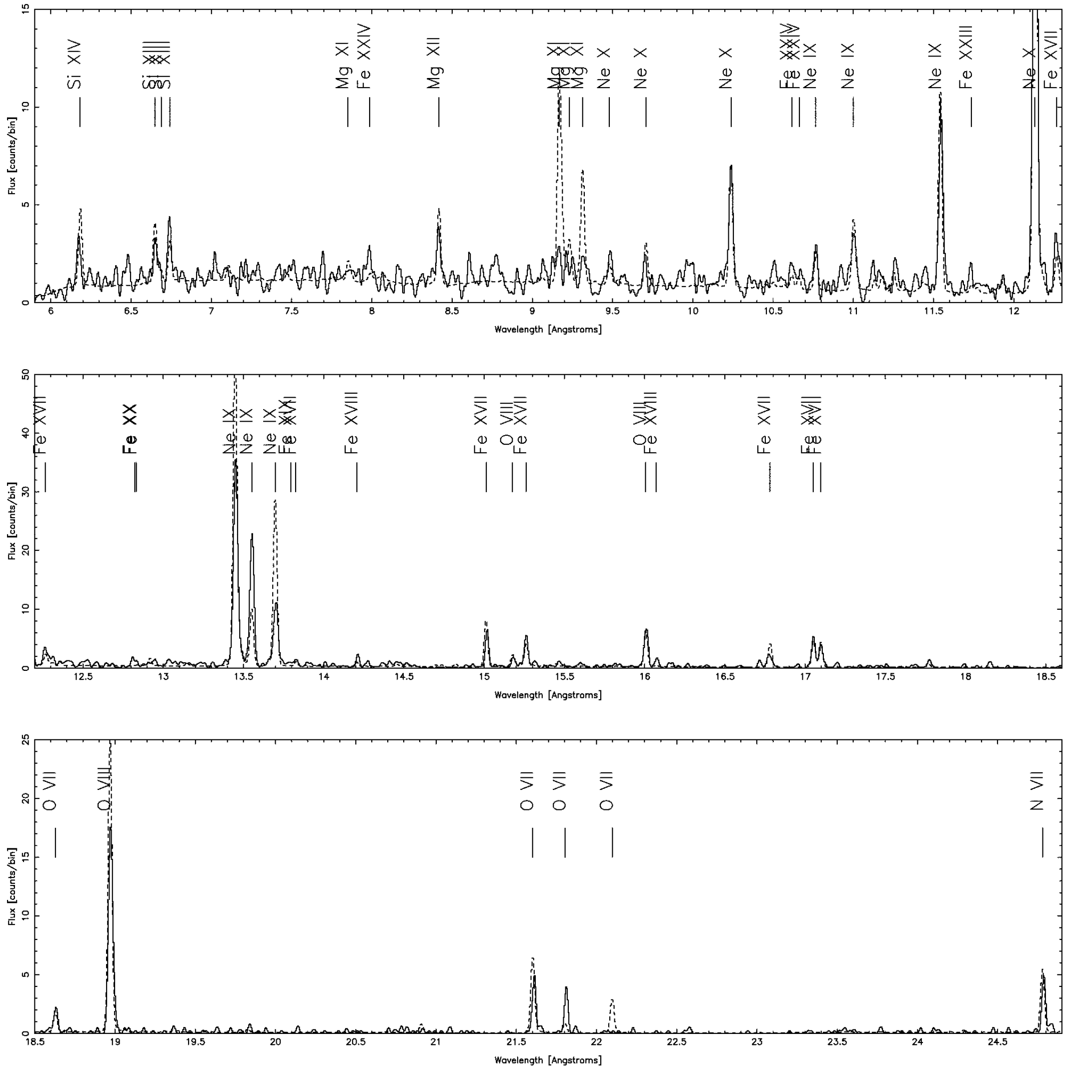


Figure 6.1: Selected regions of the X-ray spectrum of TW Hya (solid curve). The observed spectrum is overlaid with an emission measure model (dashed curve) that best fits temperature-sensitive line intensities[3].

6.3 Equal-Width Bands

The width of spectral bands in the soft X-ray region was set to 500 eV. The width of spectral bands in the hard X-ray region was set to 1000 eV. Ranges were defined as shown in Table 6.1. All of the bands were sufficiently wide enough to avoid correlation between bands, due to the energy resolution of the ACIS-I CCDs (see Chapter 3).

This method did not yield good results. The classes were not homogeneous. Sources with dissimilar spectra were placed in the same groups (see Figure 6.2). Many X-ray emission features are grouped together in one band. For example, using this definition of X-ray spectral bandpasses, the algorithm could not distinguish between a source that had a high abundance of Ne X at ~1211 eV versus a source that had a high abundance of Mg XII at ~1472 eV because the photon counts for these two features would both be summed within band number 3.

Table 6.1: Spectral Ranges for Equal Width Bands

Band Number	Range [eV]
1	0 – 500
2	501 – 1000
3	1001 – 1500
4	1501 – 2000
5	2001 – 3000
6	3001 – 4000
7	4001 – 5000
8	5001 – 6000
9	6001 – 7000
10	7001 – 8000

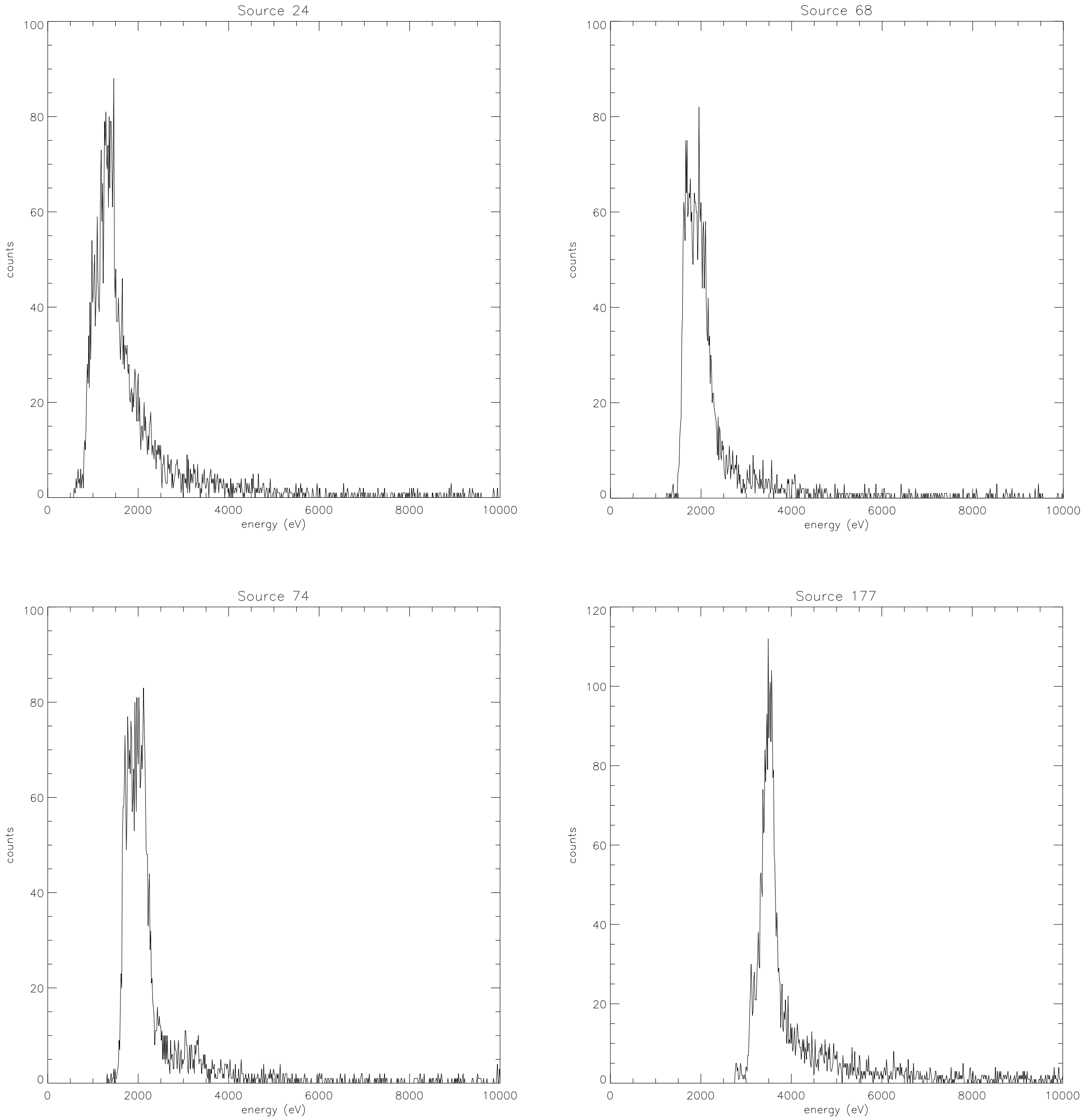


Figure 6.2: Four sources grouped into the same class when using equal-width spectral bands.

6.4 Equal Area-Under-the-Curve Bands

The spectra of all the sources in the sample were averaged to create a mean spectrum for the subset of X-ray sources. The total area under the curve was then calculated to compute the mean spectrum. This value is used to divide the spectrum into eight spectral bands, each with approximately equal area. Note that the number of spectral bands was selected arbitrarily. More or fewer bands could have been chosen. The proof of concept algorithm makes use of this technique, and it is described in more detail in Chapter 7. A disadvantage of this technique is that it is source-spectrum dependent.

6.5 Hyperspectral Bands

Multispectral systems have up to ten or twenty, non-contiguous spectral bands. Typically, each discrete band covers a wide range of energies. Hyperspectral systems have tens to hundreds of narrow, contiguous spectral bands. Spectral resolution can be defined as the smallest interval of bandwidth that can be detected in the spectrum, measured as the full-width at half of the maximum energy peak height. Multispectral systems therefore have low spectral resolution relative to hyperspectral systems.

With the expectation that most sources of interest to this work (i.e., young stars) will exhibit emission line spectra characteristic of ionized plasma, the X-ray spectrum from 0.4 keV to 8.2 keV was divided into a number of spectral bands which were chosen based on high-resolution X-ray emission line measurements from well-characterized X-ray sources[3,71]. While it is not possible to isolate every significant feature due to resolution of the device and other hardware constraints, the spectral bands were chosen to include strong lines of high-ionization species such as O VIII, Ne IX, Ne X, Mg XII, Si XIV, and Fe XXIV.

The spectral resolution of the ACIS CCDs at the nominal operating temperature of *Chandra* (-120C) was used in determining the width of the bands. The full-width half-maximum (FWHM) of the FI detectors increases with increasing energy (see Figure 3.6), so the spectral bands increase in width accordingly.

Analysis of the results of CTI correction was also used in determining the width of the spectral bands. After launch and orbital activation of *Chandra*, low energy protons that were encountered during radiation belt passages reflected off the telescope and onto the focal plane. This caused some damage to the FI detectors and increased their CTI. The ACIS instrument team developed a CTI correction algorithm to improve the spectral resolution of the FI CCDs at all energies. This algorithm was run on the *Chandra* observations used in my research.

Finally, absorption features apparent in the quantum efficiency curves of the ACIS-I CCDs (see Figure 3.7) were also considered when selecting ranges for the spectral bands. These edges were avoided when defining the bands.

The final band definitions were made by combining these hardware-imposed band constraints with knowledge of the locations of X-ray features that were considered important. Band locations and widths were selected for a set of 42 bands (Appendix A). Edges of the bands were chosen partially to avoid a feature considered important where possible, and partially to stay within the bounds of the hardware constraints. The set of 42 bands was used for the final X-ray source classification algorithm (Chapter 8).

Chapter 7

Proof of Concept

The literature review revealed that pattern recognition and multivariate statistical techniques had not been applied to X-ray observations of young stellar clusters for the purpose of clustering and classification. To test the feasibility of this idea, I developed a "proof of concept", which consisted of a preliminary algorithm, a sample dataset, and a set of input X-ray spectral band definitions.

7.1 *Chandra* Archival Observation

The *Chandra* X-ray Center (CXC) Automated Processing system generates several hundred data products derived from *Chandra* telemetry. Standard data processing is used for ACIS-I observations. Archival ACIS-I imaging of the well-studied Trapezium region of the ONC (*Chandra* Observation Id (ObsId) 1522) was used for developing the proof of concept algorithm.

7.1.1 Preprocessing

The first step was to prepare the X-ray observation dataset. For *Chandra* archive data, this involves performing pre-processing to “clean” the dataset. The initial dataset consisted of a Level 1 processed event list provided by the pipeline processing at the CXC (see Chapter 3). Additional data processing was performed as described in Ref. 5. This included:

- Astrometry correction for data aspect offsets up to 2” due to uncertainties in boresight calibration at time of processing (this is necessary for data in the *Chandra* archive)
- Application of charge transfer inefficiency (CTI) correction
- Removal of spurious events from cosmic-ray afterglows and “hot columns”

The spurious events are false events caused by flickering pixels on the CCD detectors and cosmic ray hits in the frame store area[5].

Some of the pre-processing procedures introduce a non-linear effect across the dataset. This changes the structure of the data that is used for subsequent X-ray source detection.

7.1.2 Source Detection

After pre-processing, X-ray source detection was performed on ACIS-I FI chips 0, 1, 2, and 3 to locate X-ray sources in the *Chandra* dataset. A standard, automated X-ray source detection program called WAVDETECT[72] was used. WAVDETECT is a wavelet transform source detection program that is part of the *Chandra* Interactive Analysis of Observations (CIAO)[c] software package. The user must provide a background map or use the built-in iterative background determination option.

The first step in the process is to create region files for each of the four ACIS-I front-illuminated (FI) chips. Each region file is created with a text editor and contains the *rotbox* command with the sky coordinates of the center of the chip, the x and y extent of the chip, and the roll angle. The center of the chip in sky coordinates can be obtained by running the *dmcoords* script. The x and y extent of each chip is 1024. The roll angle can be obtained by running the *dmkeypar* script with the parameter ROLL_NOM. The regions file for chip 0, "chip0.reg", is shown below.

```
# Region file format: CIAO version 1.0

rotbox(4730.10,  3603.19,  1024,  1024,  263.485)
```

[c] http://cxc.harvard.edu/ciao/

After creating the region files, I used the *dmcopy* command for each file to:

- create an image for each individual chip (see Figure 7.1)
- filter the observation event file to include only photon events with energies in the range 0.3 keV to 10 keV (any events outside this range are particle events)
- bin the data by two to obtain a better signal-to-noise ratio and also to ensure the file size would be small enough to run with WAVDETECT

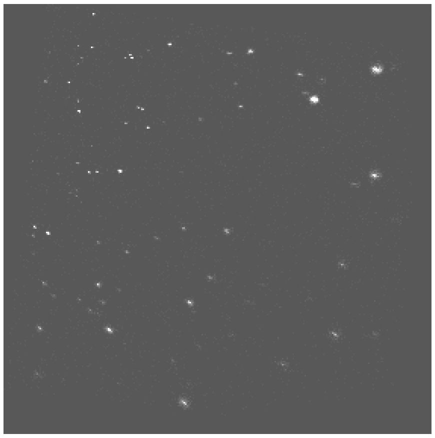


Figure 7.1: Image created from ACIS-I chip 0.

To create the exposure map for each chip, the peak energy, the sky grid coordinates, the aspect histogram, and the instrument map are needed for each of the chips. To compute the peak energy for the chip, the brightest source on the chip was identified and *dmextract* was used to extract that source's histogram. Next, I used *dmstat* to determine the maximum count rate from the histogram, followed by running *dmlist* to determine the energy at which that maximum count rate occurred. The sky grid coordinates are needed so that the exposure map that is created is the same size as the image created from the event list. To compute the sky grid coordinates, I used the *get_sky_limits* program for each chip. The next step was to create the aspect histogram file. The aspect, or aspect solution, is the pointing position of the Chandra telescope versus time. Star positions from astrometric surveys are used to put the aspect solution onto a reference frame. The aspect histogram is a binned histogram for the chip, detailing the aspect history of the observation. It gives the amount of time the Chandra optical axis dwelled on each part of the sky. The *asphist* script is used to create the aspect histogram, using parameter files from the ObsId 1522 distribution, including the aspect solution file.

Next, the instrument map was created for each chip. It is in detector coordinates, must describe the chip with full resolution, and provides the instantaneous effective area for the chip. It is basically the mirror effective area projected onto the detector surface and includes detector quantum efficiency, bad pixels, non-uniformities across the face of the detector, and mirror vignetting. The *mkinstmap* script was used, which requires at a minimum, the detector number, the pixel grid, and the peak energy. The instrument map (Figure 7.2) was required in the subsequent step to make the exposure map.

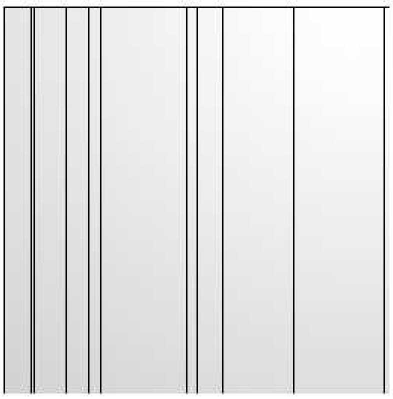


Figure 7.2: Instrument map for ACIS-I chip 0.

Finally, I created an exposure map for the observation by using the *mkexpmap* script, the sky grid coordinates, and the aspect information stored in the histogram to project the instrument map onto the sky. The exposure map is then the product of the aspect histogram and the instrument map. This exposure map (Figure 7.3) is used by WAVDETECT for source detection.

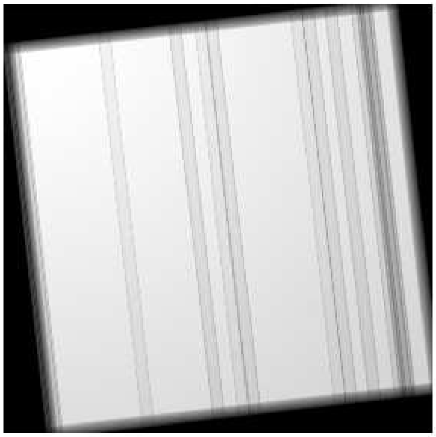


Figure 7.3: Exposure map for ACIS-I chip 0.

WAVDETECT repeats the source detection process using the Mexican Hat wavelet for a set of user-defined wavelet scales. The more scales used, the more time and memory the process can take. I worked with CFA personnel to determine optimal wavelet scales. WAVDETECT was run several times to fine-tune the selection. The scales used for the final source detection were: 2.0, 4.0, 8.0, 16.0, and 24.0 pixels.

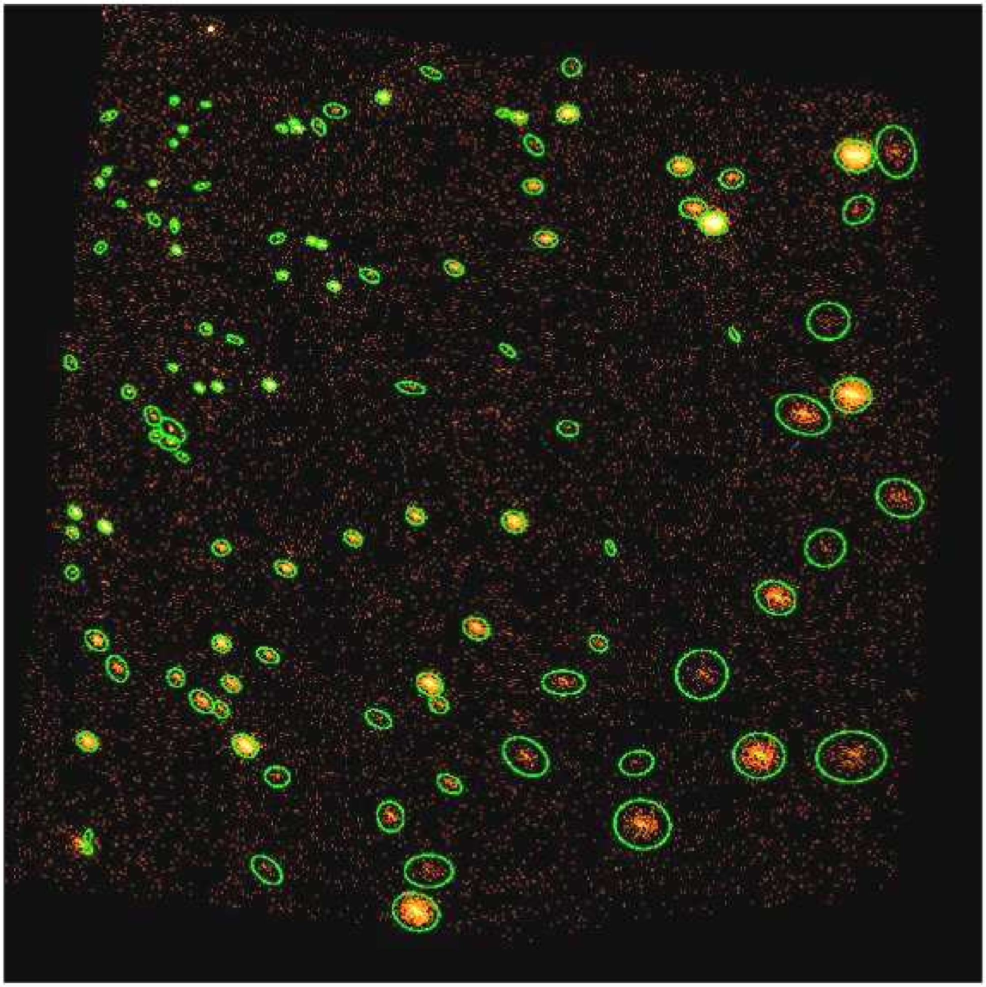


Figure 7.4: Example of detected sources for one ACIS-I chip 0 (ellipses represent 3σ).

For each source candidate, the detection with the highest correlation maximum for all of the runs was selected. WAVDETECT works well in crowded regions of sources and also in situations where there is a point source on top of an area of extended emission. WAVDETECT can also handle edge-of-field and vignetting effects. Figure 7.4 shows the results of the source detection phase for one of the ACIS CCD arrays. The ellipses encircle each detected source, with a standard deviation of 3σ.

A total of 1153 X-ray sources were detected for ObsId 1522. Detected sources with fewer than 400 total counts were eliminated, to limit the faint sources with poor photon counting statistics and to reduce the size of the dataset to a reasonable size for iterative testing of the preliminary algorithm. This resulted in 204 detected X-ray sources. These sources were sorted by number of counts and sequential numbers were assigned to each source, from brightest to faintest. A subset of the brightest sources was then eliminated due to the potential for photon pileup (sources with greater than 7600 counts). The remaining 185 detected sources (sources 20 through 204) were used in the analysis. Interactive Data Language (IDL) programs and standard CIAO tools were then used to extract the X-ray source spectra from the source detection output for each of the 185 sources. These X-ray sources were crosschecked against a table of known sources in Orion[5] and their optical and infrared attributes recorded. Figure 7.5 shows the spectra for two of the detected sources.

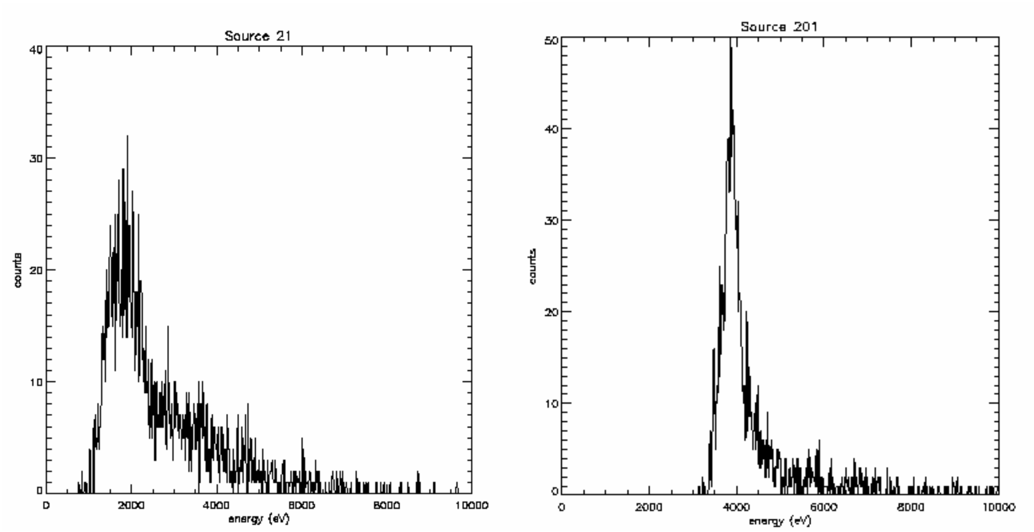


Figure 7.5: Spectra for two example sources in the testbed dataset.

7.2 X-ray Spectral Band Selection

The spectra of the 185-source test set were averaged to create a mean spectrum over all the X-ray sources (Figure 7.6). The total area under the curve was calculated for the mean spectrum. I used this value to divide the spectrum into eight spectral bands, each with approximately equal area (Figure 7.7). A multispectral approach was desired, however, the number of spectral bands selected was somewhat arbitrary. The resulting band ranges are shown in Table 7.1. The correlation matrix for the 8 bands and 185 sources was calculated using the Pearson correlation coefficient. The matrix is shown in Table 7.2.

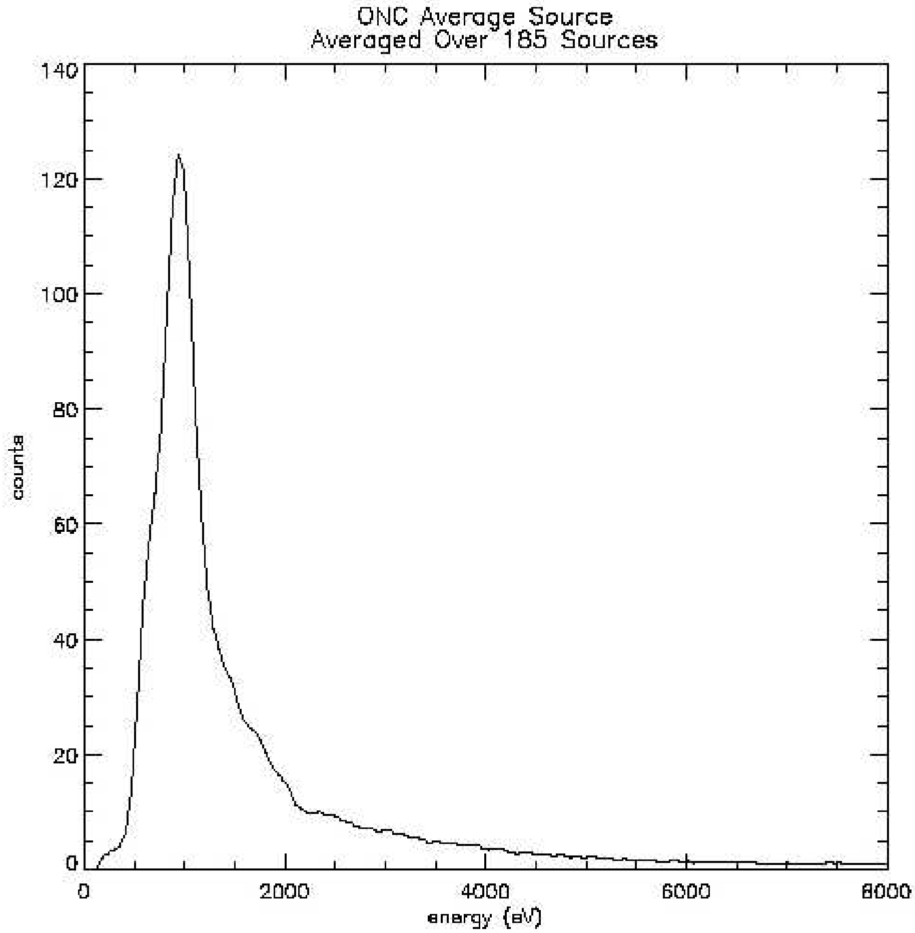


Figure 7.6: Mean X-ray spectrum created from 185 detected sources in Orion.

Table 7.1: X-ray Spectral Band Ranges

Band Number	Energy Range (eV)
1	0.00 – 759.2
2	760.2 - 934.4
3	935.4 – 1051.2
4	1052.2 - 1226.4
5	1227.4 - 1576.8
6	1577.8 - 2277.6
7	2278.6 - 4263.2
8	4264.2 – 10000.00

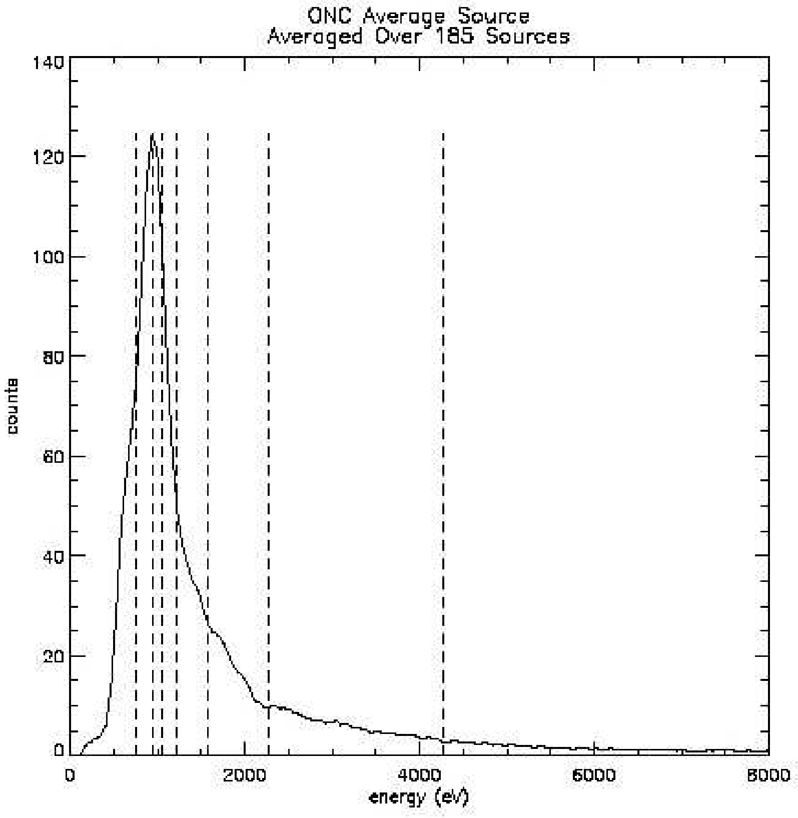


Figure 7.7: Mean source spectrum showing eight bands with equal area.

Table 7.2: Correlation Matrix for X-ray Spectral Bands

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7	Band 8
Band 1	1.000							
Band 2	0.933	1.000						
Band 3	0.862	0.973	1.000					
Band 4	0.804	0.855	0.909	1.000				
Band 5	0.476	0.538	0.580	0.744	1.000			
Band 6	0.265	0.340	0.365	0.438	0.824	1.000		
Band 7	0.157	0.208	0.218	0.223	0.476	0.819	1.000	
Band 8	0.687	0.833	0.871	0.753	0.485	0.483	0.529	1.000

It can be seen from the table that the following bands are highly correlated:

• band 1, band 2	0.933
• band 1, band 3	0.862
• band 2, band 3	0.973
• band 2, band 4	0.855
• band 3, band 4	0.909
• band 3, band 8	0.871

This strong correlation suggests the PCA would be effective in removing the redundancy in the data prior to attempting to group the sources into classes.

7.3 Principal Component Analysis[d]

For the statistical analysis, each of the eight X-ray spectral bands was considered a variable and the observations were the detected X-ray sources. I ran PCA using the correlation matrix for the X-ray spectral data. The resulting eigenvalues and eigenvectors are shown in Table 7.3. The eigenvectors determine the directions of maximum variability and can be interpreted as measuring the importance of the corresponding variable to each principal component. The eigenvalues represent the variances for each principal component.

[d] See section 5.1 for a general description of PCA.

Table 7.3: Eigenanalysis of the Correlation Matrix

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
Band 1	-0.367	0.299	-0.026	0.739	0.207	0.288	-0.172	0.273
Band 2	-0.400	0.266	-0.096	0.184	-0.382	-0.165	0.324	-0.669
Band 3	-0.406	0.239	-0.056	-0.223	-0.258	-0.540	0.074	0.601
Band 4	-0.399	0.149	0.287	-0.310	0.640	-0.213	-0.315	-0.294
Band 5	-0.335	-0.296	0.630	-0.084	-0.032	0.340	0.508	0.149
Band 6	-0.277	-0.570	0.147	0.137	-0.410	-0.095	-0.612	-0.082
Band 7	-0.211	-0.590	-0.485	0.180	0.398	-0.256	0.341	0.012
Band 8	-0.384	0.024	-0.502	-0.465	-0.087	0.603	-0.112	0.029

Eigenvalue	5.2926	1.6899	0.6246	0.2363	0.1019	0.0265	0.0245	0.0037
Proportion	0.662	0.211	0.078	0.030	0.013	0.003	0.003	0.000
Cumulative	0.662	0.873	0.951	0.980	0.993	0.996	1.000	1.000

7.3.1 Stopping Rules

The following stopping rules were used to determine the number of components to retain for further analysis: the percent of variance explained, the fair share (mineigen) criteria, and the scree plot.

Percent of Variance Explained

For this stopping rule, one retains the number of principal components required to reach a particular threshold for the amount of variance explained in the data. In the literature and in some software packages, 95% is the default threshold for cumulative variance explained[73,59]. However, there is no mathematical basis for choosing any particular fixed percentage of variance. This metric is very subjective and 95% is an arbitrary value.

If this stopping rule were used for the ObsId 1522 subset, the first three components would be retained. The first three components together explain 95.1% of the variance (see Table 7.3). The first four components together explain 98.0% of the variance. Therefore, according to this stopping rule, a sufficient amount of the data structure can be captured in three underlying dimensions. The remaining principal components account for a very small percentage of the variability and are less important.

Fair Share Criteria

The fair share is equal to the total variance divided by the number of variables, and therefore is equal to unity since the correlation matrix was used. Hence, components with eigenvalues greater than or equal to one should be retained. This method suggests that only the first two components should be retained.

Scree Plot

A scree plot[73,74] is a graph of the eigenvalues in decreasing order of variance explained. Scree is defined as a slope of loose rock debris at the base of a cliff or steep incline. Cattell (1966) named this the “scree plot” because the retained eigenvalues appear as a cliff and the deleted ones are the slope of loose rock debris at the base. An “elbow”, bend, or break in the scree plot shows the location after which the eigenvalues are relatively small and of relatively equal value. The components prior to this elbow are retained[73]. Some authors also retain the component at the location of the bend[75]. The scree plot is shown in

Figure 7.8. The elbow is at component number 3. This suggests that three components may effectively summarize the sample variability.

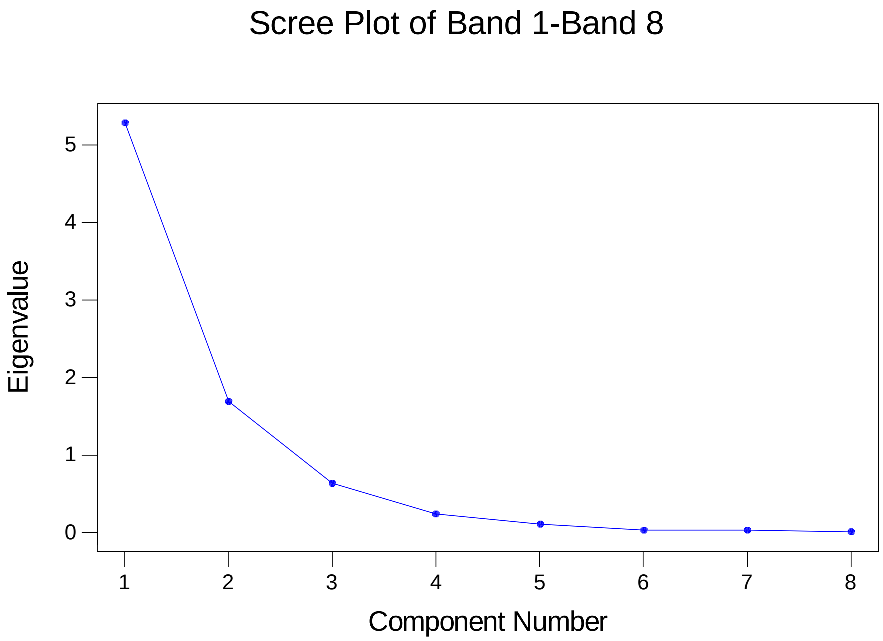


Figure 7.8: Scree plot for the eight principal components.

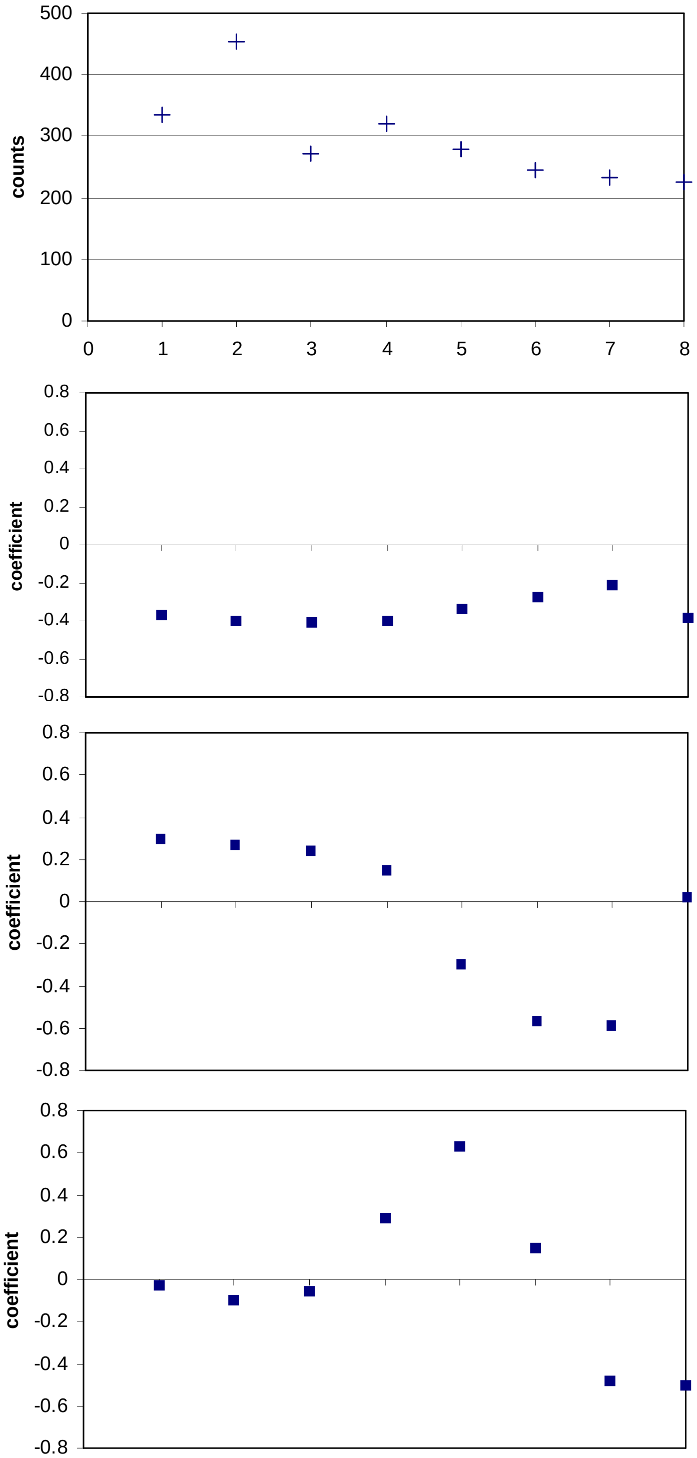


Figure 7.9: The top panel gives the average number of counts in each of the 8 bands. The bottom panels are eigenvector plots for the first three principal components.

The first three principal components were retained for the ensuing clustering steps. I chose to retain three components although one of the stopping rules suggested retaining two. It is less detrimental to retain more principal components than needed (within reason) than to delete some that represent some of the inherent variability in the dataset.

The average number of counts in each of the eight bands and the eigenvectors for the first three principal components are plotted in Figure 7.9. Principal component 1 is similar to an average of each of the spectral bands. Principal component 2 could be interpreted as an indicator of spectral hardness since soft X-rays have a positive value and hard X-rays have a negative value, with the exception of band 8 (0.02). There is no obvious interpretation for principal component 3, although it has a peak at band 5 (1.23 keV to 1.58 keV) suggesting it is related to spectral hardness.

7.4 Agglomerative Hierarchical Clustering[e]

An agglomerative hierarchical clustering method based on Euclidean distance and complete linkage was used on the first three principal components generated from the detected X- ray sources. The method started with each source as its “own cluster” and similarities between each individual source and all other individual sources were calculated. The similarity level at any step between two clusters, i and j, is the percent of

[e] See section 5.2 for a detailed description of the agglomerative hierarchical clustering method used.

the minimum distance at that step relative to the maximum inter-observation distance in the data:

$$s_{ij} = 100(1- d_{ij}) / d(max)$$

where,

d_{ij} is the Euclidean distance between cluster i and cluster j

d(max) is the maximum value in the original distance matrix

Close groups (i.e., similar groups) were successively merged based on this statistical similarity measure. Cluster merging continued until there was only one large cluster containing all the sources. At this point, the similarity level for each of the intermediate clustering steps was manually examined to find a large jump between amalgamations to estimate the number of source classes. The similarity matrix created from the clustering is shown in Appendix B.

As can be seen in the similarity matrix in Appendix B, the similarity level decreases in increments of approximately 2 or less at each step until the step between eight clusters and seven clusters, at which point it decreases by almost 8 units. This large jump indicates that eight clusters should be reasonably sufficient for a final partition of the X-ray sources. The resulting dendrogram is shown in Figure 7.10. Each resulting cluster is shown in a different color in the dendrogram. The horizontal line at a similarity level of approximately 65 illustrates where the dendrogram has been cut to obtain eight clusters.

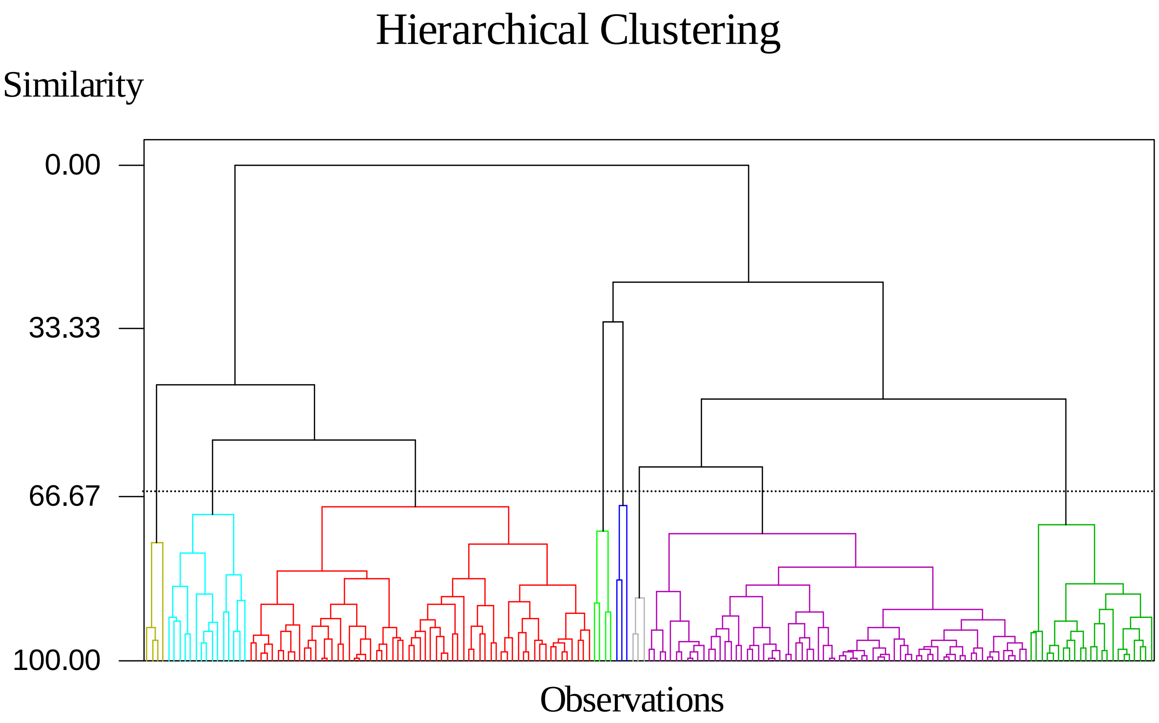


Figure 7.10: Dendrogram resulting from hierarchical clustering.

One disadvantage of hierarchical clustering is that the selection of the final number of classes (i.e., the location at which to cut the dendrogram) is somewhat heuristic. There is no mathematical basis for choosing a similarity level. A second disadvantage of hierarchical clustering is that it cannot transfer a source from one cluster to another if the source was grouped incorrectly in an earlier step. Therefore, I used the K-means algorithm to fine-tune the clusters obtained from the hierarchical clustering algorithm.

7.5 K-means Clustering[f]

K-means does not assume multivariate normality of the data. The class assignments resulting from the hierarchical clustering were used as the initial partition for the K-means clustering. Therefore, "K," the number of classes, was set to eight. Each source was examined and assigned to the cluster with the nearest centroid (mean). In some cases, this resulted in the source being reassigned to a new cluster. The centroids were then recalculated for the cluster receiving the reassigned source and for the cluster losing the source. This process was repeated until no more reassignments took place. The final number of X-ray sources in each cluster is shown in Table 7.4. The cluster numbers have no physical meaning. Appendix C shows a comparison of the cluster assignments resulting from agglomerative hierarchical clustering and K-means clustering. The K-means algorithm moved 36 of the sources (19.5%) to different clusters. The resulting clusters contained sources that were statistically similar based on the features passed to the initial clustering algorithm.

Table 7.4: Number of Sources Per Cluster

Cluster	**Number of Sources**
1	8
2	4
3	3
4	30
5	33
6	30
7	72
8	5

[f] See section 5.3 for a detailed description of the K-means clustering algorithm used.

The source spectra for all the classes were plotted. Some of the sources appeared to be outliers, based on their spectra alone. However, they were forced into one of the eight classes. Most of the classes contained at least one source that appeared as if it did not belong to that class. Class 7 contained a mixture of source spectra, i.e., it consisted of all the sources that did not fit neatly into one of the other classes.

The source spectra for the four smallest classes (1, 2, 3, and 8) are shown in Figure 7.11, Figure 7.12, Figure 7.13, and Figure 7.14, respectively. From visual inspection, it can be seen that there are strong spectral similarities within a given class. Also, strong fundamental spectral differences can be seen between classes.

7.6 Conclusions

From visual inspection of the class spectra, it was evident that the classes contained source spectra that had much stronger within class similarities than between class similarities. The algorithm isolated subtle differences between the strengths of key spectral features when grouping sources. There were also trends in the optical properties of the data. Most of the sources in classes 1 and 2 have counterparts in the optical wavelength range, while none of the sources in class 8 do[5].

Preliminary results from this proof of concept clustering on ONC sources showed promise for development of a model-independent, unsupervised method that could be used to group X-ray sources with similar spectra into classes. No a priori knowledge of the nature of each source was used to accomplish the source groupings. This algorithm was used as a baseline for development of a more sophisticated and robust X-ray source classification algorithm. To improve the results, additional work was done to better determine the number of classes necessary and to optimize the definition of the X-ray spectral bands.

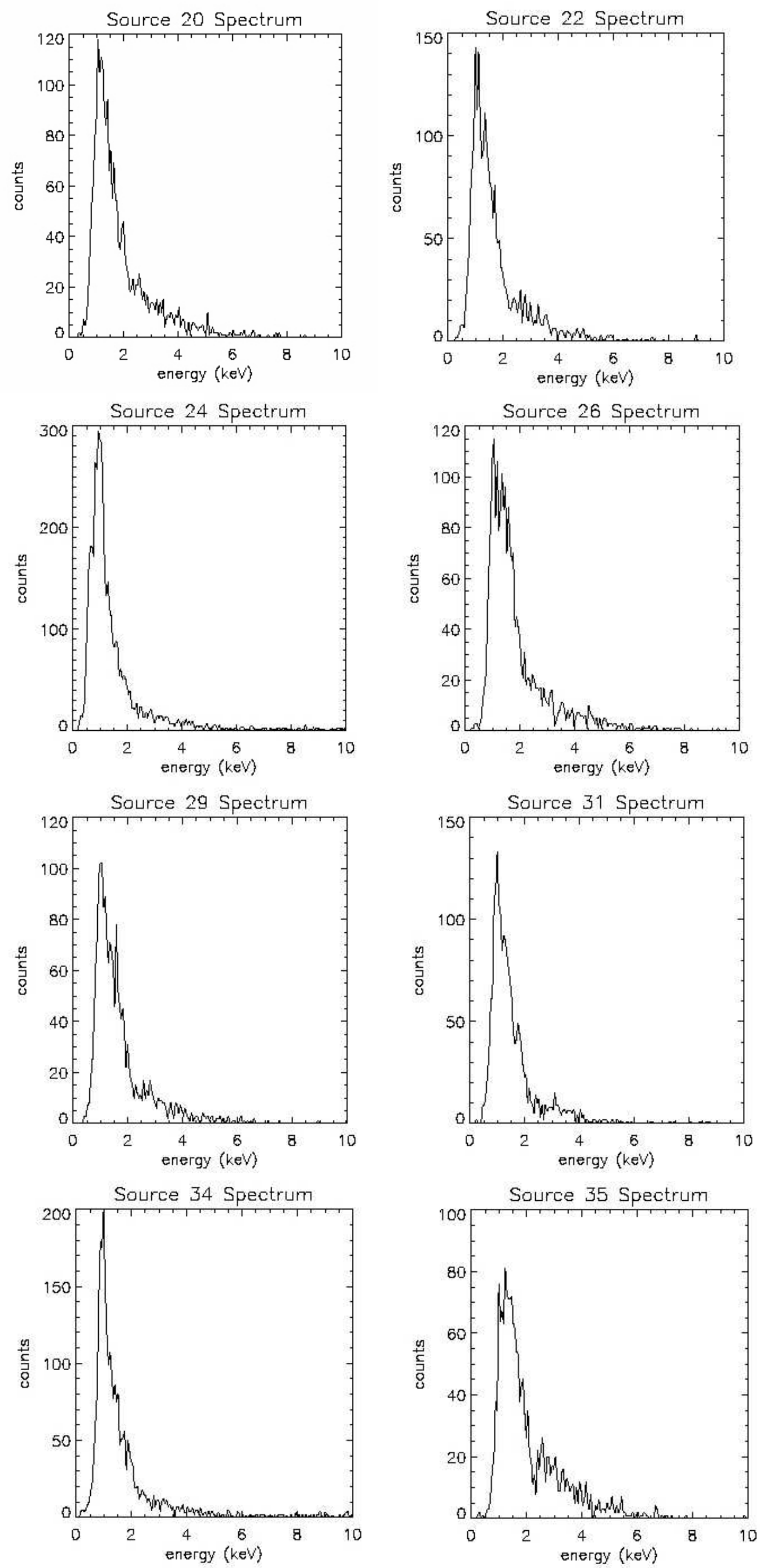


Figure 7.11: Spectra for All Sources in Class 1.

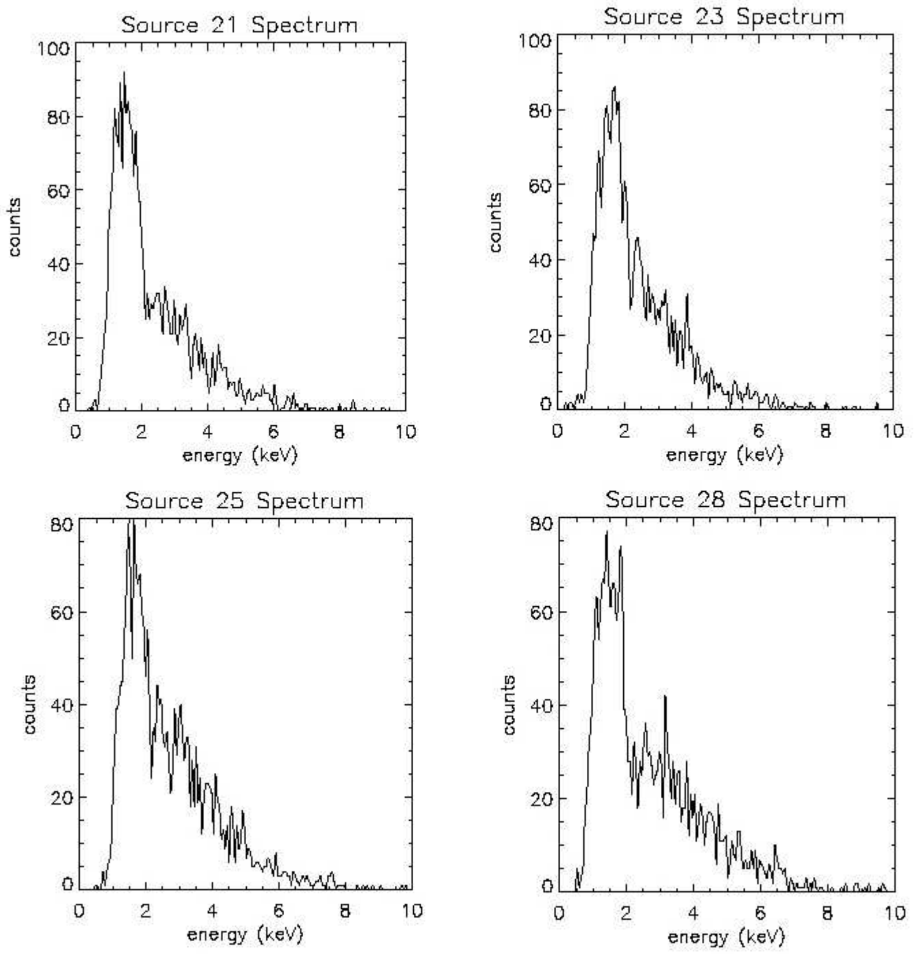


Figure 7.12: Spectra for All Sources in Class 2.

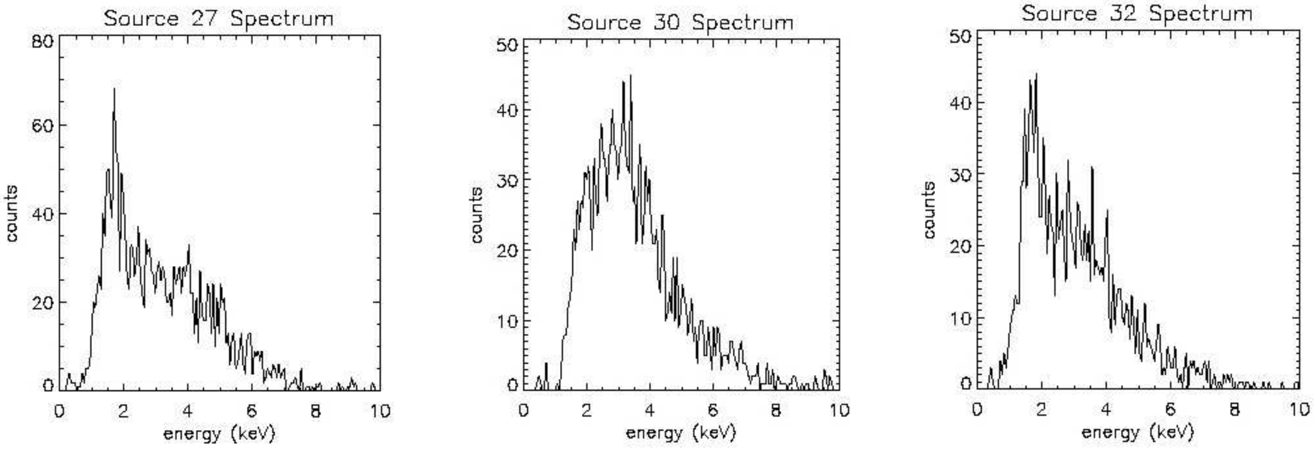
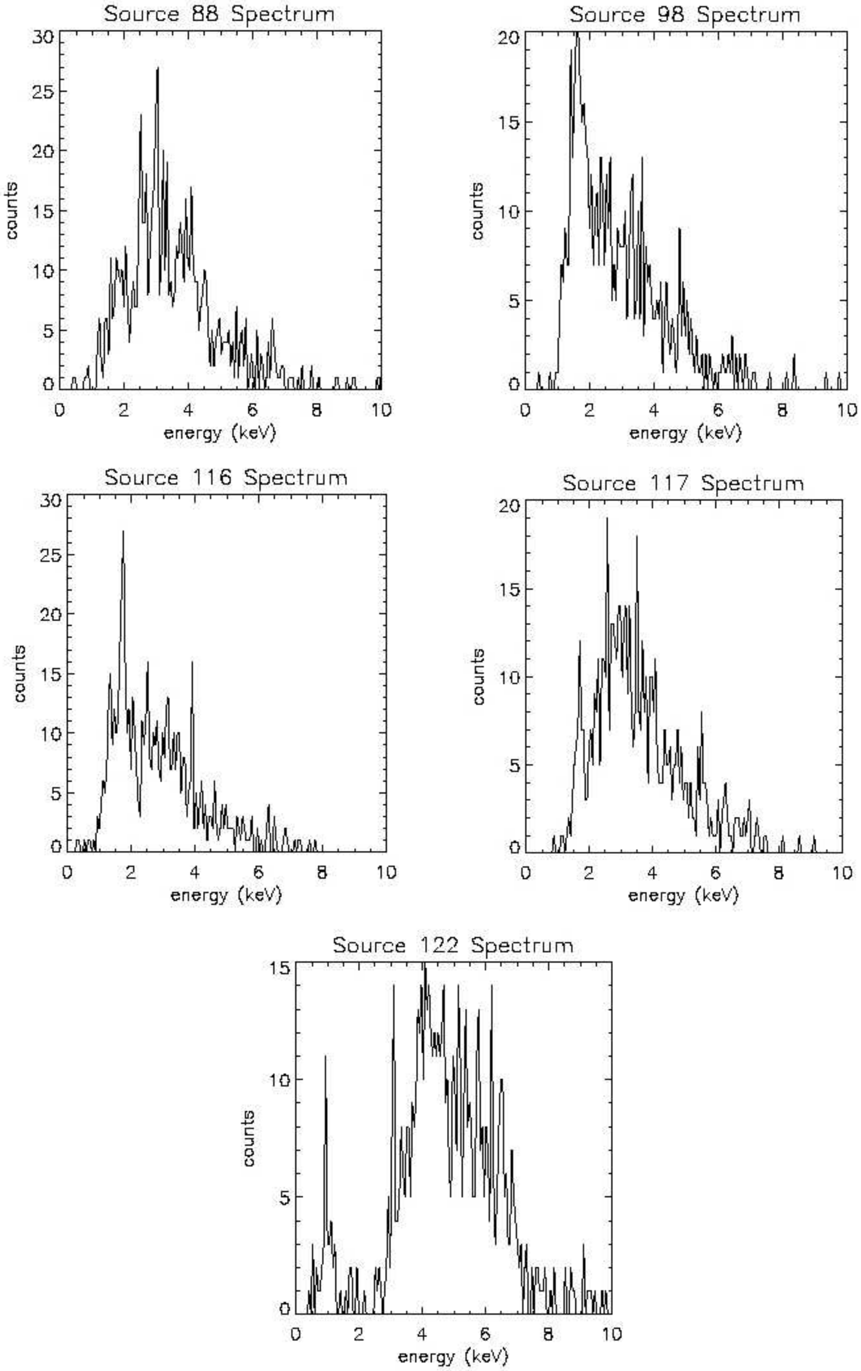


Figure 7.13: Spectra for All Sources in Class 3.



Source 88 Spectrum

Source 98 Spectrum

Source 116 Spectrum

Source 117 Spectrum

Source 122 Spectrum

Figure 7.14: Spectra for All Sources in Class 8.

Chapter 8

X-ray Source Classification Algorithm

8.1 *Chandra* Orion Ultradeep Project

Data from the *Chandra* Orion Ultradeep Project[12] (COUP) observation (ObsIds 4395, 3744, 4373, 4374, 4396, and 3498) obtained in January 2003 was used as the input dataset for development of the X-ray source classification algorithm. The 838 ks total exposure consists of six consecutive exposures obtained over a nearly-continuous period of 13.2 days. There is a gap of 29 ks between exposures due to removal of the ACIS (see Chapter 3) from the focal plane during five passages thru the Van Allen belts during this period. The COUP dataset represents the most sensitive and comprehensive description of X-ray emission from a PMS star cluster[12]. The dataset was released to the international COUP team of researchers by the COUP Data Reduction and Catalog group in November of 2003 and is summarized in Getman et al. (2005). Examples of spectra for two of the sources detected are shown in Figure 8.1.

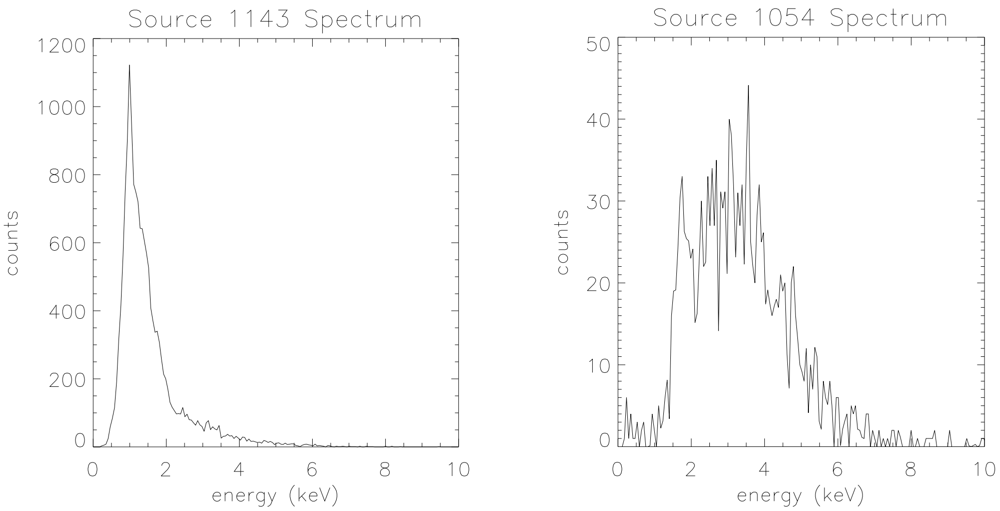


Figure 8.1: Examples of **s**oft (left) and hard (right) X-ray spectra among sources detected in the ONC.

8.1.1 Data Reduction

The Data Reduction and Catalog group of the international COUP team of researchers reduced the COUP data in a similar manner to that described for the input dataset for the prototype algorithm (see section 7.1.1), extracting valid events, locating sources, deriving X-ray properties, and constructing scientifically useful publishable tables, atlases and data files[12].

Table 8.1: Source detection problems in the COUP observation.

Number of Sources	Source Detection Problem
74	source existence is uncertain
422	double source (90% point spread function overlap)
65	pileup source (photon surface brightness > 0.003 ct/s/pix)
251	source extraction region crosses a bright source CCD readout trail
656	source in wings of a bright source with > 20000 counts or source with offaxis < 2 arcmin
556	source with inhomogeneous or low exposure map

More than 1600 sources were detected in the COUP dataset. A number of the detected sources in the COUP observation were flagged as having “source detection” problems (Table 8.1). A detected source can be flagged with more than one source detection problem.

8.1.2 Selection of Subset

A subset of the COUP observation was selected for use in developing the algorithm. Sources in the COUP dataset that were flagged as having the following source detection problems were eliminated: double source, pileup source, and source extraction region crossing a bright source readout trail. Faint sources, considered to be any source with less than 300 counts, were also eliminated. This resulted in a sample size of 444 sources for which high quality ACIS spectra could be drawn from the COUP dataset.

8.1.3 Background Correction

The long exposure of the COUP observation resulted in significant accumulation of uniform surface brightness background. The percentage of background for an individual source is calculated as follows:

$$\begin{aligned}\%\ \text{background} &= \text{BkgCts} / (\text{Total Counts}) * 100 \\ &= \text{BkgCts} / (\ \text{NetCts} + \text{BkgCts}\ ) * 100\end{aligned}$$

where,

BkgCts is the photon counts due to background radiation

NetCts is the net photon counts for the detected source

Values for BkgCts and NetCts have been provided for each source by the COUP team in the distributed data files. For my subset of 444 COUP sources, the average percentage of background is 4.41% (see Appendix D). There are ~50 sources with greater than 9% background. Photon counts due to the background radiation had to be subtracted from the photon counts for each detected source. Construction of local background spectra for each source was performed by the COUP Data Reduction and Catalog group. The process began with removal of the sources from the observation. The observation's exposure map was modified in exactly the same way, so that it accurately represented which regions have background data and which are masked out. This resulted in a data set and corresponding exposure map that look like "Swiss cheese" due to all the holes where sources were detected. A software tool called ACIS Extract[76] (AE) was then used to construct a local background spectrum for each source. AE found the smallest circular region around each detected and extracted source that contained at least the minimum number of background counts specified. If a high minimum number of background counts is specified for a region with relatively low background then relatively large background regions will result. For the COUP dataset, the minimum number of background counts for the smallest circular region around each detected and extracted source was set at 100 counts.

Since a region larger than the source extraction region was used to estimate the background spectrum, the background spectrum had to be normalized to the size of the source region. I did this by multiplying the background spectrum by a scale factor, equivalent to the ratio of source to background extraction region area, to adjust for the difference in size of the background region as compared to the size of the source region. Then the scaled background spectrum was subtracted from the source spectrum.

The spectrum of source 1067, before and after background correction, is shown in Figure 8.2. The Au Lα line feature at 9.7 keV in the original spectrum is due to fluorescence of material in the telescope. The feature has effectively been removed by the background-subtraction procedure.

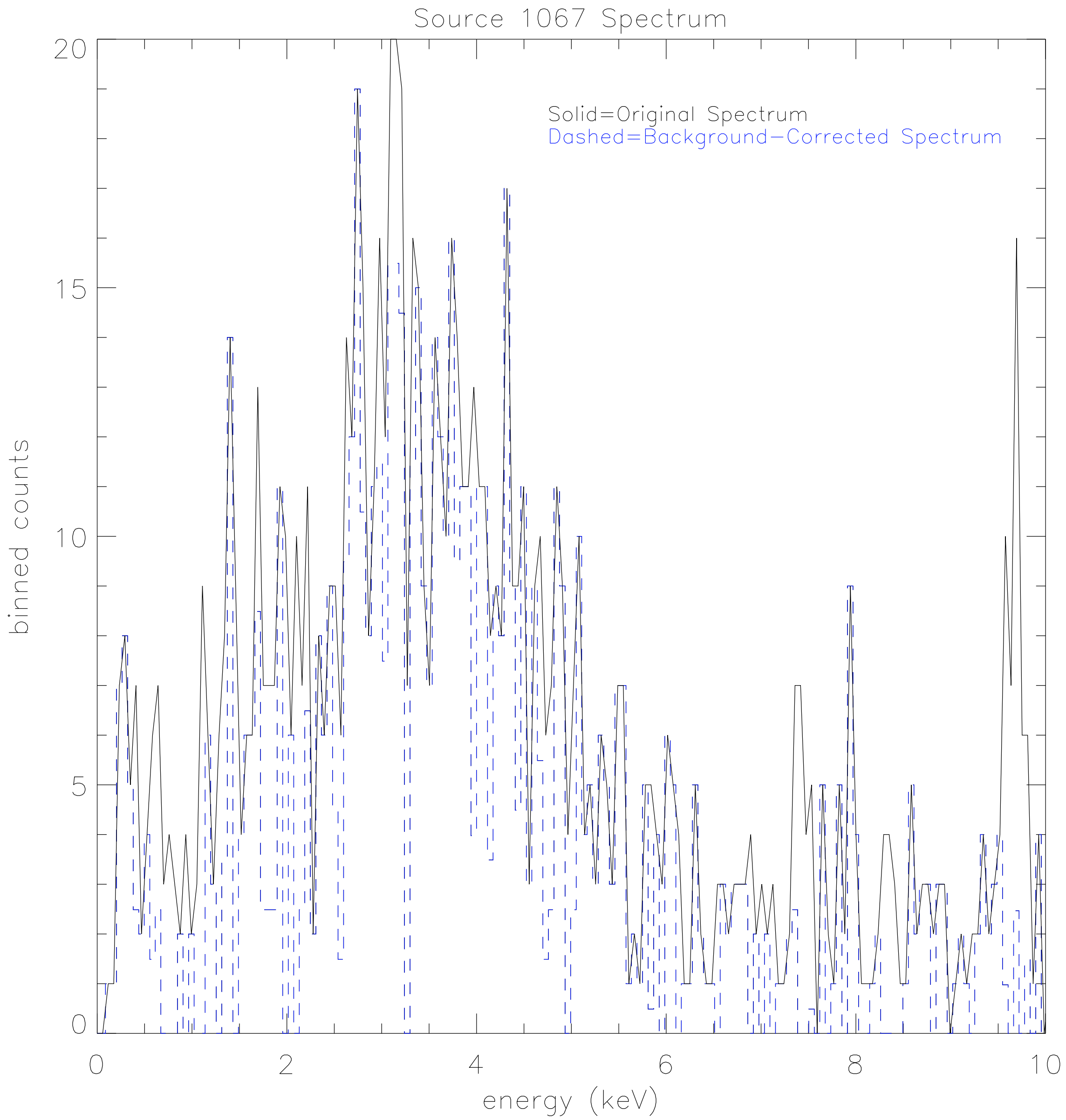


Figure 8.2: Original (solid black line) and background-corrected (dashed blue line) spectra for COUP source 1067.

The subset of 444 sources detected in the COUP observation (see section 8.1) was used for development and testing of the X-ray source classification algorithm. The high energy X-ray spectrum was divided into 42 bands (see section 6.5 and Appendix A) using the following factors to guide my choice of band locations and widths:

- high-resolution emission line data
- spectral resolution of the X-ray CCD detector arrays at -120 deg C, the nominal operating temp of *Chandra*
- quantum efficiency of the CCDs

The total number of photon counts within each of the 42 spectral bands was used as the multivariate input variables. A monotonic transformation was performed on the input data to reduce non-linearities. The correlation matrix for the resulting transformed band data is shown in Appendix E.

8.2 Principal Component Analysis[g]

PCA was used to reduce the redundancy in the transformed X-ray spectral bands. The goal of PCA is to identify a new, smaller set of uncorrelated variables, called *principal components*, which explain all or nearly all of the total variance in the dataset. Each principal component is described by:

- an eigenvector: a linear combination of the original input variables
- an eigenvalue: the variance accounted for by that component

The covariance matrix was used for PCA, rather than the correlation matrix. This is equivalent to foregoing standardization of the input variables (see section 5.1). The units of measurement of all

[g] See section 5.1 for a detailed description of PCA.

the input variables (spectral bands) are commensurate and they were not measured on widely differing scales. Therefore, standardization was not necessary.

8.2.1 Starting Rules

Two starting rules for PCA were examined. These rules aid in determining whether there is enough correlation in the dataset to warrant applying PCA. The first starting rule is a check of the bounds on the eigenvalues. Therefore, an eigenanalysis of the covariance matrix was performed to calculate the eigenvalues. The resulting eigenvectors are shown in Appendix F and the eigenvalues are shown in Appendix G.

The lower bound for the first eigenvalue (the eigenvalue corresponding to the first principal component) is the maximum variance in the sample covariance matrix[77].

$$\lambda_1 \geq \max\left(s_i^2\right) \qquad \text{for i = 1, 2, …, p} \qquad (1.1)$$

where p is the number of eigenvalues. For the COUP observation subset, this becomes

$$\lambda_1 \geq 1.114913\text{E} + 12 \qquad (1.2)$$

true, since λ_1 = 5.4858E + 12

The values for λ_1 are large because the covariance matrix rather than the correlation matrix was used. The upper bound for the first eigenvalue is the maximum of the row sums of the absolute values of the covariance matrix.

$$\lambda_1 \leq \max \sum_{j=1}^{p} \left| r_{ij} \right| \qquad \text{for i = 1,2,…, p} \qquad (1.3)$$

where p is the number of eigenvalues.

For the COUP observation subset, this becomes

$$\lambda_1 \leq 1.22881\text{E}+13 \tag{1.4}$$

true, since $\lambda_1 = 5.4858\text{E} + 12$

Another starting rule involves redundancy in the input dataset. Gleason and Staelin[78] calculated a single number from the correlation matrix to determine the level of correlation among the variables (see Equation 1.5). They called this a measure of redundancy.

$$\varphi = \sqrt{\frac{\|R\|^2 - p}{p(p-1)}} \tag{1.5}$$

where,

$$\|R\|^2 = \sum \sum r_{ij}^2 = \sum \lambda_i^2 \qquad \text{for i,j = 1,2, … , p} \tag{1.6}$$

This coefficient has the same range as a multiple correlation coefficient. If the variables are perfectly correlated then $\varphi = 1$. If there is no correlation among the variables then $\varphi = 0$. A Monte Carlo study done by Gleason and Staelin[78] show that PCA is not useful when φ gets below 0.16. However, the distribution of φ is not known and interpretation is guided by experience. For the COUP observation subset, $\varphi = 0.682$, indicating that there is correlation among the 42 variables. The correlation is high enough to indicate PCA may be of use in eliminating redundancy in the data.

8.2.2 Stopping Rules

Stopping rules are used to determine the number of components to retain. The amount of total sample variance explained, the relative variances of the components, and possible interpretations of the components have been used to determine the number of components to retain. There are over 20 stopping rules detailed in the literature. Some of these stopping rules are statistical significance tests for homogeneity of the eigenvalues. The first of these tests was developed by Bartlett[79]. Additional statistical significance tests were developed for unique population eigenvalues[80], and for small sample sizes and non-normal data[81]. Many of the other stopping rules are, necessarily, ad-hoc, testing physical significance for a specific application area. A subset of stopping rules was selected to use on the COUP subset. The stopping rules were selected based on computation speed, ease of automation, and acceptance in the literature.

The following stopping rules were computed to determine the number of components to retain for the clustering analysis. Several stopping rules have been included simply because they are commonly used. Before using the statistical significance test, several ad-hoc stopping rules were used, to get an approximation for the number of components to retain.

8.2.2.1 Scree Test

A scree plot[73,74] is a graph of the eigenvalues in decreasing order of variance explained. Scree is defined as a slope of loose rock debris at the base of a cliff or steep incline. Cattell named this the ‘scree plot’ because the retained eigenvalues appear as a cliff and the deleted ones are the slope of loose rock debris at the base[74]. An “elbow”, bend, or break in the scree plot shows the location after which the eigenvalues are all relatively small and of relatively equal value. The components prior to this elbow are retained[73]. Some authors also retain the component at the location of the bend[75]. The scree plot for the COUP observation subset is shown in Figure 8.3. It can be seen from the figure that there is a well-defined bend in the scree plot. The first three components are retained.

The scree plot is a nice visual aid for determining the number of components to retain. However, it is not very useful for an automated algorithm due to the variety of breaks that could arise, including multiple breaks or bends.

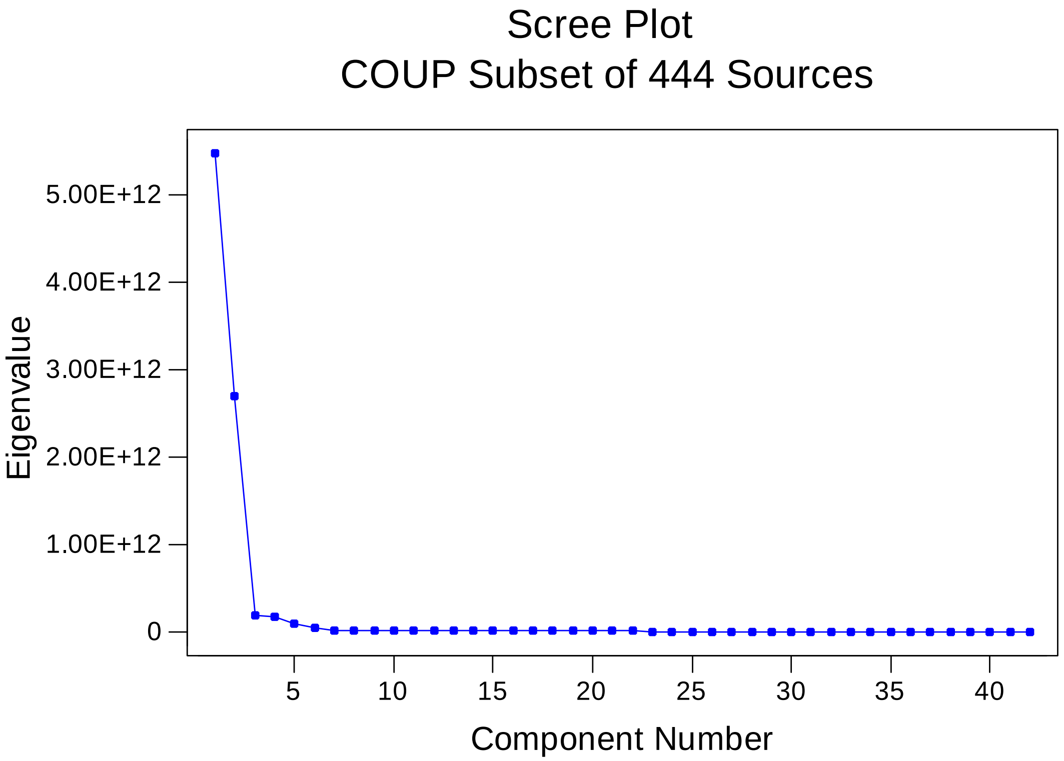


Figure 8.3: Scree Plot for COUP Subset

8.2.2.2 Horn's Stopping Rule

Horn suggested generating a normally-distributed random dataset that is the same size as the real dataset. The mean and standard deviation of the original dataset is used in generating the random dataset. The eigenvalues of this random dataset are plotted against the eigenvalues of the original dataset[82]. Where the plot from the random data crosses the plot from the real data indicates the point that separates the retained and deleted principal components.

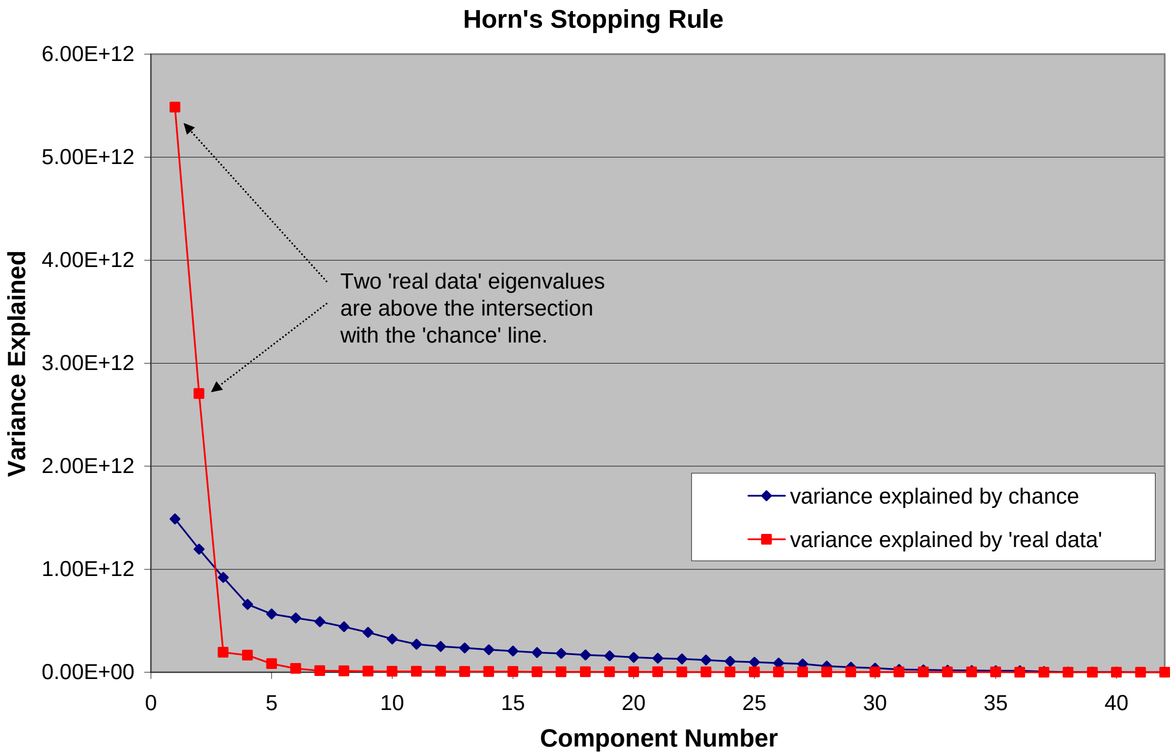


Figure 8.4: Depiction of Horn’s Stopping Rule

K matrices of random variables were generated, each matrix of order 42 by 444. The covariance matrix for each of the K matrices was computed. Then an eigenanalysis of each of the covariance matrices was performed. This resulted in a set of 42 eigenvalues for each of the K random matrices. The first eigenvalues were averaged over K, the second eigenvalues were averaged over K, and so forth. The plot of the resulting averaged eigenvalues is shown in blue in Figure 8.4. The plot of the real data (from the COUP subset) is shown in red. Two eigenvalues from the real data are above the intersection with the line obtained from the random data. These two eigenvalues are larger than they would be by chance alone. Therefore, Horn's stopping rule indicates that two principal components should be retained.

8.2.2.3 Broken Stick

The broken stick is a simple stopping rule proposed by Jolliffe[83] as a quick estimation of Horn's stopping rule. If a line of unit length is randomly broken up into p segments, the expected length of the kth-longest segment is:

$$g_k \quad = \quad \frac{1}{p}\sum_{i=k}^{p}\left(\frac{1}{i}\right) \qquad\qquad \text{for k = 1,2,…, p} \qquad\qquad (1.7)$$

The proportion of variance explained is calculated for each of the p principal components. Retain any principal component that explains a greater proportion of the variance than the corresponding value of g_k. In Figure 8.5, the blue line is a plot of equation 1.7 for the COUP subset. The red line is a plot of the proportion of variance explained for each component. The first two components should be retained, according to this method.

8.2.2.4 Average Eigenvalue

The Average Eigenvalue method is a quick technique that retains components with eigenvalues greater than the average eigenvalue[84]. The average eigenvalue is given by

$$\overline{\lambda} \;=\; \left(s_1^2 + s_2^2 + \cdots + s_p^2\right) \;/\; p$$

where s_i^2 is the variance of variable i and p is the total number of eigenvalues. For the COUP observation dataset with 42 variables, the average eigenvalue, $\overline{\lambda}$, is 2.0964E+11. The first two components, with eigenvalues of 5.49E+12 and 2.70E+12 respectively, would be retained.

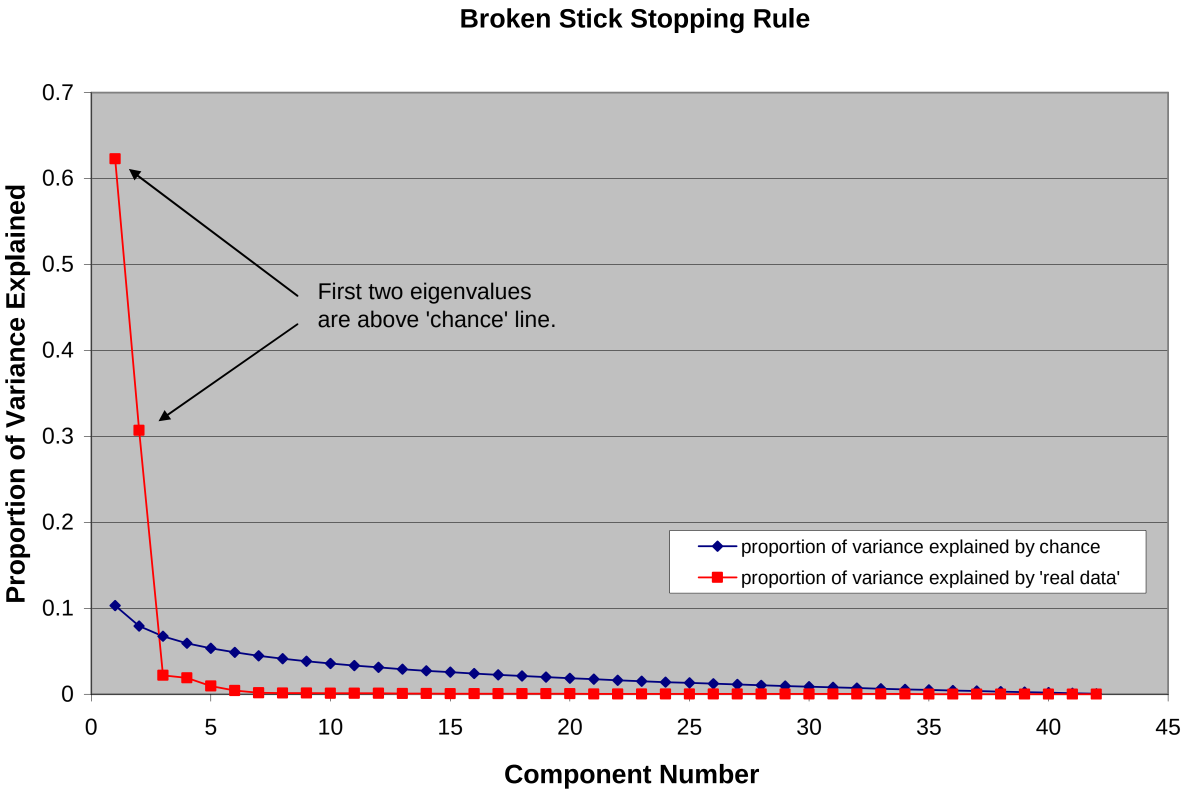


Figure 8.5: Depiction of Broken Stick stopping rule

Jolliffe[85] ran PCA on simulated data and maintained that the Average Eigenvalue method does not allow for sampling variability and therefore, retains too few components. He modified the Average Eigenvalue method by reducing the value of the average eigenvalue, prior to comparison with the component eigenvalues (see Equation 1.8). Jolliffe chose to reduce the value of the average eigenvalue by 70%, based on simulation studies on 587 sets of artificial data[85].

$$\overline{\lambda} = \mathbf{0.70} \times \left[\left(s_1^2 + s_2^2 + \cdots + s_p^2\right) / p\right] \qquad (1.8)$$

Applying this technique to the COUP subset results in a modified average eigenvalue of 1.4675E+11. The third and fourth components have eigenvalues of 1.93E+11 and 1.67E+11 respectively. Hence, they would also be retained. Therefore, a total of 4 components would be retained by use of this stopping rule.

8.2.2.5 Statistical Significance Tests

A statistical significance test is used to determine if the eigenvalues of the discarded components are not significantly different from each other. The null hypothesis, H_0, is that the eigenvalues of the last (p – k) eliminated components are approximately equal. This test allows for non-distinct eigenvalues:

$$\lambda_{p-k} \leq \lambda_{p-k+1} \leq \cdots \leq \lambda_p \tag{1.6}$$

where p is the number of eigenvalues (p also equals the number of variables) and k is the number of components retained. Bartlett's significance test is not robust for non-normal datasets[79]. Levene[81] developed a significance test that can be used for data that come from continuous, but not necessarily normal distributions. It can also be used with small sample sizes.

The initial value of k was determined from the stopping rules listed in the previous sections (Table 8.2).

Table 8.2: Comparison of Stopping Rules

Scree Plot	**Broken Stick**	**Average Root**	**Jolliffe 70% of Average Root**	**Horn Average Eigenvalues**
3	2	2	4	2

If the discarded components really have equal, or approximately equal variances, what is the chance that randomly selected samples would have variances as far apart or more so (i.e. have an F-test value as large or larger) as observed in the sample dataset? The P value answers this question. The P value is a probability that represents the lowest *level* of significance at which the observed value of the test statistic is significant. Any observed value of the test statistic is considered significant if the null hypothesis is rejected at the specified level of significance. If the P value is small, we can conclude that the variances (and thus the eigenvalues) are significantly different and it is unlikely that the difference observed is due to a coincidence of random sampling. We can reject the idea that the difference is a coincidence and conclude instead that the principal components have different variances. Therefore, the null hypothesis is rejected. If the P value is large, there is insufficient evidence, based on the data, that the eigenvalues of the discarded components differ.

The sample size of the COUP subset and the number of levels of the categorical variable (spectral bands) are so large, that there are large degrees of freedom in both the numerator and denominator of the F-test statistic. The test has extremely high power to detect statistically different variances, which may in fact have no practical significance. In other words, nearly any difference in the eigenvalues of the discarded components would be judged significant. A difference may be detected that is not considered significant for the determination of the number of components to discard.

A random subset of 44 sources (10%) was drawn from the COUP subset to lower the power of the test. Since Levene’s test works well for small sample sizes, it was used on the random subset. Table 8.3 below shows the results of Levene’s statistical significance test for homogeneity of variance for the random sample of 10% of COUP observation subset. A small significance probability, Pr > F, indicates that some linear function of the parameters is significantly different from zero. It is important to note that “statistically significant” is not the same as “physically or scientifically important”. It can be seen from Table 8.3 that the power of the test is still too high. The test suggests retaining 40 components. I tried reducing the sample size again, to 22 observations (5% of the COUP subset), but the resulting score plot (principal component 2 versus principal component 1) did not resemble the original score plot, suggesting the relationships in the original dataset were not represented accurately by the 5% sample. Therefore, the statistical significance test could not be run on my COUP subset.

8.2.3 Stopping Rule Conclusions

Four principal components were retained for the ensuing clustering steps. I chose to retain four components, although three of the stopping rules suggested retaining two (Table 8.2). It is less detrimental to retain more principal components than needed (within reason) than to delete some components that may represent some of the inherent variability in the dataset. Retaining too many principal components increases the dimensionality of the dataset and may result in preserving unwanted redundancy and/or noise.

Table 8.3: Significance Probabilities From Levene's Test

Number of Components Retained	f	Pr (F > f)
3	5.88	0.0001
4	4.43	0.0001
5	3.51	0.0001
6	2.98	0.0001
7	3.93	0.0001
8	8.36	0.0001
9	9.26	0.0001
10	10.76	0.0001
11	9.63	0.0001
12	9.21	0.0001
13	9.67	0.0001
•	•	•
•	•	•
•	•	•
34	7.95	0.0001
35	7.78	0.0001
36	4.21	0.0011
37	4.30	0.0023
38	4.21	0.0011
39	7.71	0.0010
40	1.39	0.2412

8.2.4 Eigenvector and Score Plots

Plots of the eigenvectors that correspond to the first four principal components are shown in Figure 8.6. PC 1 could be interpreted as an indicator of sources with soft X-ray spectra with an energy peak around 1 keV. PC 2 also indicates sources with soft X-ray spectra, however, the energy peak is shifted to the right, peaking around 1.5 keV. PCs 3 and 4 could be indicators of sources that have both a soft X-ray component and a hard X-ray component. The hard X-ray component indicated by PC 4 is broader and farther to the right than that of PC 3. The score plot, a plot of PC 2 versus PC 1, is shown in Figure 8.7. The overall shape of this score plot is curved, rather than aligned somewhat linearly along one of the axes or randomly scattered. This effect is sometimes seen in ecological studies of species and environmental gradients[86]. It occurs generally when the following conditions are found in the dataset:

- objects have unimodal distributions along gradients
- input variables all have the same units
- data are approximately on the same scale

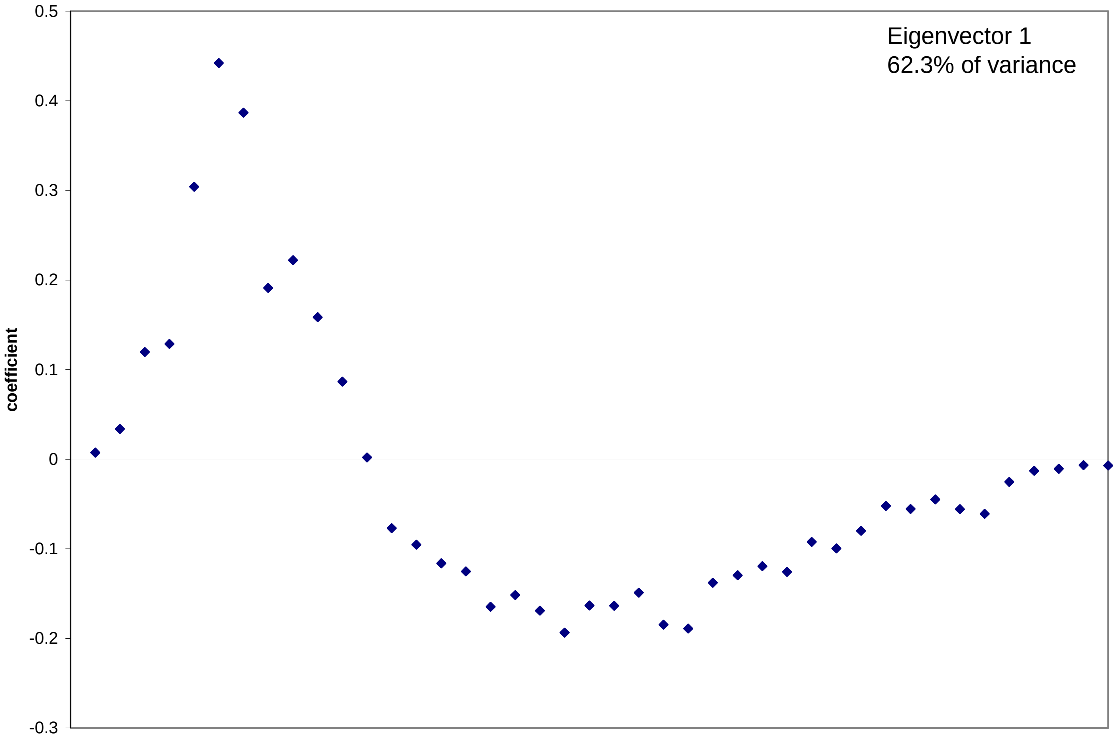
The effect of the gradient on the distance relationship between the input variables (i.e., spectral bands), calculated from the count data, is non-linear. This non-linearity shows up as a curve in the score plot. The shape of the curve can range from a bow, to an arch, to a horseshoe (one or both ends curve inwards). The shape of the score plot shown in Figure 8.7 is a horseshoe due to the incurving of the ends.

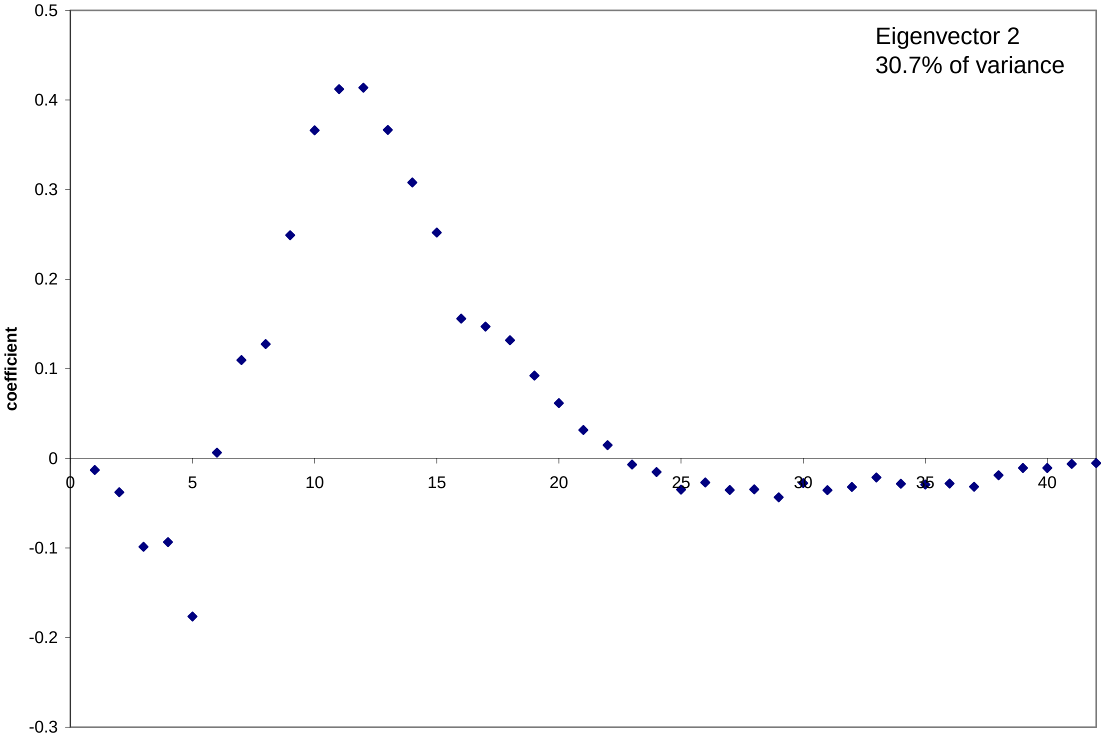
8.3 Agglomerative Hierarchical Clustering[h]

I used the unsupervised methods of agglomerative hierarchical clustering and K-means clustering for my research because one goal was to find “true” groupings of X-ray sources in the ONC, without attempting to fit the sources to any pre-defined models or groupings. The clustering techniques used find a “natural” partitioning of the data set into a relatively homogeneous number of groups, K. An agglomerative hierarchical clustering method based on Euclidean distance and complete linkage was used on the first four principal components generated from the detected X-ray sources.

Similar groups were successively merged based on the Euclidean distance measure. Cluster merging continued until there was only one large cluster containing all the sources. At this point, the Euclidean distance for each of the intermediate clustering steps was manually examined to find a large jump between amalgamations to estimate the number of source classes. The final number of clusters chosen was based on the distances between successive cluster mergers and application knowledge. The resulting dendrogram is shown in Figure 8.8. The horizontal dashed line at a distance level of approximately 2.0E+06 illustrates where the dendrogram has been cut to obtain 17 clusters. If this line were lowered, more clusters would be obtained. Table 8.4 lists the number of sources per class. Refer to Appendix H for a list of class membership as a result of running agglomerative hierarchical clustering.

[h] See section 5.2 for a detailed description of the agglomerative hierarchical clustering method used.


Eigenvector 1
62.3% of variance
coefficient
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3


Eigenvector 2
30.7% of variance
coefficient
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
0
5
10
15
20
25
30
35
40

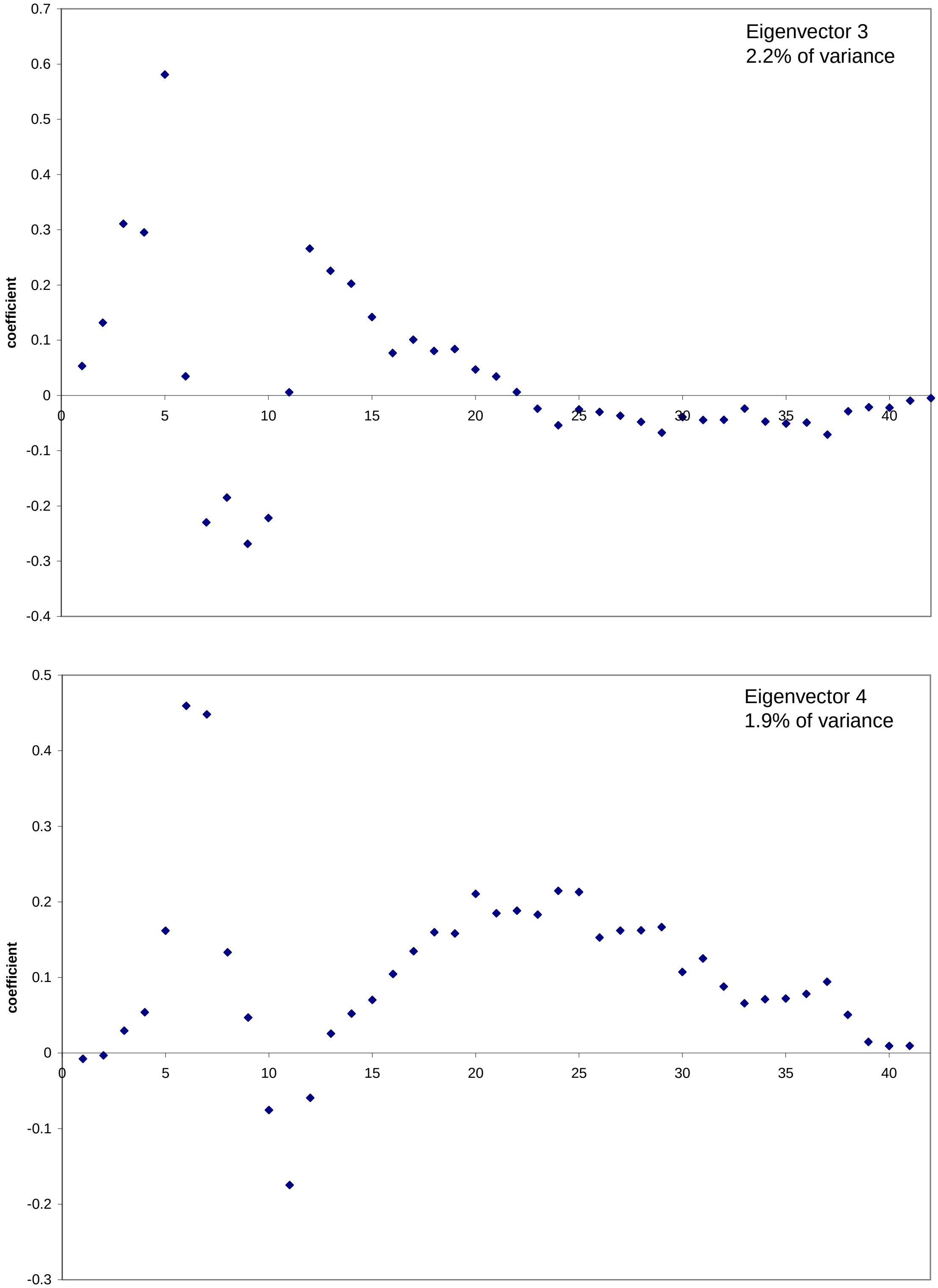


Figure 8.6: Eigenvector plots for the first four principal components.

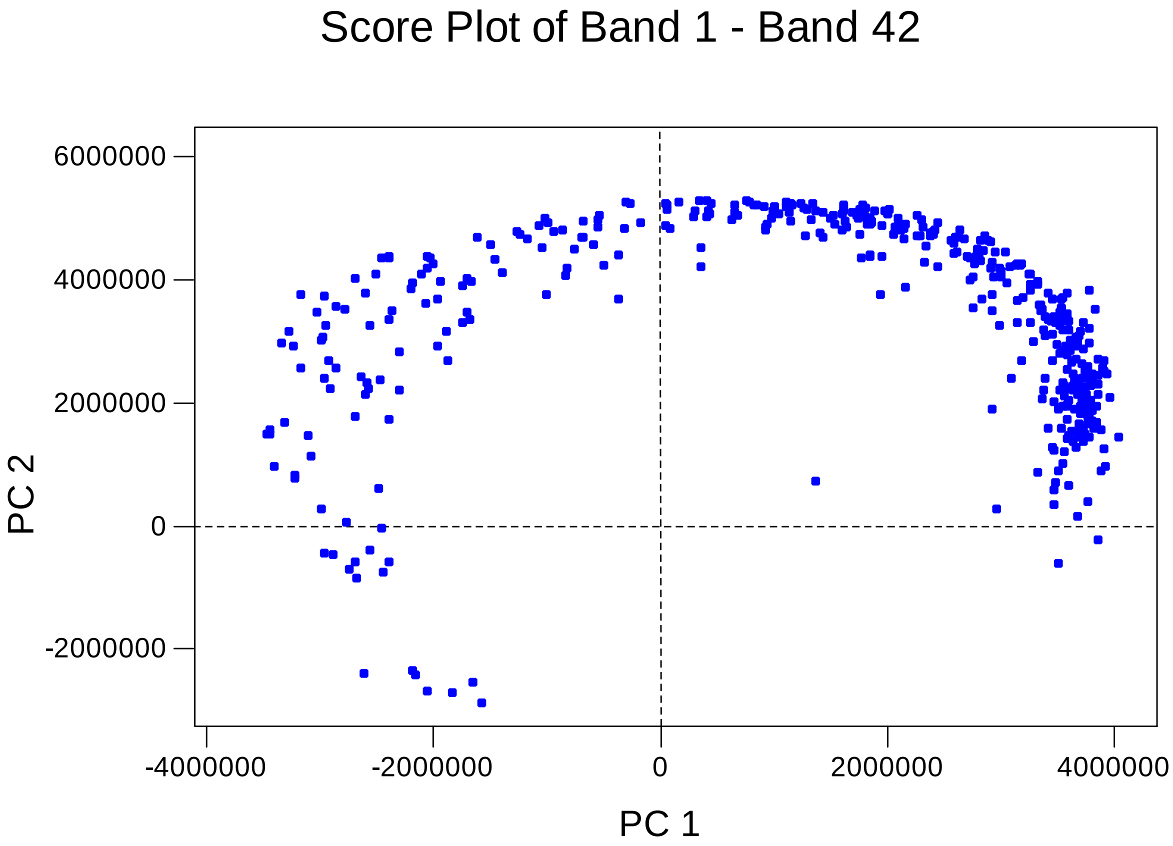


Figure 8.7: Score plot of PCs 1 and 2 computed from the X-ray spectral band data.

Table 8.4: Number of Sources Per Class After Agglomerative Hierarchical Clustering

Class	**Number of Sources**
1	7
2	12
3	9
4	18
5	2
6	9
7	24
8	21
9	12
10	14
11	68
12	44
13	32
14	108
15	49
16	14
17	1

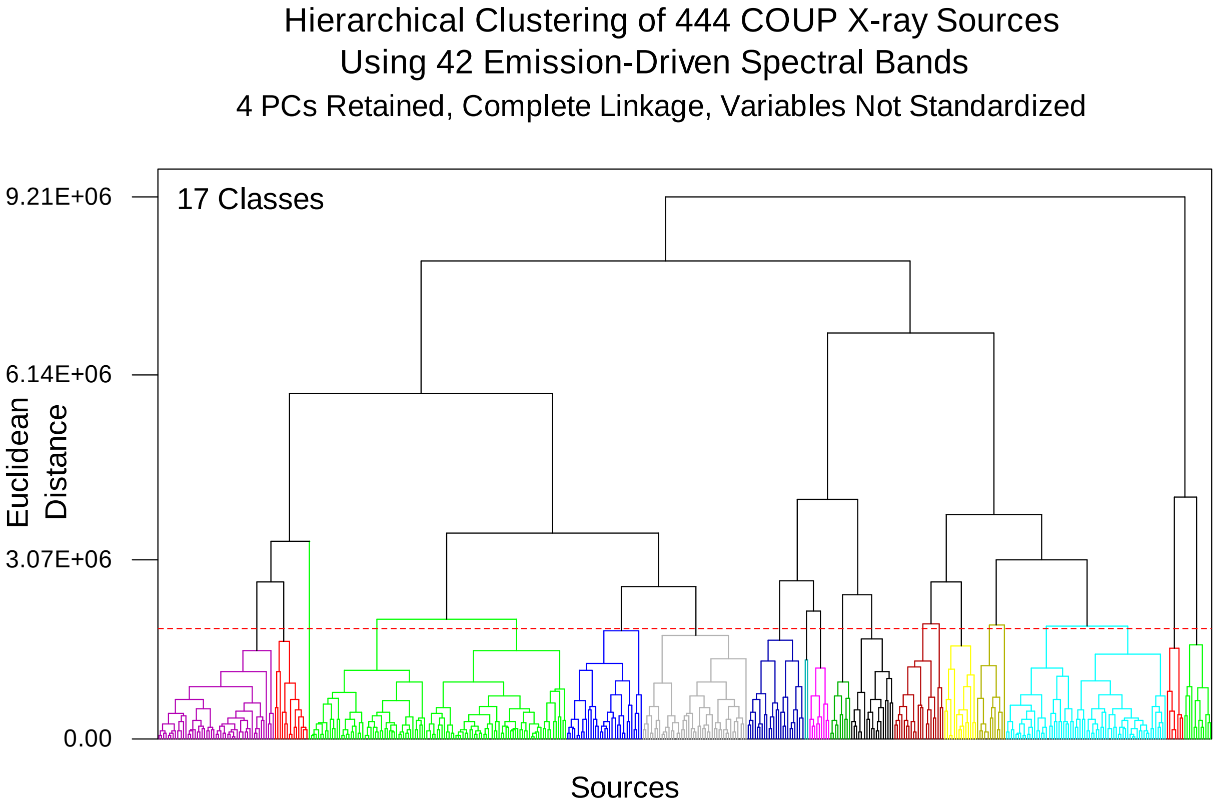


Figure 8.8: Dendrogram resulting from hierarchical clustering on COUP 444 subset, using Euclidean distance with complete linkage. The dashed line shows where the dendrogram was cut, resulting in 17 classes. Each class of sources is represented by a different color.

8.4 K-means Clustering[i]

The class assignments resulting from the hierarchical clustering were used as the initial partition for the K-means clustering. K, the number of classes, then becomes 17 by default. Each source was examined and assigned to the cluster with the nearest centroid (mean). In some cases, this resulted in the source being reassigned to a new cluster. The centroids were then recalculated for the cluster receiving the reassigned source and for the cluster losing the source. This process was repeated until no more reassignments took place. The final number of X-ray sources in each

[i] See section 5.3 for a detailed description of K-means clustering algorithm used.

cluster is shown in Table 8.5. The K-means algorithm moved 123 of the sources (27.7%) from one cluster to another during the course of the algorithm's iterations. Table 8.6 shows a 2-way cross-tabulation of the cluster membership after agglomerative hierarchical clustering (rows) and after K-means clustering (columns). Cell contents are counts. The counts on the diagonal represent sources that did not switch clusters during the K-means algorithm. The sources that did switch clusters did not move far from their initial cluster assignment. Appendix H details which sources moved to a different cluster during the K-means algorithm.

Table 8.5: Number of Sources Per Class After K-means Clustering

Class	**Number of Sources**
1	7
2	12
3	9
4	19
5	2
6	14
7	18
8	21
9	22
10	37
11	54
12	30
13	30
14	61
15	88
16	19
17	1

Table 8.6: Two-way cross-tabulation of the class membership after agglomerative hierarchical clustering (rows) and K-means clustering (columns).

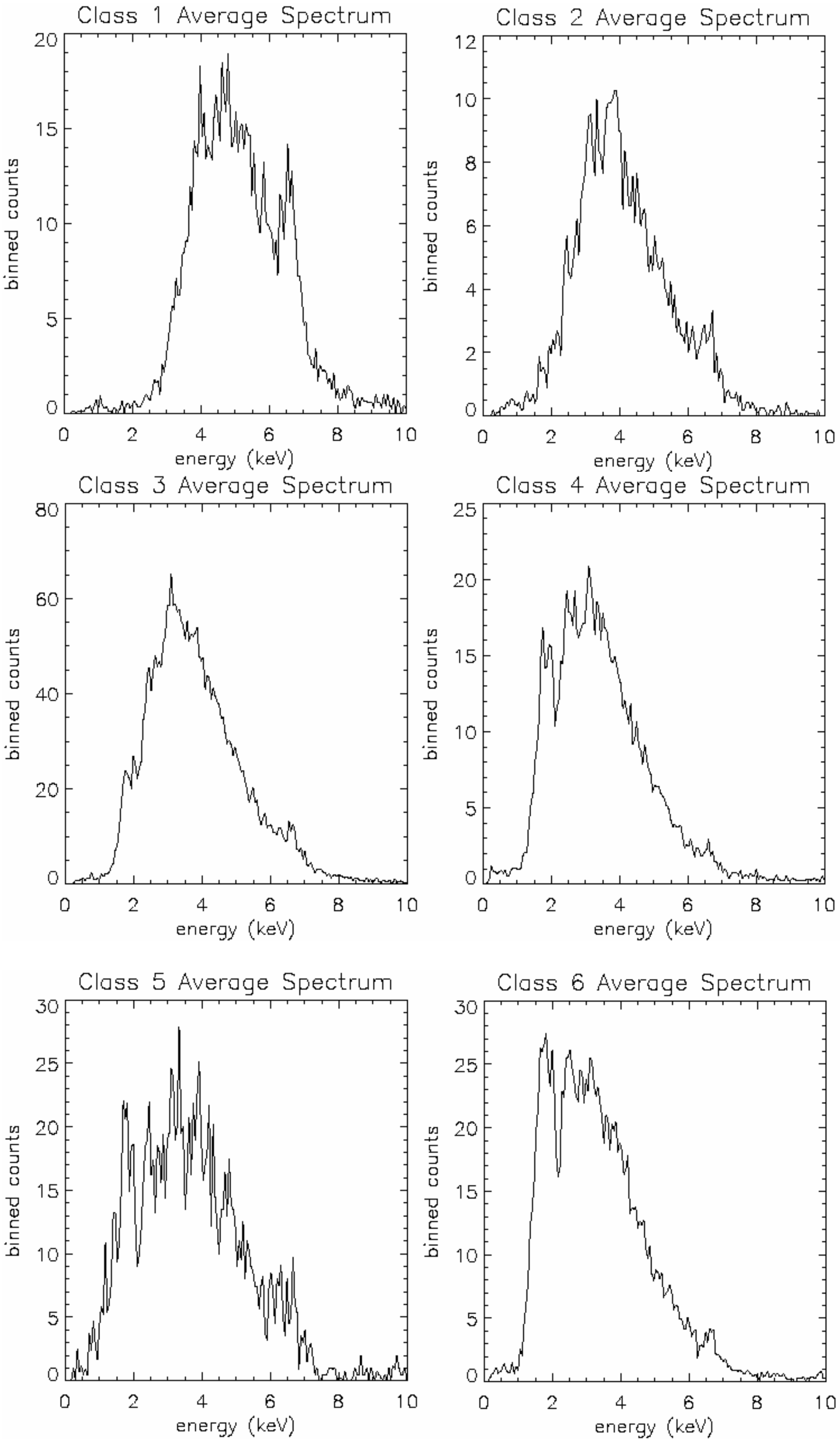
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
1	7																	7
2		12																12
3			9															9
4				18														18
5					2													2
6						9												9
7				1		5	18											24
8								20	1									21
9									12									12
10								1	9	4								14
11										33	35							68
12											16	28						44
13											3	2	27					32
14													3	61	44			108
15															44	5		49
16																14		14
17																	1	1
	7	12	9	19	2	14	18	21	22	37	54	30	30	61	88	19	1	444

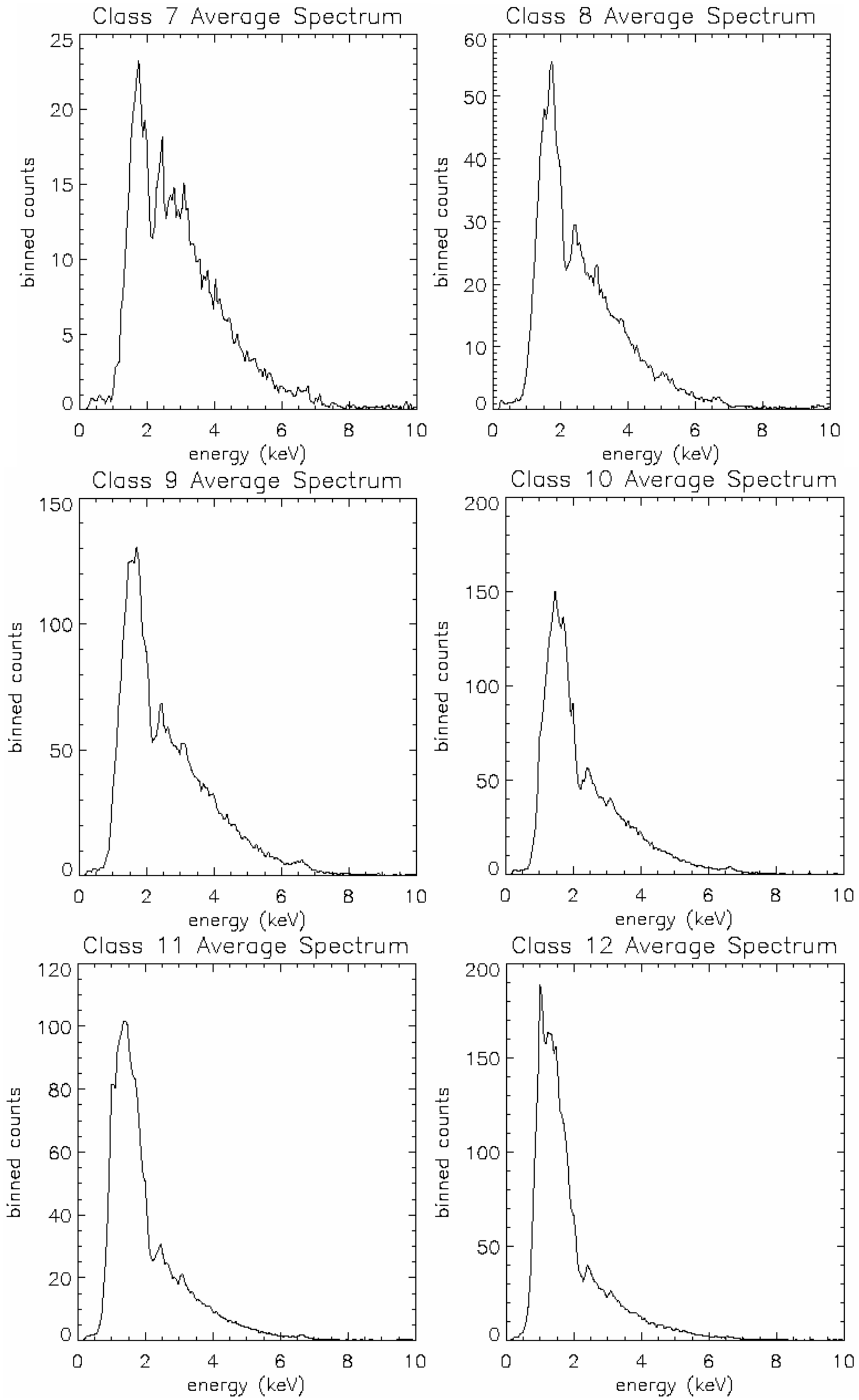
Chapter 9

Results Analysis

9.1 PCA Score Plots and Class Average Spectra

The source spectra were averaged for each of the 17 classes. The results are shown in Figure 9.1. The plot of the first two principal components for each source was recreated, this time color-coded by class (see Figure 9.2). The progression of classes moving clockwise around the horseshoe in Figure 9.2 forms a sequence of decreasing spectral hardness. The lowest numbered classes contain sources with the hardest spectra. These classes are also more easily separated visually in this plot of only the first two principal components. The highest numbered classes can be seen to be a continuum of sources with increasingly softer spectra. Finally, the lone source in Class 17 is an outlier that stands out in the plot.


Class 1 Average Spectrum
Class 2 Average Spectrum
Class 3 Average Spectrum
Class 4 Average Spectrum
Class 5 Average Spectrum
Class 6 Average Spectrum
binned counts
energy (keV)
0
2
4
6
8
10
5
15
20
12
25
30
40
60
80



Class 7 Average Spectrum

Class 8 Average Spectrum

Class 9 Average Spectrum

Class 10 Average Spectrum

Class 11 Average Spectrum

Class 12 Average Spectrum

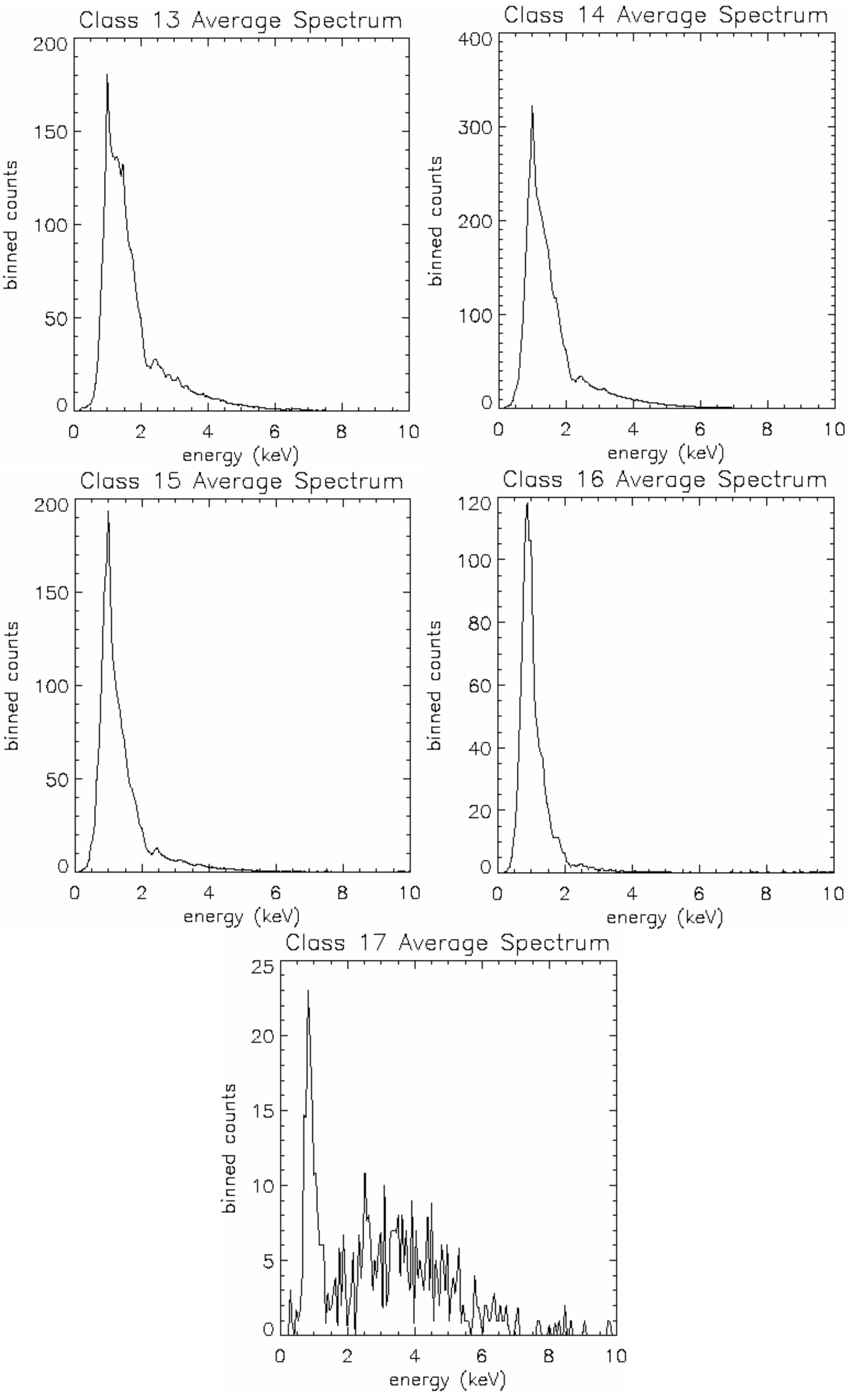


Figure 9.1: Average spectra for each of the 17 classes.

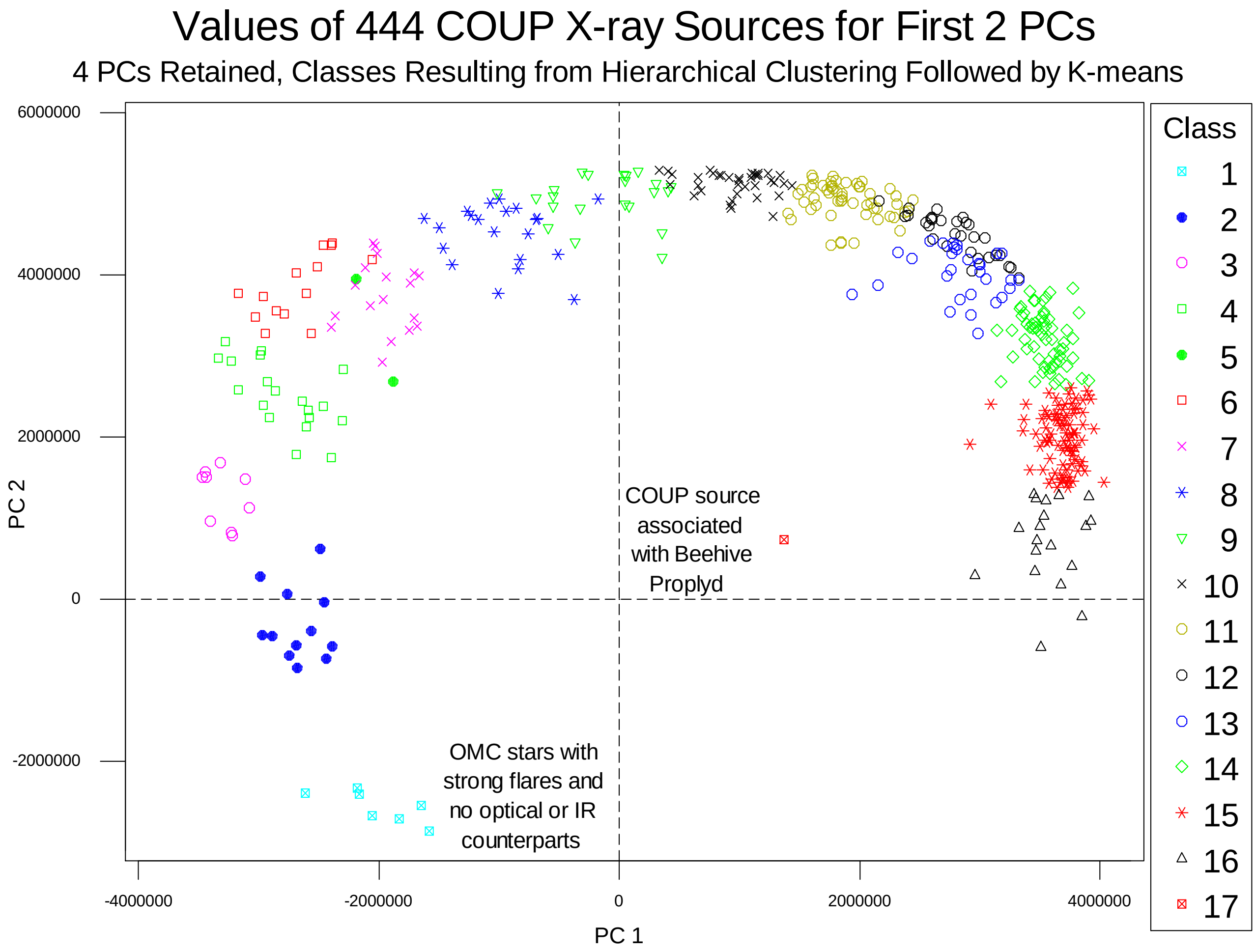


Figure 9.2: Plot of the first 2 principal components with the source classes shown. The class numbers increase clockwise around the horseshoe-shaped curve.

The plot of the first two principal components typically holds the most information about the clustering, since principal components 1 and 2 explain most of the variance in the data. For my data, principal components 1 and 2 account for 93% of the variance in the data. However, the first four principal components were used in the clustering algorithm. Principal components 3 and 4 contribute an additional 4.1% of the variance. Since it is not possible to envision a plot of all four principal components simultaneously in 4-D space, pairs of the retained principal components were plotted for further insight into the clustering assignments. A plot of PC 3 versus PC 1 is shown in Figure 9.3, PC 4 versus PC 1 is shown in Figure 9.4, PC 3 versus PC 2 in Figure 9.5, PC 4 versus PC 2 in Figure 9.6, and PC 4 versus PC 3 in Figure 9.7.

Upon examining the plot of principal component 3 versus principal component 1 (Figure 9.3), it can be seen that the separation between the classes containing sources with harder spectra (lowest numbered classes) is still apparent in this plot. Better separation between some of the classes containing sources with softer spectra can be seen in this plot, also. The lone X-ray source in Class 17 is an outlier in this plot, too.

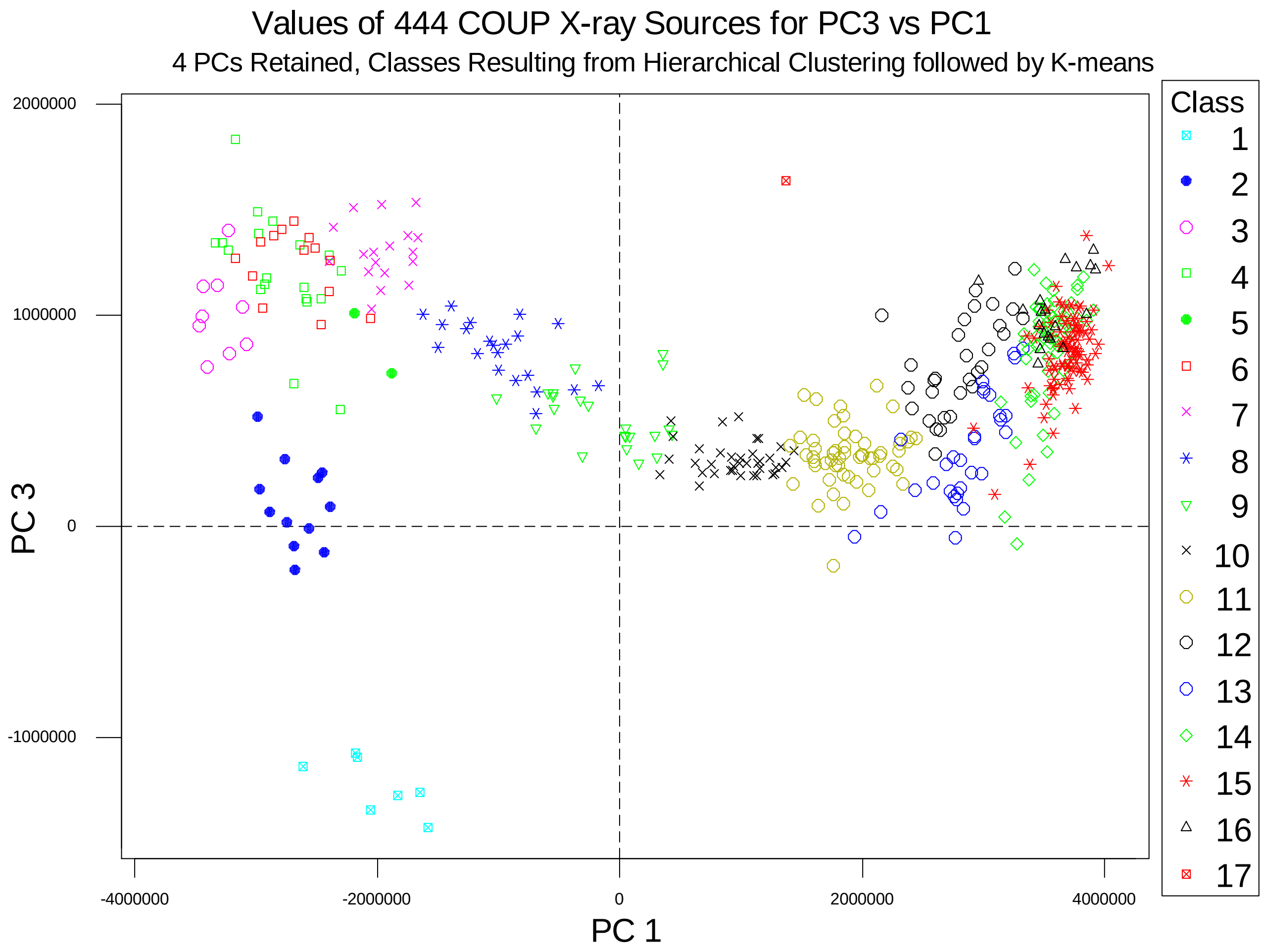


Figure 9.3: Plot of principal components 3 versus 1 with source classes color-coded.

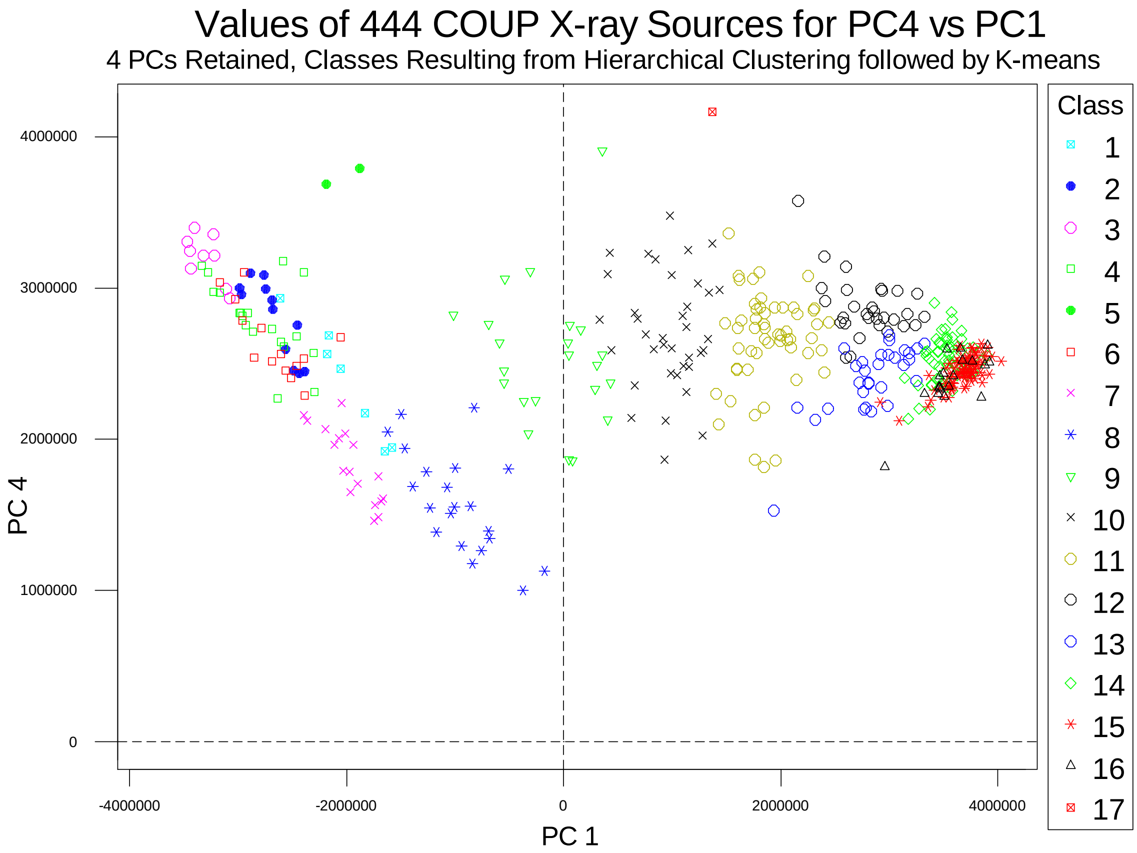


Figure 9.4: Plot of principal components 4 versus 1 with source classes color-coded.

These principal component plots show how, for the most part, the same objects appear in the same clusters in more than one of the plots. Also the outliers and the tightly clustered groups are consistent across the six plots. This is to be expected, since these first four principal components were used to create the cluster assignments and also the plots.

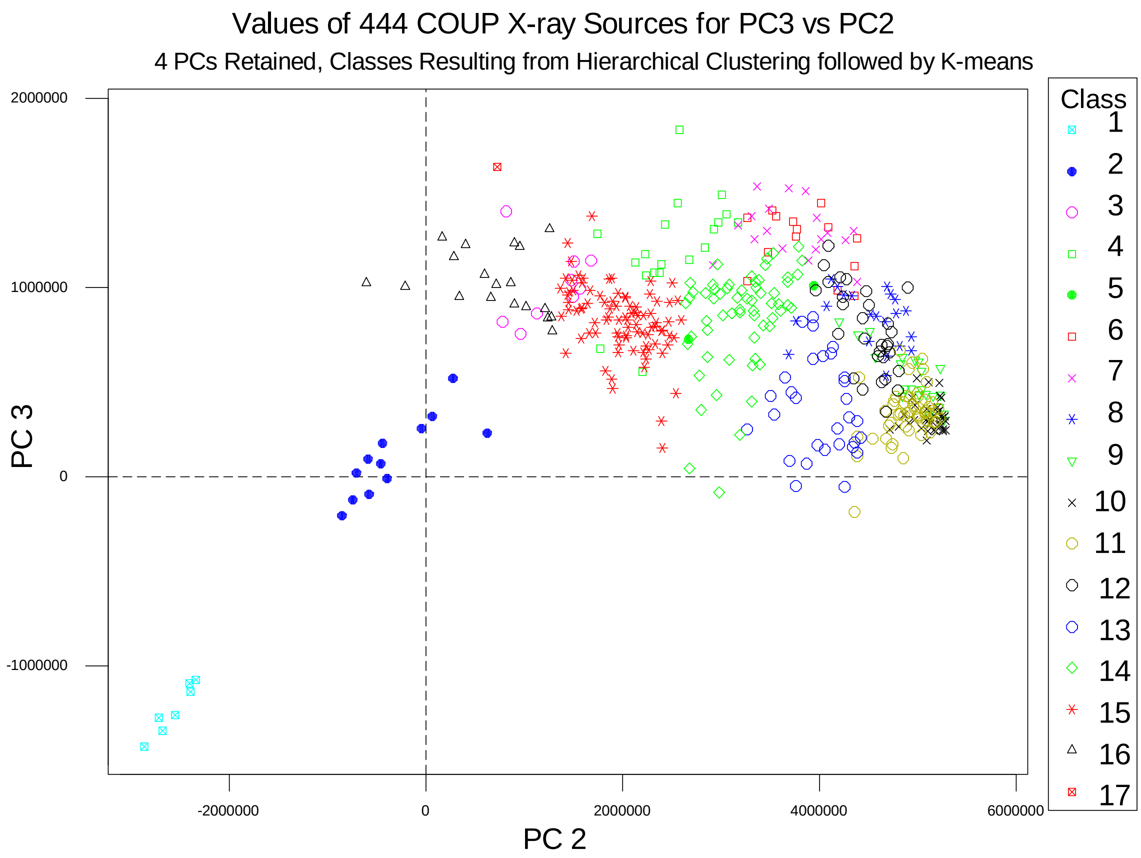


Figure 9.5: Plot of principal components 3 versus 2 with source classes color-coded.

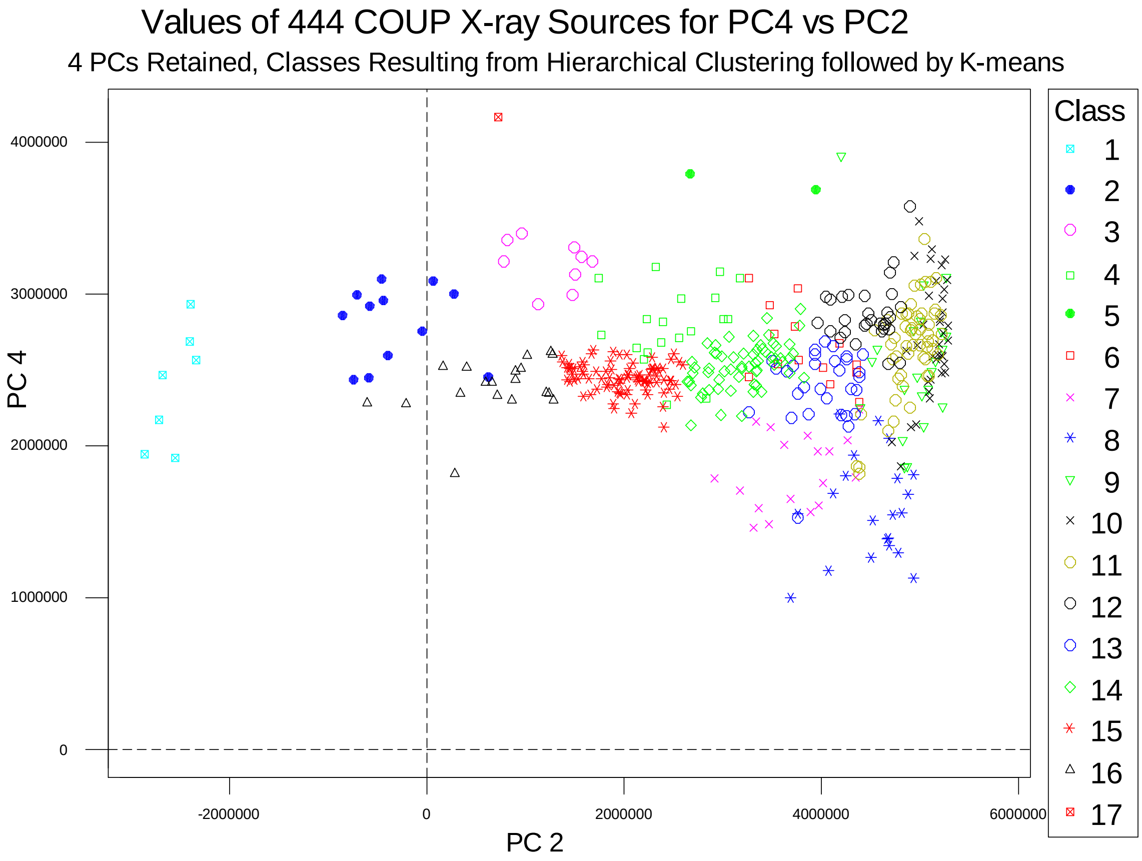


Figure 9.6: Plot of principal components 4 versus 2 with source classes color-coded.

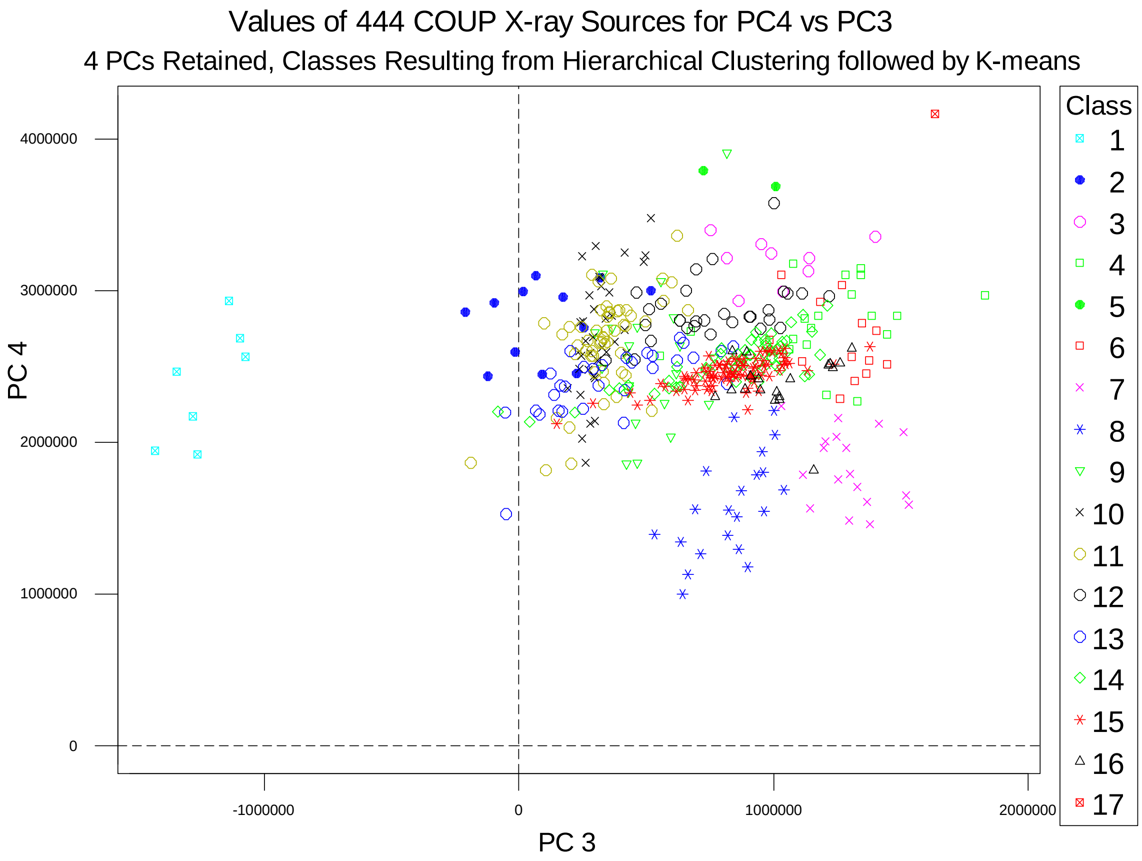


Figure 9.7: Plot of principal components 4 versus 3 with source classes color-coded.

Examples of sources drawn randomly from one of the lowered-numbered classes (Class 2) and one of the higher numbered classes (Class 14) are shown for comparison purposes in Figure 9.8 and Figure 9.9, respectively. These spectra are plotted on a log-log scale. The sources in Class 2 have harder spectra than the sources in Class 14. From visual inspection, it can be seen that there are strong spectral similarities within a given class. Also, strong fundamental spectral differences can be seen between the classes.

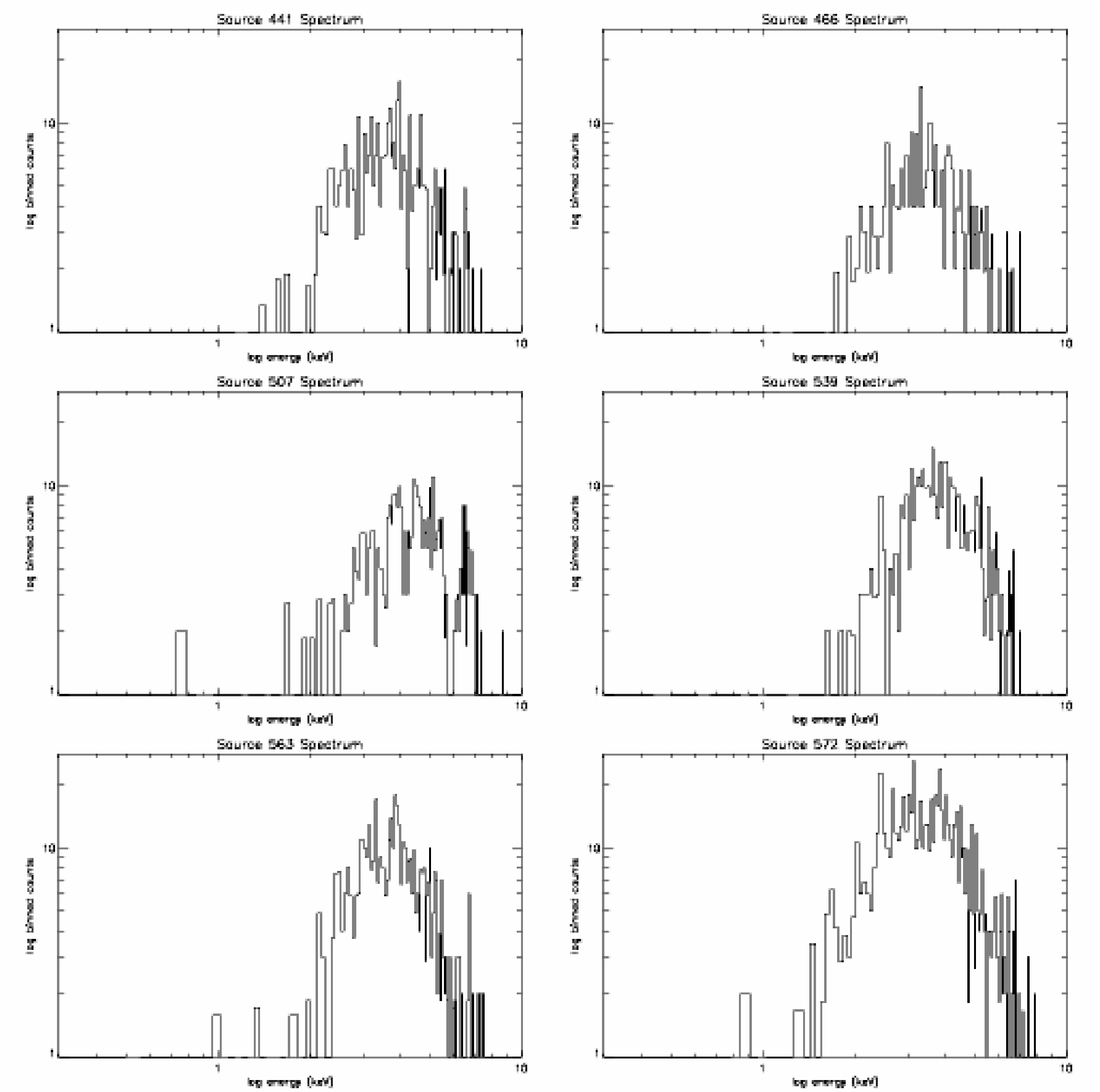


Figure 9.8: Six example sources from Class 2.

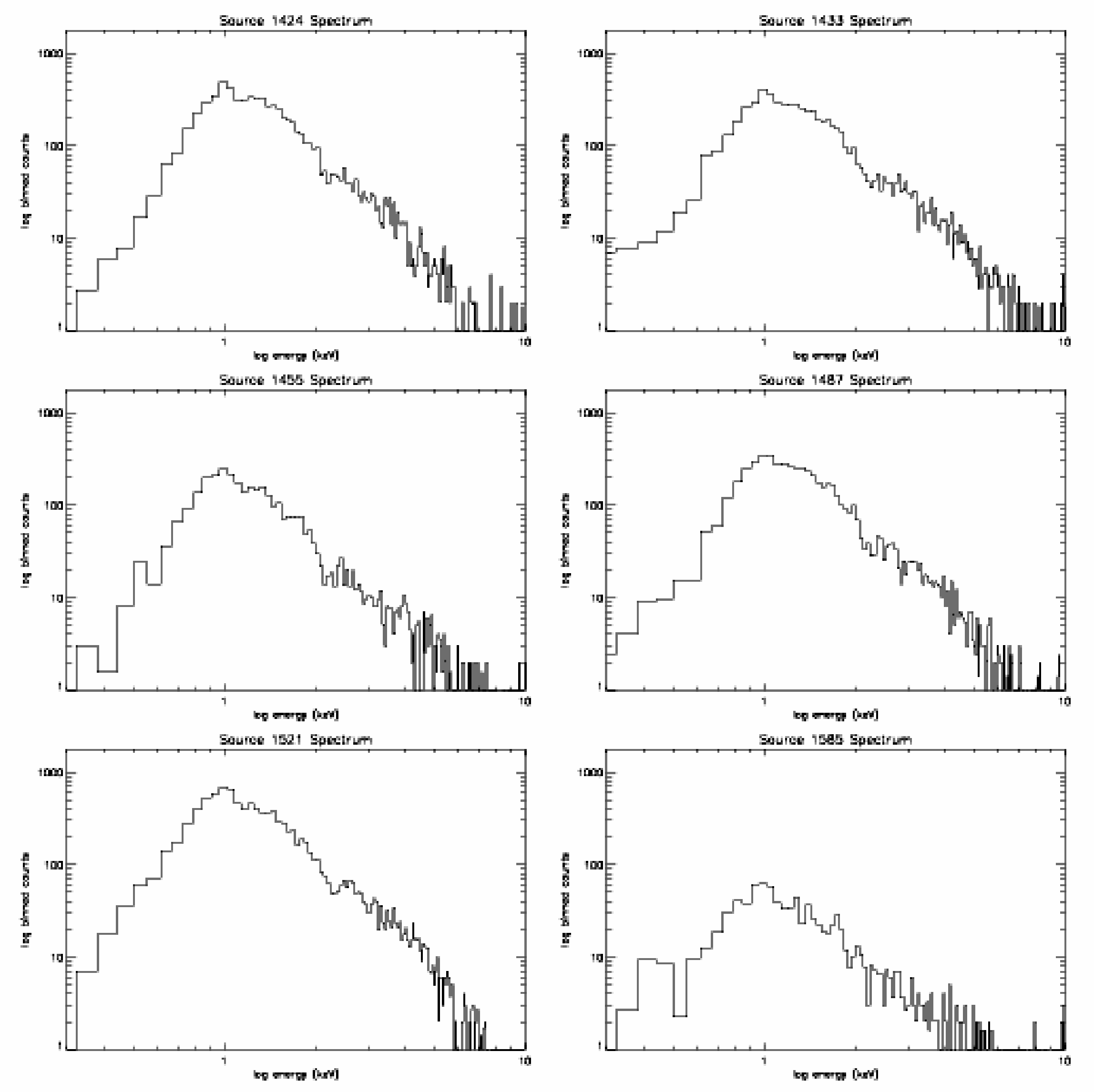


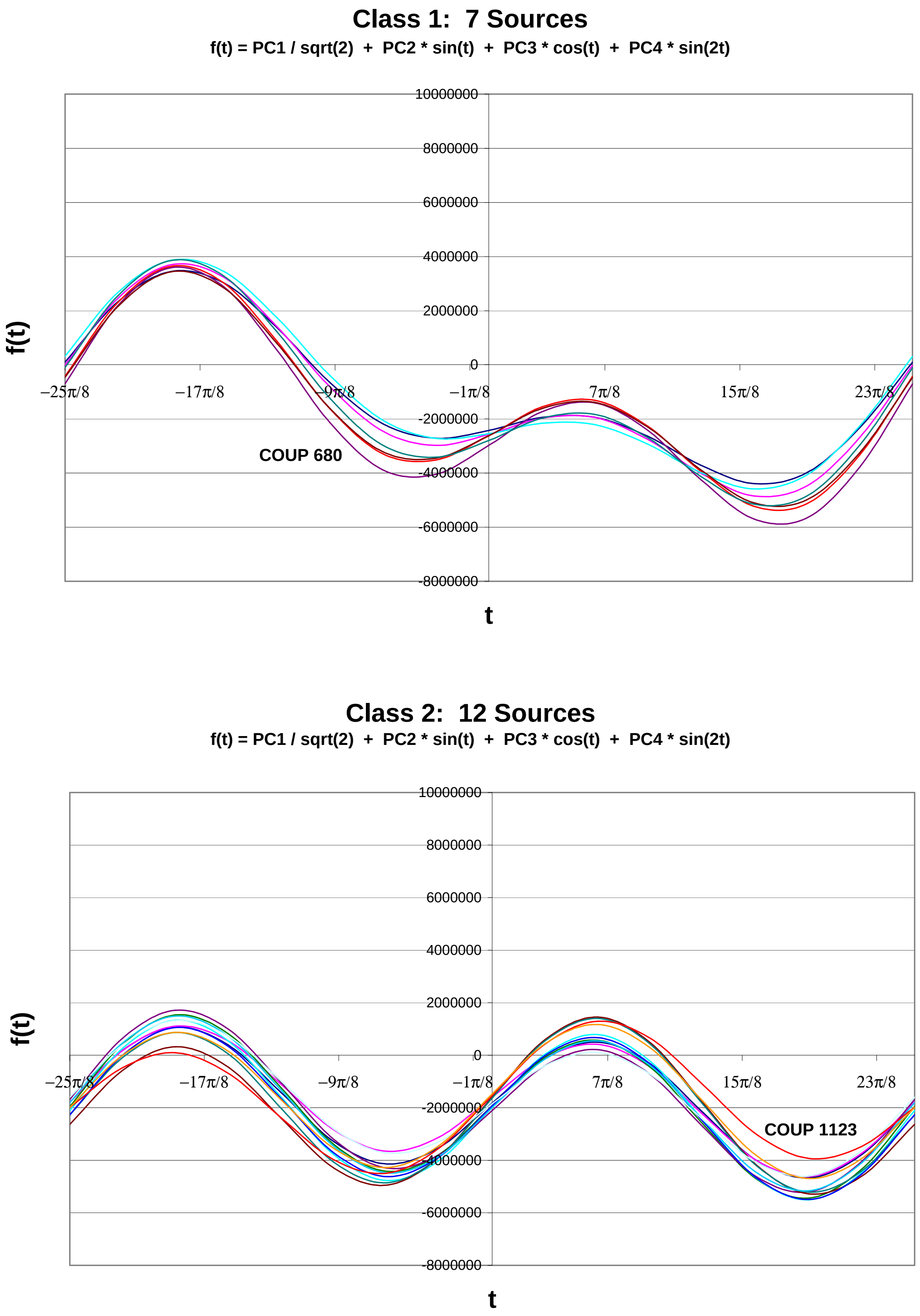
Figure 9.9: Six example sources from Class 14.

9.2 Class Homogeneity

Homogeneity of the classes was checked by plotting Andrews' curves[87]. The curves are based on a Fourier series representation. For each source, the following curve has been plotted:

$$f(t) = \frac{y_1}{\sqrt{2}} + y_2 \sin(t) + y_3 \cos(t) + y_4 \sin(2t)$$

where y_1, y_2, y_3, and y_4 are values of the first four PCs for the source being plotted. The curve is defined for the range of t from -π to π, inclusive. These profiles of the data preserve the "distance" between objects[88]. Andrews' curves were plotted separately for each of the X-ray source classes. Figure 9.10 shows the curves for the 17 classes. It can be seen immediately that different classes have different amplitude and/or different shaped curves, showing the variation between the classes. Within a class, the curves fall into fairly tight, narrow bands. Narrower bands of curves for a particular class imply greater homogeneity for that class[89]. Overall, the curves are tight for each class. Some of the classes contain sources with curves that stray a small amount from the main group of curves for that class. Also, the values for some of the curves in the intermediate-numbered classes overlap, meaning a source could potentially have been placed into the preceding class or the subsequent class. However, the shape of the curve still differs, especially the curvature near π and –π.

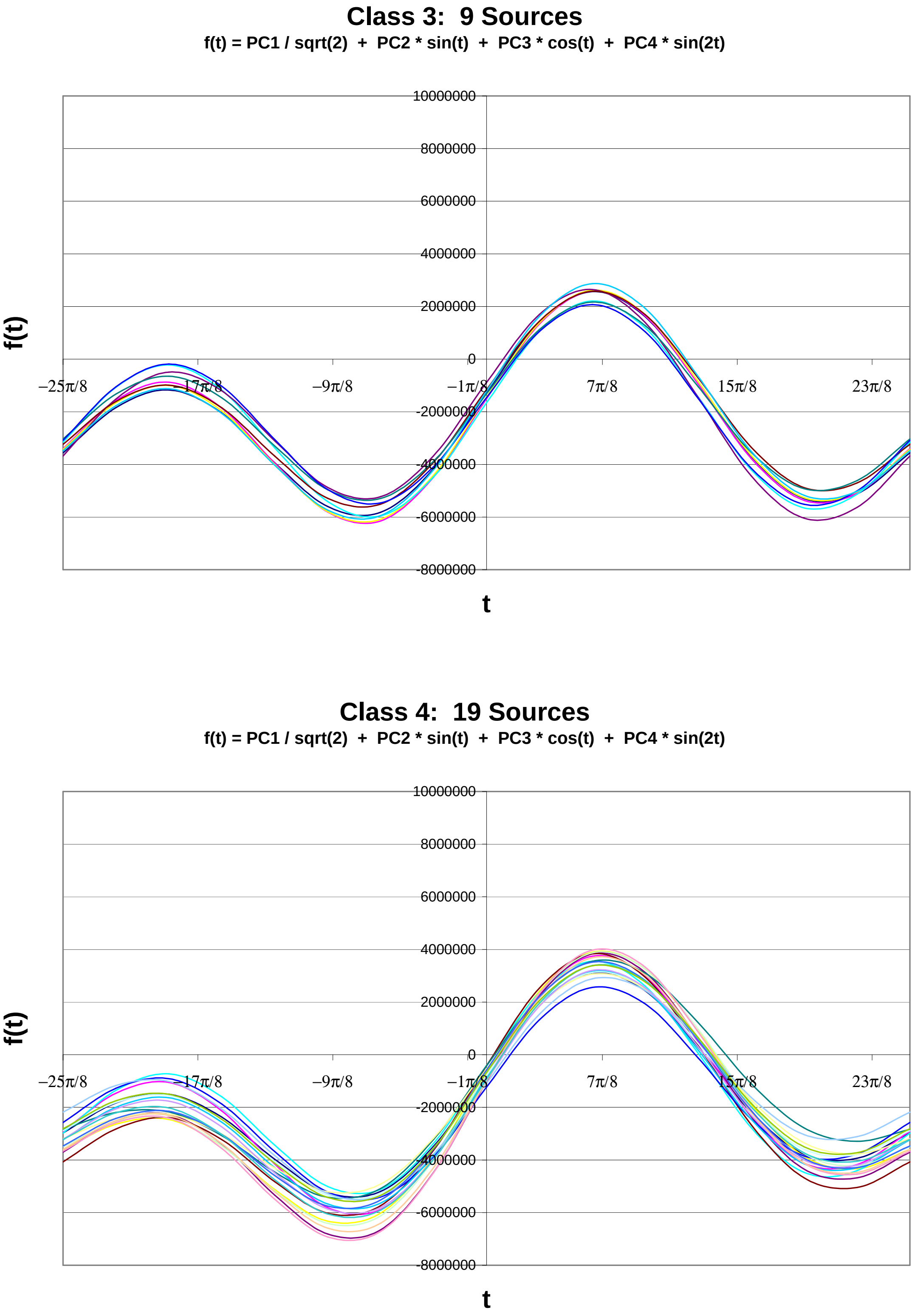

Class 1: 7 Sources

f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

Class 2: 12 Sources

f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

Figure 9.10: Andrews' curves for the 17 classes resulting from the clustering algorithm.



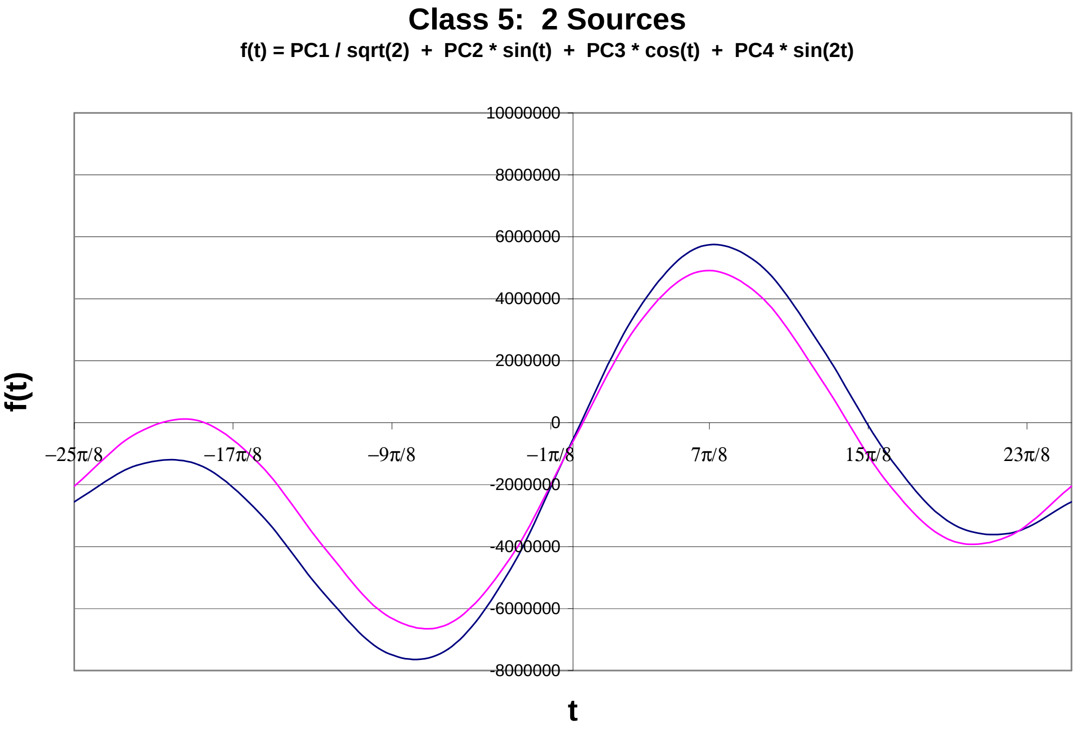
Class 3: 9 Sources

f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

Class 4: 19 Sources

f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

Figure 9.10 (cont.)



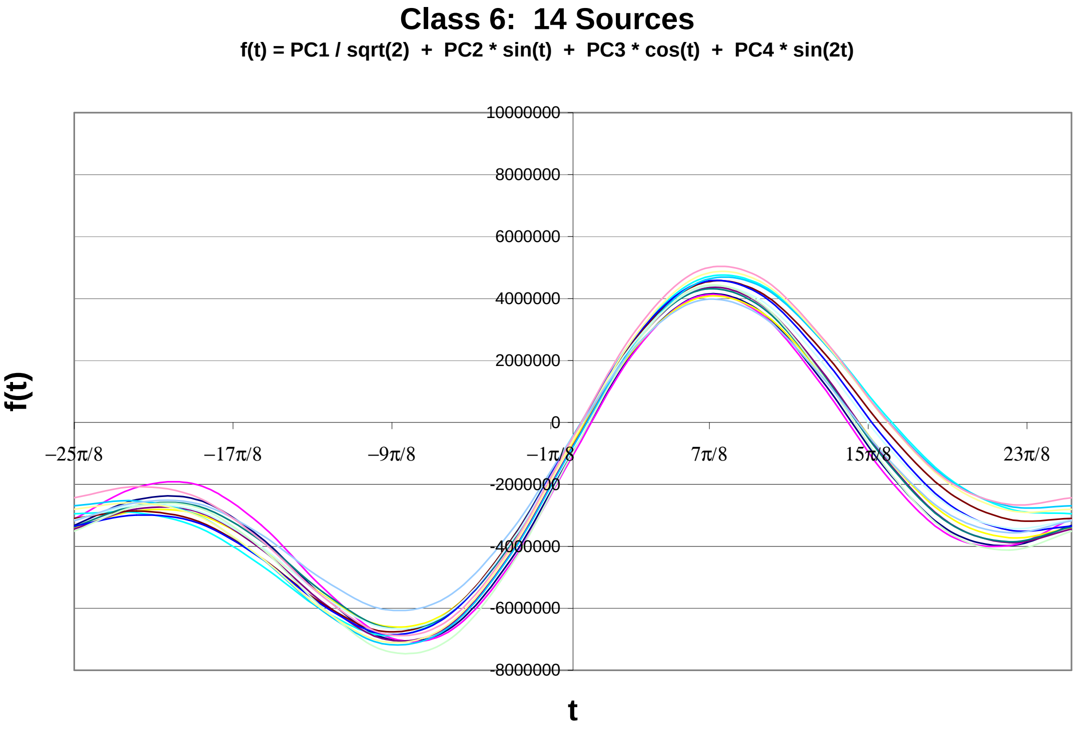
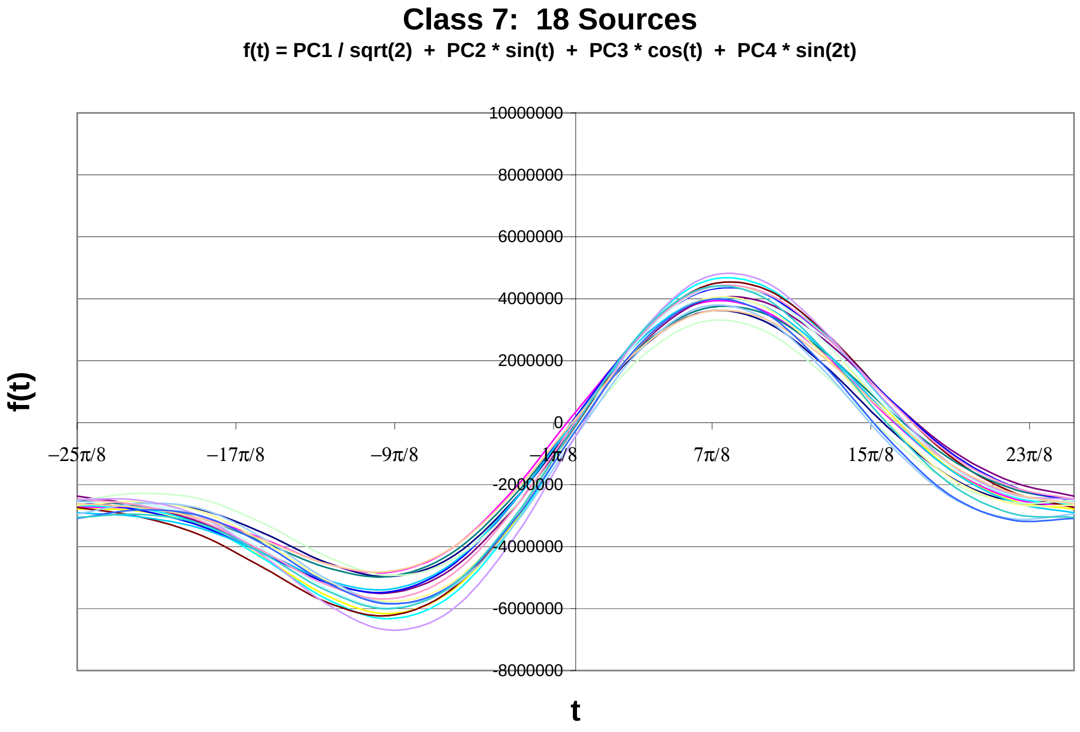
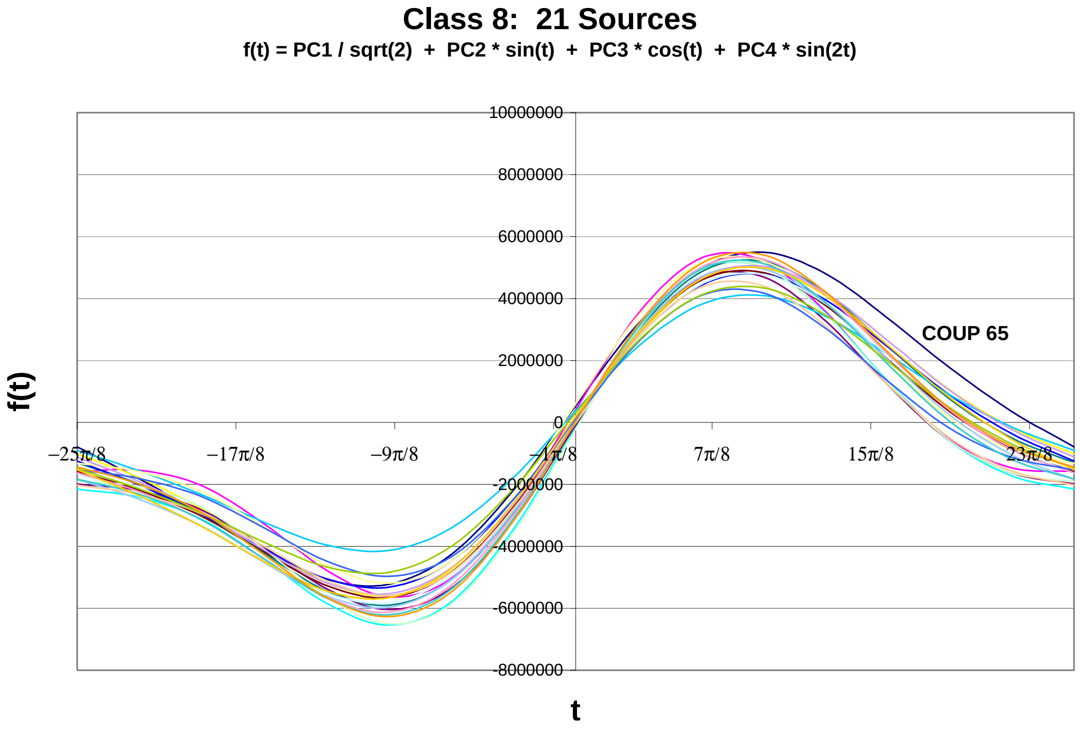


Figure 9.10 (cont.)



Class 7: 18 Sources

f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)



Class 8: 21 Sources

f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

Figure 9.10 (cont.)

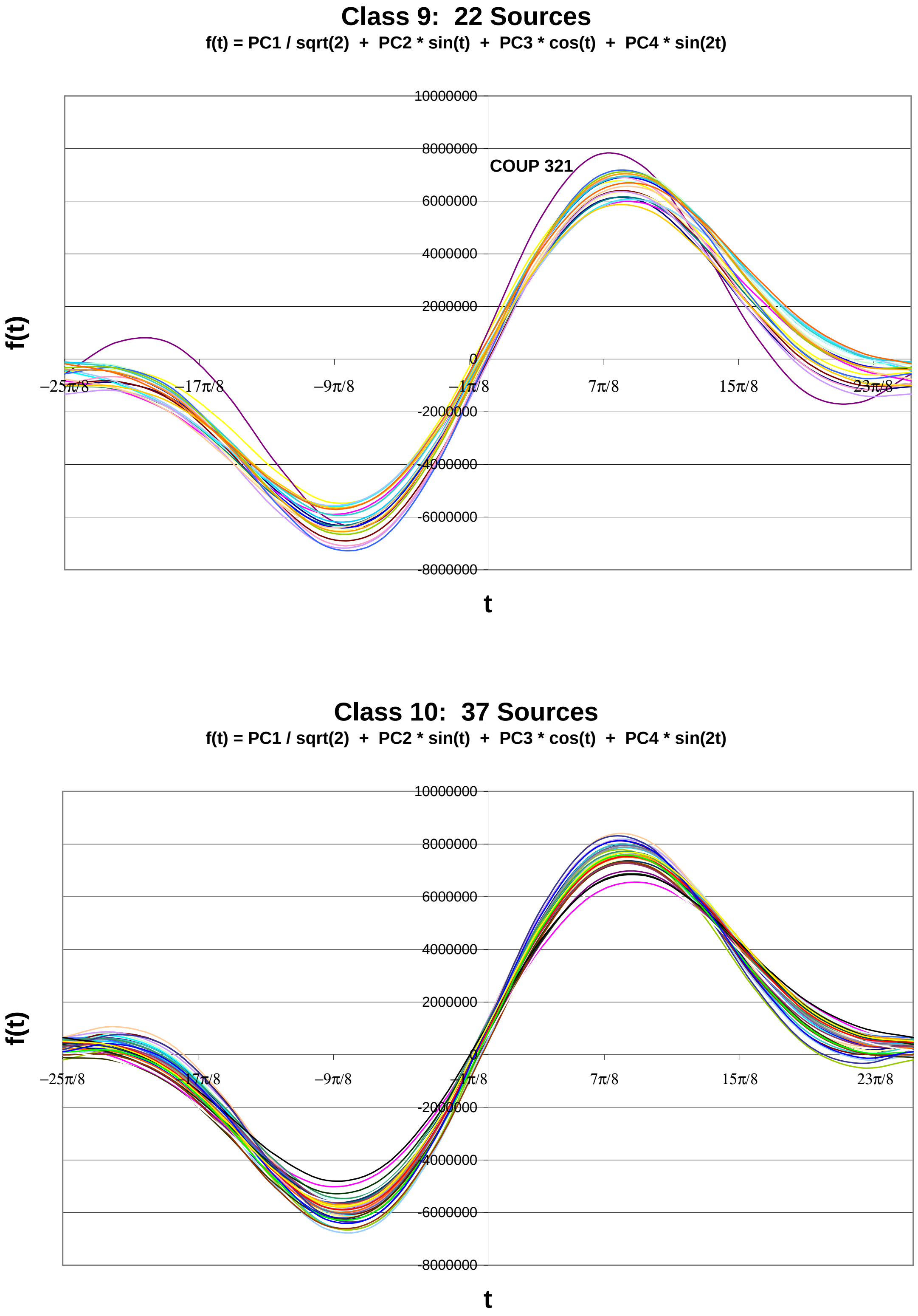
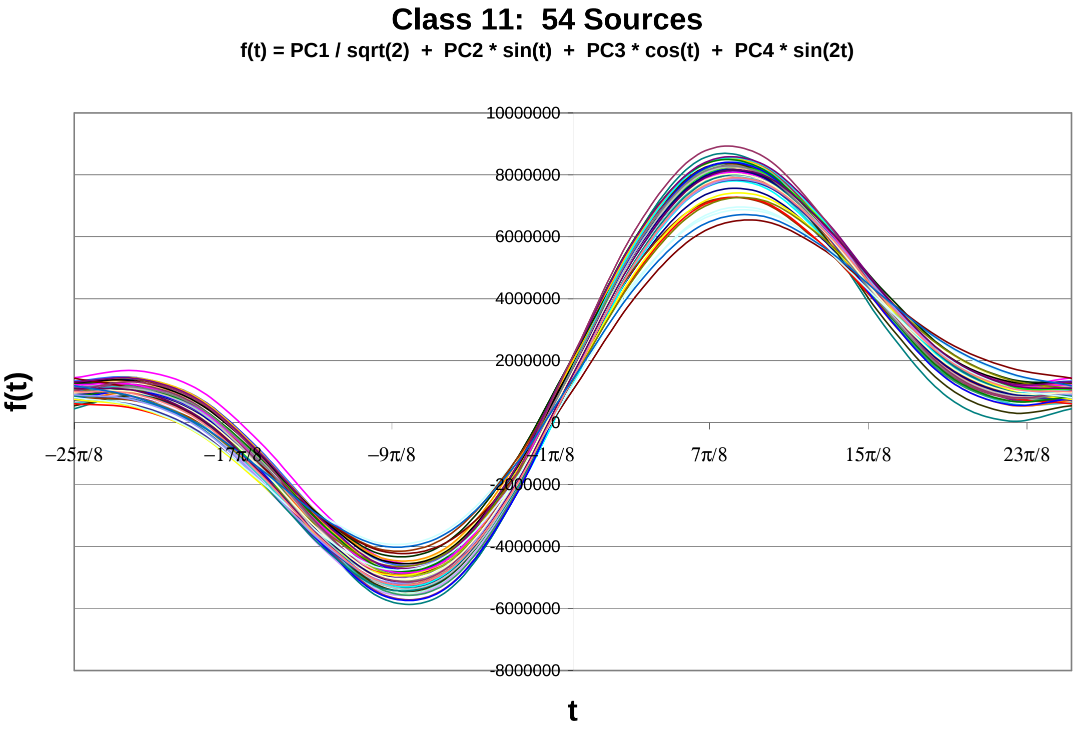


Figure 9.10 (cont.)



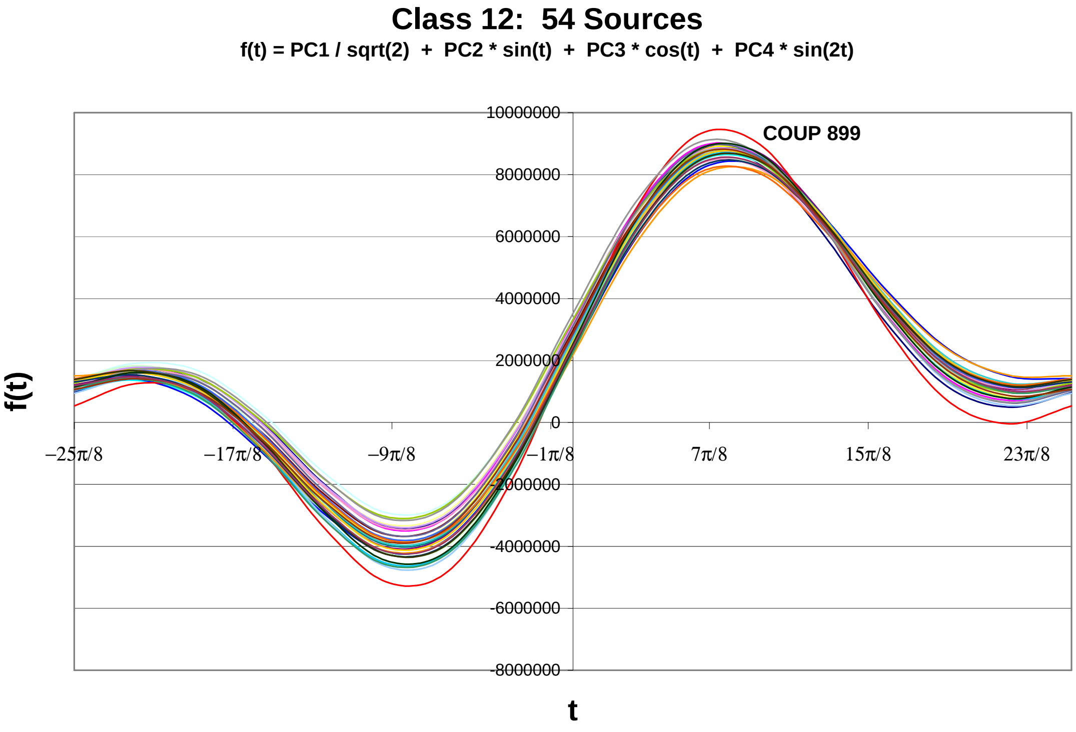
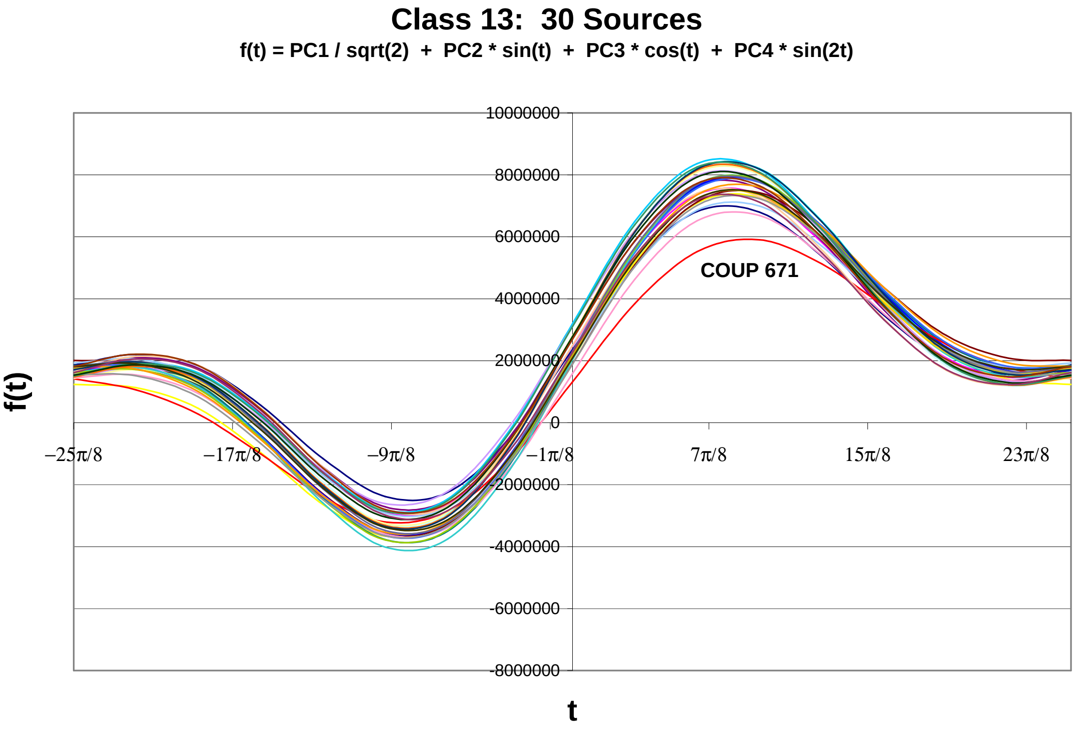


Figure 9.10 (cont.)



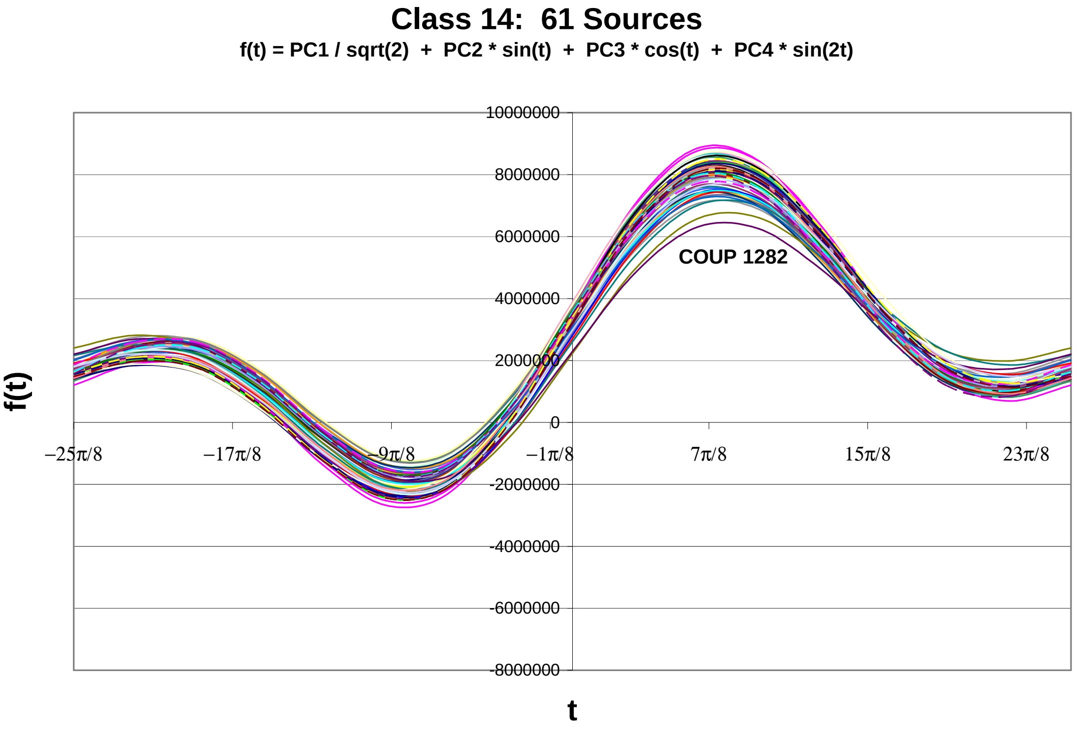
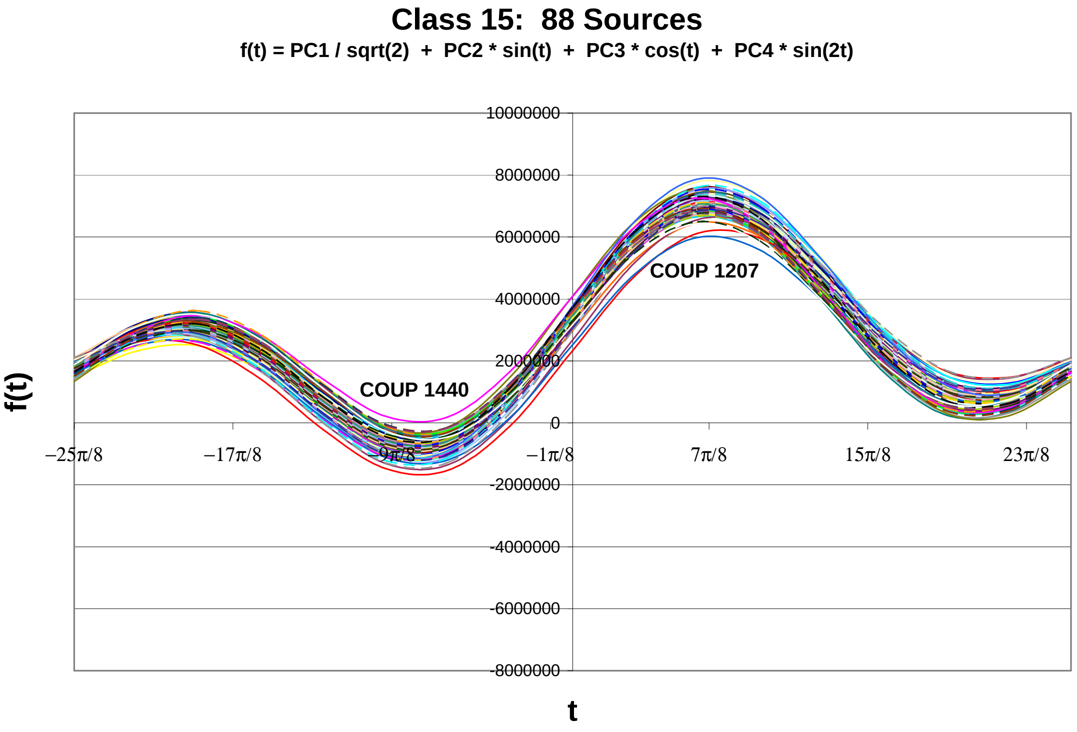


Figure 9.10 (cont.)



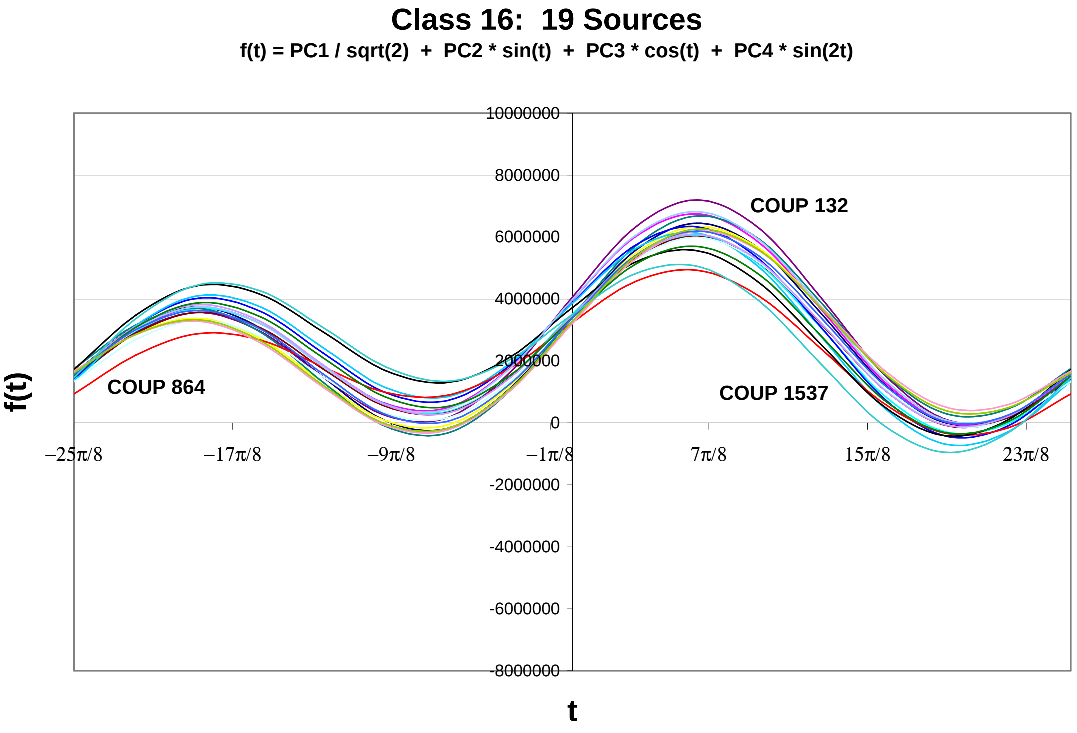
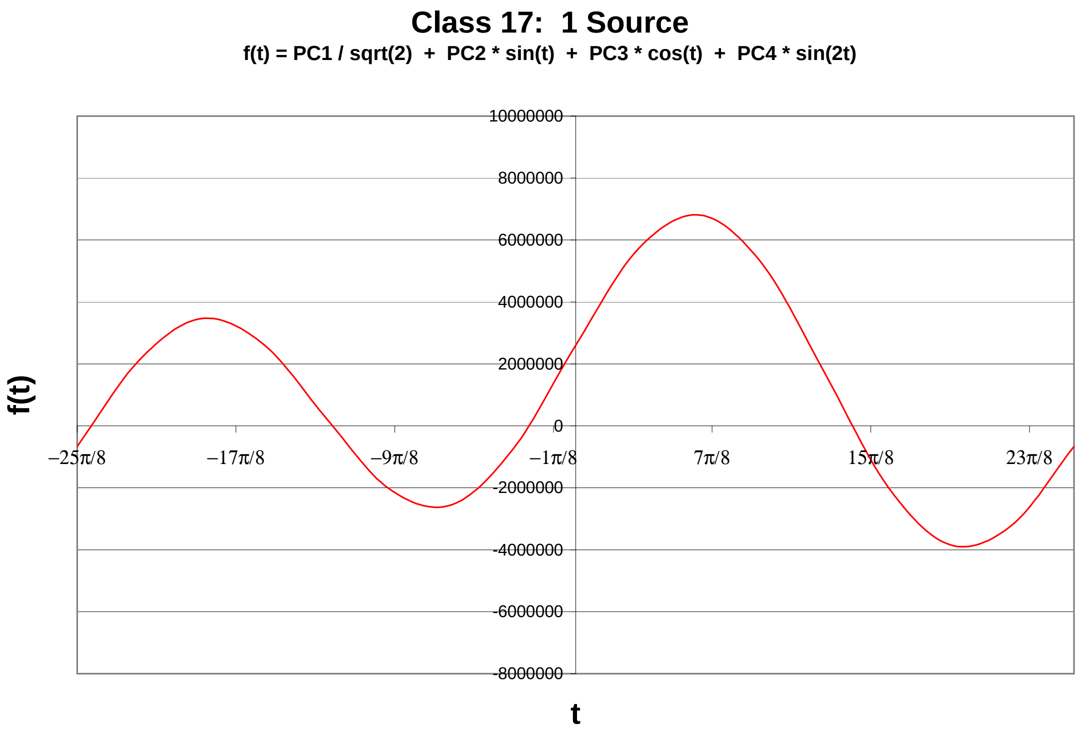


Figure 9.10 (cont.)



Class 17: 1 Source

f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

Figure 9.10 (cont.)

9.3 Omission of Agglomerative Hierarchical Clustering Step

As part of the results analysis, K-means clustering was run again without running hierarchical clustering first, to determine whether the hierarchical clustering step improved the source groupings. The corresponding score plot, color-coded by class, is shown in Figure 9.11. In comparing this plot to Figure 9.2, it can be seen that the outlier in Class 17 has now been incorrectly placed within a large class of sources with soft spectra (see plot of the X-ray spectrum for Class 17 in Figure 9.1). Also, Classes 1 and 2 from the previous clustering algorithm have now been combined into one, less-homogeneous class consisting of, for example, sources with large flares versus sources without flares, and sources with prominent Fe K-α lines versus sources without prominent Fe K-α lines. Andrews' curves were plotted for Class 1 and Class 17 (see Figure 9.12). The new Class 1 plot appears to contain two groupings of curves, signifying that the class is not as homogeneous as the classes obtained from the first clustering algorithm that included agglomerative hierarchical clustering. In the new Class 17 plot, it is interesting to note that the peaks and valleys of the curve for COUP 948 are located at the same values of **t** as the other curves in the new Class 17 plot. However, the COUP 948 curve has much larger amplitude, indicating that this source does not fit well into the new Class 17. Overall, the curves in Figure 9.10 and Figure 9.12 suggest that homogeneity of the resulting classes is greater as a result of running agglomerative hierarchical clustering prior to K-means clustering.

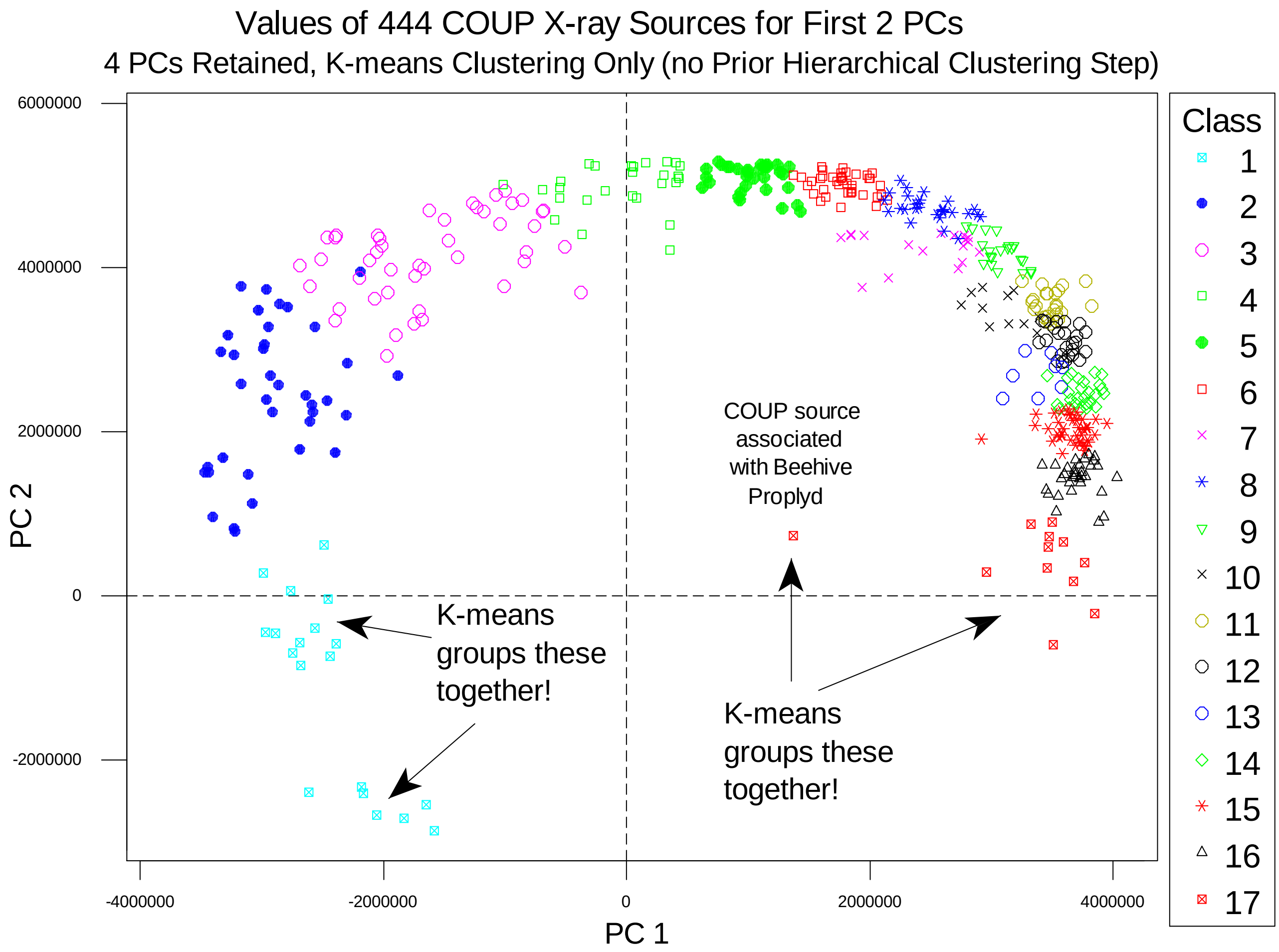
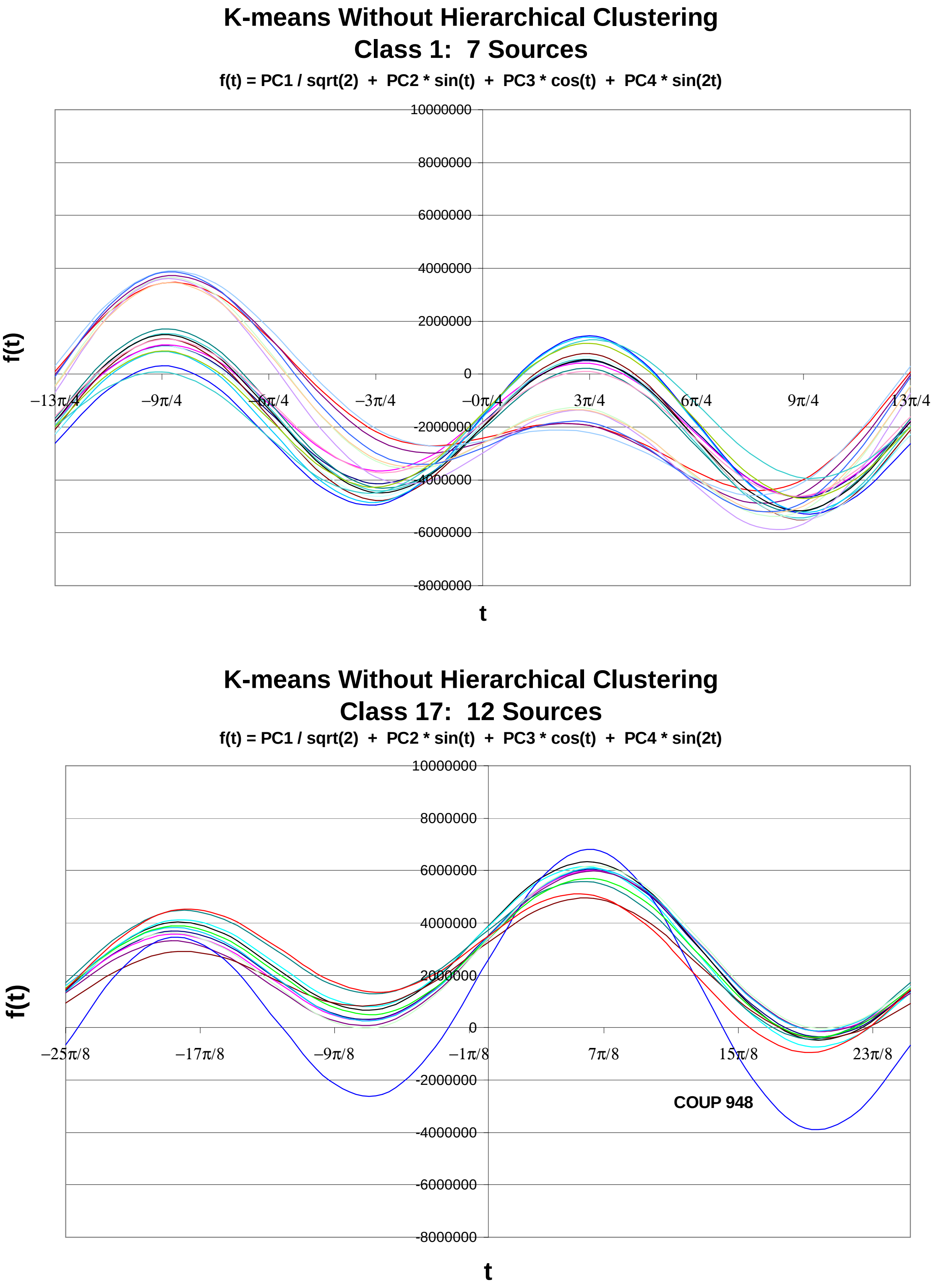


Figure 9.11: Results of running PCA followed by K-means clustering. Hierarchical clustering was not run prior to running K-means clustering.



K-means Without Hierarchical Clustering
Class 1: 7 Sources
f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

K-means Without Hierarchical Clustering
Class 17: 12 Sources
f(t) = PC1 / sqrt(2) + PC2 * sin(t) + PC3 * cos(t) + PC4 * sin(2t)

Figure 9.12: Andrews' curves for Classes 1 and 17 created from PCA followed by K-means clustering.

9.4 Hertzsprung-Russell Diagram

The Hertzsprung-Russell (H-R) diagram for the COUP 444 dataset is shown in Figure 9.13. The sources are color-coded with their X-ray spectral classes as determined by the X-ray source classification algorithm. The H-R diagram axes can be the optical stellar properties of luminosity or mass versus spectral type or (decreasing) effective photospheric temperature. The band of stars running from the upper left to the lower right of the H-R diagram is called the main sequence, where stars burn hydrogen in their cores[90]. For stars on the main sequence, the hotter the star is, the brighter it is. Stars located near the top of the diagram are more massive than stars at the bottom of the diagram, whether they are on the main sequence or not. There are six X-ray sources in the COUP 444 dataset that correspond to high mass A-type or B-type stars. These sources are labeled on the H-R diagram with their COUP source number. All six of these sources are found in X-ray spectral classes with softer spectra: classes 14, 15, and 16. Five of these sources, numbered 100, 113, 869, 1360, and 1415, have been included in a study by Stelzer et al. of the X-ray properties of O, B, and A stars[91]. For 4 out of these 5 sources (100, 113, 1360, and 1415), they conclude that the X-rays are being emitted by low-mass companions to the A-type and B-type stars found in optical spectroscopy. They base their conclusions on X-ray variability of these sources. For these 4 sources, my analysis shows the same conclusion from running my model-independent algorithm on the X-ray spectral data alone. For the fifth source (COUP 869), they studied the X-ray spectral and variability properties and concluded that the X-ray emission is from the massive B-type star itself. My X-ray source classification algorithm places this source, COUP 869, into class 16: the class with the lowest average mass and the softest spectra. The X-ray spectrum for COUP 869 (see Figure 9.14) has a similar spectral shape to the other members of X-ray spectral class 16 and also to the average class spectrum for class 16. The Andrews' curve for COUP 869 is within the group of Andrews' curves for class 16; it is not an outlier. This argues that X-ray emission from COUP 869 is coming from a low-mass companion to the high-mass B star.

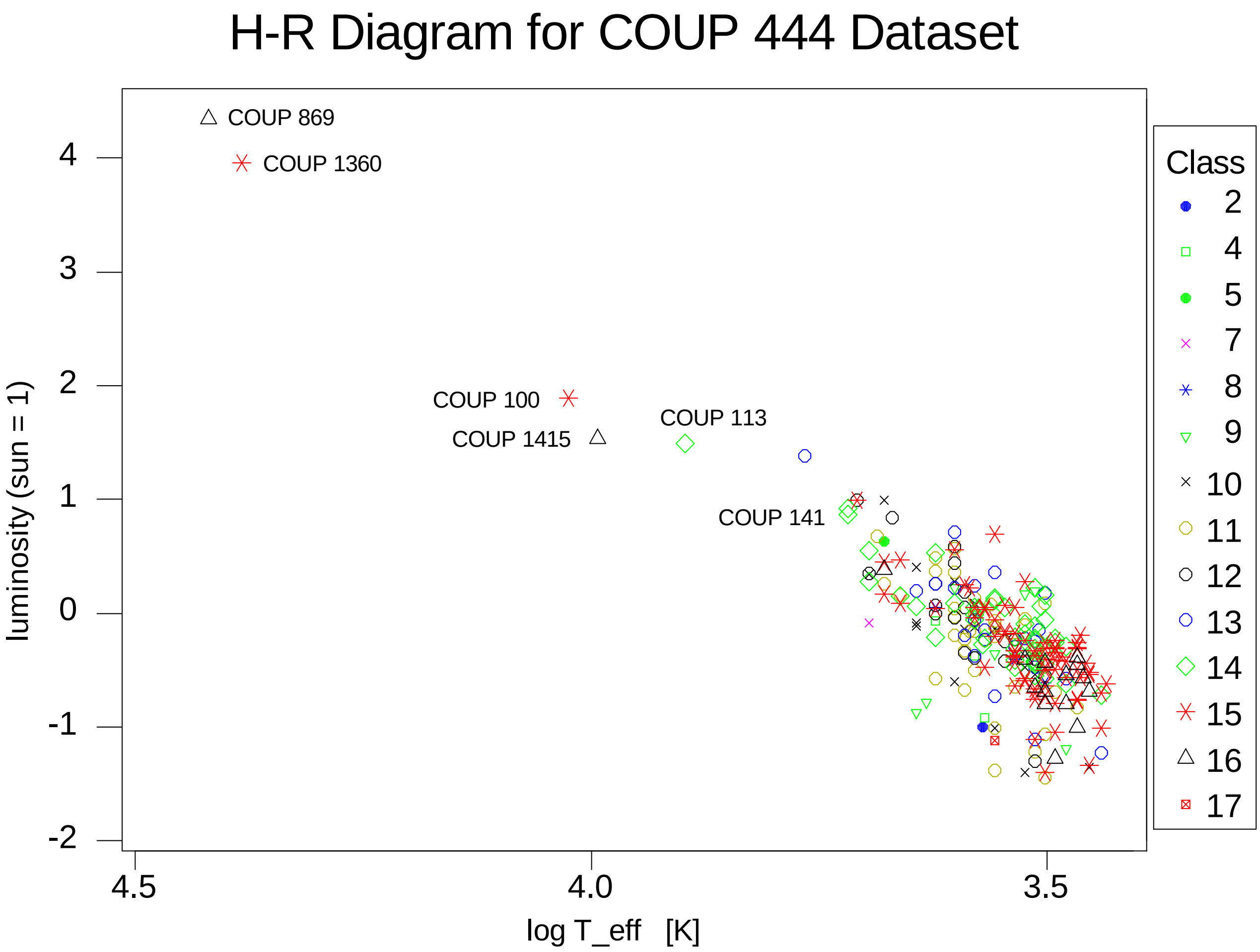


Figure 9.13: Hertzsprung-Russell diagram of COUP 444 dataset color-coded by X-ray spectral class. The A-type and B-type stars are labeled with their corresponding COUP source number.

The x-axis scale of the H-R diagrams in Figure 9.15 and Figure 9.16 was restricted to focus on the main group of sources, which are of spectral types K and M. The H-R diagram for classes 11 thru 13 is shown in Figure 9.15. The H-R diagram for classes 14 thru 16 is shown in Figure 9.16. These three classes appear to occupy slightly different regions in the H-R diagram. The sources in class 16 are clumped in the lower-right part of the main sequence. *These diagrams show a trend of increasing spectral softness with decreasing T_{eff} for X-ray sources in the ONC.*

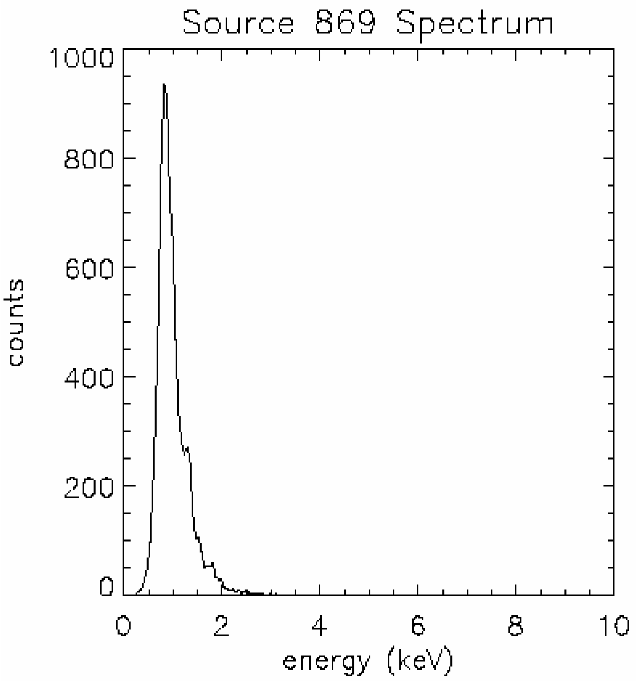


Figure 9.14: X-ray spectrum for COUP 869.

9.5 X-ray Properties Versus ONIR Properties

The source spectra, as well as the ONIR properties for the classes obtained from the K-means clustering algorithm were examined to assess the algorithm's ability to identify groups of sources that share common attributes. Table 9.1 lists the mean values for hydrogen column density (N_H), effective photospheric temperature (T_{eff})[92], stellar mass[93], stellar age[93], visual extinction[92] (A_V), and Δ(I-K) near-infrared excess[94] of the ONIR counterparts of the members of the 17 X-ray classes. The numbers in parentheses in Table 9.1 are the errors on the mean. These results were compiled from data available for the X-ray-emitting ONC population[12]. A-type and B-type stars were not included in the mean calculations based on optically-derived properties (i.e., T_{eff}).

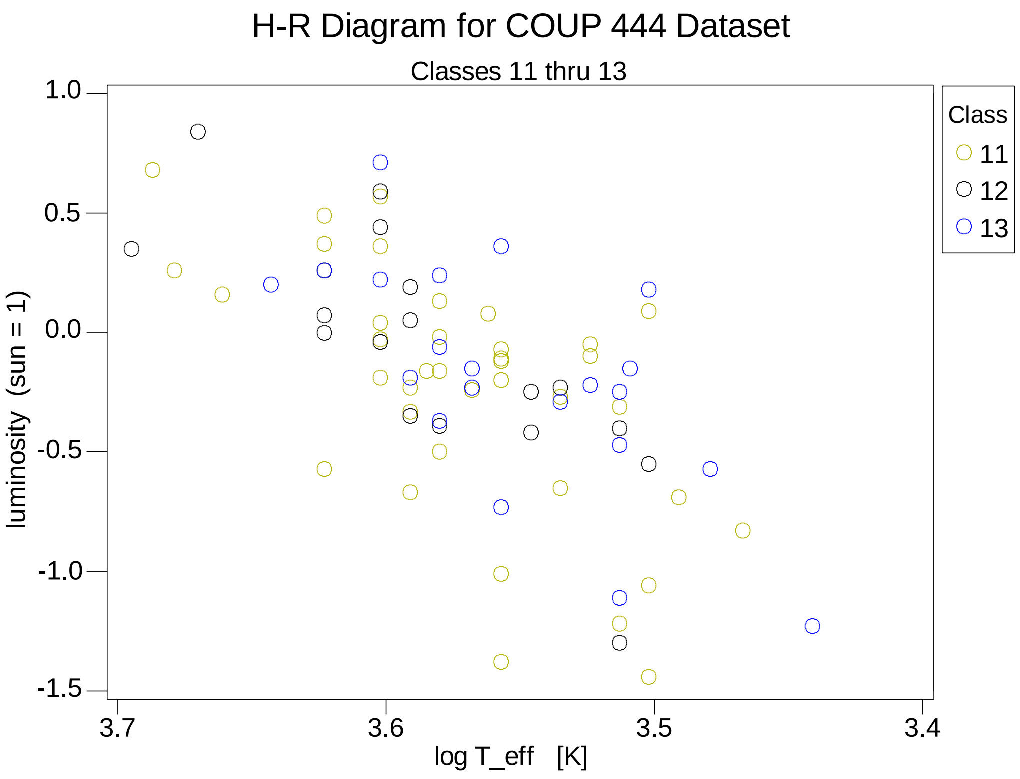


Figure 9.15: Hertzsprung-Russell diagram for soft X-ray spectrum classes 11, 12, and 13.

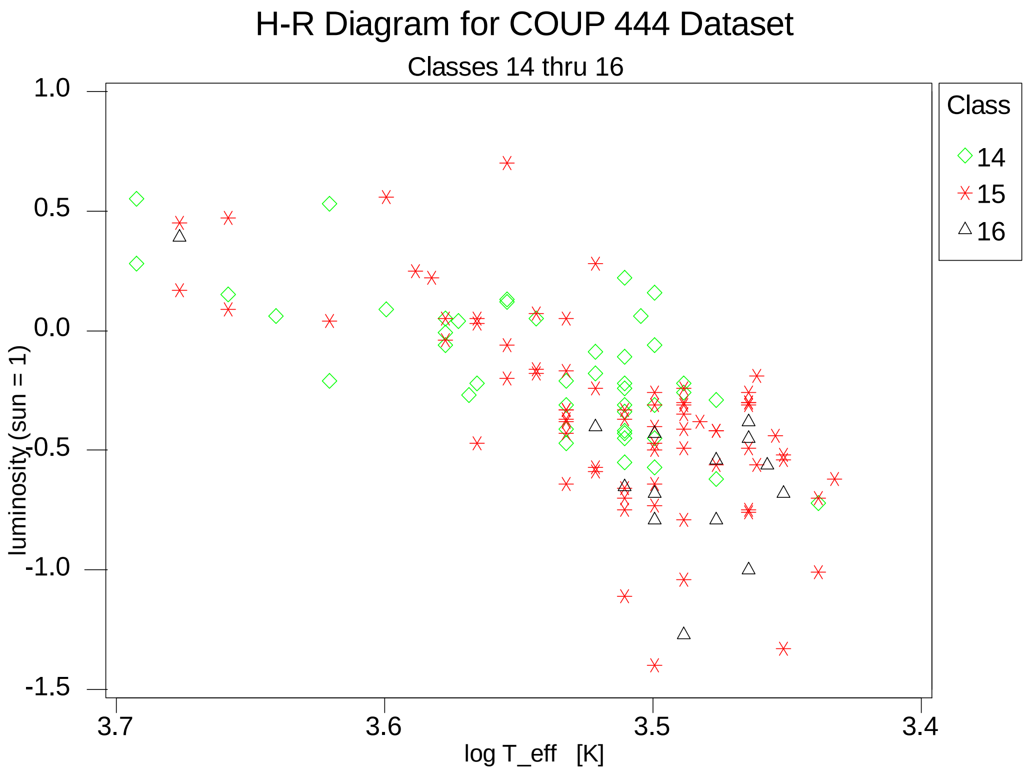


Figure 9.16: Hertzsprung-Russell diagram for the softest X-ray spectral classes: 14, 15, and 16.

N_H decreases monotonically from class 1 to class 16 (Figure 9.17). The large N_H characteristic of classes 1 through 8 is reflected in small fractions of ONIR counterparts. The mean visual extinction is observed to decrease monotonically for the classes 11 to 16 (Figure 9.18). The mean near-infrared excess is observed to decrease monotonically for the soft spectra classes 10 to 16 (Figure 9.19), suggesting a generally decreasing accretion rate. For classes 12 through 16, which have relatively large fractions of ONIR counterparts and softer X-ray emission, the mean T_{eff} decreases as the X-ray spectra gets softer (Figure 9.20). This was also shown by the H-R diagrams in section 9.4.

The stellar mass and stellar age decrease almost monotonically with increasing spectral softness for classes 10 thru 16. However, these properties are determined by comparing the source's T_{eff} and luminosity with evolution models of young stars. Mass depends directly on T_{eff} and age depends directly on luminosity and at the same time are highly model-dependent and therefore potentially uncertain.

Classes form sequences in N_H, A_V, near-IR K-band excess, stellar mass, and stellar age demonstrating that the algorithm efficiently sorts young stars into physically meaningful groups. These trends show a strong correlation between X-ray and ONIR properties of young stars in the ONC.

Table 9.1: ONIR properties of the resulting 17 X-ray classes. Values in parentheses represent error on the mean. The six A-type and B-type stars in the COUP 444 dataset have not been included in mean calculations based on optically-derived properties.

Class	Number of Sources	< log N_H > [cm^{-2}]	N	<logT_{eff} > [K]	N	Mass [solMass]	N	Age	N	< A_V > mag	N	< Δ(I-K) > Mag	N
1	7	23.40 (0.06)	7		0						0		0
2	12	22.96 (0.03)	12	3.57 :	1	0.47 :	1	7.21 :	1		0		0
3	9	22.79 (0.02)	9		0						0		0
4	19	22.66 (0.01)	19	3.59 :	3	0.64 :	3	6.67 :	3	1.34 :	3	1.66 :	3
5	2	22.52 (0.05)	2	3.68 :	1	1.91 :	1	6.27 :	1	3.67 :	1	2.61 :	1
6	14	22.48 (0.02)	14		0						0		0
7	18	22.46 (0.02)	18	3.70 :	1	1.10 :	1	7.28 :	1	3.52 :	1	0.98 :	1
8	21	22.30 (0.02)	21	3.55 :	3	0.49 :	3	6.19 :	3	1.52 :	3	0.30 :	2
9	22	22.18 (0.01)	22	3.56 (0.02)	7	0.41 (0.10)	6	5.99 (0.49)	6	1.77 (0.99)	7	1.10 :	3
10	37	22.03 (0.02)	37	3.58 (0.01)	21	0.73 (0.12)	19	6.34 (0.10)	19	2.60 (0.45)	20	1.31 (0.18)	18
11	54	21.90 (0.02)	54	3.57 (0.01)	38	0.57 (0.06)	38	6.23 (0.09)	38	2.69 (0.31)	38	0.91 (0.12)	30
12	30	21.66 (0.03)	30	3.59 (0.01)	20	0.79 (0.14)	19	6.20 (0.07)	19	1.57 (0.29)	19	0.80 (0.14)	16
13	30	21.61 (0.03)	30	3.56 (0.01)	22	0.58 (0.12)	21	5.95 (0.12)	21	1.44 (0.27)	22	0.62 (0.11)	18
14	61	21.32 (0.03)	61	3.55 (0.01)	45	0.51 (0.07)	43	5.88 (0.08)	43	1.16 (0.16)	44	0.49 (0.08)	38
15	88	20.79 (0.05)	86	3.52 (0.01)	75	0.39 (0.05)	71	5.80 (0.07)	70	0.65 (0.11)	72	0.25 (0.07)	62
16	19	20.28 (0.11)	19	3.50 (0.01)	14	0.29 (0.11)	14	5.95 (0.15)	14	0.32 (0.14)	16	0.11 (0.05)	12
17	1	20.88 :	1	3.56 :	1	0.39 :	1	7.21 :	1	0.34	1		0

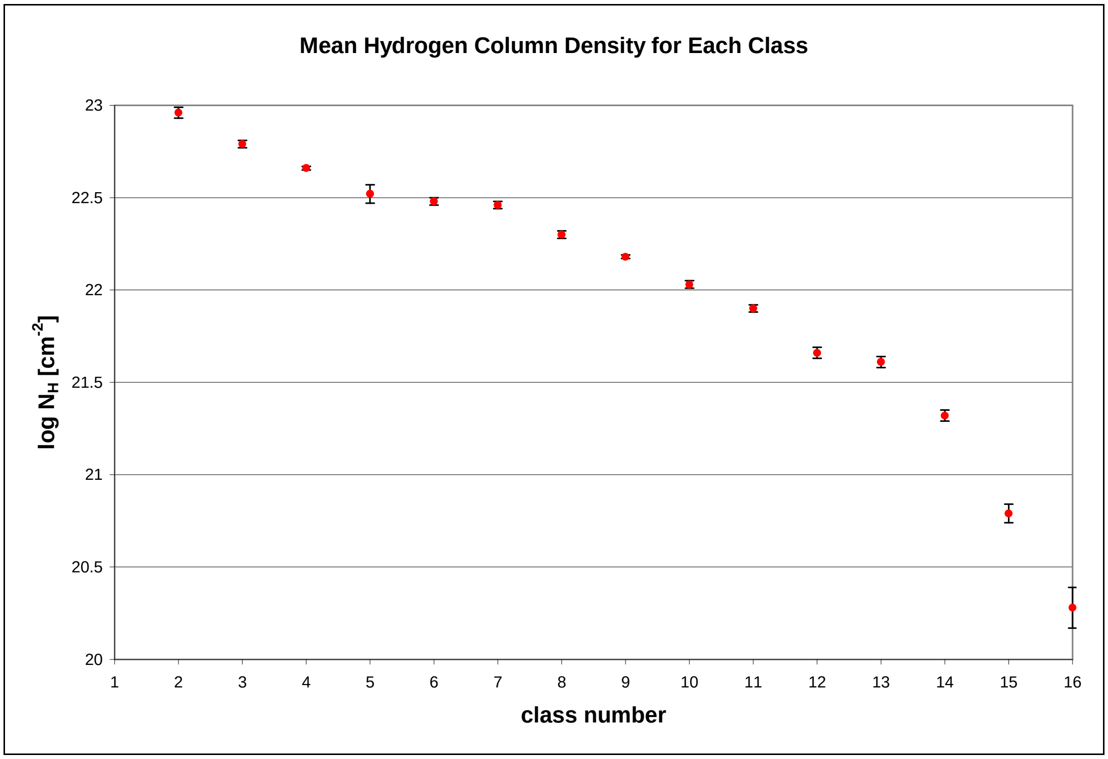


Figure 9.17: Mean hydrogen column density plotted for each class.

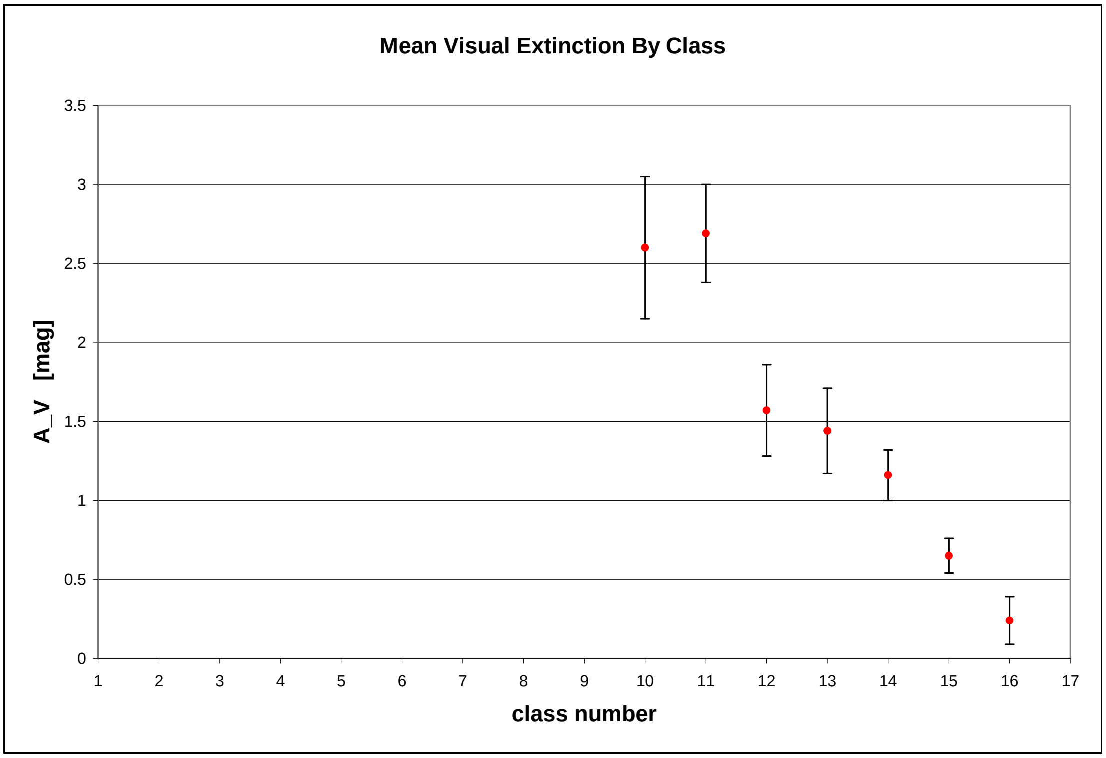


Figure 9.18: Mean visual extinction plotted by class.

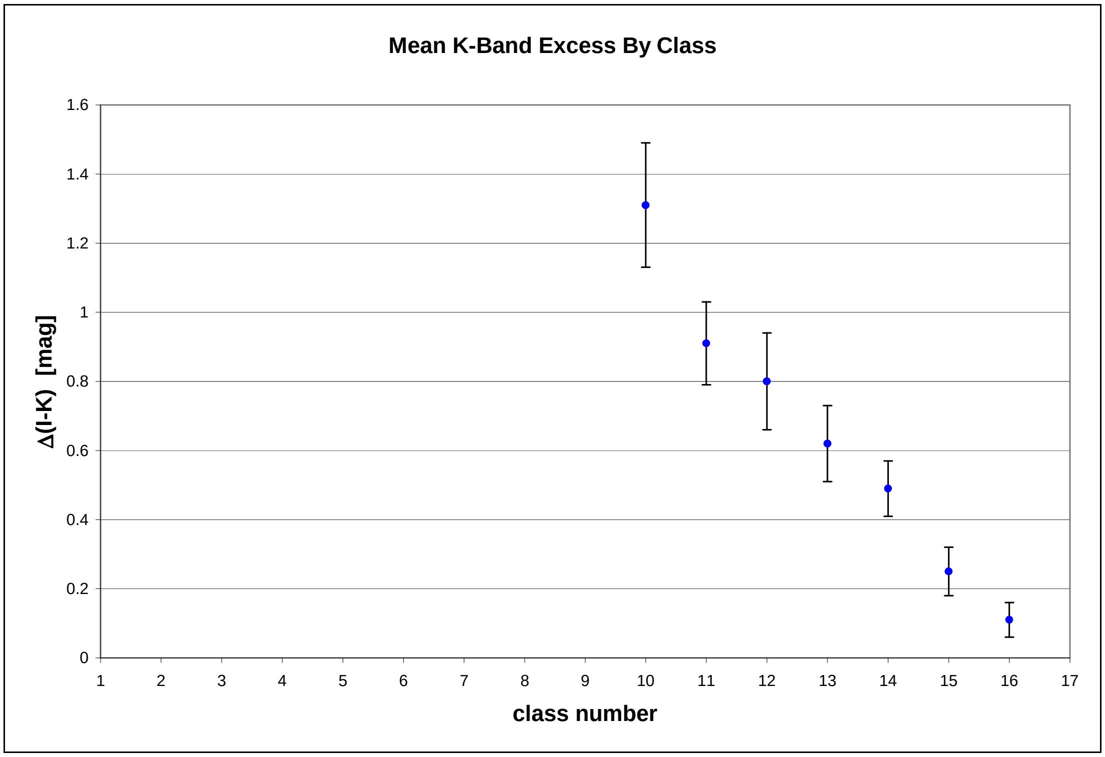


Figure 9.19: Mean near-IR K-band excess plotted by class.

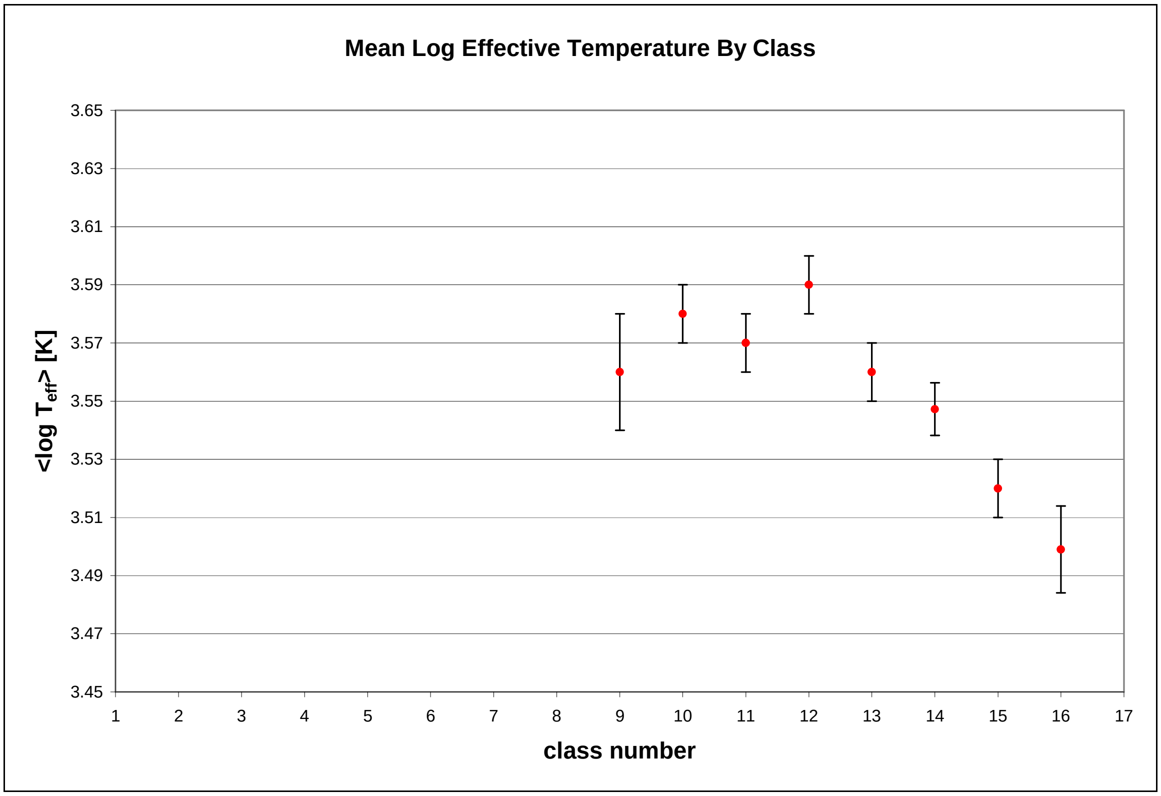


Figure 9.20: Mean log effective photospheric temperature plotted by class.

9.6 Very Deeply Embedded Protostars

Sources in class 1 are easily identified as a distinct X-ray spectral group by my classification technique. These sources lack ONIR counterparts and all have high-amplitude fast-rise X-ray flares[68]. An analogous sample of deeply embedded, flaring protostars is discussed in Tsujimoto et al.[95]. The strong Fe line emission at 6.4-6.7 keV seen in the Class 1 average spectrum attests to their high X-ray emission temperatures. These objects are likely very young protostars deeply embedded in the Orion Molecular Core.

9.7 Beehive Proplyd

COUP 948 is isolated by the source classification algorithm. It is associated with a jet source called the Beehive Proplyd (see Figure 9.21). It has an elliptical silhouette disk at the center and jets protruding along the minor axis of the ellipse[96]. The X-ray spectrum of COUP 948 has distinct hard and soft components with the soft component peaking at around 0.85 keV and the hard component with a main arc from 3.0 keV to 4.5 keV. This unique, double-peaked X-ray spectral distribution is indicative of strong shocks in the jet collimation region[96].

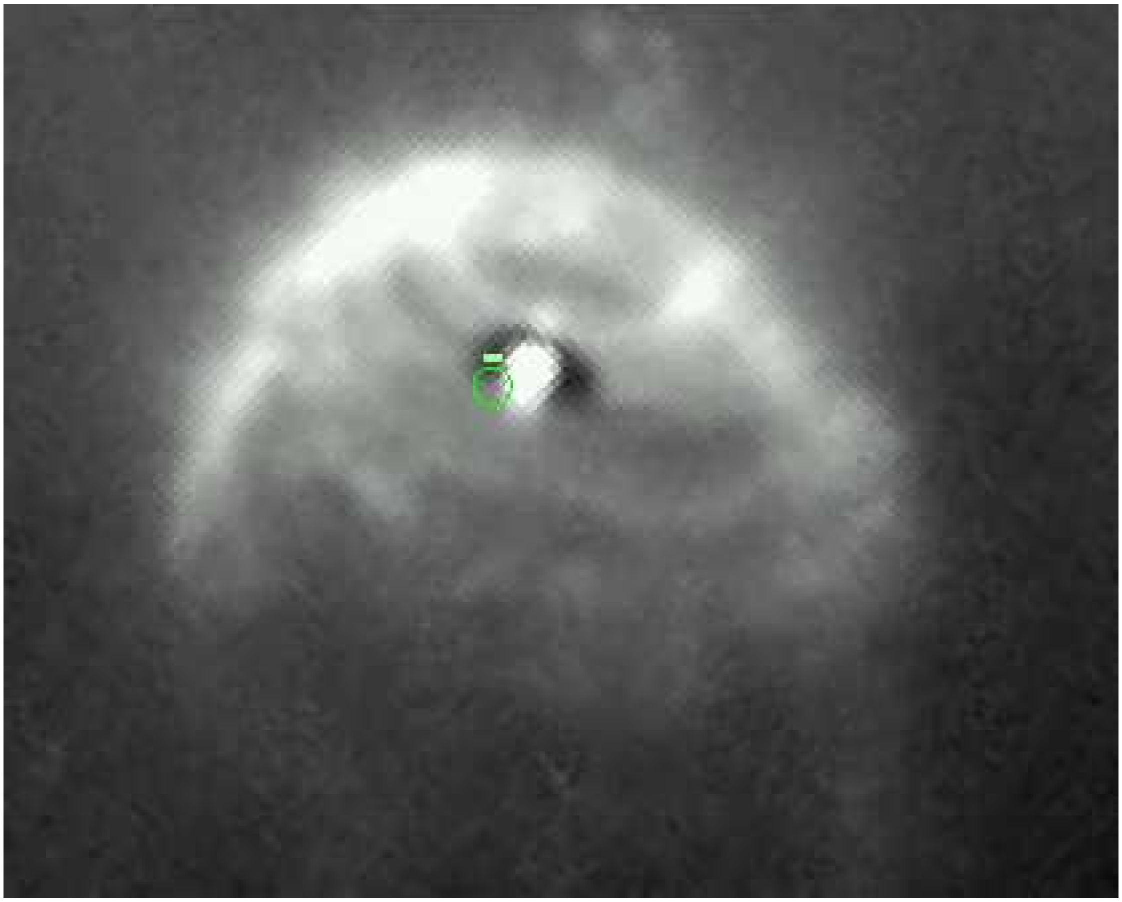


Figure 9.21: Hubble Space Telescope image of the Beehive Proplyd[96]. The position of the associated COUP source (COUP 948) is shown by the green circle.

9.8 Hardness Ratio Diagram

A common practice in X-ray astronomy is to examine X-ray spectral properties by analysis of the X-ray hardness ratio for a group of sources. To compute a hardness ratio, the full-range of the X-ray energy band is divided into sub-bands, and the source's photons are totaled for each sub-band. The ratio of the counts in one band to the counts in another band is defined as a X-ray hardness ratio. For example, if the full-energy range is divided into three energy sub-bands, labeled s (soft), m (medium), and h (hard), then a hardness ratio can be defined as HR = (h / m). Alternate hardness ratios can be defined as HR = (h / s) and HR = (h - s) / (h + s). The sub-band definitions are not the same for all X-ray missions because their definition is dependent upon the energy range over which the CCDs involved are sensitive.

For COUP, four sub-bands were defined: 0.5 to 1.7 keV (s_1), 0.5 to 2.0 keV (s_2), 1.7 to 2.8 keV (m), and 2.0 to 8.0 keV (h)[12]. The three COUP hardness ratios are defined as HR1 = (s_2 / h), HR2 = (s_1 / m), and HR3 = (m / h). HR1 represents the traditional hardness ratio definition; while HR2 is used to measure the softer part of the X-ray spectrum and HR3 the harder part of the X-ray spectrum[12].

The plot of HR3 versus HR2 for the 444 subset is shown in Figure 9.22. COUP sources 510 and 647 have not been included in this plot due to the lack of HR2 data for them in the COUP table. This plot shows that, although the X-ray spectral classes are a sequence in spectral hardness, there are additional, more subtle aspects of the classes that do not appear in the hardness ratio plot. The ordering of the sources on the HR diagram does not match the ordering of the X-ray spectral classes determined from the algorithm and the groupings obtained from the algorithm could not be obtained from this plot alone. For example, class 8 covers a wide range in HR2. Also, classes 15 and 16 are blurred in HR space. The ordering of the sources on this HR diagram is not representative of their order in X-ray spectral space.

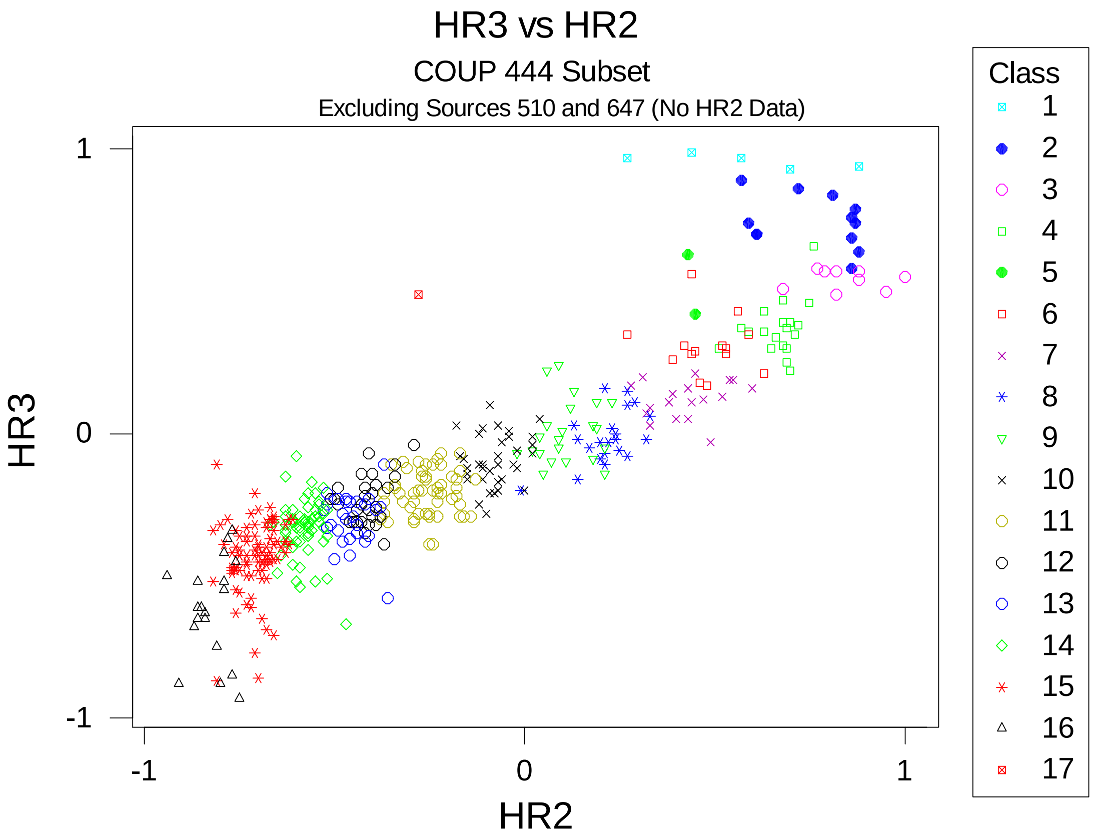


Figure 9.22: Hardness Ratio diagram for the COUP 444 subset.

Chapter 10

Summary and Future Work

10.1 Summary

I have developed an algorithm and corresponding input variable definition that groups X-ray sources based solely on observables. The method is non-parametric. It is an improvement over other methods that rely on empirical measures of X-ray spectral properties, such as hardness ratios, because it incorporates a technique that accounts for the variance in the data. Source groupings are then determined by examining the four principal components that represent the most variance in the data.

Classification results reveal that my spectral clustering technique can be used to efficiently identify very young X-ray sources that:

- lack optical and near-infrared counterparts
- display strong Fe Kα line emission
- display large-amplitude, fast-rise flares

The plot of the first two principal components (Figure 9.2) contains a horseshoe-shaped curve. The spectral hardness of the classes going clockwise around the horseshoe decreases. Extreme outliers do not fall along the horseshoe-shaped curve, but in the space surrounding the curve. In addition, the Andrews Curves for classes 16 and 17 in Figure 9.10 confirm the outlier status of the source in class 17. These results demonstrate that the algorithm can be used to place the sources in order of decreasing spectral hardness and can be used to identify outliers with unusual spectra.

Trends between X-ray spectral parameters and stellar parameters have been found for very low-mass, soft spectra, young sources. Also, there are clear correlations between the softer X-ray spectral classes and the classical optical spectral types in the cluster H-R diagram. These trends and correlations are of significance to astronomers interested in star formation and the mechanisms causing X-ray emission in young stellar clusters.

10.2 Future Work

Future work includes using the source classification algorithm to place the remaining ~1000 sources in the COUP data set into the existing X-ray spectral classes. The relationships between the X-ray spectral classes and fundamental stellar parameters found by my research may or may not be unique to the ONC. X-ray sources in other star formation regions could be grouped into clusters based on the source groupings from the ONC to determine whether candidate young stars in these nearby star formation regions fit into the previously established statistical groupings from the ONC.

Once it is determined whether or not the results from the ONC generalize to other star forming regions, the algorithm could be extended for use with ‘unknown’ X-ray source datasets, i.e., a field of X-ray sources whose mixture of foreground stars, young stars, distant AGNs, and/or other unusual sources is far less well-determined than the Orion region.

It may be useful to do a separate analysis and clustering on the temporal data using the X-ray light curves (time series plots of intensity) for each detected source. An example X-ray light curve for one of the detected X-ray sources is shown in Figure 10.1. The black line shows the full energy band (0.5 – 8.0 keV) light curve, binned according to the values in Table 10.1.

Table 10.1: Light curve bin sizes.

In-Band Source Counts	Bin Length
< 200	~ 6.3 hours
< 500	~ 3.2 hours
< 10000	~ 1.59 hours
< 20000	~ 1.1 hours
< 40000	~ 47.6 minutes
>= 40000	~ 23.8 minutes

The red line in Figure 10.1 shows the light curve in the soft energy band (0.5 – 2.0 keV). The blue line shows the light curve in the hard energy band (2.0 – 8.0 keV).

A flare can be seen in the center of the light curve. Multivariate features describing the variability of the sources and the power spectrum of the time series could possibly be used to aid in the clustering analysis. Previous astronomical studies on temporal analysis could be investigated to determine input variables that best describe the variability of the data. Finally, temporal inputs could be combined with the spectral inputs and the clustering algorithm re-run on the ONC to determine the effects of adding source variability to the algorithm. Such a hybrid method might be particularly effective when attempting to understand the robustness of the spectral classifications.

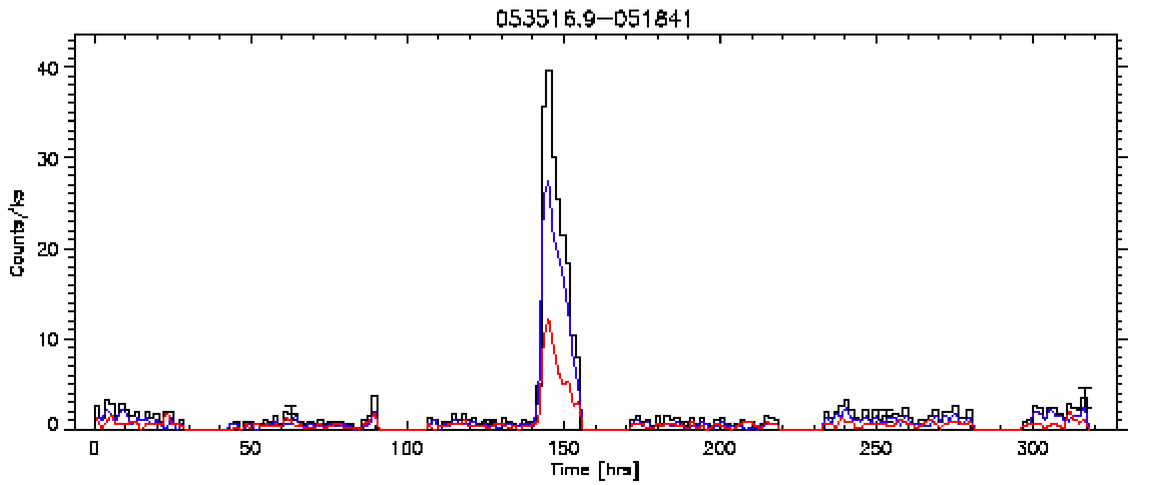


Figure 10.1: Example of a time series plot for one X-ray source.

Appendix A

42 X-ray Spectral Bands

Band Number	Low (eV) - High (eV)[a]	Ion	E (eV)	Theoretical λ (Å)[b]	Model Line Flux[c]
1	425.00 - 525.00	N VII	500.345	24.782	137.20
2	545.00 - 631.00	O VII	561.117	22.098	76.97
		O VII	568.735	21.802	18.70
		O VII	574.000	21.602	128.90
3	632.00 - 721.00	O VIII	653.640	18.970	858.80
		O VII	665.676	18.627	15.36
		Fe XVIII	703.601	17.623	55.56
4	722.00 - 815.00	Fe XVII	725.290	17.096	210.10
		Fe XVII	727.204	17.051	232.30
		Fe XVII	738.948	16.780	193.60
		Fe XVIII	767.347	16.159	31.16
		Fe XIX	769.681	16.110	43.53
		Fe XVIII	771.548	16.071	77.00
		O VIII	774.682	16.006	127.10
		Fe XVIII	781.320	15.870	18.03
		Fe XVIII	783.592	15.824	33.99
		Fe XVIII	793.571	15.625	55.81
		Fe XVII	812.499	15.261	124.40
5	816.00 - 912.00	O VIII	817.050	15.176	40.88
		Fe XIX	822.306	15.079	33.30
		Fe XVII	825.866	15.014	441.60
		O VIII	836.621	14.821	17.90
		Fe XVIII	853.141	14.534	41.11
		Fe XX	869.107	14.267	26.94
		Fe XVIII	869.778	14.256	40.82
		Ne IX	905.143	13.699	65.60
6	913.00 - 1013.00	Ne IX	914.961	13.552	19.55
		Fe XIX	917.262	13.518	99.15

		Fe XIX	918.690	13.497	44.98
		Ne IX	922.106	13.447	144.90
		Fe XXII	972.209	12.754	25.34
		Fe XX	985.970	12.576	22.67
		Fe XXI	1009.407	12.284	135.90
		Fe XVII	1010.888	12.266	45.09
7	1014.00 - 1115.00	Fe XXIII	1019.616	12.161	50.26
		Ne X	1021.801	12.135	633.70
		Fe XVII	1022.728	12.124	50.17
		Fe XXII	1053.488	11.770	70.06
		Fe XXIII	1056.540	11.736	91.19
		Ne IX	1074.112	11.544	20.24
		Fe XVIII	1075.697	11.527	12.78
		Fe XVIII	1094.787	11.326	18.61
		Fe XXIV	1109.480	11.176	76.20
8	1116.00 - 1220.00	Fe XXIV	1124.268	11.029	42.19
		Fe XXIII	1125.288	11.019	28.89
		Fe XXIII	1129.183	10.981	44.10
		Fe XIX	1146.408	10.816	11.96
		Fe XVII	1151.305	10.770	9.08
		Fe XXIV	1162.858	10.663	34.61
		Fe XXIV	1167.676	10.619	65.97
		Ne X	1211.012	10.239	89.03
9	1221.00 - 1335.00	Ne X	1277.251	9.708	28.24
		Mg XI	1331.281	9.314	26.81
10	1336.00 - 1445.00	Mg XI	1343.397	9.230	8.99
		Mg XI	1352.334	9.169	59.28
		Fe XXII	1381.566	8.975	7.57
		Fe XXII	1406.643	8.815	8.38
11	1446.00 - 1558.00	Mg XII	1472.281	8.422	102.70
		Fe XXIV	1491.048	8.316	10.09
		Fe XXIII	1493.203	8.304	8.36
		Fe XXIV	1496.627	8.285	1.85
		Fe XXIV	1506.080	8.233	4.96
		Fe XXIV	1551.690	7.991	13.82

12	1559.00 - 1673.00	Al XII	1575.147	7.872	4.53
		Mg XI	1579.561	7.850	7.66
		Al XII	1598.499	7.757	5.51
13	1674.00 - 1840.00	Al XIII	1728.884	7.172	12.00
		Fe XXIV	1729.607	7.169	4.33
		Mg XII	1744.941	7.106	14.14
		Si XIII	1839.696	6.740	20.64
14	1852.00 - 1974.00	Si XIII	1854.278	6.687	9.54
		Si XIII	1865.156	6.648	51.71
15	1975.00 - 2100.00	Si XIV	2005.427	6.183	62.34
16	2101.00 - 2400.00	Si XIV	2376.759	5.217	8.81
17	2401.00 - 2537.00	S XV	2430.332	5.102	8.78
		S XV	2448.086	5.065	5.83
		S XV	2460.717	5.039	26.65
18	2538.00 - 2676.00	S XVI	2621.470	4.730	26.05
19	2677.00 - 3045.00				
20	3046.00 - 3276.00	Ar XVII	3106.101	3.992	6.12
		Ar XVII	3124.888	3.968	2.32
		Ar XVII	3139.922	3.949	8.91
21	3277.00 - 3436.00	Ar XVII	3320.716	3.734	5.57
22	3437.00 - 3737.00	Ar XVII	3684.860	3.365	1.16
23	3738.00 - 3909.00	Ca XIX	3877.284	3.198	8.37
24	3910.00 - 4085.00				
25	4086.00 - 4266.00	Ca XX	4104.453	3.021	1.90
26	4267.00 - 4452.00				
27	4453.00 - 4643.00				
28	4644.00 - 4838.00				
29	4839.00 - 5038.00				
30	5039.00 - 5243.00				
31	5244.00 - 5454.00				
32	5455.00 - 5670.00				
33	5671.00 - 5891.00				
34	5892.00 - 6118.00				
35	6119.00 - 6351.00				
36	6352.00 - 6590.00	Fe Kα	6400		

37	6591.00 - 6834.00	Fe XXV	6662.845	1.861	25.95
38	6835.00 - 7086.00	Fe XXVI	6962.130	1.781	4.40
39	7087.00 - 7344.00				
40	7345.00 - 7609.00				
41	7610.00 - 7881.00				
42	7882.00 - 8156.00				

[a] Gaps (526-544 eV and 1841-1851 eV) due to drop in QE of ACIS-I chips

[b] From Huenemoerder, D.P., Canizares, C.R., Drake, J.J, and Sanz-Forcada, J., “The Coronae of AR Lacertae”, The Astrophysical Journal, Vol. 595, pp. 1131-1147, 2003.

[c] From the Astrophysical Plasma Emissivity Database (APED)

Appendix B

Similarity Matrix for Preliminary Dataset

Step	Number of Clusters	Similarity Level	Distance
1	184	99.66	0.03
2	183	99.59	0.035
3	182	99.59	0.036
4	181	99.51	0.042
5	180	99.43	0.049
6	179	99.41	0.05
7	178	99.37	0.054
8	177	99.35	0.056
9	176	99.16	0.072
10	175	99.14	0.074
11	174	99.12	0.076
12	173	98.98	0.088
13	172	98.92	0.093
14	171	98.89	0.096
15	170	98.81	0.103
16	169	98.77	0.106
17	168	98.76	0.107
18	167	98.75	0.108
19	166	98.7	0.112
20	165	98.69	0.113
21	164	98.63	0.118
22	163	98.56	0.125
23	162	98.4	0.138
24	161	98.4	0.138
25	160	98.39	0.139
26	159	98.36	0.141
27	158	98.34	0.144
28	157	98.31	0.145
29	156	98.26	0.15
30	155	98.23	0.153
31	154	98.22	0.154
32	153	98.18	0.157
33	152	98.17	0.158
34	151	98.17	0.158
35	150	98.11	0.163
36	149	98.1	0.164

Step	Number of Clusters	Similarity Level	Distance
37	148	98.09	0.165
38	147	97.93	0.179
39	146	97.87	0.184
40	145	97.86	0.185
41	144	97.85	0.185
42	143	97.85	0.185
43	142	97.8	0.189
44	141	97.78	0.192
45	140	97.78	0.192
46	139	97.73	0.195
47	138	97.68	0.2
48	137	97.63	0.204
49	136	97.54	0.212
50	135	97.48	0.217
51	134	97.48	0.218
52	133	97.43	0.222
53	132	97.39	0.225
54	131	97.3	0.232
55	130	97.3	0.233
56	129	97.14	0.247
57	128	97.12	0.249
58	127	97.12	0.249
59	126	97.08	0.252
60	125	97.01	0.258
61	124	96.96	0.263
62	123	96.9	0.267
63	122	96.88	0.27
64	121	96.85	0.272
65	120	96.82	0.274
66	119	96.77	0.278
67	118	96.68	0.286
68	117	96.65	0.289
69	116	96.57	0.296
70	115	96.56	0.296
71	114	96.47	0.304
72	113	96.45	0.307
73	112	96.4	0.311
74	111	96.38	0.312
75	110	96.38	0.312
76	109	96.33	0.317
77	108	96.27	0.322
78	107	96.1	0.337
79	106	96.01	0.344
80	105	96	0.345

Step	Number of Clusters	Similarity Level	Distance
81	104	95.94	0.35
82	103	95.93	0.351
83	102	95.92	0.352
84	101	95.86	0.357
85	100	95.85	0.358
86	99	95.83	0.359
87	98	95.82	0.361
88	97	95.74	0.367
89	96	95.71	0.37
90	95	95.62	0.378
91	94	95.56	0.383
92	93	95.52	0.386
93	92	95.48	0.39
94	91	95.48	0.39
95	90	95.29	0.406
96	89	95.25	0.41
97	88	95.23	0.411
98	87	95.16	0.418
99	86	94.83	0.446
100	85	94.75	0.453
101	84	94.53	0.471
102	83	94.51	0.473
103	82	94.5	0.475
104	81	94.44	0.479
105	80	94.32	0.49
106	79	94.22	0.499
107	78	94.14	0.506
108	77	94.13	0.507
109	76	94.1	0.509
110	75	94.08	0.51
111	74	94.04	0.514
112	73	93.96	0.521
113	72	93.92	0.524
114	71	93.9	0.526
115	70	93.56	0.555
116	69	93.54	0.557
117	68	93.44	0.566
118	67	93.43	0.567
119	66	93.39	0.57
120	65	93.33	0.576
121	64	93.31	0.577
122	63	93.29	0.579
123	62	93.14	0.592
124	61	93	0.604

Step	Number of Clusters	Similarity Level	Distance
125	60	92.95	0.608
126	59	92.74	0.626
127	58	92.67	0.632
128	57	92.6	0.638
129	56	92.4	0.656
130	55	92.15	0.677
131	54	92.11	0.681
132	53	92	0.69
133	52	91.76	0.711
134	51	91.68	0.718
135	50	91.64	0.721
136	49	91.6	0.725
137	48	91.2	0.76
138	47	91.13	0.765
139	46	90.92	0.783
140	45	90.36	0.831
141	44	90.31	0.836
142	43	90.19	0.846
143	42	90.12	0.853
144	41	89.7	0.889
145	40	89.68	0.89
146	39	88.95	0.953
147	38	88.7	0.975
148	37	88.69	0.976
149	36	88.6	0.984
150	35	88.4	1.001
151	34	88.05	1.03
152	33	87.86	1.048
153	32	87.38	1.088
154	31	87.24	1.101
155	30	87.12	1.111
156	29	86.69	1.148
157	28	86.66	1.15
158	27	86.11	1.198
159	26	85.15	1.281
160	25	84.74	1.316
161	24	84.71	1.319
162	23	84.53	1.334
163	22	83.72	1.405
164	21	83.5	1.423
165	20	83.36	1.435
166	19	82.66	1.496
167	18	81.83	1.568
168	17	81.25	1.617

Step	**Number of Clusters**	**Similarity Level**	**Distance**
169	16	78.36	1.867
170	15	76.47	2.03
171	14	76.29	2.045
172	13	74.37	2.21
173	12	73.88	2.253
174	11	72.57	2.366
175	10	70.54	2.541
176	9	68.87	2.685
177	8	68.76	2.695
178	7	61.09	3.357
179	6	55.46	3.842
180	5	47.21	4.554
181	4	44.38	4.798
182	3	31.71	5.891
183	2	23.61	6.59
184	1	0	8.626

Appendix C

Clustering Assignment Summary for Preliminary Dataset

Source Number	RA	DEC	Hierarchical Clustering Class Membership	K-means Class Membership	Source Changed Classes
20	83.8154	-5.3822	1	1	
22	83.7982	-5.43389	1	1	
24	83.773	-5.24785	1	1	
26	83.8601	-5.42765	1	1	
29	83.8139	-5.38228	4	1	*
31	83.8813	-5.42098	4	1	*
34	83.8804	-5.25876	4	1	*
35	83.828	-5.34258	4	1	*
21	83.8488	-5.39198	2	2	
23	83.8233	-5.29429	2	2	
25	83.8338	-5.35152	2	2	
28	83.8168	-5.397	2	2	
27	83.8532	-5.4664	3	3	
30	83.7994	-5.36358	3	3	
32	83.828	-5.387	3	3	
33	83.7409	-5.39772	4	4	
36	83.8646	-5.44099	5	4	*
37	83.7631	-5.50054	4	4	
38	83.8586	-5.42975	4	4	
39	83.825	-5.25998	4	4	
40	83.8165	-5.48127	4	4	
41	83.839	-5.41575	5	4	*
42	83.8343	-5.44405	5	4	*
43	83.8268	-5.37695	4	4	
44	83.7986	-5.28261	4	4	
45	83.7881	-5.49976	5	4	*
46	83.7737	-5.42198	5	4	*
47	83.8483	-5.31565	4	4	
48	83.8211	-5.37573	5	4	*
49	83.8478	-5.31385	5	4	*
50	83.794	-5.43832	5	4	*
52	83.8345	-5.34901	4	4	
53	83.8731	-5.27575	5	4	*
56	83.7113	-5.4002	5	4	*
57	83.8369	-5.26372	4	4	
60	83.8151	-5.42048	5	4	*
63	83.8568	-5.50533	5	4	*
69	83.7695	-5.29462	5	4	*
90	83.8873	-5.26654	4	4	

97	83.694	-5.40874	5	4	*
101	83.7353	-5.46361	5	4	*
103	83.7198	-5.4653	5	4	*
113	83.7424	-5.29376	5	4	*
144	83.694	-5.39053	5	4	*
196	83.927	-5.45818	5	4	*
55	83.7325	-5.49108	5	5	
68	83.9118	-5.298	5	5	
71	83.9139	-5.28819	5	5	
72	83.8854	-5.52139	5	5	
74	83.9144	-5.2832	5	5	
78	83.9103	-5.30409	5	5	
79	83.8733	-5.51781	5	5	
85	83.7711	-5.26869	5	5	
86	83.9125	-5.29376	5	5	
95	83.8748	-5.51176	5	5	
96	83.9493	-5.37988	5	5	
100	83.9491	-5.38001	5	5	
106	83.8757	-5.50079	5	5	
112	83.7331	-5.489	5	5	
115	83.6717	-5.44938	5	5	
119	83.7727	-5.25903	5	5	
120	83.9096	-5.30875	5	5	
129	83.9395	-5.47097	5	5	
136	83.7737	-5.25525	5	5	
140	83.8771	-5.49504	5	5	
145	83.6715	-5.44464	5	5	
146	83.8752	-5.50416	5	5	
153	83.8725	-5.52738	5	5	
156	83.7718	-5.26367	5	5	
167	83.9697	-5.3511	5	5	
168	83.701	-5.29422	5	5	
178	83.9093	-5.31335	5	5	
179	83.7333	-5.48161	5	5	
182	83.879	-5.47699	5	5	
184	83.878	-5.49061	5	5	
189	83.8784	-5.48732	5	5	
192	83.9467	-5.47901	5	5	
193	83.7351	-5.47415	5	5	
51	83.9821	-5.27145	6	6	
58	83.9357	-5.5362	6	6	
62	83.7135	-5.22303	5	6	*
66	83.9785	-5.28172	6	6	
77	83.9159	-5.26979	6	6	
89	83.915	-5.27766	5	6	*
92	83.6676	-5.49003	6	6	
99	83.9757	-5.30583	6	6	
102	83.9775	-5.29213	6	6	
104	83.918	-5.25975	6	6	
123	83.9386	-5.51897	6	6	
124	83.6702	-5.46228	5	6	*
125	83.7479	-5.22836	6	6	

135	83.7766	-5.23407	5	6	*
139	83.9082	-5.53285	6	6	
142	83.7087	-5.25025	6	6	
149	83.9651	-5.27076	6	6	
150	83.706	-5.25987	5	6	*
152	83.7095	-5.24072	6	6	
157	83.9388	-5.51031	6	6	
159	83.9423	-5.2623	6	6	
160	83.9379	-5.5237	6	6	
162	83.667	-5.47985	6	6	
165	83.9725	-5.32014	6	6	
166	83.7351	-5.22753	6	6	
175	83.9419	-5.50247	6	6	
176	83.7283	-5.50125	5	6	*
177	83.729	-5.22489	6	6	
188	83.9722	-5.32696	6	6	
199	83.6802	-5.47608	5	6	*
54	83.823	-5.38898	7	7	
59	83.807	-5.33177	7	7	
61	83.8194	-5.40153	7	7	
64	83.8783	-5.40848	7	7	
65	83.8063	-5.51534	7	7	
67	83.825	-5.3792	7	7	
70	83.7909	-5.35777	7	7	
73	83.8142	-5.37088	7	7	
75	83.785	-5.46567	7	7	
76	83.8405	-5.42389	7	7	
80	83.8172	-5.3433	7	7	
81	83.8126	-5.39408	7	7	
82	83.7548	-5.40222	5	7	*
83	83.7589	-5.44349	5	7	*
84	83.8173	-5.38511	7	7	
87	83.8659	-5.30108	7	7	
91	83.822	-5.3587	7	7	
93	83.8593	-5.33484	7	7	
94	83.8781	-5.45458	7	7	
105	83.8014	-5.39651	7	7	
107	83.8521	-5.41086	7	7	
108	83.8127	-5.36654	7	7	
109	83.8069	-5.51641	7	7	
110	83.8001	-5.34238	7	7	
111	83.8291	-5.27041	7	7	
114	83.7764	-5.36732	7	7	
118	83.8952	-5.48733	7	7	
121	83.8397	-5.52301	7	7	
126	83.8033	-5.28089	7	7	
127	83.8349	-5.51156	7	7	
128	83.8226	-5.42893	7	7	
130	83.7506	-5.38365	5	7	*
131	83.7371	-5.36008	7	7	
132	83.8113	-5.37595	7	7	
133	83.8858	-5.43588	5	7	*

134	83.8226	-5.33732	7	7	
137	83.8355	-5.39139	7	7	
138	83.8376	-5.27716	7	7	
141	83.8773	-5.40616	7	7	
143	83.878	-5.30181	7	7	
147	83.7664	-5.48473	7	7	
148	83.8246	-5.27043	7	7	
151	83.8393	-5.39575	7	7	
154	83.7976	-5.31983	7	7	
155	83.8353	-5.28701	7	7	
158	83.8124	-5.37745	7	7	
161	83.7671	-5.44355	7	7	
163	83.8201	-5.40101	7	7	
164	83.8974	-5.35731	7	7	
169	83.8431	-5.3413	7	7	
170	83.7197	-5.40073	7	7	
171	83.8548	-5.3961	7	7	
172	83.8315	-5.28428	7	7	
173	83.8148	-5.45622	7	7	
174	83.8807	-5.31552	7	7	
180	83.8735	-5.41565	7	7	
181	83.8342	-5.35919	7	7	
183	83.7943	-5.36552	7	7	
185	83.807	-5.40702	7	7	
186	83.8726	-5.42947	7	7	
187	83.804	-5.25612	7	7	
190	83.8253	-5.4931	7	7	
191	83.7177	-5.37531	7	7	
194	83.7253	-5.48084	7	7	
195	83.7928	-5.38914	7	7	
197	83.8772	-5.42719	7	7	
198	83.7507	-5.42099	7	7	
200	83.9029	-5.33595	7	7	
201	83.8173	-5.25029	7	7	
202	83.824	-5.41482	7	7	
203	83.8185	-5.40079	7	7	
204	83.8174	-5.24882	7	7	
88	83.8038	-5.3593	8	8	
98	83.8214	-5.39264	7	8	*
116	83.8144	-5.35377	7	8	*
117	83.7991	-5.42011	8	8	
122	83.828	-5.31804	8	8	

Appendix D

Background Counts Table for COUP 444 Subset

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
6	2260	372	1887	16.47
8	1349	219	1129	16.25
11	5824	123	5700	2.11
17	1126	42	1083	3.73
20	531	148	382	27.92
21	572	120	451	21.02
28	21013	149	20863	0.71
29	2349	61	2287	2.60
40	359	44	314	12.29
43	7085	63	7021	0.89
49	902	170	731	18.87
54	1640	56	1583	3.42
55	511	49	461	9.61
60	794	137	656	17.28
62	9361	66	9294	0.71
64	864	49	814	5.68
65	870	29	840	3.34
66	6266	28	6237	0.45
67	7732	41	7690	0.53
69	1009	186	822	18.45
89	2064	30	2033	1.45
90	7257	139	7117	1.92
96	1446	19	1426	1.31
100	821	141	679	17.20
109	1185	189	995	15.96
110	588	77	510	13.12
111	1020	20	999	1.96
112	7469	70	7398	0.94
113	6807	82	6724	1.20
114	471	34	436	7.23
115	6163	22	6140	0.36
117	1321	19	1301	1.44
118	407	29	377	7.14
119	737	109	627	14.81
122	4962	22	4939	0.44
128	326	18	307	5.54
132	1491	21	1469	1.41
133	341	18	322	5.29
134	322	15	306	4.67
137	523	21	501	4.02
139	6124	29	6094	0.47

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
141	13744	16	13727	0.12
154	380	12	367	3.17
164	399	9	389	2.26
165	385	13	371	3.39
169	517	96	420	18.60
172	2601	24	2576	0.92
173	12356	26	12329	0.21
174	2879	68	2810	2.36
177	5081	26	5054	0.51
179	1028	184	843	17.92
183	5776	12	5763	0.21
192	525	112	412	21.37
197	1114	67	1046	6.02
202	5150	47	5102	0.91
205	6401	17	6383	0.27
217	2331	7	2323	0.30
218	2092	9	2082	0.43
223	10306	62	10243	0.60
224	1510	6	1503	0.40
226	2684	7	2676	0.26
227	460	5	454	1.09
230	1061	5	1055	0.47
236	1189	8	1180	0.67
238	349	6	342	1.72
241	337	22	314	6.55
244	762	31	730	4.07
246	656	42	613	6.41
249	2058	8	2049	0.39
250	505	7	497	1.39
253	1395	4	1390	0.29
255	988	13	974	1.32
256	755	15	739	1.99
260	2481	10	2470	0.40
262	11551	10	11540	0.09
266	843	11	831	1.31
269	1731	62	1668	3.58
270	6655	7	6647	0.11
276	705	8	696	1.14
292	1630	49	1580	3.01
294	471	6	464	1.28
296	427	52	374	12.21
300	608	9	598	1.48
301	2167	148	2018	6.83
304	1090	6	1083	0.55
308	628	21	606	3.35
309	981	6	974	0.61
310	6189	21	6167	0.34
312	554	6	547	1.08
314	478	19	458	3.98
319	442	73	368	16.55

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
321	317	8	308	2.53
322	1636	7	1628	0.43
323	5190	43	5146	0.83
325	4972	81	4890	1.63
328	13927	71	13855	0.51
331	5934	144	5789	2.43
332	3269	10	3258	0.31
338	3205	8	3196	0.25
340	711	16	694	2.25
353	1274	27	1246	2.12
365	6499	12	6486	0.18
368	476	15	460	3.16
373	1095	134	960	12.25
376	1220	81	1138	6.64
379	743	11	731	1.48
382	5081	15	5065	0.30
385	892	6	885	0.67
387	20103	12	20090	0.06
389	878	149	728	16.99
391	1654	10	1643	0.60
395	575	9	565	1.57
404	2421	7	2413	0.29
407	606	126	479	20.83
410	490	19	470	3.89
413	3678	473	3204	12.86
414	3577	60	3516	1.68
415	2804	8	2795	0.29
418	323	14	308	4.35
424	425	12	412	2.83
427	3698	6	3691	0.16
431	20692	77	20614	0.37
435	1334	13	1320	0.98
441	417	12	404	2.88
446	1803	54	1748	3.00
454	17142	16	17125	0.09
459	8201	142	8058	1.73
466	312	7	304	2.25
468	1637	93	1543	5.68
470	12580	10	12569	0.08
471	522	8	513	1.54
472	505	3	501	0.60
481	3431	72	3358	2.10
483	707	5	701	0.71
485	4253	18	4234	0.42
488	3409	13	3395	0.38
489	2273	75	2197	3.30
490	6772	89	6682	1.31
498	547	9	537	1.65
499	5490	108	5381	1.97
507	428	13	414	3.04

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
510	415	11	403	2.66
513	336	18	317	5.37
514	6871	21	6849	0.31
515	4406	12	4393	0.27
517	611	29	581	4.75
520	857	50	806	5.84
533	372	11	360	2.96
536	1909	15	1893	0.79
539	513	10	502	1.95
545	3111	7	3103	0.23
548	2354	38	2315	1.61
550	902	13	888	1.44
553	2330	6	2323	0.26
554	14056	10	14045	0.07
557	1965	10	1954	0.51
561	13686	23	13662	0.17
563	509	17	491	3.35
565	3915	9	3905	0.23
566	1175	23	1151	1.96
572	813	18	794	2.22
585	520	38	481	7.32
595	633	8	624	1.27
598	482	13	468	2.70
599	862	15	846	1.74
602	1775	5	1769	0.28
604	312	6	305	1.93
610	733	135	597	18.44
612	2332	26	2305	1.12
614	1978	24	1953	1.21
616	1102	5	1096	0.45
624	3768	20	3747	0.53
625	681	22	658	3.24
626	5565	53	5511	0.95
627	502	22	479	4.39
631	2001	10	1990	0.50
640	448	24	423	5.37
645	4484	12	4471	0.27
647	829	20	808	2.42
648	24456	42	24413	0.17
649	5106	29	5076	0.57
653	727	11	715	1.52
655	6361	31	6329	0.49
658	2567	25	2541	0.97
660	2985	246	2738	8.24
663	1807	41	1765	2.27
664	1279	25	1253	1.96
665	1068	31	1036	2.91
666	670	15	654	2.24
667	383	6	376	1.57
671	447	118	328	26.46

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
672	7586	21	7564	0.28
680	1402	17	1384	1.21
695	660	22	637	3.34
697	6001	16	5984	0.27
700	1838	30	1807	1.63
710	644	7	636	1.09
711	3623	11	3611	0.30
712	404	11	392	2.73
713	1716	48	1667	2.80
723	593	7	585	1.18
726	480	13	466	2.71
737	4338	11	4326	0.25
739	686	21	664	3.07
750	2773	26	2746	0.94
751	378	40	337	10.61
753	5890	160	5729	2.72
754	390	49	340	12.60
756	749	18	730	2.41
763	879	15	863	1.71
776	910	28	881	3.08
780	1708	354	1353	20.74
783	441	11	429	2.50
789	1612	9	1602	0.56
790	1896	28	1867	1.48
797	1736	20	1715	1.15
798	982	13	968	1.33
801	12296	12	12283	0.10
807	1472	24	1447	1.63
817	681	13	667	1.91
823	2260	14	2245	0.62
837	2107	27	2079	1.28
849	360	9	350	2.51
852	593	15	577	2.53
856	3328	14	3313	0.42
857	786	22	763	2.80
862	516	45	470	8.74
864	346	8	337	2.32
865	352	16	335	4.56
869	7942	40	7901	0.50
878	378	64	313	16.98
885	3404	13	3390	0.38
888	455	11	443	2.42
892	943	36	906	3.82
896	1278	10	1267	0.78
897	2100	16	2083	0.76
899	1945	29	1915	1.49
902	984	12	971	1.22
903	378	15	362	3.98
914	415	2	412	0.48
919	328	13	314	3.98

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
921	1549	33	1515	2.13
924	592	3	588	0.51
936	1981	7	1973	0.35
937	658	13	644	1.98
939	10473	181	10291	1.73
948	505	17	487	3.37
949	3639	333	3305	9.15
960	5280	53	5226	1.00
966	1346	9	1336	0.67
969	1562	22	1539	1.41
970	1419	71	1347	5.01
972	1225	11	1213	0.90
974	2602	27	2574	1.04
976	3895	11	3883	0.28
986	1856	15	1840	0.81
992	1609	9	1599	0.56
998	413	17	395	4.13
1000	324	8	315	2.48
1007	352	9	342	2.56
1008	3086	11	3074	0.36
1009	319	13	305	4.09
1019	1923	8	1914	0.42
1028	5057	12	5044	0.24
1035	7558	67	7490	0.89
1041	603	12	590	1.99
1045	4281	7	4273	0.16
1053	1013	171	841	16.90
1054	1810	179	1630	9.89
1056	375	15	359	4.01
1058	1355	6	1348	0.44
1062	392	13	378	3.32
1066	3186	196	2989	6.15
1067	898	406	491	45.26
1070	4746	12	4733	0.25
1071	17079	27	17051	0.16
1074	402	2	399	0.50
1075	468	2	465	0.43
1076	1991	4	1986	0.20
1081	1051	14	1036	1.33
1086	337	7	329	2.08
1095	662	10	651	1.51
1097	2564	45	2518	1.76
1100	2715	9	2705	0.33
1101	3759	5	3753	0.13
1103	2118	12	2105	0.57
1104	1934	26	1907	1.35
1110	1757	11	1745	0.63
1111	7430	18	7411	0.24
1112	2799	10	2788	0.36
1117	1623	13	1609	0.80

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
1120	337	13	323	3.87
1121	1291	14	1276	1.09
1123	321	8	312	2.50
1126	358	41	316	11.48
1127	5680	81	5598	1.43
1128	899	11	887	1.22
1131	597	12	584	2.01
1132	630	14	615	2.23
1134	5075	152	4922	3.00
1135	373	33	339	8.87
1137	644	15	628	2.33
1139	312	4	307	1.29
1140	7044	12	7031	0.17
1141	881	6	874	0.68
1143	15904	8	15895	0.05
1147	323	10	312	3.11
1149	4512	20	4491	0.44
1150	592	9	582	1.52
1151	24113	18	24094	0.07
1154	823	7	815	0.85
1155	353	4	348	1.14
1158	8525	16	8508	0.19
1161	9283	12	9270	0.13
1165	4534	16	4517	0.35
1167	348	12	335	3.46
1169	465	11	453	2.37
1172	877	11	865	1.26
1177	4200	192	4007	4.57
1191	559	8	550	1.43
1193	3495	5	3489	0.14
1199	4097	8	4088	0.20
1200	1663	11	1651	0.66
1202	3535	11	3523	0.31
1206	934	19	914	2.04
1207	486	69	416	14.23
1210	3803	8	3794	0.21
1212	2097	10	2086	0.48
1216	645	7	637	1.09
1223	421	27	393	6.43
1231	2364	8	2355	0.34
1233	1155	98	1056	8.49
1234	4831	48	4782	0.99
1235	349	3	345	0.86
1236	4528	148	4379	3.27
1242	367	11	355	3.01
1245	1172	7	1164	0.60
1246	7641	25	7615	0.33
1258	615	8	606	1.30
1261	5744	8	5735	0.14
1264	1077	29	1047	2.70

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
1275	575	10	564	1.74
1276	587	8	578	1.37
1279	1616	9	1606	0.56
1282	1033	47	985	4.55
1284	817	3	813	0.37
1290	1938	7	1930	0.36
1291	410	3	406	0.73
1292	2552	70	2481	2.74
1296	510	13	496	2.55
1297	505	14	490	2.78
1298	329	9	319	2.74
1302	357	5	351	1.40
1306	719	5	713	0.70
1308	425	13	411	3.07
1311	5114	6	5107	0.12
1316	1132	102	1029	9.02
1336	1892	8	1883	0.42
1344	1184	8	1175	0.68
1355	5930	51	5878	0.86
1356	1023	97	925	9.49
1357	587	8	578	1.37
1360	1299	5	1293	0.39
1364	529	27	501	5.11
1369	314	12	301	3.83
1373	307	5	301	1.63
1374	5438	128	5309	2.35
1382	10291	73	10217	0.71
1384	25451	17	25433	0.07
1387	1721	9	1711	0.52
1388	2925	43	2881	1.47
1391	14398	55	14342	0.38
1398	3060	22	3037	0.72
1399	463	92	370	19.91
1404	691	72	618	10.43
1407	847	6	840	0.71
1409	6390	8	6381	0.13
1410	8210	55	8154	0.67
1411	1476	45	1430	3.05
1415	883	16	866	1.81
1419	3527	91	3435	2.58
1423	3565	11	3553	0.31
1424	7046	124	6921	1.76
1429	5538	10	5527	0.18
1430	1997	20	1976	1.00
1432	538	80	457	14.90
1433	6432	124	6307	1.93
1438	3206	12	3193	0.37
1439	1194	8	1185	0.67
1440	726	22	703	3.03
1447	698	12	685	1.72

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
1449	5758	53	5704	0.92
1450	1224	14	1209	1.14
1454	2987	112	2874	3.75
1455	3443	50	3392	1.45
1456	7532	13	7518	0.17
1457	905	17	887	1.88
1462	13136	152	12983	1.16
1463	8257	42	8214	0.51
1464	1222	30	1191	2.46
1466	5284	12	5271	0.23
1469	733	21	711	2.87
1471	384	8	375	2.09
1474	739	26	712	3.52
1475	596	105	490	17.65
1478	927	57	869	6.16
1480	928	13	914	1.40
1485	1579	125	1453	7.92
1487	5728	67	5660	1.17
1492	6801	19	6781	0.28
1503	1149	30	1118	2.61
1507	854	50	803	5.86
1512	931	25	905	2.69
1516	7798	60	7737	0.77
1521	10093	31	10061	0.31
1524	651	35	615	5.38
1529	1126	34	1091	3.02
1531	1670	38	1631	2.28
1535	489	79	409	16.19
1537	363	39	323	10.77
1539	1250	39	1210	3.12
1543	2004	50	1953	2.50
1544	3339	107	3231	3.21
1546	1984	49	1934	2.47
1550	2088	41	2046	1.96
1553	2305	61	2243	2.65
1561	3260	39	3220	1.20
1564	692	103	588	14.91
1570	4259	112	4146	2.63
1571	664	49	614	7.39
1572	636	46	589	7.24
1579	605	57	547	9.44
1585	957	41	915	4.29
1588	578	57	520	9.88
1591	1998	94	1903	4.71
1594	700	66	633	9.44
1595	6922	60	6861	0.87
1603	4334	135	4198	3.12
1607	815	116	698	14.25
1608	9368	269	9098	2.87
1609	465	118	346	25.43

Source Number	Source Counts	Bkg Counts	Net Counts	% Bkgnd
1610	1277	60	1216	4.70
1612	1128	152	975	13.49
1616	562	109	452	19.43

Appendix E

Correlation Matrix for COUP 444 Subset

Cell Contents: Pearson Correlation Coefficient

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7	Band 8
Band 2	0.626							
Band 3	0.572	0.857						
Band 4	0.530	0.808	0.970					
Band 5	0.500	0.786	0.952	0.980				
Band 6	0.275	0.554	0.750	0.799	0.861			
Band 7	0.154	0.406	0.603	0.661	0.733	0.970		
Band 8	0.041	0.245	0.435	0.499	0.576	0.870	0.946	
Band 9	-0.045	0.116	0.274	0.329	0.410	0.741	0.846	0.948
Band 10	-0.189	-0.110	0.002	0.046	0.124	0.491	0.633	0.792
Band 11	-0.246	-0.245	-0.181	-0.150	-0.081	0.264	0.415	0.604
Band 12	-0.295	-0.364	-0.371	-0.348	-0.294	0.015	0.170	0.374
Band 13	-0.359	-0.526	-0.582	-0.572	-0.534	-0.258	-0.098	0.117
Band 14	-0.377	-0.570	-0.650	-0.642	-0.614	-0.363	-0.209	-0.001
Band 15	-0.419	-0.627	-0.738	-0.735	-0.715	-0.502	-0.352	-0.142
Band 16	-0.424	-0.680	-0.825	-0.836	-0.836	-0.695	-0.566	-0.381
Band 17	-0.406	-0.668	-0.834	-0.849	-0.860	-0.778	-0.668	-0.501
Band 18	-0.407	-0.670	-0.827	-0.844	-0.858	-0.771	-0.663	-0.503
Band 19	-0.361	-0.642	-0.820	-0.849	-0.877	-0.863	-0.781	-0.648
Band 20	-0.349	-0.620	-0.797	-0.832	-0.871	-0.898	-0.834	-0.728
Band 21	-0.304	-0.596	-0.767	-0.799	-0.843	-0.905	-0.856	-0.765
Band 22	-0.304	-0.573	-0.754	-0.795	-0.842	-0.918	-0.877	-0.805
Band 23	-0.269	-0.536	-0.709	-0.752	-0.803	-0.907	-0.883	-0.824
Band 24	-0.277	-0.537	-0.715	-0.756	-0.809	-0.915	-0.892	-0.841
Band 25	-0.249	-0.494	-0.671	-0.718	-0.773	-0.910	-0.904	-0.870
Band 26	-0.260	-0.489	-0.663	-0.711	-0.770	-0.907	-0.900	-0.865
Band 27	-0.248	-0.470	-0.635	-0.677	-0.739	-0.889	-0.890	-0.867
Band 28	-0.240	-0.461	-0.627	-0.672	-0.731	-0.881	-0.881	-0.860
Band 29	-0.235	-0.443	-0.610	-0.658	-0.718	-0.873	-0.876	-0.863
Band 30	-0.239	-0.465	-0.628	-0.671	-0.731	-0.883	-0.886	-0.860
Band 31	-0.221	-0.433	-0.600	-0.648	-0.707	-0.872	-0.879	-0.863
Band 32	-0.217	-0.421	-0.588	-0.637	-0.698	-0.864	-0.877	-0.871
Band 33	-0.211	-0.420	-0.572	-0.620	-0.681	-0.852	-0.867	-0.864
Band 34	-0.200	-0.395	-0.543	-0.585	-0.646	-0.821	-0.839	-0.839
Band 35	-0.202	-0.349	-0.484	-0.528	-0.587	-0.757	-0.786	-0.811
Band 36	-0.214	-0.389	-0.530	-0.582	-0.640	-0.805	-0.826	-0.832
Band 37	-0.206	-0.369	-0.524	-0.579	-0.634	-0.783	-0.799	-0.804
Band 38	-0.154	-0.307	-0.426	-0.474	-0.528	-0.699	-0.737	-0.779
Band 39	-0.135	-0.274	-0.398	-0.445	-0.497	-0.662	-0.699	-0.735
Band 40	-0.101	-0.225	-0.323	-0.371	-0.417	-0.587	-0.623	-0.659
Band 41	-0.070	-0.191	-0.290	-0.341	-0.390	-0.553	-0.597	-0.645
Band 42	-0.036	-0.209	-0.330	-0.375	-0.408	-0.565	-0.607	-0.630

	Band 9	Band 10	Band 11	Band 12	Band 13	Band 14	Band 15	Band 16
Band 10	0.914							
Band 11	0.761	0.935						
Band 12	0.551	0.781	0.916					
Band 13	0.315	0.599	0.784	0.939				
Band 14	0.198	0.497	0.703	0.884	0.973			
Band 15	0.060	0.367	0.586	0.804	0.931	0.959		
Band 16	-0.185	0.135	0.374	0.628	0.825	0.880	0.942	
Band 17	-0.323	-0.021	0.220	0.501	0.730	0.800	0.880	0.966
Band 18	-0.328	-0.032	0.206	0.485	0.715	0.784	0.868	0.958
Band 19	-0.487	-0.195	0.048	0.340	0.597	0.681	0.791	0.923
Band 20	-0.590	-0.318	-0.089	0.199	0.473	0.569	0.695	0.861
Band 21	-0.638	-0.383	-0.162	0.116	0.395	0.496	0.626	0.808
Band 22	-0.691	-0.443	-0.226	0.046	0.332	0.439	0.571	0.772
Band 23	-0.725	-0.501	-0.303	-0.047	0.234	0.345	0.484	0.699
Band 24	-0.745	-0.521	-0.322	-0.077	0.206	0.322	0.464	0.687
Band 25	-0.790	-0.582	-0.383	-0.136	0.148	0.264	0.412	0.637
Band 26	-0.784	-0.579	-0.385	-0.143	0.139	0.253	0.397	0.626
Band 27	-0.799	-0.608	-0.427	-0.194	0.086	0.205	0.349	0.583
Band 28	-0.791	-0.607	-0.432	-0.207	0.071	0.183	0.334	0.570
Band 29	-0.802	-0.627	-0.459	-0.245	0.033	0.147	0.298	0.540
Band 30	-0.796	-0.611	-0.430	-0.215	0.065	0.177	0.326	0.562
Band 31	-0.802	-0.634	-0.463	-0.247	0.027	0.141	0.293	0.532
Band 32	-0.813	-0.644	-0.474	-0.271	-0.004	0.112	0.259	0.503
Band 33	-0.805	-0.641	-0.481	-0.274	-0.004	0.112	0.256	0.496
Band 34	-0.795	-0.655	-0.509	-0.330	-0.074	0.037	0.191	0.429
Band 35	-0.790	-0.671	-0.554	-0.394	-0.147	-0.035	0.108	0.348
Band 36	-0.781	-0.639	-0.504	-0.325	-0.070	0.045	0.187	0.427
Band 37	-0.760	-0.627	-0.503	-0.333	-0.082	0.031	0.171	0.408
Band 38	-0.770	-0.667	-0.563	-0.424	-0.178	-0.076	0.060	0.302
Band 39	-0.728	-0.631	-0.529	-0.419	-0.211	-0.120	0.003	0.230
Band 40	-0.654	-0.603	-0.533	-0.458	-0.267	-0.186	-0.045	0.154
Band 41	-0.639	-0.554	-0.463	-0.386	-0.224	-0.125	-0.022	0.167
Band 42	-0.620	-0.526	-0.424	-0.311	-0.129	-0.058	0.065	0.238

	Band 17	Band 18	Band 19	Band 20	Band 21	Band 22	Band 23	Band 24
Band 18	0.974							
Band 19	0.960	0.962						
Band 20	0.916	0.915	0.977					
Band 21	0.878	0.883	0.950	0.977				
Band 22	0.846	0.851	0.934	0.972	0.980			
Band 23	0.779	0.785	0.886	0.937	0.948	0.970		
Band 24	0.771	0.780	0.879	0.934	0.948	0.974	0.978	
Band 25	0.729	0.738	0.850	0.912	0.930	0.961	0.965	0.981
Band 26	0.715	0.729	0.839	0.904	0.923	0.955	0.968	0.973
Band 27	0.679	0.693	0.808	0.879	0.907	0.939	0.961	0.966
Band 28	0.669	0.680	0.798	0.875	0.903	0.934	0.957	0.965
Band 29	0.640	0.651	0.770	0.854	0.885	0.924	0.944	0.958
Band 30	0.661	0.672	0.789	0.867	0.893	0.926	0.948	0.956
Band 31	0.634	0.645	0.766	0.848	0.879	0.918	0.943	0.951
Band 32	0.602	0.614	0.742	0.825	0.859	0.899	0.929	0.939
Band 33	0.603	0.611	0.740	0.824	0.860	0.897	0.922	0.934
Band 34	0.534	0.544	0.682	0.772	0.815	0.857	0.892	0.908
Band 35	0.453	0.467	0.606	0.708	0.747	0.805	0.844	0.862
Band 36	0.527	0.539	0.675	0.766	0.805	0.851	0.884	0.897
Band 37	0.510	0.519	0.653	0.744	0.777	0.829	0.878	0.888
Band 38	0.408	0.424	0.560	0.660	0.701	0.756	0.812	0.822
Band 39	0.325	0.329	0.469	0.567	0.605	0.672	0.732	0.742
Band 40	0.240	0.253	0.363	0.466	0.515	0.570	0.629	0.649
Band 41	0.252	0.254	0.380	0.472	0.509	0.574	0.629	0.641
Band 42	0.319	0.322	0.427	0.499	0.522	0.568	0.610	0.624

	Band 25	Band 26	Band 27	Band 28	Band 29	Band 30	Band 31	Band 32
Band 26	0.979							
Band 27	0.971	0.978						
Band 28	0.974	0.970	0.976					
Band 29	0.965	0.968	0.969	0.979				
Band 30	0.960	0.969	0.964	0.965	0.974			
Band 31	0.963	0.961	0.966	0.969	0.972	0.967		
Band 32	0.953	0.957	0.957	0.961	0.963	0.960	0.965	
Band 33	0.948	0.948	0.957	0.957	0.957	0.950	0.953	0.951
Band 34	0.926	0.925	0.935	0.941	0.947	0.938	0.942	0.944
Band 35	0.888	0.889	0.903	0.907	0.914	0.913	0.916	0.926
Band 36	0.918	0.918	0.930	0.942	0.937	0.923	0.932	0.935
Band 37	0.900	0.903	0.911	0.923	0.927	0.911	0.920	0.931
Band 38	0.841	0.850	0.872	0.874	0.889	0.872	0.875	0.885
Band 39	0.768	0.789	0.803	0.805	0.826	0.813	0.829	0.849
Band 40	0.675	0.692	0.698	0.719	0.736	0.720	0.732	0.743
Band 41	0.672	0.678	0.691	0.716	0.728	0.708	0.716	0.727
Band 42	0.643	0.649	0.657	0.673	0.662	0.672	0.673	0.696

	Band 33	Band 34	Band 35	Band 36	Band 37	Band 38	Band 39	Band 40	Band 41
Band 34	0.950								
Band 35	0.915	0.938							
Band 36	0.933	0.933	0.939						
Band 37	0.912	0.918	0.921	0.937					
Band 38	0.890	0.897	0.909	0.909	0.909				
Band 39	0.816	0.842	0.864	0.835	0.863	0.857			
Band 40	0.734	0.773	0.777	0.763	0.756	0.791	0.831		
Band 41	0.731	0.733	0.762	0.767	0.761	0.779	0.772	0.751	
Band 42	0.673	0.684	0.695	0.696	0.679	0.707	0.699	0.669	0.697

Appendix F

Eigenvectors for COUP 444 Subset

Band	PC1	PC2	PC3	PC4	PC5	PC6
1	0.007	-0.013	0.053	-0.008	0.013	0.019
2	0.034	-0.038	0.132	-0.003	0.062	0.037
3	0.120	-0.099	0.311	0.030	0.210	0.079
4	0.129	-0.094	0.295	0.054	0.167	0.116
5	0.304	-0.177	0.581	0.162	0.297	0.180
6	0.442	0.006	0.035	0.460	-0.135	-0.273
7	0.387	0.110	-0.230	0.448	-0.241	-0.164
8	0.191	0.127	-0.185	0.133	-0.038	0.241
9	0.222	0.249	-0.268	0.047	0.193	0.567
10	0.158	0.366	-0.222	-0.075	0.435	0.135
11	0.087	0.412	0.006	-0.175	0.408	-0.384
12	0.002	0.414	0.266	-0.059	-0.023	-0.228
13	-0.077	0.366	0.226	0.026	-0.127	-0.066
14	-0.096	0.308	0.202	0.052	-0.159	-0.088
15	-0.116	0.252	0.142	0.070	-0.153	0.098
16	-0.125	0.156	0.077	0.104	-0.072	0.086
17	-0.165	0.147	0.101	0.135	-0.145	0.234
18	-0.152	0.132	0.081	0.160	-0.124	0.205
19	-0.169	0.092	0.084	0.158	-0.048	0.144
20	-0.194	0.062	0.047	0.211	-0.005	0.129
21	-0.163	0.032	0.034	0.185	0.042	0.136
22	-0.164	0.015	0.006	0.188	0.061	0.020
23	-0.149	-0.007	-0.024	0.183	0.126	-0.004
24	-0.185	-0.015	-0.054	0.215	0.153	-0.037
25	-0.189	-0.035	-0.025	0.213	0.147	-0.117
26	-0.138	-0.027	-0.030	0.153	0.136	-0.082
27	-0.130	-0.035	-0.037	0.162	0.138	-0.076
28	-0.119	-0.035	-0.048	0.162	0.152	-0.050
29	-0.126	-0.044	-0.067	0.167	0.165	-0.074
30	-0.092	-0.028	-0.039	0.107	0.114	-0.051
31	-0.100	-0.036	-0.044	0.125	0.134	-0.043
32	-0.080	-0.032	-0.044	0.088	0.117	-0.083
33	-0.052	-0.021	-0.024	0.066	0.076	-0.030
34	-0.056	-0.028	-0.047	0.071	0.104	-0.044
35	-0.045	-0.029	-0.051	0.072	0.087	-0.071
36	-0.056	-0.028	-0.049	0.078	0.110	-0.049
37	-0.061	-0.032	-0.071	0.094	0.120	-0.061
38	-0.025	-0.019	-0.029	0.051	0.052	-0.044
39	-0.013	-0.011	-0.021	0.015	0.035	-0.044
40	-0.011	-0.011	-0.022	0.009	0.030	-0.017
41	-0.007	-0.006	-0.009	0.010	0.026	-0.030
42	-0.007	-0.005	-0.005	0.007	0.015	-0.018

Band	PC7	PC8	PC9	PC10	PC11	PC12
1	-0.020	-0.028	-0.125	0.056	-0.026	-0.005
2	-0.095	-0.041	-0.392	0.006	-0.278	-0.218
3	-0.170	-0.021	-0.504	0.237	-0.313	-0.062
4	-0.066	0.011	-0.010	0.154	-0.079	0.155
5	0.087	0.014	0.389	-0.100	0.166	0.089
6	0.214	-0.171	-0.051	-0.210	0.132	-0.296
7	-0.022	0.155	-0.111	0.166	-0.188	0.239
8	-0.342	0.196	0.254	0.291	-0.104	0.207
9	-0.298	-0.055	0.026	-0.109	0.099	-0.269
10	0.370	-0.244	-0.257	-0.188	0.104	0.179

11	0.222	0.200	0.239	0.340	-0.243	-0.081
12	-0.329	0.456	-0.097	-0.415	0.126	-0.055
13	-0.213	-0.066	-0.166	0.148	0.384	0.041
14	-0.148	-0.619	0.069	0.314	0.110	0.102
15	-0.052	-0.243	0.256	-0.399	-0.605	0.118
16	0.142	-0.058	0.056	-0.017	-0.170	-0.050
17	0.209	0.079	0.090	0.166	0.043	-0.225
18	0.236	0.181	0.033	0.213	-0.013	-0.508
19	0.162	0.133	-0.115	-0.066	-0.045	0.006
20	0.140	0.151	-0.175	-0.092	-0.072	0.318
21	0.163	0.155	-0.034	0.060	0.160	0.287
22	0.065	0.071	-0.120	0.054	0.092	0.177
23	-0.080	-0.023	-0.063	-0.077	0.112	0.102
24	0.074	-0.032	-0.010	-0.001	-0.041	0.003
25	-0.076	-0.059	0.065	-0.110	-0.069	-0.073
26	-0.129	-0.005	-0.020	0.026	0.016	-0.051
27	-0.101	-0.056	0.058	0.021	0.069	-0.008
28	-0.100	0.031	0.039	-0.037	0.017	-0.049
29	-0.101	-0.003	0.031	0.070	-0.043	-0.044
30	-0.048	0.017	0.081	0.091	0.003	-0.030
31	-0.128	0.021	0.082	0.015	-0.027	-0.072
32	-0.078	-0.005	0.041	0.002	-0.010	-0.087
33	-0.032	-0.024	0.014	-0.005	0.038	-0.009
34	-0.081	0.009	0.089	-0.031	-0.022	-0.008
35	-0.080	-0.066	0.040	-0.035	0.008	-0.036
36	-0.091	-0.080	0.004	-0.062	0.033	-0.052
37	-0.145	-0.111	-0.040	-0.053	-0.001	-0.122
38	-0.026	-0.047	0.010	0.008	0.022	-0.050
39	-0.048	-0.013	0.013	-0.007	0.010	-0.036
40	-0.042	-0.024	0.037	-0.007	-0.036	-0.043
41	-0.022	-0.020	0.010	-0.005	-0.016	-0.022
42	-0.010	-0.011	0.018	-0.018	-0.011	-0.033

Band	PC13	PC14	PC15	PC16	PC17	PC18
1	0.123	0.051	0.088	-0.091	0.043	0.166
2	0.056	-0.100	0.258	0.061	0.008	0.209
3	-0.011	0.070	-0.090	0.013	0.013	-0.236
4	-0.077	-0.082	-0.199	0.026	-0.051	-0.023
5	-0.060	-0.048	0.103	-0.051	-0.040	0.160
6	0.140	0.171	-0.008	0.009	0.175	-0.255
7	-0.108	-0.138	0.028	0.076	-0.165	0.311
8	-0.128	-0.103	-0.034	-0.260	0.081	-0.345
9	0.337	0.201	0.074	0.077	-0.024	0.143
10	-0.358	-0.214	-0.113	-0.015	-0.090	-0.047
11	0.238	0.180	0.103	-0.010	0.109	-0.026
12	0.093	-0.321	-0.089	0.100	-0.132	-0.040
13	-0.399	0.457	0.149	-0.203	0.138	0.160
14	0.315	-0.244	-0.086	0.114	-0.180	-0.055
15	-0.124	0.182	-0.098	0.096	0.091	0.057
16	-0.048	-0.042	0.063	-0.091	0.194	0.134
17	-0.162	-0.369	0.531	0.169	0.022	-0.222
18	-0.125	-0.003	-0.524	-0.034	-0.134	0.099
19	0.090	0.083	-0.038	-0.181	0.057	-0.173
20	0.103	0.132	0.124	0.031	0.061	-0.263
21	0.275	0.246	-0.091	0.376	-0.123	-0.001
22	0.114	0.063	0.055	0.081	-0.089	0.122
23	0.196	-0.232	-0.082	-0.222	0.584	0.092
24	0.204	-0.213	0.045	-0.421	-0.075	0.272
25	-0.014	0.159	0.152	-0.432	-0.546	-0.119
26	-0.066	0.023	-0.221	-0.055	0.069	-0.040
27	-0.082	-0.136	-0.297	0.123	0.093	-0.025
28	-0.120	-0.005	0.030	0.106	-0.075	0.074
29	-0.215	0.011	0.098	0.272	0.020	0.277

30	-0.112	0.076	0.010	0.146	0.209	-0.029
31	-0.051	0.011	0.061	0.136	0.041	0.110
32	0.005	0.006	0.028	0.100	0.042	-0.059
33	-0.010	0.010	0.002	0.095	-0.051	-0.048
34	-0.001	0.046	0.041	0.078	-0.049	-0.063
35	-0.083	0.024	0.011	0.076	-0.063	-0.187
36	-0.029	0.036	0.015	0.135	-0.112	-0.198
37	-0.033	-0.097	0.110	0.007	0.061	-0.180
38	-0.084	0.018	0.017	0.038	0.069	-0.073
39	-0.057	-0.014	0.029	0.036	0.046	-0.020
40	-0.063	0.024	0.016	0.044	0.021	0.030
41	-0.012	-0.019	0.020	0.037	-0.003	-0.018
42	-0.040	0.009	0.026	0.005	0.031	-0.027

Band	PC19	PC20	PC21	PC22	PC23	PC24
1	-0.228	0.019	0.093	0.043	-0.012	0.016
2	-0.422	0.435	-0.055	0.127	0.025	-0.141
3	0.114	-0.300	0.036	-0.016	0.012	0.123
4	0.176	-0.262	-0.064	0.041	-0.169	-0.068
5	0.001	0.259	0.082	-0.070	0.033	0.007
6	-0.163	-0.165	-0.192	0.034	0.074	-0.055
7	0.243	0.121	0.220	0.017	-0.113	0.111
8	-0.423	0.043	-0.117	-0.060	0.097	-0.144
9	0.251	-0.032	-0.035	0.067	-0.033	0.006
10	-0.141	0.025	-0.047	-0.037	0.006	0.080
11	0.061	0.038	0.095	0.086	0.028	-0.021
12	-0.024	-0.068	-0.106	-0.084	-0.014	-0.017
13	-0.002	-0.045	0.110	0.030	-0.043	-0.026
14	0.034	0.142	-0.075	-0.089	0.144	0.040
15	-0.116	-0.116	0.133	0.030	-0.020	0.136
16	0.141	-0.054	-0.255	-0.013	-0.310	-0.492
17	0.035	-0.212	0.029	0.240	-0.107	0.169
18	-0.084	0.087	0.121	-0.214	0.113	0.083
19	0.162	0.327	0.161	0.032	0.022	-0.002
20	0.293	0.228	-0.199	-0.132	0.357	-0.061
21	-0.431	-0.209	0.109	0.043	-0.136	-0.078
22	-0.084	0.050	-0.028	-0.065	-0.271	0.040
23	-0.046	-0.080	0.241	0.071	0.068	0.303
24	-0.021	-0.310	-0.048	-0.228	-0.031	-0.170
25	-0.052	-0.048	-0.073	0.198	0.020	0.192
26	0.076	0.233	-0.345	0.192	-0.284	0.029
27	0.042	0.092	0.063	0.586	0.003	-0.214
28	0.028	-0.112	0.222	0.194	0.471	-0.210
29	0.006	-0.141	-0.279	-0.189	0.328	-0.135
30	-0.017	0.057	-0.260	-0.238	0.006	0.189
31	-0.040	0.037	-0.115	-0.049	0.031	0.429
32	0.011	0.125	-0.008	-0.148	-0.215	0.100
33	0.039	-0.002	0.023	0.058	-0.058	0.044
34	-0.039	-0.003	0.058	-0.014	-0.162	0.018
35	0.010	0.096	0.073	-0.168	-0.248	0.024
36	0.079	0.053	0.257	-0.072	-0.050	-0.225
37	-0.002	0.042	0.405	-0.370	-0.088	-0.222
38	0.025	0.027	0.136	-0.074	-0.046	-0.084
39	0.008	0.041	0.016	-0.040	-0.065	0.001
40	-0.004	-0.025	0.031	-0.031	-0.071	-0.008
41	0.008	0.000	0.033	-0.026	0.013	-0.004
42	-0.018	0.021	0.042	-0.021	-0.023	-0.015

Band	PC25	PC26	PC27	PC28	PC29	PC30
1	-0.010	-0.142	-0.214	-0.285	-0.246	-0.353
2	0.138	-0.130	-0.035	0.118	0.099	0.243
3	-0.179	0.204	0.185	-0.076	-0.220	-0.076
4	0.237	-0.425	-0.323	0.072	0.434	0.078

5	-0.061	0.164	0.078	0.005	-0.143	-0.014
6	0.034	-0.063	0.042	-0.005	0.106	-0.013
7	-0.009	0.043	-0.088	0.012	-0.110	0.006
8	-0.067	-0.004	0.089	-0.045	0.033	-0.016
9	0.061	0.007	0.000	0.019	0.015	0.013
10	-0.046	-0.021	-0.038	-0.026	-0.028	-0.039
11	0.038	0.036	0.019	0.039	0.069	0.038
12	-0.020	-0.009	-0.006	-0.060	-0.036	-0.041
13	0.093	-0.066	0.046	0.118	-0.031	0.051
14	-0.041	-0.003	-0.043	-0.066	-0.016	-0.010
15	0.149	0.089	0.060	0.051	0.037	-0.003
16	-0.540	-0.102	-0.092	-0.098	-0.083	0.022
17	0.135	0.099	-0.023	0.031	-0.015	-0.004
18	-0.037	0.007	-0.004	0.044	-0.017	0.144
19	0.131	-0.162	-0.018	-0.387	0.168	-0.418
20	0.045	-0.080	0.020	0.271	-0.151	0.164
21	-0.037	0.058	-0.195	0.139	-0.156	-0.059
22	-0.094	0.174	0.419	-0.162	0.499	0.098
23	-0.147	0.095	-0.166	-0.105	0.108	0.265
24	0.347	-0.099	0.250	0.158	-0.249	0.007
25	-0.159	0.032	-0.191	-0.010	0.096	0.003
26	0.160	0.494	-0.239	0.099	-0.060	0.008
27	0.108	-0.158	0.344	0.155	-0.106	-0.215
28	-0.169	-0.048	-0.229	-0.066	0.031	0.201
29	0.089	0.191	-0.003	-0.262	0.190	-0.231
30	0.223	-0.111	-0.124	-0.140	-0.127	0.044
31	-0.365	-0.368	0.272	0.232	0.032	-0.196
32	-0.155	-0.166	-0.274	0.230	-0.078	-0.136
33	0.065	-0.059	0.036	-0.148	-0.169	0.049
34	0.168	-0.158	0.005	-0.355	-0.204	0.144
35	0.056	-0.205	0.116	-0.135	-0.082	0.324
36	-0.112	0.031	0.104	-0.150	-0.011	0.226
37	0.050	0.168	-0.102	0.339	0.199	-0.314
38	0.062	0.007	0.046	-0.075	-0.080	-0.007
39	-0.006	-0.024	-0.012	0.013	-0.029	-0.094
40	-0.018	0.045	-0.020	-0.034	-0.129	0.063
41	-0.009	0.023	-0.010	-0.058	-0.002	-0.005
42	-0.005	-0.031	-0.006	-0.019	-0.095	0.006

Band	PC31	PC32	PC33	PC34	PC35	PC36
1	0.497	-0.380	0.290	-0.070	-0.010	-0.073
2	-0.065	0.131	-0.043	0.032	-0.047	0.066
3	-0.109	0.009	-0.085	-0.071	0.047	-0.051
4	0.128	-0.025	0.101	0.111	-0.051	0.064
5	-0.025	0.019	-0.020	-0.042	0.003	-0.019
6	-0.012	-0.067	0.037	0.049	0.024	-0.002
7	0.029	0.080	-0.050	-0.033	-0.019	-0.002
8	-0.055	-0.071	0.017	0.035	-0.017	0.050
9	0.019	0.010	-0.005	-0.066	-0.004	-0.018
10	-0.013	-0.013	-0.003	0.018	0.019	-0.000
11	0.029	0.008	0.031	0.011	-0.021	-0.011
12	0.045	0.007	-0.008	-0.041	0.003	0.017
13	-0.064	0.008	-0.005	0.035	-0.033	0.002
14	-0.038	0.013	-0.018	0.024	0.040	-0.019
15	0.002	-0.106	-0.005	-0.014	-0.012	0.025
16	0.032	0.178	-0.013	-0.050	0.094	0.022
17	0.045	-0.072	-0.005	0.021	-0.038	-0.040
18	0.056	-0.000	0.170	-0.050	-0.020	0.004
19	-0.211	0.059	-0.313	0.184	-0.012	-0.016
20	0.138	-0.060	0.319	-0.036	-0.027	-0.037
21	-0.035	0.218	-0.148	0.071	-0.067	-0.041
22	0.027	-0.328	0.104	-0.197	0.216	0.077
23	-0.016	0.159	0.092	-0.055	-0.169	-0.014

24	-0.082	-0.138	-0.175	0.136	-0.025	0.027
25	0.083	0.235	0.052	-0.230	-0.102	0.029
26	0.103	-0.153	-0.033	0.380	0.099	-0.068
27	0.055	0.075	0.050	-0.316	-0.064	-0.067
28	-0.096	-0.328	-0.218	0.085	0.427	-0.058
29	-0.119	0.135	0.171	-0.004	-0.343	-0.134
30	0.254	0.127	-0.449	-0.409	0.213	0.205
31	0.149	0.079	-0.048	0.447	0.079	-0.005
32	-0.530	-0.439	0.018	-0.246	-0.267	0.030
33	-0.164	0.070	0.272	0.037	0.150	0.720
34	-0.270	0.204	0.356	0.192	0.270	-0.235
35	0.085	-0.022	-0.051	-0.120	-0.044	-0.482
36	0.284	-0.110	-0.228	0.242	-0.484	0.216
37	0.138	0.219	0.126	-0.055	0.296	-0.013
38	0.041	0.063	0.150	0.051	-0.168	0.151
39	0.004	-0.047	0.051	0.031	0.027	-0.036
40	0.046	-0.059	0.073	0.078	0.026	0.050
41	0.066	-0.050	0.011	0.017	-0.001	0.090
42	0.038	-0.146	0.005	0.028	0.008	0.074

Band	PC37	PC38	PC39	PC40	PC41	PC42
1	-0.059	0.137	0.049	-0.069	-0.017	0.007
2	0.016	-0.028	-0.011	-0.000	0.005	-0.006
3	-0.006	-0.003	0.012	0.006	0.018	-0.002
4	0.017	-0.046	-0.019	0.013	-0.005	0.016
5	-0.005	0.019	0.002	-0.004	0.002	-0.009
6	-0.009	-0.000	-0.019	-0.018	-0.014	0.003
7	-0.013	0.001	0.007	0.020	0.011	0.002
8	0.012	0.016	-0.012	-0.018	-0.009	-0.004
9	0.036	-0.019	0.013	0.018	0.014	-0.002
10	-0.032	-0.018	0.002	-0.007	-0.008	-0.001
11	0.020	0.033	-0.014	-0.001	0.004	0.012
12	0.021	-0.029	0.026	-0.008	-0.011	0.020
13	-0.048	0.035	-0.028	0.010	-0.017	-0.038
14	0.022	-0.054	-0.000	0.003	0.003	0.028
15	-0.016	0.040	0.032	-0.017	0.032	-0.013
16	0.003	0.026	0.028	0.017	0.008	-0.018
17	-0.024	0.022	0.011	-0.013	-0.005	0.003
18	-0.048	-0.002	-0.020	0.005	0.031	-0.019
19	0.117	0.004	-0.069	-0.014	-0.065	0.034
20	-0.040	-0.034	-0.033	-0.009	0.000	-0.014
21	0.106	-0.007	-0.016	0.036	0.019	-0.012
22	-0.046	-0.112	0.039	-0.017	0.010	0.020
23	0.006	0.019	-0.057	0.026	0.018	0.008
24	0.009	0.010	0.001	-0.019	0.014	-0.030
25	0.020	-0.092	-0.015	0.022	0.022	0.016
26	0.021	0.090	0.080	0.011	-0.006	-0.008
27	-0.064	-0.040	-0.055	0.014	-0.046	-0.007
28	0.088	0.010	0.108	-0.062	0.027	0.010
29	0.060	0.121	-0.081	0.061	-0.020	0.052
30	-0.160	-0.124	0.045	-0.037	0.016	0.004
31	0.002	0.020	0.067	0.002	-0.046	0.006
32	-0.126	-0.019	0.046	-0.044	-0.067	-0.033
33	0.208	0.437	-0.102	-0.026	0.029	0.045
34	-0.464	-0.201	0.045	0.039	0.014	-0.069
35	0.445	0.392	-0.037	0.019	-0.050	0.026
36	-0.385	0.024	-0.080	-0.055	0.063	-0.003
37	-0.077	0.098	-0.039	0.026	-0.072	0.039
38	0.395	-0.430	0.689	-0.114	0.030	0.031
39	0.202	-0.233	-0.358	-0.169	0.825	-0.158
40	0.235	-0.452	-0.536	-0.324	-0.462	0.257
41	0.162	-0.141	-0.135	0.286	-0.257	-0.872
42	0.101	-0.216	-0.121	0.862	0.082	0.358

Appendix G

Eigenvalues for COUP 444 Subset

Component	Eigenvalue	Proportion	Cumulative
1	5485800000000	0.623	0.623
2	2703700000000	0.307	0.930
3	193090000000	0.022	0.952
4	166530000000	0.019	0.971
5	83088000000	0.009	0.980
6	36108000000	0.004	0.984
7	14359000000	0.002	0.986
8	12044000000	0.001	0.987
9	10638000000	0.001	0.989
10	9161318129	0.001	0.990
11	8661918888	0.001	0.991
12	8278939401	0.001	0.992
13	6213038245	0.001	0.992
14	5936007079	0.001	0.993
15	5559488440	0.001	0.994
16	5075589349	0.001	0.994
17	4989070817	0.001	0.995
18	4084620225	0.000	0.995
19	3981628161	0.000	0.996
20	3518224813	0.000	0.996
21	3273573251	0.000	0.996
22	3071969487	0.000	0.997
23	2683702815	0.000	0.997
24	2638508387	0.000	0.997
25	2544445558	0.000	0.998
26	2336962812	0.000	0.998
27	2193127824	0.000	0.998
28	1932914749	0.000	0.998
29	1867531668	0.000	0.999
30	1689332097	0.000	0.999
31	1529087747	0.000	0.999
32	1418222255	0.000	0.999
33	1344162080	0.000	0.999
34	1239097892	0.000	0.999
35	1066209823	0.000	1.000
36	918758024	0.000	1.000
37	721826174	0.000	1.000
38	647750008	0.000	1.000
39	424973629	0.000	1.000
40	300182087	0.000	1.000
41	193401860	0.000	1.000
42	153394869	0.000	1.000

Appendix H

Class Assignments After Each Clustering Method

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
471	1	1	
510	1	1	
625	1	1	
647	1	1	
680	1	1	
723	1	1	
797	1	1	
441	2	2	
466	2	2	
507	2	2	
539	2	2	
563	2	2	
572	2	2	
598	2	2	
599	2	2	
640	2	2	
667	2	2	
1123	2	2	
1139	2	2	
376	3	3	
554	3	3	
614	3	3	
655	3	3	
780	3	3	
1075	3	3	
1223	3	3	
1297	3	3	
1364	3	3	
241	4	4	
246	4	4	
385	4	4	
407	4	4	
415	4	4	
424	4	4	
533	4	4	
595	4	4	
817	4	4	
1054	4	4	
1062	4	4	
1067	4	4	
1234	4	4	
1298	4	4	
1439	4	4	

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
1469	4	4	
1480	4	4	
1535	4	4	
986	5	5	
1053	5	5	
111	6	6	
224	6	6	
292	6	6	
548	6	6	
896	6	6	
970	6	6	
1041	6	6	
1128	6	6	
1165	6	6	
314	7	7	
319	7	7	
353	7	6	*
391	7	7	
427	7	7	
472	7	4	*
712	7	7	
713	7	7	
849	7	7	
852	7	7	
878	7	7	
892	7	7	
919	7	7	
1056	7	7	
1081	7	6	*
1137	7	7	
1200	7	6	*
1276	7	7	
1284	7	7	
1296	7	7	
1308	7	7	
1579	7	7	
1607	7	6	*
1609	7	6	*
165	8	8	
192	8	8	
332	8	8	
418	8	8	
435	8	8	
481	8	8	
520	8	8	
610	8	8	
627	8	8	
653	8	8	

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
751	8	8	
1035	8	8	
1112	8	8	
1147	8	8	
1154	8	9	*
1167	8	8	
1399	8	8	
1430	8	8	
1471	8	8	
1544	8	8	
1561	8	8	
8	9	9	
321	9	9	
331	9	9	
338	9	9	
365	9	9	
561	9	9	
658	9	9	
697	9	9	
837	9	9	
1071	9	9	
1140	9	9	
1456	9	9	
65	10	8	*
137	10	10	
172	10	9	*
230	10	10	
238	10	9	*
269	10	9	*
483	10	9	*
536	10	9	*
664	10	9	*
902	10	9	*
1110	10	9	*
1258	10	10	
1357	10	10	
1603	10	9	*
55	11	11	
90	11	10	*
115	11	11	
183	11	10	*
223	11	10	*
227	11	11	
236	11	11	
250	11	10	*
260	11	10	*
262	11	10	*
301	11	11	

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
310	11	11	
322	11	10	*
323	11	11	
373	11	10	*
414	11	11	
454	11	10	*
485	11	10	*
514	11	11	
515	11	11	
612	11	11	
624	11	10	*
645	11	10	*
649	11	10	*
660	11	10	*
789	11	11	
790	11	10	*
823	11	10	*
897	11	10	*
921	11	11	
939	11	10	*
949	11	11	
976	11	11	
998	11	11	
1000	11	11	
1008	11	10	*
1028	11	10	*
1045	11	11	
1070	11	10	*
1074	11	11	
1097	11	11	
1104	11	10	*
1120	11	11	
1141	11	11	
1158	11	11	
1191	11	11	
1210	11	11	
1231	11	10	*
1245	11	10	*
1275	11	10	*
1290	11	10	*
1292	11	11	
1302	11	10	*
1316	11	10	*
1344	11	10	*
1356	11	11	
1391	11	10	*
1407	11	11	
1409	11	10	*

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
1410	11	11	
1419	11	11	
1450	11	11	
1474	11	11	
1485	11	10	*
1503	11	11	
1529	11	10	*
1531	11	11	
1550	11	11	
11	12	12	
49	12	12	
110	12	11	*
117	12	12	
174	12	12	
217	12	12	
256	12	11	*
304	12	12	
308	12	11	*
368	12	11	*
404	12	12	
446	12	12	
490	12	11	*
550	12	11	*
566	12	12	
626	12	11	*
663	12	12	
737	12	12	
753	12	12	
756	12	11	*
776	12	11	*
801	12	12	
857	12	11	*
885	12	12	
899	12	12	
992	12	11	*
1019	12	12	
1086	12	11	*
1100	12	12	
1111	12	12	
1117	12	12	
1127	12	12	
1193	12	11	*
1246	12	12	
1261	12	12	
1264	12	11	*
1374	12	12	
1382	12	11	*
1449	12	12	

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
1462	12	12	
1464	12	12	
1466	12	11	*
1478	12	12	
1570	12	12	
17	13	13	
29	13	13	
154	13	13	
177	13	13	
226	13	13	
244	13	13	
312	13	11	*
431	13	13	
499	13	13	
648	13	13	
671	13	13	
710	13	13	
750	13	13	
783	13	13	
856	13	12	*
903	13	13	
1058	13	13	
1101	13	13	
1103	13	13	
1132	13	13	
1149	13	12	*
1155	13	13	
1161	13	13	
1172	13	13	
1206	13	13	
1216	13	13	
1235	13	11	*
1336	13	13	
1369	13	13	
1447	13	13	
1475	13	13	
1588	13	11	*
6	14	14	
28	14	14	
40	14	15	*
43	14	14	
54	14	14	
62	14	14	
66	14	14	
67	14	15	*
96	14	14	
109	14	15	*
112	14	14	

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
113	14	14	
122	14	14	
134	14	14	
139	14	14	
141	14	14	
173	14	15	*
179	14	15	*
197	14	14	
202	14	14	
205	14	14	
218	14	14	
253	14	14	
270	14	15	*
309	14	14	
325	14	15	*
379	14	14	
382	14	14	
387	14	14	
410	14	15	*
413	14	14	
459	14	15	*
470	14	15	*
488	14	15	*
489	14	14	
498	14	15	*
513	14	14	
517	14	15	*
545	14	14	
557	14	14	
565	14	14	
602	14	14	
604	14	14	
616	14	15	*
631	14	13	*
665	14	15	*
666	14	14	
672	14	14	
700	14	14	
711	14	14	
726	14	15	*
739	14	15	*
754	14	14	
763	14	15	*
798	14	14	
807	14	14	
862	14	15	*
865	14	13	*
888	14	15	*

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
914	14	15	*
924	14	14	
936	14	14	
960	14	14	
969	14	15	*
972	14	14	
1007	14	14	
1009	14	14	
1076	14	14	
1095	14	15	*
1121	14	14	
1126	14	15	*
1131	14	14	
1134	14	14	
1135	14	15	*
1143	14	14	
1150	14	15	*
1151	14	14	
1169	14	15	*
1177	14	15	*
1202	14	15	*
1212	14	15	*
1233	14	15	*
1236	14	15	*
1242	14	14	
1279	14	14	
1282	14	14	
1291	14	13	*
1306	14	15	*
1311	14	15	*
1355	14	15	*
1384	14	14	
1387	14	14	
1388	14	15	*
1398	14	15	*
1423	14	15	*
1424	14	14	
1429	14	15	*
1432	14	15	*
1433	14	14	
1455	14	14	
1463	14	15	*
1487	14	14	
1492	14	15	*
1521	14	14	
1546	14	15	*
1585	14	14	
1608	14	14	

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
1610	14	15	*
20	15	15	
21	15	15	
69	15	15	
89	15	15	
100	15	15	
114	15	16	*
118	15	15	
119	15	15	
132	15	16	*
133	15	15	
169	15	16	*
249	15	15	
255	15	15	
266	15	15	
276	15	15	
296	15	15	
328	15	15	
340	15	15	
389	15	15	
395	15	15	
468	15	15	
553	15	15	
695	15	15	
937	15	15	
966	15	15	
974	15	15	
1066	15	15	
1207	15	15	
1360	15	15	
1373	15	16	*
1404	15	15	
1411	15	15	
1438	15	15	
1440	15	15	
1454	15	15	
1512	15	15	
1516	15	15	
1524	15	15	
1539	15	15	
1543	15	15	
1553	15	15	
1564	15	15	
1571	15	16	*
1572	15	15	
1591	15	15	
1594	15	15	
1595	15	15	

COUP Source Number	Hierarchical Clustering Class Membership	K-Means Class Membership	Source Changed Classes Marked With *
1612	15	15	
1616	15	15	
60	16	16	
64	16	16	
128	16	16	
164	16	16	
294	16	16	
300	16	16	
585	16	16	
864	16	16	
869	16	16	
1199	16	16	
1415	16	16	
1457	16	16	
1507	16	16	
1537	16	16	
948	17	17	

REFERENCES

[1] Weisskopf, M.C., Brinkman, B., Canizares, C., Garmire, G., Murray, S., and Van Speybroeck, L.P., “An Overview of the Performance and Scientific Results from the Chandra X-ray Observatory,” Pub of the Astronomical Society of the Pacific, Vol. 114, pp. 1-24, 2002.

[2] Feigelson, E.D. and Montmerle, T., “High-Energy Processes in Young Stellar Objects,” Annual Review of Astronomy and Astrophysics, Vol. 37, pp. 363-408, 1999.

[3] Kastner, J.H, Huenemoerder, D.P., Schulz, N.S., Canizares, C.R., and Weintraub, D.A., “Evidence for Accretion: High-Resolution X-Ray Spectroscopy of the Classical T Tauri Star TW Hydrae,” The Astrophysical Journal, Vol. 567, pp.434-440, 2002.

[4] Kastner, J.H., Crigger, L., Rich, M., and Weintraub, D., “Rosat X-ray Spectral Properties of Nearby Young Associations: TW Hydrae, Tucana-Horologium, and the Beta Pictoris Moving Group,” The Astrophysical Journal, Vol. 585, pp. 878-884, 2003.

[5] Feigelson, E.D., Broos, P., Gaffney III, J.A., Garmire, G., Hillenbrand, L.A., Pravdo, S.H., Townsley, L., and Tsuboi, Y., “X-Ray Emitting Young Stars in the Orion Nebula,” The Astrophysical Journal, Vol. 574, Issue 1, pp. 258-292, 2002.

[6] Flaccomio, E., Damiani, F., Micela, G., Sciortino, S., Harnden, F.R., Murray, S.S., and Wolk, S.J., “Chandra X-Ray Observation of the Orion Nebula Cluster. II. Relationship between X-Ray Activity Indicators and Stellar Parameters,” The Astrophysical Journal, Vol. 582, pp. 398-409, 2003.

[7] Stassun, K.G., Mathieu, R.D., Vaz, L.P.R., Stroud, N., and Vrba, F.J., “Dynamical Mass Constraints on Low-Mass Pre-Main-Sequence Stellar Evolutionary Tracks: An Eclipsing Binary in Orion with a 1.0 Msolar Primary and a 0.7 Msolar Secondary,” The Astrophysical Journal Supplement Series, Vol. 151, pp. 357-385, 2004.

[8] Garmire, G., Feigelson, E.D., Broos, P., Hillenbrand, L.A., Pravdo, S.H., Townsley, L., and Tsuboi, Y., “Chandra X-ray Observatory study of the Orion Nebular Cluster and BN/KL region,” The Astronomical Journal, Vol. 120, No. 3, pp. 1426-1435, 2000.

[9] Schulz, N.S., Canizares, C.R., Huenemoerder, D.P., Kastner, J.H., Taylor, S.C., and Bergstrom, E., “Chandra Observations of Variable Embedded X-ray Sources in Orion. I. Resolving the Orion Trapezium,” The Astrophysical Journal, Vol. 549, pp. 441-451, 2001.

[10] Feigelson, E.D., Gaffney, J.A.III, Garmire, G., Hillenbrand, L.A., and Townsley, L., “X-Rays in the Orion Nebula Cluster: Constraints on the Origins of Magnetic Activity in Pre-Main-Sequence Stars,” The Astrophysical Journal, Vol. 584, Issue 2, pp. 911-930, 2003.

[11] Preibisch, T., Kim, Y., Favata, F., Feigelson, E.D., Flaccomio, E., Getman, K., Micela, G., Sciortino, S., Stassun, K., Stelzer, B., Zinnecker, H., “The Origin of T Tauri X-ray Emission: New Insights from the *Chandra* Orion Ultradeep Project,” The Astrophysical Journal Supplement, in press, 2005.

[12] Getman, K.V., Flaccomio, E., Broos, P.S., Grosso, N., Tsujimoto, M., Townsley, L., Garmire, G.P., Kastner, J., Li, J., Harnden, Jr., F.R., Wolk, S., Murray, S.S., Lada, C.J., Muench, A.A., McCaughrean, M.J., Meeus, G., Damiani, F., Micela, G., Sciortino, S., Bally, J., Hillenbrand, L.A., Herbst, W., Preibisch, T., and Feigelson, E.D., “*Chandra* Orion Ultradeep Project: Observations and Source Lists”, The Astrophysical Journal Supplement, in press, 2005.

[13] Tucker, W. and Giacconi, R., The X-ray Universe, Harvard University Press, Cambridge, MA, pp. 23-24, 1985.

[14] Giacconi, R., Murray, S., Gursky, H., Kellogg, E., Schreier, E., and Tananbaum, H., The Astrophysical Journal, Vol. 178, pp. 281-308, 1972.

[15] Charles, P.A. and Seward, F.D., Exploring the X-ray Universe, Cambridge University Press, Cambridge, UK, pp. 370-372, 1995.

[16] Kastner, J.H., Crigger, L., Rich, M., and Weintraub, D., “Rosat X-ray Spectral Properties of Nearby Young Associations: TW Hydrae, Tucana-Horologium, and the Beta Pictoris Moving Group,” The Astrophysical Journal, Vol. 585, pp. 878-884, 2003.

[17] Charles, P.A. and Seward, F.D., Exploring the X-ray Universe, Cambridge University Press, Cambridge, UK, pp. 374, 1995.

[18] Schlegel, E.M., The Restless Universe: Understanding X-ray Astronomy in the Age of Chandra and Newton, Oxford University Press, New York, pp. 8-10, 2002.

[19] Charles, P.A. and Seward, F.D., Exploring the X-ray Universe, Cambridge University Press, Cambridge, UK, pg. 37, 1995.

[20] Wachter, K., Leach, R., and Kellogg, E., “Parameter Estimation in X-ray Astronomy Using Maximum Likelihood,” The Astrophysical Journal, Vol. 230, pp. 274-287, 1979.

[21] Shu, F.H., Shang, H., Glassgold, A.E., and Lee, T., “X-rays and Fluctuating X-Winds from Protostars,” Science, Vol. 277, pp. 1475-1479, 1997.

[22] Feigelson, E.D., “*Chandra* Studies of Star Forming Regions,” X-ray Astronomy 2000, ASP Conference Series, R. Giacconi, L. Stella, and S. Serio Eds., 2001.

[23] Tucker, W. and Giacconi, R., The X-ray Universe, Harvard University Press, Cambridge, MA, pg. 48, 1985.

[24] Chandra Proposers’ Observatory Guide, Rev. 4.0, pg. 73, Dec. 2004.

[25] Townsley, L.K., Broos, P.S., Garmire, G.P., and Nousek, J.A., “Mitigating Charge Transfer Inefficiency in the Chandra X-ray Observatory Advanced CCD Imaging Spectrometer,” The Astrophysical Journal, Vol. 534, pp. L139-L142, 2000.

[26] van Dyk, D.A., “Highly-Structured Statistical Models in High-Energy Astrophysics,” PHYSTAT2003, SLAC, Stanford, CA, 2003.

[27] Davis, J.E., “Event Pileup in Charge-coupled Devices,” The Astrophysical Journal, Vol. 562, pp. 575-582, 2001.

[28] Wheeler, J.A. and Zurek, H., Quantum Theory and Measurement, Princeton University Wheeler and Zurek, eds., Press, Princeton, 1983 (contains translation into English of Heisenberg, W. “Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik,” Zeitschrift für Physik, Vol. 43 pp. 172-198, 1927.)

[29] Plummer, D. and Subramanian, S., “The Chandra Automatic Data Processing Infrastructure,” in ASP Conf. Ser., Astronomical Data Analysis Software and Systems X, eds. F.R. Harnden, Jr., F.A.Primini, & H.E. Payne, Vol. 238, San Francisco, 2001.

[30] Fu, K. and Rosenfeld, A., “Pattern Recognition and Image Processing,” IEEE Transactions on Computers, Vol. 25, pp. 1336-1346, 1976.

[31] Argialas, D.P. and Harlow, C.A., “Computational Image Interpretation Models: An Overview and a Perspective”, Photogrammetric Engineering and Remote Sensing, Vol. 56, No. 6, pp. 871-886, 1990.

[32] Duda, R.O., Hart, P.E., and Stork, D.G., Pattern Classification, Second Edition, John Wiley & Sons, Inc., New York, pg. 582, 2001.

[33] Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., and Black, W.C., Multivariate Data Analysis, Prentice-Hall Inc., New Jersey, 1998.

[34] Wozniak, P.R., Akerlof, C., Amrose, S., Brumby, S., and 14 more authors, “Classification of ROTSE Variable Stars using Machine Learning”, AAS 199th Meeting, 130.04, 2001.

[35] Huber, R. and Dutra, L.V., “Classifier Combination and Feature Selection for Land-Cover Mapping from High-Resolution Airborne Dual-Band SAR Data,” Proc World Multiconf Systemics, Cybernetics and Informatics, Vol. V, pp.370-375, 2000.

[36] Bazell, D. and Aha, D.W., “Ensembles of Classifiers for Morphological Galaxy Classification,” The Astrophysical Journal, Vol. 548, pp. 219-223, 2001.

[37] Whitmore, B.C., “An objective classification system for spiral galaxies I. The two dominant dimensions,” The Astrophysical Journal, Vol. 278, pp. 61-80, 1984.

[38] Burda, P. and Feitzinger, J.V., “Galaxy classification using pattern recognition methods,” Astronomy and Astrophysics, Vol. 261, pp. 697-705, 1992.

[39] Hodge, P. and Kennicutt, R.C., AJ 88/296, PAPS Doc. ANJOA 88-296-300, Physics Auxiliary Publication Service, American Institute of Physics, New York, 1983.

[40] Garcia, A., Molina, R., and Perez de la Blanca, N., “Automatic Characterization of Spiral and Elliptical Galaxies from Digital Images,” Pattern Recognition Letters, Vol. 15, No. 9, pp. 861-869, 1994.

[41] Sandage, A., The Hubble Atlas of Galaxies, Carnegie Institution of Washington, 1961.

[42] Hanslmeier, A., Veronig, M., Steinegger, M., Brunner, G., Gonzi, S., Temmer, M., Otruba, W., and Messerotti, M., “Solar Activity Monitoring - a New Approach Using Combined Datasets, Pattern Recognition and Neural Networks,” Hvar Observatory Bulletin, Vol. 23, No. 1, pp. 31-39, 1999.

[43] Mashchenko, S., “On the Analysis of HI Distributions Using a Pattern Recognition Approach,” Intl. Symposium on Astrophysics Research and Science Education, Vatican Observatory, C.D. Impey, Ed., pp. 277-281, 1999.

[44] Eyer, L. and Blake, C., “Automated classification of variable stars for ASAS data,” ASP Conference Series: Radial and Nonradial Pulsations as Probes of Stellar Physics, C. Aerts, T. Bedding, and J. Christensen-Dalsgaard, Eds., Vol. N, 2002.

[45] Cheeseman, P. and Stutz, J., “Bayesian Classification (AutoClass): Theory and Results”, in Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., AAAI Press/MIT Press, 1996.

[46] Buccheri, R., DiGesu, V., Maccarone, M.C., and Sacco, B., “High resolution cluster method for topological studies of the light curves of gamma-ray pulsars,” Astronomy and Astrophysics, Vol. 201, pp. 194-198, 1988.

[47] Heck, A., Albert, A., Defays, D., and Mersch, G., “Detection of Errors in Spectral Classification by Cluster Analysis,” Astronomy and Astrophysics, Vol. 61, pp. 563-566, 1977.

[48] Bailer-Jones, C. A., “Neural Network Classification of Stellar Spectra,” Publications of the Astronomical Society of the Pacific, Vol. 109, p. 932, 1997.

[49] Vieira, E.F and Ponz, J.D., “Automated Spectral Classification Using Neural Networks,” Astronomical Data Analysis Software and Systems VII, ASP Conference Series, Vol. 145, 1998.

[50] Hauck, B. and Lindemann, E., Astronomy and Astrophysics Suppl., Vol. 11, p. 119, 1973.

[51] Yin, L.I., Trombka, J.I., Seltzer, S.M., Johnson, R.G., and Philpotts, J.A., “Possible use of pattern recognition for the analysis of Mars rover X-ray fluorescence spectra,” Journal of Geophysical Research, Vol. 94, No. B10, pp. 13611-13618, 1989.

[52] Avdyushin, S.I., Berlyand, B.O., Dernshteyn, P.B., and Burov, B.A., “Classification of regions of solar activity based on methods of pattern recognition theory,” Space Biol. and Aerospace Med., Vol. 17, No. 3, 1983.

[53] Collura, A., Micela, G., Sciortino, S., Harnden Jr., F.R., and Rosner, R., “An Objective Multicolor Method for the Characterization of Low-Resolution X-ray Spectra,” The Astrophysical Journal, Vol. 446, pp. 108-114, 1995.

[54] Babu, G.J. and Feigelson, E.D., Astrostatistics, Chapman & Hall, London, pg. 138, 1996.

[55] Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Fourth Edition, Prentice-Hall Inc., New Jersey, pp. 204-206, 1998.

[56] Manly, B.F., Multivariate Statistical Methods: A Primer, 2nd Edition, Chapman & Hall, London, pp. 132-133, 1994.

[57] Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Fourth Edition, Prentice-Hall Inc., New Jersey, pg. 752, 1998.

[58] Flury, B., Common Principal Components and Related Multivariate Models, John Wiley & Sons, Inc., New York, pp. 1, 1988.

[59] Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Fourth Edition, Prentice-Hall Inc., New Jersey, pg. 458, 1998.

[60] Jackson, J.E., A User's Guide To Principal Components, John Wiley & Sons, Inc., New York, pg. xv, 1991.

[61] Babu, G.J. and Feigelson, E.D., Astrostatistics, Chapman & Hall, London, pp. 129-147, 1996.

[62] Wall, J.V. and Jenkins, C.R., Practical Statistics for Astronomers, Cambridge University Press, Cambridge, UK, pp. 70-74, 2003.

[63] Babu, G.J. and Feigelson, E.D., Astrostatistics, Chapman & Hall, London, pg. 135, 1996.

[64] Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Fourth Edition, Prentice-Hall Inc., New Jersey, pg. 459, 1998.

[65] Babu, G.J. and Feigelson, E.D., Astrostatistics, Chapman & Hall, London, pg. 136, 1996.

[66] Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Fourth Edition, Prentice-Hall Inc., New Jersey, pp. 728-729, 1998.

[67] Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Fourth Edition, Prentice-Hall Inc., New Jersey, pg. 755, 1998.

[68] Getman, K.V., Feigelson, E.D., Grosso, N., McCaughrean, M.J., Micela, G., Broos, P., Garmire, G., Townsley, L., “Membership of the Orion Nebula Population from the *Chandra* Orion Ultradeep Project,” The Astrophysical Journal Supplement, in press, 2005.

[69] Arnaud, K., “Abundances in the Intra-cluster Medium,” Science Presentation, http://lheawww.gsfc.nasa.gov/users/kaa/abund_conf/arnaudk.html, 1995.

[70] Fabian, A.C., Iwasawa, K., Reynolds, C.S., and Young, A.J., Publications of the Astronomical Society of the Pacific, Vol. 112, pp. 1145-1161, 2000.

[71] Huenemoerder, D.P., Canizares, C.R., Drake, J.J, and Sanz-Forcada, J., “The Coronae of AR Lacertae”, The Astrophysical Journal, Vol. 595, pp. 1131-1147, 2003.

[72] Freeman, P.E., Kashyap, V., Rosner, R., and Lamb, D.Q., “A Wavelet-Based Algorithm for the Spatial Analysis of Poisson Data”, The Astrophysical Journal Suppl, Vol. 138, pp. 185-218, 2002.

[73] Jackson, J.E., A User's Guide To Principal Components, John Wiley & Sons, Inc., New York, pg. 44-47, 1991.

[74] Cattell, R.B., “The scree test for the number of factors,” Multivariate Behavioral Research, Vol. 1, pp. 245-276, 1966.

[75] Cattell, R.B. and Jaspers, J. “A general plasmode (No. 30-10-5-2) for factor analytic exercises and research,” Multivariate Behavioral Research Monographs, 67-3, pp. 1-212, 1967.

[76] Broos, P.S., Townsley, L.K., Getman, K., and Bauer, F.E., ACIS Extract, An ACIS Point Source Extraction Package, University Park: Pennsylvania State University, 2002.

[77] Jackson, J.E., A User's Guide To Principal Components, John Wiley & Sons, Inc., New York, pp. 84-85, 1991.

[78] Gleason, T.C. and Staelin, R., “A Proposal for handling missing data”, Psychometrika, Vol. 40, pp. 229-252, 1975.

[79] Bartlett, M.S. “Tests of significance in factor analysis,” Br. J. Psych. Stat. Sec., Vol. 3, pp. 77-85, 1950.

[80] Lawley, D.N., “Tests of significance for the latent roots of covariance and correlation matrices,” Biometrika, Vol. 43, pp. 128-136, 1956.

[81] Levene, H., Contributions to Probability and Statistics, Stanford Univ. Press, CA, pp 278-292, 1960.

[82] Horn, J.L., “A rationale and test for the number of factors in factor analysis,” Psychometrika, Vol. 30, pp. 179-185, 1965.

[83] Jolliffe, I.T., Principal Component Analysis, Springer-Verlag, New York, pg. 95, 1986.

[84] Jackson, J.E., A User's Guide To Principal Components, John Wiley & Sons, Inc., New York, pp. 47-48, 1991.

[85] Jolliffe, I.T., “Discarding variables in principal component analysis. I: Artificial data,” Appl. Stat., Vol. 21, pp. 160-173, 1972.

[86] Legendre, P. and Legendre, L., Numerical Ecology, Second English Edition, Elsevier Science B.V., Amsterdam, pg 465, 1998.

[87] Andrews, D.F., “Plots of high-dimensional data,” Biometrics, Vol. 28, pp. 125-136, 1972.

[88] Manly, B.F., Multivariate Statistical Methods: A Primer, 2nd Edition, Chapman & Hall, London, pg. 33, 1994.

[89] Jolliffe, I.T., Principal Component Analysis, Springer-Verlag, New York, pg. 91, 1986.

[90] Schlegel, E.M., The Restless Universe: Understanding X-ray Astronomy in the Age of Chandra and Newton, Oxford University Press, New York, pp. 116-118, 2002.

[91] Stelzer, B., Flaccomio, E., Montmerle, T., Micela, G., Sciortino, S., Favata, F., Preibisch, T., and Feigelson, E.D., “X-ray emission from early-type stars in the Orion Nebula Cluster,” The Astrophysical Journal Supplement, in press, 2005.

[92] Hillenbrand, L.A., “On the Stellar Population and Star-Forming History of the Orion Nebula Cluster,” The Astronomical Journal, Vol. 113, pp. 1733-1768, 1997.

[93] Siess, L., Dufour, E., and Forestini, M., “An internet server for pre-main sequence tracks of low- and intermediate-mass stars,” Astronomy and Astrophysics, Vol. 358, pp. 593-599, 2000.

[94] Hillenbrand, L.A., Strom, S.E., Calvet, N., Merrill, K.M., Gatley, I., Makidon, R.B., Meyer, M.R., and Skrutskie, M.F., “Circumstellar Disks in the Orion Nebula Cluster,” The Astronomical Journal, Vol. 116, pp. 1816-1841, 1998.

[95] Tsujimoto, M., Feigelson, E.D., Grosso, N., Micela, G., Tsuboi, Y., Favata, F., Shang, H., and Kastner, J.H., “Iron fluorescent line emission from young stars in the Orion Nebula,” The Astrophysical Journal Supplement, in press, 2005.

[96] Kastner, J.H., Franz, G., Grosso, N., Bally, J., McCaughrean, M.J., Getman, K., Feigelson, E.D., and Schulz, N., “X-ray Emission from Orion Nebula Cluster Stars with Circumstellar Disks and Jets,” The Astrophysical Journal Supplement, in press, 2005.