Data Representation and Basis Selection to Understand Variation of Function Valued Traits

Travis L. Gaydos

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2008

Approved by

Advisor: Dr. J. S. Marron

Reader: Dr. J. G. Kingsolver

Reader: Dr. D. G. Kelly

Reader: Dr. Haipeng Shen

Reader: Dr. M. R. Kosorok

ABSTRACT

TRAVIS L. GAYDOS: Data Representation and Basis Selection to Understand Variation of Function Valued Traits
(Under the direction of J. S. Marron)

Many fields, including evolutionary biology, collect data in which a curve corresponds to each individual. Therefore a curve is the statistical atom of analysis, which is an area in statistics known as Functional Data Analysis (FDA). A common goal in FDA is to understand the variation of curves. Often Principal Components Analysis (PCA) is a useful tool to do this. But PCA will often yield undesirable results if the amount of variation explained by directions is not significantly different.

Directions of low variation do not often explain significantly differing amounts of variation. Therefore in subspaces of low variation it is difficult to separate biological signal from noise using variation measures. In this dissertation a way to separate biological signal from noise by quantifying the simplicity structure of curves in a subspace of low variation is shown. Also asymptotic properties of subspaces of low variation and subspaces of biological signal are developed.

The results of PCA are highly dependent on the representation of the data as well. In this dissertation a representation of data curves similar to shape statistics is produced by exploiting the developmental stages of insects. This representation allows for variation, that is usually most efficiently modeled using non-linear methods when using typical FDA grid based representations of the data, to be modeled using linear PCA.

Also in the dissertation is a method to simultaneously visualize multiple t-tests to understand the slope structure of data sets.

CONTENTS

List of Figures **vi**

List of Tables **viii**

1 Introduction **1**

1.1 Introduction to Point Cloud and Object Space Using A Toy Example . . 2

2 Estimation of G and E **12**

2.1 Simulated Toy Data Set 13

2.2 Fixed Effects ANOVA on the Toy Data Set 16

2.2.1 Estimation of Phenotypic Variance 16

2.2.2 Estimation of Genetic Variance 19

2.2.3 Estimation of Environmental Variance 22

2.3 Random Effects ANOVA on the Toy Data Set 25

2.3.1 Estimation of Genetic Variance 25

3 Finding Genetic Constraints: A Simple Curve Basis Of A Nearly Null Space **28**

3.1 Introduction to the Simple Curve Basis of the Nearly Null Space 29

3.2 Measure of Simplicity 32

3.3 Method To Derive Simple Curve Basis 35

3.3.1 Methods Relation to PCA 35

3.3.2 Mathematical Derivation of F_{P_N} 38
3.4 Simple Curve Basis for Unevenly Spaced Environment Levels 40
3.5 Variance-Simplicity View of a Direction 42
3.6 Example: Caterpillar Growth Rate 43
3.6.1 Simple Curve Basis of $\mathbb{R}^6$ 44
3.6.2 PC basis 46
3.6.3 Simple Curve Basis of the Nearly Null Space 47

4 Principal Components For Developmental Stage Data 53
4.1 Introduction to Developmental Stage Landmark Data 54
4.2 PCA for Developmental Stage Landmark Data 56
4.3 Representation for a Differing Number of Developmental Stages 69
4.4 Results for the Manduca sexta Data Set 74
4.4.1 Original Raw Data Scale Landmark PCA 75
4.4.2 Landmark PCA on the Correlation Matrix 78
4.4.3 Landmark PCA on Trace Standardized Data 80

5 Hypothesis Test for Line Segment Slopes and Visualization 83
5.1 Introduction to the Data Set and Hypothesis Test for Slope Equality . . 84
5.2 Visualization of Results Of Multiple Slope Comparisons 87
5.3 Temperature Adjustment 93
5.4 Multiple Slope Comparisons for Multiple Temperature Data 98
5.5 Multiple Segment Length Comparisons 101

6 Mathematical Background 103
6.1 Geometric Introduction to Canonical Angles 103
6.2 Canonical Angles and Relation to CCA 111
6.2.1 CCA Calculations in Terms of Canonical Angles 111
6.2.2 CA calculations in Terms of CCA 121

6.3 Gap Metric 124
6.4 Euclidean Sine metric 127

7 Mathematical Statistic Investigation 131
7.1 Study of Nearly Null Space Asymptotic Properties 132
7.1.1 Definition of Nearly Null Space 132
7.1.2 Estimated Nearly Null Space Dimension Convergence 134
7.1.3 Convergence In Probability of the Nearly Null Space 144
7.2 Asymptotic Properties of the Interesting Genetic Constraint Space 150
7.2.1 Definition of the Interesting Genetic Constraint Space 150
7.2.2 Estimated Genetic Constraint Space Space Dimension Convergence 154
7.2.3 Convergence of the Estimated Interesting Genetic Constraint Space 159
7.3 Hypothesis Test for a Given Subspace contained in S 162

A Algebraic Justification 166
A.1 Algebraic Justification of F^{full} 166
A.2 Algebraic Justification of F 166

Bibliography 169

LIST OF FIGURES

1.1 Point Cloud and Object Space View of Toy Example 3

1.2 PC 1 of Toy Example 6

1.3 PC 2 of Toy Example 7

1.4 Smooth Curve Direction 1 for Toy Example 9

1.5 Smooth Curve Direction 2 for Toy Example 10

2.1 Toy data Set for Estimating Genetic and Environmental variation 13

2.2 True Genetic Curves of Toy Data Set 14

2.3 True Environmental Curves of Toy Data Set 15

2.4 Center Curves of Toy Data Set 17

2.5 Side by Side PCA of $\tilde{P}$ and P 18

2.6 Estimated Group Mean Curves of Toy Data Set 21

2.7 Side by Side PCA of $\tilde{G}$ and G 22

2.8 Estimated Individual Curves of Toy Data Set 23

2.9 Side by Side PCA of $\tilde{E}$ and E 24

2.10 Side by Side PCA of $\hat{G}$ and G 26

3.1 Measure of Simplicity in Object Space 33

3.2 Uneven Environment Levels Toy Example 41

3.3 Simple curve basis for full space 44

3.4 PCA basis 46

3.5 Simple curve basis when null space is 2-d 48

3.6 Simple curve basis when null space is 1-d 50

3.7 Simple curve basis when null space is 3-d 52

4.1 Manduca sextas' Growth Trajectories 54
4.2 Toy Data For Grid and Landmark Representation 58
4.3 PCA of Grid Based Representation 60
4.4 PCA scores scatterplot 63
4.5 Parallel Coordinates View of Landmark Representation 65
4.6 Landmark PCA on toy data set 68
4.7 Biological Correspondence of Landmarks 72
4.8 Added pseudo-landmarks for All Curves 73
4.9 PCA on Original Scale 76
4.10 PCA on Correlation Matrix 79
4.11 PCA of Trace Standardized Data 81

5.1 Manduca sextas' Growth Trajectories up to Peak 84
5.2 Manduca sextas' Growth Trajectories Highlighted Line Segments 86
5.3 Visualization of Multiple T-test Comparisons 87
5.4 Highlighted Visualization of Multiple T-test Comparisons 90
5.5 Manduca sextas' Growth Trajectories for 20 and 25 94
5.6 Manduca sextas' Growth Trajectories combined data 95
5.7 Visualization of Multiple T-test Comparisons Combined Cyan Curves . . 97
5.8 Visualization of Multiple T-test Comparisons Combined Data 99

6.1 Geometry of Canonical Angles Between 1-d Subspaces 105
6.2 Geometry of Canonical Angles Between 2-d Subspaces 108
6.3 Geometry of Canonical Angles Between a 1-d Subspace and a 2-d subspace 110
6.4 Toy Data for CCA as CA 112
6.5 Dual Space Representation 118
6.6 A_{XY} in Primal Space of X and Y 120
6.7 Mapping of Subspaces' by Cosine of Angles 123

CHAPTER 1

Introduction

Evolutionary biologists study changes in populations from one generation to the next. A common way to study populations is through *phenotypes* of individuals. A phenotype is an observable characteristic of an individual, see Lynch and Walsh (1998) for a more detailed discussion of quantitative genetics. Examples of a phenotype are growth rate of a caterpillar or height of a plant. A common practice is to view the phenotypic value of an individual over several environment levels. These environment levels could be different temperatures or densities of plants.

Those traits with continuous phenotypic values with respect to the different environment levels, are called *function valued traits* (FVT), see Kingsolver *et al.* (2001) for an introduction to function valued traits. Examples of FVT include *Pieris rapae* caterpillars growth rate as a function of temperature, see Kingsolver *et al.* (2004), and mass as a function of age for Manduca sexta hornworms, see Gilbert *et al.* (2000), Nijhout (1994), Riddiford *et al.* (2003). The phenotypic values with respect to different environment levels can now be thought of as functions, i.e. curves.

Statistical analysis on curves is an area in statistics known as Functional Data Analysis (FDA). FDA is based on the curve being the statistical atom of analysis, i.e. each individual is associated with a curve. In the the area of FDA, statistical analyses have results based on the curves of the individuals. For this case each individual is associated with a FVT and the statistical analyses treat the FVT as the statistical atom. For a

more detailed discussion of FDA, see Ramsay and Silverman (2005).

A common approach to Functional Data Analysis is to discretize the curves. Statistical analysis can be performed on vectors that contain the discretized values, see Ramsay and Silverman (2002). The discretization, i.e. representation, of the curves can determine which statistical method is most efficient at answering a biological question, see Sections 4.1 and 4.2. If the curve is discretized into d values, then the statistical analysis is performed on points in the Euclidean space $\mathbb{R}^d$. Although the analysis is done in $\mathbb{R}^d$, i.e. the *point cloud space*, the results are often more easily interpretable if they are shown as curves, i.e. in the *object space*. A toy example in 2-d space may help in the understanding of the point cloud and object space.

1.1 Introduction to Point Cloud and Object Space Using A Toy Example

A toy example is provided to show a representation of the curves, as well as demonstrating the important concepts of the object space and point cloud space. The left hand side of Figure 1.1 shows the data in the object space. The curves, $f_i(E)$ $i = 1, 2, ..., n$, are FVT, but each curve can be summarized by two numbers because of the curves' special simple structure.

Each curve has the same value for each environment level until it reaches the environment level denoted e_1. Then the FVT changes linearly until the environment level named e_2 is reached. Once environment level e_2 is reached each curve has the same attribute value for the remaining environment levels. All of the curves follow the pattern of only changing between environment levels e_1 and e_2. The curve, $f_i(E)$, can be summarized by the trait value of the curve at environmental level e_1, i.e. $f_i(e_1) = a_i$, and the trait vale at environmental level e_2, i.e. $f_i(e_2) = b_i$. If the curves are discretized using the trait values at e_1 and e_2, then these curves can always be perfectly recovered in a piecewise linear fashion. The curve $f_i(E)$ is recovered by plotting the trait value of the curve equal

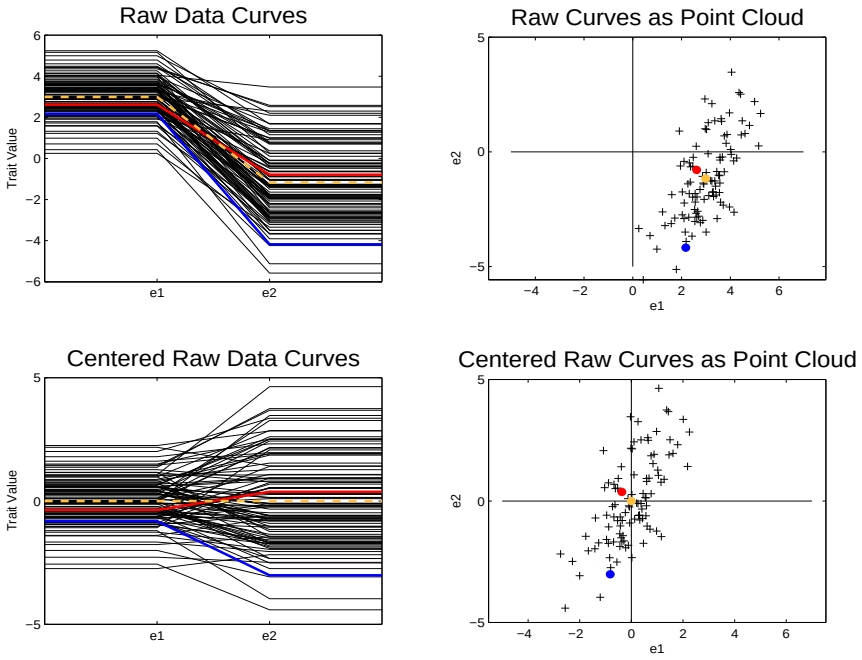

Figure 1.1: *Object view is on the left while point cloud view is on the right. Each curve can be represented by a point in the point cloud space and each point can be represented as a curve in the object space.*

to a_i from the first environmental point to e_1, then the curve is plotted as a piecewise linear line with trait values from a_i to b_i in the region from e_1 to e_2, and finally the curve is plotted as having trait value b_i from e_2 to the last environmental point. Since e_1 and e_2 are common across curves, it follows that to compare curves, $f_i(E)$ $i = 1, ..., n$, only the pairs $\{(a_i, b_i), ..., (a_n, b_n)\}$ need to be analyzed.

Now that the curves have been discretized and summarized by two trait values, each curve can be represented as a point in $\mathbb{R}^2$. The curves represented in $\mathbb{R}^2$, i.e. the point cloud space, are shown on the right hand side of Figure 1.1. To show exactly how a curve is represented as a point in $\mathbb{R}^2$, a curve is highlighted in blue in the upper left hand panel. The blue curve has a trait value of around 2 at e_1 and a trait value of around -4 at e_2. The corresponding point is highlighted in blue in the upper right panel, i.e. point cloud

view, as well. This point has a value of 2 along the horizontal axis, i.e. e_1, and a value of -4 along the vertical axis, i.e. e_2. Each curve can be represented in the point cloud view in this same manner.

Also in the point cloud view is a yellow point. This yellow point is the arithmetic mean of the horizontal axis and vertical axis, i.e. $(\overline{a},\overline{b}) = (\frac{1}{n}\sum_i a_i, \frac{1}{n}\sum_i b_i)$. The yellow dot is the arithmetic mean of the points in $\mathbb{R}^2$. But this point can also be shown as a corresponding curve in the object space. The yellow curve is this corresponding mean curve. Although this curve did not exist in the original data set, it can still be viewed in the object space by the correspondence between the point cloud view and the object space. So not only can any curve be represented by a point, but any point can be represented by a curve. Usually calculations are best understood in the point cloud space, however deeper understanding of the results can be gained by viewing the corresponding curves in the object space.

The lower portion of the figure shows the data in the object space and point cloud space with the mean removed. The yellow dot is now at (0,0) and the yellow curve is now a straight line at trait value zero. The curves and points now represent how the data differs from the mean. After removing the mean it is natural to focus the analysis on variation. Variation can be summarized by a covariance matrix. The total variation of all of the data is known in evolutionary biology as the *phenotypic variance*. A useful decomposition of the variance, in terms of evolutionary biology, is into genetic variance and environmental variance, see Chapters 2 and 3. This is a useful decomposition because the genetic variation determines the evolutionary response, an important evolutionary biological concept. Evolutionary response, $\Delta\overline{Z}$, is the change in mean phenotype from one generation to the next.

A good way to understand variation is to find an orthonormal basis in the point cloud space. This is the same as finding a set of orthogonal linear directions which can represent all of the variation of the data points. In the point cloud space a linear direction can be

thought of as a line that passes through the origin. Two directions that are orthogonal will be at right angles with respect to each other, i.e. the Euclidean inner product is 0. Once this orthonormal basis is defined, the data can be projected onto the directions of this basis to see how much variation each direction explains. These projections can also be viewed in the object space to understand the variation in that direction.

One such othonormal basis is the identity matrix of size d. In our toy example a 2×2 identity matrix is an orthonormal basis for the data. That would correspond to the horizontal and vertical axis passing through the origin. In Figure 1.1 these are the horizontal and vertical black lines in the point cloud view. The projections onto these directions would simply be the actual trait values at the corresponding environment levels. The identity matrix is one orthonormal basis used to understand variation of the data, but there are infinitely many orthonormal bases that could be used to understand the data. The orthonormal basis used to understand variation is often selected such that it optimizes some criterion.

One useful basis is that found by Principal Component Analysis (PCA). This basis is such that the first direction is the one that explains the most variation. The next direction is orthogonal to the first and explains the next most amount of variation, etc. The PCA basis is such that it explains the most variation in the least number of directions, and hence has been invaluable in many fields, including signal compression. See Section 4.1 as well as Anderson (1984), Muirhead (1982), and Ramsay and Silverman (2005) for more details about PCA.

The results of a PCA performed on the 2-d toy data set is shown in Figures 1.2 and 1.3. First the focus is on Figure 1.2, which displays the first direction of the PCA basis. The upper left panel of Figure 1.2 shows the centered data curves in the object space, the same as in the lower left panel of Figure 1.1. The panel in the upper right corner is the point cloud view of the centered data. This panel is the same as the panel in the bottom right of Figure 1.1, but now there is a green line that passes through the origin.

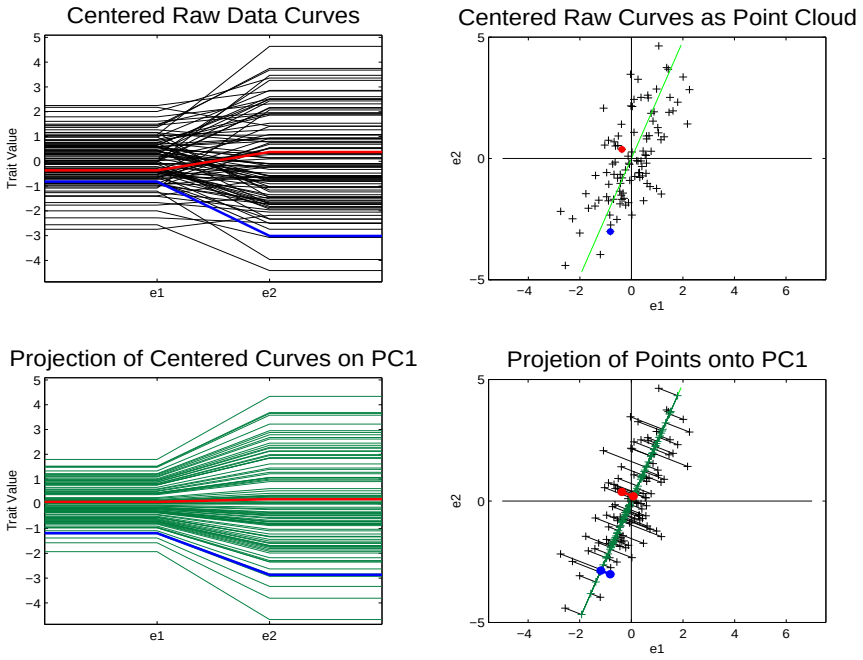

Figure 1.2: *Object view is on the left while point cloud view is on the right. Green line in point cloud space is the PC 1 direction. Lower right shows projected data onto the PC 1 direction. Projected points are along a line and quite spread out. Projected curves, in lower left panel, are multiples of each other and represent, the data quite well.*

This green line is the first PC direction. This line is equivalent in the point cloud view to the direction where the points are most spread out, which can be seen in the figure. The lower right panel shows the projection of the points onto the PC 1 direction. The projection of a point onto a direction corresponds to finding the point along the line that is closest to the original data point. This is equivalent in the 2-d case to drawing a perpendicular line from the data point to the PC 1 direction. Each of the curves has a projection point associated with it. Although these are not points from the original data set they can still be visualized in the object space. In the lower left hand corner are the projected curves associated with the PC 1 direction. The projected points fall along a line in the point cloud space, which corresponds to the lines being multiples of each other in the object space, i.e. they are a one dimensional approximation of the original data.

Notice that the curves in the lower left hand corner look similar to the curves in the upper left corner. Also notice that the perpendicular lines that connect the projected data points to the actual data points are small. These perpendicular lines are the residuals, i.e. each data point minus its corresponding projected point.

Figure 1.3 shows the PC 2 direction. Again the upper left panel shows the centered data curves. The red line in the upper right panel is the PC 2 direction. This direction is at a right angle to the PC 1 direction. The PC 2 direction is thus orthogonal, which implies that it explains a different mode of variation. In the lower right corner is the projection of the data points onto the PC 2 direction. The projected points are much

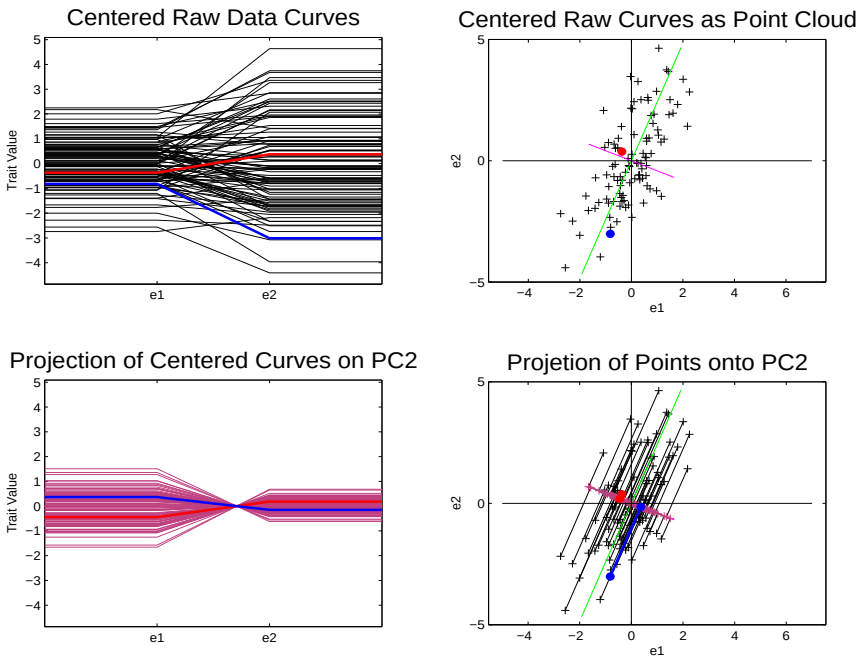

Figure 1.3: *Object view is on the left while point cloud view is on the right. Red line in point cloud space is the PC 2 direction. Lower right shows projected data onto the PC 2 direction. Projected points are along a line and not spread out. Project curves, in lower left panel, are multiples of each other and do poor job of representing the data.*

closer together for this direction. That means that this direction explains less variation than the PC 1 direction. The curves shown in the lower left panel are the projected

curves associated with the PC 2 direction. The curves do a worse job representing the actual data curves than PC 1, i.e. the curves look less similar to the curves in the upper left. The residuals are also larger for the PC 2 direction showing again that the PC 2 direction does a worse job of representing the data in only one dimension than does the PC 1 direction. Since this is only 2 dimensional data the PC 1 direction and the PC 2 direction explain all of the variation of the data points.

Also, because PC 1 and 2 are orthogonal and explain all of the variation, if the projections of PC1 and 2 are added together they will yield the actual data points. This can be seen in the point cloud view in that the residuals for PC 1 are the same as the projected points for PC 2. Also this can be seen in the object space, by the fact that if the curves are added together they will yield the actual data curves.

The principal component basis provides directions with an evolutionary biological interpretation. This interpretation is that the first PC direction calculated from genetic variation is the direction that produces the most evolutionary response when selected upon. Selecting upon a direction means to choose individuals with high absolute projection scores in that direction. Then each following principal component direction can be thought of as producing less and less evolutionary response when selecting in that direction. The lower principal components will eventually produce so little evolutionary response when selected upon, that they can be considered genetic constraints. For a more detailed discussion of genetic constraints see Section 3.1.

Other criteria, besides variation of the data can be used to construct an orthonormal basis. For example, the selection process of the basis can ignore the variation of the data completely and try to find directions which yield the smoothest possible projected curves in the object space. One measure of *smoothness* can be defined in terms of minimizing squared differences of adjacent points along the curves, standardized to have length 1. I.e. one seeks the curve that has trait values changing the least between adjacent environment levels. Figure 1.4 shows the results of this *smooth basis*.

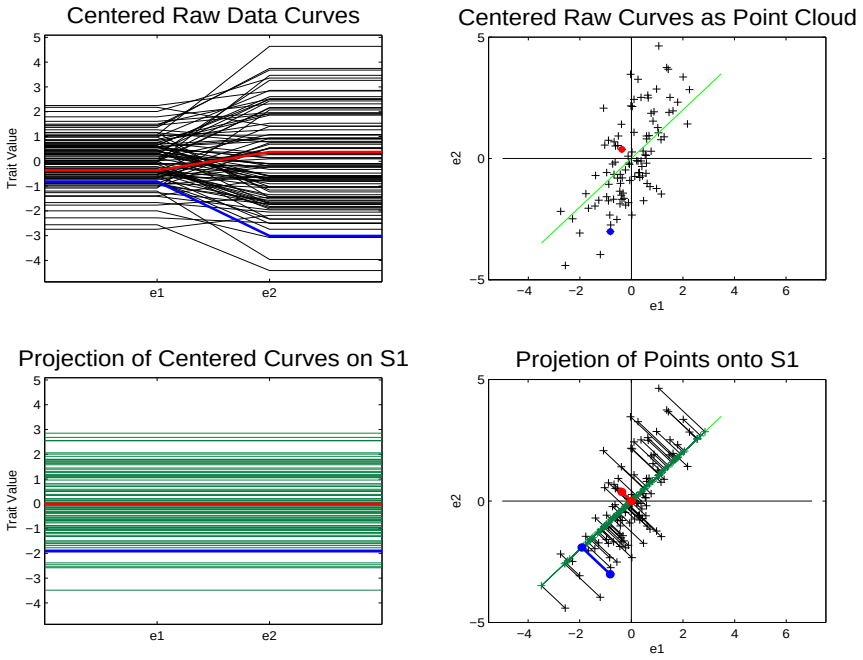

Figure 1.4: *Object view is on the left while point cloud view is on the right. Green line in point cloud space is Smooth basis direction 1. Lower right shows proejcted data onto smooth direction 1. Projected points are along a line. Project curves, in lower left panel, are multiples of each other and never cross zero.*

Figure 1.4 shows the first direction of the smooth basis, i.e. the direction that produces the smoothest projected curves. The line in the upper right panel is smooth direction 1. The lower right panel shows the projection of the data onto smooth direction 1 and the lower left panel shows the projected curves. Each projected curve has the same trait value for e_1 and e_2. Therefore the trait values of these curves have changed the least amount as possible between environment levels e_1 and e_2. Notice that the direction explains less variation than PC 1 and has larger residuals. This can be seen in the object space in that the curves do a worse job of representing the actual data curves. But smooth direction 1 explains more variation than PC 2, i.e. the projected curves do a better job of representing the actual data curves.

Figure 1.5 shows smooth direction 2. Notice that smooth direction 2 is orthogonal

to smooth direction 1. The lower right panel shows the projection of the data points onto smooth direction 2. Notice that this direction explains less variation than the PC

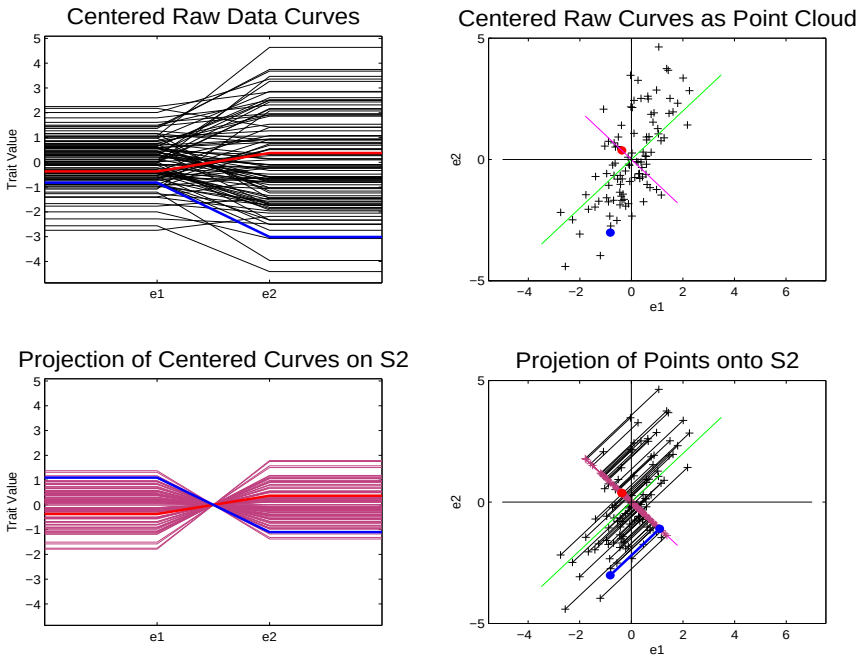

Figure 1.5: *Object view is on the left while point cloud view is on the right. Red line in point cloud space is smooth basis direction 2. Lower right shows projected data onto smooth direction 2. Projected points are along a line. Project curves, in lower left panel, are multiples of each other and cross zero once. These are the least smooth curves.*

1 direction, but more than the PC 2 direction. The residuals are also larger than PC 1 but smaller than the PC 2. This direction's projected curves are shown in the lower left panel. Each of these projected curves has $f_i(e_1) = -f_i(e_2)$, i.e. the curves have trait values changing the most between e_1 and e_2.

These are only two choices for orthonormal bases used to understand the variation of the data. One could always look at a compromise between smoothness and amount of variation explained, see Section 3.1. Also one could find a basis that optimizes a completely different criterion.

In evolutionary biology a key goal is to find a basis that separates genetic variation

from environmental variation. If it is possible to separate these different types of variation into orthogonal subspaces, then each subspace can have a basis which maximizes different criteria. Then the bases of these two subspaces can be viewed together to understand the variation as a whole. This separation of genetic and environmental variation is not always possible. So a simpler goal is to find a basis that explains the genetic variation only. Then define the directions orthogonal to this basis as genetic constraints, and find a basis to understand this genetic constraint space.

In chapter 2 of this dissertation will be a more detailed discussion of the estimation of genetic and environmental variation. Chapter 3 covers genetic constraints and the selection of a basis to view genetic variation. The basis chosen to view genetic variation will be a compromise between the PCA basis and the basis which is chosen based on smoothness of curves. Chapter 4 discusses curve representation in the context of principal components. An analysis of the slope structure of the data set introduced in Chapter 4 is presented in Chapter 5. Chapter 6 builds some background mathematical material used in Chapter 7. Some asymptotic properties of methods introduced in Chapter 3 are stated and proven in Chapter 7.

CHAPTER 2

Estimation of G and E

An important aspect of evolutionary biology is to model genetic and environmental variation. Generally individuals in an evolutionary biological study are grouped by genetic similarity. Genetic variation is then the variation between these groups while the environmental variation is the variation of individuals within these groups. A straightforward potential approach to estimating genetic variation is to use the sample covariance matrix based on sample group mean curves. This approach is heavily dependent on the fact that the true mean of the individual curves in each group is 0. Often these groups have a small number of individuals in each group, so the sample mean of the individual curves is not 0. This leads to environmental variation being misclassified as genetic variation. The procedure of estimating genetic variation via group means can be viewed as fixed effects Analysis of Variance (ANOVA). Random effects ANOVA shrinks the misclassification of environmental variation compared to fixed effects ANOVA. Random effects ANOVA does this by accounting for the inaccuracy of sample mean estimates to true means, due to the small number of individuals in each group. For a more detail discussion of the definition and differences between fixed and random effects ANOVA, see Searle *et al.* (1992).

In this chapter a toy example will be used to explore the difference between fixed effects ANOVA and random effects ANOVA. In Section 2.1 is a description of the simulated data set. In Section 2.2 the results of a fixed effects ANOVA on this toy data

set are shown where the misclassification is evident. The results of the Random effects ANOVA, which lessen this misclassification, are shown in Section 2.3.

2.1 Simulated Toy Data Set

A simulated toy data set is presented in this section, which is used to display the differences between the fixed and random effects ANOVA methods. The random effects ANOVA lessens the misclassification of environmental variation as genetic variation.

The toy data set is simulated to have 2500 curves of 11 dimensions, see Figure 2.1. Each curve is the sum of a random group, i.e. genetic, curve and a random individual, i.e. environmental, curve. Each group and individual curve has a normally distributed random error, i.e. noise curve, added to them. Our data consists of 500 groups with 5 individuals in each group. In Figure 2.1 the first 100 curves are shown. Only these

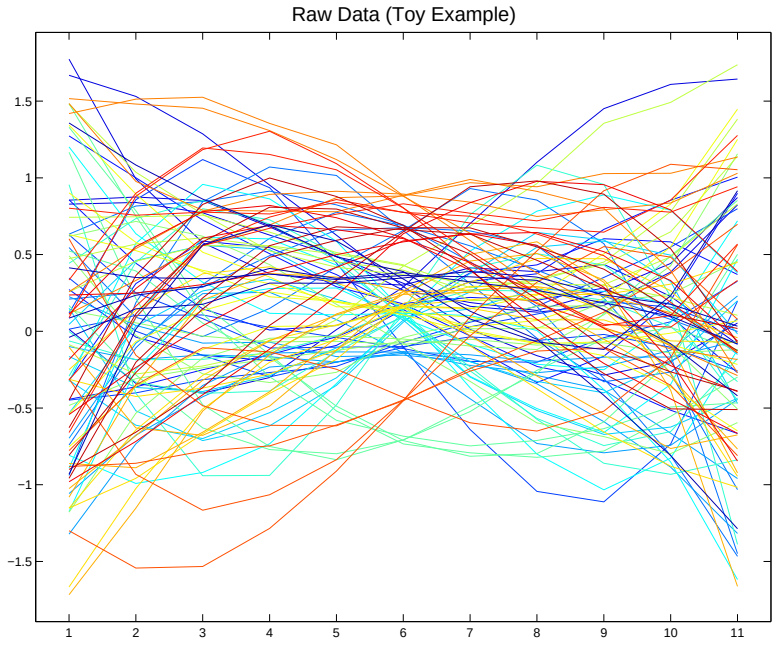

Figure 2.1: *Displayed are the first 100 data curves in the toy data set. Individuals in the same group have the same color curve.*

100 curves are shown to avoid over-plotting. In future plots of the toy data set only 100 curves will be included for the same reason.

Each member of the same group has the same group curve, i.e. genetic curve. The group curves are generated as the sum of normally distributed multiples of flat lines and parabolic curves, see Figure 2.2. Thus each curve is a random parabola with a random vertical shift. The curves appear in tight groups of 5 of the same color. This is because

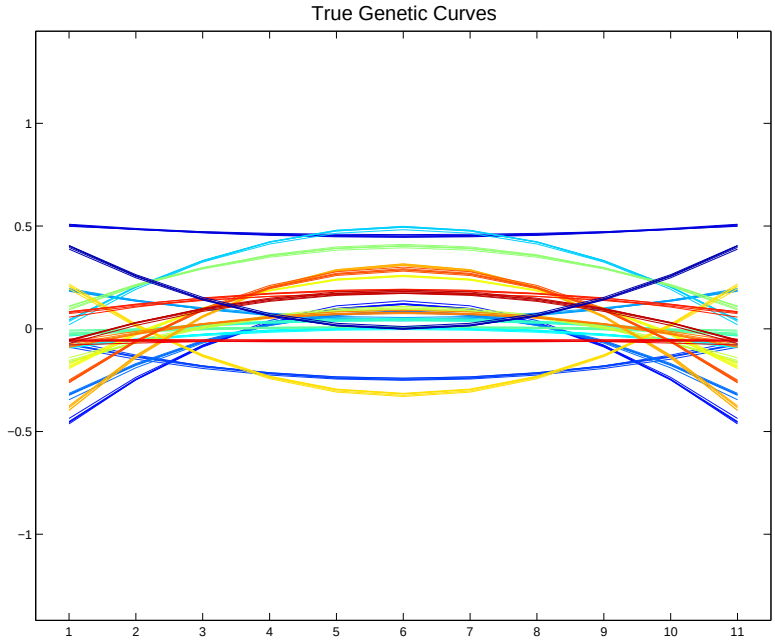

Figure 2.2: *The group curves of the first 100 individuals, as shown in Figure 2.1 using the same colors. Notice that the curves are in groups of five which correspond to the five individuals in each group. The curves are random parabolas with a random vertical shift.*

each group has 5 individuals and each individual from the same group has the same genetic curve, except for some small normally distributed random noise.

Each individual has its own distinct individual curve, i.e. members of the same group have differing individual curves but the same group curves. The individual curves are the sum of normally distributed multiples of linear and cubic curves, see Figure 2.3. The

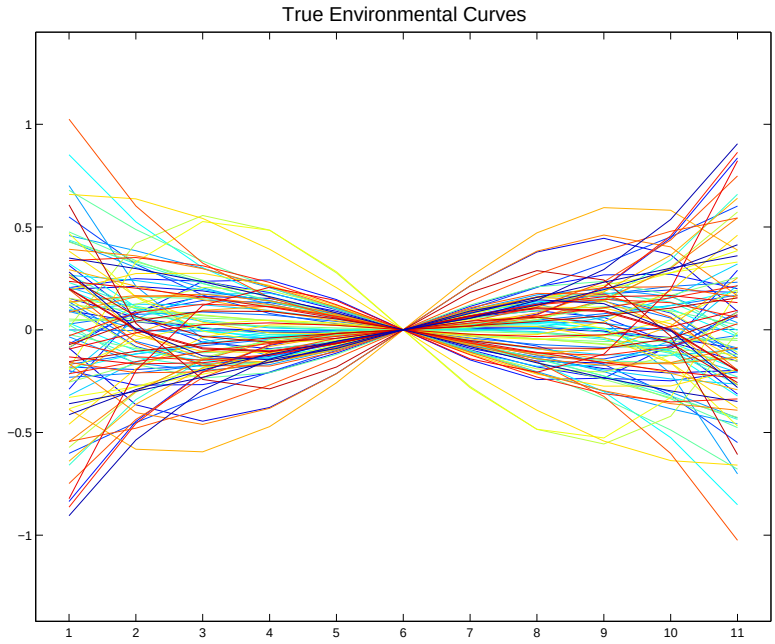

Figure 2.3: *The individual curves of the first 100 individuals, as shown in Figure 2.1 using the same colors. Notice that the curves are made up of a random combination of linear and cubic terms. The curves are no longer in tight groups of 5 of the same color.*

environmental curves are not in tight groups of 5 because each individual has its own distinct individual curve.

These components have been carefully chosen to be orthogonal, and all explain differing amounts of the total variation.

The phenotypic variation is reflected by all of these curves. The linear component explains 36.8% of the phenotypic variation. While the cubic component explains 13.2% of the phenotypic variation. The flat line component explains 36.8% of the phenotypic variation and the parabolic component explains 13.2% of the variation.

The genetic variation is reflected by only the group curves, i.e. flat line and parabolic modes of variation. The genetic variation is 50% (36.8% + 13.2%) of the total phenotypic variation. The flat line mode explains 73.5% of the genetic variation and the parabolic

mode explains 26.5% of the genetic variation.

The environmental variation is reflected by only the individual curves, i.e. the linear and cubic modes of variation. The environmental variation is 50% (36.8% + 13.2%) of the total phenotypic variation. The linear mode accounts for 73.5% of the environmental variation and the cubic mode accounts for 26.5% of the environmental variation.

The raw data curves, which are a mixture of genetic and environmental curves, are the only curves which will be explicitly viewed. We want to find a procedure which will partition the phenotypic variance into the genetic variance and environmental variance, i.e we hope to recover the curves from Figure 2.2 and Figure 2.3.

2.2 Fixed Effects ANOVA on the Toy Data Set

2.2.1 Estimation of Phenotypic Variance

This section focuses on how to use fixed effects ANOVA to estimate the phenotypic variation from the observed curves. The P matrix is a summary of phenotypic variation, so the above goal is the same as estimating the P matrix. The first step in estimating the P matrix, in the case of the fixed effects approach, is to subtract the mean for all 11 dimensions of the curves, see Figure 2.4. The centered curves are arranged into an 11 by 2500 matrix, $X_c = (X - \overline{X})$. Outer product multiplication, $X_c X_c^T$, is performed on the centered curve matrix and each entry of the matrix produced is divided by $(n-1)$. These operations produce an 11×11 empirical covariance matrix,

$$\tilde{P} = \frac{(X - \overline{X})(X - \overline{X})^T}{n-1}$$

which is the estimated P matrix.

Now that P has been estimated by $\tilde{P}$, the next question is how accurate the estimate is. A good way to visualize the variation summarized by matrices is through PCA. To compare the two matrices, P and $\tilde{P}$, we will look at side by side PCAs, see Figure 2.5.

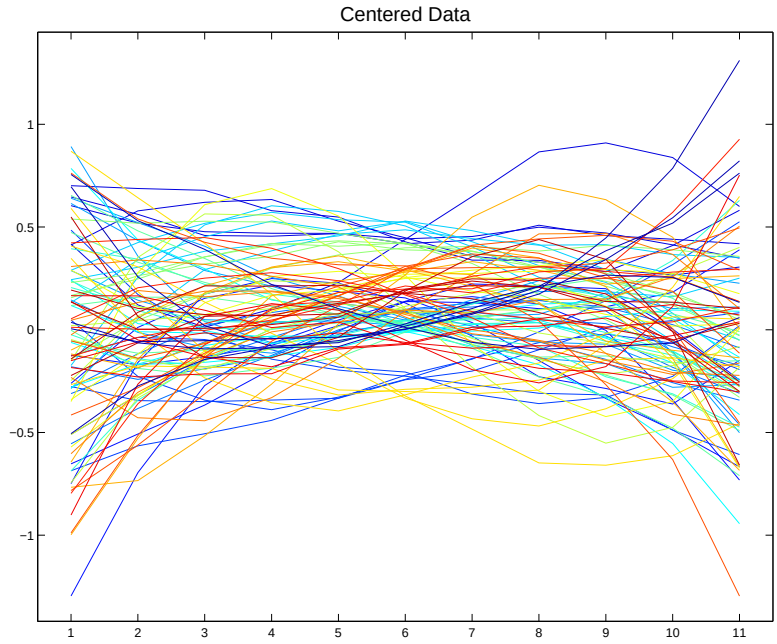

Figure 2.4: *The first 100 centered data curves with the mean for each point on the x-axis subtracted. Individuals in the same group have the same color curve. One color corresponds to more than one group.*

In Figure 2.5 the first column shows the first six PC directions of $\tilde{P}$, viewed as the corresponding unit length curves in the object space. The second column shows the first six principal component directions of the theoretical P matrix for this simulation if there were no errors, viewed as the corresponding unit length curves in the object space. The thick blue line and number at the top of each panel represents the percentage of variation of the matrix explained by each curve, i.e. each direction. The number in parentheses represents the sum of squares of the data that the direction explains. The similarity of the two columns indicates the accuracy of $\tilde{P}$ as an estimate of P.

The first panel of the first column is the PC 1 direction of $\tilde{P}$, which explains 38.6% of the variation. The PC 1 direction is a linear curve, which corresponds to an environmental mode of variation. In the second panel of the first column is the PC 2 direction of $\tilde{P}$,

which explains 34.8% of the variation. This is a flat line which corresponds to a genetic mode of variation. The line is not exactly flat because there is some mixing with the

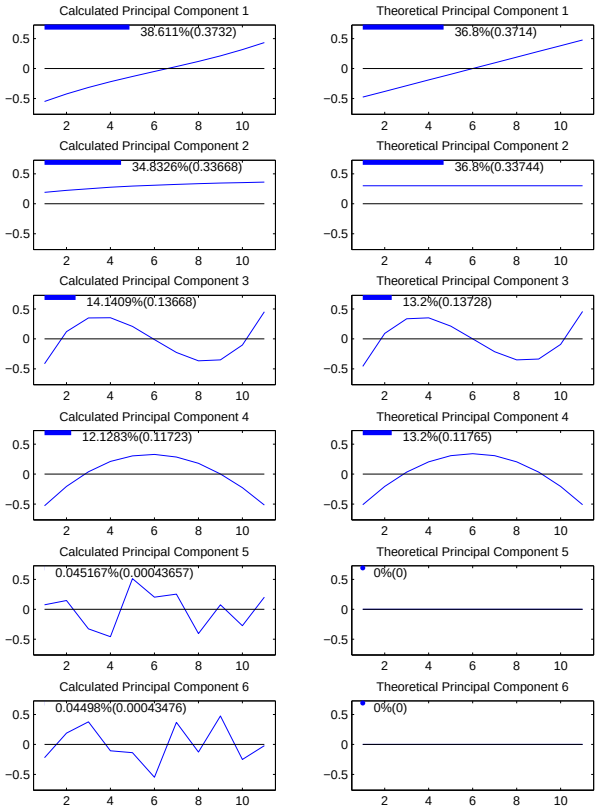

Figure 2.5: *Left hand side panels are empirical principal components of $\tilde{P}$. Right hand side panels are theoretical principal components of P. Bars and numbers at top of panels represent amount of variation of P matrix explained. The columns are similar indicating $\tilde{P}$ is a good estimate of P.*

linear mode of variation, due to the fact that they explain a similar amount of phenotypic variation. This is a weakness of PCA. If the amount of variation explained by modes of variation is similar, then the PCA results in a linear mixing of these modes of variation. The third panel is a cubic curve which is an environmental mode of variation while the fourth column is a parabolic curve which is a genetic mode of variation.

It can be seen from this figure that P is a mixture of genetic modes of variation, summarized by the matrix G, and environmental modes of variation, summarized by the matrix E. Recall that the first and third principal component directions explain variation summarized by E, i.e environmental variation, and the second and fourth principal component directions explain variation summarized by G, i.e. genetic variation. For both columns the shape of the curves and amount of variation explained are similar.

The first column appears to differ from the second column in the last two panels. For the theoretical case, i.e. column 2, the curves are flat lines at zero. This is because the P matrix is of rank 4, so the first 4 PC directions explain all of the variation of P. For the empirical case, i.e. column 1, the last two panels are random error directions. Notice that the amount of variation explained by these directions is nearly zero. Similarly the principal component directions 7-11 are just directions of random error and explain nearly zero variation. These directions are not included in the figure. Because the left hand column is for the theoretic non-error case, the columns are similar except for the error components.

This figure indicates that the estimate of the P matrix is quite good. This can be seen from the fact that the two columns are so similar. The P matrix was estimated the same way for both fixed and random effects ANOVA, so the principal component decomposition looks the same for both methods. Therefore this figure will only appear in this section of the chapter.

2.2.2 Estimation of Genetic Variance

The estimation of genetic variation using fixed effects ANOVA is presented in this section. Also the results of using this approach to estimate the genetic variation of the toy data set will be shown. Similar to the phenotypic variation case, there is a matrix G which summarizes the genetic variation. Our goal is to estimate this G matrix accurately.

The fixed effects ANOVA estimate of G depends on sample group means being used

as estimates of the true genetic curves. Often these estimates are not accurate, due to the small number of individuals in each group. The empirical covariance matrix, $\tilde{G}$, summarizing the variation of these curves is the fixed effects ANOVA estimate of G. If $\overline{X}_g$ is an $11 \times n_g$ matrix of the sample group means and n_g, i.e. 500 for this toy example, is the number of groups then

$$\tilde{G} = \frac{(\overline{X}_g - \overline{X})(\overline{X}_g - \overline{X})^T}{n_g - 1}$$

is the fixed effects ANOVA estimate of G.

The sample group mean curves for the toy example are shown in Figure 2.6. The sample group mean curves should look like the true genetic curves, shown in Figure 2.2, since they are estimates of the true genetic curves. But the sample group mean curves also include some linear and cubic modes of variation, i.e. environmental modes, along with vertically shifted parabolas, i.e. genetic modes. Also they do not appear to be in groups of 5, because each group member is estimated to have the same group mean curve. Therefore the curves are overlayed directly on top of each other. The sample group mean curves are not accurate estimates of the true genetic curves, which leads to an inaccurate estimate of G.

To view the inaccuracy of $\tilde{G}$ as an estimate of G, side by side PCAs are shown in Figure 2.7. The first column of Figure 2.7 shows the first four PC directions of $\tilde{G}$, viewed as the corresponding unit length curves in the object space. The second column shows the PC directions of the theoretical G matrix, viewed as the corresponding unit length curves in the object space.

Fixed effects ANOVA does not do a good job of estimating G, which can be seen by Column 1 and Column 2 being different. For this method the linear curve, appearing in row 3 of column 1, and cubic curve, appearing in row 4 of column 1, are showing up as PC directions of $\tilde{G}$, when these curves are actually directions of environmental variation.

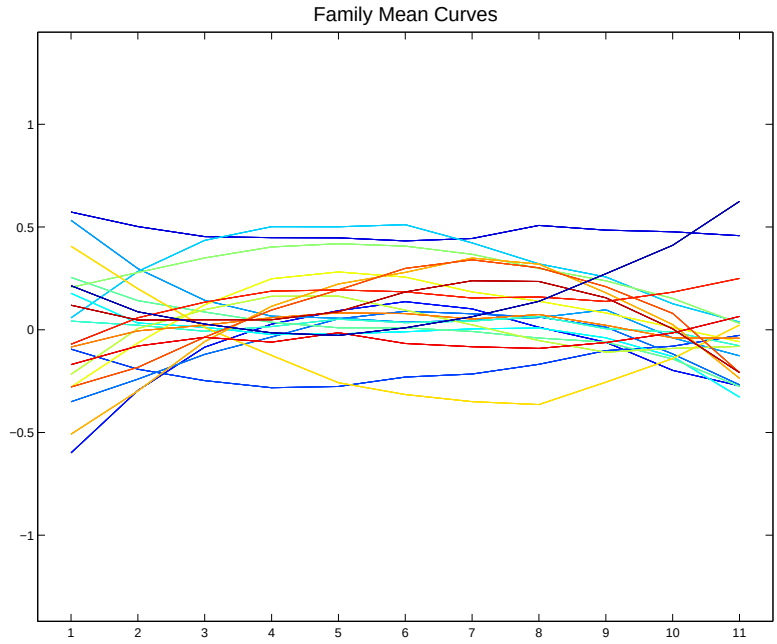

Figure 2.6: *The first 100 group mean curves. Notice that the cubic and linear part can still be seen, which leads to incorrect estimates of G. These curves do not look like the group curves from Figure 2.2*

These two modes of variation explain about 18% of the genetic variation, when they should theoretically explain almost 0% of the variation. Indicated by the fact that PC directions 3 and 4 are flat lines of height 0 in column 2 since all the variation is explained by the first two principal components in the non-error theoretical case. Fixed effects ANOVA produces estimates of G with more variation than just genetic variation. This over estimation of genetic variation causes the percentage of variation of the flat line and parabolic modes of variation to be less than in the theoretic case.

Again for this figure the last several PC directions are not pictured since they are random error directions that explain nearly zero amount of the variation, and are therefore accurate estimates of the true genetic variation.

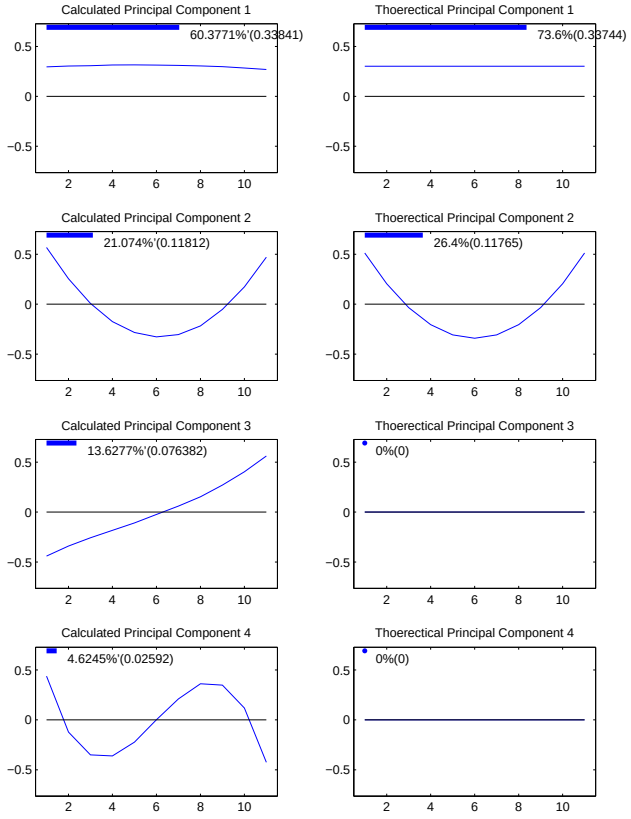

Figure 2.7: *Left hand side panels are empirical principal components of $\tilde{G}$. Right hand side panels are theoretical principal components of G. Bars and numbers at top of panels represent amount of variation of G matrix explained. Notice that environmental modes(linear and cubic)of variation are being classified as genetic variation.*

2.2.3 Estimation of Environmental Variance

This section describes the estimation of environmental variation using fixed effects ANOVA. This is done by finding an estimate of the E matrix. The fixed effects ANOVA estimate of E involves the same sample group means, calculated in the case of $\tilde{G}$. In this case the sample group means are subtracted from the original curves, which yield an estimate of the environmental curves. These estimated environmental curves can be

used to form an empirical environmental covariance matrix,

$$\tilde{E} = \frac{(X - \overline{X}_g)(X - \overline{X}_g)^T}{n_g(n_e - 1)}$$

which is an estimate of E. Where all of the notation is the same as used in earlier sections and n_e is the number of individuals in each group, i.e. 5 for this toy example.

For the toy data set these estimated environmental curves are shown in Figure 2.8. The estimated environmental curves are similar to the true environmental curves, shown in Figure 2.3. The estimated curves reflect the linear and cubic components but none of

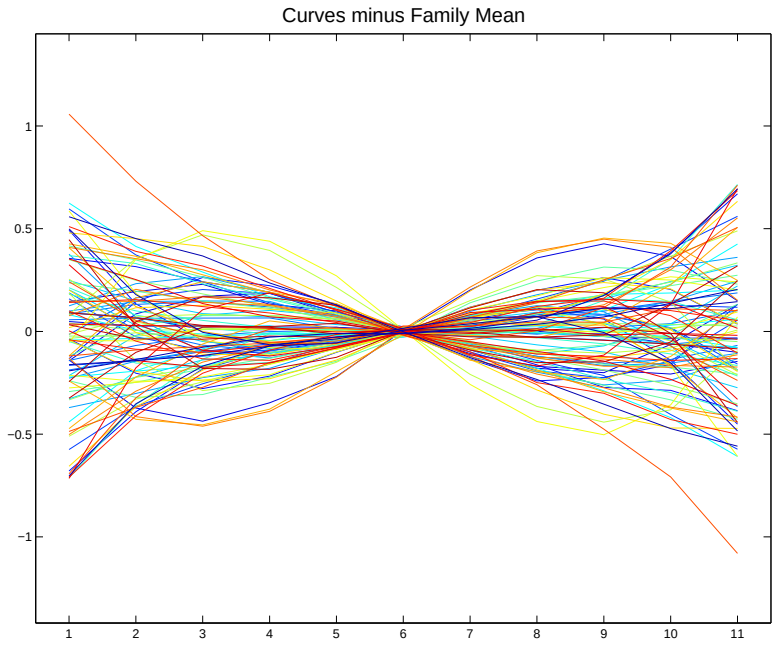

Figure 2.8: *The curves are the centered data with the group means subtracted out, the first 100 are shown. Notice that this figure is similar to the individual curves in Figure 2.3 so the curves will produce a good estimate of E.*

the flat line or parabolic components. Their accuracy yields a good estimate of E.

The accuracy of $\tilde{E}$ as an estimate of E is shown through side by side PCAs. In Figure 2.9 the panels of the first column are the PC directions of $\tilde{E}$, viewed as the corresponding

unit length curves in the object space. While the panels of the second column are the PC directions of the theoretical E, viewed as the corresponding unit length curves in the object space. In this figure the first and second columns look similar, which indicates $\tilde{E}$ is a good estimate. The first and second rows of column 1 are linear and cubic modes of

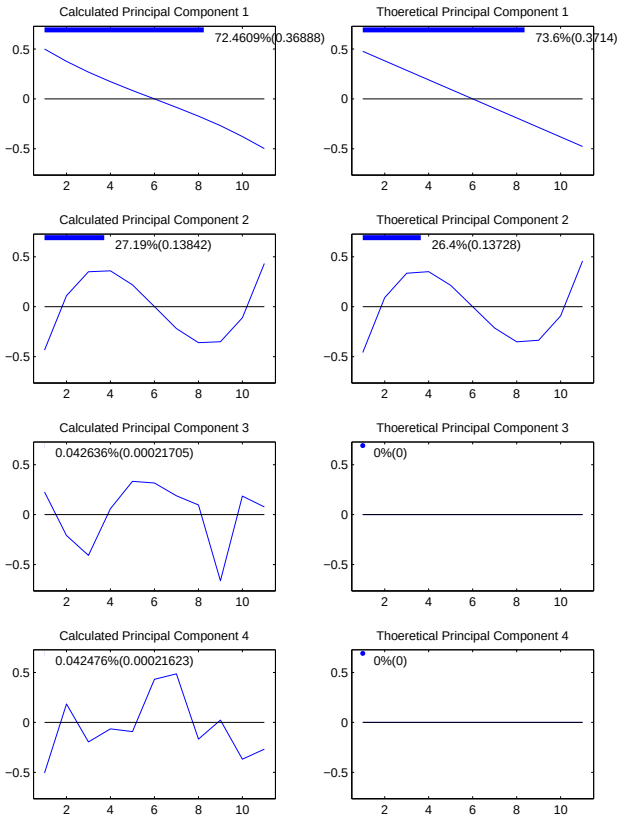

Figure 2.9: *Left hand side panels are empirical principal components of E estimated by group means. Right hand side panels are theoretical principal components of E. Bars and numbers at top of panels represent amount of variation of E matrix explained*

variation, which are the same modes of variation as the theoretical environmental modes. The third and fourth principal component directions do look different. This is just due to random error since the theoretical E is of rank two, i.e. has only two directions that explain all of the variation. The other PC directions were not included since they are

directions of random error and explain nearly no amount of the environmental variation.

2.3 Random Effects ANOVA on the Toy Data Set

Section 2.2 showed the fixed effects ANOVA method, while the details of the random effects ANOVA are given in this section. Only the estimation of G is discussed. The estimation of P and E is not covered, since both the P and E matrices are estimated in the same way as the fixed effects approach.

2.3.1 Estimation of Genetic Variance

In Section 2.2, it is shown that fixed effects ANOVA yields an estimate of the genetic covariance matrix that actually has environmental variation as well as genetic variation, see Figure 2.7. Random effects ANOVA is a method that finds an estimate of G, $\hat{G}$, which lessens the amount of environmental variation that is summarized by $\tilde{G}$, the fixed effects estimate of G. The case where all groups have the same number of individuals is described in this section to help in the understanding of how random effects ANOVA lessens the misclassification.

As can be seen from Figure 2.9, $\tilde{E}$ provides a good estimate of E. The matrix $\tilde{E}$ is an unbiased estimator of E, which provides theoretical evidence to accompany the empirical evidence. So random effect ANOVA tries to remove the environmental variation from $\tilde{G}$ by subtracting $c_{n_e} * \tilde{E}$, where c_{n_e} is a constant that depends on the number of individuals in each group.

The exact method of random effects ANOVA, when all groups have the same number of individuals, is to let

$$\hat{G} = \tilde{G} - (\frac{1}{n_e})\tilde{E}$$

be the estimate of G. This is because $(\frac{1}{n_e})\tilde{E}$ is the expected amount of environmental variation that is misclassified as genetic variation based on expected values of the covariance matrices estimated by fixed effects ANOVA. Notice that as $n_e \to \infty$, the random

effects estimate converges to the fixed effects estimate. So the random effects estimate is correcting for the bias of the sample group means due to a small sample size, and as the sample size becomes larger less of a correction is needed.

The accuracy of $\hat{G}$ as an estimate of G is shown through side by side PCAs in Figure 2.10. Shown in the first column of Figure 2.10 are the PC directions of $\hat{G}$, viewed as

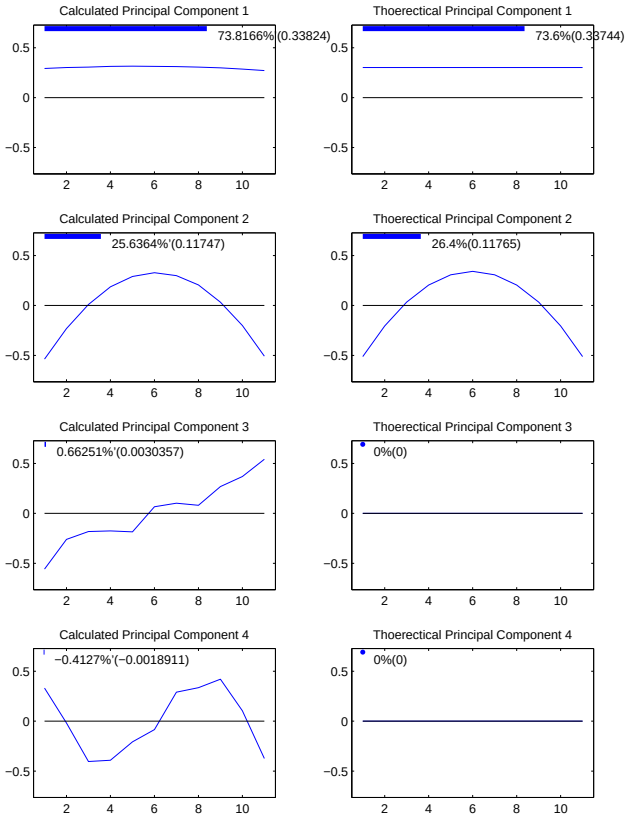

Figure 2.10: *Left hand side panels are empirical principal components of G estimated by random ANOVA. Right hand side panels are theoretical principal components of G. Bars and numbers at top of panels represent amount of variation of G matrix explained*

the corresponding unit length curves in the object space. In the second column is the PC directions of the theoretical G, viewed as the corresponding unit length curves in the object space.

Random effects ANOVA provides a much better estimate of G then fixed effects ANOVA, which can be seen by the fact that although there still seems to be some misclassification of environmental variation as genetic it is to a much lesser degree. Notice that the curves in rows 3 and 4 of column 1 are linear and cubic but the percentage of variation explained is less than 1%. But one problem that does arise from random effects ANOVA is that it can produce a covariance matrix that is not positive definite, even though covariance matrices should always be positive definite. This can be seen by the fact that the fourth direction is supposedly explaining a negative portion of genetic variation.

For the case when groups do not all have the same number of individuals the random effects estimates can found by Restricted Maximum Likelihood estimation (REML), see Searle *et al.* (1992). The REML method gets its name because it restricts the maximization to the part of the likelihood which is invariant to the mean parameters of the model. For the case when all groups have the same number of individuals the REML estimates are exactly the same as the random effects estimates described in this section. The REML estimates are unbiased but again can produce estimates that are not positive definite. Also the asymptotic distributions of the estimators can be difficult to calculate. A software package known as DFREML, see Meyer (1988), Meyer (1989), and Meyer (1998), has been used to successfully find REML estimates for many biological studies.

Also random effects estimates can be found by maximum likelihood (ML). For the case of the groups all having the same number of individuals the ML estimates are not always equal to the estimate described above. Also the estimates are not always unbiased, i.e. environmental variation can be misclassified. But the ML estimates are always positive definite and have well defined asymptotic distributions, see Searle *et al.* (1992).

CHAPTER 3

Finding Genetic Constraints: A Simple Curve Basis Of A Nearly Null Space

This chapter discusses a method to find genetic constraints of biological interest. This is done by not only measuring the amount of variation that a direction explains, but also measuring the simplicity of the corresponding curves in the object space. A genetic constraint of biological interest is a direction with low variation that is also associated with simple curves, i.e. a high *simplicity score*.

An introduction to the nearly null space and a basis based on simplicity of directions is provided in Section 3.1. The next section, Section 3.2, details how we will measure the simplicity of the curves associated with a direction, i.e calculate the direction's simplicity score. A way to derive the simple curve basis, mentioned in Section 3.1, using an eigendecompostion of a *simplicity matrix* is described in Section 3.3. Section 3.4 details how to modify this method when environment levels are not evenly spaced. How to use the idea of simplicity and amount of variation to determine if a basis yields genetic constraints of biological interest is described in Section 3.5. A way to visualize the simplicity score and amount of variation explained is shown in Section 3.6.

3.1 Introduction to the Simple Curve Basis of the Nearly Null Space

Variation of an observed characteristic of an individual is referred to as Phenotypic variation. The Phenotypic variation is modeled as lying in a vector subspace $S_P \subseteq \mathbb{R}^d$. The variation in S_P is summarized by a covariance matrix P. Useful insight about phenotypic variation comes from viewing it as a mixture of genetic variation and environmental variation. The genetic variation is modeled as lying in a subspace $S_G \subseteq S_P$. The variance in S_G is summarized by a covariance matrix G. Directions of S_P which explain little genetic variation, will also produce little genetic response when selected upon. These directions are considered to be *genetic constraints*.

A straight forward way to define genetic constraints is via the subspace, $S_N \subseteq S_P$, orthogonal to S_G. This orthogonal subspace S_N is defined to be the *nearly null space*. All genetic variation lies in S_G, so any direction in S_N explains no genetic variation because S_N is orthogonal to S_G. Therefore any direction in S_N is a genetic constraint.

The estimate of a basis of S_G is calculated from the estimated genetic covariance matrix $\hat{G}$, see Chapter 2 for details on estimating the genetic covariance matrix. Then the subspace $\hat{S}_G$ generated by this basis is the estimate of S_G. Once S_G is estimated, a nearly null space estimate, $\hat{S}_N$, can be found.

The basis of $\hat{S}_G$ should be as small as possible, i.e. having the least number of directions to explain genetic variation, in order to produce a rich orthogonal subspace of genetic constraints. A natural way to find the least number of directions that generates $\hat{S}_G$ is to perform Principal Component Analysis(PCA) on $\hat{G}$, as follows.

The numerical calculations that drive PCA of the genetic space $\hat{S}_G$ is the eigendecomposition of $\hat{G}$. The eigenvector corresponding to the the largest eigenvalue of $\hat{G}$ is the first PC direction. The eigenvector corresponding to the second largest eigenvalue is the second PC direction, etc. The first PC direction explains the most genetic variation. The

second PC direction explains the most genetic variation not explained by the first PC direction, in the sense that the directions are orthogonal. The eigenvectors are ordered in this manner to define the remaining PC directions. All eigenvectors being orthogonal is a property of the eigendecomposition, so all of the directions will explain different modes of genetic variation.

If PCA is performed on $\hat{G}$ then the first PC direction is viewed as the direction of greatest evolutionary response, i.e. least evolutionary resistance. As the PC directions explain less of the genetic variation they are viewed to have more evolutionary resistance until they can be considered genetic constraints. There are several ways to define the boundary between *response* and *constraint*. One way is to consider the set of lower PC directions, whose combined percentage of genetic variation explained is less than *constraint threshold* c_{prop}, to be the basis for $\hat{S}_N$.

The initial basis of the nearly null space, i.e. the lower PC directions, often provide directions that are hard to interpret. This is because the biological signal is weak, i.e. explains little variation, so the lower PCs are a mixture of the biological signal and random noise. As suggested by Nancy Heckman and Mark Kirkpatrick, deeper understanding of the nearly null space can be gained through another basis comprised of directions that are more interpretable.

Smooth orthogonal directions, i.e. simple curves, are often easily interpretable. A rotation of the initial basis to another orthonormal basis, that tries to find the simplest curves, yields an appealing opportunity to find insightful genetic constraints. Where the simplicity of a curve is measured by the squared vertical difference of given adjacent points along a curve in the object space. The number of simple orthogonal curves will be equal to the number of PC directions in the nearly null space, i.e. the dimensions are equal. This *simple curve basis of the nearly null space* is more interpretable but still explains the same amount of variation as the initial basis. Any single direction will explain less genetic variation than the largest PC direction in the nearly null space as

well.

This measure of simplicity provides a way to find a basis of interpretable biological directions. Viewing the measure of simplicity of directions of a basis has statistical advantages as well.

One such statistical advantage is when we would like to partition the nearly null space into a subspace of biological interest, i.e. *interesting genetic constraint space*, and one of random noise. The interesting genetic constraint space is generated by particular directions of the nearly null space.

The directions of the nearly null space can be distinguished from directions not in the nearly null space based on their percentage of variation explained. This is because the gap between the other directions percentages of variation explained and the nearly null space directions percentage of variation explained is larger than the random error. But for directions within the nearly null space the gap between percentage of variation explained is not larger than random error. Therefore it is almost impossible to differentiate between directions which generate the interesting genetic constraint space and random error directions based on amount of genetic variation. But by viewing the directions measure of simplicity there is a large enough gap to distinguish between directions within the nearly null space.

Because of this statistical advantage the directions of the simple curve basis of the nearly null space are more stable, from sample to sample, than that of the PCA basis of the nearly null space. One way to generate the interesting genetic constraint space is by choosing directions of an estimated basis. So if the estimated basis is more stable from sample to sample then the estimation of the subspace is also more stable.

Further intuition into the stability of the basis, is gained by thinking of estimating the same subspace for multiple samples. If the same subspace is estimated for multiple samples, the PCA basis can be different for each sample, even if the directions all explain differing theoretical amounts of variation. Because although the same subspace is

estimated for each sample, this does not imply that the subspace has the same estimated covariance structure for each sample, i.e estimated covariance matrix. Since the directions are explaining very similar amounts of theoretical variation, due to random error it is quite easy to have a random reordering of the directions.

But if the same subspace is estimated for multiple samples then the simple curve basis is always the same, given the directions have different simplicity scores. Because if the same subspace is estimated for each sample, then the simplicity structure is always the same, see Section 3.3.1. The simplicity structure is the same because a direction always has the same simplicity score.

3.2 Measure of Simplicity

Once the nearly null space is estimated, we would like to separate it into a subspace of biological interest and a subspace of random noise. This is accomplished by finding the simple curve basis of the nearly null space. Those directions of the simple curve basis which are considered to be simple enough will generate the *interesting genetic constraint space*. Therefore a method to calculate the simple curve basis is needed. But before describing this method, see Section 3.3, the measure of simplicity needs to be better understood, as well as characterized in vector and matrix notation.

The *measure of simplicity* that is being used is the sum of the differences squared of trait values of adjacent environmental levels along a unit length curve. This is analogous to minimizing $\int f'^2$ in the continuous case. An example of how to calculate this simplicity measure for a given direction will aid in the understanding of this measure. Assume that all environment levels are equally spaced, i.e $(e_1 - e_2) = (e_2 - e_3) = \ldots = (e_{d-1} - e_d)$ where $e_1 \ldots e_d$ are the ordered environment levels. Let β be the $d \times 1$ vector which contains the discretized values of a unit length curve, i.e. $\|\beta\| = 1$.

One particular unit length curve of a direction, β, is represented in the object space by the black line in Figure 3.1. The measure of simplicity is then the sum of the squared

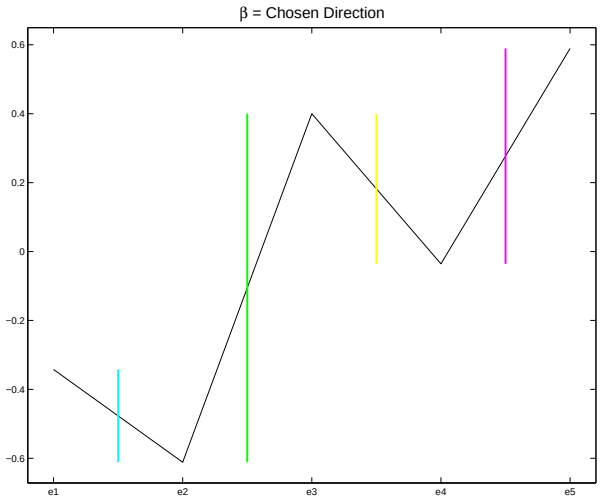

Figure 3.1: *The unit length curve of a chosen direction (β), black line, is shown in the figure. The lengths of the cyan, green, yellow, and magenta line segments are the absolute difference between trait values of adjacent environmental levels. The sum of the squared lengths of these lines is our measure of simplicity.*

differences of trait values between adjacent environment levels. For this particular β, the absolute difference between the trait values at the first and second environment level is the length of the cyan line. The absolute difference between the second environment level and third environment level trait values is the length of the green line, etc. So our simplicity measure is the sum of the squared lengths of the cyan, green, yellow, and magenta line segments. This is the interpretation of the simplicity measure in terms of the object space. But we would like a way to calculate this simplicity measure in the point cloud space, i.e. by vector and matrix multiplication.

The differences between the trait values of adjacent environment levels is calculated by taking the transpose of β and post multiplying it by a difference matrix D, i.e. $\beta^T D$.

Where D is the $d \times d-1$ matrix

$$D = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ 0 & \cdots & 0 & -1 & 0 \\ 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix}.$$

But for the *simplicity measure*, the squares of these differences are considered. Therefore the simplicity measure, m_β, is

$$m_\beta = (\beta^T D)(\beta^T D)^T.$$

This simplicity measure has a low score for the simplest directions. However for interpretation purposes, we would like to have a *simplicity score* which is high for the simplest directions. To achieve this the simplicity measure m_β is subtracted from a constant. In this case the constant is 4, since m_β is always less than 4, see Schatzman (2002). Therefore the simplicity score being used is

$$s_\beta = 4 - (\beta^T D)(\beta^T D)^T. \tag{3.1}$$

Now that the simplicity score is defined, the directions of the nearly null space can be ordered by their simplicity score. This allows for the simple curve basis to be found. An algorithm for finding the simple curve basis by an eigendecompostion of an appropriate matrix is shown in Section 3.3.

3.3 Method To Derive Simple Curve Basis

3.3.1 Methods Relation to PCA

Some intuition into the method to derive the simple curve basis is gained by studying its relationship to PCA. PCA orders orthogonal directions by the amount of variation explained. The directions are found by performing an eigendecompostion of the covariance matrix, $\hat{\Sigma}_{P_N}$. The matrix $\hat{\Sigma}_{P_N}$ summarizes the covariance structure of the subspace. For the simple curve basis analysis we are ordering orthogonal directions by their simplicity scores. The directions are found by performing an eigendecompostion on the simplicity matrix, $\hat{F}_{P_N}$. The matrix $\hat{F}_{P_N}$ summarizes the simplicity structure of the subspace.

Before describing the methods to finding the PCA basis and simple curve basis of a subspace, we first have to define the covariance matrix and simplicity matrix of the full space because the methods are highly dependent upon these. The covariance matrix of the full space is the empirical covariance matrix of the data

$$\hat{\Sigma} = \frac{1}{n-1}(X - \overline{X})(X - \overline{X}),$$

where X is a $d \times n$ data matrix and $\overline{X}$ is a $d \times n$ matrix with each column being the mean of the rows of X.

The simplicity matrix of the full space is

$$\hat{F}^{full} = 4I_d - DD^T,$$

where D is the difference matrix introduced in Section 3.2 and I_d is the identity matrix of size $d \times d$. Some intuition for why F^{full} has this from is gained by extending the definition of the simplicity score for multiple directions of a basis.

The analogous quantity of m_β for the case of a basis is to replace β with I_d, which

is a basis of the full space. This leaves the simplicity measure for multiple directions of the full space as DD^T. But for interpretability purposes we would like simple directions to be associated with high simplicity score. The analogous calculation of subtracting m_β from 4 is to subtract DD^T from $4I_d$. Therefore these steps produce the simplicity matrix $\hat{F}^{full}$ which is analogous to the simplicity score for multiple directions.

In order to find the PCA basis of the full space an eigendecompostion of $\hat{\Sigma}$ is performed. In order to find the simple curve basis of the full space an eigendecompostion of $\hat{F}^{full}$ is performed. But we would also like to find the PCA basis and simple curve basis of a subspace as well. The PCA basis and simple curve basis are found by an eigendecompostion of

$$\hat{\Sigma}_{P_N} = \hat{P}_N \hat{\Sigma} \hat{P}_N$$

and

$$\hat{F}_{P_N} = \hat{P}_N F^{full} \hat{P}_N$$

respectively.

Insight into the form of $\hat{\Sigma}_{P_N}$ is gained by thinking of PCA of projected data. Let P_N be the projection matrix of the nearly null space. The centered data, i.e. $X - \overline{X}$, projected onto the nearly null space is then $P_N(X - \overline{X})$. PCA of the nearly null space is then PCA using this projected data. The covariance matrix of the projected data is

$$\Sigma_{P_N} = \frac{1}{n-1} P_N(X - \overline{X})[P_N(X - \overline{X})]^T = P_N \hat{\Sigma} P_N,$$

which is the covariance matrix of the full space pre and post multiplied by the projection matrix of the nearly null space. Since the nearly null space is not usually known the PCA basis of the estimated nearly null space is found. To do this P_N is replaced by the projection matrix of the estimated nearly null space. This implies that an eigendecomposition of

$$\hat{\Sigma}_{P_N} = \hat{P}_N \hat{\Sigma} \hat{P}_N$$

yields the PCA basis of the estimated nearly null space. The PCA basis of the estimated nearly null space is the eigendirections of $\hat{\Sigma}_{P_N}$ which correspond to the d_N smallest eigenvalues, where d_N is the dimension of $\hat{P}_N$.

To find the simple curve basis an eigendecompostion of

$$F_{P_N} = P_N F^{full} P_N,$$

which is the simplicity matrix of the full space pre and post multiplied by the projection matrix of the nearly null space. To find the simple curve basis of the estimated nearly null space the eigendecompostion of

$$\hat{F}_{P_N} = \hat{P}_N F^{full} \hat{P}_N$$

is performed. The simple curve basis of the estimated nearly null space is the eigendirections of $\hat{F}_{P_N}$ which correspond to the d_N largest eigenvalues, where d_N is the dimension of $\hat{P}_N$. For a more mathematical derivation of F_{P_N} see Section 3.3.2.

A further investigation of $\hat{\Sigma}_{\hat{P}_N}$ and $\hat{F}$ shows an interesting property of the simplicity matrix for different samples. If the same nearly null space is estimated for multiple samples, then $\hat{P}_N$ is the same for each of the samples. This implies that $\hat{F}$ is the exact same for those samples, since F^{full} is the same for every sample. F^{full} is the same for every sample because the projection matrix of the full space is always I_d. Thus the simple curve basis is the same. But for PCA, the matrix $\hat{\Sigma}_{\hat{P}_N}$ is not necessarily the same, because $\hat{\Sigma}$ could be different for each sample. Thus the PCA basis of the nearly null space could be different for each sample, even though the estimated nearly null space is the same.

3.3.2 Mathematical Derivation of F_{P_N}

How to derive the simple curve basis of the nearly null space is defined in Section 3.3 using its relation to PCA for an intuitive understanding of the procedure. This Section provides a more mathematical derivation of the method.

To derive the simplicity matrix a calculation similar to Equation 3.1 is performed. Ideally we would wish to replace β by the basis matrix B. But more careful consideration must be taken when subtracting the simplicity score from 4. Before the simplicity matrix is defined another characterization of s_β is needed. In order for the characterization to be given, first note that

$$s_\beta = 4 - (\beta^T D)(\beta^T D)^T = 4 - \beta^T D D^T \beta.$$

Next we would like to replace DD^T by another matrix which will produce the simplicity score, with out having to subtract from 4. This is done by using the characterization

$$s_\beta = 4 - \beta^T D D^T \beta = \beta^T (4I_d - DD^T)\beta,$$

where I_d is the identity matrix of size $d \times d$.

Based on this characterization of s_β, the simplicity matrix can be defined by simply replacing β by the basis matrix B. Therefore the simplicity matrix is

$$F_B = B^T (4I_d - DD^T)B,$$

where $B = [b_1, b_2, \ldots, b_{d_N}]$ is any $d \times d_N$ basis matrix. Notice that along the diagonal are the simplicity scores of each direction of the basis. Also notice that this matrix is $d_N \times d_N$. Therefore an eigendecompostion of this matrix leads to eigenvectors of size $d_N \times 1$. The eigenvector which corresponds to the largest eigenvalue, is the linear combination of the basis directions with the highest simplicity score. This result is best understood in the

object space, but a vector of size $d \times 1$ is needed. This vector is found by post multiplying B by this eigenvector of F_B. The eigenvector which corresponds to the second largest eigenvalue, is the linear combination of the basis directions orthogonal to the first with the next highest simplicity score, etc.

The eigendecompostion of this F_B leads to the results of ordering orthogonal directions of the nearly null space by their simplicity score. But the basis B must be multiplied by the eigenvectors of F_B to get interpretable results. But the results of this come directly from an eigendecomposition of F_B pre-multiplied by B and post-multiplied by B^T, i.e.

$$F_{P_N} = BF_{d_N}B^T = BB^T(4I_d - DD^T)BB^T = P_N(4I_d - DD^T)P_N,$$

where P_N is the projection matrix onto the nearly null space. The simplest direction of the nearly null space is the eigenvector of F which corresponds to the largest eigenvalue. The direction orthogonal to the first which is next simplest is the eigenvector corresponding to the second largest eigenvalue of F_{P_N}, etc. The matrix F_{P_N} has d_N eigenvalues larger than 0. The simple curve basis is the eigenvectors which correspond to eigenvalues larger than 0.

Since the nearly null space is not usually known, the analysis will consist of an eigendecomposition of an estimate of F_{P_N}. The matrix F_{P_N} is estimated by

$$\hat{F}_{P_N} = \hat{P}_N(4I_d - DD^T)\hat{P}_N,$$

where $\hat{P}_N$ is the projection matrix of the estimated nearly null space, $\hat{S}_N$.

Also notice that F_β is dependent upon the basis, but F_{P_N} depends on the projection matrix. Therefore if different bases are used to define a subspace then F_β will be different for each basis. But P_N is always the same for a subspace so therefore F_{P_N} is always the same. It will then have mathematical advantages to use F_{P_N} as the definition of the simplicity matrix as well, see Chapter 7.

3.4 Simple Curve Basis for Unevenly Spaced Environment Levels

A method to calculate the simple curve basis by performing an eigendecompostion of a simplicity matrix is shown in Section 3.3. But the method assumed that the environment levels are all evenly spaced. But cases exist when environment levels are unevenly spaced, as in Kingsolver *et al.* (2004). For these cases a diagonal weight matrix, W, is added to the analysis. Let e be the vector that contains the ordered values of the environment levels. Then the diagonal weight matrix is

$$W = \begin{pmatrix} \sqrt{\frac{\min_j(e_{j+1}-e_j)}{(e_2-e_1)}} & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\frac{\min_j(e_{j+1}-e_j)}{(e_3-e_2)}} & 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \sqrt{\frac{\min_j(e_{j+1}-e_j)}{(e_d-e_{d-1})}} \end{pmatrix}.$$

Now define

$$F^w = P_N(4I_d - DWW^TD^T)P_N,$$

which is similar to F from Section 3.3 only with a weight matrix included. The same analysis as in Section 3.3 is now performed except with F replaced by F^w. Notice that in the case of evenly spaced points W is equal to the identity matrix and the definition of F^w is the same as F.

A toy example of unevenly spaced environment levels, both with the W matrix included in the analysis and without the W matrix included, is shown to illustrate the purpose and effects of the W matrix. For the toy example, measurements were assumed to be taken at 10, 17, 23, 27, 30, 33, 37, 43, and 50 which corresponds to intervals of 7, 6, 4, 3, 3, 4, 6, and 7. For this toy example, the initial basis is a 9×9 identity matrix. First the above analysis is performed with the F matrix, i.e. all environment levels are

considered to have the same horizontal distance between them, and second with the F^w matrix. The results of the two analyses are shown in Figure 3.2.

In Figure 3.2, the simple curves found when W is not included in the simple curve basis analysis are shown in red. The simple curves found when W is included in the analysis are shown in blue. The curves' simplicities decrease from left to right and top to bottom. The simplest curves for both analyses are the same, a flat line. The second simple curves differ though. The blue curve is more linear than the red. The red curve is steeper in the middle while the blue curve is more uniform throughout. The red curve,

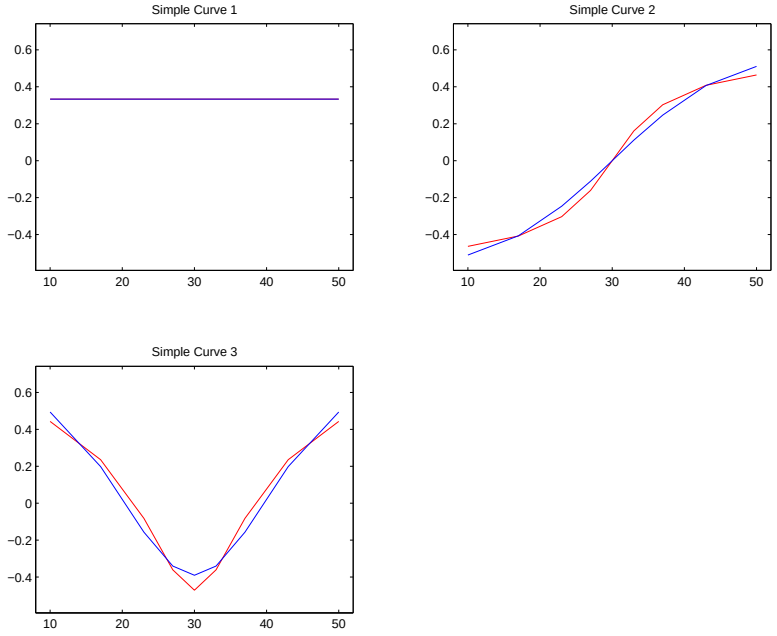

Figure 3.2: *Blue curves are the simple curves with uneven environment levels intervals adjustment. Red curves are the simple curves with out uneven environment levels interval adjustment. Both set of curves plotted on unevenly spaced environment levels. Blue curves are closer to approximate polynomials expected if environmental levels were evenly spaced.*

i.e. when W is not included, treats the environment level intervals as the same and therefore the points along the curve are the same vertical distance apart. But once these vertical points are plotted at the actual environment levels, those environment levels in

the middle are closer together horizontally causing the slope to be larger than at the ends. This is because slope is the vertical distance divided by the horizontal distance. So the differing slopes between intervals is a consequence of the vertical intervals being the same but the horizontal intervals differing. When the W matrix is included in the analysis, it causes smaller horizontal environment level intervals to also have smaller vertical intervals. This makes the blue curves have a more uniform slope, since those with smaller vertical intervals are divided by smaller horizontal intervals. Also notice that the blue parabola in the lower right panel is smoother than the red parabola. This is especially noticable at the peak where the blue parabola is less pointed than the red. Less simple blue curves may appear rougher than the red curves because of corrections to the earlier simpler curves, and the fact that both are bases of the same subspace.

3.5 Variance-Simplicity View of a Direction

A natural way to gain information about data using a basis is by viewing the amount of variation that each direction of the basis explains. This is done by projecting the data points onto a direction and finding the variance of these projected points. If the basis is orthonormal then the directions will explain 100% of the variation with each direction explaining a different mode of variation. So it is often useful to look at what percentage of the variation each direction explains. For the case of genetic constraints we are interested in directions that explain a small percentage of the variation.

But we are also interested in interpretability of directions, i.e. the simplicity of the curve in the object space which corresponds to a given direction. For the case of genetic constraints we are interested in directions which correspond to simple curves. Thus the simplicity score described in Section 3.2 is useful in determining the interpretability of a direction.

For the case of genetic constraints it is not only useful to look at the amount of variation that a direction explains but also the simplicity of the corresponding curves.

Each direction has a percentage of variation explained, as well as a simplicity score associated with it. Interesting genetic constraints are directions that explain a small percentage of variation but also have a large simplicity score. Therefore if we view either only the simplicity score or percentage of variation, we can not determine if the direction is an interesting genetic constraint. A natural way to see if a direction is an interesting genetic constraints is to plot its simplicity score vs the percentage of variation that it explains. A way to visualize these simplicity scores and variation of a basis, along with the unit length curves in object space associated with the directions of the basis, is shown in the next section.

3.6 Example: Caterpillar Growth Rate

This example consists of data of the relative growth rate of *Pieris rapae* caterpillars measured at 6 different temperatures, see Kingsolver *et al.* (2004). The temperatures were 11, 17, 23, 29, 35, and 40°C. Notice that all environment levels are six degrees apart except for the last two which are 5°C apart. So the case of unevenly spaced environment levels will be applied to this data set for the nearly null space analysis.

Three bases will be looked at in the variance-simplicity view to find genetic constraints. The first basis considered is the simple curve basis of $\mathbb{R}^6$, see Figure 3.2 for an example of the simple curve basis of $\mathbb{R}^9$. This basis naturally yields simple directions, so we will see if one of the simple directions explains a small percentage of the variation, i.e is an interesting genetic constraint. The next basis considered will be the PC basis. This basis can produce directions that explain a small percentage of the variation, which are natural genetic constraints. We will then see if one of the directions which explains a small percentage of the variation is simple. The third basis will be a compromise between these two bases. We will consider a subspace generated by the directions of the PCA basis which explain the least amount of variation. Then the simple curve basis of this subspace will be found. This way any direction in the subspace will explain a small

amount of variation, and also the simplest directions in this subspace are found. This simple curve basis of the nearly null space can be combined with a basis of the model space to be a basis of the full space. In this case we will use the the 4 directions of the PCA basis explaining the most variation to be the basis of the model space.

3.6.1 Simple Curve Basis of $\mathbb{R}^6$

The simple curve basis of $\mathbb{R}^6$ is found by the eigendecompostion of F^{full}. The basis formed by the eigenvectors of this solution generate the full space. The large eigenvalues correspond to directions with the simplest curves and the small eigenvalues correspond to directions with less simple curves.

Figure 3.3 shows the simple curve basis of $\mathbb{R}^6$. On the left hand side of the graphic are the unit length curves, shown in the object space, corresponding to the directions of this basis. The curves are numbered with the simplest being number 1 and the least

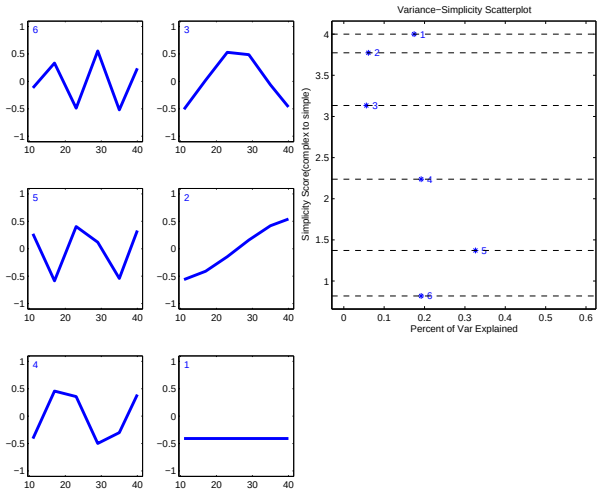

Figure 3.3: *Simple curve basis of full space. Curves are simple but ignore percent of genetic variation explained. Left hand columns are curves in object space corresponding to basis directions. Upper right box is variance-simplicity scatterplot of basis.*

simple being number 6.

One way to view these directions is as curves, but a second way to gain meaning of the directions is to look at how simple they are as well as how much variation each direction explains. The box in the upper right hand corner displays this in a *variance-simplicity scatterplot*. It summarizes each direction as a data point in this particular 2-d space. The vertical dimension is the directions simplicity score, while the horizontal dimension is the percent of variation that the direction explains. Each data point corresponds to a curve on the left and the numbers show this correspondence. The numbers decrease as the vertical axis increases. This is because the directions are ordered by simplicity and the vertical axis is the directions simplicity score. Those with larger simplicity scores are the simplest. The percent of variation, represented by the position on the horizontal axis, that each direction explains does not show an ordered pattern. For instance the simplest direction explains the third most variation while the second explains the least amount of variation. This random relationship will often be the case, since this basis ignores the amount of variation explained. The simplicity scores of these directions will be marked by dashed lines. These same dashed lines will be helpful on similar future plots for other bases as a way to gauge simplicity of directions, in terms of this simple curve basis of $\mathbb{R}^6$.

To find an interesting genetic constraint, we are looking for a direction which explains a small percentage of the variation but also has a large simplicity score, i.e. a point in the top left of these axes. Using this basis the point labeled 2 is the best candidate to be an interesting genetic constraint. This is the direction which corresponds to approximately a linear curve in the object space. This point is quite simple but it explains too much variation. It explains approximately 5 to 6% of the variation. But we defined a genetic constraint as a direction that explains less than 1% of the variation. So this basis has provided a simple curve but the direction explains too much variation to be a true genetic constraint.

3.6.2 PC basis

The PC basis is the eigendecompostion of $\hat{G}$, i.e. the empirical genetic covariance matrix. The basis formed by the eigenvectors of this solution generates the full space. The large eigenvalues correspond to directions explaining a large percentage of the variation and the small eigenvalues correspond to directions explaining a small percentage of the variation.

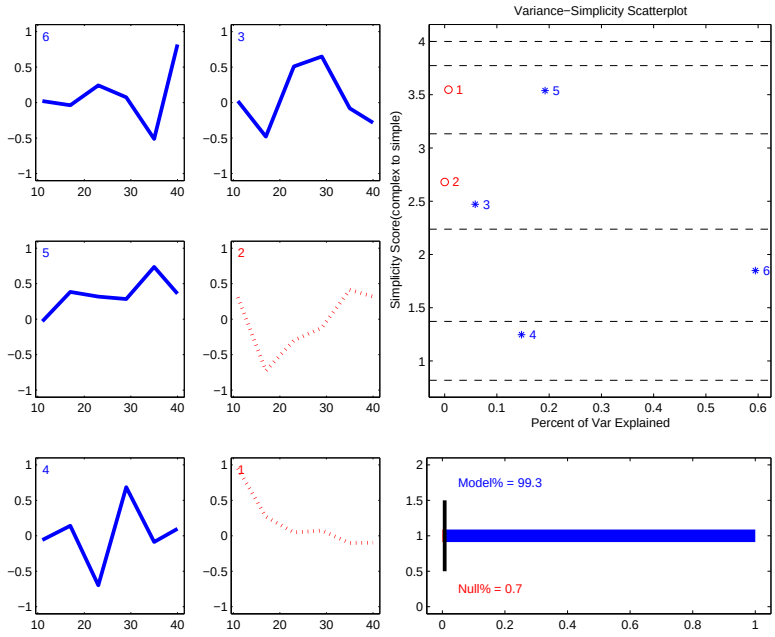

Figure 3.4: *PCA basis for full data set. Directions depend on amount of variation explained and ignore simplicity. Left hand columns are curves in object space corresponding to basis directions. Upper right box is variance-simplicity scatterplot of basis. Bottom right rectangle is amount of variation that model(blue) and null(red) spaces explain. Blue corresponds to model space directions while red corresponds to null space directions.*

Figure 3.4 has the same structure as Figure 3.3, except now we are viewing the PC basis rather than the simple curve basis of $\mathbb{R}^6$. Also the curves are numbered such that 6 explains the most amount of variation while 1 and 2 explain the least amount of variation. The last two PC directions generate the nearly null space. This subspace explains less

than 1% of the variation. This can be seen by the bottom right rectangle which shows the amount of genetic variation explained by those directions in the nearly null space and in the model space. The whole line represents 100% of the genetic variation. The red part of the line is the percent of genetic variation that the nearly null space explains, which is also written in the bottom left corner. In this case the nearly null space explains 0.7% of the genetic variation. This red portion of the bar is difficult to see since it is such a small percentage of the whole bar. The blue portion of the bar represents the amount of genetic variation explained by the model space, which is 99.3% for this basis.

The null space directions are numbered according to simplicity score, with 1 being the simplest, but the model space directions are still numbered by amount of variation explained. The point marked 1 is the best candidate to be an interesting genetic constraint. It explains less than 1% of the variation and has a simplicity score that is about half way between the second and third simplest curves of the simple curve basis of $\mathbb{R}^6$. We can tell this by the fact that the point lies half way between the second and third dashed lines from the top. Recall that these dashed lines mark the simplicity score of the simple curve basis directions. Although this curve is simple there may be a simpler curve that still explains less than 1% of the genetic variation, as derived in the next section.

3.6.3 Simple Curve Basis of the Nearly Null Space

The basis described in this section is a compromise between the bases shown in Sections 3.6.1 and 3.6.2. This basis combines the simple curve basis of the nearly null space with the PCA basis of the model space. A way to visualize the directions of this basis, their simplicity scores, and percentages of variation explained is shown in Figure 3.5. This figure has the same structure as Figures 3.3 and 3.4.

The curves which correspond to this basis are shown on the left hand side. Notice that curves 3-6 have not changed from Figure 3.4. This is because we are using the same basis to generate the model space. The curves labeled 1 and 2 are different from Figure

3.4. This is because the nearly null space basis is the simple curve basis instead of the PCA basis. This simple curve basis still generates the same subspace as the last two PC directions. So the subspace still explains less than 1% of the variation. This can be seen

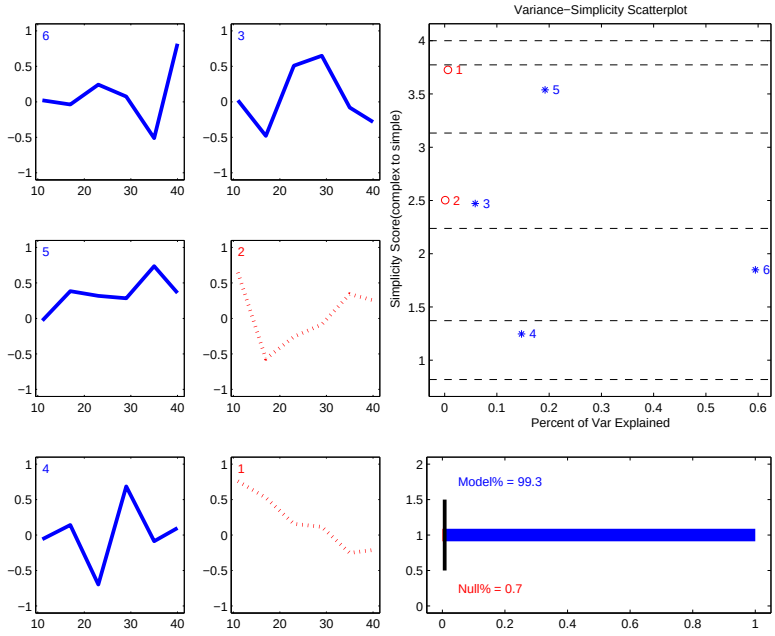

Figure 3.5: *Simple curve basis of 2-d nearly null space. Basis yields best opportunity to find interesting genetic constraints since finding simplest directions in subspace of small genetic variation. Left hand columns are curves in object space corresponding to basis directions. Upper right box is variance-simplicity scatterplot of basis. Bottom right rectangle is amount of variation that model(blue) and null(red) spaces explain. Blue corresponds to model space directions while red corresponds to null space directions.*

in that the bottom right rectangle is still exactly the same as in Figure 3.4. But the basis now has one direction that is simpler than either PC direction of the nearly null space and one curve which is less simple than either PC direction. But the percentages of variation explained by both curves will be between the percentage of variation explained by the last two PC directions.

These facts can best be seen by the box in the upper right corner, i.e. the variance-simplicity scatterplot. Again the points labeled 3-6 are exactly the same as Figure 3.4.

The point labeled 1 is higher vertically than either point 1 or 2 of Figure 3.4. This shows that the curve corresponding to the point labeled 1 is simpler than either PC direction. Also the point labeled 2 is lower vertically than either point 1 or 2 of Figure 3.4. Also points 1 and 2 are horizontally between points 1 and 2 of Figure 3.4. This is harder to see since the points are so close together horizontally. But the sum of the horizontal values is the same as the sum of the horizontal values of Figure 3.4, showing that the nullspace generated by both bases explain the same amount of variation.

Point 1 is the best candidate to be an interesting genetic constraint. This point is much closer to the second dashed line from the top, saying that it is approximately as simple as the second simplest curve in the simple curve basis of $\mathbb{R}^6$. But it still explains less than 1% of the variation. So we have found a simpler curve than the PC basis, which still explains a small percentage of the variation.

To see variations of this type of null space analysis ranging over 0 to 6 lower PCA directions generating the null space go to Gaydos (2007). At this website is a movie, named CatNull in the folder NullSpace, in which each frame is similar to Figure 3.5 but with different number of principal component directions generating the null space. Each frame has one more principal component direction generating the null space than the frame before it. Figure 3.5 is a snapshot from this movie, that was included because the nearly null space generated by the last two PC directions is considered the most reasonable estimate of the nearly null space.

The method for deciding the dimension of the nearly null space is to view the simple curve basis of the nearly null space of dimension d_N, i.e. the subspace generated by the last d_N PC directions, using the visualization tool of Figure 3.5. Then compare that to the simple curve basis of the nearly null space of dimension $d_N + 1$ using the same visualization tool. If the proportion of variance of the nearly null space of dimension $d_N + 1$ is much larger than the nearly null space of dimension d_N, then the nearly null space is considered to be of dimension d_N. If the proportion of variation of the $d_N + 1$

nearly null space is not much greater than that of d_N then the nearly null space should be of dimension $d_N + 1$. If it is decided that the nearly null space is of dimension $d_N + 1$, then the subspace of dimension $d_N + 1$ is compared to the subspace of dimension $d_N + 2$, etc.

For the example above the simple curve basis of the nearly null space of dimension 1 is shown in Figure 3.6. We will compare this to Figure 3.5 in order to determine if the nearly null space should be of dimension 1 or 2. Notice that this is the same as the PC

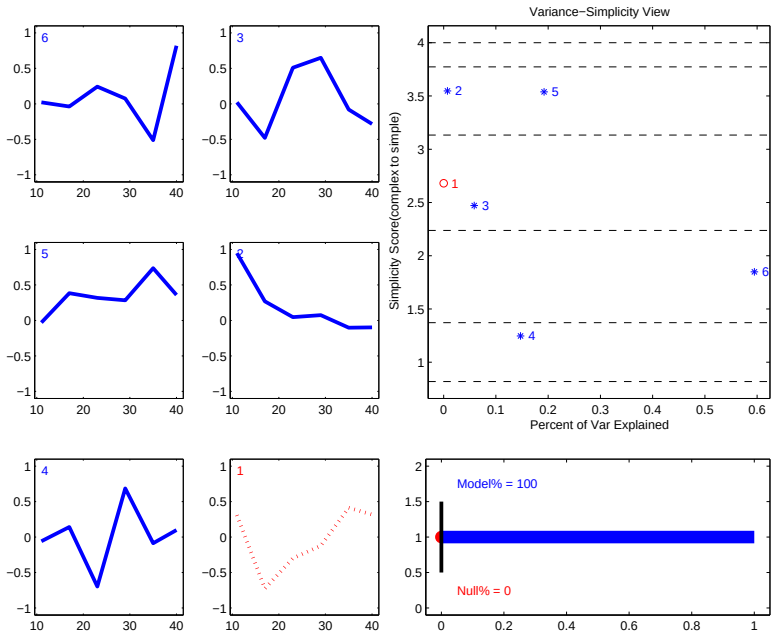

Figure 3.6: *Simple curve basis of 1-d nearly null space. Basis yields directions of small variation but not a large enough space to give an opportunity to find simple directions. Left hand columns are curves in object space corresponding to basis directions. Upper right box is variance-simplicity scatterplot of basis. Bottom right rectangle is amount of variation that model(blue) and null(red) spaces explain. Blue corresponds to model space directions while red corresponds to null space directions.*

basis, only with the last PC direction identified as the nearly null space, i.e. is colored red. All of the curves are the exact same as the PC basis. This is because the nearly null space is of dimension 1, i.e. has only one direction. Therefore the simplest direction

of this one dimensional space is also the direction which explains the least amount of variation. This nearly null space explains 0.0% of the genetic variation, see bottom right rectangle. This is a small amount of variation and can be considered a genetic constraint space. But there may be a larger space, in which all directions can still be considered to be genetic constraints. Therefore let us compare this Figure to Figure 3.5.

Figure 3.5 shows the simple curve basis of the nearly null space of dimension 2. It can be seen from the bottom right rectangle that this nearly null space explains 0.7% of the genetic variation. Therefore it is reasonable to assume every direction of this subspace is a genetic constraint. This subspace is larger than the subspace generated by only 1 PC direction and therefore gives a better chance of finding a set of interesting genetic constraints, i.e. directions of high simplicity and low variation.

Next we will investigate the subspace generated by the last 3 PC directions. Figure 3.7 shows the simple curve basis of this space. This space explains 6.6% of the genetic variation. This is a large amount of genetic variation. Since this space is larger than the subspace generated by the last 2 PC directions we are more likely to find directions of high simplicity. But since this subspace explains such a large amount of genetic variation it is also more likely that the directions of the simple curve basis are not genetic constraints, i.e. will explain too much genetic variation. So therefore the nearly null space of dimension 2 gives us the best chance of finding a set of interesting genetic constraint.

Also 2 movies, named JWNullshade and JWNullSun at the website Gaydos (2007), are the same simple curve basis analysis for 2 different 6 dimensional data sets. These data set consist of height measurements of a group of Jewel Weed plants. The height was measured at 18, 26, 33, 39, 47, and 57 days. The jewel weed plants were separated into two groups, with one group growing in the sun and the other in the shade. The shade jewel data has a nearly null space of 5 dimensions, while the sun jewel data has a nearly null space of 4 dimensions. Although sun jewel weed data could also be considered to

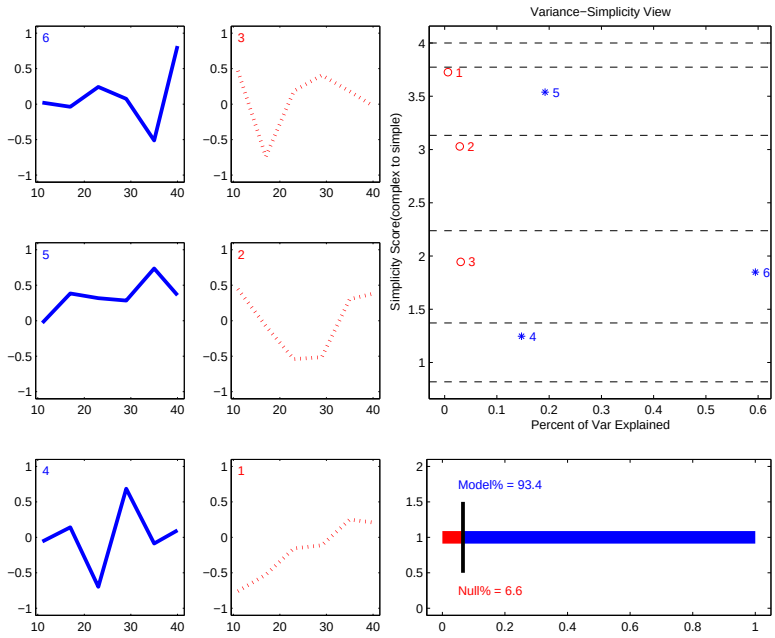

Figure 3.7: *Simple curve basis of 3-d nearly null space. Larger subspace to find simple directions but directions likely to explain too much genetic variation. Left hand columns are curves in object space corresponding to basis directions. Upper right box is variance-simplicity scatterplot of basis. Bottom right rectangle is amount of variation that model(blue) and null(red) spaces explain. Blue corresponds to model space directions while red corresponds to null space directions.*

have a 5 dimensional nearly null space.

CHAPTER 4

Principal Components For Developmental Stage Data

In this chapter the biological notion of *developmental stages* is used to motivate a new representation of functional data, similar to classical shape statistics. The new data representation transforms conventional non-linear variation into easily analyzed linear modes of variation.

Section 4.1 provides an introduction of Developmental Stage Landmark Data. Section 4.2 will give an overview of how PCA is performed on a grid based representation of the curves and then how PCA is performed on the representation of the curves using developmental stage landmarks. We will assume that all individuals have the same number of developmental stages for simplicity in explanation. Since the developmental stage landmarks represent horizontal variation as linear modes of variation, PCA a tool to understand linear modes of variation will provide easier to interpret results for the developmental stage landmark representation. Section 4.3 will describe how to handle the situation when some developmental stage landmarks do not correspond between some individuals in the population with a focus on the Manduca sexta data set, described in Section 4.1. In Section 4.4 the results of landmark PCA will be shown for the Manduca sexta data set.

4.1 Introduction to Developmental Stage Landmark Data

For data sets that have an attribute as a function of time the common Functional Data Analysis practice is to measure the attribute at a set of given, usually evenly spaced, time points. But for some data sets it is more natural to have the time points be variable as well. One such example is the growth trajectories of the tobacco hornworm, Manduca sexta. These Manduca sexta go through the process of molting and metamorphosis, i.e. *developmental stages*, see Gilbert *et al.* (2000), Nijhout (1994), Riddiford *et al.* (2003). It is natural to measure the mass of the individual as well as the time of these developmental stages. The Manduca sextas' growth trajectories are shown in Figure 4.1. The log of

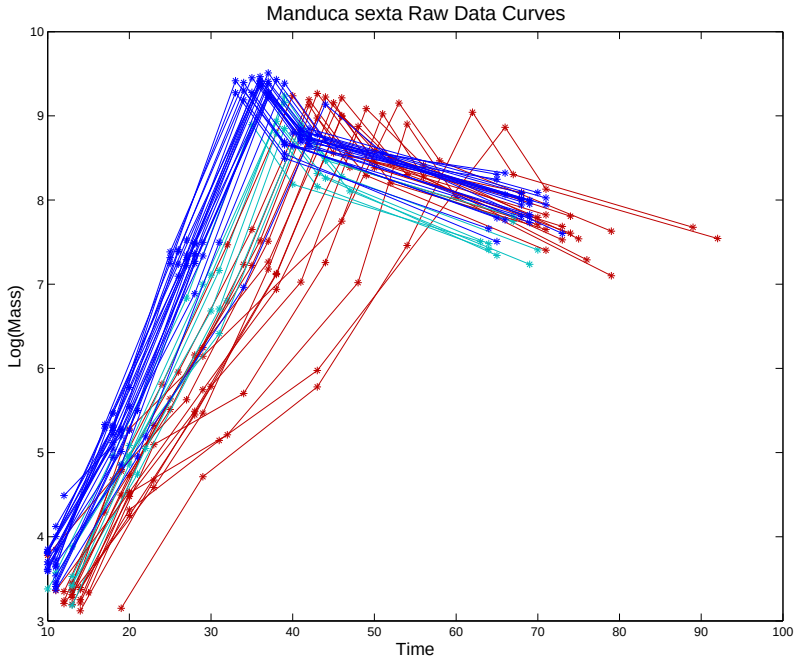

Figure 4.1: *Manduca sexta growth trajectories. Blue curves are lab caterpillars with 6 stages, cyan are field caterpillars with 6 stages, and red are field caterpillars with 7 stages. Asterisks represent developmental stages. Notice that the curves have horizontal, i.e time, modes of variation.*

mass is on the vertical axis and time is on the horizontal axis. The individuals in this data set either come from a lab population or a field population. The lab individuals will be colored blue and always have 6 stages. The developmental stages for all individuals are shown as asterisks. For the field population some individuals have 6 stages while others have 7 stages. Those field individuals with 6 stages will be colored cyan while the field individuals with 7 stages will be colored red. For this data set the time points are variable, i.e. the asterisks for each developmental stage are not in a straight vertical line.

The conventional approach to varying time points involves interpolation, see Ramsay and Silverman (2002) for a more detailed discussion on this topic. Information can be lost using this conventional interpolation approach, when there are interesting modes of variation in the time direction. The Manduca sextas' growth trajectories shown in Figure 4.1 have interesting modes of time variation, such as an overall horizontal shift of the curves. Using the conventional approach the horizontal, i.e time, modes of variation have been successfully modeled using nonlinear methods, see Izem and Kingsolver (2005).

Here we account for the time variation by using developmental stage landmarks. This developmental stage representation of the data has an advantage of representing interesting modes of variation, that were previously modeled in a nonlinear way, in a linear fashion. These developmental stage landmarks have a connection to classical shape statistics.

Landmarks form the basis of a large, and well developed, theory of statistical shape analysis. See Kendall (1999), Bookstein (1978), and see Dryden and Mardia (1998) for a particularly accessible introduction. In this approach to shape statistics, landmarks are points of correspondence on an object, i.e. a curve, that match between and within populations. The developmental stages along the curve are analogous to anatomical landmarks in shape statistics. In many biological applications anatomical landmarks are points assigned by an expert which correspond between organisms in some biologically meaningful way, see Dryden and Mardia (1998). In shape statistics, landmarks allow a

statistician to analyze how an object changes shape horizontally and vertically. In this same manner, the developmental stages allow for a statistician to see how the curves change with respect to both the attribute and time.

Because of this developmental stage representation linear methods, such as PCA, can be used to understand horizontal variation. The next section shows how to use PCA to understand the variation of the curves using the grid based and developmental stage representations.

4.2 PCA for Developmental Stage Landmark Data

The common practice, for functional data sets that involve an attribute value as a function of time, is for the experimental design to be such that the attribute is measured at a set of d given time points. A conceptual advantage of this *grid based experimental design* is that each curve can be thought of as a point in $\mathbb{R}^d$. See Section 1.1 for an introduction to this correspondence of the curves, i.e. object space, and $\mathbb{R}^d$, i.e. the point cloud space. The attribute value at each given time point is a coordinate value in $\mathbb{R}^d$. If one wants to reconstruct the curve from the data point it is important to keep track of the order along the curve from where the attribute measurement is associated. Only the order is important since the time points are common for all individuals. Vector notation allows us to do this easily. A point in $\mathbb{R}^d$ corresponds to a $d \times 1$ vector. Each individual correpsonds to a vector, x_i, which is thought of as a discretization of an individual curve. The first entry in the vector is the measurement that was taken at the first time point along the curve. The second entry is the attribute value taken at the second time point along the curve, etc. The data vectors are usefully grouped together into a $d \times n$ data matrix, X. The data matrix X corresponds to n points in $\mathbb{R}^d$.

A good way to understand the variation of these points is through PCA. Traditional PCA has a requirement that the data vectors have the same length, which is guaranteed by this design. This is why it is common to use the grid based experimental design when

planning to use PCA to understand variation, even if interpolation must be used, e.g. for varying time points.

A limitation of this experimental design is that only vertical variation is explicitly modeled. But for Figure 4.1 it is clear that horizontal, i.e. time, variation is also of interest. But PCA is heavily dependent on the input vectors, i.e. data representation. If different representations of the data are used then PCA will give different modes of variation. For the grid based representation, if there is time variation in the curves then PCA will try to model the time variation in terms of mass variation. This is since the PCA input vectors only have mass values. So if the input vectors do not have time values which vary, then PCA can not explicitly model time variation.

When the data consists of developmental stage landmarks, each curve has a measurement of time and mass at d developmental stages. Each curve can be represented as a point in $\mathbb{R}^{2d}$. The dimension of the space is twice the number of landmarks, because not only is there an attribute measurement but also a time measurement. Each attribute value is again a coordinate value but also each time point is a coordinate value as well. No longer are all individuals associated with the same given time points, therefore the time measurements also have variation. Now PCA can be used to understand both vertical and horizontal variation explicitly, since both the variation of mass and time are reflected directly by the data vectors. A major advantage of this data representation is that biologically important modes of variation, that were nonlinear using the classical grid based representation, now become linear.

To reconstruct the curves from the data points in $\mathbb{R}^{2d}$, it is important to know which landmark each attribute and time measurement is associated with. Again the vector notation allows us to do this easily. Each individual corresponds to a $2d \times 1$ vector, x_i^L, which is a discretization of it's curve. By design, the first entry corresponds to the attribute value of the first landmark and the $d+1$ entry is the time point associated with the first landmark. The second entry is the attribute value associated with the second

landmark and the $d+2$ entry is the time point associated with the second landmark, etc. These vectors can be grouped together into a $2d \times n$ data matrix, X^L, which corresponds to the n points in $\mathbb{R}^{2d}$.

A toy example is given to highlight the important point about a landmark representation turning nonlinear modes of variation into linear modes of variation. In this example it is assumed that the curves have 7 developmental stage landmarks. The toy example consists of parabolas that are shifted along the horizontal axis. The curves are rainbow colored with curves having early peaks being colored blue to the curves with later peaks being colored red. The actual data curves are shown in upper left panel of Figure 4.2.

The grid based experimental design measures the curves at a set of given time points. For this toy example, 30 grid points were chosen to be evenly spaced between -3 and 3.

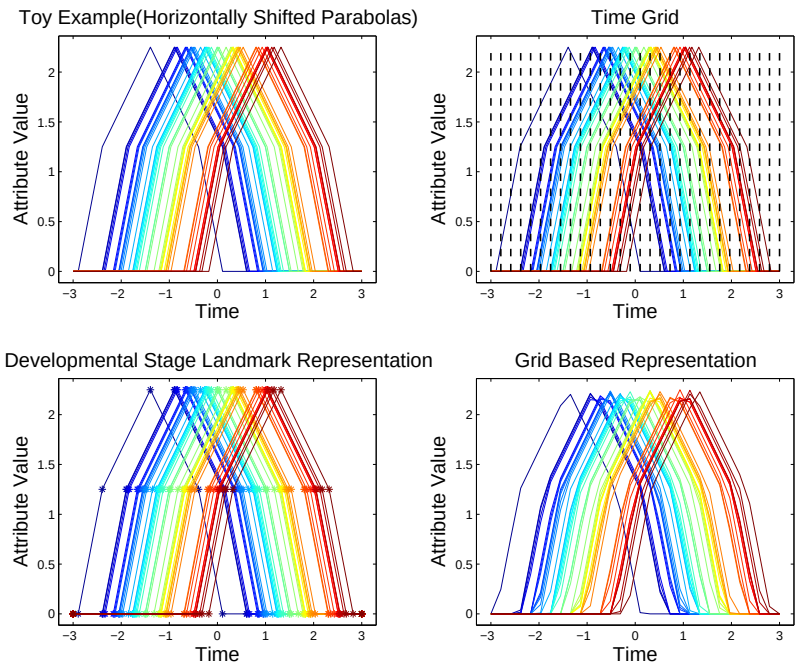

Figure 4.2: *Top left are toy data curves. Dashed lines in top right are the given evenly spaced time points. Bottom right is the grid based representation of curves. Bottom left are the landmark representation of curves. Notice that landmark representation does much better job of representing actual curves than grid based representation.*

This range was chosen so that all curves have an attribute value at each given time point, i.e all curves start at -3 and end at 3. The given time points are shown in the upper right of Figure 4.2 as dotted lines. Where the dotted lines intersect a curve are the attribute values for that curve. Each curve is represented by a point in $\mathbb{R}^{30}$, which corresponds to a 30×1 vector, x_i. This is because there are 30 attribute values associated with each curve at the 30 given time points. The bottom right shows the representation of the curves based on these grid points and attribute values. The grid based representation of the curves is produced by plotting the attribute values at the given time points. The first entry of x_i is plotted at given time point 1 and the second entry of x_i is plotted at given time point 2, etc. The points are then linearly interpolated in between the given time points. Notice that these curves look similar to the original curves but they are not exactly symmetric and have corners which are different. For instance around the peaks the lines before and after the peaks are not mirror images of each other as in the original curves.

The developmental stage landmark representation of the curves is shown in the bottom left of Figure 4.2. Each developmental stage is represented by an asterisk. At each asterisk the attribute value and time is measured. Now each curve is represented as a point in $\mathbb{R}^{14}$, which corresponds to a 14×1 vector, x_i^L. The mass measurements are recorded in the first half of the vector and the time measurements are recorded in the second half of the vector. The developmental stage representation of the curves are produced by plotting the attribute value at the landmarks vs the time at the landmarks and linearly interpolating the attribute value and time points in between landmarks. The 1^{st} entry of x_i^L, i.e. the attribute value at the first landmark, is plotted vs the eigth entry of x_i^L, i.e the time value at the first landmark, and the second entry of x_i^L is plotted vs the ninth entry of x_i^L, etc. For this toy example, the curves are represented perfectly by the developmental stage landmark representation. Notice that the top left and bottom left panels of the figure have curves which are exactly the same.

Since each curve is only the mean curve shifted along the horizontal axis, intuitively one expects the variation of the curves to be one dimensional in nature. Figure 4.3 shows the results of a PCA when the common grid based representation is used. In Figure 4.6 are the results of a PCA when the developmental stage landmark representation is used. These figures will help to compare and contrast the two representations of the data.

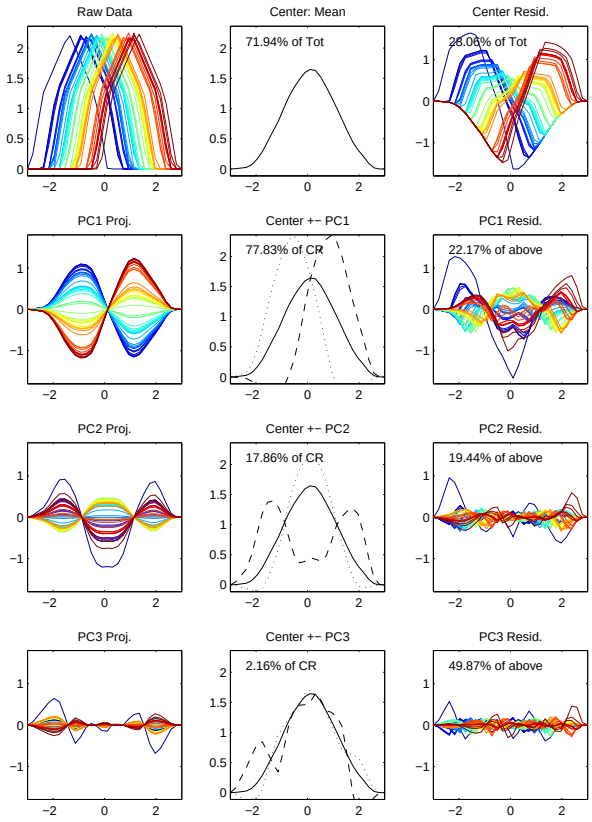

Figure 4.3: *PCA of Grid based representation, not all variation explained by one linear dimension due to representation. Horizontal shift is modeled as vertical modes of variation. Top row from left to right are the grid based representation of the curves, mean curve for this representation, and residual curves. Second Row from left to right are projected data curves of PC 1, mean curve plus and minus multiple of PC 1 curve, and residual curves. Rows 3 and 4 have same structure as Row 1 except with PC 2 and 3 instead of PC 1.*

Row 1 Column 1 of Figure 4.3 shows the grid based representation of the curves. The

curves are the same as in the bottom right of Figure 4.2. In Row 1 Column 2 is the mean of these curves. The mean curve at each time value is the mean of attribute values at that time point, i.e. vertical dashed line in the right of Figure 4.2. The peak of this curve is not at the vertical height of all of the peaks of the other curves. This is because the peak of the mean curve is the mean of the measurements at a specific vertical line, not the mean height of all of the peaks. The mean curve is more bell shape than linear line segments, which differs from the curves in the top left. This mean curve being a different shape than all of the curves indicates an important sense in which it does not accurately reflect the center of the curves. In Row 1 Column 3 are each individual's curve after the mean has been removed, i.e the mean residuals.

Row 2 Column 1 shows the projection of each individual onto the PC 1 direction vector called v_1. This corresponds to plotting the projections of the data onto v_1, denoted as $p_{1,i}$, vs the given time points, where

$$p_{1,i} = \langle v_1, x_i \rangle v_1 = v_1^T x_i v_1.$$

This is the representation of the curves, after the mean has been removed, if only one linear direction was used. The mode of variation described by PC 1 is about the blue curves having a larger mass at early time points and the red curves having higher masses at later time points.

This mode of variation is due to the representation of the data. For this representation PCA is trying to explain horizontal shift via vertical variation. The blue curves have peaks before time point 0 while the red curves have peaks after time point 0. The mean curve has its peak around time point 0. PCA of this representation only explicitly models vertical variation from the mean. So the blue curves have their peaks early and are then vertically higher than the mean curve for early times, where as the red curves have their peaks later and are vertically higher than the mean curve for later times. The red and

blue curves are most prominent in PC 1 due to the fact that the farther a peak is from time point 0 the more vertical variation from the mean curve. This is because the mean curve's height decreases as it moves away from time point 0. The shape of the curves are reminiscent of the second Hermite basis function.

An important point is that PC 1 does not explain all of the variation from the mean. Row 2 column 2 shows the extreme positive and negative multiples of v_1 added to the mean curve and shows the percent of variation that PC 1 explains in the top left corner. PC 1 only explains around 78% of the variation, but it was expected that all of the variation could be explained by one direction. Row 2 Column 3 shows the residuals after the mean and PC 1 have been removed. Again these curves show that there is a significant amount of variation that is not explained by PC 1.

Row 3 has the same structure as Row 2 except now all of the panels involve PC 2 instead of PC 1. The third column is the residuals after the mean, PC1, and PC 2 have been removed. The rows continue with this same structure. PC 2 explains a significant amount of the variation, about 18%. The mode of variation of PC 2 is about the curves with peaks in the middle time range vs curves with peaks at the starting or ending time ranges. This mode of variation is reminiscent of the third Hermite basis function. Again this is because the obviously horizontal mode of variation is being modeled via vertical modes of variation.

The data was simulated to have, in some intuitive sense, only one dimension of variation, Figure 4.3 shows that the data lie in a more than one dimensional subspace using this representation. This is because PCA only finds directions of linear variation but the family of curves has non-linear variation, that was imposed by the grid based representation. This non-linear, yet one dimensional variation can be understood by looking at a scatter plot of the PC scores, which is shown in Figure 4.4.

Figure 4.4 is a graphic of a 4×4 matrix of the scatter plots of PC scores. This graphic allows us to visualize a 4 dimensional subspace through projections of the data

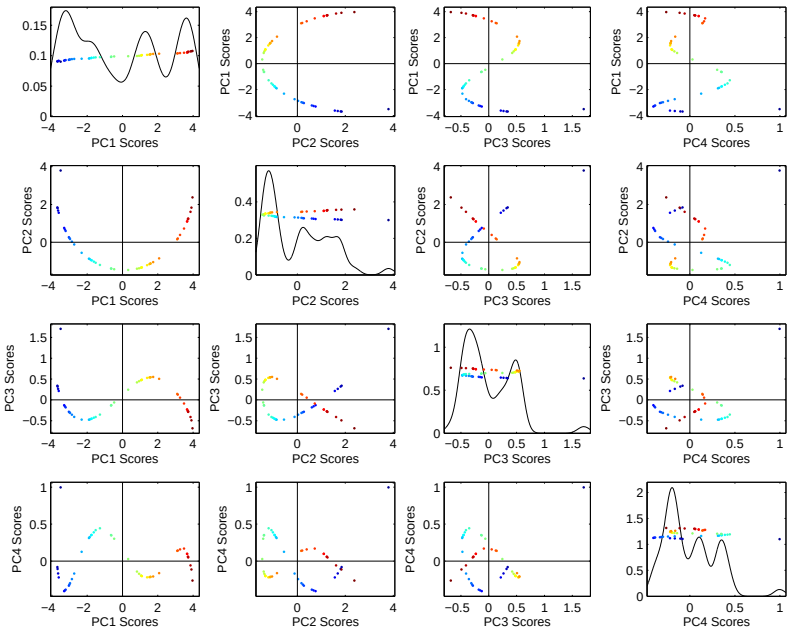

Figure 4.4: *Along diagonal are projections of data onto 1-d subspaces generated by PC directions. Off diagonal are data projections onto 2-d subspaces generated by PC directions. Notice that all of the data lies along a line in each of the 2-d off diagonal subspaces, which indicates non-linear variation.*

onto 1 dimensional and 2 dimensional subspaces generated by the PC directions. Along the diagonal is the plot of the projections onto the 1 dimensional subspaces generated by PC directions 1-4. Row 1 column 1 is the projection of the data onto PC direction 1, and row 2 column 2 is the projection of the data onto PC direction 2, etc. Also in the diagonal panels are a smoothed density curve of the PC scores. These are kernel density estimates, see Wand and Jones (1995) for a good introduction. Off diagonal are the projections of the data onto a 2-dimensional subspace generated by pairs of PC directions. Row 1 has PC1 projections along the vertical axis for columns 2-4 and the projection onto PC2-4 along the horizontal axis respectively. Row 2 has projections onto PC 2 direction along the vertical axis for columns 1, 3, and 4 and the projections of the data onto PC directions 1, 3, and 4 respectively. The rows continue in this way.

Row 2 column 1 shows the projection of the data onto the subspace generated by the PC 1 and PC 2 directions. All of the data points line along a parabola in this scatter plot. This shows that indeed there is non-linear variation in the data set, i.e. this "one-dimensional" data set lies in at least a 2-dimensional subspace. All of the other scatter plots show that the data points lie along some curve as well, again highlighting the non-linear variation. In particular the data lies in at least a 4-dimensional subspace.

The data are in some sense 1 dimensional since all of the data points lie along a curve in a 4-d space. This 1 dimension is explained by a horizontal shift, but the grid based representation tried to model this horizontal shift through vertical variation, i.e the horizontal variation is more efficiently modeled as nonlinear. The landmark representation allows PCA to explicitly model both vertical and horizontal variation, so the horizontal shift can now be modeled through horizontal variation rather than vertical variation, i.e horizontal variation can now be efficiently modeled as linear.

If the data have developmental stage landmarks, then each individual is represented by a data point that can have variation in both mass and time. For this toy example, the curves are parabolas shifted horizontally and not changed vertically. Because of this the curves have no mass variation and all of the variation is in a linear direction of time variation.

The upper left panel of Figure 4.5 shows the landmark representation of the data set in the mass vs time view, it is the same as the lower left panel of Figure 4.2. The upper right panel shows the landmark representation of the data viewed in parallel coordinates. Each view of the data can be used to gain useful insight. Certain aspects of the data may be better understood in the mass vs time view than parallel coordinates view or vice versa.

Parallel coordinates is a multidimensional visualization tool, where each entry of a vector is plotted vs its entry number, see Inselberg (1985). For this case because of the way we grouped the data measurements in the vector, all of the mass measurements

will be plotted first and then all of the time measurements will be plotted next. To gain understanding of the data through the parallel coordinates view, the ordering of the measurements in the vector is essential. To form the parallel coordinates plot the values

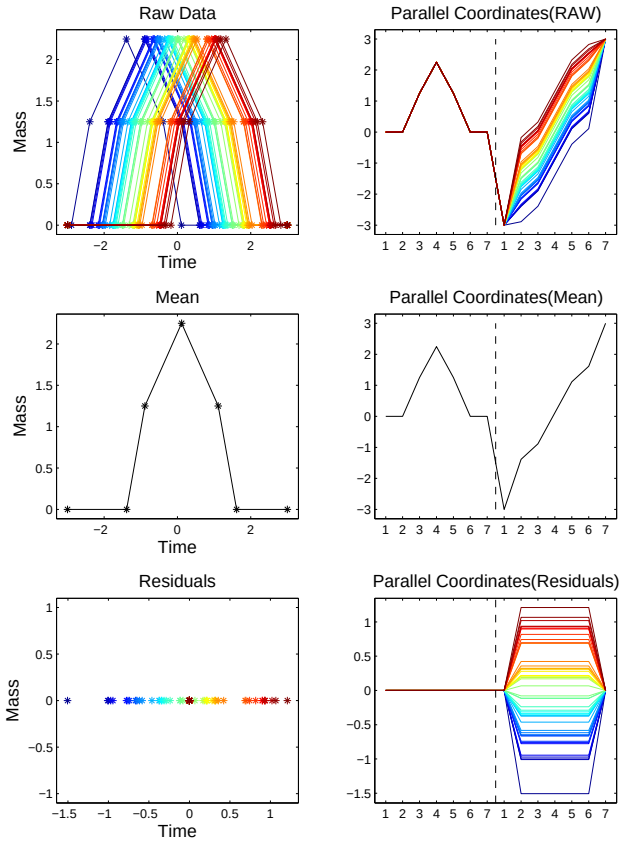

Figure 4.5: *Left hand side is mass vs time view and right hand side is parallel coordinates view. Top row is landmark representation of curves in mass vs time view, left, and parallel coordinates view, right. Row 2 is mean curve shown in mass vs time view, on left, and parallel coordinates view, on right. Bottom row is mean residuals shown in mass vs time view, on left, and parallel coordinates view, on right. All lines are multiples of each other in the column 2 residual plot indicating that variation in data can be explained by one linear direction.*

for the mass of each developmental stage landmark are plotted first and then the time of each developmental stage landmark next. This also corresponds to plotting each column vector of the data matrix, due to the chosen structure of the vectors. Notice that the

horizontal axis shows numbers 1-7 twice. The numbers represent which developmental stage each value corresponds to, and the numbers are there twice since each developmental stage corresponds to both a mass, shown on the left, and time measurement, shown on the right. The dotted line separates mass values from time values.

For the toy example the mass measurements form a parabola over the developmental stages, while the time measurements form lines that increase as the developmental stage increases. Notice that each individual has the same left hand side, i.e mass portion of the curve, but different right hand sides, i.e. time portions of the curve. In row 2 is the mean curve viewed in the mass vs time view, i.e. column 1, and the parallel coordinates view, i.e column 2. The mean curve in the mass vs time view is a parabola. The mean curve, denoted by the vector $\overline{X}^L$, has entries that are the mean of all individuals for that specific entry of the vectors, i.e. the mean mass or time of each landmark. This parabola has the same shape as all of the raw data curves. All of the data curves are simply a horizontal shift of this curve. The parallel coordinate view shows the same thing. This already shows that this mean curve is an intuitively more sensible representation of the mean, than the mean curve of the grid representation.

Row 3 is the residuals after the mean has been removed. In the mass vs time view the picture does not give us much information. This indicates that the residuals should be viewed added to the mean in the mass vs time view, for easier interpretation. Each point, not at (0,0), is actually 5 points, because each developmental stage is shifted the same amount horizontally from the mean except the first and last stage. The first and last stage of all individuals is now at the point (0,0), since all individuals have the same value for the first and last developmental stages. All stages have a vertical value of zero since they are never shifted vertically from the mean but only horizontally.

The parallel coordinates view of the residuals is more easily interpreted with the mean removed than the mass vs time view. Now it can be seen that the mass measurements have no variation from the mean, since they are all flat lines along zero. It is also seen

that the time measurements for stages 2-6 are the same distance from the mean for each color, i.e. individual. The first and last stages have no variation. All of these lines are parallel which shows that the data has only 1 linear direction of variation for this representation. This is only time variation, since only the right hand side of the curves have spread to them.

Figure 4.6 shows the landmark PCA for this data set. The first two columns of Figure 4.6 are visualizations of the data in the mass vs time view, while the last two columns are visualizations of the data in parallel coordinates view. Row 1 shows the raw data in column 1 using the mass vs time view. This is the same developmental stage representation of the curves as shown in Figures 4.1 and 4.5. This is entries 1-7 of X_i^L plotted vs entries 8-14 of X_i^L. This representation shows the curves are parabolas, shifted along the horizontal axis and are exactly the same as the true data curves. Row 1 column 2 shows the mean of the data in the mass vs time view. The mean is the parabola which is shifted along the horizontal axis to generate the curves.

The third column shows the same data set, with the mean removed, in the parallel coordinates. This is the same view that is used in the second column of Figure 4.5. In this view it is easy to see that the mass measurements have no variation, i.e. are all flat lines at 0. The time measurements are parallel flat lines. All the lines being parallel is an indication that the variation can be explained by one linear direction using the developmental stage representation. This tells us that the middle 5 landmarks for each individual are shifted horizontally from the mean by the same amount, suggesting that the variation is linear in this representation. The first and last stages have the same time measurements for all individuals and therefore have no variation.

Row 2 column 1 shows the projection of the data onto the PC 1 direction, v_1^L, for this representation as a function of mass versus time. This is achieved by adding the projection vector to the mean vector, $p_1^L + \overline{X^L}$. Then plot entries 1-7 versus entries 8-14.

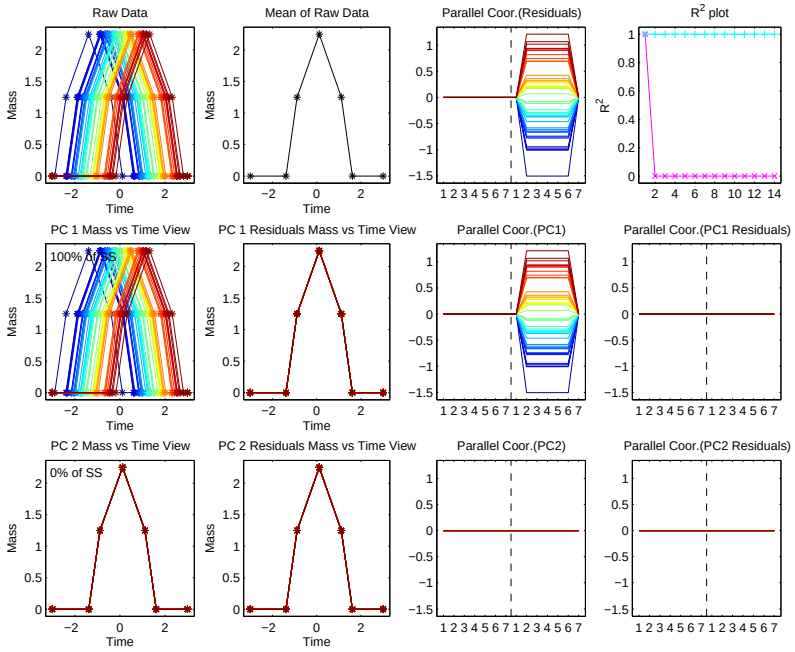

Figure 4.6: *Landmark PCA of Toy Data, now all of of the variation is explained by one linear direction. Left 2 columns are mass vs time view. Right 2 column are parallel coordinates view. Row 1 is raw data, row 2 is PC 1 projections, and row 3 is PC 2 projections.*

The vector

$$p^L_{1,i} = \langle v^L_1, x^L_i \rangle v^L_1$$

is the projection of individual i onto the PC 1 direction. This is the best representation of the curves if only one linear direction is used. In this case the projection curves are the exact same as the original curves. This is because PC 1 explains all of the variation of the curves. This can also be seen in the upper left corner of the panel, where the number, 100%, tells us how much variation is explained by this direction.

Row 2 column 3 shows the projection of each individual onto the PC 1 direction, with the mean removed, using parallel coordinates. This again corresponds to plotting the projection vector,

$$p^L_1 = \langle v^L_1, x^L_i \rangle v^L_1$$

of each individual for the PC 1 direction. Once again these curves are the exact same as the curves in the panel above it, indicating that one linear direction explains all of the variation.

Row 2 column 2 shows the residuals, after PC 1 has been removed, added to the mean plotted in mass vs time view. This panel shows the mean curve which indicates there is no variation left to explain. In row 2 column 4 the residuals are shown, after PC 1 is removed and with out the mean, plotted in parallel coordinates view. The residual curves are flat lines at zero. This is another indication that there is no variation left to explain after PC 1. Row 3 has the same structure as row 2 except with PC 2 instead of PC 1. PC 2 explains 0% of the variation, since there was no variation left to explain after PC 1.

4.3 Representation for a Differing Number of Developmental Stages

PCA is a tool for understanding linear variation and therefore has been useful for many data sets, for reasons stated in Sections 1.1 and 4.1. A weakness of PCA, for analyzing developmental stage data, is it requires data vectors of the same length. So a classical analysis requires a common number of developmental stages. But experiments may not always consist of individuals with the same number of developmental stages, such as the Manduca sexta data set introduced in Section 4.1. PCA can be used when an appropriate modification is made to the developmental stage landmark representation. This section will provide one option for such a modification. The new representation will be presented through the Manduca sexta example.

Manduca sexta individuals historically have 5 instars and 3 other developmental stages but through evolution some now have 6 instars and 3 other developmental stages. For this experiment the individuals were not measured until the third instar. For this reason those individuals with 5 instars were only measured at 6 developmental stages and will

be referred to as 6 stage individuals. While those with 6 intstars were measured at 7 developmental stages and will be referred to as 7 stage individuals. The stages also will be referred to as stage 1 starting with the third instar and counting up from there. For a visualization of the curves from the third instar on see Figure 4.1. Due to the population having individuals with differing number of developmental stages the developmental stage representation described in Section 4.2 will have to be modified before traditional PCA can be performed.

If it was biologically evident which developmental stage was extra then those individuals with 6 stages could have a pseudo-landmark added, i.e. a seventh developmental stage, through interpolation. If for instance, the extra stage was between the original first and second stage then a developmental stage could be added between these two stages through linear interpolation. By embedding those curves of individuals with 6 stages into a higher dimensional space, all curves have an equal number of landmarks, which allows simple direct PCA, see Vapnik (1995) and Scholkopf *et al.* (1998).

However for the Manduca sexta data set, it is not biologically evident which stage is the extra developmental stage. What is evident is that some of the stages for worms with 7 stages correspond to some of the stages for worms with 6 stages. Figure 4.7 shows the biological correspondence of stages between worms with 7 stages and worms with 6 stages.

The upper left panel shows the curve of one worm from the population of worms with 7 stages, shown as the red curve, and the curve of one worm from the population of worms with 6 stages, shown as the blue curve. The asterisks represent original developmental stages. Those stages connected by a dotted line correspond biologically. The table below shows this biological correspondence of the stages.

7 Stage Curve	6 Stage Curve
1	1
2	2
3	2
4	3
5	4
6	5
7	6

As can be seen from the figure and table, stages 1,4,5,6, and 7 of the worm with 7 stages correspond to stages 1,3,4,5, and 6 of the worm with 6 stages, i.e. are connected by a dotted line. But also stages 2 and 3 of the worm with 7 stages correspond with only stage 2 of the worm with 6 stages. The second blue asterisk has two dotted lines connected to it. If there was one stage for the 7 stage individuals that was an extra stage then there would be one red asterisk not connected to any blue asterisk by a dotted line. Since all the red asterisks are connected to a blue asterisk, where to add an extra stage for the blue curve is not obvious. Since it is not obvious where to add a landmark to the curve of the worm with 6 stages, both the data points representing the the curve of the worm with 6 stages and the data point representing the curve of the worm with 7 stages will be embedded into a higher dimensional space.

The worm with 7 stages will have one stage added to it's curve between it's second and third stage, while the worm with 6 stages will have two stages added to it's curve, one before and one after it's second stage. The added stages can be seen in the upper right panel of Figure 4.7. The added stages are represented by green circles. The stages are added as a linear interpolation of the midpoint between the appropriate adjacent stages. Now with the added stages both worms have 8 developmental stage landmarks, i.e data vectors of equal length.

Now both curves have the same number of landmarks. But the correspondence of

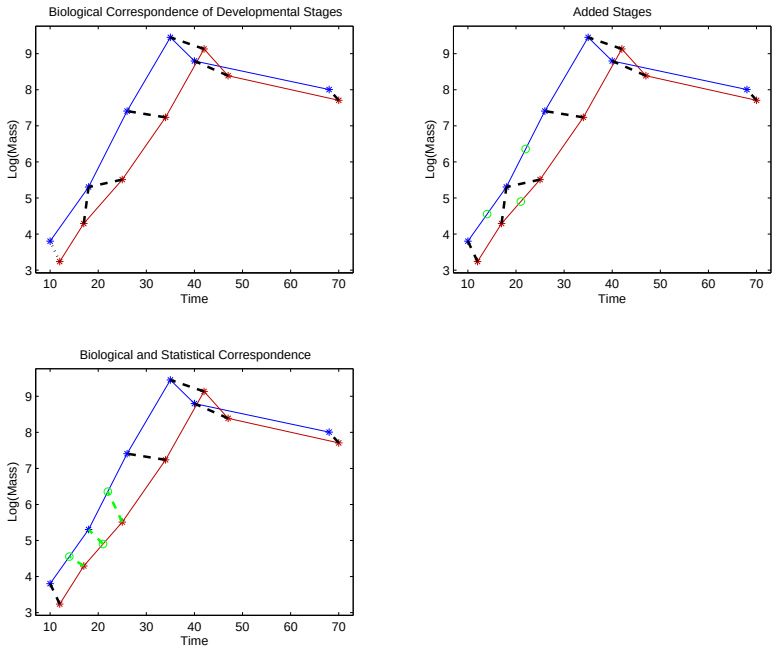

Figure 4.7: *Top left show the biological correspondence of a worm with 6 stage and a worm with 7 stages. Top right shows how an extra stage was added to the curve of the worm with 7 stages and 2 stages were added to the curve of the worm with 6 stages. Bottom left shows the correspondence between both developmental stage landmarks and pseudo-landmarks.*

landmarks still has to be established. The correspondence of the landmarks are shown in the bottom left panel of Figure 4.7 as dotted lines. The stages that had a one to one biological correspondence are considered corresponding landmarks. Remember that a biological correspondence was shown as a black dotted line. A one to one biological correspondence is when only one red asterisk is connected by a dotted line with only one blue asterisk. That is stages 1,4,5,6, and 7 of the worm with 7 stages corresponds to stages 1,3,4,5, and 6 of the worm with 6 stages. Also stage 2 of the worm with 7 stages corresponds to the stage added before the second stage of the worm with 6 stages. This correspondence is shown as a green dotted line connecting the first green circle along the blue line to the second red asterisk. This time the correspondence is shown as a green dotted line because it is not a true biological correspondence, but a

correspondence assumed for statistical purposes. Another statistical correspondence is between the added stage of the worm with 7 stages and the second stage of the worm with 6 stages. This correspondence is shown as a green dotted line between the second blue asterisk and the green circle along the red curve. Finally the third stage of the worm with 7 stages corresponds statistically to the stage added after stage 2 of the worm with 6 stages. This correspondence is shown as a green dotted line between the second green circle along the blue curve and the third red asterisk. Now both worms have 8 stages and all of the correspondences, represented by dotted lines, are established.

Figure 4.8 shows how a stage was added to the full population of worms with 7 stages on the right hand side. The upper right panel shows the the full population of 7 stage worms with only the original developmental stages, represented by red asterisks. The

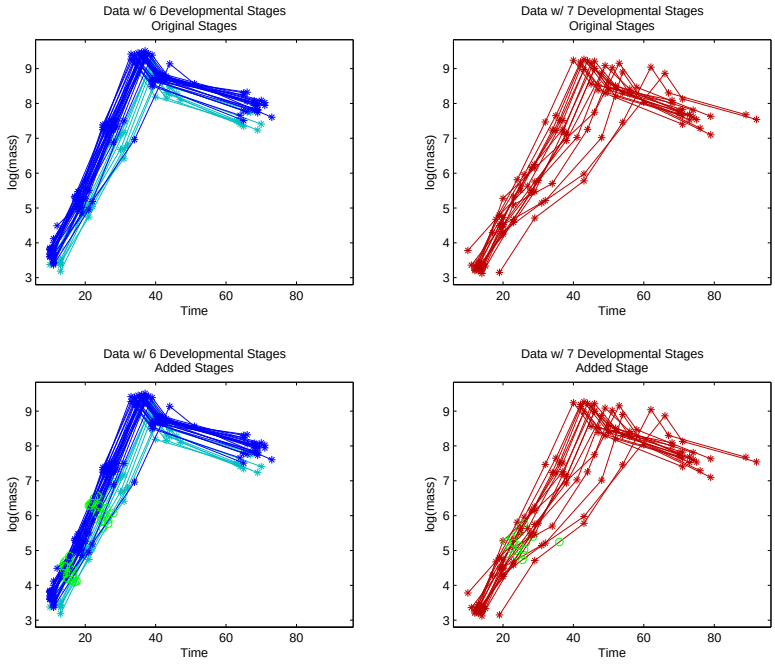

Figure 4.8: *Top left is Data with 6 stages and no stages added. Top right is data with 7 stages and no stages added. Bottom left is Data with 6 stages and 2 stages added. Bottom right is data with 7 stages and 1 stage added. Added stages are in green.*

bottom right panel shows the same curves again, only now with a stage added between the second and third stages of each curve. The added stages are again shown as green circles.

The left hand side of Figure 4.8, shows how two stages were added to the worms with 6 stages. The upper left panel shows the curves of the 6 stage worms with only the original developmental stages. The bottom left panel shows the full population of 6 stage worms with a stage added before and after the second developmental stage for each curve. The added stages are represented by green circles.

Now all individuals have 8 developmental stage landmarks along each curve. Each individual having the same number of developmental stage landmarks along the curve allows PCA to be performed on the full data set.

4.4 Results for the Manduca sexta Data Set

Section 4.2 described how PCA is performed on developmental stage landmark data. This section shows the results for landmark PCA of the Manduca sexta data set. First the results of PCA will be shown for the data on the original raw data scale without any standardization, shown in Section 4.4.1. These results are strongly driven by horizontal variation, because of differing horizontal and vertical scales. To better understand vertical variation a landmark PCA of the correlation matrix was performed, shown in Section 4.4.2. Both horizontal and vertical variation can be understood through landmark PCA of the correlation matrix, but a substantial drawback of this approach is that the covariance structure within mass and time is lost. An approach that gives a more useful decomposition of the variation of the data is standardizing the mass measurements using the trace of mass covariance matrix and the time measurements using the trace of the time covariance matrix, is shown in Section 4.4.3.

4.4.1 Original Raw Data Scale Landmark PCA

Figure 4.9 shows the results of landmark PCA on the original raw Manduca sexta data curves. The graphic has the same structure as Figure 4.6. The 2 left most columns are visualizations of the data using the mass vs time view while the 2 right most columns are visualizations of the data using the parallel coordinates view. Row 1 column 1 shows the original raw Manduca sexta data set with mass plotted versus time. The markers along each curve are the developmental stage landmarks. Each curve has eight developmental stage landmarks because of the extra stage correction described in Section 4.3. The curves for this Figure have the same color scheme as Figure 4.1.

Shown in the panel in row 1 column 2 is the mean of the data using the mass vs time view. The mean for this data set is an increasing line for stages 1 to 6. Where the peak of the mean curve is reached at developmental stage 6. After stage 6 the developmental stages decrease in mass while increasing in time. The third column in row 1 shows the data with the mean removed, using parallel coordinates. For this view, first the mass measurements of each stage are plotted on the left and the time measurements for each stage are plotted on the right. The dotted line separates mass from time. Notice that there are two of each number from 1-8 on the horizontal axis. This is because each developmental stage landmark has both a mass and time associated with it. Also it is evident from this plot that time, and the variation in time, is on a much larger scale than is mass. Time being on a larger scale than mass can be seen using the mass vs time view, in the top left panel, as well. Notice that the horizontal, i.e. time, axis ranges form 20 to 80 while the vertical, i.e. mass, axis ranges from 4 to 8.

Row 2 column 1 is a visualization of the data projected onto the PC 1 direction added to the mean, for ease of interpretation, and plotted as a function of mass vs time. The best way to understand variation using this view is to notice how the markers are spread for each developmental stage. Notice that the blue markers in general have a smaller time measurement for each stage while the red markers have a larger time measurement.

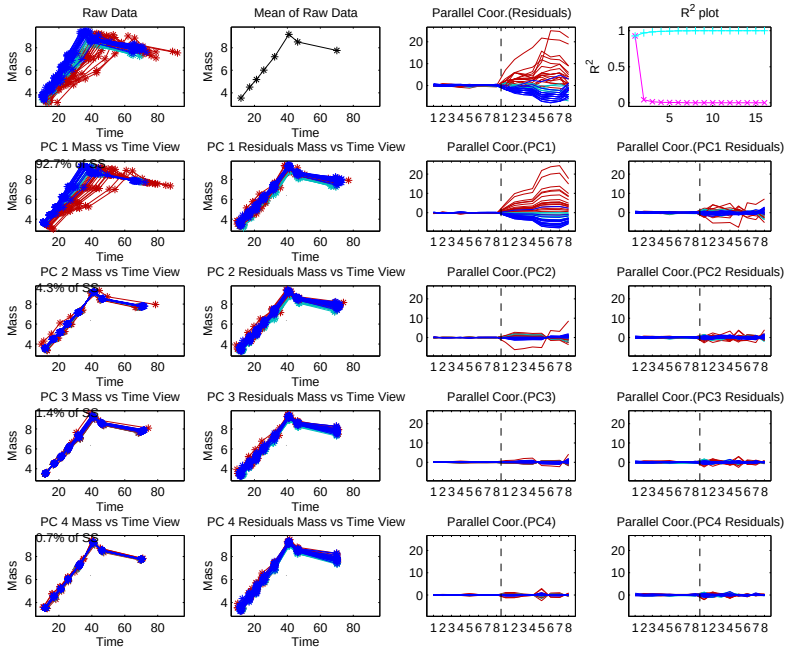

Figure 4.9: *PCA on originally scaled data, time variation is much larger than mass variation and therefore dominates PCA calculations. Mass vs time view of results shown in left 2 columns while parallel coordinates view of results shown in right 2 columns. Notice in parallel coordinates view how the time portion of curves fans out for higher developmental stages in the raw data and PC 1.*

Also notice the markers fall along a line for each developmental stage. This is because PCA only finds directions of linear variation. Also these curves are quite similar to the actual data curves, which is due to the fact that the PC 1 direction explains over 92% of the variation.

The power spectrum of the PCA is shown in row 1 column 4. This is a plot of the the percent of variation explained by PC j vs j. This plot shows that there is a drop after PC 1 and then the curve levels off for the rest of the PC's, i.e. there is a knee at PC 2. For this plot it is common to look for a knee in the curve and then the number of PC's to the left of this knee is viewed as the effective dimension of the data. In this case the data variation is explained almost completely by PC 1.

Row 2 column 2 shows the residuals, after PC 1 has been removed, added to the mean, again for interpretability purposes. Notice that these curves have little variation. This again shows that PC 1 explains most of the variation.

Shown in Row 2 column 3 is the projection of the data onto PC 1 in the parallel coordinates view but not added to the mean. The parallel coordinates can provide another way to understand the variation of PC 1. In this view it can be seen that PC 1 consists mostly of time variation by the left hand side, i.e. mass portion, of the curves being much less spread out than the right hand side, i.e. time portion, of the curves. Also the time portion of the curves fans out as the developmental stages increase. This tells us that the time variation increases as the developmental stages do. Also notice that the time portion of the blue curves are mostly below zero while the time portion of the red curves are mostly above zero. This correlates with what was seen using the mass vs time view when it was pointed out that most blue individuals have a lower time value for each developmental stage than the red individuals.

Row 2 column 4 shows the residuals, after PC 1 and with out the mean, using the parallel coordinates view. The curves in this panel are not spread out much again showing that PC 1 explains most of the variation.

Row 3 and Row 4 are the same as Row 2 except with PC 2 and PC 3 respectively, instead of PC 1. Again these PC directions mostly explain time variation which can best be seen in the parallel coordinates view by the right hand side of the curves being more spread out than the left hand side.

The mode of variation described by PC 2 is about how time at lower stages is different from time at higher stages. This can be seen in the parallel coordinates view by the time portion of the curve being low at first and then switching to high at the other end. This also can be seen in the mass vs time view by the fact that those individuals with lower time values at the lower developmental stages have higher time values at the higher developmental stages. PC 2 seems to be driven mostly by an extreme point. Notice

in the mass vs time view that there is one curve that is not grouped with the rest of the curves. That same curve is not grouped with the rest of the curves in the parallel coordinates view as well.

Because time is on such a larger scale compared to mass, PCA is going to find directions that reflect mostly time variation. Since mass information is likely to be relevant as well, an adjustment of the mass and time scales to be similar allows PCA to find directions reflecting both mass and time variation. A simple, commonly used approach to this is tried in Section 4.4.2. An improved approach appears in Section 4.4.3.

4.4.2 Landmark PCA on the Correlation Matrix

A common approach to adjusting the mass and time scales is standardizing each coordinate to have a variance of one. This is the familiar PCA on the correlation matrix.

Figure 4.10 shows the PCA results based on the correlation matrix. This graphic has the same structure as Figure 4.9, only in the 2 right most columns are now visualizations of the standardized data and projections of the standardized data using parallel coordinates. The first 2 columns show the data and projections as curves in the mass vs time view. The projections are recalibrated to be on the original scale, in mass vs time view, by multiplying the mass and time of each developmental stage by its original variance and then adding these to the mean vector.

The representation of the raw data using the mass vs time view, i.e. row 1 column 1, is the same as Figure 4.9. The mean is the same as well, shown in row 1 column 2. The mean residual data using the parallel coordinates view is shown as the standardized data in row 1 column 3. The scale is much smaller than originally and both mass and time variation can be seen. The left side and right side of the curve has about similar spread. But the curves no longer fan out as the developmental stages increase in the time coordinates, unlike in row 1 column 3 of Figure 4.9. This is because all coordinates are made to have a variance of 1 which is the same as making the spread of the curves

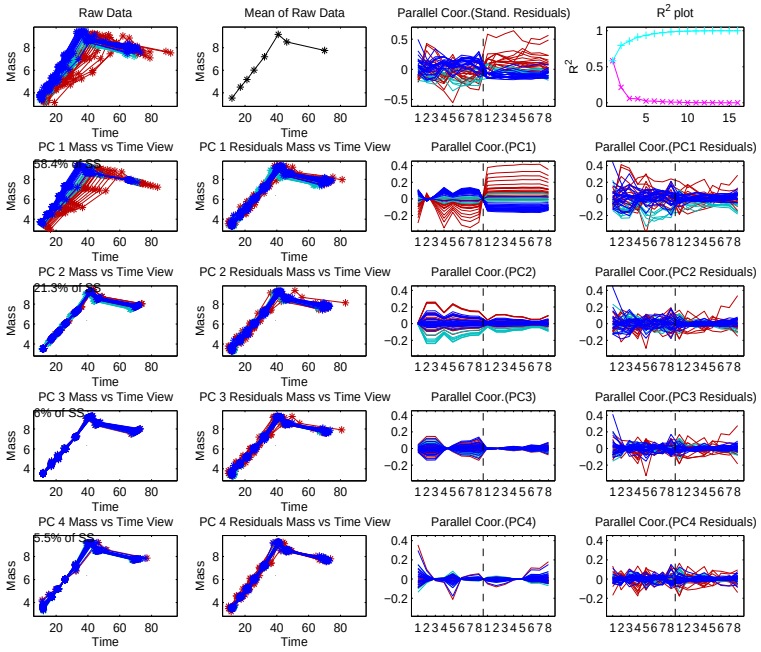

Figure 4.10: *PCA on Correlation Standardized Data, time and mass are on similar scales but covariance structure, within time coordinates, is lost. The loss of covariance structure can be seen best in parallel coordinates PC 1 projection by there no longer being a fanning of the curves as developmental stages get larger.*

similar for each coordinate. The power spectrum, shown in row 1 column 4, now has a knee at PC 3. So most of the variation of the data is explained by PC 1 and PC 2.

The mode of variation described by PC 1, shown in row 2, is interpreted as those individuals that have a smaller mass also take a longer time to reach that mass or those with a larger mass take less time to reach that mass. The color indicates that PC 1 separates those individuals with 6 stages from those with 7 stages. This is seen using both the mass vs time view, in row 2 column 1, and parallel coordinates view, in row 2 column 3. The red curves generally have lower mass values but higher time values while the blue curves have higher mass values but lower time values. PC 1 no longer has the fan structure in the time coordinates such as was evident when the data was analyzed on the original scale, which is best seen using the parallel coordinates view, i.e. row 2 column 3.

This is because all of the coordinates were made to have variance 1 so the variance does not increase as developmental stages increase. This tells us that the covariance structure within the time coordinates has been lost using this standardization technique.

The mode of variation described by PC 2, shown in row 3, is about individuals with a smaller mass taking less time to reach that mass and vice versa. PC 2 separates those individuals in the field group that have only 6 stages from the rest. This can best be seen using the parallel coordinates view, shown in row 3 column 3. The cyan curves have lower mass and time values than the other curves, i.e. the cyan curves are always below zero in the parallel coordinates view. The cyan curves form a different group than the other curves using the mass vs time view as well, shown in row 3 column 1.

4.4.3 Landmark PCA on Trace Standardized Data

There is a second option to make the scales of mass and time similar, but preserve the covariance structure within each of mass and time. This is done by not standardizing coordinate by coordinate but rather divide all mass measurements by one constant and all time measurements by another larger constant. This way both mass and time are on similar scales but the coordinates within mass and time are allowed to have different variances.

The method we chose to standardize mass and time separately, upon suggestion by Nancy Heckman, is to divide all mass measurements by the square root of the sum of all the mass coordinates variances and the time measurements by the square root of the sum of all the time coordinates variances. This is analogous to dividing all mass measurements by the square root of the trace of the covariance matrix formed by them and the time measurements by the square root of the trace of the covariance matrix formed by the time measurements. In this case the time coordinates have more variance than the mass coordinates, therefore the time coordinates will be divided by a larger constant. Dividing the time coordinates by a larger constant will adjust the mass and

time scales to be similar.

Figure 4.11 shows the PCA results from this standardization technique. Again this

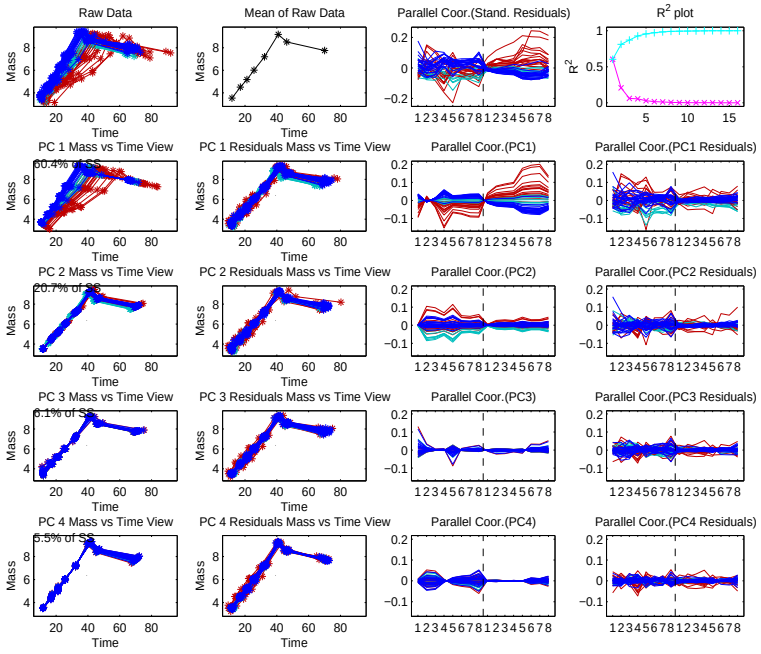

Figure 4.11: *PCA on Trace Standardized Data, time and mass are on the same scale and covariance structure is conserved. The curves fan out as developmental stages increase in the time coordinates*

graphic has the same structure as Figure 4.9 and Figure 4.10. The 2 right most columns are now the trace standardized data along with projections of the trace standardized data in the parallel coordinates view. The 2 left most columns show the data along with the projection in the mass vs time view. The projections are recalibrated to be on the original scale by multiplying by the square root of the trace and adding to the mean.

PC 1 and PC 2 yield the same general information, i.e. separating out the groups, as done by the correlation matrix PCA. Notice that rows 1-3 look similar to Figure 4.10. Also the power spectrum is similar to that of the correlation standardized data. But viewing PC 1 using parallel coordinates view, shown in row 2 column 3, the curves

again spread out more as the developmental stages increases, i.e. fan out in the time coordinates. So the covariance structure within time and mass coordinates is not lost. The same understanding of the group structure is gained as in the correlation matrix PCA but the covariance structure within mass and time is still visible.

CHAPTER 5

Hypothesis Test for Line Segment Slopes and Visualization

This chapter will focus on an analysis of the developmental stage landmark data first introduced in Section 4.1, see Figure 5.1. Several t-tests are preformed to test which sets of line segments, between landmarks, have equal slopes to better understand the slope structure of the data. Also a way to simultaneously visualize the results of the multiple t-tests in one graphic is presented. This visualization provides a way for each test to not only be interpreted individually, but for multiple tests to be interpreted as a group.

In Section 5.1 is an introduction to the data set being analyzed as well as a general overview of the t-test that is being used to test equality of slopes. A tool to visualize the results of multiple t-tests for this data set is presented in Section 5.2. Section 5.3 describes a method to combine data sets measured at different temperatures. The data sets are adjusted to minimize the effect temperature has on the slopes. The combined data set is then analyzed in Section 5.4 using the tools of the previous sections. A similar analysis involving the length of sets of line segments instead of slopes is discussed in Section 5.5.

5.1 Introduction to the Data Set and Hypothesis Test for Slope Equality

A truncated version of the Manduca sexta data set, i.e. the developmental stage landmark data introduced in Section 4.1, is focused on for this analysis, see Figure 5.1. The analysis is going to provide a tool to better understand the slope structure of the data. For this chapter, only the landmarks up to the peak of the curves, i.e. highest mass value, is considered. The landmarks of the data are represented by asterisks. Also it will be assumed that all of the curves start at the origin. The Manduca sexta data set is further truncated to only include the individuals from the field population. For Figure 5.1 only individuals measured at 20°C are shown.

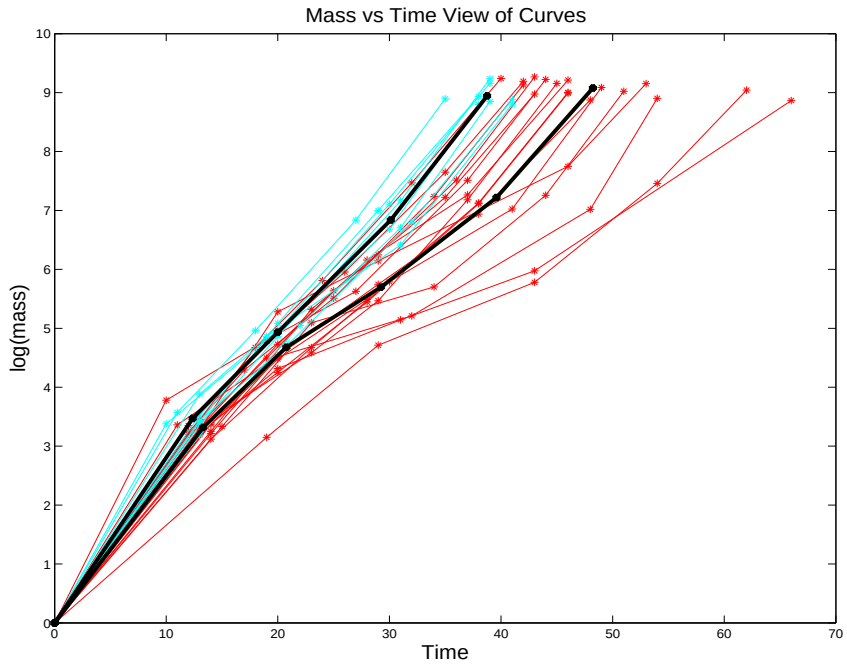

Figure 5.1: *Manduca sexta growth trajectories up to peak. Cyan are field caterpillars with 6 original stages and red are field caterpillars with 7 original stages. Asterisks represent developmental stages. Thick black lines are mean curves of the cyan and red curves.*

The thick black line on the left, corresponds to the mean curve for the population of individuals with 6 original developmental stages, i.e. cyan curves. While the thick black line on the right, corresponds to the mean curve of the population with 7 original developmental stages, i.e. red curves.

One question of biological interest is how the slopes of the line segments relate within each group. But also of interest is how the slope of line segment slopes relate between groups. One way to help understand these questions is by testing if the mean slope of a set of line segments is equal to the mean slope of another set of line segments. Here a set of line segments is the collection of line segments between adjacent landmarks for each group. For this data set there are a total of 9 sets of line segments. Each cyan curve has 4 line segments, while each red curve has 5 line segments. Therefore there will be a total of $\binom{9}{2} = 36$ comparisons, since we wish to test all pairs of sets of line segments.

In order to test if the mean slope of one set of line segments is equal to the mean slope of a second set of line segments a t-test is performed. A total of 36 t-tests are performed, with each test having a similar procedure. A detailed description of one test is explained in the rest of this section.

Without loss of generality, the test for comparing the mean slope of the first set of line segments of the cyan curves with the mean slope of the second set of line segments for the red curves is described. Figure 5.2 shows the data set being considered with the sets of line segments being compared highlighted in different colors, to aid in the explanation of sets of line segments and the t-test procedure. The first set of line segments of the cyan curves is highlighted in grey, while the second set of line segments of the red curves is highlighted in magenta.

The first step in the test is to calculate the slope of each grey and magenta line segment. Then the sample mean, $\overline{b}_g$, and sample variance, var_g, of the slopes of the grey line segments are calculated. Also the sample mean, $\overline{b}_m$, and sample variance, var_m, of the slopes of the magenta line segments are calculated. The number of grey line segments,

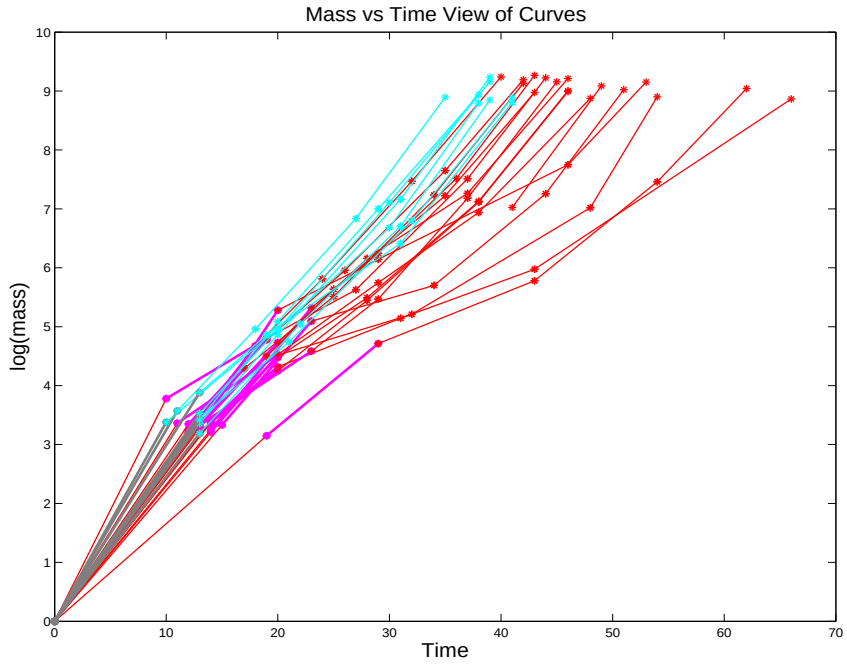

Figure 5.2: *Manduca sexta growth trajectories up to peak. Grey portions of each cyan curve make up the first set of lines segments of the cyan curves. Magenta portions of each red curve makes up the second set of line segments of the red curves.*

n_g, and the number of magenta line segments, n_m, are counted. Then a two sample t-test statistic is calculated. For the grey and magenta line segments this is

$$T = \frac{\bar{b}_g - \bar{b}_m}{\sqrt{\frac{var_g}{n_g} + \frac{var_m}{n_m}}}.$$

Then a p-value for T is calculated. The p-value is equal to

$$\text{p-value} = 2\mathbb{P}(|T| > t_{\min(n_1,n_2)}),$$

where $t_{\min(n_1,n_2)}$ follows a t-distribution with degrees of freedom equal to $\min(n_1, n_2)$. If the p-value is small enough then the hypothesis that the mean slopes are equal is rejected.

This testing procedure is performed for all 36 comparisons. Since there is such a large

number of comparisons a way to visualize the results of all the tests at once is helpful, which is shown in the next section.

5.2 Visualization of Results Of Multiple Slope Comparisons

The above section described how to test if the mean slope of two sets of line segments are equal. For the data set, a t-test for each pair of sets of line segments is performed. A way to simultaneously view the results of all of the t-tests in one graphic is shown in Figure 5.3. By viewing all of the results using one graphic, the results can be understood

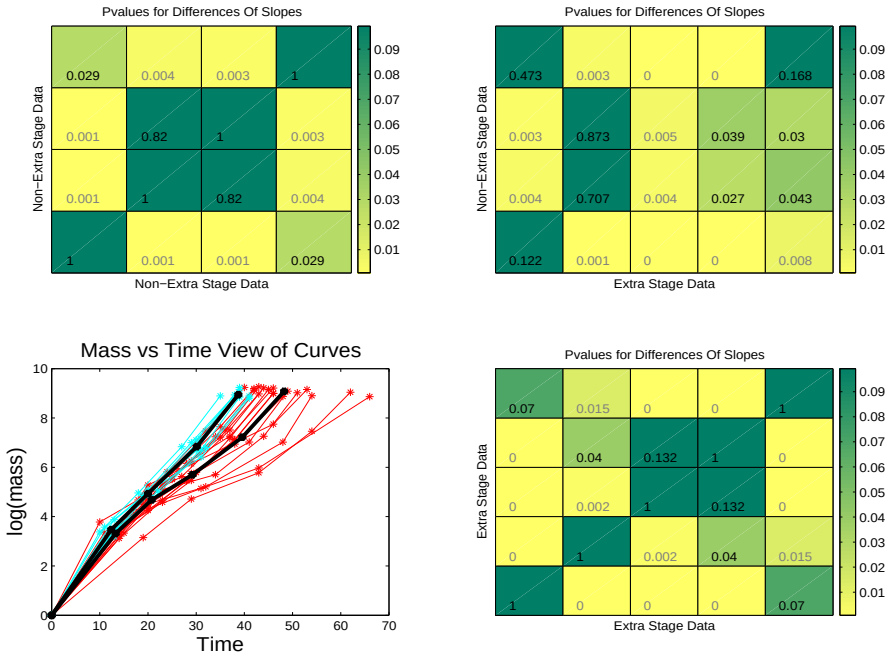

Figure 5.3: *Bottom left displays data along with mean curves of cyan and red curves added. Matrix in top left shows results of t-tests for within cyan group. Matrix in bottom right shows results of t-tests for within red group. Matrix in top right displays results of t-test between cyan and red groups.*

individually or as a group. Also the display allows for interpretation within groups of

curves, i.e. colors, as well as between groups of curves.

We will start by analyzing the slope structure within the cyan curves. The lower left of Figure 5.3 is the same as Figure 5.1. One question is if the mean slope of the first set of line segments of the cyan curves is equal to the mean slope of the second set of line segments of the cyan curves. A t-test of the type described in Section 5.1 is performed. Notice that in the figure in the bottom left that the mean slope of the first set of line segments looks visually different than the mean slope of the second set of line segments. Therefore a low p-value is expected.

The matrix in the top left displays the p-value of the t-test of mean slope equality for every pair of sets of line segments of the cyan curves. The specific p-value of the test of the first set of line segments vs the second set of line segments is shown in the box that is in the bottom row second from the left. The p-value for this test is 0.001. This p-value matches the intuition seen from the picture in the bottom left. The box is also colored according to the p-value. The coloring, shown in the color bar to the right of the matrix, goes from yellow to green, with yellow representing a low p-value and green representing a high p-value. Therefore those that are yellow are considered to have significantly different mean slopes.

The coloring scheme above is chosen based on 0.05 being the level of significance. Any box with a p-value of 0.05 is colored 50% yellow and 50% green. But 36 tests are being performed on one data set. Therefore a Bonferroni correction level of significance could be used as well, i.e. $\frac{0.05}{36}$. The p-value of all 36 tests are displayed in one of the matrices, so the p-value can be compared to $\frac{0.05}{36}$ to see if the mean slopes are significantly different based on the Bonferroni correction. The coloring scheme can be adjusted as well, so that any box with a p-value of $\frac{0.05}{36}$ is colored 50% yellow and 50% green.

The bottom row displays the results of the comparison of the first set of line segments to the rest of the line segments of the cyan curves. Figure 5.4 is provided to aid in the understanding of how the results are displayed in the matrix in the upper left. This figure

highlights the first mean line segment of the cyan curves in grey. Notice also that a grey box has been drawn around the bottom row of the matrix in the upper left. This box indicates that the bottom row of this matrix corresponds to the results of the first set of line segments vs the rest of the line segments of the cyan curves. Also notice that the horizontal and vertical axis of the matrix is labeled as non-extra stage data, which tells us that the results shown in this matrix only pertain to the cyan curves. The left most box in the bottom row shows the p-value for the t-test of the first set of line segments vs itself. This test is going to say that the slopes are always equal, so therefore the p-value is 1. The box second to the left is the p-value of the comparison of the first set of line segments to the second set of line segments. This test yielded a p-value of 0.001. The box third to the left is the test of the first set of line segments to the third set of line segments, with a p-value of 0.001 as well, etc.

The mean curve of the cyan curves, shown in the bottom left, is also colored to help illustrate which box matches with which set of line segments. The first mean line segment is colored grey. All of the results include this set of line segments. The next mean line segment is colored the same yellow as the box second from the left. This indicates that the first set of line segments is being compared to the second set of line segments. The third mean line segment is colored the same yellow as the third box, etc.

The matrix and colored line segments tells us that the mean slope of the first set of line segments is different than the mean slopes of the last 3 sets of line segments. Notice in the bottom left of the figure that the mean slope of the first set of line segments is larger than the other 3. The difference between the mean slope of the first set of line segments and the mean slopes of the second and third set of line segments is about the same, which is reflected in that both tests have a p-value of 0.001. While the difference between the mean slope of the first set of line segments and the mean slope of the last set of line segments is less, which is reflected by a larger p-value equal to 0.029.

The second row from the bottom is the results of the tests of the second line segment

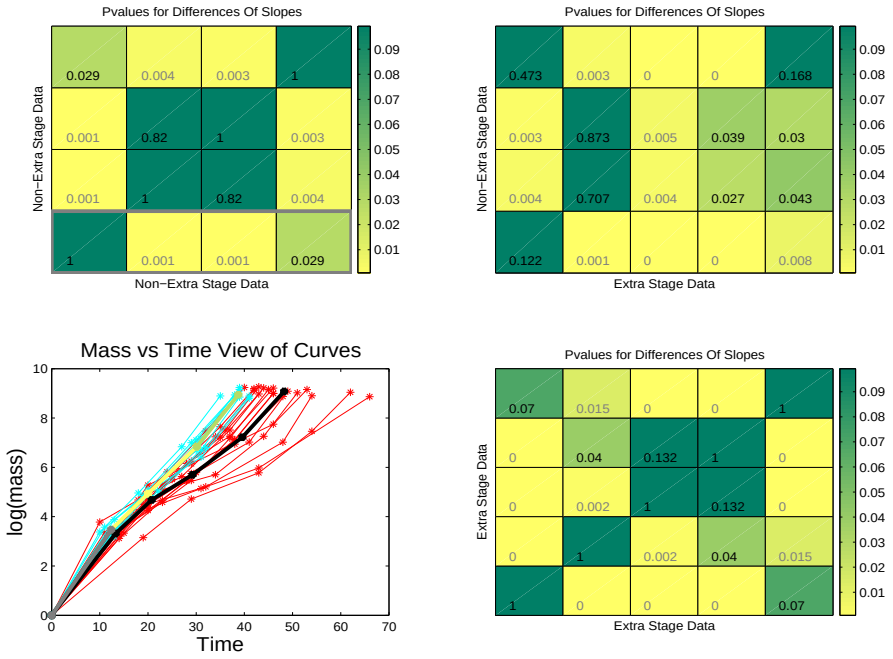

Figure 5.4: *Bottom left displays data along with mean curves of cyan and red curves added. Matrix in top left shows results of t-tests for within cyan group. Matrix in bottom right shows results of t-tests for within red group. Matrix in top right displays results of t-test between cyan and red groups. First mean line segment highlighted in grey to match box in upper left matrix. Grey indicates that the first set of line segments is being compared to rest of cyan set of line segments, with results shown in bottom row of upper left matrix.*

vs the rest of the line segments of the cyan curves. The box most to the left has the same p-value and color as the box second to the left in the bottom row. This is because the same two sets of line segments are being compared. The matrix is symmetric by the same reasoning. Also the matrix is always going to have 1 along the diagonal from bottom left to upper right. Because if the set of line segments is compared to itself, then the mean slopes are always equal. The third row from the bottom is the results of the third set of line segments vs the rest of the sets of line segments of the cyan curves etc.

From investigating this matrix, some statements can be made about the slope structure of the cyan curves. The main lesson about the slope structure is that the mean slopes are all significanlty different except for the second and third set of line segments.

Therefore the mean growth rate is similar, in the sense of no statistically significant difference, for the second and third segments.

The mean slope of the second and third line segments being equal gives evidence that the interpolation procedure performed in Section 4.3 is appropriate. This is since landmarks were added by linear interpolation between these three landmarks. Since the slopes between these three landmarks is the same, this says that adding a landmark by linear interpolation is appropriate because the landmarks are linear related.

These facts are seen in the bottom left of the figure as well. The mean slope of the first set of line segments is different than the other mean slopes. Also the last set of line segments appears to be different than the rest. While the mean slope of second and third line segments appear to be similar to each other but different than the rest.

The matrix in the bottom right of Figure 5.3 shows the corresponding test results for the comparisons of sets of line segments of the red curves. The matrix has the same structure as the one in the upper left, but is a 5×5 matrix. This is because there are now 5 sets of line segments to be compared rather than 4. Again there is always 1 along the diagonal from bottom left to upper right. Also the matrix is symmetric. Again the p-values for each test are displayed and the color scheme is the same as for the upper left matrix.

The mean slopes of all of the line segments are significantly different except for the third and fourth set of line segments. Also the first set of line segments may have a mean slope the same as the fifth set of line segments. For the red curves the growth rate is similar for the third and fourth line segments.

These results are again supported by viewing the mean curve of the red curves in the bottom left of the figure. The third and fourth sets of line segments have mean slopes which are most similar. While the mean slope of the first and last are similar to each other but quite different from the others. Also the mean slope of the second set of line segments appears different from the rest.

The results of the comparisons between groups is shown in the matrix in the upper right corner of Figure 5.3. This is a 4×5 matrix, and therefore not symmetric nor does it have ones along the diagonal. This is because the matrix displays the results for the t-tests between groups. Notice that the horizontal axis is labeled extra stage data and the vertical axis is labeled non-extra stage data. The vertical axis corresponds to sets of line segments from the cyan curves, while the horizontal axis corresponds to sets of line segments of the red curves. Therefore there are 4 boxes vertically since there are 4 sets of line segments associated with the cyan curves. Also there are 5 boxes horizontally since there are 5 sets of line segments associated with the red curves.

The box in the lower left hand corner displays the p-value for the test between the mean slope of the first set of line segments of the cyan curves vs the mean slope of the first set of line segments of red curves. The rest of the bottom row shows the p-values for the tests of the first set of line segments of the cyan curves vs the rest of the sets of line segments of the red curves. While the rest of the first column displays the p-values of the tests of the first set of line segments of the red curves vs the rest of the sets of line segments of the cyan curves. The color scheme is the same as the matrices in the upper left and lower right.

This matrix displays how the slope structure between the cyan curves and red curves is related. The first two sets of line segments for both the red and cyan curves have similar mean slopes. The third set of line segments of the cyan curves has a similar mean slope to the second set of line segments for both curves. While the third and fourth set of line segments of the red curves has a mean slope that is different. The last set of line segments for both curves have a similar mean slope to each other. The mean relative growth rate for both sets of curves is similar at the beginning, i.e. the first and second set of line segments, and at the end, i.e. the last set of line segments. But the mean growth rate is different in the middle, i.e. third of cyan is different than third and fourth of red. This difference between the groups fits biologically, since this is the point where

it is believed that a developmental stage is added for the red curves.

A similar analysis is performed for individuals measured at 25°C. The results of the mean slope comparisons, shown in a graphic similar to Figure 5.3, which can be viewed by opening the file Ttest25 at Gaydos (2007) in the folder T-test.

5.3 Temperature Adjustment

The above sections analyzed data consisting only of individuals that were measured at 20°C, but there are also individuals measured at 25°C. Figure 5.5 shows the growth rate curves for both temperatures up to the peak mass value. Individuals with 6 and 7 original developmental stages, represented by cyan and red curves respectively, were measured at both temperatures. The solid cyan and red curves are the individuals measured at 20°C, while the dashed cyan and red curves are the individuals measured at 25°C. This graphic shows that temperature has a significant effect on the overall slope of the curves. But if the overall slope difference due to temperature differences is adjusted for, then how do the slopes of the cyan and red curves relate?

The objective of the adjustment is for the solid and dashed cyan curves to lie on top of each other. Also the dotted and dashed red curves should lie on top of each other. One way to do this adjustment is to multiply all of mass measurements of the cyan individuals measured at 20°C by a constant. Also the time measurements of these individuals are multiplied by another constant. But the constants are chosen, such that the peak mass and time values of the mean curve of the cyan individuals measured at 20°C is shifted to be exactly on top of the peak mass and time values of the mean curve of the cyan individuals measured at 25°C. Then a similar adjustment is performed on the red curves.

An adjustment with overall means, i.e. both colors included, was performed and similar results were obtained. The results of this adjustment can be seen at Gaydos (2007) in the folder overallMeanAdjustment. The files of interest are NonExtra, Extra, and All, which shows the cyan curves, red curves and all curves adjusted by the overall means.

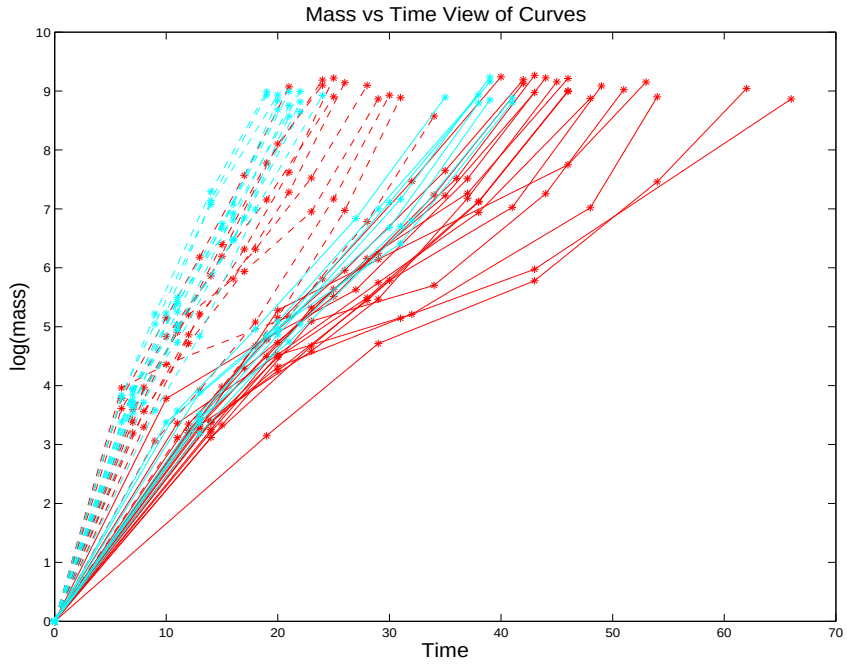

Figure 5.5: *Manduca sexta growth trajectories up to peak are shown. Cyan are field caterpillars with 6 original stages and red are field caterpillars with 7 original stages. Asterisks represent developmental stages. Solid lines are individuals measured at 20° C, while dashed lines are individuals measured at 25° C.*

It was decided to perform the adjustment color by color, since that way within color temperature effect would be removed. This would not necessarily be true with an overall mean adjustment. Also it was decided to multiply the landmarks by a constant, since we are dealing with slope and the curves are all assumed to start at the origin. Therefore the multiplication is easily interpreted in the mass vs time view, as a rotation and enlarging or shrinking of the curves. If mass and time of the landmarks were subtracted by a constants then the curves would no longer start at zero. Another option is to subtract a constant from all of the slopes of line segments, but then a separate adjustment is needed when length of line segments is considered, see Section 5.5. The multiplication adjustment provides a way to adjust both the slope and length of the line segments at once.

The data curves after this adjustment is performed are shown in Figure 5.6. The solid and dashed cyan curves lie roughly in the same area, as do the solid and dashed red curves. Also notice that the horizontal axis in Figure 5.5 includes values 0 to 70, while the axis in Figure 5.6 includes only values from 0 to 40. This is because the time values of the solid lines landmarks were multiplied by a constant which shifted them to the left. Also the mass values were multiplied by a constant. But the effect of the multiplication on mass was much less than for the time values, since the original mass values of the solid and dashed lines are similar.

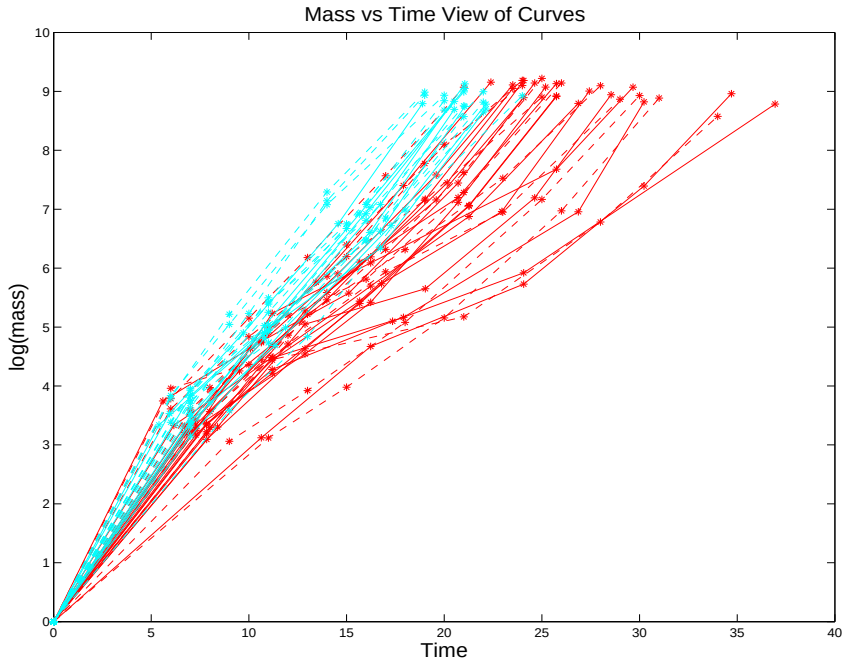

Figure 5.6: *Manduca sexta growth trajectories adjusted for temperature effect on overall slope are shown. Cyan are field caterpillars with 6 original stages and red are field caterpillars with 7 original stages. Asterisks represent developmental stages. Solid lines are individuals measured at 20° C, while dashed lines are individuals measured at 25° C.*

Once this adjustment is completed a question arises as to how do the slope structures between the 20 and 25°C individuals relate? Ideally if we wish to compare between red and cyan curves for both temperatures, then the two groups of cyan curves should have

similar slope structures. Also the same should be true of the red curves. The tools described in earlier sections of this chapter will be used to analyze if the slope structures between temperatures within each color are similar.

First we will restrict our analysis to the cyan curves. The same t-test procedure will be performed but the between group comparison is now between temperatures, i.e. comparing dotted and dashed curves, rather than differing number of stages. The results are shown in Figure 5.7. The bottom left of the figure displays the adjusted curves and their group means. Notice that the last landmarks of the two mean curves lie directly on top of each other, due to the adjustment.

The upper left and lower right are the comparisons for within groups, i.e. line types (temperatures). Notice that the lower right matrix of Figure 5.7 and the upper left of Figure 5.3 are the same. This is because although the slopes of the individuals have changed, all of the slopes have only been multiplied by the same constant. Therefore by the nature of the t-test all of the p-values for the tests are the exact same.

The matrix in the upper left is the comparison of slope structure within the dashed, i.e. 25°, curves. The dashed curves were never multiplied by a constant in the combination of the two line types. Therefore this matrix is the same as in the upper left of Ttest25 at Gaydos (2007).

The matrix in the upper right is the t-test p-values for between group comparisons. If the adjustment is good then the two temperatures should have similar slope structures, i.e. sets of line segments in the same order along the curves will have the same mean slope. A way to tell if they have similar slope structure is to view the p-values along the diagonal of the matrix in the upper right, from lower left to upper right. If the adjustment is good then the p-values will be large along the diagonal, i.e green boxes. In this case the adjustment works well, except that the last set of line segment mean slopes are significantly different. The mean curves in the bottom left visually display this result. The two curves are relatively parallel, except the last line segments. The last line

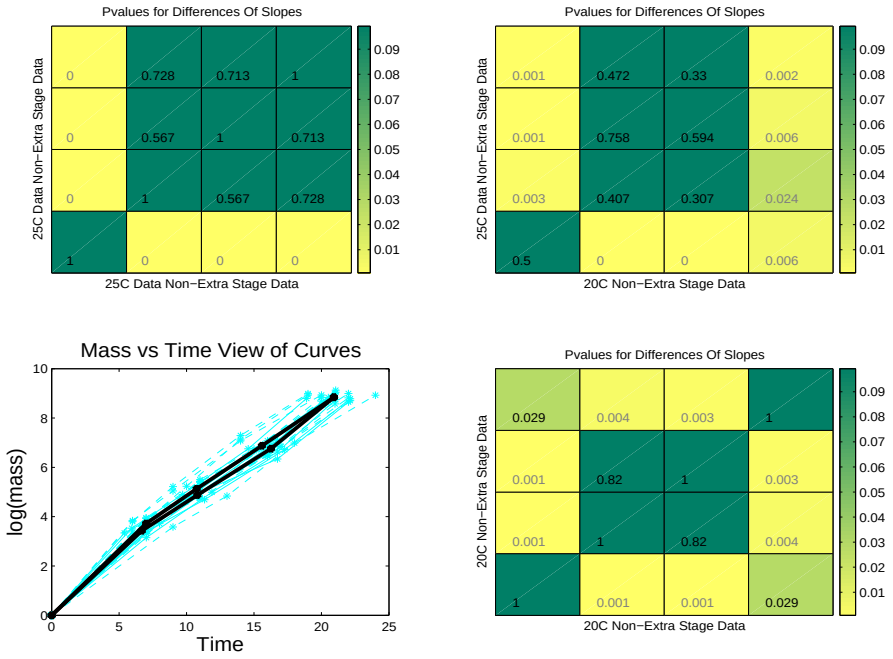

Figure 5.7: *Bottom left displays data along with mean curves of solid and dashed cyan curves added. Matrix in top left shows results of t-tests for within dashed curves group. Matrix in bottom right shows results of t-tests for within solid curves group. Matrix in top right displays results of t-test between solid and dashed cyan groups. Notice that diagonal from bottom right to upper left is green boxes with high p-values, except for top right box. This indicates that the adjust is good.*

segments start at different points, but end at the same point. These mean curves both end at the same point due to the adjustment performed earlier in this section. Forcing the mean curves to end at the same point may be the cause of the last sets of line segments having differing slopes.

A similar analysis for the red curves was performed and a similar result was rendered. Again only the last set of line segments did not have equal mean slopes, i.e. the top right box in the matrix was yellow with a small p-value. The analysis can be viewed by opening the file Extra at Gaydos (2007) in the folder groupadjusted.

By using the tools of this section it is judged that the adjustment works well. Therefore we will proceed with the adjusted data. All cyan curves are considered to be one

group and all red curves are considered the other group. Now insight into the structure of all the cyan curves and red curves can be gained. Also how the slope structures are related between the red and cyan curves can be investigated. The next section shows the results of the t-tests for this combined data set.

5.4 Multiple Slope Comparisons for Multiple Temperature Data

A way for the data of individuals measured at 20 and 25°C to be combined is described in Section 5.3. Now we wish to analyze the slope structure within the cyan curves and also within the red curves for this combined data set, with the goal of increased statistical power. Also the question of how do the slope structures between the cyan and red curves relate is analyzed. These questions will be investigated using the t-tests described in Section 5.1 and the visualization tool described in Section 5.2.

The results of the analysis of the adjusted data set, with both temperatures included, are shown in Figure 5.8. The adjusted data along with the mean curves for the cyan and red curves are shown in the bottom left of the figure. The upper left matrix displays the results for the t-tests of sets of line segments within the cyan curves. The lower right matrix displays results of the t-tests for sets of line segments within the red curves. While the upper right matrix displays the results of the t-tests between the cyan and red curves' sets of line segments.

In viewing the upper left matrix, it can be seen that within the cyan curves the second and third sets of line segments have similar mean slopes. It also suggests that the last set of line segments has a similar mean slope to the second set of line segments. But due to the fact that the solid and dashed curves had different mean slopes for the last set of line segments after adjustment, this result is questionable. All other line segments have differing mean slopes. The mean curve of the cyan curves, in the bottom left corner, appears to have slopes which are equal after the first line segment. The difference in

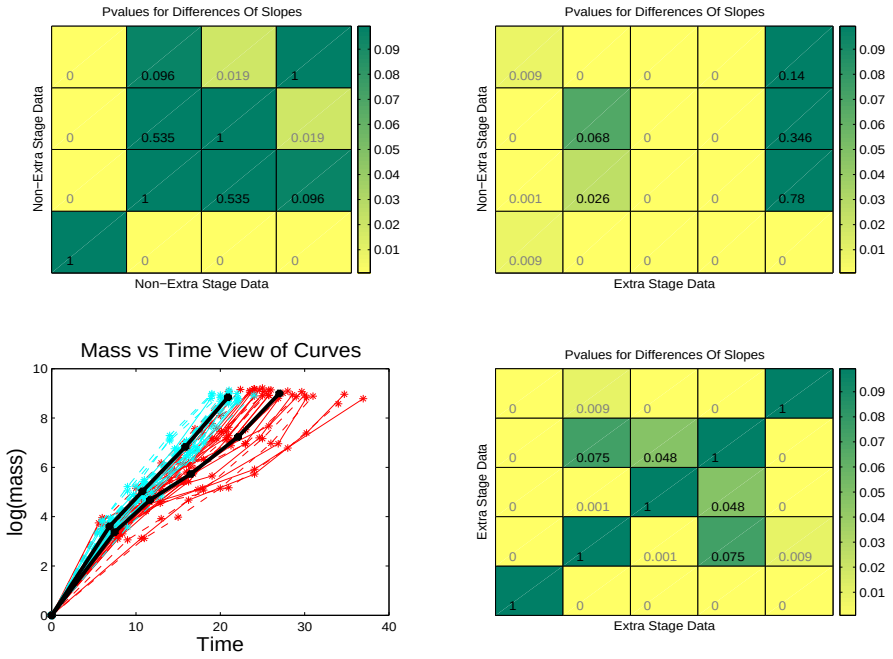

Figure 5.8: *Bottom left displays adjusted data of 20 and 25° along with mean curves of cyan and red curves added. Matrix in top left shows results of t-tests for within cyan group. Matrix in bottom right shows results of t-tests for within red group. Matrix in top right displays results of t-test between cyan and red groups.*

slopes of the third and fourth line segment does not appear visually, but does appear in the results of the t-test.

This matrix is similar to the separate 20 and 25° analysis in that the second and third set of line segments have slopes which are not significantly different. Also the first set of line segments has a mean slope significantly different than the rest. The rest of the tests for the combined data seems to compromise between the two separate temperature analyses. For instance the second and fourth set of line segments had significantly different mean slopes for 20° but not for 25°, while the combined data has a p-value which indicates borderline significance. Similar tests with differing results for 20 and 25°, yield test test results for the combined data set which are a compromise of these 2 results. This is expected since they have about an equal number of cyan curves for both temperatures.

Temperature 20° has 8 cyan curves while 25° has 12 cyan curves.

The matrix in the bottom right analyzes the slope structure within the red curves. The second and fourth sets of line segments, as well as the third and fourth may have similar mean slopes. The p-values for these comparisons are close to the border for significance, and therefore it is hard to distinguish if the mean slopes are similar or not. But the other p-values are much lower and therefore these mean slopes are much more similar than the others. The mean curve of the red curves, in the bottom left corner, reinforces these results. The middle three line segment slopes are visually much more similar to each other, than to the first and last. The beginning and end of this mean curve has a mean slope steeper than in the middle.

The combined data has mean slopes which are borderline significantly different for the third and fourth set of line segments. While the separate temperature analyses yield means slopes which were not significantly different for the third and fourth set of line segments. The rest of the matrix yields test results much similar to results of 20° than 25°. This is expected since the 20° has 17 red curves while the 25° has only 11 red curves.

How the slope structures between groups relate is investigated by viewing the matrix in the upper right. This matrix suggests that the second, third and fourth line segments of the cyan curves have similar mean slopes to the last line segments of the red curves. But again due to the fact that the adjustment did not cause the last set of line segments of the solid and dashed lines to have equal mean slopes, this result is questionable. Also the mean slopes of the second and third sets of line segments of the cyan curves could be similar to the mean slope of the second set of line segments of the red curves. Although these p-values are borderline to be considered to have significantly differing slopes, the p-values are much larger than the rest of the p-values, except for the ones mentioned earlier. Again the slope structures of the mean curves can be investigated visually in the bottom left corner. The second line segment for the mean curve on the right appears to be parallel to the second and third line segments of the mean on the left.

For the between color analysis the statistical power seems to increase with the combined data set over the separate temperature analyses. Overall the p-values are lower except for the last column. Also the first sets of line segments have significantly different mean slopes for the combined data set, which was not true for either of the separate temperature analyses. Also the combined data results matrix has more p-values close to 0 than either of the other separate temperature analyses.

5.5 Multiple Segment Length Comparisons

A similar analysis can be performed on the lengths of the line segments. But instead of comparing means of slopes of sets of line segments the mean lengths of sets of line segments are compared. The same visualization tool is used. For this analysis go to the website Gaydos (2007). The folder Ttest contains another folder Lentgh Test. In the folder LengthTest are the files which show these results.

This analysis has one further complication, the fact that the time and mass scales are very different. Notice in Figure 5.1 that the horizontal axis ranges form 0 to 60 while the vertical axis ranges from 0 to 10. If the original scales of mass and time are used then the portion of length attributed to time is going to dominate the length calculation. So the analysis will be more strongly influenced by time differences than mass differences. Therefore a way to standardize the scales of mass and time is necessary. To see results of the length test on unstandardized data, i.e. raw scale, see the files 20unstandardized and 25unstandardized for 20 and 25 degrees respectively.

A similar standardization to the one performed for the PCA analysis of landmark data, see Section 4.4.3, is done. The standardization is performed by dividing the mass measurements by the square root of the trace of a matrix measuring sums of squares of mass measurements. While the time measurements are divided by the square root of the trace of a matrix measuring the sums of squares of time measurements. In the PCA analysis this was the covariance matrix of mass measurements and time measurements

respectively. The covariance matrix measures sums of squares from a mean curve. For this analysis the sums of squares of the mean curve must be taken into account. Therefore the mass measurements are divided by the trace of a matrix measuring the total sums of squares from zero of the mass measurements and the time measurements are divided by the trace of a matrix measuring the total sums of squares from zero of time measurements. To see results on the standardized data see files 20standardized and 25standardized. Also at the website in this same folder is the results for when the combined data is analyzed, i.e. 20 and 25 degree data is included. To see the results open the file Combined. Only the standardized data results are shown for this case.

CHAPTER 6

Mathematical Background

In chapter 3 there is a discussion of estimating subspaces. The distance between the estimated subspace, $\hat{N}$, and the true subspace, N, provides a measure of how accurate the estimate is and thus provides the foundation of a mathematical statistical analysis. But the idea of distance between subspaces is not straightforward. In this chapter several metrics from the literature will be defined and then compared. These metrics will define what is meant by distance between two subspaces. But before the metrics can be defined the idea of canonical angles must be introduced. Section 6.1 contains an introduction to the geometry of canonical angles, described through a series of low dimensional toy examples. Section 6.2 consists of an introduction to Canonical Angles and their relation to the common statistical procedure Canonical Correlation Analysis (CCA). Section 6.3 contains a discussion of a metric from the literature known as the *gap metric*. Section 6.4 has a discussion of a second metric from the literature known as the *Euclidean sine metric*.

6.1 Geometric Introduction to Canonical Angles

This section gives a geometric introduction to canonical angles, since all of the metrics discussed later are based on canonical angles. For a more detailed discussion of canonical angles see Stewart and Sun (1990) or Kato (1966). The material is presented through a series of low dimensional toy examples, so that the written text can be accompanied by graphics that are easily visualized. The first toy example is the definition of a canonical

angle between 2 lines in a 2 dimensional Euclidean space. The canonical angle is the angle between these lines. The second toy example is the canonical angles between two planes in a 3 dimensional Euclidean space. Finally a third toy example is the canonical angles between a line and a plane in a 3 dimensional Euclidean space. This is presented after the 2 plane example because additional consideration is needed to handle the unequal dimensions of the subspaces.

Canonical angles define how one subspace is related to another. In the case of 2 lines in a 2 dimensional Euclidean space, the angle between the lines is the canonical angle. These lines are 1 dimensional subspaces in a 2 dimensional space. So the angle between these lines is between 0 and 90 degrees. An angle of 0° indicates they are the same subspaces and 90° indicates they are orthogonal. Figure 6.1 shows several different canonical angles between 1 dimensional subspaces. The subspaces are represented by the lines. The horizontal line parallel to the x-axis represents the subspace N, in following with the notation from above. While the other line represents the subspace $\hat{N}$. The top left panel shows two 1-d subspaces with a canonical angle of 0°. Notice that the two lines are directly on top of each other, i.e. they are the same. The top right panel shows 2 1-d subspaces with a canonical angle of 30°. The first subspace is the horizontal line while the other subspace is a line 30° degrees from it. The bottom left panel shows 2 1-d subspaces with a canonical angle of 60°. The bottom right shows 2 1-d subspaces with a canonical angle of 90°. Notice that as the canonical angle gets larger the two subspaces are in some sense farther apart.

One way to think about the distance between the subspaces is by the angle between them, but a second way is via a right triangle formed between them. This right triangle representation is important for the metrics below as well as to gain a general understanding of canonical angles. This triangle is defined completely by the angle. Since their are actually infinitely many triangles the convention is to use triangles that have a hypotenuse of 1. This way the cosine of the angle is exactly equal to the Euclidean inner

product. To form the triangle between the subspaces, a point that is distance 1 from the origin is found along one of the lines. In Figure 6.1 we will use the line, $\hat{N}$, which is not parallel to the x-axis to find a point at distance 1 from the origin. Finding this point is exactly the same as finding an orthonormal basis of the 1-d subspace $\hat{N}$. This point is then projected onto the other subspace. The triangle is then formed by the origin, the basis point of $\hat{N}$, and its projection onto N.

Let us focus on the top right corner panel of Figure 6.1, the triangle is formed by the magenta, cyan, and black line segments. The black line is the hypotenuse of the triangle,

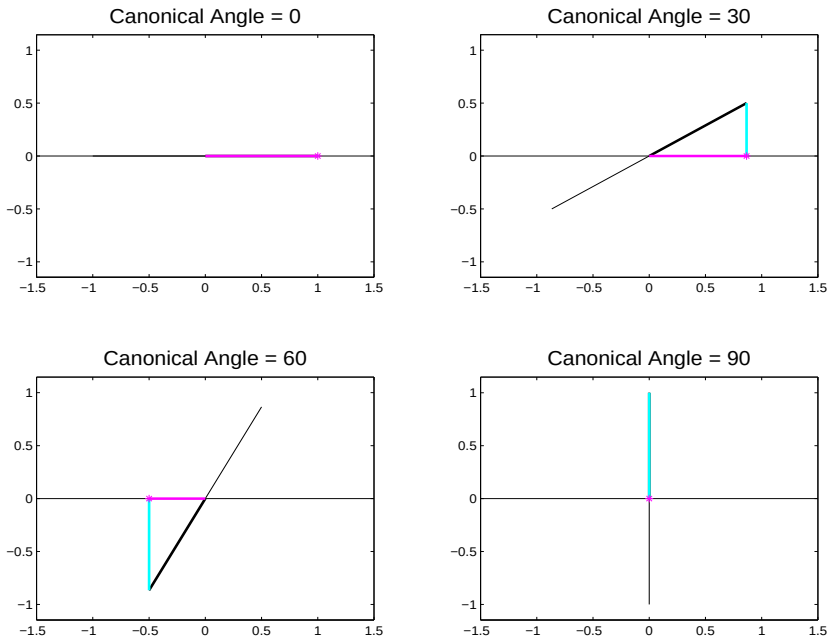

Figure 6.1: *Upper left shows canonical angle of 0 between 2 1-d subspaces, upper right shows canonical angle of 30, bottom left shows canonical angle of 60, and bottom right shows canonical angle of 90. Magenta line gets smaller while cyan line gets larger as canonical angles get larger.*

i.e. the line segment from the origin to the basis point. The magenta line is the line segment from the origin to the projection of the basis point. Finally the cyan line is the line segment between the basis point and its projection onto N. The magenta line has

length equal to the cosine of the canonical angle while the cyan line has length equal to the sine of the canonical angle. From Figure 6.1 note that the magenta line's length gets smaller as the canonical angle gets larger while the cyan line's length gets larger. These are all right triangles with a hypotenuse of 1. By the Pythagorean theorem the sum of the squared lengths of the magenta and cyan lines is 1. So the magenta and cyan lines' lengths get larger and smaller deterministically based on each other.

The top left shows the case when the canonical angle is $0°$, i.e the subspaces are the same. In this case the magenta line length is 1, i.e the cyan line length is 0. This tells us that the basis point is equal to its projection. This only happens if the basis point is in both spaces. The bottom right is the opposite. In this case the cyan line length is 1 and the magenta line length is 0. So here the subspaces are orthogonal. For the case of the canonical angle of $30°$, the magenta line is smaller than in the top left but larger than in the other 2 panels. Also the subspaces are farther apart than in the top left but closer than in the rest of the panels. In the bottom left panel the magenta line is larger than in the bottom right panel but smaller than the magenta line in the rest. This corresponds to the subspaces being closer in the bottom left only when compared with the bottom right panel.

Let us move on to the more complicated case of 2 planes in a 3 dimensional Euclidean space. Now it is not obvious where to form the triangles at, as well as what the canonical angles are between subspaces, since the subspaces have infinitely many angles between them. Figure 6.2 shows an example of 2 planes in a 3-dimensional space. The two subspaces chosen for this toy example are shown in the top left of Figure 6.2. The one subspace, N, is represented by the pink plane, while the other subspace, $\hat{N}$, is represented by the green plane. So first we will start with two arbitrary orthonormal basis points of one subspace, represented by the black lines, and project them onto the other subspace. This will allow us to form triangles at right angles to each other, in the sense that both hypotenuses are at $90°$ of each other, and so are the line segments on N. But there are

many such pairs of angles which define these triangles, so more thought is needed to define the canonical angles.

The panels in the top row of Figure 6.2 display two example of pairs of triangles that are 90° from each other. The top left is a pair of triangles formed from an arbitrary basis, while the top right is a pair of triangles formed from the canonical angle basis. The first, i.e. smallest, canonical angle is taken from the triangle with the smallest possible cyan line segment length, i.e. the direction of $\hat{N}$ with the smallest distance between its basis point and the projection of the basis point onto N. For the current example, the first canonical angle is 0°. The top right panel shows the triangles formed by the canonical angles. Notice that one triangle has a cyan line segment length of zero. The sine of the next canonical angle corresponds to the length of the cyan line segment of the triangle which is 90° from the first one and has the smallest cyan line segment length. The cyan line segment in the top right panel corresponds to the triangle with the next smallest cyan line segment length that is 90° from the triangle with a cyan line segment length of 0. In general, the sine of the third canonical angle corresponds to the length of the cyan line segment of the triangle 90° from both of the other triangles which has the smallest cyan line segment length, etc. The sine of the canonical angles are equal to the length of the cyan line segments. For the toy example there are only two canonical angles because the subspaces are of dimension 2. In general the number of canonical angles is equal to the largest dimension of the two subspaces.

These triangles of increasing cyan line segments, i.e. sines of angles, can be calculated by solving an eigenproblem. First an arbitrary orthonormal basis of one subspace is projected onto the other. Then finding the minimum cyan line segment length is the same as finding the maximum magenta line segment length. Finding the maximum magenta line segment length is the same as finding the direction of N, which maximizes the sum of squares of the basis points of $\hat{N}$ projected onto N. The eigenvalues define the sum of squares of the eigendirections. Sums of squares define the magenta line segments

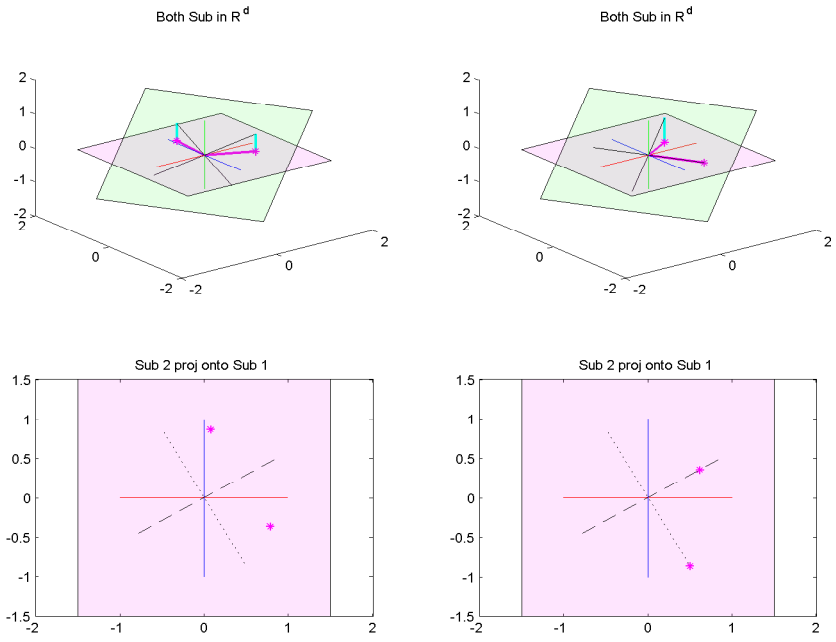

Figure 6.2: *Upper right and left shows N, pink plane, and $\hat{N}$, green plane, in $\mathbb{R}^3$. Top left is arbitrary basis of $\hat{N}$. Top right is $\hat{N}$ with basis that matches the canonical angles, with cyan line segments of 0 and $\sqrt{2}/2$, i.e. sine of 0 and 45. Bottom left is N with canonical directions shown. Bottom right is N with canonical directions shown, as dotted and dashed lines, and square root of eigenvalues marked by magenta points.*

squared lengths in these directions and therefore the triangles. Once the triangles are established the angles that define them are the canonical angles.

Figure 6.2 provides help in understanding of how the eigenanalysis solves for the canonical angles. The one subspace, N, is represented by the pink plane, while the other subspace, $\hat{N}$, is represented by the green plane. An arbitrary orthonormal basis of $\hat{N}$ is represented by the black lines in the top left panel. There are two triangles that can be formed from the basis points of these directions. These triangles both have cyan lines with a length larger than 0. But since these are planes in a 3 dimensional space they have at least one direction in common, hence there is a triangle that can be formed between these two subspaces with a cyan line segment with length 0. This corresponds

to a canonical angle of 0°. Also these subspaces are not exactly the same, so there should be a canonical angle greater than 0°. This is also true since no matter what orthonormal basis is chosen the total sum of the squared lengths of the cyan line segments will be the same.

To find the directions which yield the canonical angles in the subspace N, an eigen-analysis of the projected basis points of $\hat{N}$ onto N is performed. In the bottom left corner of Figure 6.2 is the subspace N in a 2-dimensional view with the projected basis points of $\hat{N}$, i.e the magenta asterisks. The dotted line on this plot is the direction which has the largest sum of squares of these points, i.e. the eigendirection which corresponds to the largest eigenvalues. The dashed line is the direction which has the next largest sum of squares of the points, i.e. the other eigendirection. But since this is a 2 dimensional example, the dashed line has the least sum of squares of any direction as well. The top right shows the basis of $\hat{N}$, which has its basis points' projections onto N lying on the dotted and dashed lines. What you will notice is that the dotted line direction is in both subspaces, and hence has no cyan line and a magenta line with a length of 1. This implies that there is a canonical angle of 0° between these 2 subspaces. The direction associated with the dashed line, now has a larger cyan line segment than the original basis triangles. This is because in this case it is the last canonical angle and will therefore have the largest cyan line segment possible, i.e. the shortest magenta line segment.

In the bottom right of Figure 6.2 is the subspace N but now with the new canonical angle basis of $\hat{N}$ projections shown. Notice that the projections fall onto the eigendirections. The magenta point on the largest eigenvalue direction has a length of 1 from the origin. To find the canonical angle find the angle which has a cosine of 1. The other point happens to be a distance of $\sqrt{2}/2$ from the origin in this example, and to find the second canonical angle find the angle with a cosine of $\sqrt{2}/2$. The eigenvalues are equal to the sum of squares of these points. So the cosine of the canonical angles are equal to the square root of the eigenvalues, based on the projection of basis points of $\hat{N}$ onto N.

The canonical angles define orthonormal bases of N and $\hat{N}$, which have angles between directions which are smallest, then next smallest, etc. This is the same as finding the eigenvalues and eigendirections of an orthonormal basis of $\hat{N}$ projected onto N and vice versa.

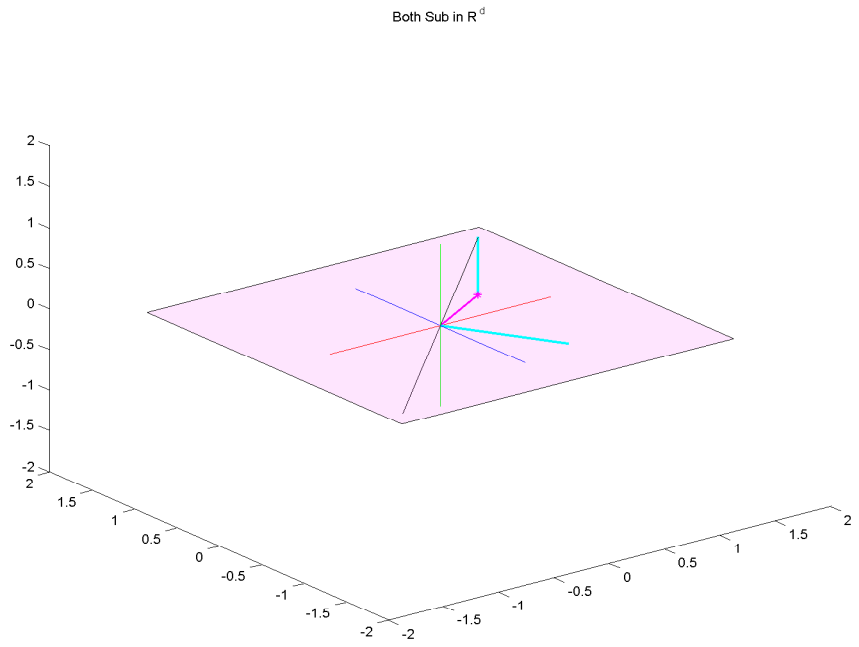

Figure 6.3: *1-d subspace, $\hat{N}$ represented by the line, while 2-d subspace, N, represented by the pink plane. There is a cyan line of length $\sqrt{2}/2$ and 1. So the canonical angles are 45 and 90. One canonical angle is 90, since subspaces are not of equal dimension.*

Finally we have to define the canonical angles in the case where the subspaces have differing dimensions. This will be explained through a toy example with a line, i.e 1 dimensional subspace, and a plane, i.e. 2 dimensional subspace, in 3 dimensions. Two such subspaces of differing dimension are shown in Figure 6.3. The pink plane will represent N and the line will represent $\hat{N}$. Again we take a basis of $\hat{N}$ and project it onto N. This will correspond to the smallest canonical angle. The cosine of the smallest canonical angle will be equal to the length of the magenta line, which is $\sqrt{2}/2$. But since

one subspace is 2 dimensional there should be 2 triangles defining these subspaces, i.e. 2 canonical angles. If we look at the orthogonal direction of the magenta line, then there are no basis points of $\hat{N}$ to project. We then assume that this triangle has magenta line segment of length 0 and a cyan line segment of length 1. This triangle is shown in Figure 6.3 by the cyan line with a length of 1. In this case the subspaces have canonical angles of 45° and 90° degrees between them.

6.2 Canonical Angles and Relation to CCA

Canonical angles are closely related to the calculation of Canonical Correlation Analysis. CCA, developed by Hotelling (1936), is useful when n observations correspond between two paired data sets. Also see Anderson (1984) for more detailed discussion of CCA than is presented in this chapter. Both data sets are standardized to have mean 0 and variance of 1 in every direction. CCA tries to find a direction in one standardized data set with projection coefficients that are most correlated with projection coefficients of a direction from the other standardized data set. It then finds a second orthogonal direction of the first standardized data set with projection coefficients that are most correlated with projection coefficients of a second orthogonal direction of the other standardized data set, etc.

The calculation of CCA is related to canonical angles, which is shown in Section 6.2.1. For an introduction to the geometry of CCA see Kuss and Graepel (2003). In section 6.2.2 it is shown how to solve for canonical angles between 2 subspaces in the framework of CCA. This will be done so that the intuition of CCA can be used to understand canonical angles. A toy example will be used throughout to help visualize the written explanation.

6.2.1 CCA Calculations in Terms of Canonical Angles

CCA is used to understand correlation between two paired data sets. Let one data set be represented by X, a $p \times n$ matrix. Let the other data set be represented by Y, a

$q \times n$ matrix. The n observations are assumed to correspond between the two data sets. In the *primal space* of X the columns are treated as data, i.e. the data are represented as n points in $\mathbb{R}^p$. For the primal space of Y, the data are represented as n points in $\mathbb{R}^q$.

For the toy example let $p = q = 2$ and $n = 4$, so X is a 2×4 matrix as is Y. The primal space representation of X is shown in the upper left corner of Figure 6.4, while the primal space representation of Y is shown in the upper right corner of Figure 6.4. The points with the same color are assumed to correspond between X and Y. Both

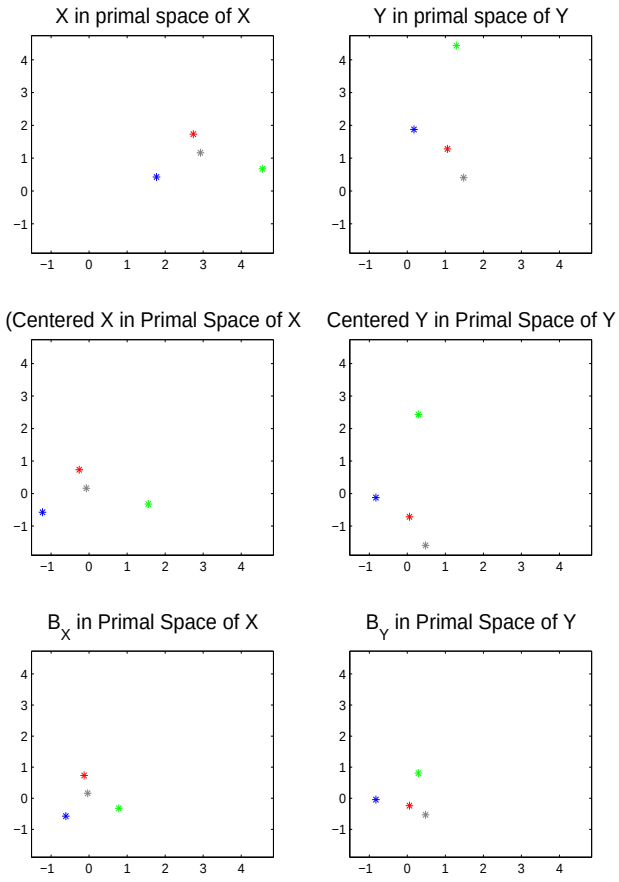

Figure 6.4: *Top row is X and Y represented in the primal spaces. Second row is row sample mean centered X and Y. Notice that points now approximately centered around origin. Third row is full standardization of X and Y using empirical covariance matrix. The covariance structure is no longer the same as in rows above.*

matrices are represented as 4 points in 2 dimensional Euclidean spaces. Although both spaces are 2 dimensional, X and Y are not shown overlayed on the same axes because the measurements of X and Y are not necessarily of the same variables, e.g. they could be on completely different scales. For example X could be height measurements while Y is weight measurements, so it is not appropriate to look at X and Y on the same axes. In general for CCA it is not necessary for p and q to be equal.

Similar to PCA which is based on the covariance matrix, that summarizes the variance structure, CCA is based on the matrix denoted A_{XY}, which summarizes the cross correlation structure of X and Y.

In order to calculate A_{XY} we first have to standardize X and Y, much like in the univariate case of correlation. The first step in the standardization process is to subtract the sample mean of each row of X from X and like wise for Y. The sample mean centered version of X, i.e. $X - \overline{X}$, in the primal space of X is shown in row 2 column 1 of Figure 6.4. The sample mean centered version of Y, i.e. $Y - \overline{Y}$, viewed in the primal space of Y is shown in row 2 column 2 of Figure 6.4. The points in both spaces are now centered around the origin.

The next step in the standardization process is to make projections of the data in all possible directions have a sum of squares of 1. This is done by multiplying the centered data by the square root of the inverse of the empirical covariance matrix and a constant. The standardized version of X is

$$B_X = \frac{1}{\sqrt{n-1}} \hat{\Sigma}_X^{-\frac{1}{2}} (X - \overline{X})$$

where $\overline{X}$ is the sample mean of the rows of X and $\hat{\Sigma}_X = \frac{1}{n-1}(X - \overline{X})(X - \overline{X})^T$ is the empirical covariance matrix of X. The corresponding standardized version of Y is

$$B_Y = \frac{1}{\sqrt{n-1}} \hat{\Sigma}_Y^{-\frac{1}{2}} (Y - \overline{Y})$$

where $\overline{Y}$ is the sample mean of the rows of Y and $\hat{\Sigma}_Y = \frac{1}{n-1}(Y - \overline{Y})(Y - \overline{Y})^T$ is the empirical covariance matrix of Y. The standardized version of X in it's primal space is shown in the bottom left corner of Figure 6.4. The standardized version of Y in it's primal space is shown in the bottom right corner of Figure 6.4. This standardized data has a sum of squares of 1 when projected in any direction. Therefore the data points do not have the same covariance structure as the plots above them.

Now that B_X and B_Y are defined, we will define

$$A_{XY} = B_X B_Y^T,$$

to be the matrix which summarizes the cross correlation structure between X and Y. It is important to note that this is not the usual cross correlation matrix generally denoted by R_{XY}. The calculation of R_{XY} is similar to A_{XY}, except instead of multiplying $(X - \overline{X})$ by the inverse of the covariance matrix it is multiplied by the inverse of the diagonal elements of the covariance matrix with the off diagonals equal to 0 and a corresponding calculation is performed for Y. In order to calculate R_{XY}, let

$$\tilde{B}_X = \frac{1}{\sqrt{n-1}} \tilde{\Sigma}_X^{-\frac{1}{2}} (X - \overline{X})$$

and

$$\tilde{B}_Y = \frac{1}{\sqrt{n-1}} \tilde{\Sigma}_Y^{-\frac{1}{2}} (Y - \overline{Y})$$

where

$$\tilde{\Sigma}_X = \begin{pmatrix} \hat{\Sigma}_X(1,1) & 0 & 0 & \cdots & 0 & \\ 0 & \hat{\Sigma}_X(2,2) & & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots & \\ 0 & \cdots & 0 & 0 & \hat{\Sigma}_X(p,p) & \end{pmatrix}$$

and

$$\tilde{\Sigma}_Y = \begin{pmatrix} \hat{\Sigma}_Y(1,1) & 0 & 0 & \cdots & 0 \\ 0 & \hat{\Sigma}_Y(2,2) & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots & \\ 0 & \cdots & 0 & 0 & \hat{\Sigma}_Y(q,q) \end{pmatrix}.$$

The usual cross correlation matrix can now be characterized in terms of $\tilde{B}_X$ and $\tilde{B}_Y$, i.e.

$$R_{XY} = \tilde{B}_X \tilde{B}_Y^T.$$

There is a case where $A_{XY} = R_{XY}$ and that is when the original coordinates of X and Y have a diagonal covariance matrix. In that case $A_{XY} = R_{XY}$ because $\hat{\Sigma}_X$ and $\hat{\Sigma}_Y$ are already diagonal matrices. For the rest of this section we will assume that X and Y are data matrices such that this is true for simplicity in explanation and interpretation purposes.

In Section 6.2.2 we will show how to frame the calculations of canonical angles in terms of CCA. In order to do this we will assume that X and Y are orthonormal bases. If $\overline{X}$ and $\overline{Y}$ are assumed to be 0 then $\hat{\Sigma}_X$ and $\hat{\Sigma}_Y$ are the identity matrices. This implies that when X and Y are orthonormal bases $A_{XY} = R_{XY}$. Also when X and Y are orthonormal bases they are already in the standardized form needed for CCA calculations.

If X and Y have diagonal covariance matrices then the ij^{th} entry of A_{XY} is the empirical correlation between the projection coefficients of the i^{th} original primal coordinate direction of X and the projection coefficients of the j^{th} original primal coordinate direction of Y. This is also the univariate correlation between the i^{th} row of X and the j^{th} row of Y. If X and Y are not assumed to have diagonal covariance matrices then the calculations to follow are still applicable but the interpretation of A_{XY} as usual correlations no longer holds.

For the toy example, we need to only concentrate on the third row of Figure 6.4 to

demonstrate the calculation of A_{XY}. The horizontal value of each same colored point in the primal spaces of X and Y are multiplied together and then summed. This is entry (1,1) of A_{XY}. If the horizontal values of the points in Y are multiplied by the vertical values of the same colored points in X and summed, then this is entry (2,1) of A_{XY}. If the vertical values of the points in Y are multiplied by the horizontal values of the same colored points in X and summed, then this is entry (1,2) of A_{XY}. If the vertical values of the points in Y are multiplied by the vertical values of the same colored points in X and summed, then this is entry (2,2) of A_{XY}.

If we consider the $p \times q$ matrix A_{XY} from the perspective of the primal space of X, then A_{XY} can be thought of as q points in a p dimensional space, i.e. the columns of A_{XY} are viewed as data. The value of a point of A_{XY} along a coordinate in the primal space of X reflects the correlation of a row of Y with a row of X, where each data point corresponds to a row of Y. The first CCA direction of X is the same as the direction which maximizes the sum of squared distance from the origin of the q points of A_{XY} in the primal space of X. This direction is then found by an eigenanalysis of $A_{XY}A_{XY}^T$ or Singular Value Decomposition (SVD) of A_{XY}. The direction which corresponds to the largest eigenvalue is called the first CCA direction of X. This direction is called the first CCA direction of X and not Y, since the calculations are performed in the primal space of X. The directions which correspond to the decreasing eigenvalues are the corresponding CCA directions of X.

To find the CCA directions of Y we think about A_{XY} in the primal space of Y, i.e. as p points in a q dimensional space. We are now considering the rows of A_{XY} as data points. So the directions found by the eigenanalysis of A_{XY} in the primal space of Y, i.e. the eigenanalysis of $A_{XY}^T A_{XY}$, are the CCA directions of Y.

The sum of squares of the points of A_{XY} are the same for both the primal space of X and Y. Since we are looking at the same matrix the eigenvalues for both spaces will be the same. The eigenvalues represent the square of the correlation coefficients between

directions. The only change is the directions corresponding to the eigenvalues in each space, i.e. the CCA directions of Y are the eigendirections which correspond to $A_{XY}^T A_{XY}$ but the CCA directions of X correspond to $A_{XY} A_{XY}^T$. In each space, CCA is finding the directions most correlated between standardized versions of X and Y.

To understand how canonical angles play a role in the calculation of CCA, we think about X and Y in their *dual spaces*, i.e. treat their rows as data. In the dual space X is represented as p points in $\mathbb{R}^n$ and Y is represented as q points in $\mathbb{R}^n$. Remember that the n observations are assumed to correspond between data sets, so then we can think of the p points associated with X and the q points associated with Y in the same $\mathbb{R}^n$ coordinate system. But again for correlation we are interested in the standardized versions of X and Y, so B_X and B_Y are thought of in the dual spaces of X and Y.

The dual space of X and Y in our toy example is 4 dimensional. In order to calculate B_X and B_Y the sample mean is subtracted from the data. This tells us that B_X has data points which are orthogonal to the direction where the n individuals all have the same value for a dimension, i.e. the direction [1 1 1 1]. The same is true for B_Y. Therefore we can look at the data in the remaining three directions orthogonal to this one without any loss of information, since B_X and B_Y will have coefficients of 0 in the direction [1 1 1 1].

The top left of Figure 6.5 is B_X represented in the dual space of X, and in the top right is B_Y represented in the dual space of Y. Both data are represented as 2 points in the 3 dimensional space orthogonal to [1 1 1 1]. The space is also rotated such that the points of B_X lie on the x and y axes. This is done so that the view of the data points can be better seen and related to each other.

Notice that $B_X B_X^T = I_p$, so B_X^T is an orthonormal basis for a p dimensional subspace of $\mathbb{R}^n$. Also notice that $B_Y B_Y^T = I_q$, so B_Y^T is an orthonormal basis for a q dimensional subspace of $\mathbb{R}^n$. We can then think about the two subspaces generated by these two bases in $\mathbb{R}^n$.

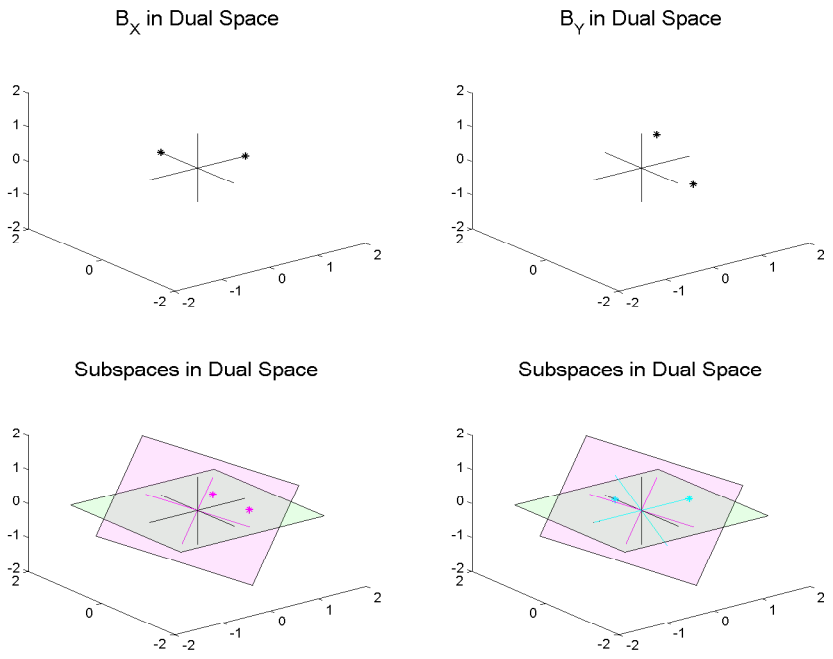

Figure 6.5: *Upper left is B_X in the dual space of X, while upper right is Y in the dual space of Y. B_X and B_Y are represented as 2 points in a 3-d space. Lower left is N and $\hat{N}$ in $\mathbb{R}^3$. Notice that the planes intersect but also are not the same. So the subspaces N and $\hat{N}$ have canonical angles of 0 and, for this example, 45 degrees.*

The subspaces generated by B_X^T and B_Y^T for the toy example are shown in the bottom left corner of Figure 6.5. The subspace, N, that is parallel to the x and y axis is the subspace generated by B_X^T and is represented by the green plane. While the other subspace, $\hat{N}$, is generated by B_Y^T and is represented by the pink plane. The magenta lines on $\hat{N}$ represent the basis B_Y^T. The magenta points on N represent the projection onto N of the unit length points corresponding to the directions of B_Y^T, i.e. the basis points. It was shown in Section 6.1 that by solving an eigenproblem formulated from these points that the canonical angles could be found. Also notice that the projection coefficients are $A_{XY} = B_X B_Y^T$. So the eigenproblem to solve for the CCA directions is the same eigenproblem that solves for the canonical angles. The CCA directions are already shown to equal the eigendirections of the outer and inner product of A_{XY}, i.e.

$A_{XY}A_{XY}^T$ and $A_{XY}^T A_{XY}$, for X and Y respectively. To get the CA basis directions of N, B_X^T is multiplied by the eigendirections of the outer product of A_{XY}. To get the CA basis directions of $\hat{N}$, B_Y^T is multiplied by the eigendirections of the inner product of A_{XY}. Also the squared cosine of the canonical angles are equal to the eigenvalues, which represents the squared correlation of the first CCA directions of X and Y.

The bottom right of Figure 6.5 shows the canonical angle bases of N and $\hat{N}$. The canonical angle basis of $\hat{N}$ is represented by the cyan lines. The canonical angle basis of N are the lines on N which pass through the cyan points and the origin, i.e. the x and y axes. Notice that the x and y axes are the directions of the basis B_X^T. Therefore the canonical angle basis of N is B_X^T.

The toy example is useful to better understand why the same eigenanalysis to find the canonical angles is related to CCA. The upper left of Figure 6.6 shows A_{XY}, for the toy example, as data points thought of in the primal space of X, represented by the magenta points. Point 1 is about at (0.7,0.5) which says that the horizontal axis projection coefficients of Y, i.e. row 1 of Y, is more correlated with the horizontal axis projection coefficients of X, i.e. row 1 of X, than the vertical axis projection coefficients of X, i.e. row 2 of X. Point 2 is about at (0.7,-0.5) which says that the vertical axis projection coefficients of Y, i.e. row 2 of Y, is more correlated with the horizontal axis projection coefficients of X as well. The dotted line is the direction of X which maximizes the sum of squares of the two points associated with A_{XY}, i.e. eigendirection. This panel is the same as if the green plane is shown as a 2-d object. Notice that the magenta points are in the same position relative to the directions of B_X.

This eigendirection is the first CCA direction of X, i.e. the direction with standardized projection coefficients most correlated with standardized projection coefficients of a direction of Y. The dashed line is the direction in the primal space of X which next maximizes the sum of squares of these points and is orthogonal to CCA direction 1 of X. This eigendirection is the second CCA direction of X. The dotted and dashed lines are

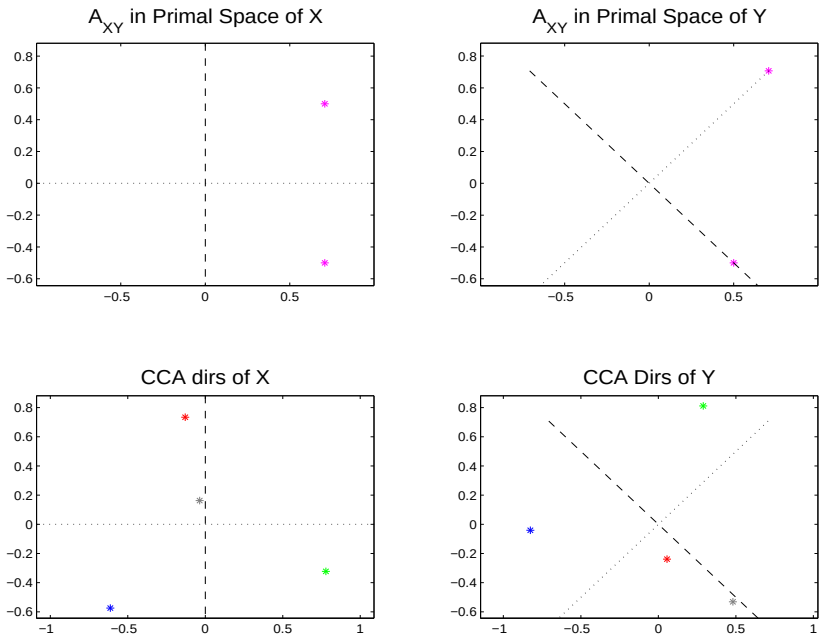

Figure 6.6: *Upper left is A_{XY} in primal space of X, while upper right is A_{XY} in primal space of Y Dotted lines are CCA direction 1 which maximizes sums of squares, i.e. correlation between B_X and B_Y. Dashed lines are CCA direction 2 which next maximizes sums of squares. Bottom left is CCA directions of X with B_X and bottom right is CCA directions of Y with B_Y*

the x and y axis in this panel. This is because the directions of B_X are the canonical angle directions, i.e. CCA directions. In the bottom left of Figure 6.6 is the CCA directions of X shown with B_X in the primal space of X. Since both X and Y are 2 dimensional, all of the correlation is explained by these directions.

The display on the right side of Figure 6.6, for the Y-space, are similar to those on the left for the X-space. In this case the panel is the magenta plane shown as a 2-d object. The magenta lines would correspond to the x and y axis of this figure. The dotted and dashed lines are the first and second CCA directions of Y. This would be the same as the cyan lines in the bottom right of Figure 6.5. In the bottom right of Figure 6.6 is the CCA directions of Y shown with B_Y in the primal space of Y.

For the toy example one eigenvalue is 1, and the dotted line is the direction associ-

ated with that eigenvalue. This says that if you were to project the data points of the standardized version of X onto the dotted line in the bottom left panel of Figure 6.6, then they would be perfectly correlated, i.e have r= ± 1, with the points of standardized version of Y projected onto the dotted line in the bottom right of Figure 6.6. The second eigenvalue is 0.5, so the projections onto the dashed lines have a correlation of r=$\pm\sqrt{(0.5)} = \pm 0.7071 = \pm\sqrt{2}/2$.

So finding the canonical angles corresponds to maximizing the sum of squares of these projections coefficients. So therefore if the canonical angles are found then the calculations for CCA are complete. The link between the two is that the cosine of the canonical angles are equal to the square root of the eigenvalues of A_{XY}, which is equal to the amount of correlation between the CCA directions of the standardized versions of X and Y.

For the toy example, the cosine of the canonical angles are $\cos(0) = 1$ and $\cos(45) = \sqrt{2}/2$, see Figure 6.5. So 1 and $\sqrt{2}/2$ is the distance of the cyan points from the origin. This tells us that the cosine of the canonical angles is equal to the correlation coefficient. We can use the cyan points distance since that is the canonical angle basis of $\hat{N}$, and therefore the distance is equal to the square root of the eigenvalues.

6.2.2 CA calculations in Terms of CCA

Section 6.2.1 showed how the calculations of CCA in the primal space could be derived in terms of canonical angles in the dual space. In this section it is shown how to start with two subspaces in $\mathbb{R}^n$, and calculate the canonical angles through CCA. We want to go from the case of subspaces as shown in the bottom left of Figure 6.5 to data as shown in row 3 of Figure 6.4.

Let N be of dimension p, and let B_X^T be an orthonormal basis of N represented by p points in $\mathbb{R}^n$. Let $\hat{N}$ be of dimension q, and let B_Y^T be a basis of $\hat{N}$ represented as q points in $\mathbb{R}^n$. Each subspace is then being compared in the same space $\mathbb{R}^n$. In CCA we

went from thinking about n points in two different spaces, to thinking about a p and a q dimensional subspace in $\mathbb{R}^n$, i.e from the row 1 of Figure 6.4 to the lower left of Figure 6.5. We can apply the above connection in reverse now. We will think about B_X^T as n points in a space that is p dimensional and we will think of B_Y^T as n points in a different q dimensional space.

Each of the n points correspond between spaces, by the angle between the basis directions and the coordinate direction of $\mathbb{R}^n$. Let N_i be a $p \times 1$ vector that corresponds to a point in $\mathbb{R}^p$, then the entries of this vector are equal to the cosine of the angles between the basis directions of N and the ith coordinate of $\mathbb{R}^n$. Let $\hat{N}_i$ be a $q \times 1$ vector that corresponds to a point in $\mathbb{R}^q$, then the entries of this vector are equal to the cosine of the angles between the basis directions of $\hat{N}$ and the ith coordinate of $\mathbb{R}^n$. This is the same as projecting a point that is distance 1 from the origin along a coordinate of $\mathbb{R}^n$ onto the subspace N and onto $\hat{N}$, respectively. This is the same by the relation of the Euclidean inner product and the cosine of an angle between directions. This operation is the same as plotting B_X as n points in a p dimensional space and B_Y as n points in a q dimensional space.

How to map the subspaces in the bottom of Figure 6.5 into spaces that look like the bottom row of Figure 6.4 using the cosine relationship will now be shown. The green, red, and blue lines represent the coordinate directions of $\mathbb{R}^3$, shown in the upper left of Figure 6.7. The x and y coordinate axes, represented by the red and blue lines, are the basis chosen to generate N. The black lines represent the basis chosen to generate $\hat{N}$.

Each of these bases can be mapped into a different 2 dimensional Euclidean space. The mapping is done by taking the cosines of the angles between a direction of the basis and the original coordinates of $\mathbb{R}^3$. In this case each basis is 2 directions in a 3 dimensional space, so there will be 3 points in a 2 dimensional space for N and $\hat{N}$.

The mapping of B_X^T into a 2-d space is shown in the bottom left panel. Notice that the subspace N is orthogonal to the green line in the bottom of Figure 6.7, i.e. each

direction of the basis is 90° from this direction. Also $\cos(90°) = 0$, so the green point in the bottom left of Figure 6.7 is at $(\cos(90°), \cos(90°)) = (0, 0)$. Now let's look at the basis with relation to the x-axis, i.e. red line. One of the directions of the basis is this red line, i.e. has an angle of 0° with respect to it, while the other direction is orthogonal to it. So the red point is at $(\cos(0°), \cos(90°)) = (1, 0)$ in the bottom left of Figure 6.7. Finally one direction of the basis is the blue line while the other is orthogonal to it. So the blue dot is at the point $(\cos(90°), \cos(0°)) = (0, 1)$ in the bottom left of Figure 6.7. The mapping of B_Y^T into a 2-d space is shown in the bottom right panel. The color of

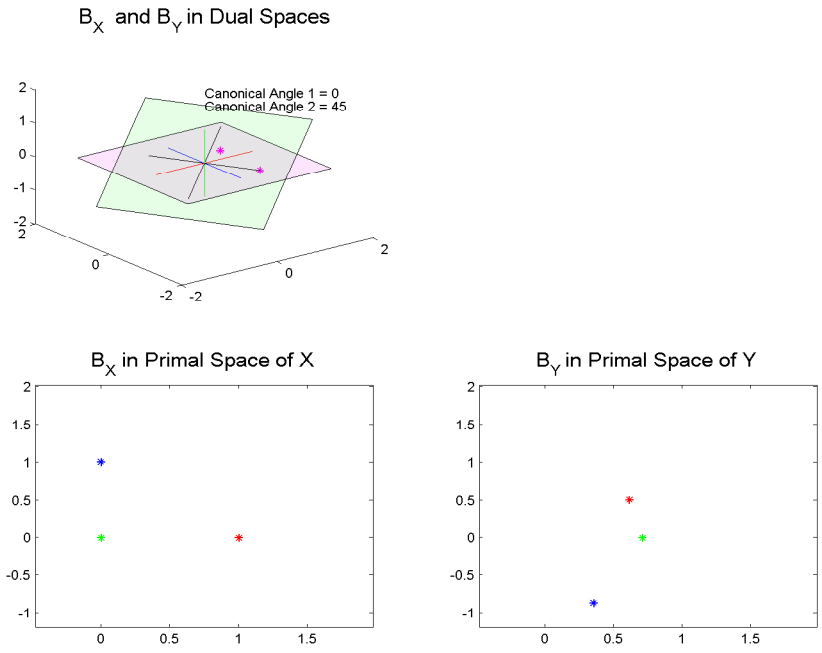

Figure 6.7: *Upper left is N and $\hat{N}$ in $\mathbb{R}^3$. Lower left is mapping of N to a 2-d Euclidean space by using cosine. Lower right is mapping of $\hat{N}$ to a 2-d Euclidean space by using cosine. Notice that the same colored points correspond between spaces and to the coordinates of $\mathbb{R}^3$.*

the points is the same as the axis colors in the top panel, with the value being the cosine of the angle that each direction of the basis is away from that color line in $\mathbb{R}^3$.

If a direction is found in each space such that these n projection coefficients are per-

fectly correlated, then these directions of N and $\hat{N}$ have the same angles with relation to the original $\mathbb{R}^n$ directions, i.e. the two subspaces have a common direction. So as the n points are less correlated the directions are, in some sense, farther apart in $\mathbb{R}^n$. If the n projection coefficients have a correlation of 0 then N and $\hat{N}$ have directions which are orthogonal. So by finding directions which have n projection coefficients correlated it is the same as finding directions which have similar angles with relation to the original directions of $\mathbb{R}^n$. This exact problem is solved by CCA analysis for data that is standardized to have mean 0 and variance of 1 in every direction. Orthonormal bases have sums of squares, i.e. variance, of 1 in every direction. However the rows do not have mean 0. But we are interested in the relationship of these points from 0, not the sample mean. Therefore it is assumed that the mean is 0. Therefore if the orthonormal bases of N and $\hat{N}$ are thought of as the standardized versions of X and Y then the calculations of Section 6.2.1 can be followed to find the squared sines of the canonical angles.

This allows us to think of canonical angles in terms of the intuition of CCA. Canonical angles find directions of each subspace where the cosine of angles between these directions and the original coordinates are most correlated, i.e. the angles are similar. So the directions with the most correlated cosine of angles have the smallest canonical angles associated with them. So this allows us a way to view the correlation of angles.

6.3 Gap Metric

The Gap Metric, D_{gap}, is a common metric from the literature for the distance between two subspaces, see Stewart and Sun (1990) or Kato (1966). The Gap Metric is defined as the sine of the largest canonical angles between two subspaces.

$$D_{gap}(N, \hat{N}) = \sin\theta_1$$

where $\{\theta_1, \theta_2, \ldots, \theta_{d_N}\}$ are the canonical angles between N and $\hat{N}$ in decreasing order. The gap metric can also be expressed in terms of the eigenvalues of the inner or outer

product of A_{XY}, i.e. $A_{XY}^T A_{XY}$ or $A_{XY} A_{XY}^T$, the matrix introduced in Section 6.2.1. This characterization of the gap metric is,

$$D_{gap}(N, \hat{N}) = \sqrt{1 - \min_i \lambda_i^A},$$

where λ_i^A are the eigenvalues of the inner or outer product of A_{XY}. The gap metric seems to get its name from the fact that it is the largest distance between a point that is a distance of 1 from the origin in $\hat{N}$ and its projection onto N, i.e. the sine of the largest canonical angle.

There are two other important characterizations of the gap metric based on projection matrices see Stewart and Sun (1990) Theorem 5.5. These characterizations will be important in Chapter 7, when the focus will be on the asymptotic properties of an estimated subspace. This estimated subspace is studied mainly through its projection matrix. Let P_N be the projection matrix of N and $\hat{P}_N$ be the projection matrix of $\hat{N}$. Both of the following characterizations of the gap metric only hold for the case when N and $\hat{N}$ are of equal dimension. Assume N and $\hat{N}$ both are of dimension d_N and subspaces of $\mathbb{R}^d$. The first characterization is based on the eigenvalues of the matrix P_{one}, defined as

$$P_{one} = (P_N(I_d - \hat{P}_N))(P_N(I_d - \hat{P}_N))^T = P_N - P_N \hat{P}_N P_N,$$

which are denoted by $\{\lambda_1^{P_{one}}, \ldots, \lambda_d^{P_{one}}\}$ in descending order. The eigenvalues of P_{one} are equal to $\{sin^2\theta_1, \ldots, sin^2\theta_k, 0, \ldots 0\}$, where

$$k = d_N \text{ if } 2d_N \leq d \text{ or}$$
$$d - d_N \text{ if } 2d_N > d$$

Therefore the first characterization is

$$D_{gap}(N, \hat{N}) = \sqrt{\max_i \lambda_i^{P_{one}}}.$$

The next characterization is also based on the projection matrices of N and $\hat{N}$. This characterization is based on the eigenvalues of

$$P_{two} = (P_N - \hat{P}_N)(P_N - \hat{P}_N)^T,$$

which are denoted by $\{\lambda_1^{P_{two}}, \ldots, \lambda_d^{P_{two}}\}$ in descending order. The eigenvalues of P_{two} are also related to the canonical angles. But in this case the eigenvalues are equal to

$$\{sin^2\theta_1, sin^2\theta_1, \ldots, sin^2\theta_k, sin^2\theta_k, 0, \ldots 0\}$$

, where k is the same as in the case of P_{one}. So for this characterization

$$D_{gap}(N, \hat{N}) = \sqrt{\max_i \lambda_i^{P_{two}}}.$$

Several characterizations of the gap metric are given above, now some properties of the gap metric are presented. The distance defined by the Gap metric is strongly influenced by the largest canonical angle. Because $D_{gap}(N, \hat{N})$ depends more heavily on θ_1, then the other θ_i's are discounted, which can be seen from the following example. When the dimension of $\hat{N}$ is $d_N \pm 1$ and the dimension of N is d_N then $D_{gap}(N, \hat{N}) = 1$. Also let $\tilde{N}$ be of dimension $d_N \pm m$, when m is an integer > 1, and N is of dimension d_N then $D_{gap}(N, \tilde{N}) = 1$ as well. Therefore the gap metric says two subspaces that are off by one dimension are the same distance apart as two subspaces that are off by more than one dimension.

Also the gap metric says that two subspaces with canonical angles of $\theta_1 = 80$, and

$\theta_2 = 10$ are the same distance apart as two subspaces with canonical angles of $\theta_1 = 80$ and $\theta_2 = 70$. In both cases $D_{gap} = \sin 80$. This is again because the gap metrics value is directly related only to θ_1.

Another matrix from the literature described in the next section will depend more heavily on all of the canonical angles. Therefore this metric will be more sensitive to changes in the canonical angles besides the largest.

6.4 Euclidean Sine metric

Another common metric from the literature is

$$D_{sine}(N, \hat{N}) = \sqrt{\sum \sin^2 \theta_i} \ = \sqrt{\max(\dim N, \dim \hat{N}) - \sum \lambda_i^A},$$

see Stewart and Sun (1990) or Kato (1966). There are also two other characterizations of the Euclidean sine metric based on P_N and $\hat{P}_N$, the projection matrices of N and $\hat{N}$ respectively, through the matrices P_{one} and P_{two}. Again it has to be assumed that N and $\hat{N}$ are of the same dimension. The first characterization is

$$D_{sine}(N, \hat{N}) = \sqrt{trace(P_N - P_N \hat{P}_N P_N)} = \sqrt{\sum_{i=1}^{d} \lambda_i^{P_{one}}},$$

which is the square root of the sum of the eigenvalues of P_{one}. It may seem as though more attention needs to be paid to the number of eigenvalues being summed, since there are more eigenvalues than there are canonical angles. But in this case all eigenvalues not corresponding to a canonical angle are 0, and therefore do not contribute anything to the calculation of the trace. The second characterization

$$D_{sine}(N, \hat{N}) = \sqrt{\frac{1}{2} trace[P_{two}]} = \sqrt{\frac{1}{2} trace[(P_N - \hat{P}_N)^T (P_N - \hat{P}_N)]},$$

is based on projection matrices as well.

Although sometimes we are only summing $2 \times (d - d_N)$ squared sines of canonical angles this characterization of the Euclidean sine metric is still correct. This is because the remaining canonical angles are 0, so the Euclidean sine metric is not changed if the summation of these values are included or not. This is best understood by thinking of the subspaces orthogonal to N and $\hat{N}$, i.e. $N^{\perp}$ and $\hat{N}^{\perp}$. The subspaces N and $\hat{N}$ should be the same distance apart as $N^{\perp}$ and $\hat{N}^{\perp}$. Therefore if $d < 2d_N$ then the above metric can be thought of as summing the squared sines of the canonical angles between $N^{\perp}$ and $\hat{N}^{\perp}$, which have $d - d_N$ canonical angles.

Several characterizations of the Euclidean sine metric are given above, some intuition into how it measures the distance between subspaces is given here. The Euclidean sine metric is equal to the square root of the sum of squared length of all line segments between the basis points of an orthonormal basis of $\hat{N}$ and their projections onto N. So not just the largest line segment appears directly in the calculation of the metric but all of the line segments do, i.e. the largest canonical angle is not as influential as in the gap metric.

This metric is more sensitive to the dimensions of the two subspaces. For instance, the distance between two subspace with dimensions that differ by one would be less than two subspaces with dimensions that differ by a larger amount. If N is of dimension d_N and $\hat{N}$ is of dimension $d_N + 1$ and $N \subseteq \hat{N}$ then $D_{sine}(N, \hat{N}) = 1$ as does $D_{gap}(N, \hat{N})$. But if N is of dimension d_N and $\tilde{N}$ is of dimension $d_N + m$, m>1, and $N \subseteq \tilde{N}$ then $D_{sine}(N, \tilde{N}) = \sqrt{m}$ while $D_{gap}(N, \tilde{N})$ still equals 1. Also unlike the gap metric though the sine metric is smaller when $\theta_1 = 80$ and $\theta_2 = 10$ than $\theta_1 = 80$ and $\theta_2 = 70$. In the first case $D_{sine} = 1$ while for the second case $D_{sine} = 1.3612$. Also if N and $\hat{N}$ are orthogonal to each other then $D_{sine} = \sqrt{\max(\dim N, \dim \hat{N})}$.

But if the sine metric is large you are unable to tell if it is due to one large angle or many small angles. This is best understood by an example. Let there be two pairs of subspaces. For the first pair of subspaces, one subspace is of dimension 4 while the other

is of dimension 2 and therefore $\theta_1 = 90$ and $\theta_2 = 90$ and let $\theta_i = 0$ for $3 \leq i \leq 4$. In this case $D_{sine} = \sqrt{(2)}$. The other pair of subspaces are both of dimension 4 and $\theta_i = 30$ for $1 \leq i \leq 4$. So for this pair of subspaces $D_{sine} = \sqrt{(2)}$ as well.

For the Euclidean sine metric no eigenanalysis is needed. As long as you find a basis of N and $\hat{N}$ the Euclidean sine metric can be calculated. In section 6.1 and 6.2.1 it is shown that the cosines of the canonical angles are equal to $\sqrt{\lambda_i^A}$. Also $A_{X,Y} = B_X B_Y^T$, where B_X and B_Y are any orthonormal bases of $\hat{N}$ and N. The Euclidean sine metric is equal to $\sqrt{\max(\dim N, \dim \hat{N}) - \sum \lambda_i^A}$. But $\sum \lambda_i^A = \text{trace}(A_{X,Y} A_{X,Y}^T)$, so the canonical angles do not have to be found, i.e. an eigenanalysis is not needed. For the gap metric the canonical angles must be found in order to find the largest possible cyan line segment between N and $\hat{N}$, see Figure 6.2, or eigenanalysis of a matrix performed.

In the next chapter we will study the asymptotic properties of an estimated subspace. In order for an estimated subspace to be a good estimate we would like for it to converge to the true subspace. Therefore we will be interested in the relation of these two metrics as they converge to 0. The metrics have the following relationship, if $D_{sine}(N, \hat{N}) \to 0$ then $D_{gap}(N, \hat{N}) \to 0$ where $\hat{N}$ is a sequence of subspaces which converges to N in the Euclidean sine metric. The first step in proving this relationship is noticing that

$$(\sin\theta_1)^2 \leq \sum_{i=1}^{d_N} (\sin\theta_i)^2.$$

This is since all of the canonical angles are between 0 and 90°, i.e. $\sin\theta_i \geq 0$ for all i. The above implies that

$$\sqrt{(\sin\theta_1)^2} \leq \sqrt{(\sum_{i=1}^{d_N} \sin\theta_i)^2}.$$

The proof is complete by noticing that $\sqrt{(\sin\theta_1)^2} = D_{gap}(N, \hat{N})$ and $\sqrt{(\sum_{i=1}^{d_N} \sin\theta_i)^2} = D_{sine}(N, \hat{N})$.

It also holds that if $D_{gap}(N, \hat{N}) \to 0$ then $D_{sine}(N, \hat{N}) \to 0$. The proof starts by

noticing that for all i,

$$(\sin\theta_i)^2 \leq (\sin\theta_1)^2,$$

since θ_1 is the largest canonical angle. Therefore this implies that

$$\sum_{i=1}^{d_N}(\sin\theta_i)^2 \leq d_N(\sin\theta_1)^2.$$

If the square root of both sides are taken, i.e.

$$\sqrt{\sum_{i=1}^{d_N}(\sin\theta_i)^2} \leq \sqrt{d_N(\sin\theta_1)^2},$$

then by the characterizations given above this is the same as stating

$$D_{sine}(N,\hat{N}) \leq \sqrt{d_N}D_{gap}(N,\hat{N}).$$

Therefore if $d_N < \infty$ and $D_{gap}(N,\hat{N}) \to 0$, then $D_{sine}(N,\hat{N}) \to 0$. This relationship of the two metrics converging to 0 is important for chapter 7.

CHAPTER 7

Mathematical Statistic Investigation

This section gives a discussion of the work for the dissertation in the area of mathematical statistics. The focus of this analysis is on the asymptotic properties of the nearly null space and the interesting genetic constraint space. Section 7.1.2 studies the convergence of the dimension of the estimated nearly null space to the dimension of the true nearly null space. The next section, 7.1.3, analyzes the deeper property of the convergence of the estimated nearly null space to the true nearly null space. Section 7.2.2 considers the convergence of the dimension of the estimated interesting genetic constraint space to the dimension of the true interesting genetic constraint space. Section 7.2.3 follows this with an investigation of the convergence of the estimated interesting genetic constraint space to the true interesting genetic constraint space. Section 7.3 describes a hypothesis test of whether $S_0 \subseteq S$, where S is the interesting genetic constraint space and S_0 is a given vector subspace.

Before moving on to the precise mathematical statistics that are going to be investigated, first several distributional assumptions are stated to establish the mathematical setting. For the following sections it is assumed that there is a sequence, $X = [X_1, ...,X_n]$, of independent identically distributed d dimensional normal random vectors. Each vector X_i is assumed to have covariance matrix Σ, which has eigenvalues λ_1 to λ_d in descending order. The eigenvalues will be assumed to have the structure of $\lambda_m > \lambda_{m+1} = \cdots = \lambda_d > 0$. With the larger eigenvalues having an unspecified structure.

Also Σ will be estimated by an empirical covariance matrix $\hat{\Sigma} = \frac{1}{n-1}(X - \overline{X})(X - \overline{X})^T$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. $\hat{\Sigma}$ has the estimated eigenvalues $\hat{\lambda}_1$ to $\hat{\lambda}_d$, in descending order. See Anderson (1984) for a more detailed discussion about multivariate analysis.

7.1 Study of Nearly Null Space Asymptotic Properties

7.1.1 Definition of Nearly Null Space

Before the mathematical statistical properties of the nearly null space are investigated, the nearly null space first must be defined. One way to define the nearly null space, which was discussed in Chapter 3, is the subspace generated by the eigenvectors which correspond to the smallest eigenvalues of Σ, the covariance matrix, such that the proportion of variance of these small eigenvalues is less than a given constant, c_{prop}. This definition will be referred to as the *proportion threshold nearly null space*, i.e N_{prop}. See Anderson (1984) for more discussion on why this threshold is commonly used. The estimate of this nearly null space, $\hat{N}_{prop}$, is the subspace generated by the eigenvectors which correspond to the small eigenvalues of $\hat{\Sigma}$, the empirical covariance matrix, such that the proportion of variance of these small eigenvalues is less than c_{prop}.

An alternative definition is to define the nearly null space to be the subspace generated by the eigenvectors which correspond to eigenvalues of Σ that are less than a given threshold, c_{thresh}. This definition will be referred to as the *eigenvalue threshold nearly null space*, i.e. N_{thresh}. The subspace N_{thresh} is estimated by $\hat{N}_{thresh}$, which is the subspace generated by the eigenvectors which correspond to the eigenvalues of $\hat{\Sigma}$ that are less than c_{thresh}. There are other definitions that could be used to define this nearly null space, but the above two definitions will be focused on here.

The proportion threshold nearly null space, $\hat{N}_{prop}$, has the advantage over the eigenvalue threshold nearly null space, $\hat{N}_{thresh}$, in that it is more interpretable. Often when

thinking about a model space, one looks at the total percentage of variation explained by the first several principal components. The proportion threshold nearly null space can be easily thought of as the orthogonal subspace to this, i.e. as the subspace generated by the smallest principal components that explain a small percentage of the total variation. But the proportion threshold nearly null space requires careful treatment when the eigenvalues are tied at the proportion threshold limit. For instance let $\lambda_j = \lambda_{j+1}$, where $\sum_{i=j+1}^{d} \lambda_i < c_{prop} \sum_{i=1}^{d} \lambda_i < \sum_{i=j}^{d} \lambda_i$. The eigendirections corresponding to these values are random directions in a 2-d space since $\lambda_j = \lambda_{j+1}$, so the subspace N_{prop} will not be clearly defined. The eigenvalue threshold nearly null space does not have this problem because if $\lambda_j = \lambda_{j+1}$, then either both will be included in the nearly null space or both will not be included in the nearly null space. So N_{thresh} is clearly defined. We also know that any direction in N_{thresh} will explain less variation than c_{thresh}, but N_{thresh} could explain a large percentage of the total variation if a large portion of the total variation is equally distributed over many directions. Due to the limitations of each definition of the nearly null space, the mathematical statistical properties of both are investigated.

Although N_{prop} was introduced in earlier chapters, in general for the mathematical statistics N_{thresh} will be analyzed before N_{prop}. This is because in general the assumptions used for both are similar, except that N_{prop} will have the added assumption that $\lambda_j \neq \lambda_{j+1}$. This extra assumption allows N_{prop} to be uniquely defined.

Although if the nearly null space is assumed to be the subspace generated by eigenvectors which correspond to $\lambda_{m+1} \ldots \lambda_d$, then both N_{prop} and N_{thresh} will have the same assumption. This assumption is that $\lambda_m \neq \lambda_{m+1}$. This will be the case when the interesting genetic constraint space is being considered.

For this analysis c_{thresh} and c_{prop} will be defined in two ways. First c_{thresh} and c_{prop} are viewed as constants that are not equal to one of the eigenvalues of Σ or the proportion of the lower eigenvalues of Σ, respectively. Next c_{thresh} is viewed as a sequence which depends on n that converges to an eigenvalue. For this case we will let $c(n)_{thresh} =$

$\lambda_{m+1} + n^{-\alpha}$ and study the asymptotic properties of the nearly null space as $n \to \infty$. Also we will let $c(n)_{prop} = \frac{\sum_{i=m+1}^{d} \lambda_i}{\sum_{i=1}^{d} \lambda_i} + n^{-\alpha}$ and investigate the asymptotic properties of the nearly null space.

The sequences $c(n)_{thresh}$ and $c(n)_{prop}$ are defined so that the convergence asymptotically can be studied in a similar manner to that of contiguity in statistical inference. Contiguity studies probability measures which live "on top of each other" in the limit, see van der Vaart (1998). In this same way $c(n)_{thesh}$ and $\hat{\lambda}_j$ both are equal in the limit to λ_j. So the question is which converges quicker to λ_j. The probability of $\hat{\lambda}_j$ being less than $c(n)_{thesh}$ for different α's gives an answer to this question and therefore information about the convergence of $\hat{N}_{thresh}$ to N_{thesh}. This same type of asymptotic analysis is performed for the case of $c(n)_{prop}$.

7.1.2 Convergence In Probability Of the Dimension of the Estimated Nearly Null Space

This section studies the convergence in probability of the dimension of the estimated nearly null space to the dimension of the true nearly null space. It outlines several conditions necessary to have $\dim(\hat{N}_{thresh})$ converge in probability to $\dim(N_{thresh})$ and to have $\dim(\hat{N}_{prop})$ converge in probability to $\dim(N_{prop})$, as the sample size n grows. First the nearly null space defined as N_{thresh} is investigated, then an investigation is done for N_{prop}.

Theorem 1 Given the definition of N_{thresh}, $\hat{N}_{thresh}$, and c_{thresh} from above, the dimension of $\hat{N}_{thresh}$ converges in probability to the dimension of N_{thresh}, i.e.

$$\lim_{n \to \infty} \mathbb{P}(dim(\hat{N}_{thresh}) = dim(N_{thresh})) = 1$$

In order to show that the dimension of $\hat{N}_{thresh}$ converges in probability to the dimension of N_{thresh}, the behavior the sample eigenvalues with relation to c_{thresh} are studied. Without loss of generality it is assumed that $\lambda_m > c_{thresh} > \lambda_{m+1}$. So in order to prove

Theorem 1 it is shown that

$$\mathbb{P}(\hat{\lambda}_i > c_{thresh} \text{ for } i \leq m \text{ and } \hat{\lambda}_j > c_{thresh} \text{ for } j \geq m+1)$$

converges in probability to 1.

Proof of Theorem 1 Estimated eigenvalues which correspond to true eigenvalues which are different are asymptotically independent, see Anderson (1963). Therefore as $n \to \infty$

$$\lim_{n\to\infty} \mathbb{P}(\hat{\lambda}_i > c_{thresh} \text{ for } i \leq m \text{ and } \hat{\lambda}_j > c_{thresh} \text{ for } j \geq m+1) \tag{7.1}$$

$$= \lim_{n\to\infty} \prod_{r=1}^{R} \mathbb{P}(\hat{\Lambda}_r > c_{thresh}) \prod_{R+1} \mathbb{P}(\hat{\Lambda}_{R+1} < c_{thresh}) \tag{7.2}$$

Where $\hat{\Lambda}_r$ is a set of estimated eigenvalues which corresponds to a set of true eigenvalues which are equal. For our case we will assume that there are $R+1$ distinct true eigenvalues, labeled λ_1^{dist} to λ_{R+1}^{dist} in descending order. Each λ_r^{dist} is assumed to have multiplicity q_r. If each piece of the product in equation 7.2 is shown to converge in probability to 1, then the Theorem will be proven.

In order to make a probability statement about a set of eigenvalues the asymptotic distribution of the set is useful. The focus need only be on the distribution of each set separately and not the distribution of all sets together, since distinct eigenvalues are asymptotically independent.

In Anderson (1963) it is shown that $\sqrt{n}(\hat{\Lambda}_r - \Lambda_r)$ converges in distribution to H. Let $h_{i,r} = \sqrt{n}(\hat{\lambda}_{i,r} - \lambda_r^{dist})$, where $\lambda_{i,r}$ is the ith largest sample eigenvalue in the set $\hat{\Lambda}_r$. If the eigenvalue has a multiplicity of 1, then H converges to a normal distribution with mean 0 and variance $2(\lambda_r^{dist})^2$. If the eigenvalue has a multiplicity of size q_r, then H converges in distribution to f($\sqrt{n}(\hat{\Lambda}_r - \Lambda_r)$;$q_r$,$\lambda_r^{dist}$).

Where f($\sqrt{n}(\hat{\Lambda}_r - \Lambda_r)$;$q_r$,$\lambda_r^{dist}$)

$$= K(q_r, \lambda_r^{dist}) exp(\frac{\sum_i^{q_r} h_{i,r}^2}{4(\lambda_r^{dist})^2}) \prod_{i<j} (h_{i,r} - h_{j,r}).$$

In the above density $K(q_r, \lambda_r^{dist})$ is a constant that ensures that the density integrates to 1.

Now that the asymptotic distribution of each set is defined, the asymptotic probability of the set can be studied. Finding the probability for each set of sample eigenvalues less then c_{thresh} will be similar. Therefore finding the probability of only one set of estimated eigenvalues less than c_{thresh} will be shown in detail. So for this case there is only one set of sample eigenvalues which should be less than c_{thresh}. The set contains q_r sample eigenvalues. We want to investigate the probability that all of the sample eigenvalues are less than the constant c_{thresh}.

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(\hat{\Lambda}_{R+1} < c_{thresh}) \\
&= \lim_{n\to\infty} \mathbb{P}(\sqrt{n}(\hat{\Lambda}_{R+1} - \Lambda_{R+1}) < \sqrt{n}(c_{thresh} - \lambda_{R+1}^{dist})) \\
&= \lim_{n\to\infty} \mathbb{P}(H < \sqrt{n}(c_{thresh} - \lambda_{R+1}^{dist}))
\end{aligned}$$

Notice that $(c_{thresh} - \lambda_{R+1}^{dist})$ is positive. Therefore as $n \to \infty$, $\sqrt{n}(c_{thresh} - \lambda_{R+1}^{dist}) \to \infty$. So as $n \to \infty$

$$\begin{aligned}
&= \lim_{n\to\infty} \mathbb{P}(\sqrt{n}(\hat{\Lambda}_{R+1} - \Lambda_{R+1}) < \sqrt{n}(c_{thresh} - \lambda_{R+1}^{dist})) \\
&= \mathbb{P}(H < \infty))
\end{aligned}$$

Therefore $\mathbb{P}(\hat{\Lambda}_{R+1} < c_{thresh})$ converges in probability to 1.

A similar argument can be made for sets of eigenvalues greater than c_{thresh}. Again only one set of eigenvalues are shown in detail. For this case we will look at the set of eigenvalues closet to c_{thresh}.

$$\begin{aligned}&\mathbb{P}(\hat{\Lambda}_R > c_{thresh})\\&= \mathbb{P}(\sqrt{n}(\hat{\Lambda}_R - \Lambda_R) > \sqrt{n}(c_{thresh} - \lambda_R^{dist}))\end{aligned}$$

For this case $(c_{thresh} - \lambda_R^{dist})$ is negative. Therefore as $n \to \infty$, $\sqrt{n}(c_{thresh} - \lambda_R^{dist}) \to -\infty$. So as $n \to \infty$, $\mathbb{P}(\hat{\Lambda}_R > c_{thresh})$ is equal in limit to $\mathbb{P}(H > -\infty)$. Therefore $\mathbb{P}(\hat{\Lambda}_R > c_{thresh})$ converges in probability to 1.

Since the probability of each term in Equation 7.2 converges in probability to 1, it follows that the product as a whole converges in probability to 1. Therefore the dimension of $\hat{N}_{thresh}$ converges in probability to the dimension of N_{thresh}. □

Defining N_{thresh} and $\hat{N}_{thresh}$ by letting $c(n)_{thresh} = \lambda_{m+1} + n^{-\alpha}$, allows for a deeper understanding of the asymptotic convergence of the dimension of $\hat{N}_{thresh}$ to N_{thresh}.

Theorem 2 Given the definition of N_{thresh} and $\hat{N}_{thresh}$ from above as well as $c(n)_{thresh} = \lambda_{m+1} + n^{-\alpha}$, the dimension of $\hat{N}_{thresh}$ converges in probability to the dimension of N_{thresh} when $0 < \alpha < \frac{1}{2}$, i.e.

$$\lim_{n\to\infty} \mathbb{P}(dim(\hat{N}_{thresh}) = dim(N_{thresh})) = 1$$

The proof for this Theorem is similar to the case of c_{thresh}. Again the probability will be studied as a product of several probabilities. Each set $\hat{\Lambda}_r$ will be studied separately, and if all sets converge to above or below the threshold respectively, then the estimated dimension will converge in probability to the true dimension. Without loss of generality

it will be assumed that $c(n)_{thresh} = \lambda_{m+1} + n^{-\alpha}$, where $\lambda_{m+1} \subseteq \Lambda_{R+1}$.

Proof of Theorem 2 First a set that corresponds to an eigenvalue larger that λ_{R+1}^{dist} is chosen and the probability that the set is large than $c(n)_{thresh}$ is studied.

$$\begin{aligned}
&\mathbb{P}(\hat{\Lambda}_R > c(n)_{thresh}) \\
&= \mathbb{P}(\hat{\Lambda}_R > \lambda_{R+1}^{dist} + n^{-\alpha}) \\
&= \mathbb{P}((\hat{\Lambda}_R - \Lambda_R) > (\lambda_{R+1}^{dist} - \lambda_R^{dist}) + n^{-\alpha}) \\
&= \mathbb{P}(\sqrt{n}(\hat{\Lambda}_R - \Lambda_R) > \sqrt{n}(\lambda_{R+1}^{dist} - \lambda_R^{dist}) + n^{\frac{1}{2}-\alpha})
\end{aligned}$$

Notice that $(\lambda_{R+1}^{dist} - \lambda_R^{dist})$ is negative. This implies that as $n \to \infty$, $\sqrt{n}(\lambda_{R+1}^{dist} - \lambda_R^{dist}) \to -\infty$. Also since $\alpha < \frac{1}{2}$ then $n^{\frac{1}{2}-\alpha} \to \infty$. But $\sqrt{n}(\lambda_{R+1}^{dist} - \lambda_R^{dist})$ is on a large scale than is $n^{\frac{1}{2}-\alpha}$, so $\sqrt{n}(\lambda_{R+1}^{dist} - \lambda_R^{dist}) + n^{\frac{1}{2}-\alpha} \to -\infty$. Therefore as $n \to \infty$

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(\hat{\Lambda}_R > c(n)_{thresh}) \\
&= \mathbb{P}(H > -\infty)
\end{aligned}$$

which implies that $\mathbb{P}(\hat{\Lambda}_R > c(n)_{thresh})$ converges in probability to 1. This same argument can be made for all sets Λ_r such the true eigenvalues which correspond to the set are greater than λ_{R+1}^{dist}.

Next the set $\hat{\Lambda}_{R+1}$ is investigated.

$$\begin{aligned}
&\mathbb{P}(\hat{\Lambda}_{R+1} < c(n)_{thresh}) \\
&= \mathbb{P}((\hat{\Lambda}_{R+1} - \Lambda_{R+1}) < (\lambda_{R+1}^{dist} - \lambda_{R+1}^{dist}) + n^{-\alpha}) \\
&= \mathbb{P}((\hat{\Lambda}_{R+1} - \Lambda_{R+1}) < n^{-\alpha})
\end{aligned}$$

$$= \mathbb{P}(\sqrt{n}(\hat{\Lambda}_{R+1} - \Lambda_{R+1}) < n^{\frac{1}{2}-\alpha})$$

Therefore as $n \to \infty$

$$\lim_{n\to\infty} \mathbb{P}(\hat{\Lambda}_{R+1} < c(n)_{thresh})$$
$$= \mathbb{P}(H < \infty)$$

Therefore $\mathbb{P}(\hat{\Lambda}_{R+1} < c(n)_{thresh})$ converges in probability to 1.

Now all terms of the product are shown to converge in probability to 1. So this implies that the dimension of $\hat{N}_{thresh}$ converges in probability to the dimension of N_{thresh} when $c(n)_{thesh}$ is used as the eigenvalue threshold. □

Next when the definition of the nearly null space is changed to N_{prop}, the asymptotic properties of the convergence of the dimension of $\hat{N}_{prop}$ to N_{prop} are investigated. Similar to the case of N_{thresh}, first c_{prop} is assumed to be a given fixed constant not equal to the proportion of variance of the lower eigenvalues. It is also assumed that $\sum_{i=j+1}^{d} \lambda_i < c_{prop} \sum_{i=1}^{d} \lambda_i < \sum_{i=j}^{d} \lambda_i$, where $\lambda_j \neq \lambda_{j+1}$. With out loss of generality it is assumed that $j = m$.

Theorem 3 Given the definition of N_{prop}, $\hat{N}_{prop}$, and c_{prop} from above, the dimension of $\hat{N}_{prop}$ converges in probability to the dimension of N_{prop} , i.e.

$$\lim_{n\to\infty} \mathbb{P}(dim(\hat{N_{prop}}) = dim(N_{prop})) = 1$$

The behavior of the proportion of variance of a set of eigenvalues will be examined in order to show that the estimated dimension converges to the true dimension. So in order

for Theorem 3 to be proved it can be shown that

$$\mathbb{P}(f_{m+1}(\hat{\lambda}) > c_{prop} \text{ or } \ f_m(\hat{\lambda}) < c_{prop})$$

converges in probability to 0, where $f_j(\hat{\lambda}) = \frac{\sum_{i=j}^{d} \hat{\lambda}_i}{\sum_{i=1}^{d} \hat{\lambda}_i}$. Also $f_j(\lambda)$ will represent an analogous quantity, with the estimated eigenvalues replaced by the true eigenvalues.

Proof of Theorem 3 Similar to the case of N_{thresh}, if the asymptotic distribution of $f_j(\hat{\lambda})$ is understood then probability statements can be made. It is shown in Anderson (1984) that

$$\sqrt{n}(f_j(\hat{\lambda}) - f_j(\lambda))$$

converges in distribution to N_j, which is normally distributed with mean 0 and variance

$$2(\frac{\sum_{i=j}^{d} \lambda_i}{\sum_{i=1}^{d} \lambda_i})^2(\sum_{i=1}^{j-1} \lambda_i^2) + 2(\frac{\sum_{i=1}^{j-1} \lambda_i}{\sum_{i=1}^{d} \lambda_i})^2(\sum_{i=j}^{d} \lambda_i^2).$$

Now that the asymptotic distribution of $f_j(\hat{\lambda})$ is understood, we can examine the probability of $f_j(\hat{\lambda})$ being greater than or less than a constant.

$$\begin{aligned}
&\mathbb{P}(f_{m+1}(\hat{\lambda}) > c_{prop} \text{ or } \ f_m(\hat{\lambda}) < c_{prop}) \\
&\leq \mathbb{P}(f_{m+1}(\hat{\lambda}) > c_{prop}) + \mathbb{P}(f_m(\hat{\lambda}) < c_{prop})
\end{aligned}$$

Now examine each piece of the sum separately. First focus on showing that $\mathbb{P}(f_{m+1}(\hat{\lambda}) > c_{prop})$ converges to 0.

$$\mathbb{P}(f_{m+1}(\hat{\lambda}) > c_{prop})$$

$$= \mathbb{P}((f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > (c_{prop} - f_{m+1}(\lambda)))$$
$$= \mathbb{P}(\sqrt{n}(f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > \sqrt{n}(c_{prop} - f_{m+1}(\lambda)))$$

Notice that $(c_{prop} - f_{m+1}(\lambda))$ is positive, so therefore $\sqrt{n}(c_{prop} - f_{m+1}(\lambda)) \to \infty$. This implies that as $n \to \infty$

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > \sqrt{n}(c_{prop} - f_{m+1}(\lambda)))$$
$$= \mathbb{P}(N_{m+1} > \infty)$$

So then $\mathbb{P}(f_{m+1}(\hat{\lambda}) > c_{prop})$ converges in probability to 0.

Next show that $\mathbb{P}(f_m(\hat{\lambda}) < c_{prop})$ converges in probability to 0.

$$\mathbb{P}(f_m(\hat{\lambda}) < c_{prop})$$
$$= \mathbb{P}((f_m(\hat{\lambda}) - f_m(\lambda)) < (c_{prop} - f_m(\lambda)))$$
$$= \mathbb{P}(\sqrt{n}(f_m(\hat{\lambda}) - f_m(\lambda)) < \sqrt{n}(c_{prop} - f_m(\lambda)))$$

Notice that $(c_{prop} - f_m(\lambda))$ is negative, so therefore $\sqrt{n}(c_{prop} - f_m(\lambda)) \to -\infty$. This implies that as $n \to \infty$

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(f_m(\hat{\lambda}) - f_m(\lambda)) < \sqrt{n}(c_{prop} - f_m(\lambda)))$$
$$= \mathbb{P}(N_m < -\infty)$$

So then $\mathbb{P}(f_m(\hat{\lambda}) < c_{prop})$ converges in probability to 0. $\square$

Again defining N_{prop} and $\hat{N}_{prop}$ by letting $c(n)_{prop} = \frac{\sum_{i=m+1}^{d} \lambda_i}{\sum_{i=1}^{d} \lambda_i} + n^{-\alpha}$, allows for a deeper understanding of the asymptotic convergence of the dimension of $\hat{N}_{prop}$ to N_{prop}.

Theorem 4 Given the definition of N_{prop}, $\hat{N}_{prop}$, and $c(n)_{prop}$, the dimension of $\hat{N}_{sum}$ converges in probability to the dimension of N_{sum} when $0 < \alpha < \frac{1}{2}$, i.e.

$$\lim_{n\to\infty} \mathbb{P}(dim(\hat{N}_{prop}) = dim(N_{prop})) = 1$$

Proof of Theorem 4 This proof is quite similar to the proof of c_{prop} as the threshold limit.

$$\begin{aligned}
&\mathbb{P}(f_{m+1}(\hat{\lambda}) > c(n)_{prop} \text{ or } \ f_m(\hat{\lambda}) < c_{prop}) \\
&\leq \mathbb{P}(f_{m+1}(\hat{\lambda}) > c(n)_{prop}) + \mathbb{P}(f_m(\hat{\lambda}) < c(n)_{prop})
\end{aligned}$$

Now examine each piece of the sum separately. First focus on showing that $\mathbb{P}(f_{m+1}(\hat{\lambda}) > c_{prop})$ converges in probability to 0.

$$\begin{aligned}
&\mathbb{P}(f_{m+1}(\hat{\lambda}) > c(n)_{prop}) \\
&= \mathbb{P}((f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > (c(n)_{prop} - f_{m+1}(\lambda))) \\
&= \mathbb{P}(\sqrt{n}(f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > \sqrt{n}(c(n)_{prop} - f_{m+1}(\lambda)))
\end{aligned}$$

Notice that $c(n)_{prop} = f_{m+1}(\lambda) + n^{-\alpha}$. So then

$$\mathbb{P}(\sqrt{n}(f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > \sqrt{n}(c(n)_{prop} - f_{m+1}(\lambda)))$$

$$= \mathbb{P}(\sqrt{n}(f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > \sqrt{n}(f_{m+1}(\lambda) + n^{-\alpha} - f_{m+1}(\lambda)))$$
$$= \mathbb{P}(\sqrt{n}(f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > n^{\frac{1}{2}-\alpha})$$

Since $\alpha < \frac{1}{2}$, $n^{\frac{1}{2}-\alpha} \to \infty$, this implies that as $n \to \infty$

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(f_{m+1}(\hat{\lambda}) - f_{m+1}(\lambda)) > n^{\frac{1}{2}-\alpha})$$
$$= \mathbb{P}(N_{m+1} > \infty)$$

So then $\mathbb{P}(f_{m+1}(\hat{\lambda}) > c(n)_{prop})$ converges in probability to 0.

Next show that $\mathbb{P}(f_m(\hat{\lambda}) < c_{prop})$ converges in probability to 0.

$$\mathbb{P}(f_m(\hat{\lambda}) < c(n)_{prop})$$
$$= \mathbb{P}((f_m(\hat{\lambda}) - f_m(\lambda)) < (c(n)_{prop} - f_m(\lambda)))$$
$$= \mathbb{P}(\sqrt{n}(f_m(\hat{\lambda}) - f_m(\lambda)) < \sqrt{n}(c(n)_{prop} - f_m(\lambda)))$$
$$= \mathbb{P}(\sqrt{n}(f_m(\hat{\lambda}) - f_m(\lambda)) > \sqrt{n}(f_{m+1}(\lambda) + n^{-\alpha} - f_m(\lambda)))$$
$$= \mathbb{P}(\sqrt{n}(f_m(\hat{\lambda}) - f_m(\lambda)) > \sqrt{n}(f_{m+1}(\lambda) - f_m(\lambda)) + n^{\frac{1}{2}-\alpha})$$

Notice that $(f_{m+1}(\lambda) - f_m(\lambda))$ is negative, so therefore $\sqrt{n}(f_{m+1}(\lambda) - f_m(\lambda)) \to -\infty$. But $\alpha < \frac{1}{2}$, which implies $n^{\frac{1}{2}-\alpha} \to \infty$. But $\sqrt{n}(f_{m+1}(\lambda) - f_m(\lambda))$ is on a larger scale than is $n^{\frac{1}{2}-\alpha}$, so $\sqrt{n}(f_{m+1}(\lambda) - f_m(\lambda)) + n^{\frac{1}{2}-\alpha}) \to -\infty$

This implies that as $n \to \infty$

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(f_m(\hat{\lambda}) - f_m(\lambda)) < \sqrt{n}(c(n)_{prop} - f_m(\lambda)))$$

$$= \mathbb{P}(N_m < -\infty)$$

So then $\mathbb{P}(f_m(\hat{\lambda}) < c(n)_{prop})$ converges in probability to 0. $\square$

7.1.3 Convergence In Probability of the Nearly Null Space

The next step in our mathematical statistical analysis is to show that the estimated nearly null space will converge to the true nearly null space. First a metric of distance between subspaces must be decided upon. In Section 6.3 and Section 6.4 the gap metric, D_{gap}, and the Euclidean sine metric, D_{sine}, were introduced respectively. Although there are other metrics to measure the distance between subspaces, we will concentrate on these two metrics. For a discussion of the advantages and disadvantages of the two metrics see Sections 6.3 and 6.4.

Both of these metrics can be written in terms of the matrices which project onto the given subspaces. Therefore let P_N be the matrix which projects onto N_{thresh} and $\hat{P}_N$ be the matrix which projects onto $\hat{N}_{thresh}$.

Before proceeding a lemma about true and estimated projection matrices of eigenspaces, from Tyler (1981), will be stated. This lemma will be used in several of the proofs. Let P_0 be the matrix which projects onto the subspace generated by the eigenvectors which correspond to the true eigenvalues λ_{m+1} to λ_d. Also let $\hat{P}_0$ be the subspace generated by the estimated eigenvectors which correspond to the estimated eigenvalues $\hat{\lambda}_{m+1}$ to $\hat{\lambda}_d$. Notice that $P_0 = P_N$, but it is not always true that $\hat{P}_0 = \hat{P}_N$. The estimated cases are not always equal, since $\hat{P}_N$ may have a larger or smaller dimension than $\hat{P}_0$. But in the case when the dimension of $\hat{N}_{thresh}$ is equal to N_{thresh} then $\hat{P}_0 = \hat{P}_N$. In the lemma below P_N can always replace P_0. Also $\hat{P}_N$ can replace $\hat{P}_0$ when the estimated dimension is equal to the true dimension.

If $a_n vec(\Sigma - \hat{\Sigma})$ converges in distribution to a multivariate normal then the below can

be stated.

Lemma 1 Define the norm $\| B \|$=[max eigenvalue $(B'B)]^{\frac{1}{2}}$. If $\| \Sigma - \hat{\Sigma} \| \leq \frac{\lambda_m - \lambda_{m+1}}{2}$ then there is a sequence a_n, with $a_n \to \infty$, so that

$$a_n(P_0 - \hat{P}_0) \to N_0$$

where vec(N_0) is distributed multivariate normal with mean 0 and covariance matrix Σ_0.

With vec being an operation that stacks columns of a matrix into a vector. Let $B = [b_1, b_2, \cdots, b_m]$, where B is a $d \times m$ matrix and each b_i a d dimensional vector. Then

$$vec(B) = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

where vec(B) is a dm dimensional vector.

A fact concerning the condition $\| \Sigma - \hat{\Sigma} \| \leq \frac{\lambda_m - \lambda_{m+1}}{2}$ is stated, which is useful in allowing Lemma 1 to be used in the proof of the following theorem.

Fact 1 Let $E_n = \text{event}\{\| \Sigma - \hat{\Sigma} \| \leq \frac{\lambda_m - \lambda_{m+1}}{2}\}$. Since $vec(\Sigma - \hat{\Sigma})$ converges in probability to 0, it follows that

$$\lim_{n \to \infty} \mathbb{P}(E_n) = 1.$$

Theorem 5 Given the definition of N_{thresh}, $\hat{N}_{thresh}$, and c_{thresh} from above, it follows that for $\forall\ \epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) = 0$$

Proof of Theorem 5 For the proof of this Theorem, Lemma 1 will be helpful. First $\mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon)$ is broken into where E_n occurs and where it does not occur.

$$\mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n) + \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n^c).$$

In order to use Lemma 1 we would like to only concentrate on the first probability in the sum. Therefore it is shown that the second probability in the sum converges to 0.

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n^c)$$
$$\leq \lim_{n\to\infty} \mathbb{P}(E_n^c) = 0$$

The convergence to 0 can be seen by Fact 1. This convergence to 0 implies

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) = \lim_{n\to\infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n)$$

Also Lemma 1 is only helpful in the case that $\hat{P}_0 = \hat{P}_N$. So the next step is to break $\mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n)$ into cases when $\hat{P}_0 = \hat{P}_N$ and $\hat{P}_0 \neq \hat{P}_N$. This is the same as the case when the $dim(N_{thresh}) = dim(\hat{N}_{thresh})$ and when $dim(N_{thresh}) \neq dim(\hat{N}_{thresh})$.

$$\mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n)$$
$$= \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n | dim(N_{thresh}) = dim(\hat{N}_{thresh}))\mathbb{P}(dim(N_{thresh}) = dim(\hat{N}_{thresh}))$$
$$+ \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n | dim(N_{thresh}) \neq dim(\hat{N}_{thresh}))\mathbb{P}(dim(N_{thresh}) \neq dim(\hat{N}_{thresh}))$$
$$= \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n | \hat{P}_0 = \hat{P}_N)\mathbb{P}(\hat{P}_0 = \hat{P}_N)$$
$$+ \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n | \hat{P}_0 \neq \hat{P}_N)\mathbb{P}(\hat{P}_0 \neq \hat{P}_N)$$

Since $dim(\hat{N}_{thesh}) \to dim(N_{thresh})$ in probability, it follows that $\mathbb{P}(\hat{P}_0 \neq \hat{P}_N)$ converges in probability to 0. Which implies that $\mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n | \hat{P}_0 \neq \hat{P}_N)\mathbb{P}(\hat{P}_0 \neq$

$\hat{P}_N$) converges in probability to 0. Also since $dim(\hat{N}_{thesh}) \rightarrow dim(N_{thresh})$ in probability, then $\mathbb{P}(\hat{P}_0 = \hat{P}_N)$ converges in probability to 1.

Therefore as $n \rightarrow \infty$

$$\lim_{n\rightarrow\infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n)$$
$$= \lim_{n\rightarrow\infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n | \hat{P}_0 = \hat{P}_N)$$

Now only the terms where $\hat{P}_N = \hat{P}_0$ need to be examined. This allows Lemma 1 and the continuous mapping theorem to be used to show that $\mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon \cap E_n | \hat{P}_0 = \hat{P}_N)$ converges in probability to 0.

But first an alternate characterization of D_{sine} is needed, where

$$D_{sine} = \sqrt{\frac{1}{2} trace[(P_0 - \hat{P}_0)((P_0 - \hat{P}_0)^T]}$$

when the dimension of P_0 and $\hat{P}_0$ are equal.

Noting lemma 1, it can be seen that $(P_0 - \hat{P}_0) \rightarrow 0$ in probability. Then by the continuous mapping theorem $\sqrt{\frac{1}{2} trace[(P_0 - \hat{P}_0)((P_0 - \hat{P}_0)^T]} \rightarrow 0$ in probability. This implies that $\mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) \rightarrow 0$ in probability. □

Theorem 6 Given the definition of N_{thresh}, $\hat{N}_{thresh}$, and c_{thresh}, it follows that for $\forall$ $\epsilon > 0$

$$\lim_{n\rightarrow\infty} \mathbb{P}(D_{gap}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) = 0$$

Proof of Theorem 6 The proof of Theorem 6, is to notice that if D_{sine} converges to zero then so does D_{gap}. This is because if the trace of a matrix converges to zero then the largest eigenvalue also will converge to zero, for a positive definite matrix. This relationship is shown in Section 6.4. □

The next step is to assume $c(n)_{thresh} = \lambda_{m+1} + n^{-\alpha}$ and to show that $\hat{N}_{thresh}$ converges

in probability to N_{thesh}. The same assumptions as when the threshold was a constant, along with the additional assumption that $0 < \alpha < \frac{1}{2}$ are needed to show convergence in both D_{gap} and D_{sine}.

Theorem 7 Given the definition of N_{thresh} and $\hat{N}_{thresh}$ from above and letting $c(n)_{thresh} = \lambda_{m+1} + n^{-\alpha}$, it follows that for $0 < \alpha < \frac{1}{2}$ and for $\forall\ \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) = 0$$

Proof of Theorem 7 For this proof the steps of proving Theorem 5 are followed exactly. Since again the dimension of the estimated nearly null space converge in probability to the true nearly null space. Then the problem can be rewritten as a probability of $P_0 - \hat{P}_0$. □

Theorem 8 Given the definition of N_{thresh} and $\hat{N}_{thresh}$ from above and letting $c(n)_{thresh} = \lambda_{m+1} + n^{-\alpha}$, then for $0 < \alpha < \frac{1}{2}$ and for $\forall\ \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(D_{gap}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) = 0$$

Proof of Theorem 8 This Theorem is proven, by noticing that Theorem 8 is the same as Theorem 7 except with D_{sine} replaced by D_{gap}. But if D_{sine} converges then this implies that D_{gap} will converge. □

The above should also be shown in the case of N_{prop} being the definition of the nearly null space and $\hat{N}_{prop}$ being the estimated nearly null space. First it should be shown when c_{prop} is a given constant not equal to the proportion of variance of the lower eigenvalues.

Theorem 9 Given the definition of N_{prop}, $\hat{N}_{prop}$, and c_{prop} it follows that for $\forall\ \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(N_{prop}, \hat{N}_{prop}) > \epsilon) = 0$$

Proof of Theorem 9 The proof is exactly the same as the proof of Theorem 5,

only with N_{thresh} replaced by N_{prop}. This is the case since the proof only depends on the dimension of the estimated subspace converging to the true subspace implying that $\hat{P}_0 = \hat{P}_N$, which occurs for $\hat{N}_{prop}$.

Theorem 10 Given the definition of N_{prop}, $\hat{N}_{prop}$, and c_{prop} from above it follows that for $\forall\ \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(D_{gap}(N_{prop}, \hat{N}_{prop}) > \epsilon) = 0$$

Proof of Theorem 10 Follows from the proof of Theorem 9, since if $D_{sine}(N_{prop}, \hat{N}_{prop})$ converges in probability to 0, then so does $D_{gap}(N_{prop}, \hat{N}_{prop})$.

Also for the case when $c(n)_{prop} = \frac{\sum_{i=m+1}^{d} \lambda_i}{\sum_{i=1}^{d} \lambda_i} + n^{-\alpha}$ the convergence in probability of $\hat{N}_{prop}$ to N_{prop} should be shown. All of the assumptions of above are needed along with the additional assumption that $0 < \alpha < \frac{1}{2}$.

Theorem 11 Given the definition of N_{prop}, $\hat{N}_{prop}$,and $c(n)_{prop}$ from above it follows that for $0 < \alpha < \frac{1}{2}$ and for $\forall\ \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(N_{prop}, \hat{N_{prop}}) > \epsilon) = 0$$

Proof of Theorem 11 Follows directly from the proof of Theorem 7, using the same reasoning that the proof of Theorem 9 follows from the proof of Theorem 5.

Theorem 12 Given the definition of N_{prop}, $\hat{N}_{prop}$,and $c(n)_{prop}$ from above it follows that for $0 < \alpha < \frac{1}{2}$ and for $\forall\ \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(D_{gap}(N_{prop}, \hat{N_{prop}}) > \epsilon) = 0$$

Proof of Theorem 12 Follows from the proof of Theorem 11, since if $D_{sine}(N_{prop}, \hat{N}_{prop})$ converges in probability to 0, then so does $D_{gap}(N_{prop}, \hat{N}_{prop})$.

7.2 Asymptotic Properties of the Interesting Genetic Constraint Space

7.2.1 Definition of the Interesting Genetic Constraint Space

Similar to the case of the nearly null space, before the asymptotic properties of the interesting genetic constraint space, described in Section 3.1, can be studied, this space must first be defined. First some intuition into what is the interesting generic constraint space is developed. Then later in the section comes a more rigorous mathematical definition.

The interesting genetic constraint space, S, is defined in a similar manner to N_{thresh}. The set of directions which generate N_{thresh}, is chosen such that the direction that explains the least amount of variation is included in the set if the amount of variation explained is less than c_{thresh}. Then the direction which is orthogonal to the first and explains the least amount of variation is included in the set if the amount of variation explained is less than c_{thresh}. This process continues until a direction explains more variation than c_{thresh}. This subspace is also generated by the eigendirections which correspond to eigenvalues of Σ, i.e. the covariance matrix, that are less than c_{thresh}.

The set of directions which generate S is chosen such that the direction in the nearly null space that has the highest simplicity score is included in the set if the simplicity score is greater than c_{simp}. Then the direction in the nearly null space which is orthogonal to the first and has the highest simplicity score is included in the set if the simplicity score is greater than c_{simp}. This process continues until a direction in the nearly null space has a simplicity score less than c_{simp}. An alternate way of generating this subspace is using eigendirections which correspond to eigenvalues of a matrix. But in this case we will focus on the eigenvalues and eigendirections of a simplicity matrix, rather than a covariance matrix. The simplicity matrix defines simplicity scores of directions in the nearly null

space. Therefore S is generated by eigendirections which correspond to eigenvalues of the simplicity matrix larger than c_{simp}.

Note that the interesting genetic constraint space is not defined using a threshold of the proportion of total energy of a matrix, such as was the case for the nearly null space.

The above paragraphs gave intuition into how S is going to be defined. The rest of the section defines S in a more rigorous mathematical fashion.

The definition of the interesting genetic constraint space, S, is highly dependent on the definition of the nearly null space. This is because S is always going to be a subspace of the nearly null space. So we must first consider the nearly null space before proceeding to the definition of S. The nearly null space is defined in two ways in Section 7.1.1, labeled as N_{thresh} and N_{prop}. The subspaces N_{thresh} and N_{prop} are equal if the constants c_{thresh} and c_{prop} are chosen such that the eigendirections which generate the two subspaces correspond to the same eigenvalues of Σ. For this section the nearly null space is denoted as N and it will be assumed that c_{prop} and c_{thresh} are chosen such that the subspace is generated by the eigendirections which correspond to $\lambda_{m+1} \dots \lambda_d$. The estimated nearly null space is also defined in two ways in Section 7.1.1, labeled as $\hat{N}_{thresh}$ and $\hat{N}_{prop}$. For this section it can be assumed that N is estimated by either $\hat{N}_{thresh}$ or $\hat{N}_{prop}$. Therefore the estimated nearly null space is denoted by $\hat{N}$. It was shown in Section 7.1.2 and Section 7.1.3 that $\hat{N}_{thresh}$ and $\hat{N}_{prop}$ have similar asymptotic properties.

Now that the null space is defined, a way to find orthogonal directions of maximum simplicity is needed. One way to do this is to define a simplicity matrix which summarizes the simplicity scores of the directions of the null space. This will be analogous to the case of principal components where a covariance matrix summarizes the variance of directions. Then similar to principal components, an eigendecomposition of this simplicity matrix will yield orthogonal directions of maximum simplicity.

Before a simplicity matrix of the nearly null space is found, first a simplicity matrix of the full space is defined. Our definition of simplicity from Section 3.2, is that

the simplest direction is the one that corresponds to the vector with minimal value for the first differences squared of adjacent entries. One possible simplicity matrix is to use DD^T, where D is a matrix which produces first differences. This matrix produces eigendirections which correspond to directions which fit our ideal of simplicity. But the matrix produces eigenvalues which have two disadvantages. The first disadvantage is that of interpretability. The simplest direction will have the smallest eigenvalue. But our emphasis is on maximizing simplicity, so therefore we would like to have the simplest direction correspond to the largest eigenvalue. The second problem with the eigenvalues of this matrix is that there is an eigenvalue of 0. This is undesirable since when we consider the nearly null space, those directions not in the nearly null space will have a simplicity score of 0. Therefore if the direction from the full space with a simplicity score of 0 is in the nearly null space, we will be unable to separate it from the directions not in N.

A simplicity matrix, F^{full}, which is a function of DD^T and overcomes the problems mentioned above is considered. The matrix F^{full} should have eigendirections which are the same as DD^T, since those are the directions which match our idea of simplicity. Define, the matrix

$$F^{full} = 4I_d - DD^T$$

Note that F^{full} has the same eigendirections as DD^T, because $-DD^T$ has the same eigendirections as DD^T. F^{full} has a constant simplicity score added to every direction in the full space. Therefore the ordering of the simplicity scores of every direction is the same, but the simplicity scores are all shifted by a constant. For an algebraic proof of the eigendirections being the same see Appendix A. The simplicity matrix F^{full} does not encounter the problem of having eigenvalues equal to 0. Since DD^T always has eigenvalues between 0 and 4, with the largest eigenvalue never reaching 4, see Schatzman (2002). The problem of DD^T having eigendirections ordered incorrectly, i.e. from least simple to most simple, is addressed by the negative sign in the definition of F^{full}.

Now that the simplicity matrix of the full space is defined the simplicity matrix of the nearly null space can be defined. The simplicity matrix of the nearly null space can be thought of as a projection of the simplicity matrix of the full space into the nearly null space. Let P_N be the matrix which projects onto the nearly null space. Then the simplicity matrix of the nearly null space is

$$F = P_N F^{full} P_N.$$

The simplicity matrix F has similar properties to F^{full}. First is that the eigendirections of F will have the same eigendirections as $P_N D D^T P_N$, see Appendix A. Also F will have the same number of eigenvalues greater than 0 as the dimension of P_N and all other eigenvalues will be 0. Finally the simplest direction in the nearly null space will correspond to the eigendirection associated with the largest eigenvalue of F. The next simplest direction of the nearly null space orthogonal to the first will correspond to the eigendirection associated with the second largest eigenvalue, etc.

Therefore S is defined as the subspace generated by eigendirections of F which correspond to eigenvalues larger than c_{simp}. The eigenvalues of F are denoted by λ_1^F to λ_d^F in descending order.

The *estimated interesting genetic constraint space*, $\hat{S}$, is defined by a simplicity matrix of the estimated nearly null space. Let $\hat{P}_N$ be the matrix which projects onto $\hat{N}$. Our simplicity matrix of the estimated nearly null space is

$$\hat{F} = \hat{P}_N F^{full} \hat{P}_N.$$

The definition of $\hat{S}$ is the subspace generated by the eigendirections of $\hat{F}$ which correspond to the eigenvalues larger than c_{simp}. The eigenvalues of $\hat{F}$ are denoted by $\hat{\lambda}_1^F$ to $\hat{\lambda}_d^F$ in descending order. The matrix $\hat{F}$ has as many non-zero eigenvalues as the dimension of $\hat{P}_N$.

7.2.2 Convergence In Probability Of the Dimension of the Estimated Interesting Genetic Constraint Space

In section 7.2.1 the interesting genetic constraint space is defined. Also an estimate, $\hat{S}$ of this space is defined. This section studies the convergence in probability of the dimension of $\hat{S}$ to the dimension of S. For the remaining sections it is assumed that c_{simp} is a constant not equal to one of the eigenvalues of F and greater than 0.

Theorem 13 Given the definition of S, $\hat{S}$, and c_{simp}, it follows that the dimension of $\hat{S}$ converges in probability to the dimension of S, i.e.

$$\lim_{n\to\infty} \mathbb{P}(dim(\hat{S}) = dim(S)) = 1$$

The proof of Theorem 13 involves studying the behavior of the eigenvalues of $\hat{F}$ with relation to c_{simp}. Without loss of generality assume $\lambda_k^F > c_{simp} > \lambda_{k+1}^F$, where $k+1 < d_N = dim(N)$. Theorem 13 is proven if it is shown that

$$\mathbb{P}(\hat{\lambda}_1^F < c_{simp} \text{ or } \ldots \hat{\lambda}_k^F < c_{simp} \text{ or } \hat{\lambda}_{k+1}^F > c_{simp} \text{ or } \ldots \hat{\lambda}_d^F > c_{simp})$$

converges to 0.

Proof of Theorem 13 The first step of the proof is to break the above probability into the case where the estimated eigenvalues correspond to true eigenvalues of 0 and eigenvalues greater than 0.

$$\begin{aligned}
&\mathbb{P}(\hat{\lambda}_1^F < c_{simp} \text{ or } \ldots \hat{\lambda}_k^F < c_{simp} \text{ or } \hat{\lambda}_{k+1}^F > c_{simp} \text{ or } \ldots \hat{\lambda}_d^F > c_{simp}) \\
&\leq \mathbb{P}(\hat{\lambda}_1^F < c_{simp} \text{ or } \ldots \hat{\lambda}_k^F < c_{simp} \text{ or } \hat{\lambda}_{k+1}^F > c_{simp} \text{ or } \ldots \hat{\lambda}_{d_N}^F > c_{simp}) \\
&+ \mathbb{P}(\hat{\lambda}_{d_N+1}^F > c_{simp} \text{ or } \ldots \hat{\lambda}_d^F > c_{simp})
\end{aligned}$$

For the second part of the sum, none of the estimated eigenvalues are greater than 0 unless the dimension of $\hat{P}_N$, i.e. the dimension of $\hat{N}$, is greater than $d_N = dim(N)$. This implies that

$$\begin{aligned}
&\mathbb{P}(\hat{\lambda}^F_{d_N+1} > c_{simp} \text{ or } \ldots \ \hat{\lambda}^F_d > c_{simp}) \\
&\leq \mathbb{P}(dim(\hat{N}) > dim(N))
\end{aligned}$$

It was shown in Section 7.1.2 that as $n \to \infty$, $\mathbb{P}(dim(\hat{N}) > dim(N))$ converges to 0. Because of this we need only concentrate on the behavior of $\hat{\lambda}^F_1, \ldots, \hat{\lambda}^F_{d_N}$. In other words

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(\hat{\lambda}^F_1 < c_{simp} \text{ or } \ldots \ \hat{\lambda}^F_k < c_{simp} \text{ or } \hat{\lambda}^F_{k+1} > c_{simp} \text{ or } \ldots \ \hat{\lambda}^F_d > c_{simp}) \\
&= \lim_{n\to\infty} \mathbb{P}(\hat{\lambda}^F_1 < c_{simp} \text{ or } \ldots \ \hat{\lambda}^F_k < c_{simp} \text{ or } \hat{\lambda}^F_{k+1} > c_{simp} \text{ or } \ldots \ \hat{\lambda}^F_{d_N} > c_{simp}) \\
&\leq \lim_{n\to\infty} \mathbb{P}(\hat{\lambda}^F_1 < c_{simp}) \ + \ \ldots \ + \ \lim_{n\to\infty} \mathbb{P}(\hat{\lambda}^F_k < c_{simp}) \\
&+ \ \lim_{n\to\infty} \mathbb{P}(\hat{\lambda}^F_{k+1} > c_{simp}) \ + \ \ldots \ + \ \lim_{n\to\infty} \mathbb{P}(\hat{\lambda}^F_{d_N} > c_{simp})
\end{aligned}$$

Since there is a finite number of terms in the sum, if it is shown that all of the terms converge to 0 then the proof will be completed. First focus on an estimated eigenvalue which corresponds to a true eigenvalue of F greater than c_{simp}. Without loss of generality, consider $\hat{\lambda}^F_k$.

$$\begin{aligned}
&\mathbb{P}(\hat{\lambda}^F_k < c_{simp}) \\
&= \mathbb{P}(\hat{\lambda}^F_k - \lambda^F_k < c_{simp} - \lambda^F_k) \\
&\leq \mathbb{P}(|\hat{\lambda}^F_k - \lambda^F_k| > |c_{simp} - \lambda^F_k|)
\end{aligned}$$

All terms involving an estimated eigenvalue that corresponds to a true eigenvalue of F greater than c_{simp} can be written in a similar manner as above. Next focus on an estimated eigenvalue less than c_{simp}. Without loss of generality, consider $\hat{\lambda}^F_{k+1}$.

$$\begin{aligned}
&\mathbb{P}(\hat{\lambda}^F_{k+1} > c_{simp}) \\
&= \mathbb{P}(\hat{\lambda}^F_{k+1} - \lambda^F_{k+1} > c_{simp} - \lambda^F_{k+1}) \\
&\leq \mathbb{P}(|\hat{\lambda}^F_{k+1} - \lambda^F_{k+1}| > |c_{simp} - \lambda^F_{k+1}|)
\end{aligned}$$

Every term of the sum is written as an absolute value between an estimated eigenvalue with its corresponding true eigenvalue being greater than a constant. Therefore every term in the sum is bounded by $\mathbb{P}(\max_{1\leq i\leq d_N} |\hat{\lambda}^F_i - \lambda^F_i| > \min_{1\leq i\leq d_N} |c_{simp} - \lambda^F_i|)$, i.e.

$$\begin{aligned}
&\mathbb{P}(|\hat{\lambda}^F_i - \lambda^F_i| > |c_{simp} - \lambda^F_i|) \\
&\leq \mathbb{P}(\max_{1\leq i\leq d_N} |\hat{\lambda}^F_i - \lambda^F_i| > \min_{1\leq i\leq d_N} |c_{simp} - \lambda^F_i|)
\end{aligned}$$

Now an inequality from Stewart and Sun (1990), see Equation 4.6, can be used in order to complete the proof. The inequality states that

$$\max_{1\leq i\leq d_N} |\hat{\lambda}^F_i - \lambda^F_i| \leq \| \hat{F} - F \| .$$

Now if it can be shown that the maximum eigenvalue of $\hat{F} - F$ converges to 0 then the proof will be completed. So in order to complete the proof, the distribution of $a_n vec(\hat{F} - F)$ is used. If $\hat{F} - F$ is characterized as projection matrices then the distribution can be calculated. The following equality is useful

$$a_n vec(\hat{F} - F) = a_n vec(\hat{P}_N F^{full} \hat{P}_N - P_N F^{full} P_N).$$

Ideally Lemma 1 and the delta method could be used in order to show that the above converges in distribution to a multivariate normal. But again there is the problem that P_N is always equal to P_0, but $\hat{P}_N$ is not always equal to $\hat{P}_0$. So therefore the above can be written as

$$\begin{aligned}
&a_n vec(\hat{P}_N F^{full} \hat{P}_N - P_N F^{full} P_N) \\
&= a_n vec(\hat{P}_N F^{full} \hat{P}_N - \hat{P}_0 F^{full} \hat{P}_0 + \hat{P}_0 F^{full} \hat{P}_0 - P_N F^{full} P_N) \\
&= a_n vec(\hat{P}_N F^{full} \hat{P}_N - \hat{P}_0 F^{full} \hat{P}_0) + a_n vec(\hat{P}_0 F^{full} \hat{P}_0 - P_N F^{full} P_N)
\end{aligned}$$

In order to get the distribution above, we will focus on the separate parts of the sum. First we we look at the second part. Using Lemma 1 and the delta method $a_n vec(\hat{P}_0 F^{full} \hat{P}_0 - P_N F^{full} P_N)$ converges in distribution to a multivariate normal. Therefore if it can be shown that $a_n vec(\hat{P}_N F^{full} \hat{P}_N - \hat{P}_0 F^{full} \hat{P}_0)$ converges in probability to 0, then $a_n vec(\hat{F} - F)$ will converge to the same multivariate normal as $a_n vec(\hat{P}_0 F^{full} \hat{P}_0 - P_N F^{full} P_N)$ by Slutzky's lemma, see van der Vaart (1998).

It is also known that $\hat{P}_N = \hat{P}_0$, if the dimension of $\hat{N}$ is equal to dimension of N. Therefore for all $\epsilon > 0$

$$\begin{aligned}
&\mathbb{P}(a_n |vec(\hat{P}_N F^{full} \hat{P}_N - \hat{P}_0 F^{full} \hat{P}_0)| > \epsilon) \\
&= \mathbb{P}(|vec(\hat{P}_N F^{full} \hat{P}_N - \hat{P}_0 F^{full} \hat{P}_0)| > \frac{\epsilon}{a_n}) \\
&\leq 1 - \mathbb{P}(|a_n vec(\hat{P}_N F^{full} \hat{P}_N - \hat{P}_0 F^{full} \hat{P}_0))| = 0)
\end{aligned}$$

$$= 1 - \mathbb{P}(dim(\hat{N}) = dim(N))$$

In Section 7.1.2 it is shown that as $n \to \infty$ $\mathbb{P}(dim(\hat{N}) = dim(N)) \to 1$. Therefore $a_n vec(\hat{P}_N F^{full} \hat{P}_N - \hat{P}_0 F^{full} \hat{P}_0)$ converges in probability to 0.

Now if the asymptotic distribution of $a_n vec(\hat{P}_0 F^{full} \hat{P}_0 - P_N F^{full} P_N)$ is found then the asymptotic distribution of $a_n vec(\hat{F} - F)$ is known. By Lemma 1 it is known that as $a_n \to \infty$, $a_n vec(\hat{P}_0 - P_N)$ converges in distribution to a multivariate normal with mean 0 and covariance Σ_0, where P_N is interchangable with P_0. Also the condition $\| \Sigma - \hat{\Sigma} \| \leq \frac{\lambda_m - \lambda_{m+1}}{2}$ is satisfied with probability 1 as $n \to \infty$, which can be seen by Fact 1. By applying the delta method, where $\phi(P_0) = P_0 F P_0$, it is shown that as $a_n \to \infty$, $a_n vec(\hat{P}_0 F^{full} \hat{P}_0 - P_N F^{full} P_N)$ converges in distribution to a multivariate normal with mean 0 and covariance $\phi'^T \Sigma_0 \phi'$. This implies that $vec(\hat{P}_0 F^{full} \hat{P}_0 - P_N F^{full} P_N)$ converges in distribution to a multivariate normal with mean 0 and covariance $\frac{1}{a_n} \phi'^T \Sigma_0 \phi'$.
Therefore

$$\mathbb{P}(\| \hat{F} - F \| > \min_{1 \leq i \leq d_N} |c_{simp} - \lambda_i^F|)$$

converges to 0.
Which implies that

$$\mathbb{P}(\max_{1 \leq i \leq d_N} |\hat{\lambda}_i^F - \lambda_i^F| > \min_{1 \leq i \leq d_N} |c_{simp} - \lambda_i^F|)$$

converges to 0.

Thus the dimension of $\hat{S}$ converges in probability to the dimension of S. $\square$

7.2.3 Convergence In Probability of the Estimated Interesting Genetic Constraint Space

This section is going to investigate the convergence in probability of $\hat{S}$ to S. This will be shown using D_{sine} and D_{gap} as metrics of the distance between subspaces.

For this convergence again the work of Tyler (1981) will be followed closely. It will be helpful to restate Lemma 1 in terms of matrices which project onto eigenspaces of F and $\hat{F}$. In Section 7.2.2 it is shown that $a_n vec(\hat{F} - F)$ converges in distribution to a multivariate normal.

Let $P_{S,0}$ be the matrix which projects onto the subspace generated by the eigenvectors which correspond to the true eigenvalues $\{\lambda_1^F, \ldots, \lambda_k^F\}$. Also let $\hat{P}_{S,0}$ be the projection matrix of the subspace generated by the estimated eigenvectors which correspond to the estimated eigenvalues $\{\hat{\lambda}_1^F, \ldots, \hat{\lambda}_k^F\}$. Let P_S be the projection matrix of S and $\hat{P}_S$ be the projection matrix of $\hat{S}$.

Notice that $P_{S,0} = P_S$, but it is not always true that $\hat{P}_{S,0} = \hat{P}_S$. The estimated cases are not always equal, since $\hat{P}_S$ may have a larger or smaller dimension than $\hat{P}_{S,0}$. But in the case when the dimension of $\hat{S}$ is equal to S then $\hat{P}_{S,0} = \hat{P}_S$.

Therefore Lemma 1 can be stated with Σ replaced by F as

Lemma 2 Define the norm $\| B \|$=[max eigenvalue $(B'B)]^{\frac{1}{2}}$. If $\| F - \hat{F} \| \leq \frac{\lambda_k^F - \lambda_{k+1}^F}{2}$ then there is a sequence a_n, with $a_n \to \infty$, so that

$$a_n(P_{S,0} - \hat{P}_{S,0}) \to N_{S,0}$$

where vec$(N_{S,0})$ is distributed multivariate normal with mean 0 and covariance matrix $\Sigma_{S,0}$. The above lemma will be used in the proof of convergence in probability of $\hat{S}$ to S.

A similar fact to Fact 1 is also stated, which will be of help in the following proof.

Fact 2 Let $E_n^2 = \text{event}\{\| \Sigma - \hat{\Sigma} \| \leq \frac{\lambda_m - \lambda_{m+1}}{2} \cap \| F - \hat{F} \| \leq \frac{\lambda_k^F - \lambda_{k+1}^F}{2}\}$. Since $vec(\Sigma - \hat{\Sigma})$ converges in probability to 0 and $vec(F - \hat{F})$ converges in probability to 0,

it follows that

$$\lim_{n\to\infty} \mathbb{P}(E_n^2) = 1.$$

Theorem 14 Given the definition of S, $\hat{S}$, and c_{simp} from above, it follows that for $\forall\ \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon) = 0$$

Proof of Theorem 14 The proof of Theorem 14 is similar to the proof of Theorem 5. Lemma 2 is helpful in proving Theorem 14. First $\mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon)$ is broken into where E_n^2 occurs and where it does not occur.

$$\mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap E_n^2) + \mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap (E_n^2)^c).$$

In order to use Lemma 2 we would like to only concentrate on the first probability in the sum. Therefore it is shown that the second probability in the sum converges to 0.

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap (E_n^2)^c)$$
$$\leq \lim_{n\to\infty} \mathbb{P}((E_n^2)^c) = 0$$

The convergence to 0 can be seen by Fact 2. This convergence to 0 implies

$$\lim_{n\to\infty} \mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon) = \lim_{n\to\infty} \mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap E_n^2)$$

Also Lemma 2 is only helpful in the case when $\hat{P}_{S,0} = \hat{P}_S$. Therefore $\mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon)$ is separated into the cases where $\hat{P}_{S,0} = \hat{P}_S$ and where $\hat{P}_{S,0} \neq \hat{P}_S$. These cases are the same as $dim(S) = dim(\hat{S})$ and when $dim(S) \neq dim(\hat{S})$ respectively.

$$\mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap E_n^2)$$

$$= \mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap E_n^2 | \hat{P}_{S,0} = \hat{P}_S)\mathbb{P}(\hat{P}_{S,0} = \hat{P}_S)$$
$$+ \mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap E_n^2 | \hat{P}_{S,0} \neq \hat{P}_S)\mathbb{P}(\hat{P}_{S,0} \neq \hat{P}_S)$$

It is shown in Section 7.2.2 that $dim(\hat{S}) \rightarrow dim(S)$ in probability, therefore $\mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap E_n^2 | \hat{P}_{S,0} \neq \hat{P}_S)\mathbb{P}(\hat{P}_{S,0} \neq \hat{P}_S)$ converges in probability to 0. Also since $dim(\hat{S}) \rightarrow dim(S)$ in probability, it follows that $\mathbb{P}(\hat{P}_{S,0} = \hat{P}_S)$ converges in probability to 1. Therefore, it need only be shown that $\mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon \cap E_n^2 | \hat{P}_{S,0} = \hat{P}_S)$ converges in probability to 0 for the proof to be completed.

Following similar reasoning as the proof of theorem 5,

$$D_{sine}(S, \hat{S}) = \sqrt{\frac{1}{2} trace[(P_{S,0} - \hat{P}_{S,0})((P_{S,0} - \hat{P}_{S,0})^T]}$$

since the dimensions of $\hat{P}_{S,0}$ and $P_{S,0}$ are the same.
From Lemma 2 it can be seen that ($\hat{P}_{S,0}$ - $P_{S,0}$) converges in probability to 0. Then by the continuous mapping theorem $\sqrt{\frac{1}{2} trace[(P_{S,0} - \hat{P}_{S,0})((P_{S,0} - \hat{P}_{S,0})^T]}$ converges in probability to 0. This implies that $\mathbb{P}(D_{sine}(S, \hat{S}) > \epsilon)$ converges to 0. □

Next the case of when the metric D_{gap} is used to measure the distance between subspaces is considered.

Theorem 15 Given the definition of S, $\hat{S}$, and c_{simp} from above, it follows that for $\forall\ \epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_{gap}(S, \hat{S}) > \epsilon) = 0$$

Proof of Theorem 15 The proof of Theorem 15, follows from the observation that if D_{sine} converges to zero then so does D_{gap}. This is because if the trace of a matrix converges to zero then the largest eigenvalue also will converge to zero, for a positive definite matrix. This relationship is shown in Section 6.4. □

7.3 Hypothesis Test for a Given Subspace contained in S

The above sections developed the idea of the interesting genetic constraint space. Also an estimate, $\hat{S}$, for this subspace is given. A question of importance to a biologist might be if a particular subspace of biological interest, B, is a subspace of S. Often times B is generated by choosing directions of biological interest that look similar to a set of eigendirections of $\hat{F}$ which are included in the set of directions which generate $\hat{S}$. For example Figure 3.5 in Section 3.6.3 shows the simple curve basis of the caterpillar example. For this case the simplest curve looks to be almost the second smoothest curve of the full space simple curve basis. So a question may be if the second simplest direction of the full space is in the interesting genetic constraint space. This section will develop a test for determining if B is a subspace of the interesting genetic constraint space.

Given a subspace B, a hypothesis test will be developed to see if $B \subseteq S$. This will be done by showing that the projection of an orthonormal basis of B, denoted by A_B, onto S is equal to A_B. Let P_S be the projection matrix of S, and A_B be an orthonormal basis of B. Then the test will involve the hypothesis

$$H_0 : P_S A_B = A_B$$

$$vs$$

$$H_1 : P_S A_B \neq A_B$$

where if H_0 is true then $B \subseteq S$. This will be done by building on the work of Tyler (1981) and Tyler (1983) which develops a test for a given basis being contained in a subspace generated by a set of eigenvectors. A similar test, but only for the 1 dimensional case, is developed by Anderson (1963). Tyler shows a general asymptotic distribution of A_B -

$\hat{P}_S A_B$, where $\hat{P}_S$ is the matrix which projects onto $\hat{S}$, and uses it to develop several tests.

The work of Tyler (1981) depends on the matrix and estimate of the matrix from which the eigendirections which generate the true and estimated subspace are chosen. Therefore first it will be shown that $\hat{F}$ and F fit into the frame work of Tyler (1981). Once this is shown then the asymptotic distribution of $a_n(A_B$ - $\hat{P}_S A_B)$ under H_0 is known.

The first property of F that must be shown is that it is symmetric. This can be seen algebraically, starting with F^T

$$
\begin{aligned}
&= (P_N F^{full} P_N)^T \\
&= (P_N)^T (F^{full})^T (P_N)^T
\end{aligned}
$$

Note that since P_N is a projection matrix it follows that $P_N^T = P_N$. So therefore $F^T = P_N (F^{full})^T P_N$, and to complete the proof it should be shown that $(F^{full})^T = F^{full}$. The matrix $(F^{full})^T$

$$
\begin{aligned}
&= (4I_d - DD^T)^T \\
&= 4I_d^T - (DD^T)^T \\
&= 4I_d - (D^T)^T (D)^T \\
&= 4I_d - DD^T \\
&= F^{full}
\end{aligned}
$$

Notice that this is also true of $\hat{F}$ by the same reasoning.

The symmetry of F and $\hat{F}$ is shown above so the next property which needs to be shown is that $a_n(\hat{F} - F)$ converges in distribution to a multivariate normal distribution. It was already shown in Section 7.2.2 that $a_n vec(\hat{F} - F)$ converges in distribution to N_F,

which has a multivariate normal distribution with mean 0 and covariance $\phi'^T\Sigma_0\phi' = \Sigma_F$, if $\| \Sigma - \hat{\Sigma} \| \leq \frac{\lambda_m - \lambda_{m+1}}{2}$. This condition holds with probability 1 as $n \to \infty$

With these two properties, along with the condition $\| F - \hat{F} \| \leq \frac{\lambda_k^F - \lambda_{k+1}^F}{2}$ that holds with probability 1 as $n \to \infty$, the work of Tyler (1981) can be followed to find the asymptotic distribution of $a_n vec(A_B - \hat{P}_S A_B)$.

In order to show the asymptotic distribution of $a_n vec(A_B - \hat{P}_S A_B)$, using Lemma 2 is helpful. Again the fact that $\hat{P}_S$ is not always equal to $\hat{P}_{S,0}$ restricts us from using Lemma 2 immediately. But a similar argument as when the asymptotic distribution of $a_n vec(\hat{F} - F)$ was found can be made. Following this argument $a_n vec(A_B - \hat{P}_S A_B)$ is equal in distribution to $a_n vec(A_B - \hat{P}_{S,0} A_B)$, since $a_n vec(\hat{P}_S A_B - \hat{P}_{S,0} A_B)$ converges in probability to 0.

The work of Tyler (1981) can be followed exactly to show that $a_n vec(A_B - \hat{P}_S A_B)$ converges in distribution to N_{A_B}, which has a multivariate distribution with mean 0 and covariance matrix Σ_{A_B}. In order to define Σ_{A_B} some notation needs to be established.

Let $\omega = (\lambda_1^F \dots \lambda_k^F)$ and $P_{\lambda^{F,dist}}$ be the projection matrix onto the eigenspace associated with the eigenvalues equal to $\lambda^{F,dist}$. Therefore $P_S = \sum_{\lambda^{F,dist} \in \omega} P_{\lambda^{F,dist}}$. The estimated analog for this notation is $\hat{\omega} = (\hat{\lambda}_1^F \dots \hat{\lambda}_k^F)$ and $\hat{P}_S = \sum_{\hat{\lambda}^F \in \hat{\omega}} P_{\hat{\lambda}^F}$.

Using this notation

$$\Sigma_{A_B} = (A_B \otimes I) C_\omega^T \Sigma_F C_\omega (A_B \otimes I)$$

where $C_\omega = \sum_{\lambda^{F,dist} \in \omega} \sum_{\mu^{F,dist} \in \omega^c} (\lambda^{F,dist} - \mu^{F,dist}) P_{\lambda^{F,dist}} \otimes P_{\mu^{F,dist}}$. But often the true values are not known so the projection matrices and Σ_F can be replaced by consistent estimators. Therefore

$$\hat{\Sigma}_{A_B} = (A_B \otimes I) \hat{C}_\omega^T \hat{\Sigma}_F \hat{C}_\omega (A_B \otimes I)$$

where $\hat{C}_\omega = \sum_{\hat{\lambda}^F \in \hat{\omega}} \sum_{\hat{\mu}^F \in \hat{\omega}^c} (\hat{\lambda}^F - \hat{\mu}^{F,dist}) P_{\hat{\lambda}^F} \otimes P_{\hat{\mu}^F}$.

A consistent estimator of $\hat{\Sigma}_F = \phi'(P_N)\Sigma_\Sigma \phi'(P_N)$ is $\phi'(\hat{P}_N)\hat{\Sigma}_\Sigma \phi'(\hat{P}_N)$, where Σ_Σ is the

asymptotic variance of $a_n vec(\hat{\Sigma} - \Sigma)$. A consistent estimator $\hat{\Sigma}_\Sigma$ is given in Tyler (1981).

Therefore the hypothesis test given in Tyler (1981) and Tyler (1983) can be used to test the hypothesis

$$H_0 : P_S A_B = A_B$$

$$vs$$

$$H_1 : P_S A_B \neq A_B.$$

Notice that the test is based on the statistic $(A_B - \hat{P}_S A_B)$. The squared lengths of these columns summed together is the Euclidean sine metric if the dimension of both spaces is assumed to be equal. The basis of one space, i.e. A_B, is projected onto the other space, i.e. $\hat{P}_S A_B$. Then the length between the basis points and the projections, i.e. $(A_B - \hat{P}_S A_B)$, is used to determine the distance between subspaces.

Also a biologist may be interested in if a chosen subspace, B_1, is a genetic constraint space, i.e. in the nearly null space. For this case the work of Tyler (1981) and Tyler (1983) can be followed exactly. In this case the biologist may use the simple curve basis as a tool. The biologist can use the simple curve basis directions to determine a subspace that is "biologically interesting". Then this subspace and be tested to see if it is in the nearly null space, i.e. a genetic constraint space.

APPENDIX A

Algebraic Justification

A.1 Algebraic Justification of F^{full}

This section shows that F^{full}, introduced in Section 3.3, has the same eigenvectors as DD^T. Let the eigendecompostion of $DD^T = \Gamma_s \Lambda_s \Gamma_s^T$. This implies that the eigendecompostion of $-DD^T = \Gamma_s(-\Lambda_s)\Gamma_s^T$. We wish to add $4 \times I_d$ to $-DD^T$ in order to have all eigenvalues be greater than 0. In fact the eigenvalues of F^{full} will be the diagonal of $(4I_d - \Lambda_s)$.

We start with F^{full} equal to the matrix

$$
\begin{aligned}
& 4I_d - DD^T \\
&= 4I_d + \Gamma_s(-\Lambda_s)\Gamma_s^T \\
&= 4\Gamma_s\Gamma_s^T + \Gamma_s(-\Lambda_s)\Gamma_s^T \\
&= \Gamma_s 4I_d \Gamma_s^T + \Gamma_s(-\Lambda_s)\Gamma_s^T \\
&= \Gamma_s(4I_d - \Lambda_s)\Gamma_s^T
\end{aligned}
$$

This implies that F^{full} has eigenvectors Γ_s and eigenvalues which correspond to the diagonal of $(4I_d - \Lambda_s)$. The value 4 can be replaced with any constant greater than 4 and the above would still hold.

A.2 Algebraic Justification of F

This section will show that the eigendirections of F, introduced in Section 3.3, are the same as $P_N DD^T P_N$ and also the eigenvalues of F are equal to the eigenvalues of

$-DD^T$ plus 4 for $i = 1 \ldots d_N$ and 0 otherwise. Where the dimension of the null space is d_N. Let the eigendecomposition of $P_N DD^T P_N$ equal JMJ^T, where J is a matrix of the eigenvectors and M has the eigenvalues along the diagonal and 0's off the diagonal.

Let the eigendecompostion of $DD^T = \Gamma_s \Lambda_s \Gamma_s^T$. This implies that the eigendecompostion of $-DD^T = \Gamma_s(-\Lambda_s)\Gamma_s^T$. We wish to add $4 \times I_d$ to $-DD^T$ in order to have all eigenvalues be greater than 0. It was shown in Section A.1 that $F^{full} = \Gamma_s(4I_d - \Lambda_s)\Gamma_s^T$.

The matrix $P_N(-DD^T)P_N = P_N[\Gamma_s(-\Lambda_s)\Gamma_s^T]P_N = J(-M)J^T$, while the matrix $F = P_N\Gamma_s(cI_d - \Lambda_s)\Gamma_s^T P_N$. We wish to show that $F = J(Z + M)J^T$, where

$$Z = \begin{pmatrix} 4I_{d_N} & 0 \\ 0 & 0 \end{pmatrix}$$

which implies that F has the eigenvalues $4I_d - M$ for $i = 1 \ldots d_N$ and 0 otherwise and the eigenvectors J. This would mean that $P_N DD^T P_N$ and F have the same eigenvectors, and the directions have simplicity scores which are shifted by a constant except for those that are 0.

We start with the characterization of F as $P_N\Gamma_s(-\Lambda_s + 4I_d)\Gamma_s^T P_N$.

$$\begin{aligned} & P_N\Gamma_s(-\Lambda_s + 4I_d)\Gamma_s^T P_N \\ & = P_N\Gamma_s(-\Lambda_s)\Gamma_s^T P_N + P_N\Gamma_s(4I_d)\Gamma_s^T P_N \\ & = J(-M)J^T + P_N(4I_d)P_N \end{aligned}$$

The last step in the above equations is done by noticing that $J(-M)J^T = P_N\Gamma_s(-\Lambda_s)\Gamma_s^T P_N$ by definition. Also notice that $\Gamma_s(4I_d)\Gamma_s^T = 4\Gamma_s\Gamma_s^T = 4I_d$. Next $4I_d$ is replaced by $J4J^T$, so

$$
\begin{aligned}
&P_N \Gamma_s(-\Lambda_s + 4I_d)\Gamma_s^T P_N \\
&= J(-M)J^T + P_N J4J^T P_N
\end{aligned}
$$

By definition of P_N the first d_N eigenvectors of J will be in the null space and the others will not. Therefore the projection of J_i for $i = 1 \dots d_N$ onto N is equal to J_i. While the projection of J_i for $i = d_N + 1 \dots d$ onto N is equal to 0. Therefore

$$
\begin{aligned}
&J(-M)J^T + P_N J4J^T P_N \\
&= J(-M)J^T + [J_1, J_2, \dots, J_{d_N}, 0, \dots, 0]4[J_1, J_2, \dots, J_{d_N}, 0, \dots, 0]^T \\
&= J(-M)J^T + [J_1, J_2, \dots, J_{d_N}, 0, \dots, 0]Z[J_1, J_2, \dots, J_{d_N}, 0, \dots, 0]^T
\end{aligned}
$$

In the above the zero vectors can be replace by $[J_{d_N+1} \dots J_d]$ since the are being multiplied by 0.

$$
\begin{aligned}
&J(-M)J^T + [J_1, J_2, \dots, J_{d_N}, 0, \dots, 0]Z[J_1, J_2, \dots, J_{d_N}, 0, \dots, 0]^T \\
&= J(-M)J^T + JZJ^T \\
&= J(-M + Z)J^T
\end{aligned}
$$

This shows the $F = J(-M + Z)J^T$, so therefore F has eigenvectors J and eigenvalues which are the diagonal of $(-M + Z)$. The value 4 can be replaced with any constant greater than 4 and the above would still hold.

BIBLIOGRAPHY

Anderson T.W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* **34**, 122–148.

Anderson T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Bookstein F.L. (1978). *Measurement of Biological Shape and Shape Change*. Lecture Notes in Biomathematics. Springer-Verlang, New York.

Dryden I.L. and Mardia K.V. (1998). *Statistical Shape Analysis*. Wiley Series in Probability and Statistics. Wiley, New York.

Gaydos T.L. (2007). Additional web material. URL `http://www.unc.edu/~tgaydos`.

Gilbert L.I., Granger N.A. and Roe R.M. (2000). The juvenile hormone: historical facts and speculations on future research directions. *Insect Biochemistry and Molecular Biology* **30**, 617–644.

Hotelling H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–377.

Inselberg A. (1985). The plane with parallel coordinates. *The Visual Computer* **1**, 69–91.

Izem R. and Kingsolver J.G. (2005). Variation in continuous reaction norms: Quantifying directions of biological interest. *The American Naturalist* **166**, 277–289.

Kato T. (1966). *Perturbation Theory for Linear Operators*. Die Grundlehren der mathematischen Wissenschaften in Einseldarstellungen ; Bd. 132. Spinger-Verlang, Berlin; New York.

Kendall D.G. (1999). *Shape and Shape Theory*. Wiley Series in Probability and Statistics. Wiley, New York.

Kingsolver J.G., Gomulkiewicz R. and Carter P.A. (2001). Variation, selection and evolution of function valued traits **112-113**, 87–104.

Kingsolver J.G., Ragland G.J. and Shlichta J.G. (2004). Quantitative genetics of continuous reaction norms: Thermal sensitivity of caterpillar growth rates. *Evolution* **58**, 1521–1529.

Kuss M. and Graepel T. (2003). The geometry of kernel canonical correlation analysis. Technical Report TR-108, Max Planck Institute for Biological Cybernetics, Tübingen, Germany. URL `citeseer.ist.psu.edu/kuss03geometry.html`.

Lynch M. and Walsh B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland,Massachesetts.

Meyer K. (1988). Dfreml: a set of programs to estimate components uncer an individual animal model. *Journal of Dairy Science* **2**, 33–34.

Meyer K. (1989). Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genetics Selection and Evolution* **21**, 317–340.

Meyer K. (1998). *DRREML, version3.0β user notes*.

Muirhead R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Nijhout H.F. (1994). *Insect Hormones*. Princeton University Press, Princeton, N.J.

Ramsay J.O. and Silverman B.W. (2002). *Applied Functional Data Analysis: methods and case studies*. Springer Series in Statistics. Springer, New York.

Ramsay J.O. and Silverman B.W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, New York.

Riddiford L.M., Hirune K., Zhou X. and Nelson C.A. (2003). Insights into the molecular basis of the hormonal control of molting and metamorphosis from manduca sexta and drosophila melanogaster. *Insect Biochemistry and Molecular Biology* **33**, 1327–1338.

Schatzman M. (2002). *Numerical Analysis: A Mathematical Introduction*. Claredon Press, Oxford.

Scholkopf B., Smola A.J. and Muller K.R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**, 1299–1319.

Searle S.R., Casella G. and McCulloch C.E. (1992). *Variance Components*. Wiley Series in Probability and Statistics. Wiley, New York.

Stewart G.W. and Sun J.g. (1990). *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, San Diego.

Tyler D.E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics* **9**, 725–736.

Tyler D.E. (1983). A class of asymptotic tests for principal component vectors. *The Annals of Statistics* **11**, 1243–1250.

van der Vaart A.W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

Vapnik V.N. (1995). *The Nature of Statistical Learning*. Statistics for engineering and information science. Springer, New York.

Wand M.P. and Jones M. (1995). *Kernel Smoothing*. Monographs on statistics and applied probability ; 60. Chapman and Hall, London.