University of California
Los Angeles

Learning a simplicial structure using sparsity

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

John Flynn

2014

Abstract of the Thesis

Learning a simplicial structure using sparsity

by

John Flynn

Master of Science in Statistics

University of California, Los Angeles, 2014

Professor Alan Loddon Yuille, Chair

We discuss an application of sparsity to manifold learning. We show that the activation patterns of an over-complete basis can be used to build a simplicial structure that reflects the geometry of a data source. This approach is effective when most of the variability of the data is explained by low dimensional geometrical structures. Then the simplicial structure can be used as a platform for local classification and regression.

The thesis of John Flynn is approved.

Frederic Paik Schoenberg

Ying Nian Wu

Alan Loddon Yuille, Committee Chair

University of California, Los Angeles

2014

to my family

Table of Contents

List of Figures

Acknowledgments

I'm very grateful to Alan Yuille, Ying Nian Wu, and Chunyu Wang for their assistance. However any errors or misconceptions are very much mine!

CHAPTER 1

Introduction

There is a vibrant literature on the use of sparsity for learning. Some well known examples include Tibshirani's lasso [Tib96] for variable selection, Olshausen and Field's [Oo96] use of a sparse coding model to describe image patches, and Tao and Candes [CT05] work on sparse signal recovery. More recently Wu et al. [WSG10] use sparse coding in generative models for images and for object detection and recognition and Wang et al. [WWL12] use sparse coding in an application to 3d pose reconstruction. These are only a few samples from a very extensive literature.

The application here is to data sources where the variability comes mostly from geometry, where the data lies close to a small number of low dimensional structures. In that situation the activations of a sparse coding are manageable and lead to a low dimensional simplicial structure that echoes the geometrical structure of the data.

The method produces a simplicial structure in the same ambient Euclidean space as the data. Similar to k-means there is an assignment of training data points to parts of the structure, but here points are assigned to simplices rather than to point centers. The simplicial structure is derived from faces of a convex polytope constructed from the bases found in a sparse coding of the training data.

We haven't explored how, in a formal sense, the topology of the simplicial structure reflects the topology of the data. Carlsson and de Silva explore topological approximation by small simplicial complexes in [CD03]. They build their simplicial structure using a neighborhood structure based on an estimate, like

isomap uses, of geodesic distances, and they make formal comparisons with other simplicial structures such as the Čech and Rips complexes. While we haven't made formal comparisons our method seems to capture the geometrical essentials of the arrangement of the data.

The practical applications of our method might be similar to those where arrangements of hyperplanes are used to capture the geometry of the data [CPR12, PRA13, EV13]. We examine similar applications, to 3d pose recovery from 2d images, and to digit classification and faces under various lighting. In these data most of the variability comes from geometry.

The simplicial structure can be used as a platform for local exploration of the data. In some applications to classification it might turn out that the simplices are fairly pure, that is, that one class predominates on each simplex, but in more complicated situations the local simplicial coordinate system might be used for regression or classification on each simplex.

Since the simplicial structure is built from an overcomplete basis there are only two parameters, that is, the usual parameter for the l_1 penalty and the number of bases. There is no need for an a priori estimate of intrinsic dimension, or an assumption that the data manifold has a single geometrical component, or that the geometrical components have the same dimension.

We apply this construction to some synthetic data as a proof of concept. In Figure 1.1 the green circles are random samples from a spiral and a plane, the blue points are a sparse basis and the green simplices are the simplicial structure coming from the activations of the sparse coding.

The document is arranged as follows. Chapter 2 has a discussion of the geometry of the l_1 penalty. Chapter 3 describes the construction of a simplicial structure and Chapter 4 discusses some applications.

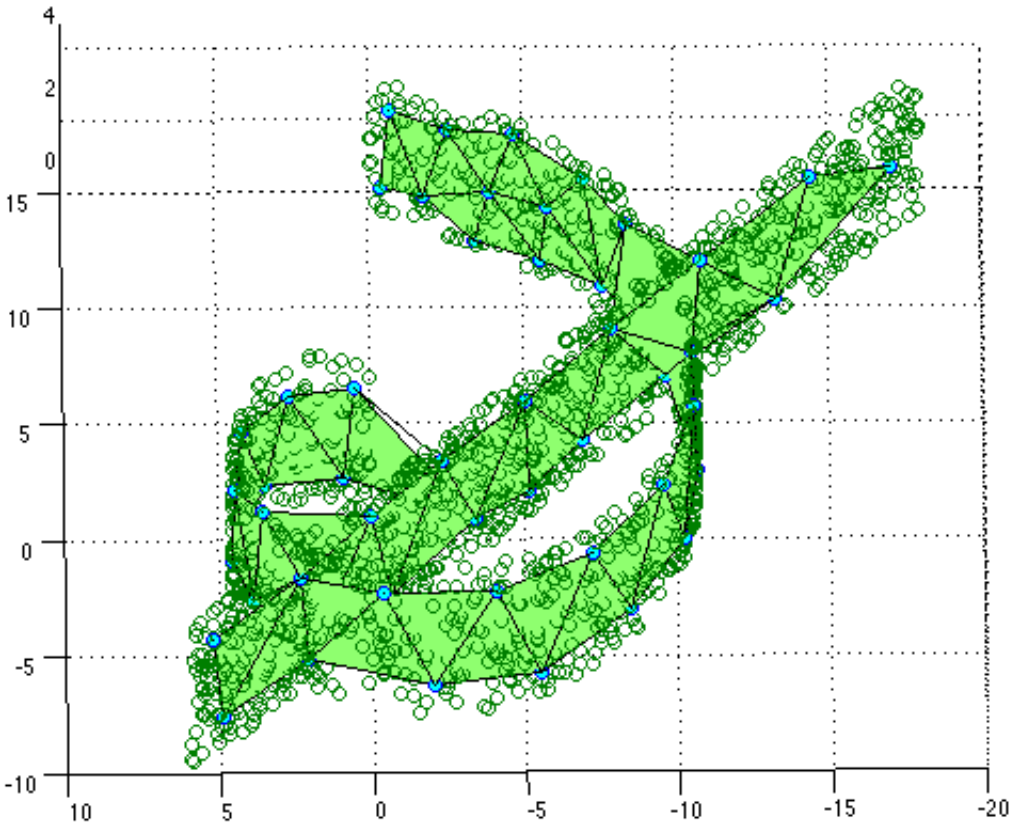


Figure 1.1: Points from a spiral and a plane, together with a simplicial structure.

CHAPTER 2

The geometry of the l_1 penalty

It is well known that the l_1 penalty induces sparsity, and that it conveys some robustness to noise. Here, we discuss some geometry associated with the penalty. There is a sophisticated discussion of this geometry in [Don05].

2.1 l_1 penalized regression

The standard l_1 penalized regression is the convex minimization problem

$$\min_{\beta} \left(\frac{1}{2} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right) \tag{2.1}$$

where $y \in \mathbb{R}^d$ is a vector of outcomes, $X \in \mathbb{R}^{d \times p}$ is a matrix of predictors, and $\lambda \in R_{\geq 0}$ is a penalty parameter. The minimizer $\hat{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients. It is common to standardize the columns of X, so that each column has l_2 norm one. We assume this standardization here.

It is well known that while the solution to the unpenalized problem may have no zero entries, the solution $\hat{\beta}$ becomes sparse as λ increases. Indeed if $\lambda > ||X^t y||_\infty$ then $\hat{\beta} = \vec{0}$, [OPT00]. Figure 2.1 shows a plot of coefficients $\hat{\beta}$ against λ — notice the increasing sparsity and shrinkage. We call the locations of the non-zero entries of $\hat{\beta}$ the activations.

Tibshirani discusses some uniqueness questions in [Tib13]. The solution $\hat{\beta}$ may not be unique; however it is when X is in general position. Thus if X is perturbed by small random noise then the solution for the perturbed system is unique with probability 1. Tibshirani establishes that, when $\lambda > 0$, all solutions have the same

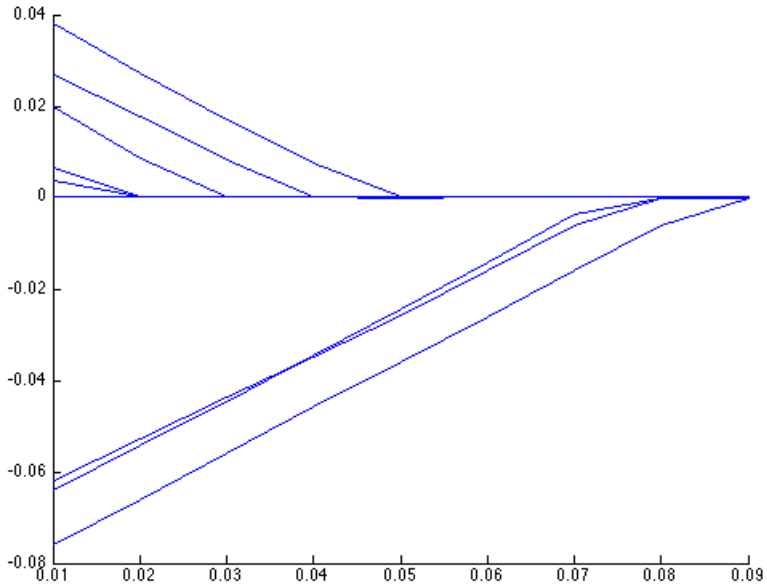


Figure 2.1: Regression coefficients $\hat{\beta}_i$ versus λ.

norm $||\hat{\beta}||_1$ and the same projection $X\hat{\beta}$. We'll revisit the uniqueness question later.

It is intuitive that $||\hat{\beta}||_1$ decreases with λ. To see this suppose that $\lambda_1 < \lambda_2$ and denote the corresponding minimizers by $\hat{\beta}_1$ and $\hat{\beta}_2$. Since $\hat{\beta}_1$ is the minimizer for $\frac{1}{2}||y - X\beta||_2^2 + \lambda_1||\beta||_1$ we obtain

$$\frac{1}{2}||y - X\hat{\beta}_1||_2^2 + \lambda_1||\hat{\beta}_1||_1 \leq \frac{1}{2}||y - X\hat{\beta}_2||_2^2 + \lambda_1||\hat{\beta}_2||_1$$

thus

$$\lambda_1(||\hat{\beta}_2||_1 - ||\hat{\beta}_1||_1) \geq \frac{1}{2}\left(||y - X\hat{\beta}_1||_2^2 - ||y - X\hat{\beta}_2||_2^2\right).$$

Similarly, since $\hat{\beta}_2$ is the minimizer for $\frac{1}{2}||y - X\beta||_2^2 + \lambda_2||\beta||_1$, we obtain

$$\lambda_2(||\hat{\beta}_1||_1 - ||\hat{\beta}_2||_1) \geq \frac{1}{2}\left(||y - X\hat{\beta}_2||_2^2 - ||y - X\hat{\beta}_1||_2^2\right).$$

Adding the last two inequalities we obtain

$$(\lambda_2 - \lambda_1)(||\hat{\beta}_1||_1 - ||\hat{\beta}_2||_1) \geq 0,$$

thus $||\hat{\beta}_1||_1 \geq ||\hat{\beta}_2||_1$.

2.2 Geometry of l_1 penalized regression

We now relate l_1 penalized regression to the geometry of the convex hull of the columns of $[X, -X]$. Denote this hull by C_X. We'll see that the projection $X\hat{\beta}$ corresponding to a minimizer $\hat{\beta}$ lies on a facet of $||\hat{\beta}||_1 C_X$.

First we observe that for any $\beta \in \mathbb{R}^p$, there is a vector z in the convex hull C_X such that

$$X\beta = ||\beta||_1 z.$$

To see this, denote the columns of X by $x_1, \ldots, x_p$. Then

$$\begin{aligned} X\beta &= \beta_1 x_1 + \cdots + \beta_p x_p \\ &= ||\beta||_1 z, \text{ where } z = \left(\sum_{i:\beta_i \geq 0} \frac{|\beta_i|}{||\beta||_1} x_i + \sum_{i:\beta_i < 0} \frac{|\beta_i|}{||\beta||_1} (-x_i) \right) \end{aligned}$$

and $z \in C_X$ since

$$\sum_{i:\beta_i \geq 0} \frac{|\beta_i|}{||\beta||_1} + \sum_{i:\beta_i < 0} \frac{|\beta_i|}{||\beta||_1} = 1.$$

Next we establish that $X\hat{\beta}$ lies on a facet of the convex polytope $||\hat{\beta}||_1 C_X$ when $\hat{\beta} \neq 0$, (when $\hat{\beta} = 0$ the polytope $||\hat{\beta}||_1 C_X$ doesn't have any facets!) Towards this, observe that C_X is centrally symmetric (that means that $-z \in C_X$ whenever $z \in C_X$), therefore the points on the facets of C_X are exactly those points that lie in C_X but not in rC_X for any $r < 1$. Similarly the facets of $||\hat{\beta}||_1 C_X$ are comprised of those points that lie in $||\hat{\beta}||_1 C_X$ but not in rC_X for any $r < ||\hat{\beta}||_1$. Finally, notice that $X\hat{\beta}$ does not line in rC_X for any $r < ||\hat{\beta}||_1$, since if $X\hat{\beta} = X\beta_1$ with $||\beta_1||_1 < ||\hat{\beta}||_1$ then

$$\begin{aligned} \frac{1}{2}||y - X\beta_1||_2^2 + \lambda||\beta_1||_1 &= \frac{1}{2}||y - X\hat{\beta}||_2^2 + \lambda||\beta_1||_1 \\ &< \frac{1}{2}||y - X\hat{\beta}||_2^2 + \lambda||\hat{\beta}||_1, \end{aligned}$$

but this a contradiction since $\hat{\beta}$ minimizes $\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$.

This gives an understanding of the possible activation patterns for penalized regressions on X. The possible activations are the combinations of signed columns of X such that the convex hull of these signed columns is contained in a facet of C_X. Combinations of columns of X where the convex hull lies in the interior of C_X are not activated by any y or λ.

In [Don05] Donoho discusses possible activation patterns with a view to sparse signal recovery. He calls a polytope k-neighborly if the convex hull of any k vertices spans a face. In that situation any pattern of k can be activated. He discusses the asymptotics of k-neighborliness for random matrices.

This geometric perspective on the l_1 penalty gives us an understanding of the possible uniqueness of $\hat{\beta}$. If X is in general position then C_X is simplicial and every point on a facet is expressible in a unique way as convex combination of vertices. To construct a matrix X such that for some y and λ the minimizer $\hat{\beta}$ is not unique, build a polytope in the sphere such that some facet is not a simplex, see Figure 2.2.

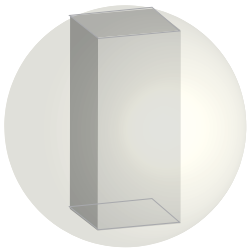


Figure 2.2: An example of C_X where there in an non unique $\hat{\beta}$ for some y

But we are wandering from our path! To summarize our geometrical understanding. We've seen that the minimizer $\hat{\beta}$ is unique for matrices that are constructed from some continuous random process. The activation pattern in $\hat{\beta}$ corresponds to the vertices on a proper face of the polytope C_X. This is depicted in Figure 2.3. This understanding of activation will be important in the sequel.

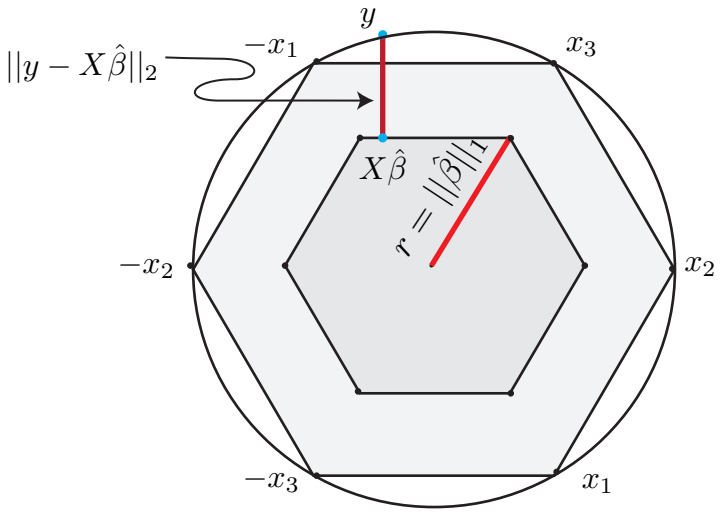


Figure 2.3: The convex hull C_X of $[X, -X]$, with the projection $X\hat{\beta}$ of y onto a facet of $||\hat{\beta}||_1 C_X$. Positions 1 and 3 are activated.

CHAPTER 3

Sparse coding and activated simplices

3.1 Sparse coding

Sparse coding is a widely used technique for compression of high dimensional data. Olshausen and Field observe in [Oo96] that the receptive fields of V1 cells in the cerebral cortex resemble an over-complete basis for image patches. Figure 3.1 shows a sparse coding of 14×14 image patches from images of scenes in nature. Baraiuik et al. [BCD10] and Huang et al. [HZM11] discuss structured sparse coding, which uses activation patterns of an overcomplete basis to improve reconstruction. One might argue that understanding activation patterns in sparse coding is key to understanding datasets in high dimensions. We build a simplicial structure on naive activation statistics. This is a useful approach for data on low dimensional structures in moderate dimensional ambient spaces.

We now establish a framework for a discussion of sparse coding. Suppose we have a set of n samples $\mathcal{Y}$ from a data source in $\mathbb{R}^d$. A sparse coding of $\mathcal{Y}$, using p sparse bases and at penalty level $\lambda > 0$, is one part, X, of an approximate solution to the minimization

$$\min_{X,\beta} \sum_{y \in \mathcal{Y}} \left(\frac{1}{2} \|y - X\beta_y\|_2^2 + \lambda \|\beta_y\|_1 \right) \tag{3.1}$$

where $X \in \mathbb{R}^{d \times p}$ is restricted in the minimization to have columns with l_2 norm bounded by 1, and β is unrestricted in $\mathbb{R}^{p \times n}$. In non-degenerate situations (where each column of X is non-zero and has a non-zero coefficient in some β_y, that is,

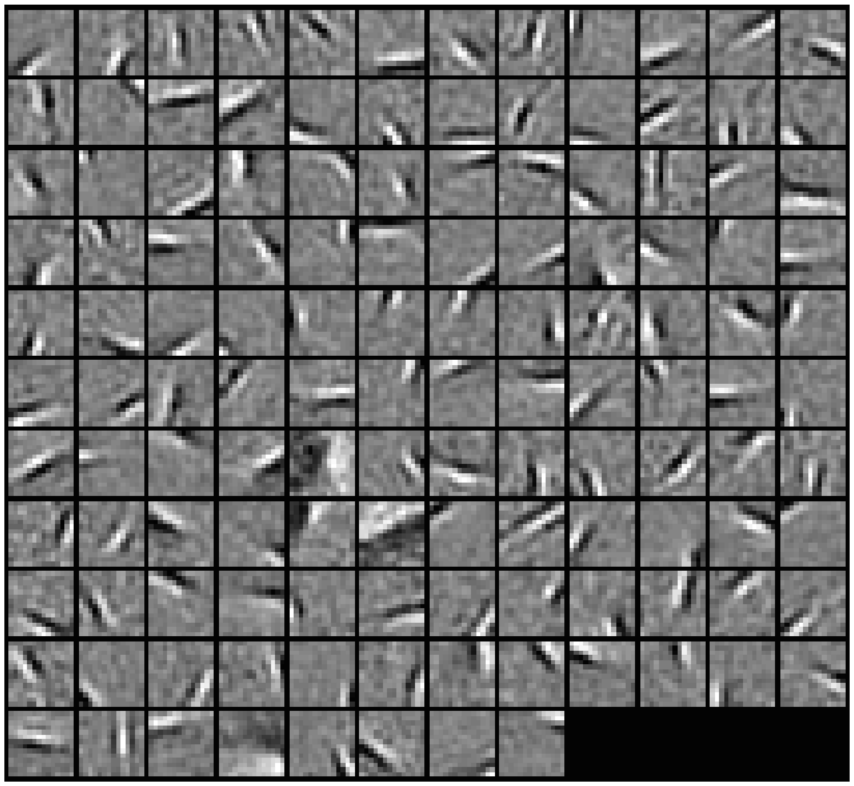


Figure 3.1: A sparse coding of 14×14 image patches.

where λ is not too large and the sparse bases are not too numerous) the minimizer X has columns with l_2 norm exactly 1.

To see this, suppose that X doesn't have this normalization; let $\tilde{X}$ be the corresponding matrix with normalized columns and let $\tilde{\beta}_y$ be the corresponding coefficient vector, but with the coefficients scaled so that $X\beta_y = \tilde{X}\tilde{\beta}_y$, then $\|y - X\beta_y\|_2 = \|y - \tilde{X}\tilde{\beta}_y\|_2$, but $\|\tilde{\beta}_y\|_1 \leq \|\beta_y\|_1$, so $\tilde{X}, \tilde{\beta}$ is a better solution to (3.1). We assume this non-degeneracy from now on. Then a sparse basis consists of p points on the hypersphere S^{d-1} in $\mathbb{R}^d$.

The minimization 3.1 is difficult — it is a non-convex problem that is convex in each variable X, β, when the other is held fixed. We'll discuss this more later.

We might think of a sparse coding as a well positioned basis that allows an efficient sparse expression for each training point y. Indeed the coefficients β_y are the l_1 penalized regression coefficients for regression of y on X. We continue to use the language of activations as before, we say that a set of bases is activated by y if the corresponding entries of β_y are non-zero.

The relationship between the geometry of X and the training data $\mathcal{Y}$ is complicated, without some further assumptions on $\mathcal{Y}$. Figure 3.2 shows a sparse basis on $S^2 \subset \mathbb{R}^3$ learned from points sampled from the union of two orthogonal ellipses in $\mathbb{R}^3$, each with eccentricity 3. Notice that the bases congregate under the major axes of the ellipses. This effect occurs because, for fixed X, the penalty

$$\min_{\beta_y} \left(\frac{1}{2} \|sy - X\beta_y\|_2^2 + \lambda \|\beta_y\|_1 \right)$$

scales super-linearly with s for $s > 1$. Thus the bases tend to congregate in the learning process to reduce errors for distant points. There are also congregating effects when a non-uniform is induced by the projection of the data onto the hypersphere. The relationship is further complicated if many folds of the data lie over the same points on the hypersphere. Figure 3.3 illustrates this.

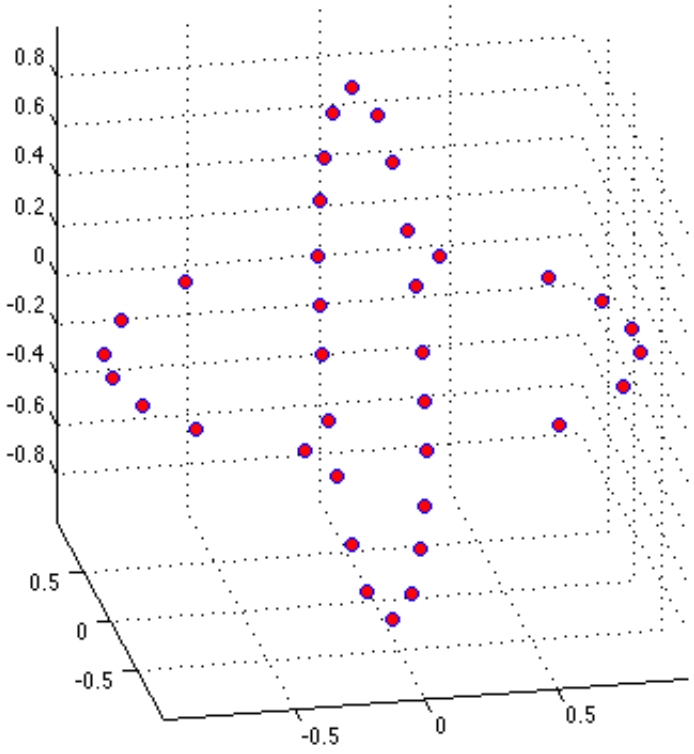


Figure 3.2: A sparse coding for data sampled from two ellipses in $\mathbb{R}^3$. The bases lie on the surface S^2 and congregate under the major axes of the ellipses.

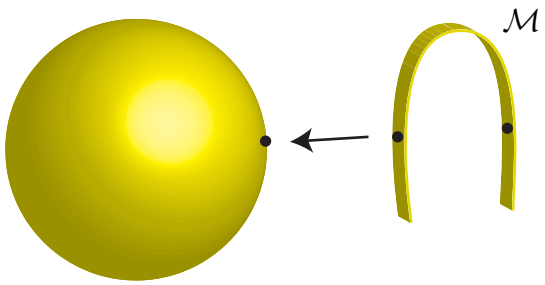


Figure 3.3: Multiple folds of the data manifold over the same point on the sphere. This complicates the relationship between a sparse basis and the data.

Assumption: To simplify the relationship between the data and sparse basis, and to ensure that the sparse basis is directly comparable to the data, we now assume that the training data $\mathcal{Y}$ lies on the sphere $S^{d-1} \subset \mathbb{R}^d$.

This is not a restrictive assumption. Many data sets are normalized in this way; for example, image patches are often contrast normalized. However in some situations a direct normalization of $\mathcal{Y}$ is inappropriate, but then, there are more benign approaches to map the data onto a hypersphere. We'll discuss something like stereographic projection below. This approach preserves the geometry of the data quite well. Figure 3.4 shows data sampled from a spiral, together with a sparse basis. The spiral was mapped into a hypersphere S^3 in $\mathbb{R}^4$ by a radial projection from a hyperplane; the sparse basis was constructed in S^3 and then mapped back into the original ambient space of the data. Notice that the bases are intermingled among the training data.

It might be tempting to speculate that the sparse bases reflect the density of the data on the hypersphere, but this isn't quite right. A better understanding might be that the bases are positioned so that the faces of the polytope C_X provide an efficient projection of the data. Areas of high curvature require many bases for efficient reconstruction, but near planar areas can be reconstructed efficiently

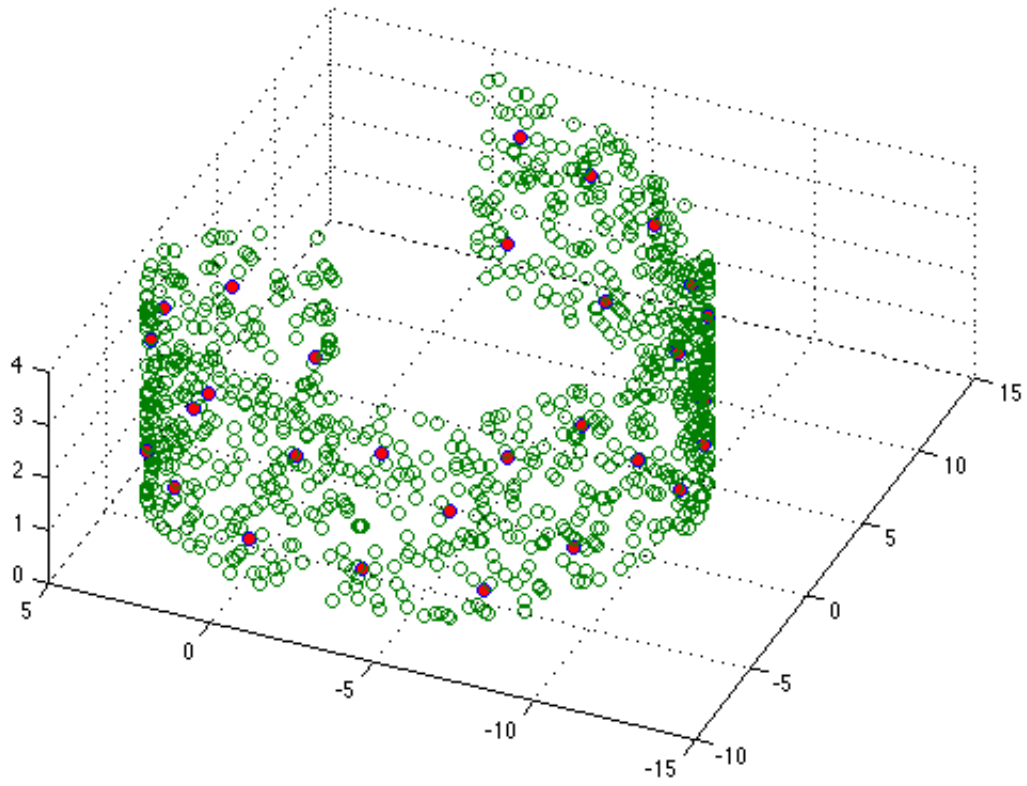


Figure 3.4: The green circles represent points sampled from a spiral; the red points are a sparse basis. The sparse basis was constructed on $S^3 \subset \mathbb{R}^4$.

with fewer bases. See Figure 3.5 for a conceptual illustration.

We would like to understand the sparse basis in terms of projection onto the polytope C_X, however it is somewhat more complicated. In fact, as we saw in the previous chapter, the projection $X\beta_y$ is onto a scaling $\|\beta_y\|_1 C_X$ of the polytope, and the amount of scaling depends on y.

We now advance an argument to show that when p is large the quantity $\|\beta_y\|_1$ is approximately constant. When the number of bases is large, the bases activated by a training point y are typically not far away. Therefore the projection error $||y - X\beta_y||_2$ is approximately $1 - \|\beta_y\|_1$. See Figure 3.6. Therefore $||\beta_y||_1$ approximately

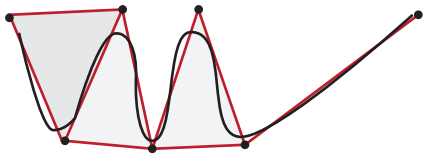


Figure 3.5: A conceptual drawing to illustrate that more bases are required at areas of high curvature. The black curve represents the data source, the black dots are vertices of C_X.

minimizes

$$\frac{1}{2}(1-r)^2 + \lambda r$$

and hence $||\beta_y||_1 \approx 1 - \lambda$ and $||y - X\beta_y||_2 \approx \lambda$. Our experiments show that these estimates are reasonable in practice.

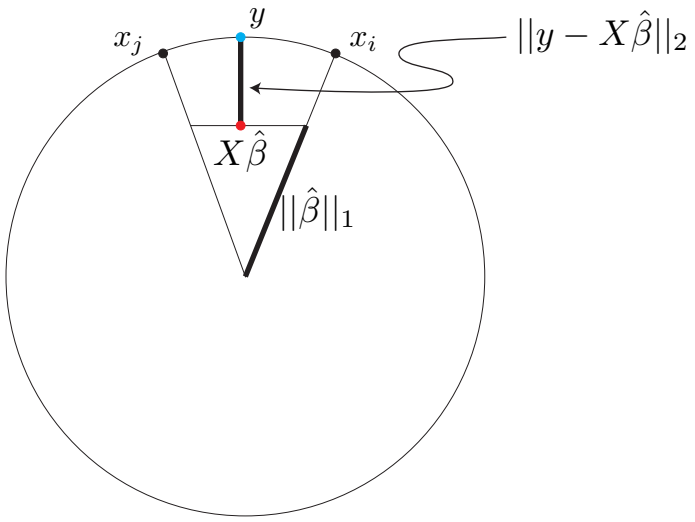


Figure 3.6: When there are many bases $\|y - X\beta_y\|_2 \approx 1 - \|\beta_y\|_1$.

Note that this approximation is invalid if y is out-of-sample. These estimates help in understanding the arrangement of sparse bases. In the learning process the polytope C_X adjusts so that low dimensional faces of λC_X are close to the training data.

3.2 Activated simplices

We're now ready to construct activated simplices from a sparse basis X learned from training data $\mathcal{Y}$.

Let y be a training data point; we say that y **activates the simplex**

$$\langle \mathrm{sgn}(\beta_{i_1}) x_{i_1}, \ldots, \mathrm{sgn}(\beta_{i_k}) x_{i_k} \rangle$$

if the non-zero locations of $\beta = \beta_y$ are $i_1, \ldots, i_k$. We know from the discussions of the previous chapter that this simplex is a face of the polytope C_X.

The **simplicial structure** built from the activation statistics is the set of activated simplices, but with some simplices dropped when they have very low activation rates.

It is hard to make precise statements about the reconstruction error in using the simplicial structure as a surrogate for the data source. We argued above that the bases X are adjusted in the learning process so that faces of C_X efficiently reflect the training data. However, it is unreasonable to expect that the training data is close to the activated simplices of C_X when λ is large, since the dimension of the activated simplices may be small in that situation. But the simplicial structure may still capture some essentials of the arrangement of the data — for example, it may be topologically equivalent to the data source. It is known that in practice sparse coding gives good a reconstruction for certain data types (for example, image patches) and since the simplicial structure is merely a description of the activation of the sparse bases, it should give good reconstruction in similar situations. In summary, the effectiveness of the simplicial structure as a reconstruction of the data may depend on the data source and λ, but the structure should be more resilient as a qualitative summary.

3.3 Projection onto hyperspheres

The problem of embedding data from $\mathbb{R}^d$ into a hypersphere is a matter of cartography. The geometry of the hypersphere is different from that of Eulcidean space, but small patches of Euclidean space can be embedded in the hypersphere without much distortion.

The standard stereographic projection (Figure 3.7) maps the Euclidean space $\mathbb{R}^d$ into the hypersphere $S^d \subset \mathbb{R}^{d+1}$. It distorts geometry only a little on small patches.

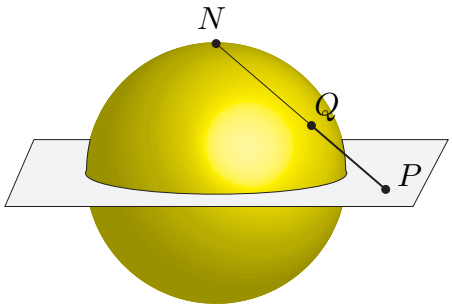


Figure 3.7: Stereographic projection from the north pole onto the equatorial plane

We've used what might be called *offset radial projection* in our experiments. See Figure 3.8. Offset radial projection is the map

$$x \in \mathbb{R}^d \mapsto (x, \text{offset}) \in \mathbb{R}^{d+1} \mapsto \frac{(x, \text{offset})}{||(x, \text{offset})||_2} \in S^d$$

It is revealing to describe this map in terms of a projection onto the tangent plane T^d of the hypersphere at the point $(0, \ldots, 0, 1)$

$$x \in \mathbb{R}^d \mapsto (x, \text{offset}) \in \mathbb{R}^{d+1} \mapsto \frac{(x, \text{offset})}{\text{offset}} \in T^d \mapsto \frac{(x, \text{offset})}{||(x, \text{offset})||_2} \in S^d$$

The map onto the tangent plane is simply a rescaling. If $\frac{\|x\|_2}{\text{offset}}$ is small then $\frac{(x,\text{offset})}{\text{offset}}$ is close to $(0, \ldots, 0, 1)$ and for such x the map from T^d to S^d distorts only a little.

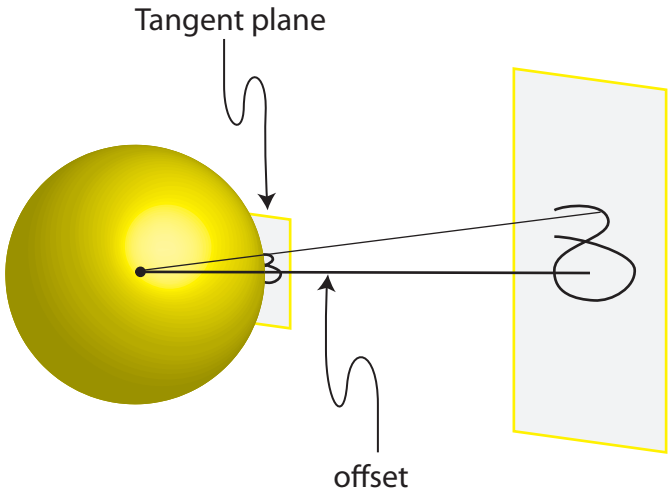


Figure 3.8: Offset radial projection from a hyperplane onto the hypersphere

If the dataset lies in a disk of radius R in $\mathbb{R}^d$, and $\frac{R}{\text{offset}}$ is small, then the dataset will be embedded in the hypersphere without much distortion. In practice nothing terrible happens when $\frac{R}{\text{offset}}$ is as large as 1.

These projections onto spheres have the disadvantage that they distort planes: a hyperplane will be bent somewhat in the projection. However since sparse basis constructions are essentially local these projections are benign.

CHAPTER 4

Computation and synthetic datasets

4.1 Synthetic Data

We performed some experiments on synthetic data as a proof of concept. We experimented with some data sampled from from the union of a hyperplane and a spiral ribbon. The data was mapped into $S^3 \subset \mathbb{R}^4$ using the offset radial projection described in Section 3.3. A sparse bases was learned for the data on the sphere, and the bases and active simplices were projected back to $\mathbb{R}^3$. The sparse bases were learned from an initialization using a subset of the training data selected uniformly at random.

We are interested in the effects of the parameters p and λ, and on the ability of the simplices to distinguish the two geometrical components of the data source. It is unsurprising that more sparse bases is better, and that a larger lambda activates lower dimensional simplices. For the dataset here, it seems that the sets of sparse bases found for different λ are qualitatively similar, but the activation patterns, and the simplicial structure that describes activations were different. It is interesting to speculate about the stability of the polytopes C_X, but this is hard to formalize; the higher dimensional activated faces of C_X seem to stable for ranges of λ.

Most activated simplices lie in a single geometrical component (which suggests that these simpicicial structures may be useful for some classification problems). The intrinsic dimension of the data is 2 and not too many 3 dimensional simplices

are activated, even at small λ.

Figure 4.1 shows the cross-section of the spiral ribbon and planar section at $z = 0$. Figure 4.2 shows the simplicial structure learned with 30 bases and $\lambda = 0.25$, and the corresponding histogram of activations. Though the intrinsic dimension of the data is 2 we do see a small number of activations of a 3-simplex. When there are only 30 bases some parts of the spiral ribbon are explained by a 1-simplex.

Figure 4.3 shows a simplicial structure learned with 50 bases and $\lambda = 0.25$, and the corresponding histogram of activations. The simplicial structure is more accurate with more bases, though it still has some difficulty separating the end of the ribbon from the planar part.

Figure 4.4 shows a simplicial structure learned with 50 bases but now with $\lambda = 0.55$. Now no 3d simplex is activated and more of the data is explained by 1d simplices.

Figure 4.5 shows a simplicial structure learned with 50 bases and with $\lambda = 0.75$. The activations of 2d simplices are now rare. Most of data is explained by 1d simplices. Interestingly it seems that the sparse bases and the polytope C_X are very similar for the various values of λ. We see a dramatic change in the activations but not in the polytope.

Finally Figure 4.6 shows a simplicial structure learned with 50 bases and with $\lambda = 0.1$. We see more activations of 3d simplices that occurred for $\lambda = 0.25$, but the differences are not dramatic.

4.2 Computation

We'll say only a little about computation since there is a well developed literature on the computational aspects of dictionary learning. There are two computational challenges — the computation may be slow, and the computation may converge on

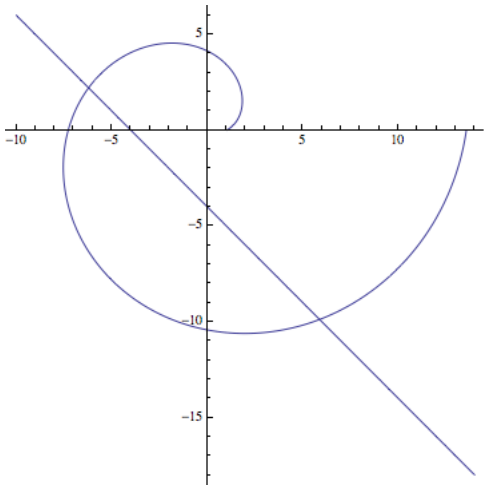
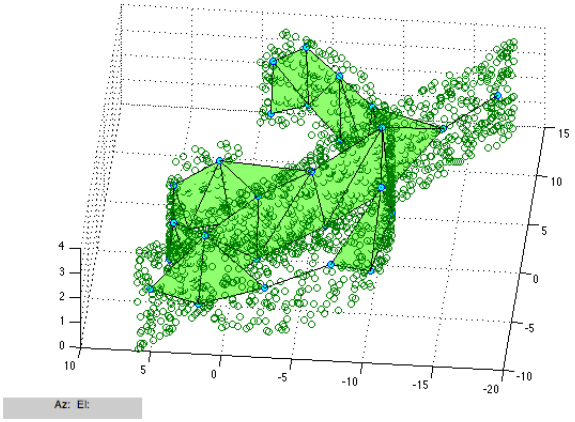
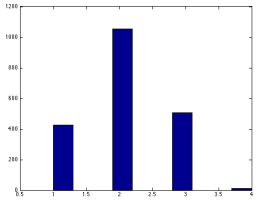


Figure 4.1: Cross section of the spiral ribbon and the hyperplane patch

an undesirable local minimum. Mairal's `SPAMS` package [MBP09] was efficient for the type of data we were interested in, but the problem of initialization, in order to avoid undesirable local minima remained. The standard initialization of the sparse code is to use a subset of the training data selected uniformly at random. A k-means++ initialization [AV07] seems to improve on this. Our intuition is that extreme positions among the training data are good locations for bases and the k-means++ initialization was better than a uniform random selection at placing bases near these locations.

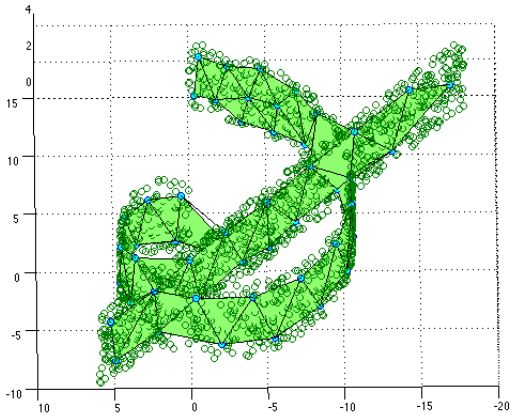


(a) Simplicial structure

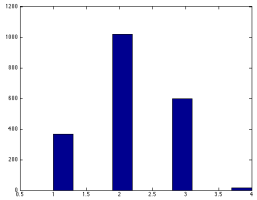


(b) Histogram of activations

Figure 4.2: Simplicial structure from 30 bases with $\lambda = 0.25$.

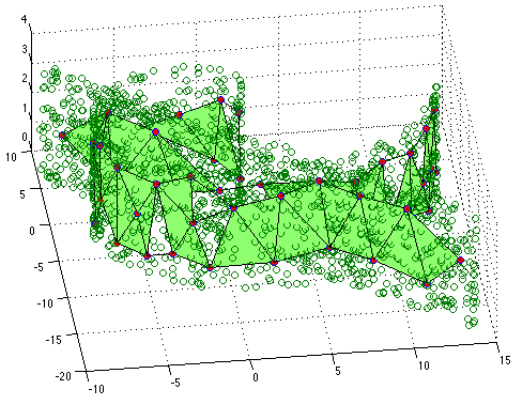


(a) Simplicial structure

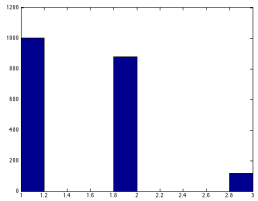


(b) Histogram of activations

Figure 4.3: Simplicial structure from 50 bases with $\lambda = 0.25$.

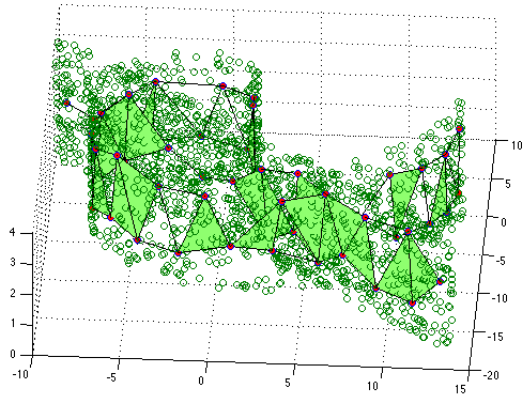


(a) Simplicial structure

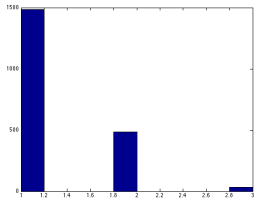


(b) Histogram of activations

Figure 4.4: Simplicial structure from 50 bases with $\lambda = 0.55$.

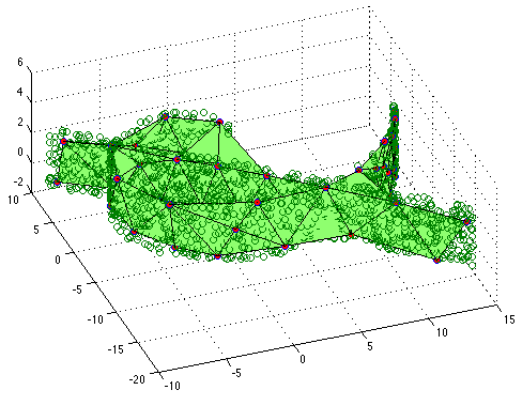


(a) Simplicial structure

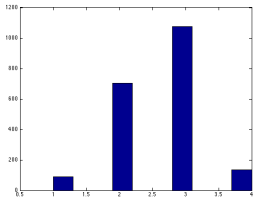


(b) Histogram of activations

Figure 4.5: Simplicial structure from 50 bases with $\lambda = 0.75$.



(a) Simplicial structure



(b) Histogram of activations

Figure 4.6: Simplicial structure from 50 bases with $\lambda = 0.1$.

CHAPTER 5

Applications

We now discuss some practical applications of the simplicial construction. Suppose we have a simplicial structure Δ, constructed from a sparse coding X of training data $\mathcal{Y}$ at l_1-penalty level λ_{cons}. We associate a new data point y with a simplex in Δ by minimizing the unpenalized l_2 distance from y to simplices in Δ:

$$\delta(y) = \arg\min_{\delta \in \Delta} d_2(y, \delta) \tag{5.1}$$

This might be a costly minimization since it involves a search over all simplices in Δ. The quantity $d_2(y, \delta)$ is

$$d_2(y, \delta) = \min_{\alpha_1, \ldots, \alpha_d} \|y - (\alpha_1 v_1 + \cdots + \alpha_d v_d)\|_2 \tag{5.2}$$

where $v_1, \ldots, v_d$ are the vertices of δ and $\alpha_1, \ldots, \alpha_d$ are constrained to be non-negative and sum to 1. But one might restrict the search to simplices δ that share some vertices with an l_1 penalized projection of y on X, projected using some small penalty λ_{search}.

The reconstruction error for Δ is estimated as the mean, over test data, of $d_2(y, \delta(y))$; in other words, it is the average distance of test data from the simplices.

In our practical applications, regression and classification are carried out by local classification or regression on the simplices of Δ. Thus, for classification applications, a data point y is classified using some local classification scheme on $\delta(y)$. This local scheme might simply be nearest neighbor classification on $\delta(y)$, or it might involve a logistic regression built on the simplicial coordinates of $\delta(y)$.

We experiment on some real data where the intrinsic dimension is reasonable and the variability is strongly explained by geometry. We experimented with datasets of 3d poses, of faces under different lighting conditions, and of handwritten digits. These applications are comparable to those in Pitelis at al. [PRA13] and the results are competitive. We only give an overview of results of these applications here because our computations are still quite raw. The details will be presented in another publication.

5.1 3d poses

In [WWL12] Wang et al. discuss the recovery of human 3d poses from a single image with unknown camera viewpoint. They use an overcomplete basis to model 3d poses and estimate a 3d pose from an image by simultaneously estimating the camera parameters and pose to minimize a distance of the projection from the 2d joint positions inferred from the image.

We experimented with restricting the activations of sparse combinations of pose bases to those that had been activated in training, that is, we used the simplicial structure Δ as a model for 3d poses. We found that this improves the reconstruction of poses in some experiments with data with known ground truth. Figure 5.1 shows histograms of the reconstruction errors when poses are restricted to the activated simplices, and of the reconstruction errors when poses are reconstructed as more general combinations of the bases.

We can understand this in terms of the polytope C_X. When the training data has low intrinsic dimension only certain low dimensional faces of the polytope are activated. However any face of C_X can occur as an outcome of l_1 penalized regression. The faces that aren't activated during training correspond to unrealistic combinations. Figure 5.2 is a conceptual illustration. It shows two red circles representing the training data, and it shows a wireframe for the polytope C_X.

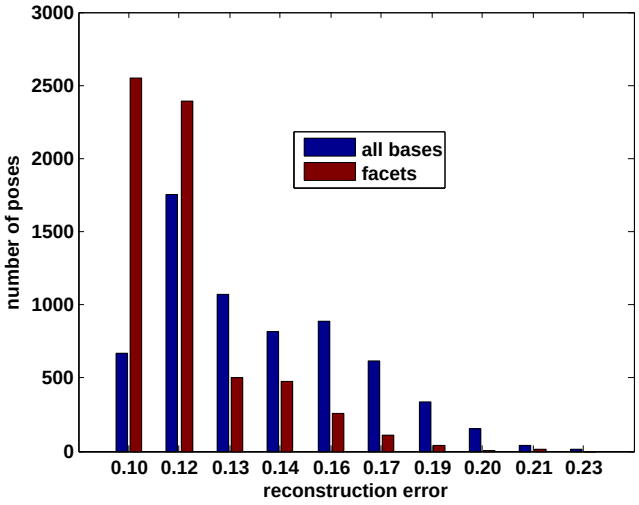


Figure 5.1: Histograms of reconstruction errors for 3d pose reconstruction, when poses are reconstructed using the simplicial structure (red), and when poses are reconstructed as penalized combinations of the bases (blue). The simplicial structure is better.

Only the simplices corresponding to line segments under the circles are activated in training. However any face of C_X can be activated by l_1 penalized regression.

5.2 Faces under varying illumination

It is well known that images generated from a single face, in a fixed pose, but with varying lighting, lie very close to a 9 dimensional hyperplane in image space. In [EV13] Vidal has a graphic showing the decay of the singular values for data coming from a single face, pose, under varying lighting. It shows that most of the variation from lighting is captured by the first few principal directions.

When images are generated from a small number of face-poses, under different lighting, the data is explained well by a union of low dimensional hyperplanes (one hyperplane per face-pose). Thus most of variability in this data can be explained by low dimensional geometrical structures.

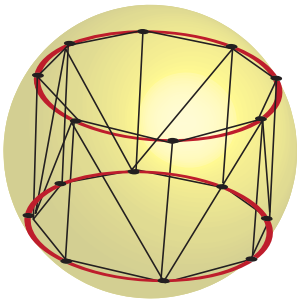


Figure 5.2: A conceptual illustration showing a polytope C_X learned from training data corresponding to the two red circles. Any face of the polytope can be activated by l_1 penalized regression, but only the one dimensional simplices under the circles correspond to the training data. The wireframe outlines the faces of the polytope.

Elhamifar et al, [EV13], Pitelis at al. [PRA13] and others explore some classification experiments on this kind of data. In similar experiments we classify images by the nearest neighbor classification on the nearest active simplex, and obtain competitive results.

5.3 Digits from Semeion

The Semeion digit dataset is a small dataset of 1593 16×16 grey scale images of handwritten digits. We explore classification experiments similar to those in [EV13] and [PRA13] and obtain competitive results, using nearest neighbor classification on the nearest active simplex.

References

[AV07] D Arthur and S Vassilvitskii. "k-means++: The advantages of careful seeding." *Proceedings of the eighteenth annual ACM- ...*, 2007.

[BCD10] R G Baraniuk, V Cevher, M F Duarte, and C Hegde. "Model-Based Compressive Sensing." *Information Theory, IEEE Transactions on*, **56**(4):1982–2001, 2010.

[CD03] Gunnar Carlsson and Vin De Silva. "Topological approximation by small simplicial complexes." Technical report, 2003.

[CPR12] G D Canas, Tomaso A Poggio, and Lorenzo Rosasco. "Learning Manifolds with K-Means and K-Flats." *NIPS*, 2012.

[CT05] E J Candes and T Tao. "Decoding by linear programming." *Information Theory, IEEE Transactions on*, **51**(12):4203–4215, 2005.

[Don05] David L Donoho. "Neighborly polytopes and sparse solutions of underdetermined linear equations." Technical report, 2005.

[EV13] E. Elhamifar and R. Vidal. "Sparse Subspace Clustering: Algorithm, Theory, and Applications." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **35**(11):2765–2781, Nov 2013.

[HZM11] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. "Learning with structured sparsity." *The Journal of Machine Learning Research*, **12**:3371–3412, 2011.

[MBP09] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. "Online dictionary learning for sparse coding." pp. 689–696, 2009.

[Oo96] B A Olshausen and others. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." *Nature*, **381**(6583):607–609, 1996.

[OPT00] Michael R Osborne, Brett Presnell, and Berwin A Turlach. "On the LASSO and its dual." *Journal of Computational and Graphical Statistics*, **9**(2):319–337, 2000.

[PRA13] N Pitelis, C Russell, and L Agapito. "Learning a Manifold as an Atlas." In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 1642–1649, 2013.

[Tib96] Robert Tibshirani. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B. Methodological*, pp. 267–288, 1996.

[Tib13] Ryan J Tibshirani. "The lasso problem and uniqueness." *Electronic Journal of Statistics*, **7**:1456–1490, 2013.

[WSG10] Ying Nian Wu, Zhangzhang Si, H Gong, and Song Chun Zhu. "Learning active basis model for object detection and recognition." *International Journal of Computer Vision*, **90**(2):198–235, 2010.

[WWL12] C Wang, Y Wang, Z Lin, Alan L Yuille, and Wen Gao. "Robust Estimation of 3D Human Poses from a Single Image." In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 4321–4328, 2012.