Fast and Accurate Estimation for Astrophysical Problems in Large Databases

Joseph Richards

Thesis Proposal

August 18, 2009

Abstract

In this proposed thesis, I will develop efficient non-parametric methods for parameter estimation in large databases of high-dimensional, noisy data. Specifically, I plan to continue exploring the efficacy of the diffusion map method of data transformation, used in conjunction with the Nyström extension, in uncovering underlying structure in complicated, high-dimensional data sets. I will explore models that effectively exploit this structure to accurately estimate scientific parameters of interest. Additionally, I will formulate multi-scale methods that generate data-driven bases within local partitions to allow for the modeling of more complicated data. The proposed methods will be tested on three research problems in astrophysics: estimation of star formation history (SFH) in galaxies using spectra from the Sloan Digital Sky Survey (SDSS), photometric redshift estimation using data from the SDSS Photometric Survey and the Canada-France-Hawaii Telescope Legacy Survey, and detection of classes of quasar and outliers using SDSS spectra. Preliminary work has already shown promise in the proposed methods for both SFH estimation (Richards et al. 2009) and photometric redshift estimation (Freeman et al. 2009).

1 Astronomical survey data

Technological advancements in observational astronomy have caused a recent flood of astronomical surveys that collect data from billions of objects.[1] Modern surveys amass data for different types of astronomical object (e.g. stars, galaxies, quasars) across the entire electromagnetic spectrum, including the radio (FIRST, Becker et al., 1995; NVSS, Condon et al., 1998), microwave (WMAP, Wright et al., 2009), infrared (2MASS, Skrutskie et al., 2006; AKARI, Murakami et al., 2007), optical (SDSS, York et al., 2000; LSST, Becla et al., 2006; Pan-STARRS, Kaiser & Pan-STARRS Team, 2002), ultraviolet (GALEX, Bianchi & The GALEX Team, 1999), X-ray (XMM-Newton, Georgakakis et al., 2003), and gamma-ray (GLAST, Gehrels & Michelson, 1999).

Several surveys are dedicated to estimating specific sets of parameters for astronomical objects. For example, the 2dF surveys (Boyle et al., 2001) are dedicated to the accurate estimation of the redshifts of 250,000 galaxies and 25,000 quasars. The DEEP2 survey (Davis et al., 2003) is designed to study the evolution of the properties of galaxies (e.g. mass, metallicity, spatial clustering) as a function of redshift. Larger surveys are often designed to estimate parameters for a broader range of astronomical objects across many astronomical disciplines. For instance, the SDSS survey has

[1] see `http://www.cv.nrao.edu/fits/www/yp_survey.html` for a listing of more than 70 current astronomical surveys with links to information and data products from each.

collected data to estimate the physical properties of galaxies, the evolution of quasars, and the structure and mixture of stellar populations of the Milky Way.

Data collected by astronomical surveys are generally high-dimensional, complicated, and noisy. Typical data from these surveys can come in several forms, including photometric fluxes with measurements in 5-10 spectral bands, high-resolution broad-band spectra with data in 1000s of spectral bins, and astronomical images with flux measurements in millions of pixels. Using survey data to draw statistical inferences about the measured astronomical objects can be a difficult task. The mechanisms that drive the underlying physics are complicated and often poorly understood, signal-to-noise ratios in the observed data are generally small, and the databases which are produced are usually immense.

My goal in this proposed thesis is to develop methods for parameter estimation in large databases of high-dimensional, noisy data such as those collected by astronomical surveys. I plan to apply these methods to three research problems in astrophysics:

1. Estimation of star formation history in galaxies using spectra from the Sloan Digital Sky Survey (SDSS).

2. Photometric redshift estimation using data from the SDSS Photometric Survey and the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS, Cuillandre & Bertin, 2006).

3. Detection of sub-populations of quasars and outliers using SDSS spectra.

All three proposed applications have an extensive amount of recent literature and great importance in the development of physical models that describe the Universe. My work on the first application is already well-developed whereas advancements on the second and third applications are in their early stages. The databases I will use for these three problems are all large: there are over 700,000 SDSS galaxy spectra, millions of objects in the SDSS photometric catalog and CFHTLS deep survey, and more than 70,000 SDSS quasar spectra.

The outline of this thesis proposal is as follows. In §2, I review some of the existing statistical methods that are utilized in the astrophysical community to analyze large databases of objects, outline the techniques I plan to develop and implement, and state the aims of the proposed thesis. In §3, I detail the proposed methodology, highlighting some recent advances. I dedicate §4 to describing the three astrophysical applications which I propose to pursue in the thesis. For each problem, I summarize some preliminary results and sketch the proposed work. I conclude with an approximate timeline in §5.

2 Existing methods and proposed aims

Many of the current problems of parameter estimation in astronomical databases are extremely complicated. Because the data-generating processes tend to be complex and non-linear and the data are high-dimensional and noisy, much of the research in this field relies on comparing observed data to well-understood systems and/or using simpler, lower-dimensional representations of the observed data to draw inferences.

In the astrophysical literature, a common technique of parameter estimation is template fitting, where observed data are compared to sets of (model or empirical) observable data, generated from each combination of a grid of parameters. See, for example, Benitez (2000); Bolzonella et al. (2000); Budavári et al. (2001); Tremonti et al. (2004); Feldmann et al. (2006) for recent applications of template fitting in astronomy. There are obvious drawbacks to this approach. Most notably, estimates are highly reliant on the choice of parameter grid as well as the quality of the templates. These techniques also tend to be computationally burdensome.

Another common strategy is to perform dimensionality reduction on the observed data and subsequently use this simpler representation of the data to draw inferences. Recently, principal components analysis (PCA) has enjoyed wide popularity in astrophysics as a method for linear dimensionality reduction (e.g., see Boroson & Green, 1992; Connolly et al., 1995; Madgwick et al., 2003; Vanden Berk et al., 2006; Rogers et al., 2007). In Richards et al. (2009a), we presented a formal method of using, e.g., the representation of a data set in a PCA basis to estimate physical parameters of interest. However, PCA is a linear method, and is not appropriate when there are non-linear relationships between data or when there exist outliers, as is usually the case in astronomical databases.

In the following, I summarize the techniques I plan to develop in the proposed thesis. The three main goals are:

Aim 1. *Estimate physical parameters for high-dimensional databases of astronomical objects using the diffusion map transformation.*

Diffusion map (Coifman & Lafon, 2006; Lafon & Lee, 2006; Lafon et al., 2006), an approach to spectral connectivity analysis (SCA, Lee & Wasserman, 2008), is a method of non-linear dimensionality reduction built on principles of weighted undirected graphs and Markov random walks. In Richards et al. (2009a), we showed that regression on the diffusion map basis representation of a set of SDSS spectra achieved better redshift predictions than those estimated by principal component regression. Subsequently, we have shown that using the diffusion map transformation for the estimation of star formation history in galaxies (Richards et al., 2009b) and photometric redshift prediction (Freeman et al., 2009) produces excellent results compared to other methods employed in the literature, such as template matching, PCA, and artificial neural networks. I propose to extend the use of diffusion map to other parameter estimation problems in astronomical databases and to more rigorously compare its performance to other techniques for high-dimensional inference in large databases.

Aim 2. *Develop multi-scale methods that generate data-driven bases within local partitions.*

In many estimation problems, richer sets of bases may be needed to achieve small prediction error. The idea of building PCA bases within local partitions has been explored in the literature by, e.g., Kambhatla & Leen (1997) and Einbeck et al. (2007), the latter of which applied the techniques to an astronomical data set. These authors show an increase of performance in estimation problems over using global PCA bases. I propose to extend this work and to devise and implement a statistically rigorous method of partitioning the data (including choosing the number of partitions to use), especially when the data are of very high dimension and there are large amounts of data. I plan to implement both local PCA and local diffusion map and to compare the performance of these bases in regression problems.

Aim 3. *Study the viability of these methods to produce fast and accurate parameter estimates for immense data sets.*

The Nyström extension (Baker, 1977) is a standard method of using the estimated eigenfunctions of a data-dependent kernel to estimate the value of the eigenfunctions for new data points (Bengio et al., 2004). When we are faced with massive data sets, estimating the diffusion map eigenvectors is computationally infeasible because it is necessary to invert an n-by-n matrix, where n is the number of data points. We have found that when our data sets exceed

$\sim 10^4$ data points, diffusion map fails due to memory and computational time limitations. In practice, we can estimate the diffusion map eigenvectors for a subset of our data and extend these estimates to the rest of our data set via the Nyström extension. In Freeman et al. (2009) we employed this technique to extend estimates to some 300,000 objects in the SDSS photometric catalog with an estimated 2.4% loss in accuracy of prediction. Another option to accelerate these computations is to first use *kd*-trees (Friedman et al., 1977; Gray et al., 2004) to organize our high-dimensional data and then to exploit this structure for fast and accurate distance computations and matrix inversions. I propose to perform further research into the performance of the Nyström extension and *kd*-trees under different situations and for a variety of applications.

3 Methodology

3.1 Diffusion map

In this section, I describe diffusion map (Coifman & Lafon, 2006; Lafon & Lee, 2006), a non-linear technique based on spectral connectivity analysis (SCA, Lee & Wasserman, 2008), that transforms data into a natural coordinate system. The diffusion map framework attempts to retain the cumulative local interactions between data points, i.e. their connectivity in the context of a fictive diffusion process over the data.

The starting point of diffusion map is a weighted graph $G = (\Omega, W)$, where the nodes in the graph are the observed data points. The weight given to the edge connecting $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$w(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{s(\mathbf{x},\mathbf{y})^2}{\epsilon}\right), \tag{1}$$

where $s(\mathbf{x},\mathbf{y})$ is a locally relevant similarity measure. For instance, $s(\mathbf{x},\mathbf{y})$ could be chosen as the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. The flexibility in the choice of $s(\mathbf{x},\mathbf{y})$ is much of the appeal of this approach: it is often simple to determine whether or not two data points are similar, and the choice of similarity measure can be selected depending on the problem and data set on hand. The tuning parameter ϵ in (1) is usually chosen small enough that $w(\mathbf{x},\mathbf{y}) \approx 0$ unless $\mathbf{x}$ and $\mathbf{y}$ are similar, but large enough such that the constructed graph is fully connected. In practice, we generally choose the ϵ that minimizes the estimated risk for the task at hand.

The next step is to use these weights to build a Markov random walk on the graph. From node (data point) $\mathbf{x}$, the probability of stepping directly to $\mathbf{y}$ is defined naturally as

$$p_1(\mathbf{x},\mathbf{y}) = \frac{w(\mathbf{x},\mathbf{y})}{\sum_{\mathbf{z}} w(\mathbf{x},\mathbf{z})}. \tag{2}$$

This probability is close to zero unless $\mathbf{x}$ and $\mathbf{y}$ are similar. Hence, in one step the random walk will move only to very similar nodes (with high probability). These one-step transition probabilities are stored in the n by n matrix $\mathbf{P}$. It follows from standard theory of Markov chains that, for a positive integer t, the element $p_t(\mathbf{x},\mathbf{y})$ of the matrix power $\mathbf{P}^t$ gives the probability of moving from $\mathbf{x}$ to $\mathbf{y}$ in t steps. Increasing t moves the random walk forward in time, propagating the local influence of a data point (as defined by the kernel w) with its neighbors.

For a fixed time (or scale) t, $p_t(\mathbf{x},\cdot)$ is a vector representing the distribution after t steps of the random walk over the nodes of the graph, conditional on the walk starting at $\mathbf{x}$. In what follows, the points $\mathbf{x}$ and $\mathbf{y}$ are close if the conditional distributions $p_t(\mathbf{x},\cdot)$ and $p_t(\mathbf{y},\cdot)$, are similar. Formally, the diffusion distance at a scale t is defined as

$$D_t^2(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{z}} \frac{\left(p_t(\mathbf{x},\mathbf{z}) - p_t(\mathbf{y},\mathbf{z})\right)^2}{\phi_0(\mathbf{z})} \tag{3}$$

where $\phi_0(\cdot)$ is the stationary distribution of the random walk, i.e., the long-run proportion of the time the walk spends at node $\mathbf{z}$. Dividing by $\phi_0(\mathbf{z})$ serves to reduce the influence of nodes which are visited with high probability regardless of the starting point of the walk. The distance $D_t(\mathbf{x},\mathbf{y})$ will be small only if $\mathbf{x}$ and $\mathbf{y}$ are connected by many short paths with large weights. This construction of a distance measure is robust to noise and outliers because it simultaneously accounts for the cumulative effect of *all* paths between the data points. Note that the geodesic distance (the shortest path in a graph), on the other hand, often takes shortcuts due to noise.

The final step is to find a low-dimensional embedding of the data where Euclidean distances reflect diffusion distances. A bi-orthogonal spectral decomposition of the matrix $\mathbf{P}^t$ gives $p_t(\mathbf{x},\mathbf{y}) = \sum_{j\geq 0} \lambda_j^t \psi_j(\mathbf{x})\phi_j(\mathbf{y})$, where ϕ_j, ψ_j, and λ_j are the left eigenvectors, right eigenvectors and eigenvalues of $\mathbf{P}$, respectively. It follows that

$$D_t^2(\mathbf{x},\mathbf{y}) \;=\; \sum_{j=1}^{\infty} \lambda_j^{2t} (\psi_j(\mathbf{x}) - \psi_j(\mathbf{y}))^2. \tag{4}$$

The proof of (4) and the details of the computation and normalization of the eigenvectors ϕ_j and ψ_j are given in Coifman & Lafon (2006) and Lafon & Lee (2006). By retaining the m eigenmodes corresponding to the m largest nontrivial eigenvalues and by introducing the diffusion map

$$\mathbf{\Psi}_t : \mathbf{x} \mapsto [\lambda_1^t \psi_1(\mathbf{x}), \lambda_2^t \psi_2(\mathbf{x}), \cdots, \lambda_m^t \psi_m(\mathbf{x})] \tag{5}$$

from the data space to $\mathbb{R}^m$, we have that

$$D_t^2(\mathbf{x},\mathbf{y}) \;\simeq\; \sum_{j=1}^{m} \lambda_j^{2t} (\psi_j(\mathbf{x}) - \psi_j(\mathbf{y}))^2 \;=\; ||\Psi_t(\mathbf{x}) - \Psi_t(\mathbf{y})||^2 \,, \tag{6}$$

i.e., Euclidean distance in the m-dimensional embedding defined by equation (5) approximates diffusion distance. See `http://www.stat.cmu.edu/~jwrichar/software.html` for an R package and sample Matlab code for diffusion maps.

The choice of the parameters m and t can either be determined by the fall-off of the eigenvalue spectrum or by minimizing an appropriate risk function for the problem at hand (e.g., clustering, classification, regression, or data visualization). An objective measure of performance should be defined and utilized to find data-driven best choices for these tuning parameters. In Richards et al. (2009b), we introduced a multi-scale diffusion map,

$$\mathbf{\Psi}_{\rm ms} : \mathbf{x} \mapsto \left[\frac{\lambda_1}{1-\lambda_1}\psi_1(\mathbf{x}), \frac{\lambda_2}{1-\lambda_2}\psi_2(\mathbf{x}), \cdots, \frac{\lambda_m}{1-\lambda_m}\psi_m(\mathbf{x})\right]. \tag{7}$$

This representation has the advantages of eliminating the tuning parameter, t, and preserving the multi-scale diffusion distance,

$$D_{\rm ms}^2(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{z}} \frac{\sum_{t=1}^{\infty}(p_t(\mathbf{x},\mathbf{z}) - p_t(\mathbf{y},\mathbf{z}))^2}{\phi_0(\mathbf{z})} \tag{8}$$

which simultaneously considers all time scales.

3.1.1 Diffusion K-means

In this section, we describe our method of clustering a high-dimensional data set and of finding an appropriate set of K prototypes from a large set of (high-dimensional) data, using the diffusion map representation. This method is fully described in Richards et al. (2009b).

We cluster data and determine a set of K prototypes by performing K-means clustering in the m-dimensional diffusion map representation (5) of a data set. The K-means algorithm is a standard machine learning method used to cluster data points into K groups. It works by minimizing the total intra-cluster variance,

$$V = \sum_{j=1}^{K} \sum_{\mathbf{x} \in S_j} ||\mathbf{\Psi}(\mathbf{x}) - \mathbf{c}(S_j)||_2^2 \tag{9}$$

where S_j is the set of points in the jth cluster and $\mathbf{c}(S_j)$ is the m-dimensional geometric centroid of the set S_j. Here (and in the remainder of the proposal) we use $\mathbf{\Psi}(\mathbf{x})$ to denote either the m-dimensional t-step diffusion map $\mathbf{\Psi}_t(\mathbf{x})$ in (5) or the m-dimensional multi-scale diffusion map $\mathbf{\Psi}_{\mathrm{ms}}(\mathbf{x})$ in (7). We choose to use K-means for its reliance on Euclidean distance, which, in diffusion space, approximates diffusion distance, as shown in (6). As described in §3.2 of Lafon & Lee (2006), for diffusion maps

$$\mathbf{c}(S_j) = \sum_{\mathbf{x} \in S_j} \frac{\phi_0(\mathbf{x})\mathbf{\Psi}(\mathbf{x})}{\sum_{\mathbf{x} \in S_j} \phi_0(\mathbf{x})}, \tag{10}$$

where ϕ_0 is the trivial left eigenvector of $\mathbf{P}$. The K-means algorithm begins with an initial partition of the data and alternately computes cluster centroids and reallocates points to the cluster with the nearest centroid. The algorithm stops when no points are allocated differently in two consecutive iterations. The final allocation of points define the clusters and the final centroids define the K prototypes.

3.2 Nyström extension

Computation of diffusion coordinates (5) requires eigen-decomposition of the n by n matrix $\mathbf{P}$. This operation is computationally intractable for data sets of $n \gtrsim 10^4$. The astronomical data sets which I will analyze in this thesis are much larger, and hence we require a more computationally efficient method for computing diffusion coordinates.

A standard method for extending eigenvalue estimates from a small set of data to new data points is the Nyström extension (Baker, 1977). The implementation is as follows: split the original data into a training set of $n(< 10^4)$ objects and a validation set of n' objects. We estimate the $n \times m$ diffusion map, $\mathbf{\Psi}$, for the training set as in §3.1. Then, we compute the $n' \times n$ weight matrix $\mathbf{W}$, where each row is a vector of weights between one object in the validation set and every object in the training set, computed via (1). We use the same ϵ in the computation of $\mathbf{W}$ as was used in the original computation of $\mathbf{\Psi}$, though other options have been proposed (e.g. Lafon et al., 2006). We row-normalize $\mathbf{W}$ by dividing each row by its row sum.

Call $\mathbf{\Lambda}$ the $m \times m$ diagonal matrix with entries $1/\lambda_i$, where λ_i is the eigenvalue corresponding to ψ_i in the decomposition of $\mathbf{P}$. The diffusion map coordinates for the n' validation objects are approximated by

$$\mathbf{\Psi}' = \mathbf{W}\mathbf{\Psi}\mathbf{\Lambda}, \tag{11}$$

which is the Nyström formula. The Nyström extension, is of order $\mathcal{O}((n' + n)n^2)$, compared to a complexity of $\mathcal{O}((n' + n)^3)$ for exact eigen-decomposition. Hence, there are clear computational advantages to using the Nyström approximation when $n' \gg n$. In Belabbas & Wolfe (2009), the authors present bounds on the Nyström approximation error, as well as methods to choose a training set to decrease the approximation error.

3.3 Adaptive regression

Our next problem is how to, in a statistically rigorous way, predict a function $y = r(\mathbf{x})$ of high-dimensional data, $\mathbf{x}$, using a sample of known pairs $(\mathbf{x}, y)$. In Richards et al. (2009a) we utilized the idea that one may use the eigenfunctions from diffusion maps as

- coordinates of the data points, or
- forming a Hilbert orthonormal basis for any function supported on the diffusion map space (including the regression function $r(\mathbf{x})$).

This latter insight allows us to formulate a general regression framework, as I describe below.

Any function r satisfying $r(\mathbf{x})^2 dx < \infty$, can be written as

$$r(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{x}) \tag{12}$$

where the sequence of functions $\{\psi_1, \psi_2, ...\}$ forms an orthonormal basis. The choice of basis functions is traditionally not adapted to the geometry of the data. Standard choices are, for example, Fourier or wavelet bases for $\mathbf{L}^2$, which are constructed as tensor products of one-dimensional bases. The latter approach makes sense for low dimensions, for example for $p = 2$, but quickly becomes intractable as p increases. In particular, note that if a wavelet basis in one dimension consists of q basis functions, and hence requires the estimation of q parameters, the naive tensor basis in p dimensions will have qp basis functions/parameters, creating an impossible inference problem even for moderate p. Because this basis is not adapted to the geometry of the data, there is little hope of finding a subset of these basis functions which will do an adequate job of modeling the response.

Our proposal is an adaptive framework where the basis functions reflect the intrinsic geometry of the data. First, we construct a data-driven basis for the lower-dimensional, possibly non-linear subset of the data space where the data lie. Let $\psi_1, \psi_2, ..., \psi_m$ be the eigenfunctions computed by spectral decomposition of the matrix $\mathbf{P}$ defined in (2). These eigenfunctions are orthogonal with respect to the stationary distribution of the Markov random walk on our data described in §3.1. Our regression function estimate $\widehat{r}(\mathbf{x})$ is given by

$$\widehat{r}(\mathbf{x}) = \sum_{j=1}^{m} \widehat{\beta}_j \psi_j(\mathbf{x}) \tag{13}$$

where the different terms in the series expansion represent the fundamental eigenmodes of the data. Note that neither the choice of t in the diffusion map (5) nor the use of the multi-scale diffusion map (7) changes the analysis, as both produce a simple rescaling in $\widehat{\beta}_j$. In our analyses, both m and the diffusion map ϵ are chosen to minimize the prediction risk, which we generally estimate using cross validation.

3.4 Multi-scale bases

Often times, the complicated nature of the application will necessitate the use of a richer set of basis functions. For instance, in estimation of the redshift of galaxies using photometric data, galaxies having different photometric colors may not only have different redshifts, but also different intrinsic properties such as age, galaxy type, star formation rate, metallicity, etc. An appropriate basis to estimate properties for one subset of objects will not generally be appropriate for parameter estimation for a different type of object. My proposal is to develop methods that quickly partition data sets based on their observable properties and then to build data-driven bases in each partition. Heuristically, the goal is to find a set of partitions where within each partition, the data relate in a simple way to the parameter(s) of interest. Then, data-driven bases within each partition should be able to more accurately predict these parameter(s).

Methods for data partitioning must be able to handle high-dimensional data and be scalable for massive data sets, since this is often the regime we operate in, especially for analysis of astronomical databases. The methods which I propose are based on the eigen decomposition of the local empirical covariance matrices of a data set, and are similar to those of Einbeck et al. (2007). As this project has begun recently, the method proposed below should be viewed as a starting point for the further development of more refined multi-scale methods. In this thesis, I plan to produce better multi-scale methods and study their performance on real data. The proposed starting method is the following.

Starting with data $\mathbf{X}$ with n observations and d variables, we compute the first p principal components (PCs) $\mathbf{v}_1, ..., \mathbf{v}_p$ from the eigen decomposition of the d-by-d empirical covariance matrix,

$$\widehat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{14}$$

where $\mathbf{V}$ is a matrix whose columns are the PCs and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues. From (14), we split the data at its empirical mean, orthogonal to the first PC, $\mathbf{v}_1$. For each resultant partition, $R^{(i)}$, we estimate the local covariance matrix,

$$\widehat{\mathbf{\Sigma}}^{(i)} = \frac{1}{n_i} \sum_{\mathbf{x} \in R^{(i)}} (\mathbf{x} - \bar{\mathbf{x}}^{(i)})(\mathbf{x} - \bar{\mathbf{x}}^{(i)})^T \tag{15}$$

and local mean $\bar{\mathbf{x}}^{(i)}$, where $n_i = \sum_{i=1}^{n} \mathrm{I}(\mathbf{x} \in R^{(i)})$. In each subsequent step, we split the partition with the maximum $|\widehat{\mathbf{\Sigma}}^{(i)}|$ at the mean of the partition, orthogonal to $\mathbf{v}_1^{(i)}$. This procedure creates a tree structure, where the optimal depth can be chosen either by splitting down to a pre-specified level $G_{\max}$, or by fitting a model in PC space and using a penalized likelihood such as Bayesian Information Criterion (BIC). We may also allow for neighboring partitions to be recombined, but preliminary work has shown that this process can be too computationally expensive for the method to be scalable to large data sets.

Once we have estimated an appropriate partitioning of our data set, $\mathbf{X}$, our goal is to determine an appropriate data-driven basis in each partition, $R^{(i)}$, with which we will make statistical inferences. Again, using a different basis within each partition gives our statistical model much flexibility over using a global data-driven basis over the entire range of data. Within each partition, bases such as PCA or diffusion map can be computed. The former technique is similar to the local PCA method of Kambhatla & Leen (1997), while the latter may produce a much more flexible model because it can capture any nonlinear structure of data within each partition. However, local diffusion map bases induce additional computational complexity because an appropriate tuning parameter, ϵ, must be estimated in each partition. An additional challenge in the use of these multi-scale bases is to derive a method to join information in neighboring partitions to produce a set of basis functions that is smooth across the data set. In Einbeck et al. (2007), the authors use a kernel method to join the information in neighboring partitions. However, in very high dimensions, this technique will fail due to the sparseness of data in high dimensional spaces (curse of dimensionality).

In this proposed thesis, I will use the simple method described above as a starting point to develop better, more sophisticated multi-scale techniques that can accurately model complex, high-dimensional data sets. I will begin by studying the performance of the simple method on real data, comparing to other methods that use global bases. Then, I will explore, for example, other methods of partitioning data, ways to combine neighboring partitions and select an optimal number of partitions, and techniques to share information across boundaries. For a more concrete list of proposed work, see section §4.2.2.

4 Astrophysical applications

In this section, I describe the three astrophysical applications which I propose to pursue in the thesis. For each application, I review relevant literature, mention the work I have already performed, and outline the proposed work for the thesis.

4.1 Star formation history (SFH) estimation for galaxies

Determining the physical parameters of large samples of galaxies is crucial to constrain knowledge of the complicated physical processes that govern the formation and evolution of these systems. Astrophysicists seek methods that quickly and effectively map observed data from galaxies to sets of parameters that describe the star formation and chemical histories of these galaxies. Due to the complexity of the mechanisms of galaxy evolution, it is critical that we adopt a model that uses few assumptions about how galaxies evolve. Moreover, full utilization of the available theoretical stellar models, avoiding ad hoc simplifications, is critical to obtaining accurate parameter estimates.

There has been a large amount of work dedicated to estimating the physical parameters of databases of galaxies using different types of data. Recently, full high-resolution, broad-band spectra (Reichardt et al., 2001; Panter et al., 2003; Cid Fernandes et al., 2004, 2005; Mathis et al., 2006; Ocvirk et al., 2006; Asari et al., 2007) have been utilized to estimate the star formation histories (SFHs) and metallicities of galaxies. Modern databases of high-quality, homogeneous galaxy spectra, such as SDSS, have made possible the approach of full high-resolution spectral fitting to infer the properties of large populations of galaxies. Recent work has shown promise in the estimation of SFH parameters using SDSS galaxy spectra (see Panter et al., 2007 and Asari et al., 2007 for reviews of results achieved by two such fitting methods). However, large-scale analyses, while possible, are not necessarily accurate, and can be computationally challenging.

4.1.1 Preliminary work

A common technique in the literature, called empirical population synthesis, is to model galaxies as mixtures of simple stellar populations (SSPs) with known physical parameters. Recent studies that have used this method are, e.g. Bica (1988), Cid Fernandes et al. (2001), and Moultaka et al. (2004). Under this model, galaxy data are treated as linear combinations of the observable properties of SSPs. The STARLIGHT spectral fitting code, introduced by Cid Fernandes et al. (2004), fits observed spectra with linear combinations of SSP spectra from the evolutionary population synthesis models of Bruzual & Charlot (2003).

In this work, we adopt the STARLIGHT model of galaxy spectra introduced in Cid Fernandes et al. (2004):

$$M_\lambda = M_{\lambda_0} \left(\sum_{j=1}^{K} x_j \xi_{j,\lambda} r_\lambda \right) \otimes G(v_*, \sigma_*), \tag{16}$$

where M_λ is the model flux in wavelength bin λ. The component ξ_j is the jth basis spectrum normalized at wavelength λ_0. The basis of SSP spectra $\xi = \{\xi_1, ..., \xi_K\}$ is chosen before performing the analysis. The scalar $x_j \in [0, 1]$ is the proportion of flux contribution of the jth component at λ_0, where $\sum_{j=1}^{K} x_j = 1$; the vector $\mathbf{x}$ with components $x_j, (j = 1, ..., K)$ is the *population vector* of a galaxy. These flux fractions can be converted to mass fractions, μ_i, using the light-to-mass ratios of each basis spectrum ξ_j at λ_0.

Other components of the model in (16) are the scaling parameter M_{λ_0}, the reddening term, $r_\lambda = 10^{-0.4(A_\lambda - A_{\lambda_0})}$, and the Gaussian convolution, $G(v_*, \sigma_*)$. The reddening term describes distortion in the observed spectrum due to foreground dust, and is modeled by the extinction law of Cardelli et al. (1989) with $R_V = 3.1$. The Gaussian convolution, $G(v_*, \sigma_*)$, accounts for movement of stars within the observed galaxy with respect to our line-of-sight, and is parametrized by a central velocity v_* and dispersion σ_*.

To fit individual galaxies, we use the STARLIGHT fitting routine. The code uses a Metropolis algorithm plus simulated annealing to find the minimum of

$$\chi^2(\mathbf{x}, M_{\lambda_0}, A_V, v_*, \sigma_*) = \sum_{\lambda=1}^{N_\lambda} [(O_\lambda - M_\lambda) w_\lambda]^2, \tag{17}$$

where O_λ is the observed flux in the wavelength bin λ, M_λ is the model flux in (16), and w_λ is the inverse of the noise in the measurement O_λ. The summation is over the N_λ wavelength bins in the observed spectrum. The minimization of (17) is performed over $K + 4$ parameters: $x_1, ..., x_K, M_{\lambda_0}, A_V, v_*$, and σ_*, where K is the number of basis spectra in ξ. The speed of the algorithm scales as K^2. For the analysis of, e.g. $> 700,000$ galaxy spectra in the SDSS database, it is crucial to use a basis with a small number of spectra.

In, for example, Cid Fernandes et al. (2005), the STARLIGHT model (16) and fitting algorithm are tested using simulations. The simulated spectra are generated from the model in (16) using $K = 45$ SSPs from Bruzual & Charlot (2003). The authors fit the simulations using the same basis of SSPs that was used to generate the simulations. From their analyses, they conclude that in the absence of noise, the algorithm accurately recovers the input parameters. Although their use of the same basis for both generating and fitting simulated spectra is an appropriate test that the algorithm works, it is not a fair assessment of the expected performance of the methods for a database of real galaxy spectra.

In reality, the population of all observed galaxies is not constrained to be mixtures of a small number of SSPs on a discrete grid of age and metallicity. A more physically accurate representation of galaxies is as mixtures of an infinitely large basis of SSPs on a continuous grid of age and metallicity and, depending on the complexity of the underlying physics, a possibly infinite grid of prescriptions for initial mass function and evolutionary track. We simulate galaxies from a large database of $\sim 10^3$ SSPs on a fine grid of age and metallicity to determine whether we can accurately describe their SFHs, metallicities, and kinematic parameters using an appropriately chosen, computationally tractable basis prototype spectra.

The key to finding accurate SFH estimates for galaxies is to determine a basis, ξ, that captures most of the variability of a large set of SSP spectra, in a small number of prototype spectra. Our approach is to use diffusion K-means to estimate a small number of prototypes from a large SSP database containing 1278 spectra from the models of Bruzual & Charlot (2003). The database is meant to represent the large population of SSPs from which observed galaxies can be composed. In Fig. 1, we plot the mapping of 1278 SSP spectra into the $m = 3$ dimensional diffusion space. The large black dots denote the K-means centroids for $K = 45$. The K prototypes capture the

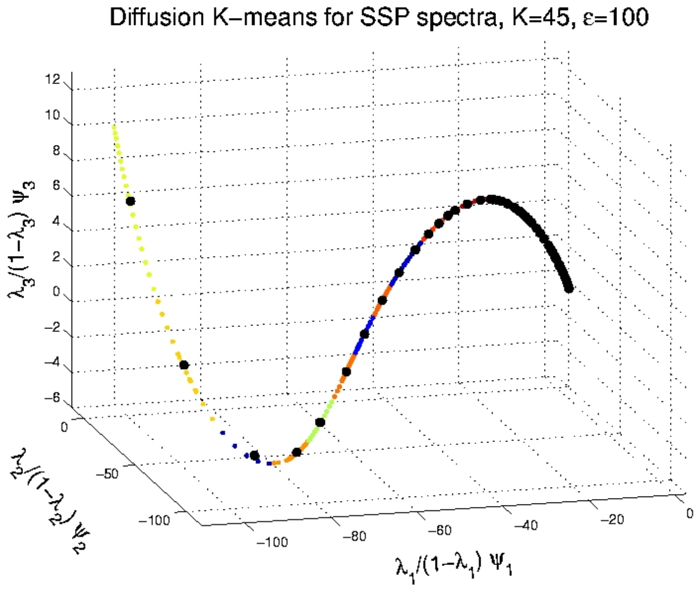


Figure 1: *Diffusion K-means for SSP spectra.* Representation of 1278 SSP spectra in 3-dimensional diffusion space. Large black dots denote the $K = 45$ centroids. Individual SSPs are colored by cluster membership. The SSPs reside on a simple, low-dimensional manifold which is captured by the K prototypes.

variability of the SSP spectra along a low-dimensional manifold in diffusion space. In Fig. 2 we plot the 45 SSP spectra in CF05 and the 45 prototype spectra found by diffusion K-means. All spectra are normalized at λ_0=4020Å and are colored by log age. The diffusion K-means prototypes spread themselves evenly over the range of spectral profiles, capturing a gradual trend from young to old spectra. On the other hand, the CF05 basis includes many similar spectra from younger populations and sparsely covers the range of spectral profiles of older populations.

In Richards et al. (2009b) we showed that our approach achieves better SFH parameter estimates than the bases used in the literature (Cid Fernandes et al., 2005, CF05, of size 45 and Asari et al., 2007, Asa07, of size 150) and those derived by other K-means methods. Table 1 shows results for one set of simulated spectra. In this paper, we also showed that diffusion K-means bases achieved a significant decrease in the age-metallicity degeneracy that is common in galaxy SFH estimation. For a sample of 3046 SDSS galaxy spectra, we found that the diffusion K-means bases obtain estimates that are consistent with independent measurements from the MPA/JHU database[2] (Tremonti et al., 2004) and show a smaller age-metallicity degeneracy than the basis of Asa07.

4.1.2 Proposed work

I am currently in the process of constructing a catalog of estimated SFH parameters for all 781,692 galaxies with SDSS spectra. The estimates are obtained using the STARLIGHT code with diffusion K-means derived basis of 150 prototype SSP spectra. To complete the analysis, the following work must be done:

- *Develop and analyze different methods of prototype selection.* Thus far, I have compared the performance of diffusion K-means to that of standard and PC K-means, using simulated

[2] http://www.mpa-garching.mpg.de/SDSS/

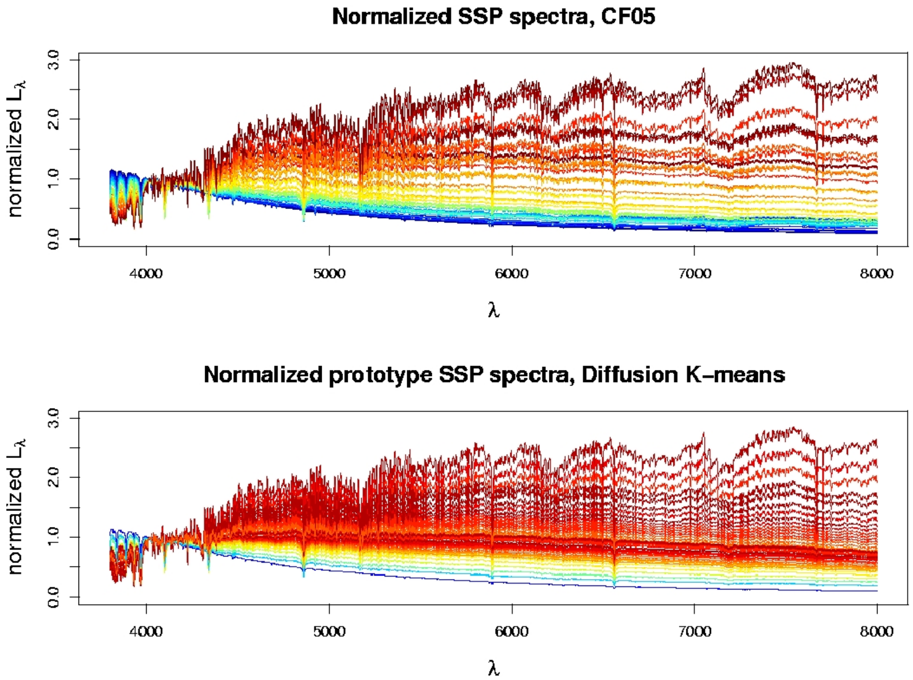


Figure 2: *Basis spectra for CF05 and Diffusion K-means, colored by* $\log t$*.* All spectra are normalized to 1 at $\lambda_0 = 4020$ Å. The diffusion K-means basis covers the spectral range in gradual increments while the CF05 basis includes many young spectra with similar spectral properties and sparsely covers the spectral range of older populations.

galaxy data. In the statistics literature, there are few studies on selecting an appropriate set of prototypes given a large set of examples. This presents an opportunity to develop a field that is both important in many scientific applications and lacking in methodology. I plan on developing other methods of prototype selection based on sparse methods that are traditionally used in a regression framework (e.g. basis pursuit, matching pursuit). I also plan to compare their performance with, for example, the performance of diffusion K-means. The former methods are linear and select a subset of the original examples, while diffusion K-means forms new combinations of examples using the data connectivity structure and clustering in eigenspaces.

- *Error estimation for STARLIGHT SFH parameters.* The STARLIGHT algorithm takes 5-15 minutes to find the maximum likelihood estimate for each galaxy. Mapping out the likelihood surface for each galaxy would be prohibitively slow. The challenge is to accurately approximate the errors on the SFH parameters in a way that is fast. If we look at this as an application of maximum likelihood, and then find the Fisher information, then we are discarding our most valuable information, namely that we have many spectra and the parameter estimates for each. Our approach should share this information across the different galaxies. For example, if two similar spectra give very different estimates of a parameter, then that parameter estimate should be given a high standard error. We can assess similarity by exploiting the natural geometry of the spectra via diffusion map. For example, in Figure 3, we find a clear trend in average stellar age (and metallicity) of each galaxy in diffusion space. However, there are some galaxies that have similar spectral characteristics to other galax-

Table 1: Summary of parameter mean-square errors for simulated spectra with exponentially decreasing SFH and random starbursts. The bases compared are: CF05–Cid Fernandes et al. (2005), Asa07–Asari et al. (2007), dM–diffusion K-means, sK–standard K-means, PC–principal components K-means. Bold represents the lowest error for each experiment.

K	Basis	$\langle \log t_* \rangle_L$	$\log\langle Z_* \rangle_L$	A_V	σ_*	$\mathbf{x}_c$	μ_c
45	CF05	0.368	0.089	0.025	369.0	0.664	0.349
	dM	**0.074**	0.024	**0.007**	**156.4**	**0.195**	0.059
	sK	0.194	0.021	0.010	171.7	0.219	**0.046**
	PC	0.222	**0.020**	0.012	174.6	0.212	**0.046**
150	Asa07	0.167	0.059	0.013	169.6	0.421	0.175
	dM	**0.093**	**0.018**	**0.007**	**153.1**	**0.213**	**0.067**
	sK	0.150	0.021	0.008	153.4	0.303	0.118
	PC	0.163	0.029	0.009	160.9	0.317	0.117

ies with inconsistent age (and metallicity) estimates. These objects should be given larger standard errors by our technique. This approach also has obvious computational advantages, as the full information matrix approach would not be feasible with the STARLIGHT model (16), and the large number of galaxies we want to handle.

- *Aperture bias correction.* Due to the small size of the spectroscopic fiber used to collect light from each galaxy in the SDSS survey, light will be sub-sampled from galaxies whose angular extent on the sky is large. At the median redshift in the SDSS survey, it is estimated that only about 1/3 of the total light from each galaxy is sampled by the spectrum. It is not appropriate to assume that the galaxy light that is missed by each spectrum has the same properties as the light that was sampled by the spectrum because in distant parts of each galaxy, different physics govern star formation processes. Each SDSS galaxy spectrum is accompanied by a set of 5 photometric fluxes, one for each of 5 filters. These fluxes are obtained by integrating the light over the entire extent of the galaxy in each filter. By comparing these fluxes (SDSS model magnitudes) with the fluxes obtained by the spectra (SDSS fiber magnitudes), we can estimate the amount of light that has been missed by the spectral fiber. Moreover, we can estimate the color of this light, which tells us about the properties (e.g. ages and metallicities) of the stars that have not been sampled by the spectral fiber.

  An empirical technique for aperture bias correction was introduced by Brinchmann et al. (2004) to estimate the current star formation rate (SFR) of each galaxy. In their correction method, they binned observed galaxies into 1845 color bins and used the distribution of SFR within the appropriate color bin to estimate the SFR of the portion of each galaxy falling outside of the spectral fiber. Their choice of dividing the color space into bins leads to several problems. For instance, they are forced to choose an ad hoc grid of colors, are not able to use all of the color information due to low galaxy counts, and often find that the color of the light missed by the fiber does not fall in any color bin. We can easily do better by treating this as a multiple regression problem, where the predictors are the 4 colors and the responses are the properties of the stars: $\langle \log t \rangle_L, \log\langle Z \rangle_L$, and L/M. We can estimate an appropriate model using the observed light from each galaxy and use this model to predict the properties of the unobserved light. These estimates (and their uncertainties) can be used in a straightforward manner to correct the aperture bias in the STARLIGHT SFH estimates.

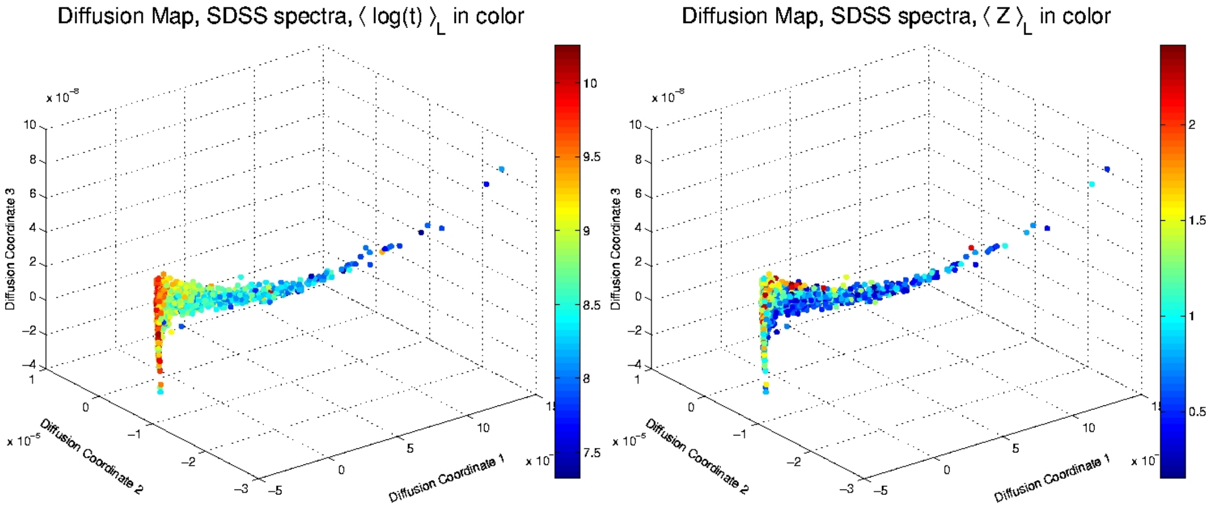


Figure 3: *Three-dimensional diffusion map representation of 3223 SDSS spectra.* Plotted in color are the average age and metallicity of the stars in each galaxy, respectively, as estimated by Cid Fernandes et al. (2005). Both parameters show strong trends across the manifold, with obvious anomalies. The anomalous galaxies should be given larger standard error in their SFH estimates.

4.2 Photometric redshift estimation

Accurate estimation of the redshifts of galaxies from photometric data (so-called photometric redshifts) is a key component to fulfilling the promise of next-generation cosmological surveys. For instance, the Large Synoptic Survey Telescope (LSST; Becla et al., 2006) is expected to collect photometry data for billions of galaxies. Compare this to, e.g., the $\sim 10^5$ galaxy spectra collected by the modern-day spectroscopic DEEP2 Galaxy Redshift Survey (Davis et al., 2003). However, any cosmological inferences made with photometric catalogs of galaxies will rely heavily on the accuracy of their redshift estimates. Prediction of accurate redshifts from the 5 measurements per galaxy that are generally collected by photometric surveys presents an immense challenge over, e.g., using full broad-band spectra, which usually contain measurements in 1000s of wavelength bins. Accordingly, there has been much recent work dedicated to photometric redshift estimation (for example, see Freeman et al., 2009 and references there-in).

In Freeman et al. (2009), we used a global diffusion map adaptive regression framework to estimate redshifts for galaxies in the SDSS and DEEP2 photometric catalogs. In that work, we showed prediction accuracies that were on par with previous analyses of the same data sets. We also demonstrated that some of the observed biases in estimation were due to attenuation bias (measurement error). In this thesis, I propose to extend this analysis using multi-scale data-driven bases within local partitions.

4.2.1 Preliminary work

I have begun to apply our proposed multi-scale techniques to the set of photometric CFHTLS data from the sample of 5,223 galaxies described in Freeman et al. (2009) to test these methods. All of these galaxies have excellent DEEP2 redshift determinations (> 99.5% confidence that the predicted DEEP2 redshift is correct). In the following, we take the DEEP2 spectroscopic redshift determinations as the truth and attempt to predict these values using the 5-band CFHTLS pho-

Method	$\widehat{R}_{\text{CV}}$	η (%)
global diffusion map	0.0551	2.29
local PCA (non-weighted)	0.0553	2.50
local PCA (weighted)	0.0547	2.24
local diffusion map (non-weighted)	0.0559	2.31
local diffusion map (weighted)	0.0546	2.22

Table 2: Preliminary results of redshift estimation for DEEP2 galaxies.

tometry data (accepting the spectroscopic redshift values as the truth is reasonable because these estimates generally have very little error compared to the magnitude of the errors on spectroscopic redshifts).

We use our partitioning algorithm to divide the data set into 24 disjoint sets. Partition size ranges from 14 to 659. In each partition, we compute both the local PCA and local diffusion map bases. Using adaptive regression of redshift on the local basis representation, we use 5-fold cross validation to optimize the R_{CV} (as defined in Freeman et al., 2009) of the fits over m for local PCA and over (m, ϵ) for local diffusion map. Due to differing densities of data in each partition, we define $\epsilon = \epsilon(\pi)$ as the median distance to the $n_i\pi$th nearest neighbor. We consider the same grid of π for each partition. For this data set, complete analysis using local diffusion map takes 30 CPU minutes, compared to 10 CPU seconds for local PCA.

Results, compared to those in Freeman et al. (2009) (global diffusion map), are in Table 2. For all local methods, we consider both non-weighted and weighted linear regression models, where weights are defined as the inverse of the redshift measurement errors. The preliminary results show that both cross-validation risk and proportion of catastrophic failures are on par with the global diffusion map results.

4.2.2 Proposed work

In this thesis, I will explore the viability of the proposed multi-scale methods to accurately estimate parameters for large databases. Here, I list some ways in which the multi-scale methods need to be further studied and improved. For each, I give concrete details about how I plan to approach the problem. I also relate this research to the estimation of photometric redshift for galaxy databases.

- *Explore other ways to partition data.* Our current method for partitioning the data is to build a tree structure by splitting perpendicularly to the first PC at the mean of each subset down to a pre-specified level. There are countless other ways in which we can partition data: we can use different prescriptions for the location and direction of each split, choose a stopping point by various different criteria, and elect to allow the algorithm to combine neighboring partitions. We want a partitioning algorithm that produces accurate results, under the constraints that it is fast and can handle high-dimensional data. A first step in the research into these partitioning methods is to investigate the behavior of different stopping criteria and assess the feasibility of allowing the method to join neighboring partitions.

- *Sharing information between neighboring partitions.* Although we build local bases within each partition, we still have a notion that neighboring partitions should have some common features. Further research into kernel methods, block bases, and other techniques to share information across boundaries needs to be done. I will begin by using a kernel method such as that in Einbeck et al. (2007) and comparing to models that do not share information between neighboring partitions.

- *Will multi-scale methods produce more accurate estimates for real problems?* Using real examples such as photometric redshift estimation, I will determine if the proposed methods have better performance over methods that use global bases. Also, I plan to compare the computational complexity of global and multi-scale methods. The primary test sets that I will use are the CFHTLS and SDSS photometric databases.

- *Produce accurate photometric redshift estimates.* Using Freeman et al. (2009) as a launching point, I will continue to pursue and develop methods that can accurately determine redshifts for galaxies using only their photometric colors. These methods will necessarily have some correction for attenuation bias, and may be driven by multi-scale bases.

4.3 Quasar outlier detection and estimation of sub-populations

Quasars (quasi-stellar objects, QSOs) are highly-energetic galaxies with active galactic nuclei powered by super-massive black holes that consume up to 1000 solar masses of material per year. Because these objects are among the most luminous in the Universe, they can be observed at very high redshift, and thus allow astronomers to probe the conditions of the early Universe. However, much is still unknown about the physics that govern QSOs, and a great deal of recent work has been dedicated to understanding the variability in databases of observed quasars and in relating observable QSO properties to their underlying physics.

The observed spectra of QSOs are characterized by large, broad emission lines. Much work has attempted to characterize the observed variability in both the spectral continua and different emission lines across samples of QSO spectra. One of the earliest efforts to characterize the relationships among observed quasar properties was performed by Boroson & Green (1992) on a set of 87 QSOs. In that work, the authors derived a set of 13 observed properties of quasar emission lines and continua from the spectra. They performed PCA on these derived properties to determine which observable information could explain most of the variability in the data. Later, Sulentic et al. (2000) performed a similar analysis using a slightly different set of observed properties, and found that three observed properties – width of a low-ionization line (FWHM Hβ), the ratio of the strengths of an iron and hydrogen line (R_{FeII}), and X-ray continuum strength (Γ_{soft}) – were the main contributors in the first PC (which they call eigenvector 1). Sulentic et al. (2002) use the main contributors to eigenvector 1 both to cluster the data into seven groups (using arbitrary bins, not by any statistical clustering method) and to identify outliers.

There are several drawbacks to the techniques used by Boroson & Green (1992), Sulentic et al. (2000), and Sulentic et al. (2002). First, they are dependent on derived properties (e.g. emission line widths and strengths, continua strengths) from the observed spectra. These properties must be carefully measured for each spectrum, a process that becomes computationally challenging for large databases of QSOs. Also, by focusing on a few derived properties, the majority of the bins in each observed spectrum is ignored. Second, the authors employ PCA, which is based on linear correlation structure. Any non-linear structure, which undoubtedly exists in databases of quasars, will be ignored. And third, all the inferences made in these papers are based only on the principal physical elements of eigenvector 1, which, in the case of Boroson & Green (1992), only explains 29% of the variability in their data. Conceivably, much useful and important information is being ignored in these studies.

At the heart of these studies are two interesting statistical challenges: how to find sub-populations in large sets of complicated, high-dimensional data and how to identify outliers in databases. Astronomers are interested in finding homogeneous sets of QSOs within large databases to learn about the variability in the physics of quasars. There is interest in creating composite spectra of each homogeneous subset of quasars to create high signal-to-noise templates which astronomers can study to learn more about the physics of each sub-population of objects. There is also much interest in QSO outlier detection within the astrophysical community in order to find objects whose physics are strange and undiscovered. Recently, Boroson & Lauer (2009) employed PCA on a 1700 Å wide region of a set of 17,500 SDSS quasar spectra and found the first known quasar with two broad-line systems, which they attributed to a binary super-massive black hole system, a previously undiscovered astrophysical phenomenon.

4.3.1 Preliminary work

In the thesis, I will study SDSS quasar spectra using diffusion map. The goals are to estimate QSO sub-populations and detect outliers. I will study the objects in the SDSS Quasar Catalog, fifth release (Schneider et al., 2007). There are a total of 77,429 objects in the catalog at redshifts ranging from 0.08 to 5.41. Thus far, the project is in the development stages. Here, I outline some of the work that I have begun.

The first challenge in this analysis is choosing an appropriate discrepancy measure, $s(\mathbf{x},\mathbf{y})$ between two QSO spectra, $\mathbf{x}$ and $\mathbf{y}$. We first shift all spectra to rest frame using their SDSS redshift estimates and re-bin to one flux measurement per Angstrom. Then, because the large redshift range in the catalog causes many rest-frame spectra to have small intervals of spectral overlap, we have divided our sample into 4 redshift bins to ensure a minimum overlap of 748 Å for each pair of quasars. Details of the 4 redshift bins are in Table 3. To eliminate absolute flux dependencies, we normalize each spectrum to the median flux within a normalization interval containing no emission features (see Table 3). We define our discrepancy measure as

$$s(\mathbf{x},\mathbf{y}) = \frac{1}{N_\mathrm{o}} \sum_{i=1}^{N_\mathrm{o}} \left(\frac{x(\lambda_i) - y(\lambda_i)}{\sqrt{\widehat{\sigma}(x(\lambda_i))^2 + \widehat{\sigma}(y(\lambda_i))^2}} \right)^2 \tag{18}$$

where N_o is the number of overlapping (unflagged) wavelength bins between the rest-frame spectra $\mathbf{x}$ and $\mathbf{y}$, $x(\lambda_i)$ is the normalized flux in the ith overlapping wavelength bin of rest-frame spectrum $\mathbf{x}$, and $\widehat{\sigma}(x(\lambda_i))$ is the estimated error in that normalized flux. The discrepancy measure (18) is the average weighted L2 distance between two spectra per overlapping wavelength bin, where the weights are the inverse of the estimated errors in the distance. It is natural to use the average distance per wavelength bin because each pair of spectra will have a different amount of spectral overlap, depending on their redshifts. In my preliminary work, I have considered three different discrepancy measures: the weighted L2 distance as defined in (18), a second distance measure that only considers continuum wavelengths where emission features occur[3], and a third measure that only considers wavelengths where no emission features occur. I use three different discrepancy measures (and thus construct three different diffusion maps) to build a more complete picture of the QSO database by, for example, finding outliers whose emission line (continua) features are anomalous and estimating sub-populations of QSO based only on emission lines (continua).

I have found the diffusion map embeddings for a sample of QSOs within each redshift bin for each of the three discrepancy measures. Starting with a training set of 2000 quasars within each redshift

[3] I follow the prescription in `http://www.sdss.org/dr2/algorithms/speclinefits.html` of having the emission line mask include all wavelengths within 30 pixels of any of the lines listed.

Table 3: Four redshift bins used in SDSS QSO analysis.

redshift	n_{qso}	min. λ overlap (Å)	normalization interval (Å)
0.078-1	20,710	1076	4155-4270
1-2	39,773	1168	2470-2700
2-3	11,572	1034	1690-1810
3-5.42	5374	748	1415-1515

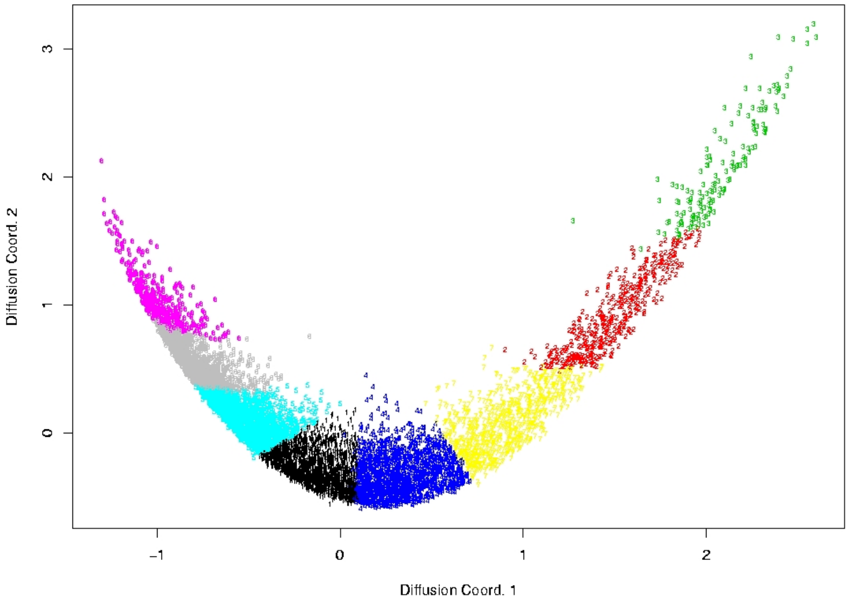


Figure 4: *Two-dimensional diffusion map embedding of 7639 quasars.* This sample of quasars in the 0-1 redshift bin lie on a relatively simple manifold in diffusion space. Here, the L2 distance measure in (18) was adopted.

bin, I extend to the remaining spectra in each redshift bin using the Nyström extension. In Figure 4, I plot the two-dimensional diffusion map representation of a sample of 7639 quasars in the 0-1 redshift bin (after outlier removal), using the L2 discrepancy measure in (18). The individual data points are colored by K-means cluster, where $K = 8$ was arbitrarily chosen. The average spectrum in each estimated sub-population, plus 90% point-wise confidence bands are plotted in Figure 5. The average spectra show significant differences between sub-populations, especially for smaller wavelengths. Also, the point-wise confidence bands tend to be small. This type of information is very useful to astronomers who attempt to understand the physics that control the observed variability in large samples of QSOs. Additionally, five outliers were found in this data set using this discrepancy measure, where I defined an outlier using diffusion coordinate-wise thresholding.

4.3.2 Proposed work

As this work is still in an early stage, here I propose a few general issues that have arisen and speculate on possible solutions to each. I also propose immediate steps to be taken.

- *Selection of the tuning parameter, ϵ.* In this application, the value of ϵ controls the number of outliers and the structure of the diffusion map. In practice, we want to keep the number of outliers small because these data will be analyzed manually to determine if the observed objects are interesting. I propose to begin by varying ϵ and observing how the results change. I will also experiment with the methods of ϵ-selection proposed in Lee & Wasserman (2008). I expect that we will see more outliers as we decrease ϵ, but also foresee that the structure of the diffusion map may change in unexpected ways. Eventually, I would like to formulate an explicit loss function to minimize with respect to ϵ.

- *Explore methods to estimate the sub-populations.* There are several different clustering techniques we could use to estimate the sub-populations. Thus far, I have used diffusion K-means, but I could also use a variety of other clustering methods in diffusion space such as mixture-model based or hierarchical linkage methods. In fact, because we have formulated our problem as the estimation of a mixture of sub-populations of objects, it seems that mixture-model based cluster methods, where the distribution in diffusion space is modelled as a mixture of, e.g., normal distributions, would be conducive to this task. Along with this is the related problem of estimating the appropriate number of sub-populations. I will explore both of these avenues.

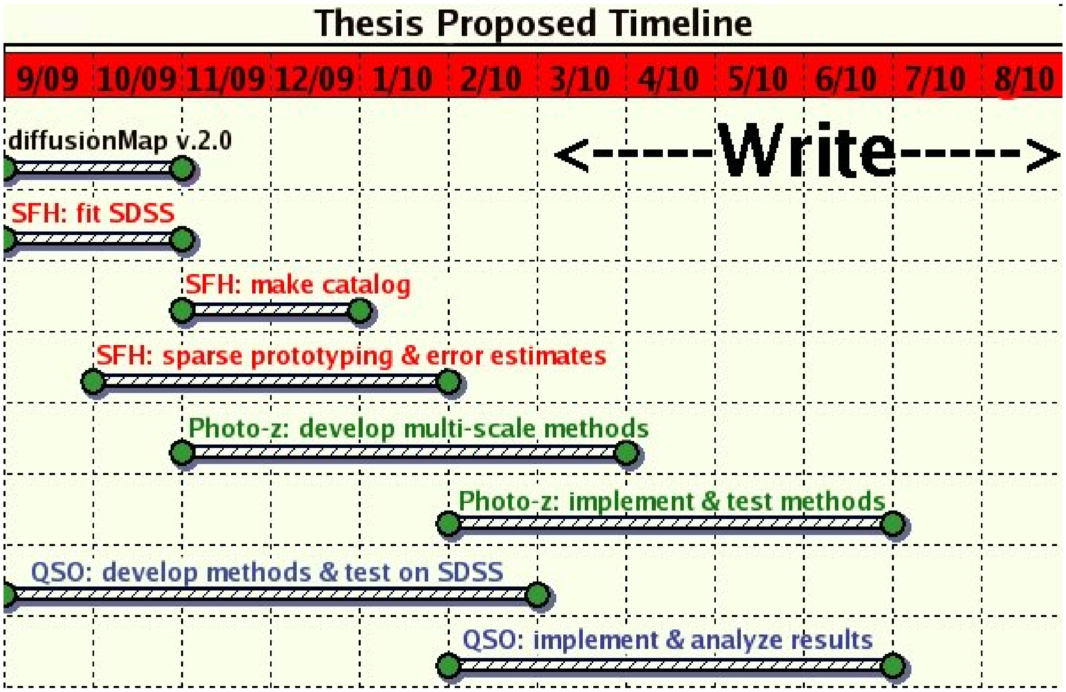
  However, this type of analysis raises another, more high-level issue. Thus far, we have operated under the assumption that the population of QSOs actually consists of some number of sub-populations whose observed spectra are all similar. However, in diffusion space, there do not appear to be any clear separations between sub-populations. Thus, would a better approach be to view the diffusion coordinates as varying due to some underlying, (physically-meaningful) continuous parameter? An obvious advantage to taking the former approach is that we can plot the spectra within each estimated sub-population and give the astronomer something that is easy to digest. However, if the source of variability can be traced to some physical parameter(s), then this sort of information will also be invaluable to the astronomer.

- *How to find outliers.* Up to now, the outlier detection routine that I have implemented only selects those objects for which at least one diffusion map coordinate is far from the median of that coordinate value. However, as exemplified by Figure 4, points often lie in center of the convex hull of the data, far from any other data. These data are most certainly outliers: from the Nyström extension, a datum far away from any of the data in the training set will have approximately equal weights to any of the points in the training set, and thus its diffusion map representation will lie in the center of the convex hull of the diffusion map. To detect these outliers, we need to employ a more sophisticated method than coordinate-wise thresholding. One option is to use the original distance matrix before applying the Nyström extension. I will begin exploring outlier detection methods that use both the distance matrix and the diffusion coordinates.

- *Interpretation of results.* After performing this analysis, do we find that the outliers are interesting objects? Are the clusters physically meaningful? Can we gain extra insight into the physics of quasars from this type of analysis? How can the results of such an analysis guide future methods for these types of data? This particularly pertains to the question of whether we want to model these QSOs either as coming from disjoint sub-populations or as varying with some set of continuous parameters.

5 Proposed timeline

Below, I give an approximate timeline to the completion of the proposed work.



References

Asari, N. V., Cid Fernandes, R., Stasińska, G., Torres-Papaqui, J. P., Mateus, A., Sodré, L., Schoenell, W., & Gomes, J. M. 2007, MNRAS, 381, 263

Baker, C. T. H. 1977, The numerical treatment of integral equations, 2nd edn. (Clarendon Press)

Becker, R. H., White, R. L., & Helfand, D. J. 1995, ApJ, 450, 559

Becla, J., Hanushevsky, A., Nikolaev, S., Abdulla, G., Szalay, A., Nieto-Santisteban, M., Thakar, A., & Gray, J. 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 6270, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series

Belabbas, M.-A., & Wolfe, P. J. 2009, Proc. Natl. Acad. Sci., 106, 369

Bengio, Y., Paiement, J., Vincent, P., Delalleau, O., Le Roux, N., & Ouimet, M. 2004, in Advances in Neural Information Processing Systems 16, ed. S. Thrun, L. Saul, & B. Schölkopf (Cambridge, MA: MIT Press), 177–184

Benitez, N. 2000, The Astrophysical Journal, 536, 571

Bianchi, L., & The GALEX Team. 1999, Memorie della Societa Astronomica Italiana, 70, 365

Bica, E. 1988, A&A, 195, 76

Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, AAP, 363, 476

Boroson, T. A., & Green, R. F. 1992, ApJS, 80, 109

Boroson, T. A., & Lauer, T. R. 2009, Nature, 458, 53

Boyle, B. J., Croom, S. M., Hoyle, F., Outram, P. J., Shanks, T., Smith, R. J., Miller, L., & Loaring, N. S. 2001, in Astronomical Society of the Pacific Conference Series, Vol. 232, The New Era of Wide Field Astronomy, ed. R. Clowes, A. Adamson, & G. Bromage, 65–+

Brinchmann, J., Charlot, S., White, S. D. M., Tremonti, C., Kauffmann, G., Heckman, T., & Brinkmann, J. 2004, MNRAS, 351, 1151

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000

Budavári, T., et al. 2001, AJ, 122, 1163

Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245

Cid Fernandes, R., Gu, Q., Melnick, J., Terlevich, E., Terlevich, R., Kunth, D., Rodrigues Lacerda, R., & Joguet, B. 2004, MNRAS, 355, 273

Cid Fernandes, R., Mateus, A., Sodré, L., Stasińska, G., & Gomes, J. M. 2005, MNRAS, 358, 363

Cid Fernandes, R., Sodré, L., Schmitt, H. R., & Leão, J. R. S. 2001, MNRAS, 325, 60

Coifman, R. R., & Lafon, S. 2006, App. Comput. Harmon. Anal., 21, 31

Condon, J. J., Cotton, W. D., Greisen, E. W., Yin, Q. F., Perley, R. A., Taylor, G. B., & Broderick, J. J. 1998, AJ, 115, 1693

Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. 1995, AJ, 110, 1071

Cuillandre, J.-C., & Bertin, E. 2006, in SF2A-2006: Semaine de l'Astrophysique Francaise, ed. D. Barret, F. Casoli, G. Lagache, A. Lecavelier, & L. Pagani, 265–+

Davis, M., et al. 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4834, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. P. Guhathakurta, 161–172

Einbeck, J., Evers, L., & Bailer-Jones, C. 2007, ArXiv e-prints

Feldmann, R., et al. 2006, MNRAS, 372, 565

Freeman, P. E., Newman, J. A., Lee, A. B., Richards, J. W., & Schafer, C. M. 2009, MNRAS, accepted

Friedman, J. H., Bentley, J. L., & Finkel, R. A. 1977, ACM Trans. Math. Softw., 3, 209

Gehrels, N., & Michelson, P. 1999, Astroparticle Physics, 11, 277

Georgakakis, A., Georgantopoulos, I., Stewart, G. C., Shanks, T., & Boyle, B. J. 2003, MNRAS, 344, 161

Gray, A. G., Moore, A. W., Nichol, R. C., Connolly, A. J., Genovese, C., & Wasserman, L. 2004, in Astronomical Society of the Pacific Conference Series, Vol. 314, Astronomical Data Analysis Software and Systems (ADASS) XIII, ed. F. Ochsenbein, M. G. Allen, & D. Egret, 249–+

Kaiser, N., & Pan-STARRS Team. 2002, in Bulletin of the American Astronomical Society, Vol. 34, Bulletin of the American Astronomical Society, 1304–+

Kambhatla, N., & Leen, T. K. 1997, Neural Comput., 9, 1493

Lafon, S., Keller, Y., & Coifman, R. R. 2006, IEEE Trans. Pattern Anal. Mach. Intell., 28, 1784

Lafon, S., & Lee, A. B. 2006, IEEE Trans. Pattern Anal. Mach. Intell., 28, 1393

Lee, A. B., & Wasserman, L. 2008, ArXiv e-prints

Madgwick, D. S., et al. 2003, ApJ, 599, 997

Mathis, H., Charlot, S., & Brinchmann, J. 2006, MNRAS, 365, 385

Moultaka, J., Boisson, C., Joly, M., & Pelat, D. 2004, A&A, 420, 459

Murakami, H., et al. 2007, PASJ, 59, 369

Ocvirk, P., Pichon, C., Lançon, A., & Thiébaut, E. 2006, MNRAS, 365, 74

Panter, B., Heavens, A. F., & Jimenez, R. 2003, MNRAS, 343, 1145

Panter, B., Jimenez, R., Heavens, A. F., & Charlot, S. 2007, MNRAS, 378, 1550

Reichardt, C., Jimenez, R., & Heavens, A. F. 2001, MNRAS, 327, 849

Richards, J. W., Freeman, P. E., Lee, A. B., & Schafer, C. M. 2009a, ApJ, 691, 32

—. 2009b, MNRAS, accepted

Rogers, B., Ferreras, I., Lahav, O., Bernardi, M., Kaviraj, S., & Yi, S. K. 2007, MNRAS, 382, 750

Schneider, D. P., et al. 2007, AJ, 134, 102

Skrutskie, M. F., et al. 2006, AJ, 131, 1163

Sulentic, J. W., Marziani, P., Zamanov, R., Bachev, R., Calvani, M., & Dultzin-Hacyan, D. 2002, ApJL, 566, L71

Sulentic, J. W., Zwitter, T., Marziani, P., & Dultzin-Hacyan, D. 2000, ApJL, 536, L5

Tremonti, C. A., et al. 2004, ApJ, 613, 898

Vanden Berk, D. E., et al. 2006, AJ, 131, 84

Wright, E. L., et al. 2009, ApJS, 180, 283

York, D. G., et al. 2000, AJ, 120, 1579

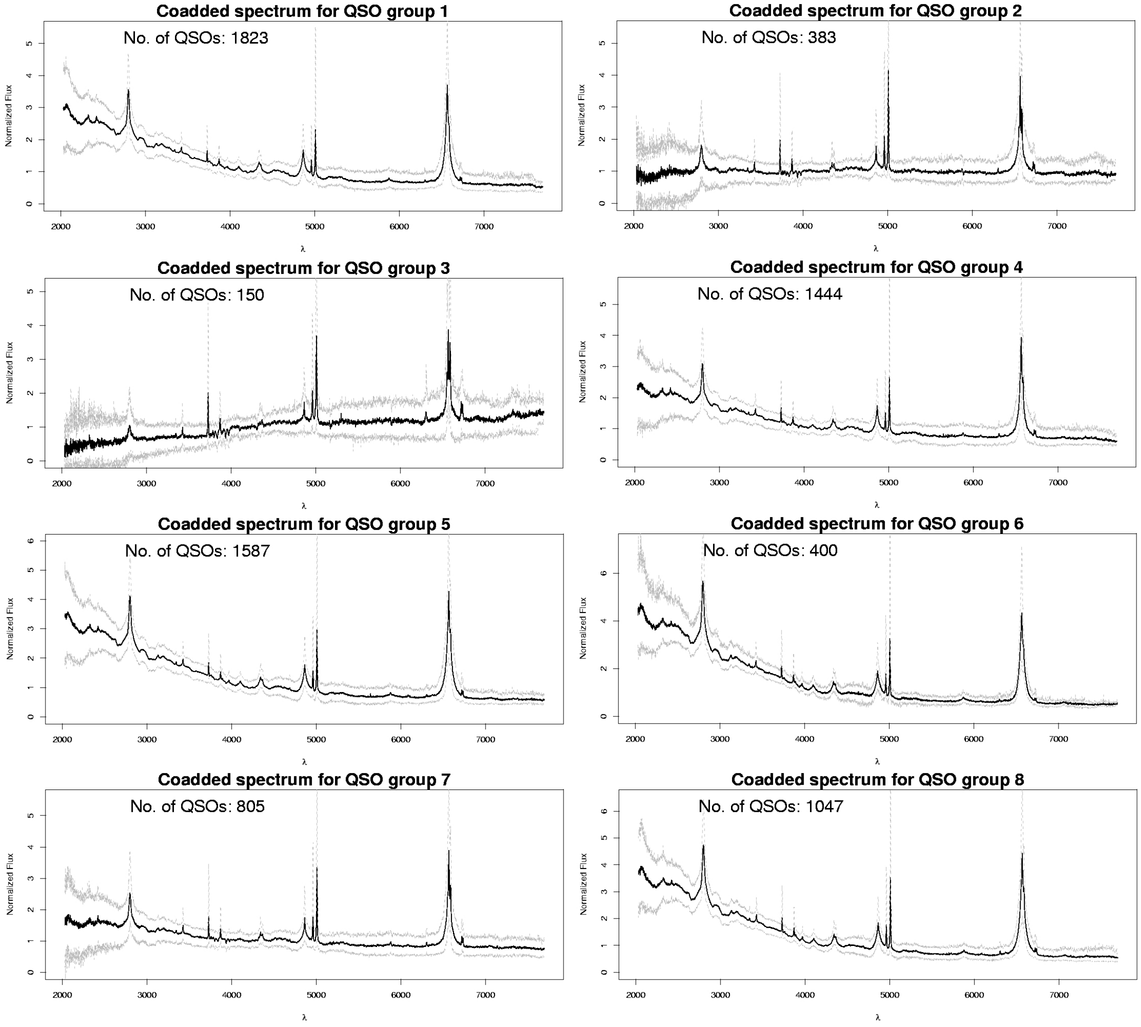


Figure 5: *Average spectrum for each of eight diffusion K-means clusters for 7639 quasars.* Plotted are the 8 mean rest-frame spectra plus 90% point-wise confidence bands. There are obvious differences between the average spectrum in each cluster, especially for shorter wavelengths. Point-wise confidence bands are generally small.