

Bayesian Models for Relating Gene Expression and
Morphological Shape Variation in Sea Urchin

Larvae

by

Daniel Erskine Runcie

Department of Statistical Science
Duke University

Date: _____

Approved:

Scott C. Schmidler, Supervisor

Sayan Mukherjee

Gregory A. Wray

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2012

ABSTRACT

Bayesian Models for Relating Gene Expression and
Morphological Shape Variation in Sea Urchin Larvae

by

Daniel Erskine Runcie

Department of Statistical Science
Duke University

Date: _____

Approved:

Scott C. Schmidler, Supervisor

Sayan Mukherjee

Gregory A. Wray

An abstract of a thesis submitted in partial fulfillment of the requirements for
the degree of Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2012

Copyright © 2012 by Daniel Erskine Runcie
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

A general goal of biology is to understand how two or more sets of traits in an organism are related - for example, disease state and genetics, physiology and behavior, or phenotypic variation and gene function. Many of the early advancements in statistical analysis dealt with relating measured traits when one could be represented as a single number. However, many traits are inherently multi-dimensional, and technologies are advancing for rapidly measuring many types of such highly complex traits. Making efficient use of these new, larger datasets requires new statistical models for biological inference. In this thesis, I develop a method for relating two very different types of traits in sea urchin larvae: morphological shape, and developmental gene expression. In particular, I develop an approach for regression modeling using *shape* as a response variable. I use this method to address the question of whether variation in the expression of regulatory genes during development predicts later morphological variation in the larvae. I propose a hierarchical random effects factor regression model with shape as a response variable for relating morphology and gene expression when the individuals in each dataset are related, but not identical. I fit an approximation to the general model by breaking it into three discrete steps. I find that gene expression can explain $\sim 25\%$ of mean symmetric form variation among cultures of related larvae, and identify several groups of related genes that are correlated with aspects of morphological variation.

Contents

Abstract	iv
List of Tables	vii
List of Figures	viii
Acknowledgements	ix
1 Introduction	1
2 Shape statistics of sea urchin larvae	5
2.1 The data	7
2.2 Statistical shape analysis	7
2.3 Analytical Methods	9
2.3.1 Generalized Procrustes Analysis	9
2.3.2 Symmetric GPA	10
2.4 Results	11
2.5 Alternative Registration Techniques	13
2.5.1 Non-isotropic variances	13
2.5.2 Reference point analysis and growth models	17
2.6 Conclusions	18
3 Estimation of culture mean shapes by a hierarchical factor model	19
3.1 Model	21
3.1.1 Identifiability of factors	23

3.2	Priors	24
3.3	Gibbs sampler	26
3.4	Choice of k	29
3.5	MCMC inference	30
3.5.1	Posterior summarization	30
3.5.2	Convergence diagnostics	30
3.6	Results	31
4	Gene expression correlates of shape variation	35
4.1	Data	36
4.2	Modeling strategy	37
4.3	Selecting the tuning parameter λ	39
4.4	Results	40
4.5	Discussion	43
5	Joint model of shape and gene expression	47
5.1	Joint model for variation in gene expression and morphological shape	48
5.2	Inference on joint model	52
5.3	Comparison of 3-step model and joint model	53
	Bibliography	56

List of Tables

4.1	Bootstrap scores for gene expression factor regression	42
-----	--	----

List of Figures

2.1	Example <i>S. purpuratus</i> larva.	6
2.2	Larval shape variation.	12
2.3	Principle Component axes of shape variation.	14
3.1	Fitted magnitudes of culture mean shape vectors	32
3.2	Robustness of factor model estimates	33
3.3	Fitted differences in culture mean shape	34
4.1	Gene expression data.	38
4.2	Shape variation factors correlated with gene expression	44

Acknowledgements

I would like to thank my Master's committee, Scott, Sayan and Greg for their help throughout this project, and Sayan in particular for encouraging me to pursue a more formal training in statistics. I would also like to thank David Garfield for leading The Big Cross project to measure genetic variation in sea urchin development from which the data in this thesis originated, as well as Courtney Babbitt, BJ Nielsen, Ralph Haygood, Alex Primos, Tonya Severson and several others for their efforts to pull off the huge experiment. Thanks also to Lisa Pfefferle and Jennifer Wygoda in my lab for their encouragement, and members of the Schmidler group for their comments and suggestions on the analyses I present here. Finally, I would like my wife, Alex, and son, Matias, for tolerating some long nights and absent periods as I pursued this work.

1

Introduction

This thesis examines the problem of finding relationships between gene expression variation and morphological shape variation.

An organism's shape - the arrangement of its morphological features in space - is one of its most recognizable characteristics. In multicellular organisms, the spatial relationships of individual cells and structures determine the shapes of organs, appendages and body plans (Alberts, 2008). Since Darwin, variations within and among species in shape have been held as some of the most dramatic examples of evolutionary change (Raff, 1996). However, beyond measuring a handful of linear lengths and proportions among structures, until relatively recently researchers found quantifying shape variation to be a challenge. The field of shape analysis was revolutionized by two principle developments: digitizing equipment capable of recording the relative positions of features on organisms, and the mathematical fields of geometric morphometrics and statistical shape analysis (Goodall and Mardia, 1991; Dryden and Mardia, 1998; Kendall, 1984; Bookstein, 1986). Over the past several decades, these tools have added a new dimension to the evolutionary study of shape (Klingenberg, 2010).

A major challenge in modern evolutionary biology is the determination of the genetic basis of phenotypic shape variation. In particular, how does molecular variation in genes and developmental processes lead to variation in morphological shape? The shape of organs and appendages can be tightly controlled through development. Developmental processes are themselves controlled by suites of genes in gene networks, and much of the fields of Developmental Biology and Systems Biology are focused on identifying how these gene networks function. Significant control of development in animals is known to be exerted through the regulation of gene expression. By turning the expression of regulatory genes on or off at particular times and in particular locations, organisms control the identity and positioning of the cells that form physical structures. Therefore, the expression of the key regulatory genes during development may provide insight into the genetic and developmental basis of phenotypic shape variation.

In this thesis, I analyze data from an experiment that characterized variation in gene expression and morphology among a naturally-derived population of sea urchin larvae. The data consist of eight morphological landmarks digitized in three dimensions on 1,110 larvae from 72 separate cultures, and corresponding gene expression measurements. The gene expression measurements were taken on pools of several hundred larvae from each of the 72 cultures at seven time points spanning the development period from the earliest cell divisions until the time that the larval morphology was measured.

This experimental system was chosen because the developmental role of a large set of regulatory genes is very well characterized in sea urchin embryos, and experimental perturbations to these genes are known to affect embryonic and larval morphology (Oliveri et al., 2008; Peter and Davidson, 2010; Su et al., 2009; Peter and Davidson, 2011; Sharma and Ettensohn, 2011). Also, the functional and ecological consequences of variation in early larval morphology of sea urchins has been studied by several

authors (Hart and Strathmann, 1994; Strathmann, 2006; McEdward and Herrera, 1999). Analysis of these data in the original study (Garfield et al, *in prep*) found several intriguing correlations between the expression of particular genes and aspects of larval morphology. Here, I have expanded upon this to develop a more thorough statistical analysis of how variation in the known sea urchin embryonic developmental pathways might lead to later variation in larval morphology.

The statistical analyses in this thesis focus on the problem of relating two multi-dimensional data sets - morphology and gene expression - when the individuals measured in each dataset are related by an experimentally defined structure. Throughout, I use the word *shape* to refer to “all the geometric information that remains when location and rotational effects have been removed,” which is more classically called *size-and-shape* or *form* (Dryden and Mardia, 1998). I first partition total shape variation into its symmetric and asymmetric components to focus on the variation most likely to be relevant at the among culture level. I then develop a model that relates both symmetric shape and gene expression to variation in underlying latent processes in sea urchin embryos, and then model variation in these latent processes across cultures, uniting the morphology and gene expression measurements. I tailor this analysis to the goals of both predicting shape variation based on gene expression, and selecting genes to validate through future experiments. I find that variation in several genes is predictive of variation in specific aspects of larval shape variation. Shape variation related to skeletal growth are more correlated with gene expression variation than aspects of variation that appear to reflect more transient (ex. muscular) processes in the larvae.

In Chapter 2, I explore methods for quantifying shape variation among larvae. I describe the Generalized Procrustes algorithm for shape registration and an extension for parsing symmetric and asymmetric shape variation. I propose two modifications to this method that may allow more accurate or biologically interpretable descriptions

of shape variation. I then briefly characterize overall shape variation in the larvae.

In Chapter 3, I derive a model to characterize variation in larval shape among the 72 cultures. I use a Bayesian hierarchical normal factor model to find a low-dimensional representation of between-culture shape variation and among-individual residual shape variation. This model fits lower-dimensional latent shape traits that capture the majority of among-culture shape variation. I use these estimated latent traits in Chapter 4 to identify correlations between gene expression and morphological variation among cultures.

In Chapter 4, I implement an iterative grouped-lasso algorithm to regress the latent shape traits of Chapter 3 on the high-dimensional gene expression measurements. This step completes the link between gene expression and morphology by identifying the genes most predictive of major aspects of variation in shape. A focus of this section is on respecting the rotational symmetry of the latent factor traits of Chapter 3 in the regression, which I accomplish with a modified Procrustes optimization algorithm.

In Chapter ??, I link the methods of Chapters 2-4 under a general joint model of morphology and gene expression, explicitly linking the shared parameters, and propose a Gibbs sampler for fitting this joint model. The joint model respects much of the intent of the separate step-wise model proposed in Chapters 2-4, but differs in several conceptual and practical points. I discuss these differences, and the associated computational challenges.

Shape statistics of sea urchin larvae

At approximately three days post-fertilization, the developing sea urchin embryo begins to form paired skeletal structures called “spicules” (Fig. 2.1). Each calcium carbonate spicule is a tri-radiate structure with three branches extending at roughly equal angles in a plane. The two spicules are calcium-carbonate crystals that form on opposite sides of the embryonic midline (animal-vegetal axis). The spicules grow over the next several days and weeks to form the larval skeleton, which gives the animal its three-dimensional shape during the larval period. This larval shape is critical for functions such as swimming and feeding, and variations in size and shape have been hypothesized to be selectively important. The skeletal shape is also developmentally plastic, responding to the nutritional status of the larva to optimize a life history strategy between maximal food collection and developmental rate.

In this study, larvae were fixed at five days post-fertilization for morphological analysis. At this stage, the initial skeletal form is clearly developed, although this form will continue to change shape, expand, and add new rods and spicules throughout the larval period. Example larvae are shown in Fig. 2.1. The three rods of each spicule are conventionally named the *body rod*, the *post-oral rod*, and the *anterolat-*

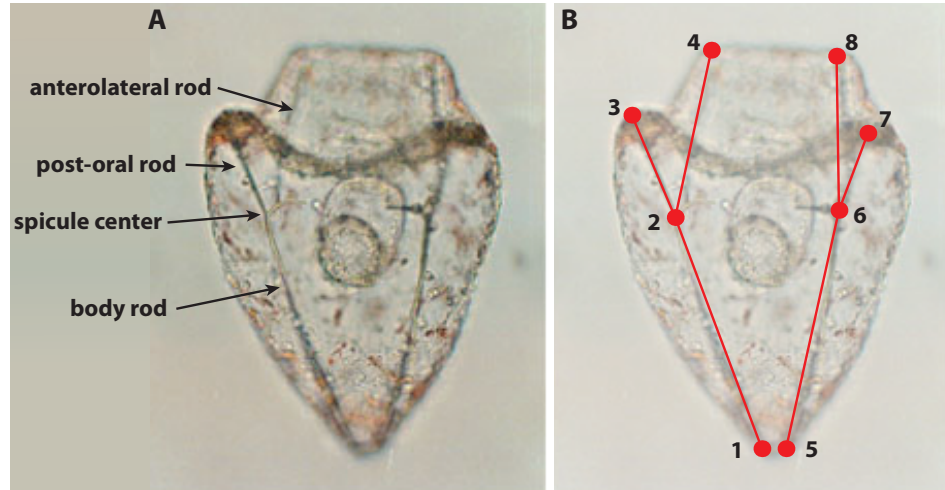


FIGURE 2.1: Example *S. purpuratus* larva. A) Photograph of a larva 5 days post fertilization. Rods are named on one spicule. B) Same image, with the 8 landmarks identified. Landmarks were located in three dimensions by photographing each larva at 2.5 or $5\mu\text{m}$ increments, and then identifying the x - and y - coordinates of the landmark in the image in which the landmark was most in-focus (z - coordinate).

eral rod, and are labeled *left* or *right* according to the side of the midline, with the *post-oral rods* in front of the *anterolateral rods*.

The goal of this chapter is to characterize variation in the size and shape of the larval skeleton among larvae. Since the larval skeleton is roughly bilaterally symmetrical, I decomposed the shape variation among larvae into a symmetric component representing variation affecting both sides of the larvae equally, and an asymmetric component representing differences between the two sides. I begin with details on how data on larval shape was recorded, followed by a brief statistical background on shape analysis. Next, I describe the methods used to quantify and visualize shape variation in the larva. I conclude with a prospective on alternative methods of shape analysis that may provide additional insight into larval growth and variation.

2.1 The data

At five days post fertilization, larval shape of 5-23 (median 17) larvae per culture (71 cultures, total 1,110 larvae) was digitized in three-dimensions by digital microscopy. Each larva was photographed at $\sim 200X$ magnification in bright-field with a Zeiss Axioskop 2 microscope, using the Axiovision v4.6 software. 15-30 photographs were taken of each larvae in vertical steps of 2.5 or $5\mu\text{m}$. Images were then imported into *ImageJ* v1.45s as a z -stack and eight points were digitized (Fig. 2.1B) representing the vertex and the tip of each rod of both spicules. For each point, the image in the z -stack in which the target point was most in-focus selected, and then the x - and y -coordinates were marked. These three coordinates (x -, y -, and z -) were then rescaled to μm based on a photograph of a scale micrometer. The scale of the z - axis was specified in the Axiovision software as the z -step size.

2.2 Statistical shape analysis

The eight digitized points on each larva are referred to as *landmarks*, which I assume to well-characterize the larva's skeletal structure. However, the raw Cartesian (x, y, z) coordinates are not directly suitable for statistical analysis, since the precise configuration of the larvae in the images is not standardized and contains no information relevant to growth or developmental state. This is a typical issue in shape analysis and can be dealt with using statistical shape analysis.

In the field of statistical shape analysis, *shape* is defined as all geometric information about an object after location, rotation and scaling have been removed (Dryden and Mardia, 1998). The related concept of *size-and-shape* refers to all geometric information after only location and rotation have been removed. I chose to analyze larval *size-and-shape* because I expect larval shape to change in a complex way with increasing size. Thus *shape* itself will hold little meaning for understanding larval

development. Below, and throughout this thesis, I will use the more familiar word *shape* to refer to *size-and-shape* as defined above.

Following the definitions in Dryden and Mardia (1998), shape is analyzed by forming the cartesian coordinates of each individual i into a configuration matrix. Let X_i be the $k \times m$ configuration matrix for larva i , representing the $k = 8$ points in $m = 3$ dimensions. The *shape* of individual i is $[X_i] = \{X_i\Gamma + \mathbf{1}_k\gamma^T : \Gamma \in SO(m), \gamma \in \mathbb{R}^m\}$. Here, Γ is a rotation matrix of dimension m such that $\Gamma^T\Gamma = \mathbf{I}_m$, and $|\Gamma| = +1$, \mathbf{I}_m is the $m \times m$ identity matrix and $\mathbf{1}_k$ is a column vector of k ones. Thus, $[X_i]$ is the equivalence class of configurations over the operations of rotation and translation. Any two larvae with configuration matrices X_1 and X_2 , have the same *shape* if there exists a rotation matrix Γ , and a translation vector $\gamma \in \mathbb{R}^m$ such that $X_1 = \Gamma X_2 + \mathbf{1}_k\gamma^T$.

To describe the differences among larval shapes, I require a measure of distance. The partial procrustes distance d_p is the appropriate metric for *shape* space (Dryden and Mardia, 1998). The partial Procrustes distance between two shapes $[X_1]$ and $[X_2]$ minimizes the Euclidean distances between their configuration matrices over translation and rotation (but not scale):

$$d_p(X_1, X_2) = \inf_{\Gamma \in SO_k, \gamma \in \mathbb{R}^m} \|X_1 - X_2\Gamma - \mathbf{1}_k\gamma^T\| \quad (2.1)$$

where $\|X\| = \sqrt{\text{trace}(X^T X)}$ is the Frobenius norm of X .

To analyze variation in *shape*, it is convenient to project shapes into the Procrustes tangent space (Goodall and Mardia, 1991), which is a linearized tangent space to *shape* space. To calculate coordinates in the Procrustes tangent space, I chose a central (mean) shape, or pole, and calculate a *registered* configuration of each larva's coordinates by minimizing the partial Procrustes distance (2.1). I call this mean shape \bar{X} (defined below (2.3)), find registered coordinates for larva i as

$\hat{X}_i = X_i \hat{\Gamma}_i + \mathbf{1} \hat{\gamma}_i^T$, and then calculate the Procrustes tangent space coordinates as:

$$X_i^* = \hat{X}_i - \bar{X} \quad (2.2)$$

2.3 Analytical Methods

2.3.1 Generalized Procrustes Analysis

I used the technique of Generalized Procrustes Analysis (GPA, Dryden and Mardia, 1998) to simultaneously find the optimal mean shape \bar{X} , and registered configurations X_i^* which minimize the total partial Procrustes distances in shape space. GPA iteratively finds a set of rotation matrices, Γ_i , translation vectors γ_i and a mean shape \bar{X} to minimize the partial Procrustes distances among configurations $i = \{1, \dots, n\}$:

$$\{\hat{\Gamma}_i, \hat{\gamma}_i, \bar{X}\} = \inf_{\Gamma_i, \gamma_i, \bar{X}} \sum_{i=1}^n \|X_i \Gamma_i - \mathbf{1}_N \gamma_i^T - \bar{X}\| \quad (2.3)$$

Following Dryden and Mardia (1998), the GPA algorithm is:

1. Calculate optimal translation vectors γ_i as the mean of the columns of X_i .

Subtract the matrix $\mathbf{1}_k \gamma_i^T$ from X_i to form the centered configuration X_i^c :

$$\gamma_i = \frac{1}{k} \sum_{j=1}^k \mathbf{x}_j^i$$

$$X_i^c = X_i - \mathbf{1}_k \gamma_i^T$$

where \mathbf{x}_j^i is the j th row of X_i .

2. Estimate the mean shape, given initial estimates of Γ_i :

$$\hat{X} = \frac{1}{n} \sum_{i=1}^n X_i^c \Gamma_i$$

3. Estimate the rotation matrices Γ_i as:

$$\hat{X}^T X_i^c = USV^T$$

$$\Gamma_i = VU^T$$

and registered configurations X'_i as

$$X'_i = X_i \Gamma_i$$

where USV^T is the singular value decomposition of $\hat{X}^T X_i^c$ such that S is diagonal, U and V are orthogonal matrices, and the m th column of V is multiplied by -1 if $|V| = -1$.

4. Repeat 2-3 until convergence of $\{\hat{\Gamma}_i, \hat{\gamma}_i, \bar{X}\}$. For the larval configuration data, convergence was achieved in 19 iterations

Finally, the registered coordinates can be projected onto the tangent space using \bar{X} as a pole as in (2.2).

2.3.2 Symmetric GPA

When shapes include a plane of symmetry such that there is a relation between corresponding landmarks on opposite sides, it can be informative to decompose total shape variation into a component that affects both sides of the figure equivalently (symmetric component), and a component which describes the differences between the sides (asymmetric component). (Mardia et al., 2000; Klingenberg et al., 2002) developed a variation on the classic GPA algorithm to quantify symmetric and asymmetric shape variation from landmark data.

In brief, a reflected dataset is created as $Y_i = QX_iH$, where $H = \text{diag}(-1 \ 1 \ 1)$ is a reflection matrix and Q is a $k \times k$ permutation matrix that re-orders the rows of X_i so that landmarks of the left side are matched to the corresponding landmarks of the right side. Here, the reflection occurs across the x axis in the original

coordinate system. However, this choice is arbitrary because it occurs before the rotations are calculated by GPA. GPA is then performed jointly on the two data sets: $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$. After registration, the symmetric shape of individual i is: $X_i^s = \frac{1}{2}(X_i^c + Y_i^c)$, and its asymmetric shape is: $X_i^a = \frac{1}{2}(X_i^c - Y_i^c)$. Symmetric and asymmetric tangent space projections for X_i^s can be calculated at poles \bar{X} and $\mathbf{0}$, respectively.

Based on the terminology of Mardia et al. (2000), the sea urchin larvae exhibit matching symmetry, since the plane of symmetry between the left and right spicules crosses part of the figure (in particular, the relative positions of the two spicules is part of the overall larval shape). I applied the matching symmetry GPA algorithm to the larval shape data. At this early stage of development, little asymmetry has developed in the larvae. Therefore, most asymmetric variation is expected to be random developmental noise among individuals and is unlikely to be in a consistent direction among similarly staged cultures. I hypothesized that symmetric shape variation would be more likely to correlate with gene expression since gene expression was measured in pooled larvae and thus averaged over hundreds of individuals.

2.4 Results

Fig. 2.2 shows the variation in each of the eight landmarks on the larva. Variation is present in all directions at each landmark and appears approximately elliptical in distribution. The clouds of points around landmark pairs (1,5) and (3,7) appear to be somewhat elongated in the direction of the body rods and post-oral rods, respectively, suggesting that some of the variation in these landmarks is related to the lengthening of the rods. Variation around landmark pair (4,8) appears spherical, while variation around the landmark pair (2,6) representing the vertexes of the spicules is slightly flattened in the direction perpendicular to the plane defined by the body rods. However, some of these patterns of variation may result from the

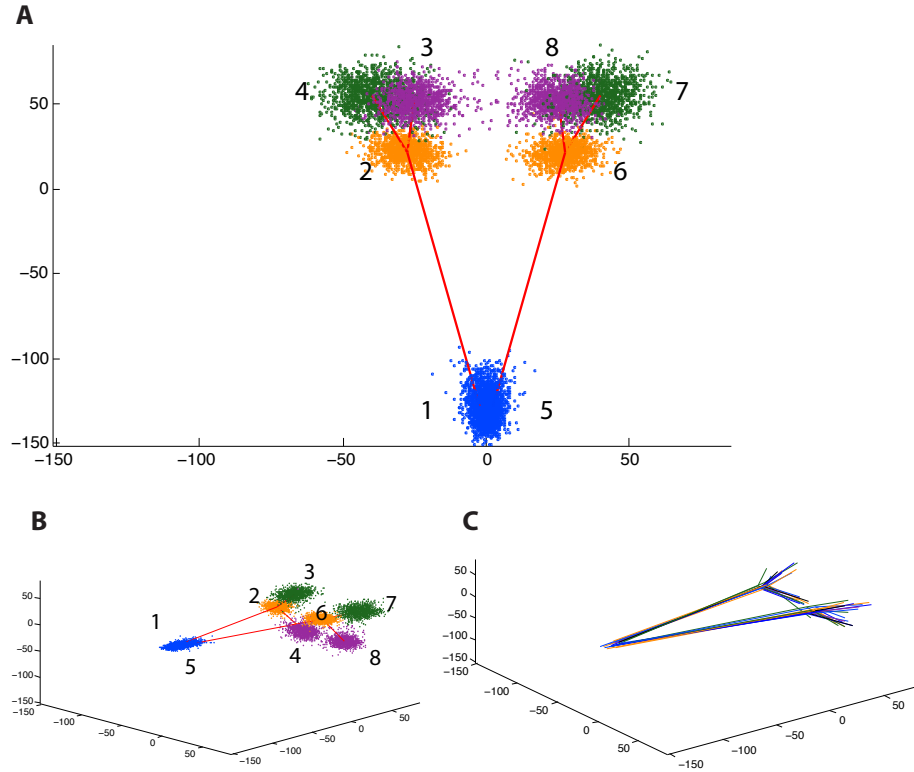


FIGURE 2.2: Representations of variation in larval shape. A) and B) show two views of the clouds variation in larval shape among the > 1000 larvae. Each point represents the registered (rotation and translation removed) coordinates of one of the eight landmarks of one larva. Points are colored according to the landmark number (see Fig. 2.1). C). Skeletal shapes are diagrammed for 10 randomly selected larvae after registration.

Procrustes registration criteria, rather than purely biological variation. For example, the lack of variation in landmarks 1 and 5 perpendicular to the midline axis of the larva may be due to the fact that these points are relatively isolated in space from all other points, and thus may have a large influence on the Procrustes fit.

To coarsely characterize *shape* variation, I used principle components analysis to assess covariation among the landmarks in the symmetric shape tangent space and of the asymmetric shape residuals. The symmetric shape tangent space coordinates have nine degrees of freedom. The first principle component of this variation ac-

counted for 33% of the variation, and the first five principle components accounted for > 90% of the variation. The asymmetric shape residuals also have nine degrees of freedom. The first principle component of this variation accounted for 44% of the asymmetric variation, and the first six principle components were needed to account for > 90% of the variation. In total, the symmetric component accounted for 77% of the total variation in shape.

Figure 2.3 shows representations of several of the predominant principle component axes of symmetric and asymmetric shape variation. To visualize principle components of symmetric shape tangent space coordinates, I used the inverse-projection of the principle component eigenvectors (\mathbf{v}_j) in the tangent space to an icon (X_{v_j}) in Cartesian coordinates:

$$X_{v_j} = \text{mat}(\mathbf{v}_j + \text{vec}(\bar{X})) \quad (2.4)$$

where $\text{mat}(\cdot)$ is the inverse vectorization operator that re-shapes a $km \times 1$ vector into a $k \times m$ matrix by column.

To visualize principle components of asymmetric variation, I simply added the inverse-vectorization of these principle component eigenvectors to \bar{X} .

2.5 Alternative Registration Techniques

Here, I propose two model extensions for improving the registration and analysis of larval shapes.

2.5.1 *Non-isotropic variances*

The GPA technique is appropriate if the shape variation is normal and isotropic:

$$\begin{aligned} X_i &= (\bar{X} + E_i)\Gamma_i^T + \mathbf{1}_k\gamma^T \\ \text{vec}(E_i) &\sim N(\mathbf{0}, \sigma^2\mathbf{I}_{km}) \end{aligned} \quad (2.5)$$

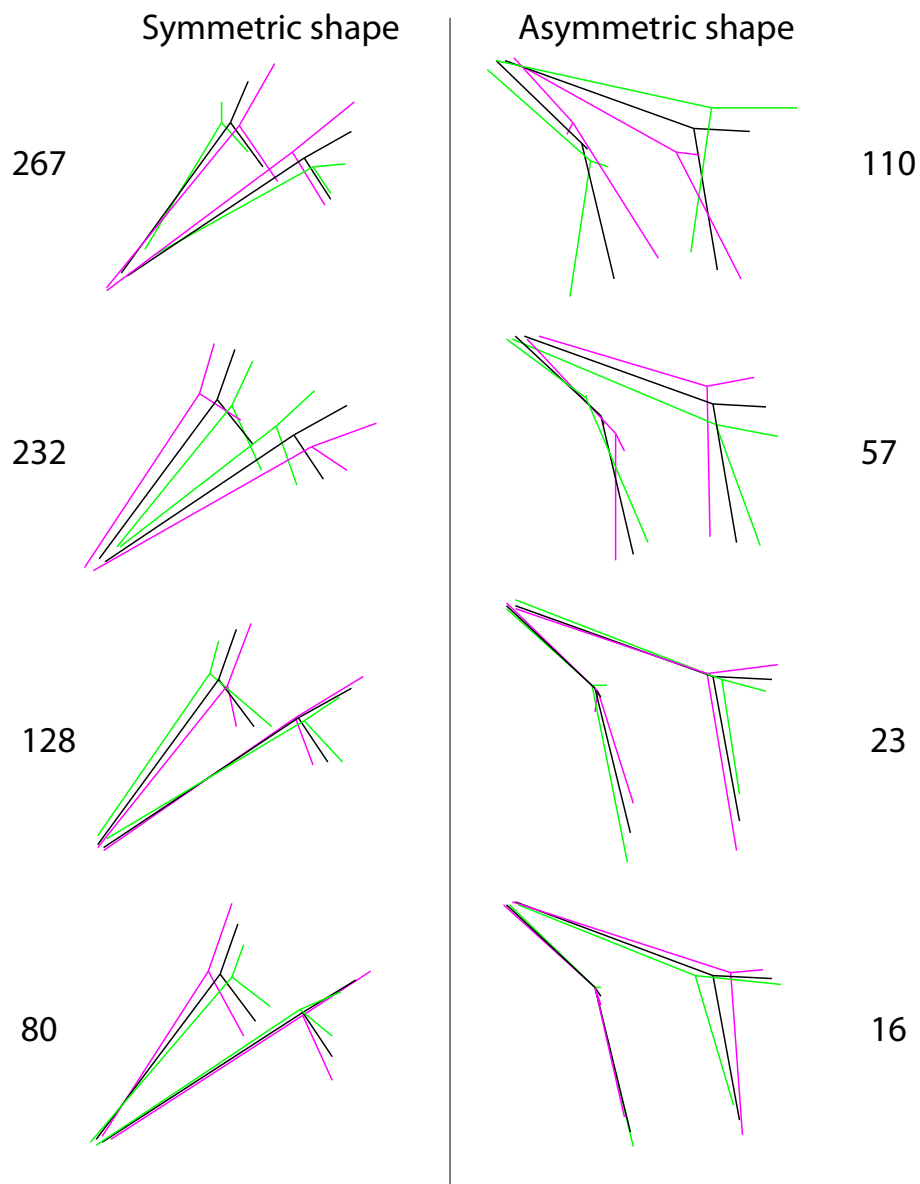


FIGURE 2.3: Representations of shape variation along the principle component axes of symmetric shape variation (left), and asymmetric shape variation (right). The variance in symmetric or asymmetric shape space of each axis is listed next to the visualization of the shape change. In each image, the black shape is the population mean shape, \bar{X} . The magenta and green shapes are 3 population standard deviations in the positive and negative directions along each principle components axis for the symmetric axes, and 5 standard deviations for the asymmetric axes. The two sets of axes are rotated differently to highlight the variation that they capture.

However, neither biological variation, nor measurement error in the digitization of larval coordinates is likely to fit this distribution. Biological variation results from the processes of growth and development in the larva, and the rates of growth at each vertex of the spicules vary over time, with earlier growth dominated by lengthening of the body rods, before switching to the postoral rods, and later, the anterolateral rods. Also, the growth of each spicule is constrained, initially by the crystal structure of the calcite mineral, and later by the tension of the ectoderm, which likely leads to non-isotropic variation. Measurement error depends on the configuration of the larva in the microscope's field, the ability of the user to accurately identify each landmark, and the resolution of the microscope. As described below, the resolution in the z direction was much lower than in the x and y directions.

A variance component model that reflects different (and possibly non-isotropic) distributions for the biological and technical variation may be more appropriate:

$$\begin{aligned}
 X_i &= (\bar{X} + E_i^b)\Gamma_i^T + \mathbf{1}_k\gamma^T + E^m \\
 \text{vec}(E_i^b) &\sim \text{N}(\mathbf{0}, \Sigma_b) \\
 \text{vec}(E_i^m) &\sim \text{N}(\mathbf{0}, \Sigma_m)
 \end{aligned} \tag{2.6}$$

where E_i^b represents biological variation around the mean shape, with covariance Σ_b , and E_i^m represents measurement error, after the larva has been rotated and translated on the microscope slide, with covariance Σ_m .

If I assume that the measurement error is due entirely to the resolution of the microscope (in particular, that the correct pixel was digitized for each landmark), then the measurement errors in the x , y , and z directions will be uncorrelated:

$$\begin{aligned}
 \Sigma_m &= D \otimes \mathbf{I}_k \\
 D &= \begin{bmatrix} \sigma_{xy}^2 & 0 & 0 \\ 0 & \sigma_{xy}^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{bmatrix}
 \end{aligned}$$

where σ_{xy}^2 is related to the pixel size, and σ_z^2 to the z -step size used for the images. The pixel size at the 200X magnification was $0.317\mu m$ and the z -step size was $2.494\mu m$. Thus, I can calculate $\sigma_{xy}^2 = \frac{1}{12}0.317^2$ and $\sigma_z^2 = \frac{1}{12}2.494^2$.

Model (2.6) can be simplified to:

$$\begin{aligned} X_i &= \bar{X}\Gamma_i^T + \mathbf{1}_k\gamma^T + E_i \\ v(E_i) &\sim N(\mathbf{0}, D \otimes \mathbf{I}_k + \Gamma_i\Sigma_b\Gamma_i^T) \end{aligned} \quad (2.7)$$

where E_i captures both biological variation and measurement error.

Given model (2.7), the appropriate distance measure of shape variation is:

$$G_{\Gamma, \gamma}^{\Sigma}(X_i, \dots, X_n) = \sum_i^n \|\Sigma_i^{-\frac{1}{2}}v((X_i - \mathbf{1}_N\gamma_i^T)\Gamma_i - \bar{X})\| \quad (2.8)$$

with $\Sigma_i = D \otimes \mathbf{I}_k + \Gamma_i\Sigma_b\Gamma_i^T$.

In general, Σ_b will be unknown and must be estimated, for example by extending the iteratively re-weighted GPA algorithm of Goodall (1991), which iterates between finding the Γ_i given an estimate $\hat{\Sigma}_b$, and then re-estimating $\hat{\Sigma}_b$. However, in (2.8), the γ_i and Γ_i can no longer be estimated as in standard GPA above. There is no explicit solution for Γ_i given an unconstrained Σ_b and γ_i , but the algorithm of Viklands (2006) can be used to find a local minima solution. The γ_i and Γ_i must be estimated together, and so an iterative algorithm can be used:

1. find optimal translation, given an estimate of the rotation:

$$\gamma_i|\Gamma_i = (\mathbf{1}_{km}^T\Sigma^{-1}\mathbf{1}_{km})^{-1}\mathbf{1}_{km}^T\Sigma^{-1}v(X_i\Gamma_i - \bar{X})\Gamma_i^T$$

2. find optimal rotation, given an estimate of the translation (Viklands, 2006):

$$\Gamma_i|\gamma_i = \arg \min_{\Gamma_i|\Gamma_i\Gamma_i^T=\mathbf{I}_k} \|\Sigma^{-1/2}((X_i - \mathbf{1}_k\gamma_i^T) \otimes \mathbf{I}_m)v(\Gamma_i) - v(\bar{X})\|$$

This algorithm (in particular the estimation of Γ_i) is not guaranteed to find the globally optimal rotation matrix, but in practice appears to work well.

By constraining the measurement part of the error variance, model (2.7) may produce a more robust shape registration when the magnitude of measurement error is large relative to biological variance. However, the empirical variance of each coordinate in each dimension from the standard GPA registration (2.3) ranged from 6.8-85.0 (mean 44.2), while the estimated measurement error in the z direction had a variance of 0.52. Thus, in this data set, the maximum possible contribution of measurement error (at least as defined here) to total variance in any coordinate was only about 7%, and on average likely much less. Therefore, model (2.7) is unlikely to produce dramatically different results than standard GPA.

Additionally, the assumption in (2.7) that biological variation E_i^b has a multivariate normal distribution is likely violated in our model because the larvae are not independent. Larvae that were grown in the same culture dish shared environmental conditions, and were genetically related, and so their shapes are correlated. An explicit model of this non-random pattern of variation among larvae will be developed in Chapter 3.

2.5.2 Reference point analysis and growth models

An alternative approach is to re-formulate the problem into a time-indexed model of growth trajectories in three-dimensional space. Since the larval skeleton grows outwards from the two spicules, using the spicules as reference points may serve to ground the observed variation in a more interpretable and biologically meaningful manner relative to the growth of the animal.

One possibility is to identify the orientation and translation of each larva by forcing a triangle defined by three key landmarks to lie in a pre-defined plane, centered at the origin. Then, variation could be measured relative to some fixed (and poten-

tially biologically meaningful) axis of larval development. One choice would be to fix the triangle defined by landmarks 1, 2, and 6 (Fig. 2.1). This triangle relates to the anterior-posterior and left-right axes of the larval development, and is thus a natural plane from a developmental standpoint. This approach is similar to using the classical Bookstein coordinates, which define rotation and scaling relative to two landmarks (Bookstein, 1986). However, like the Bookstein coordinates (discussed in Dryden and Mardia, 1998), this triangle-based method will give the appearance of strong correlations in shape variability among the non-fixed landmarks, making the interpretation of such variation more difficult.

Another possibility would be to fix the spicules themselves as origins and measure variation in two parts: 1) relative arm positioning and length within each spicule, and 2) relative spicule positioning and rotation. This approach seems promising, but I have not developed it yet.

2.6 Conclusions

The tangent space coordinates from any of the discussed methods provide an measure of larval shape variation. Differences in models of registration may provide more biologically interpretable measures of shape variation. However, for this project, the tangent space residuals of symmetric shape variation seem to be the most useful description of larval shape.

Estimation of culture mean shapes by a hierarchical factor model

Describing and predicting shape variation, such as the variation among larval skeletal shapes described in Chapter 2, is challenging because shapes are necessarily multi-dimensional. In many multivariate situations, the scale of the problem can be dramatically reduced by assuming that the majority of the high-dimensional variation can be captured by a low-dimensional space. Inference of, and in, this lower-dimensional space can often ease computational burdens, and make results easier to interpret. Factor models are a powerful, and widely used technique for dimension reduction. In a factor model, a high-dimensional $p \times n$ matrix, X , of variables (p) and observations (n) is decomposed into two lower-dimensional matrices, A and S , with dimensions $p \times k$ and $k \times n$ and a residual matrix E . In most cases, $k \ll \min(p, n)$:

$$X = AS + E \tag{3.1}$$

Here, the matrix S contains k factors, and the matrix A the factor loadings. The factors define k new variables for each observation that capture most of the variation in the original p variables, while the factor loadings describe how the new factor variables relate to the original variables. If much of the variation in the original p

variables can be captured by these k new variables, the residuals will be small, and inferences based on A and S will generalize well to the full data set.

In Chapter 2, I quantified shape variation in sea urchin larvae based on the three-dimensional coordinates of eight landmarks on the larval skeleton. Thus, each larval shape is characterized by 24 variables. However, these 24 variables are highly correlated, even after registration by Generalized Procrustes Analysis, and the covariance itself is singular. Each larval shape, although described by 24 variables, in fact only has $24 - 3 - 3 = 18$ free parameters, because three parameters are lost to an arbitrary translation, and three to an arbitrary rotation towards the mean shape. Additionally, by extracting only symmetric shape variation, the coordinates of one side of a larva fully specify those of the other side, reducing the free parameters to nine. A principle components analysis of tangent space variation in symmetric shape finds eight principle component axes with non-zero eigenvalues, as expected. However, four of these axes capture $> 90\%$ of the total variation. Therefore dimension reduction on this shape dataset is important.

In the experiment described in Chapter 2 5 – 23 larvae were measured from each of $L = 71$ cultures. Within each culture, larvae were genetically related (in fact, they were all *full-sibs*), and culture dishes may have also varied in environmental parameters such as water quality and larval density. The among-culture differences in larval shape are the primary interest in this thesis, because such differences likely result at least partly from genetic differences among the larvae, and because we measured gene expression in pools of hundreds of larvae from each culture, and thus could not correlate individual gene expression differences with shape variation.

In this Chapter, I propose a Bayesian factor model to parsimoniously model the shape variation among larvae from Chapter 2, extend the model to fit mean shape variation among larval cultures, and derive a Gibbs algorithm to sample from the posterior distribution of the factor scores. In Chapter 4, I will use the fitted factors

as “shape traits” to test for gene expression correlates of shape variation.

3.1 Model

Let \mathbf{x}_{ij} be the $p \times 1$ vector of symmetric *shape*-tangent space coordinates of the left-size landmarks (points 1-4, Fig. 2.1 $p = 12$) of larvae $j = \{1, \dots, n\}$ in culture $i = \{1, \dots, L\}$. I model the vector \mathbf{x}_{ij} using the following variance component model:

$$\mathbf{x}_{ij} = \mathbf{u}_i + \mathbf{e}_{ij} \quad (3.2)$$

where \mathbf{u}_i is a $p \times 1$ vector representing the tangent space coordinates of the mean shape for culture i , and \mathbf{e}_{ij} is a vector of residuals.

Both \mathbf{u}_i and \mathbf{e}_{ij} are random p -dimensional vectors in the tangent space, which I assume have the following distributions:

$$\mathbf{u}_i \sim N_p(\mathbf{0}_p, A_u A_u^T + \Psi_u) \quad (3.3)$$

$$\mathbf{e}_{ij} \sim N_p(\mathbf{0}_p, A_e A_e^T + \Psi_e)$$

where $N_p(\mu, \Sigma)$ is the p -dimensional multivariate normal distribution with mean vector μ and covariance Σ , and $\mathbf{0}_p$ is a p -dimensional zero-vector. Dimension subscripts are omitted below unless necessary for clarity.

Here, A_u and A_e are $p \times k_u$ and $p \times k_e$ matrices, respectively, and Ψ_u and Ψ_e are $p \times p$ diagonal matrices. When $p(p+1)/2 - p(k+1) + k(k-1)/2 \geq 0$ for $k = \{k_u, k_e\}$, the covariances of \mathbf{u}_i and/or \mathbf{e}_{ij} are constrained. This dimension reduction makes the model for \mathbf{x}_{ij} more parsimonious, and reduces the number of parameters that must be estimated.

While full rank (for convenience), the covariance matrices in (3.3) are highly structured, reflecting the expected non-independence of the coordinates in \mathbf{x}_{ij} (and \mathbf{u}_i and \mathbf{e}_{ij}). The forms of these covariances arise naturally under the assumption that a larval shape (whether a culture mean shape or an individual residual shape)

is dependent on a low number of latent traits, or factors. A simple factor model for \mathbf{u}_i and \mathbf{e}_{ij} is:

$$\begin{aligned}\mathbf{u}_i &= A_u \mathbf{y}_i + \boldsymbol{\epsilon}_i \\ \mathbf{e}_{ij} &= A_e \mathbf{f}_{ij} + \boldsymbol{\epsilon}_{ij}\end{aligned}\tag{3.4}$$

where A_u and A_e are defined as above. The $k_u \times 1$ and $k_e \times 1$ vectors \mathbf{y}_i and \mathbf{f}_{ij} are latent factor traits for the mean shape of culture i and the individual residual shape of larva j in culture i . Under this model, the variability underlying any covariation among coordinates of culture mean shape has dimension k_u , and the variability of individual residual shape has dimension k_e . The matrices A_u and A_e are loading matrices that describe how these latent factor traits relate to the original p -dimensional tangent space coordinates.

I use the standard factor model assumptions that \mathbf{y}_i and \mathbf{f}_{ij} are independent and normally distributed with unit variance:

$$\begin{aligned}\mathbf{y}_i &\sim N_{k_u}(\mathbf{0}_{k_u}, \mathbf{I}_{k_u}) \\ \mathbf{f}_{ij} &\sim N_{k_e}(\mathbf{0}_{k_e}, \mathbf{I}_{k_e})\end{aligned}\tag{3.5}$$

and that the residual vectors are also independent and normally distributed, but with unique variances:

$$\begin{aligned}\boldsymbol{\epsilon}_i &\sim N(\mathbf{0}, \Psi_u) \\ \boldsymbol{\epsilon}_{ij} &\sim N(\mathbf{0}, \Psi_e)\end{aligned}\tag{3.6}$$

with Ψ_u and Ψ_e $p \times p$ diagonal matrices.

The full model over all individuals can be written hierarchically as:

$$\begin{aligned}X &= UZ + E \\ U &= A_u Y^T + E_u \\ E &= A_e F^T + E_e\end{aligned}\tag{3.7}$$

where $X = [\mathbf{x}_{11} \dots \mathbf{x}_{Ln}]$ is formed by concatenating the individual shape vectors into a single $p \times n$ matrix, $U = [\mathbf{u}_1 \dots \mathbf{u}_L]$ is formed similarly by concatenating the mean culture shapes, $Y = [\mathbf{y}_1, \dots, \mathbf{y}_L]^T$ and $F = [\mathbf{f}_{11}, \dots, \mathbf{f}_{Ln}]^T$ contain the latent factors for culture and individual residual as columns, E the individual larva residual vectors \mathbf{e}_{ij} , and Z is a $(L \times n)$ binary incidence matrix relating individuals to cultures. The model has no intercept because the coordinates \mathbf{x}_{ij} are tangent space residuals with the mean shape as the pole.

Given the distributions of \mathbf{y}_i and \mathbf{f}_i (3.5) and $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_{ij}$ (3.6), the matrices of culture means (U) and individual residuals (E) have distributions (conditional on A_u, A_e, Ψ_u and Ψ_e):

$$\begin{aligned} U &\sim \text{N}(U \mid A_u Y^T; \mathbf{I}_L, \Psi_u) \\ E &\sim \text{N}(E \mid A_e F^T; \mathbf{I}_n, \Psi_e) \end{aligned} \quad (3.8)$$

where $\text{N}(X \mid M; \Sigma, \Psi)$ is the matrix normal distribution for parameter X with mean matrix M , row covariance Σ and column covariance Ψ .

The likelihood of the data and latent “traits”, conditional on the factor loadings and error parameters is:

$$\begin{aligned} X, U, Y, F \mid A_u, A_e, \Psi_u, \Psi_e &\sim \text{N}(\text{vec}(X) \mid \text{vec}(UZ + A_e F^T), Z^T Z \otimes \Psi_u + \mathbf{I}_n \otimes \Psi_e) \\ &\times \text{N}(U \mid A_u Y^T; \mathbf{I}_r, \Psi_u) \times \text{N}(Y; \mathbf{I}_r, \mathbf{I}_{k_u}) \times \text{N}(F; \mathbf{I}_n, \mathbf{I}_{k_e}) \end{aligned} \quad (3.9)$$

where $\text{vec}(X)$ is the vectorization operator that stacks the columns of X into a single $pn \times 1$ vector.

3.1.1 Identifiability of factors

In the scope of this thesis, the factors Y (in particular) and F are the most important parameters of (3.7) because they are the larval shape traits that I test for correlation with gene expression variation. However, as is common with standard factor models,

these parameters are only identifiable in the likelihood (3.9) up to their product with an orthogonal matrix. For any $H \in O(k_u)$:

$$A_u Y^T = A_u H^T H Y^T = A'_u Y'^T \quad (3.10)$$

with $A'_u = A_u H^T$ and $Y'^T = H Y^T$. Unless H is the identity, $Y' \neq Y$ and $A'_u \neq A_u$, yet this substitution does not change the likelihood (3.9): $p(X, U, Y, F \mid A_u, A_e, \Psi_u, \Psi_e) = p(X, U, Y, F \mid A'_u, A_e, \Psi_u, \Psi_e)$.

Identifiability of Y and F can be ensured through appropriate priors on A_u and A_e (Aguilar and West, 2000; West, 2003). I use the lower-triangle parameterization of these matrices (see section 3.2) to make prior specification more straightforward, and to make assessing MCMC convergence easier (section 3.5). However, the lower-triangle form of the loading matrices is arbitrary and has no relation to the underlying biology of shape variation.

Instead, I use model (3.7) to infer the equivalence class:

$$[Y] = \{Y H^T \mid H \in O(k_u)\} \quad (3.11)$$

which I call the *shape* of Y based on the similarity between (3.11) and the definition of *shape* in statistical shape analysis (see section 2.2), except that no translation of Y is allowed. The matrix Y is then an *icon* of $[Y]$ and I treat all such icons, as well as the corresponding tangent space loadings $A_u H$, as equally plausible prior to observing gene expression (see Chapter 4).

3.2 Priors

For Bayesian inference on Y and F , given the data, X and model (3.9), priors are needed on the loading matrices A_u and A_e and on Ψ_u and Ψ_e .

Since in the marginal likelihood (after marginalizing over Y and F), A_u and A_e

only appear as the products $A_u A_u^T$ and $A_e A_e^T$:

$$X \mid A_u, A_e, \Psi_u, \Psi_e \sim \text{N}(\text{vec}(X) \mid \mathbf{0}, Z^T Z \otimes (A_u A_u^T + \Psi_u) + \mathbf{I}_n \otimes (A_e A_e^T + \Psi_e)) \quad (3.12)$$

A_u and A_e have fewer than the maximum $p \times k_u$ or $p \times k_e$ identifiable parameters. Therefore, a convenient prior for A_u and A_e forces the matrices to be lower-triangular, with positive elements on the diagonal. I used this prior for A_u :

$$A_{u(i,j)} \sim \begin{cases} 0 & \text{if } j > i \\ \text{N}^+(0, \sigma_u^2) & \text{if } j = i \\ \text{N}(0, \sigma_u^2) & \text{if } j < i \end{cases} \quad (3.13)$$

for $i = \{1, \dots, p\}$ and $j = \{1, \dots, k_u\}$ and where $\text{N}^+(\mu, \sigma^2)$ is the univariate normal distribution with mean μ , variance σ^2 , but left-truncated at zero. This prior assumes that the individual elements of A_u are independent. The prior on A_e was similar, with the parameters σ_u^2 and σ_e^2 estimating the magnitudes of culture (and individual) residual variation, respectively. Priors on σ_u^2 and σ_e^2 were specified as:

$$1/\sigma_u^2 \sim \text{Ga}(a_u/2, b_u/2) \quad (3.14)$$

$$1/\sigma_e^2 \sim \text{Ga}(a_e/2, b_e/2)$$

where $\text{Ga}(a/2, b/2)$ is the gamma distribution with shape $a/2$ and rate $b/2$. I chose $b_u = a_u$ and $b_e = a_e$ so that the prior on σ_u^2 and σ_e^2 had mean one. I set $a_u = a_e = 4$ so that these distributions were proper, but diffuse. To assess the sensitivity to this hyperparameter, I also ran the model with a more informative prior: $a_u = a_e = 20$.

Finally, priors on the diagonal elements of the matrices Ψ_u and Ψ_e , termed $\lambda_{u1} \dots \lambda_{uk_u}$ and $\lambda_{e1} \dots \lambda_{ek_e}$ were specified as:

$$1/\lambda_{ui} \sim \text{Ga}(g_u/2, h_u/2) \quad (3.15)$$

$$1/\lambda_{ej} \sim \text{Ga}(g_e/2, h_e/2)$$

Here I set $h_u = h_e = 4$ and $g_u = g_e = 20$ so that the prior on the residual variance of each coordinate was diffuse with a prior mean equal to 10% of the total variation

in that coordinate. To assess the sensitivity to this hyperparameter, I also ran the model $g_u = g_e = \{4, 200\}$.

All these priors are conjugate to the likelihood, and so facilitate the derivation of full conditional probabilities for a Gibbs sampler.

3.3 Gibbs sampler

I estimated the posterior distributions of the unknown parameters of model (3.9) by MCMC using a Gibbs sampler. Since conjugate priors were used for each parameter, conditional posteriors were available for each update.

The posterior over all parameters, θ is proportional to:

$$p(\theta | X) \propto p(X | U, A_e, F, \Psi_e) p(A_e | \sigma_e^2, \Psi_e) \\ \times p(U | A_u, Y, \Psi_u) p(A_u | \sigma_u^2, \Psi_u) \quad (3.16)$$

By breaking up the likelihood as in (3.16), it is clear that conditional on U , the model for $X - UZ$ has an identical form to the model for U . Therefore, the Gibbs updates for all parameters with subscript e are identical to the updates for parameters with subscript u , and I will only demonstrate the latter set here.

Below, let θ_{-x} represent all model parameters except x . The Gibbs sampler iterates through samples of U , Y and A_u , and the variances σ_u^2 and Ψ_u , each conditional on current estimates of all others parameters.

First, the conditional posterior of the culture mean shape matrix U is:

$$p(U | \theta_{-U}) \propto p(X|U, A_e, F, \Psi_e) p(U|A_u, S, \Psi_u) \\ \times \exp \left\{ -\frac{1}{2}(\text{vec}(X - A_e F^T) - \text{vec}(UZ))^T (\mathbf{I}_n \otimes \Psi_e^{-1})(\text{vec}(X - A_e F^T) - \text{vec}(UZ)) \right\} \\ \times \exp \left\{ -\frac{1}{2}(\text{vec}(U) - \text{vec}(A_u Y^T))^T (\mathbf{I}_r \otimes \Psi_u^{-1})(\text{vec}(U) - \text{vec}(A_u Y^T)) \right\}$$

which can be simplified to:

$$p(U | \theta_{-U}) \propto \exp \left\{ -\frac{1}{2}(\text{vec}(U)^T((ZZ^T \otimes \Psi_e^{-1}) + (\mathbf{I}_r \otimes \Psi_u^{-1}))\text{vec}(U) \right. \\ \left. - 2(\text{vec}(X - A_e F^T)^T(Z^T \otimes \Psi_e^{-1}) + \text{vec}(A_u Y^T)^T(\mathbf{I}_n \otimes \Psi_u^{-1}))\text{vec}(U)) \right\}$$

Therefore, the conditional posterior of $\text{vec}(U)$ is the multivariate normal distribution $\text{vec}(U) | \theta_{-U} \sim N_{pr}(\mu, \Sigma)$ with parameters:

$$\Sigma = ((ZZ^T \otimes \Psi_e^{-1}) + (\mathbf{I}_r \otimes \Psi_u^{-1}))^{-1} \\ \mu = \Sigma(\text{vec}(X - A_e F^T)^T(Z^T \otimes \Psi_e^{-1}) + \text{vec}(A_u Y^T)^T(\mathbf{I}_n \otimes \Psi_u^{-1}))$$

Next, the conditional posterior of the latent factors Y is:

$$p(Y^T | \theta_{-Y^T}) \propto p(U | A_u, Y, \Psi_u) p(Y) \\ = \exp \left\{ -\frac{1}{2}(\text{vec}(U) - \text{vec}(A_u Y^T))^T(\mathbf{I}_r \otimes \Psi_u^{-1})(\text{vec}(U) - \text{vec}(A_u Y^T)) \right\} \\ \times \exp \left\{ -\frac{1}{2}\text{vec}(Y^T)^T \text{vec}(Y^T) \right\}$$

which can be simplified to:

$$p(Y^T | \theta_{-Y^T}) \propto \frac{\exp \left\{ -\frac{1}{2}(\text{vec}(Y^T)^T(\mathbf{I}_r \otimes (A_u^T \Psi_u^{-1} A_u + \mathbf{I}_k))\text{vec}(Y^T) \right\}}{\exp \left\{ \text{vec}(U)^T(\mathbf{I}_r \otimes \Psi_u^{-1} A_u)\text{vec}(Y^T) \right\}}$$

Therefore, the conditional posterior of $\text{vec}(Y^T)$ is the multivariate normal distribution $\text{vec}(Y^T) | \theta_{-Y^T} \sim N_{k_u r}(\mu, \Sigma)$ with parameters:

$$\Sigma = (\mathbf{I}_r \otimes (A_u^T \Psi_u^{-1} A_u + \mathbf{I}_k))^{-1} \\ \mu = \Sigma \text{vec}(U)^T(\mathbf{I}_r \otimes \Psi_u^{-1} A_u)$$

Here, since Σ is formed from a Kronecker product, the distribution of Y^T can be written as a Matrix Normal distribution using the inverse vectorization operator

$\text{vec}^{-1}(\cdot)$:

$$p(Y^T \mid \theta_{-Y^T}) \sim \text{N}(\text{vec}^{-1}(\mu); (A_u^T \Psi_u^{-1} A_u + \mathbf{I}_k)^{-1}, \mathbf{I}_r)$$

showing that the columns of Y^T are independent and can be sampled independently.

Next, the conditional posterior of the factor loadings A_u is:

$$\begin{aligned} p(A_u \mid \theta_{-A_u}) &\propto p(U \mid A_u, Y, \Psi_u) \prod_{i,j} p(A_{i,j} \mid \sigma_u^2) \\ &= \exp \left\{ -\frac{1}{2} (\text{vec}(U) - \text{vec}(A_u Y^T))^T (\mathbf{I}_r \otimes \Psi_u^{-1}) (\text{vec}(U) - \text{vec}(A_u Y^T)) \right\} \\ &\quad \times \prod_{i,j} p(A_{i,j} \mid \sigma_u^2) \end{aligned}$$

which can be simplified to:

$$p(A_u \mid \theta_{-A_u}) \propto \frac{\exp \left\{ -\frac{1}{2} (\text{vec}(A_u)^T (Y Y^T \otimes \Psi_u^{-1}) \text{vec}(A_u)) \right\}}{\exp \left\{ \text{vec}(U)^T (Y^T \otimes \Psi_u^{-1}) \text{vec}(A_u) \right\}} \times \prod_{i,j} p(A_{i,j} \mid \sigma_u^2)$$

Since the $p(A_{i,j} \mid \sigma_u^2)$ are independent and Ψ_u is diagonal, this factors into independent distributions for the rows of A_u . Let $\mathbf{a}_{i\cdot}$ denote row i of A_u , up to and including the diagonal if $i \leq k_u$, $\mathbf{u}_{i\cdot}$ the corresponding row of U , and Y_i the first i rows of Y if $i \leq k_u$, or the full matrix Y otherwise. Then, the conditional posterior of $\mathbf{a}_{i\cdot}$ is the multivariate normal distribution $\mathbf{a}_{i\cdot} \mid \cdot \propto \text{N}(\mu, \Sigma)$, conditional on $\mathbf{a}_{ii} > 0$ if $i \leq k$, with parameters:

$$\Sigma = (1/\sigma_{a_{i\cdot}}^2 + \lambda_{ui} Y_i Y_i^T)^{-1}$$

$$\mu = \lambda_{ui} \Sigma Y_i \mathbf{u}_{i\cdot}$$

The conditional posterior of the inverse of the variance parameter σ_u^2 is a Gamma distribution:

$$\sigma_U^2 \mid \theta_{-\sigma_i^2} \sim \text{Ga} \left(\frac{a_u + pk - k(k-1)/2}{2}, \frac{b_0 + \sum_{i=1}^p \sum_{j=1}^i A_{ij}^2}{2} \right) \quad (3.17)$$

The conditional posterior of the inverse of the diagonal matrix Ψ_u factors into independent Gamma distributions for each element λ_{ui} on the diagonal. Let $E_u = U - A_u Y^T$, and let \mathbf{e}_i be the i th column of E_u . Then:

$$\lambda_{ui} \mid \theta_{-\lambda_{ui}} \sim \text{Ga} \left(\frac{g_u + p}{2}, \frac{h_u + \sum_{i=1}^p \mathbf{e}_i^T \mathbf{e}_i}{2} \right) \quad (3.18)$$

3.4 Choice of k

A persistent challenge in factor analysis is choosing the dimension of the latent factors. Ideally, these parameters can be chosen directly based on Bayesian inference by specifying a prior on k , or through out-of-sample prediction or cross validation. However, more commonly, several values of k are selected, and factors are dropped post-analysis based on their magnitude, or the sensitivity of the results to the choice of k is assessed. Here, two such parameters must be chosen, k_u and k_e , representing the dimensions of the culture and residual factors, respectively. The maximum plausible value for either parameter is nine because that is the number of free coordinates in the p -dimensional symmetric tangent space. If $k_e = 9$, \mathbf{e}_{ij} can be fit exactly by $A_e \mathbf{f}_{ij}$ and so ϵ_{ij} will be estimated to be zero (see equation (3.4)). In my analysis, I first used the principle components analysis of Chapter 2 to guide the choice of k . Five principle components accounted for $> 95\%$ of the total symmetric tangent space variation. I thus fit model (3.7) with all combinations of $k_u = \{3, 4, 5, 6, 7\}$ and $k_e = \{0, 3, 5, 7\}$, and compared the estimates of $[Y]$ and U . Setting $k_e = 0$ is equivalent to fitting a model for X without any residual covariance. I chose $k_u = k$ if estimates of $[Y]$ and U fit with $k_u = k$ or $k_u = k + 1$ or were very similar. I chose k_e similarly.

3.5 MCMC inference

I implemented the Gibbs sampler 3.3 in MATLAB R2010b and ran MCMC chains of the sampler for a total of 4×10^5 iterations. I discarded the first 2×10^5 iterations as a burn-in period, and then collected 400 posterior samples of each parameter with a thinning rate of 500. Based on trial and error, this burn-in length and thinning rate were sufficient to achieve convergence for most MCMC chains (see section 3.5.2).

3.5.1 Posterior summarization

The factor matrix, Y is the focus of model (3.7). I calculated a posterior mean equivalence class $[\bar{Y}]$ based on the posterior samples $Y^{(1)} \dots Y^{(N)}$ by finding an icon of this class, \bar{Y} that minimized the modified partial Procrustes distances among all posterior samples:

$$\bar{Y} = \arg \min_H \sum_{i=1}^N |Y^{(i)} H_i - \bar{Y}| \quad \text{st } H_i H_i^T = \mathbf{I}_{k_u} \quad (3.19)$$

I solved (3.19) using the GPA algorithm (2.3), except no mean vector was calculated.

To calculate a posterior mean of A_u , I used the rotation matrices $H_1 \dots H_N$ calculated above (3.19) to define:

$$\bar{A}_u = \sum_{i=1}^N A_u^{(i)} H_i \quad (3.20)$$

To calculate the posterior mean of U , I simply took element-wise averages of this parameter from posterior samples because U is identifiable in the likelihood (3.9).

3.5.2 Convergence diagnostics

I assessed convergence of the MCMC chains by comparing posterior estimates of $[\bar{Y}]$ among independent chains, and by estimating the mixing properties of each chain.

To measure the similarity of two posterior estimates $[\bar{Y}]^{(1)}$ and $[\bar{Y}]^{(2)}$ from independent chains, I calculated the correlation between corresponding columns of the

icons $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}H$. I first used Ordinary Procrustes analysis (Dryden and Mardia, 1998) to find the rotation / reflection matrix $H \in O(k_u)$ that minimizes the partial Procrustes distances between $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}$:

$$H = \arg \min |\bar{Y}^{(1)} - \bar{Y}^{(2)}H| \quad \text{st } HH^T = \mathbf{I}_{k_u}$$

I then calculated the average column correlations as:

$$\frac{1}{k_u} \sum_{i=1}^{k_u} \sigma(\mathbf{y}_i, \mathbf{y}_i^*)$$

where \mathbf{y}_i^* is the i th column of $\bar{Y}^{(2)}H$.

To assess the mixing of each chain, I calculated the modified partial Procrustes distance between each posterior sample $Y^{(i)}$ and \bar{Y} , and used the MATLAB function *autocorr* to measure their autocorrelation through the chain.

3.6 Results

I used two methods to choose k_u and k_e . First, I tested if increasing number of factors accounted for more variation in the fitted larval shapes. I calculated the average magnitude of a culture mean tangent-space shape vector:

$$\frac{1}{r} \sum_{i=1}^r |\mathbf{u}_i|$$

where $|\mathbf{u}_i| = \sqrt{\mathbf{u}_i^T \mathbf{u}_i}$. As shown in Fig. 3.1, increasing the number of culture factors generally increased the magnitude of \mathbf{u}_i , and increasing the number of residual shape factors decreased the magnitude of \mathbf{u}_i . However, for $k_u \geq 5$ and $k_e \geq 5$, this measure did not change considerably, suggesting that five factors of each type were needed to account for the majority of variation in shape.

Second, I tested if the estimated factors were robust to model parameterization. I compared posterior estimates of $[\bar{Y}]$ from models fit with differing numbers of k_u

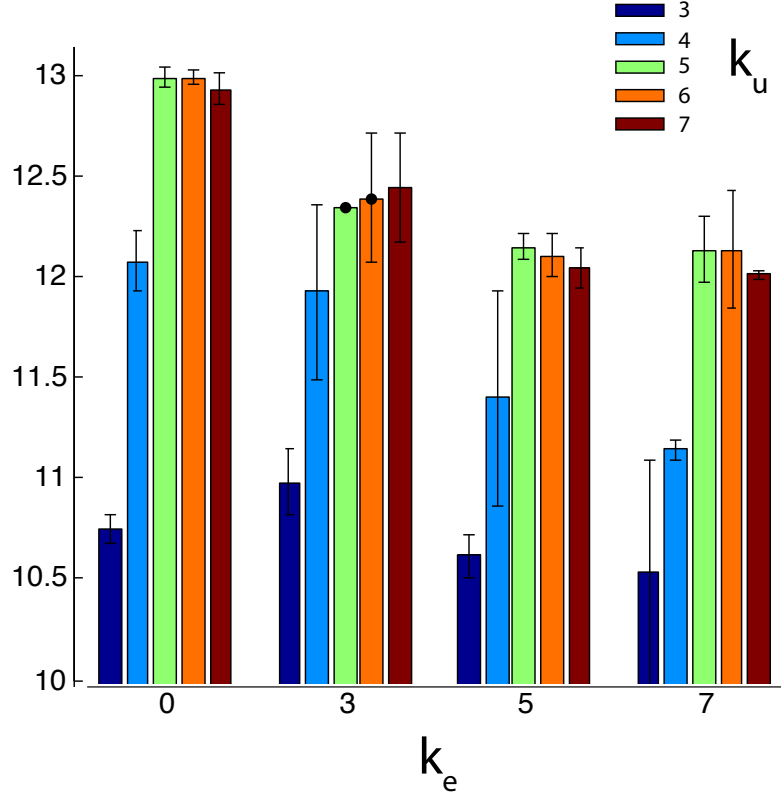


FIGURE 3.1: Fitted magnitudes of culture mean shape vectors. Each bar represents the average magnitude of the vectors representing the mean shape of each of the 72 cultures estimated with a particular number of factors of culture shape (k_u) or residual shape (k_e). Bars are the mean estimates over two independent runs of the Gibbs sampler. Error bars show 2SD.

and k_e . Analogously to the test for convergence 3.5.2, I compared two estimates $[\bar{Y}]^{(1)}$ and $[\bar{Y}]^{(2)}$ by calculating the minimum correlation between the columns of two maximally aligned icons $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}$:

$$\min_{i=1\dots k_u} \sigma(\mathbf{y}_i, \mathbf{y}_i^*) \quad (3.21)$$

where \mathbf{y}_i^* is the i th column of $\bar{Y}^{(2)}H$, and H is the matrix that minimizes:

$$H = \arg \min |\bar{Y}^{(1)} - \bar{Y}^{(2)}H| \quad \text{st } HH^T = \mathbf{I}_{k_u}$$

H is calculated as $H = VU^T$ with $USV = \bar{Y}^{(1)T}\bar{Y}^{(2)}$ the singular value decomposition of the matrix product on the RHS.

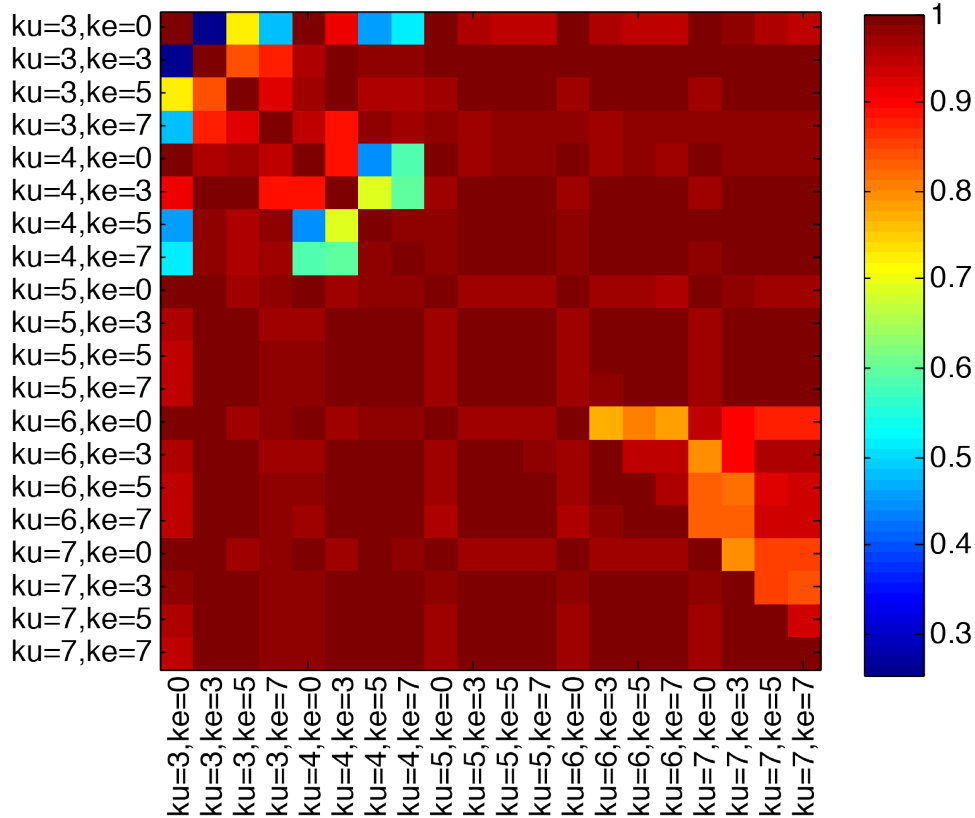


FIGURE 3.2: Robustness of factor model estimates. Heatmap showing minimum correlation between factors in optimally aligned icons of posterior means $[\bar{Y}]$. The diagonal shows the minimum correlation between columns of icons of posterior means calculated from two independent MCMC chains of the same model. Elements above the diagonal show the minimum correlation between columns of two icons $[\bar{Y}]^{(1)}$ and $[\bar{Y}]^{(2)}$. If $[\bar{Y}]^{(2)}$ has more columns than $[\bar{Y}]^{(1)}$, the minimum is taken over the lower number of columns. Elements below the diagonal are identical for $k_u \leq 5$. For greater values of k_u , the minimum only considers the five most similar columns of the two icons.

If more factors are fit than are needed to adequately explain the variation in shape, then some of the factor estimates will be highly variable. Fig. 3.2 shows that for $k_u = 5$, estimates of all factors are relatively robust to changes in k_e , but for lower or higher k_u , at least one of the estimated factors was very sensitive to model specification. However, in models with $k_u = 6$ or 7 , the first five factors were still highly robust 3.2

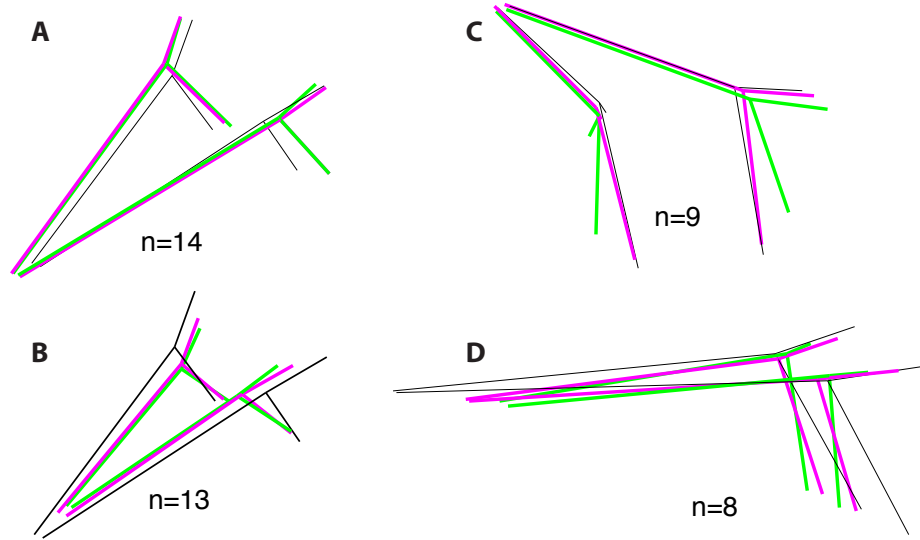


FIGURE 3.3: Fitted differences in culture mean shape. For some cultures, the factor model estimates of culture mean shape (magenta) were very similar to the OLS estimates (green). A-B) Factor model and OLS estimates of two of the most-divergent culture mean shapes were nearly identical. C-D) The two cultures where the two estimates differed most. In C), the OLS estimate is very different than the mean, while the factor model estimate nearly has the mean shape. In D), the difference between the OLS estimate and the factor model estimate is mostly concentrated in the positioning of the anterolateral rods and the length of the body rods. In all panels, shape differences are multiplied by 2 to highlight shape variation. Numbers below each figure give the number of larvae measured from each culture.

Based on these two methods, I chose $k_u = 5$ and $k_e = 5$. Figure 3.3 shows estimates of the mean shape of several cultures based on this model, and compares these estimates to ordinary least squares (OLS) estimates $\hat{U} = XZ^T(ZZ^T)^{-1}$. Note that the model (green) is always closer to the population mean \bar{X} (black) than the OLS estimate is (magenta), and that estimates differ more if fewer individuals from a particular culture were measured. This is largely because the prior on Y shrinkages culture means towards zero. Here, I have magnified the differences in shape by a factor of 2 to show the variation more clearly.

Gene expression correlates of shape variation

A central goal in biology today is explaining how genetic variation, and variation in molecular processes, leads to variation in phenotypes. Measuring variation in gene expression is a powerful technique for investigating these relationships. Much of the developmental control of higher-order phenotypes involves regulating gene expression. Tracking large numbers of genes simultaneously can thus provide insight into the functioning of molecular pathways and networks. Also, the genetic control of gene expression, while still complex, is considerably simpler than for organismal-level traits such as morphology.

In this chapter, I explore the link between gene expression and skeletal shape in sea urchin larvae. In Chapters 2 and 3, I formulated a low-dimensional representation of larval morphological variation. Here, I use regularized regression techniques to identify genes whose expression correlates with variation in larval shape.

Many genes that regulate early morphogenesis in sea urchin embryos and larvae are known. Several relatively complete transcriptional regulatory pathways have been worked out that describe how the activation or repression of a small number of transcription factors and signaling molecules is responsible for specifying cell-type

identify and morphogenesis in the embryos (Oliveri et al., 2008; Peter and Davidson, 2010; Su et al., 2009; Peter and Davidson, 2011; Sharma and Ettensohn, 2011). These regulatory genes act dynamically through early development, from soon after fertilization through the formation of the larval body plan. In addition, members of a suit of structural genes involved in the physical construction of the larval skeleton are known. In this study, we measured the expression of 74 genes at seven time points in 71 cultures. We measured the morphology of voucher larvae from the same cultures (as described in Chapters 2 and 3), and thus could assess the correlation between the expression of genes with plausible roles in embryonic morphogenesis and skeletal development, and skeletal shape.

4.1 Data

Gene expression measurements were performed on 74 genes in pooled embryos from 71 cultures at seven stages of development. From each pool of embryos, RNA was extracted, converted to cDNA and quantified using the Illumina DASL platform, a small-scale microarray suitable for the simultaneous analysis of large numbers of samples. Each gene was represented by 3-6 probes on the array. Residual probe intensities, after regressing out background probe values, were log₂-transformed and normalized to the mean expression of 4 probes targeting the control gene, *RBM8A*. The probes targeting the same gene were then tested for consistency by comparing the within-gene correlations across all samples to the among-gene correlations. Probes that were no more similar to other probes targeting the same gene than expected were removed. Finally, the log₂ intensities of all remaining probes targeting each gene were averaged.

During the sample collection and DASL processing, several samples were flagged as outliers and removed. Most of these samples were from the first two time points, and were identified because fertilization rates were low in the cultures, leaving large

numbers of un-fertilized eggs mixed with the developing embryos. Therefore, the $l \times gt$ (l cultures, g genes, and t time points) matrix, X of gene expression measures had several holes (Fig. 4.1). Since among-time point correlations of gene expression values tended to be low, rather than attempt to impute these missing values I chose in the regression of shape on gene expression to drop cultures with missing gene expression data. So as not to completely throw away much of the available data, I ran two regression analyses. The first included 68 cultures but gene expression only from time points 3-7. The second included only the 44 cultures with complete (7 time point) gene expression time series (Fig. 4.1).

As a dependent variable, I used the equivalence class of factor matrices $[Y]$ of latent culture mean shape factors described in Chapter 3. As an estimate of $[Y]$, I used the posterior mean $[\bar{Y}]$ of a model fit with $k_u = 5$ culture shape factors. To fit the regression, I selected an arbitrary icon of $[\bar{Y}]$, \bar{Y} , and then optimized over the set of rotation/reflection matrices $H \in O_{k_u}$ that together with \bar{Y} define $[\bar{Y}]$.

4.2 Modeling strategy

To identify gene expression correlates of shape, I used the following multivariate regression model:

$$YH = XB + \epsilon \tag{4.1}$$

for Y , the $(l \times k_u)$ matrix of shape factors, and B the $(gt \times k_u)$ matrix of regression coefficients and $H \in O_{k_u}$. In this model, Y had either 44 or 68 rows (samples), depending on the version of the analysis. In each sample, we measured the expression of 74 genes in each of 7 or 5 time points, respectively. Therefore, there were many more predictor variables than dependent variables, and regularization of the estimate of B was required.

A number of commonly used shrinkage estimators of B are available including

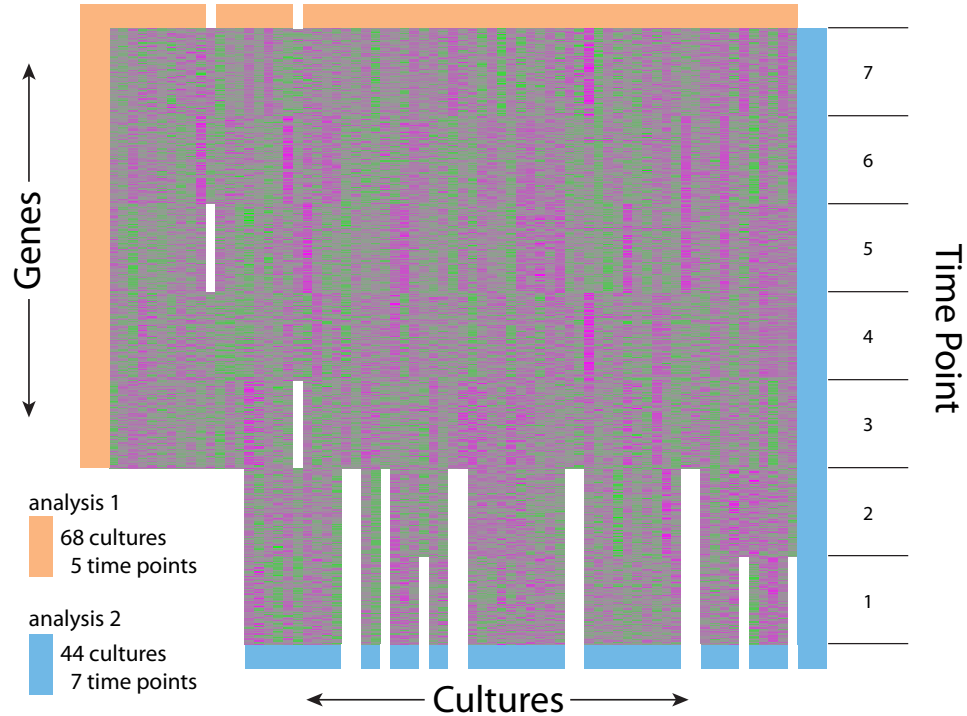


FIGURE 4.1: Representation of the gene expression data, and sample selection. The heatmap shows normalized gene expression values for each of the 74 genes at each of 7 time points from soon after fertilization until the time of morphological assessment. Colors towards purple signify higher than average expression, while colors towards green signify lower expression. Each gene is normalized to have zero mean and variance one. White holes signify missing data. Data is only missing by culture \times time point. From these gene expression data, I selected two subsets without missing data for analysis. These subsets are boxed in orange (1) and blue (2).

the Ridge and Lasso estimators, or Bayesian alternatives such as the Bayesian horseshoe and spike-slab priors (ex. Armagan et al., 2011). Since each gene enters into the regression 5 or 7 times from the different time points, and variable selection is desirable for experimental validation, I used the group Lasso algorithm, grouping the set of regression coefficients across the time points for each of the 74 genes. The group Lasso is computational efficient and leads to “variable-selection” in that many regression coefficients can be shrunk identically to zero (Yuan and Lin, 2006). While the standard Lasso imposes an L_1 penalty on all regression coefficients, the group

lasso imposes an L_1 penalty at the level of groups of coefficients, and an L_2 penalty on the coefficients of each group, effectively shrinking whole groups of coefficients to zero. The multivariate group Lasso minimizes:

$$\arg \min_{B, H \in O_{k_u}} \frac{1}{2} \|YH - XB\|_2^2 + \lambda \sum_j^{k_u} \sum_g^G \sqrt{p_g} \|B_{j,g}\|_2 \quad (4.2)$$

where $B_{j,g}$ is the $g \times t \times k_u$ vector of regression coefficients for the g genes at t times points for the $j = (1, \dots, k_u)$ factors, p_g is the length of group g and λ is a tuning parameter that must be selected. As a note, fixing H LHS of (4.1) would not affect inference of B under ordinary least squared, or with the Ridge penalty, but generally does result in a different solution for B with the group lasso penalty because the shrinkage is along defined axes in predictor space.

To fit (4.2), I used an iterative strategy. I first fixed H at an initial value, $H^{(0)}$ and then estimated $B^{(1)} | H^{(0)}$ by (4.2). Given $B = \hat{B}^{(1)}$, H only enters the first term of (4.2). Thus, finding $\hat{H}^{(1)} | \hat{B}^{(1)}$ is equivalent to the Ordinary Procrustes Problem (Chapter 2), and can be solved as in (2.4). I iterated between these two steps until $\{\hat{H}, \hat{B}\}$ converged. Depending on λ , this tended to take 20-200 iterations.

4.3 Selecting the tuning parameter λ

To select the model parameter λ , I used leave-one-out cross-validation. For each potential value of λ , I re-fit model (4.2) n times (for $n = 68$ or 44), each time leaving out one of the rows of Y and X to use as validation data. I calculated the mean square predictive error (MSPE) as:

$$MSPE(\lambda) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i \hat{H}(\lambda)_{-i} - \mathbf{x}_i \hat{B}(\lambda)_{-i}\|^2 \quad (4.3)$$

where $\|\cdot\|^2$ is the squared Frobenius norm, \mathbf{y}_i is the i th row of the rotated factors, \mathbf{x}_i is the i th row of X , and $\hat{B}(\lambda)_{-i}$ and $\hat{H}(\lambda)_{-i}$ are the fitted values of B and H when (4.2) is fitted with given λ but without the data from row i .

I chose the value of λ minimized (4.3), and then refit (4.2) using all the data to produce final estimates of B and H .

4.4 Results

I fit model (4.2) with $k_u = 5$ and $n = 68$ or 44 , for models 1 and 2, respectively. For model 1, $\lambda \in (2, 7)$ gave similar MSPE values, with the maximum predictive accuracy $\max_{\lambda} \left(1 - \frac{MSPE_{\lambda}}{MSPE_0}\right) = 0.24$ at $\hat{\lambda} = 4.14$. For model 2, the optimal $\hat{\lambda} = 3.6$ with predictive accuracy $\max_{\lambda} \left(1 - \frac{MSPE_{\lambda}}{MSPE_0}\right) = 0.25$. The similarity in predictive accuracies of the two models is surprising given the much smaller sample size of model 2. This suggests that the expression of several early-expressed genes during development are useful for predicting later larval phenotypes.

Using the results of model 2, 28/74 genes were selected by the rotated group Lasso algorithm as having non-zero relationships with one (or more) factors. Five genes were selected for two factors. To test the reliability of this variable selection, I bootstrapped the group Lasso regression (given λ and \hat{H}) by selecting 70% = 30 samples (from the original 44) with replacement, refitting (4.2) with $H = \hat{H}$. I then counted the percentage of bootstrapped runs in which each gene was given a non-zero weight in the regression. 13 genes were included by the group Lasso for one or more factors in greater than 50% of the bootstrap replicates. (Table 4.1).

To inspect which genes were correlated with what types of shape change, I first characterized the magnitude of each rotated factor as γ_i such that:

$$\sum_{i=1}^{k_u} \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|} \gamma_i \frac{\mathbf{y}_i^T}{\|\mathbf{y}_i\|} = A_u \hat{H} \hat{H}^T Y^T \quad (4.4)$$

were \mathbf{a}_i and \mathbf{y}_i are the i th columns of $A_u \hat{H}$ and $Y \hat{H}$, respectively (Table 4.1). I then plotted the shape change along each factor (Fig. 4.2). Some factors (ex. factor 5) represent mostly shape change, with little change in the lengths of the skeletal rods,

others (ex. factor 1) represent mostly uniform growth, and the remainder included various combinations of growth and shape differences.

Table 4.1: Bootstrap scores for gene expression factor regression. For each rotated factor (columns of $Y\hat{H}$, factor, the genes that were assigned a non-zero regression coefficient by the group Lasso algorithm in 50% of bootstrap replicates are listed, along with the mean regression coefficient at each of the 5 time points over the bootstrap replicates. The magnitude of variance explained by each factor γ_i is also listed. The absolute magnitude of the regression coefficients is arbitrary, but their relative magnitude (among genes and factors) is consistent.

Factor	Factor Weight	Gene	Inclusion percentage	Time point 3	Time point 4	Time point 5	Time point 6	Time point 7
1	54.4	Brn1/2/4	61.9	0.0324	0.0276	0.018	-0.008	-0.0259
		Dkk	60.0	0.0337	0.012	0.0192	-0.0101	0.0349
		Nkx2.2	75.6	0.0118	-0.0107	0.0608	0.0594	0.0071
		SM30-E	75.8	-0.0192	-0.0349	-0.0465	-0.0466	-0.018
		Tbr	96.9	-0.0259	-0.0171	-0.013	-0.0889	-0.1311
		Wnt8	52.0	0.0171	0.025	0.0214	-0.0059	0.0181
2	22.0	Fmo2	69.9	-0.0694	0.0398	0.0192	0.0157	0.037
		SM32	59.1	-0.0132	0.0619	0.0236	0.0083	0.0233
		Su(H)	86.7	-0.0657	-0.0415	-0.032	-0.0322	-0.0161
3	45.6	CyclinT	92.2	0.0018	-0.118	-0.0479	-0.0359	0.0291
		Eve	68.8	0.0259	-0.0775	-0.0062	3.05E-04	-0.0146
		RhoA	77.2	0.0583	-0.0626	-0.0301	-0.0275	-0.0088
4	21.2	Ficolin	69.9	0.003	0.0061	-0.041	0.03	0.0305
		Pax2/5/6	60.1	0.0091	-0.0078	-0.0344	-0.0407	0.0054
5	20.8	Deadringer	91.3	0.0474	0.0066	0.0687	0.0711	0.049
		Delta	57.7	0.0066	0.0309	0.0133	0.0251	0.0216
		Nodal	51.3	6.42E-06	0.0309	-0.0113	-0.0313	-0.0038
		Tel	50.9	-0.0119	0.0129	0.0071	0.0194	-0.0195

4.5 Discussion

In this chapter, I have constructed a regression model on gene expression values through later sea urchin embryonic development that can predict nearly 25% of estimated among-culture variation in larval shape, 5 days after fertilization. The results presented here demonstrate that the expression of genes in the sea urchin endomesoderm and ectoderm gene regulatory networks is informative for both the length of the calcium carbonate spicules in the larvae, and their three-dimensional conformation. The particular genes that the model choice algorithms selected are interesting given their known roles during embryogenesis. For example, two of the six genes reliably selected for Factor 1 (*Tbr* and *SM30-E*) are expressed exclusively in the cells that construct the larval skeleton, and the shape change described by this factor appears related to the overall size of the skeleton. Also, two of the three genes reliably selected for Factor 3 (*CyclinT* and *RhoA*) are ubiquitously expressed. This factor describes shape variation not directly related to skeletal growth, but to the overall conformation of the larva. Based on my observation of embryonic and larval development, dramatic shape variation of this type may relate to the health or stress level of the embryo. Finally, the genes *Delta*, *Notch* and *Su(H)* (Factors 2 and 5) are important signaling molecules that help define territory boundaries in the developing embryos, and *Nkx2.2*, *Deadringer* and *Px2/5/6* (Factors 1, 4 and 5) are all important transcription factors expressed in the ectoderm, which is known to provide signals that direct the spatial location of the skeletogenic cells.

This regression model does have limitations in terms of relating regression coefficients to sea urchin biology. I chose to use the group Lasso regularization penalty on the regression coefficients because variable (gene) selection in the regression is useful for downstream confirmatory experiments, and based on an expectation that individual genes may contribute to expression over long periods of development. If the

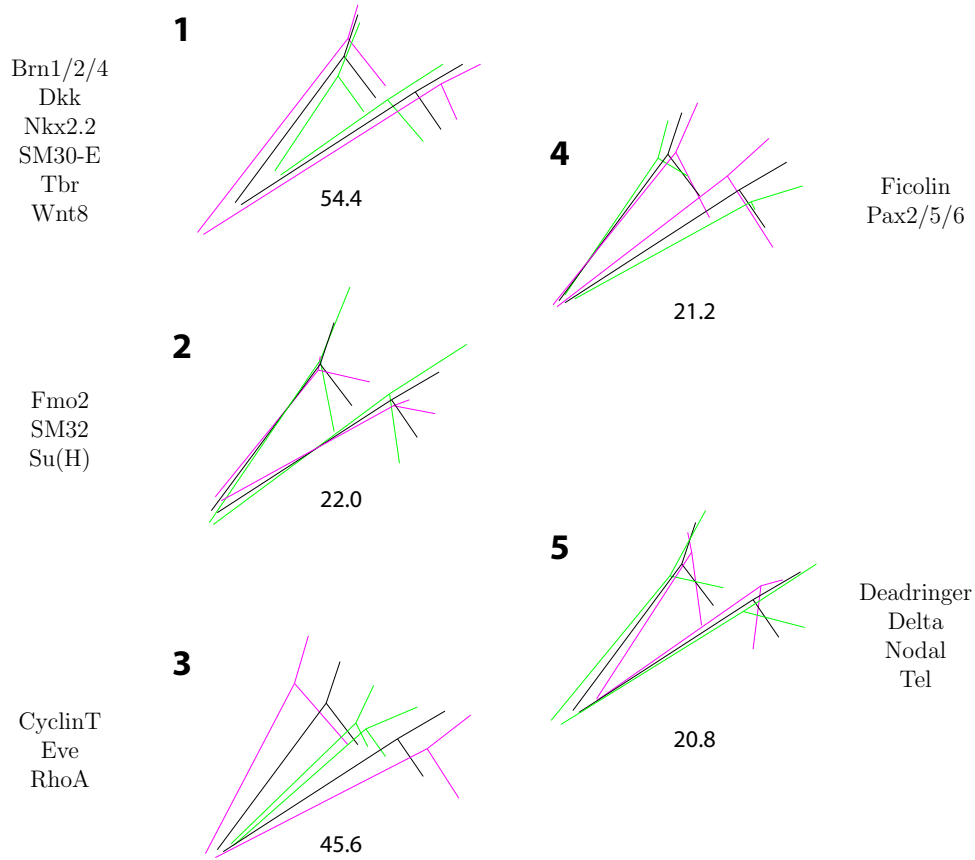


FIGURE 4.2: Shape variation represented by the five factors regressed on gene expression phenotypes. These factors were rotated so as to be maximally correlated with gene expression under the group Lasso regularization penalty. In each figure, the magenta and green figures show 10 population standard deviations of variation along each factor for clarity. The variances associated with each factor are listed below the figure. Gene with expression patterns correlated with each factor are listed (Table 4.1).

latter is true, the regression may have more power if whole expression time-courses are treated as together. In fact, the maximum cross-validation accuracy of the group Lasso regression model above was considerably higher than a similar regression with the simple Lasso penalty. However, since the group Lasso algorithm tends to deal with sets of correlated predictor variables (time-courses for each gene) by selecting only one gene from each set, the fact that a gene is not selected in my regression

model does not necessarily mean that it is not well correlated with shape variation. It could just be that other genes are correlated with both the target gene and shape. In a gene network, if gene A regulates shape, and gene B regulates gene A, gene B might individually correlate well with shape variation, but its correlation with shape would necessarily be lower than that of gene A, and thus would tend not to be selected. This is the intended behavior of the lasso algorithm, because in this situation, gene B will likely not contribute more to prediction of shape, given additional knowledge of gene A, but limits the types of downstream discovery possible from a such a regression analysis. Other techniques such as factor regression (and the related 2-Block Partial least Squares algorithm (Rohlf and Corti, 2000)) are better suited to such investigations, but I have not explored whether their predictive accuracy would suffer relative to the group Lasso.

An issue with the model selection algorithms presented here is the the cross-validation statistic that I used can be highly sensitive to un-modeled covariance among samples. In this experiment, the 72 cultures were not independent - each culture shared a male parent and a female parent with other cultures in the experiment. Thus, from the perspective of other completely un-related cultures, this cross-validation statistic is likely anti-conservative. This relatedness among the cultures in the experiment is based on a breeding design from quantitative genetics, and is designed to estimate patterns of genetic variation and covariation in a population. With large sample sizes (in particular large numbers of independent genetic backgrounds), such an experiment can well characterize the patterns of variation in a population. In this experiment, much of the observed variation was likely caused by genetics, but since only 12 genetic backgrounds were sampled, it is not realistic to build a predictive model for gene expression - shape relationships in un-sampled genetic backgrounds. Therefore, the predictive accuracy reported above only relates to this set of genetic backgrounds (and the particular culturing conditions that were

used).

Joint model of shape and gene expression

In Chapters 2-4, I proposed a 3-step model for linking a multidimensional high-level phenotype (larval shape), to a multidimensional low-level trait (gene expression), when these two sets of traits are not measured on the same individuals. The three steps can be summarized as:

1. Estimate a mean shape over all larvae and use this population mean to calculate residual deviations for each larva (Chapter 2). Simultaneously, decompose this shape variation into symmetric and asymmetric components.
2. Group larvae by culture (representing common genetic background and shared environmental conditions), and estimate the average larval shape (from step 1) in each culture with a reduced dimension variance component model. Extract a lower-dimensional representation of mean shape for each culture.
3. Use regularized multivariate regression to identify gene expression traits (measured at the level of culture) correlated with variation in culture mean shape (from step 2).

In this chapter, I outline a unified Bayesian model for joint analyses of these three components simultaneously, and to which the three steps can be considered as an approximation. I first formulate this model explicitly and then discuss how inference on this joint model may differ from inference on the 3-step model. To the extent that it is possible, I use the same notation for each sub-model as used in Chapters 2-4.

5.1 Joint model for variation in gene expression and morphological shape

The joint model is a hierarchical combination of steps 1-3 with shared parameters.

The first level of the joint model relates each larva's configuration to an icon of the population's mean shape, \bar{X} , and partitions deviations from this mean in terms of symmetric and asymmetric shape residuals. For larva j in culture i , let X_{ij} be a $k \times m$ configuration matrix representing the larva's shape. For clarity, let X_{ij} be ordered so that the larva's left-side coordinates are in rows $(1, \dots, \frac{k}{2})$, and the corresponding right-side coordinates are in rows $(\frac{k}{2} + 1, \dots, k)$. Without loss of generality, we assume that \bar{X} is centered and oriented so that its plane of symmetry is the $y - z$ plane. The symmetric partition of X_{ij} with respect to \bar{X} is:

$$X_{ij}\Gamma_{ij} = \bar{X} + X_{ij}^S + X_{ij}^A + \mathbf{1}_k\gamma_{ij}^T \quad (5.1)$$

$\Gamma_{ij} \in SO_m$ is an arbitrary rotation matrix, $\gamma_{ij} \in \mathbb{R}^k$ is an arbitrary translation vector, and X_{ij}^S and X_{ij}^A residual symmetric and asymmetric shape deviations, respectively. Since X_{ij}^S is a symmetric icon, ordered as above with the same plane of symmetry as \bar{X} , coordinates in rows $(\frac{k}{2} + 1, \dots, k)$ are mirror images of coordinates in rows

$(1, \dots, \frac{k}{2})$ and related by the reflexion matrix $H = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Thus, X_{ij}^S can

thus be represented as:

$$X_{ij}^S = \begin{bmatrix} R_{ij}^S \\ R_{ij}^S H \end{bmatrix} \quad (5.2)$$

for R_{ij}^S a $\frac{k}{2} \times m$ matrix. The asymmetric residual matrix also has redundant parameters and can be represented in reduced form as:

$$X_{ij}^A = \begin{bmatrix} R_{ij}^A \\ -R_{ij}^A H \end{bmatrix} \quad (5.3)$$

$$(5.4)$$

for R_{ij}^A also a $\frac{m}{2} \times k$ matrix.

Equations (5.2) and (5.3) show that the residual terms X_{ij}^S and X_{ij}^R are highly constrained, but R_{ij}^S and R_{ij}^A are not. I assume that all the asymmetric variation is noise and independent of culture effects on shape and gene expression, and thus focus only on modeling X_{ij}^S . These symmetric residuals may vary systematically by genetic background and environment. However, modeling R_{ij}^S instead of X_{ij}^S is easier because it is full rank.

The second level of the joint model decomposes R_{ij}^S into variation due to culture plus symmetric variation unique to the individual.

$$R_{ij}^S = U_i^S + E_{ij}^S \quad (5.5)$$

where U_i^S is the (symmetric) mean form of culture i and E_{ij}^S is the (symmetric) residual for larva j in culture i .

As in Chapter 3, (5.5) is treated as a random effects model. Therefore, $\text{vec}(U_i^S)$ and $\text{vec}(E_{ij}^S)$ are random vectors. To capture their lower-dimensional structure, these are modeled with independent factor models:

$$\text{vec}(U_i^S) = A_u \mathbf{y}_i + \boldsymbol{\epsilon}_i \quad (5.6)$$

$$\text{vec}(E_{ij}^S) = A_e \mathbf{f}_{ij} + \boldsymbol{\epsilon}_{ij}$$

where \mathbf{y}_i and \mathbf{f}_{ij} are lower dimensional ($k_u \times 1$ and $k_e \times 1$ with $k_u < k$ and $k_e < k$) vectors of latent traits underlying the culture and individual residual shape variation. Linking (5.1) and (5.5) in this way provide an advantage over the 3-step model by incorporating the registration of larval forms directly into the model for shape variation and thus making use of information about morphological variation among cultures, rather than assuming each larval form is independent.

Finally, in the third level, I relate each of the k_u latent traits in \mathbf{y}_i to the gene expression measures on each culture \mathbf{x}_i^E through a linear regression model. Here, \mathbf{x}_i^E is a vector of gene expression measures for each of the 72 genes at the 7 time points. As in Chapter 4, the model for \mathbf{y}_i is:

$$\mathbf{y}_i = \boldsymbol{\beta} \mathbf{x}_i^E + \boldsymbol{\xi}_i \quad (5.7)$$

where $g = 1 \dots G$ indexes groups of \mathbf{x}_i^E representing the same gene measured at different time points, $\boldsymbol{\beta}$ is a matrix with k_u rows and as many columns as \mathbf{x}_i^E . Unlike in Chapter 4, the regression model (5.7) is fit simultaneously with the shape factors (\mathbf{y}_i) themselves. This removes the necessity of re-rotating these shape factors *post-facto*.

Combining (5.1), (5.5), (5.6) and (5.7), the model for X_{ij} is:

$$\begin{aligned} X_{ij} \Gamma_{ij} &= \bar{X} + \mathbf{1}_k \gamma_{ij}^T + \begin{bmatrix} R_{ij}^S \\ R_{ij}^S H \end{bmatrix} + \begin{bmatrix} R_{ij}^A \\ -R_{ij}^A H \end{bmatrix} \\ R_{ij}^S &= \text{vec}^{-1} (A_u (\boldsymbol{\beta} \mathbf{x}_i^E + \boldsymbol{\xi}_i) + \boldsymbol{\epsilon}_i + A_e \mathbf{f}_{ij} + \boldsymbol{\epsilon}_{ij}) \end{aligned} \quad (5.8)$$

I place uniform priors on Γ_{ij} and γ_{ij} , and independent normal priors on each element of $\boldsymbol{\xi}_i$, $\boldsymbol{\epsilon}_i$, $\boldsymbol{\epsilon}_{ij}$, \mathbf{f}_{ij} and R_{ij}^A , with variances σ_x^2 , σ_u^2 , σ_e^2 , 1 and σ_a^2 , respectively. Priors on A_u and A_e are as described in (3.13), with hyper variance parameters $\sigma_{A_u}^2$ and $\sigma_{A_e}^2$. For $\boldsymbol{\beta}$, I place a Multi-Laplace prior over the set of coefficients for each gene in each rom of $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}_{j,g} | \lambda) \sim \lambda^{p_g/2} \exp(-\lambda \|\boldsymbol{\beta}_{j,g}\|)$$

where g indexes the columns of $\boldsymbol{\beta}$ representing the coefficients for one gene over the 7 time points, j indexes a row of $\boldsymbol{\beta}$, $p_g = 7$ is the length of $\boldsymbol{\beta}_{j,g}$ and λ is a hyperparameter that either can be modeled with a prior, or selected by cross-validation, as in Chapter 4. The maximum *a posteriori* estimate of $\boldsymbol{\beta}$ in log-space under this prior recovers the group-Lasso solution (Chapter 4) (Raman et al., 2009)

Treating $X_{ij}^A = \begin{bmatrix} R_{ij}^A \\ -R_{ij}^A H \end{bmatrix}$ as the residual, the parameters of model (5.8) are then:

$$\Theta = \{\boldsymbol{\beta}, \Gamma_{ij}, \gamma_{ij}, \boldsymbol{\xi}_i, \boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_{ij}, \mathbf{f}_{ij}, A_u, A_e, \sigma_x^2, \sigma_u^2, \sigma_e^2, \sigma_a^2, \sigma_{A_u}^2, \sigma_{A_e}^2\}$$

Since X_{ij}^A is not full rank, the likelihood of (5.8) is difficult to calculate. However, if we partition $X_{ij}\Gamma_{ij} = \begin{bmatrix} X_{ij}^1\Gamma_{ij} \\ X_{ij}^2\Gamma_{ij} \end{bmatrix}$ where X_{ij}^1 and X_{ij}^2 are the coordinates of the left and right side of larva ij , and partition $\bar{X} = \begin{bmatrix} \bar{X}^1 \\ \bar{X}^2 \end{bmatrix}$ equivalently, we can form transformed variables $X_{ij}^* = \begin{bmatrix} X_{ij}^1\Gamma_{ij} \\ X_{ij}^2\Gamma_{ij}H \end{bmatrix}$ and $\bar{X}^* = \begin{bmatrix} \bar{X}^1 \\ \bar{X}^2 H \end{bmatrix}$. Then, the transformed model is:

$$X_{ij}^* = \bar{X}^* + \begin{bmatrix} \mathbf{1}_{k/2}\gamma_{ij}^T \\ \mathbf{1}_{k/2}\gamma_{ij}^T H \end{bmatrix} + \begin{bmatrix} \mathbf{I}_{k/2} \\ \mathbf{I}_{k/2} \end{bmatrix} R_{ij}^S + \begin{bmatrix} R_{ij}^A \\ -R_{ij}^A \end{bmatrix}$$

which is equivalent to (5.8). Now, since multiplying the (rotated) right-side coordinates by H reflects them across the plane of symmetry, pre-multiplying both sides by:

$$Q = \left(\begin{bmatrix} \mathbf{I}_{k/2} & \mathbf{I}_{k/2} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{k/2} \\ \mathbf{I}_{k/2} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{I}_{k/2} & \mathbf{I}_{k/2} \end{bmatrix}$$

averages the left and right coordinates and gives:

$$QX_{ij}^* = Q\bar{X}^* + \frac{1}{2}(\mathbf{1}_{k/2}\gamma_{ij}^T + \mathbf{1}_{k/2}\gamma_{ij}^T H) + R_{ij}^S \quad (5.9)$$

since $Q \begin{bmatrix} R_{ij}^A \\ -R_{ij}^A \end{bmatrix} = 0$. Here, the asymmetric residuals R_{ij}^A are eliminated by Q and the transformation to X_{ij}^* after rotation.

By (5.8), the prior density of R_{ij}^S is:

$$\text{vec}(R_{ij}^S) \sim N_{\frac{k}{2}m} \left(A_u(\boldsymbol{\beta}\mathbf{x}_i^E + \boldsymbol{\xi}_i) + \boldsymbol{\epsilon}_i + A_e\mathbf{f}_{ij}, \sigma_e^2 \mathbf{I}_{\frac{k}{2}m} \right) \quad (5.10)$$

Combining (5.9) and (5.10), the likelihood of the data $(X_{1,1} \dots X_{r,n})$ is

$$\begin{aligned} L(X_{1,1} \dots X_{r,n} \mid \Theta) = \\ \prod_{i=1}^r \prod_{j=1}^n N_{\frac{k}{2}m} \left(\text{vec} \left(Q \begin{bmatrix} X_{ij}^1 \Gamma_{ij} \\ X_{ij}^2 \Gamma_{ij} H \end{bmatrix} \right) \mid \frac{1}{2} \text{vec} (\bar{X}^1 + \bar{X}^2 H + \mathbf{1}_{k/2} \gamma_{ij}^T + \mathbf{1}_{k/2} \gamma_{ij}^T H) \right. \\ \left. + A_u(\boldsymbol{\beta}\mathbf{x}_i^E + \boldsymbol{\xi}_i) + \boldsymbol{\epsilon}_i + A_e\mathbf{f}_{ij}, \sigma_e^2 \mathbf{I}_{\frac{k}{2}m} \right) \quad (5.11) \end{aligned}$$

Finally, the posterior $\Pi(\Theta \mid X_{1,1}^* \dots X_{r,n}^*)$ is:

$$\begin{aligned} \Pi(\Theta \mid X_{1,1} \dots X_{r,n}) = \\ L(X_{1,1} \dots X_{r,n} \mid \Theta) \pi(\bar{X}, \Gamma_{ij}, \gamma_{ij}) \pi(A_u, A_e) \pi(\boldsymbol{\xi}_i) \pi(\mathbf{e}_i) \pi(\mathbf{f}_i) \pi(\boldsymbol{\beta}) \quad (5.12) \end{aligned}$$

where the priors are defined as above.

5.2 Inference on joint model

The goal of the joint model is to identify genes with expression variation correlated with aspects of larval variation in morphological form. From this perspective, the most informative parameters of the joint model are: $\boldsymbol{\beta}$, the matrix of regression coefficients and A_u , the matrix that relates variation in underlying latent factors to symmetric *shape* space.

Simultaneous inference of all these parameters is possible from this joint model with a Gibbs sampler. To do so, the posterior (5.12) can be expanded by reintroducing R_{ij}^S as in (5.9) and (5.10) to the likelihood. The Gibbs updates of the factor

model (5.10) are as described in Chapter 3. Gibbs updates of β under the Bayesian group Lasso prior are described in (Raman et al., 2009). The conditional posterior for Γ_i has the form of a Matrix Bingham-von-Mises Fisher family distribution (Hoff, 2009). Finally, the conditional posterior for \bar{X} and γ_{ij} will each be multivariate normal.

Since the configuration matrices X_{ij} represent shapes, the variables Γ_{ij} and γ_{ij} are defined only up to a global average orientation and rotation of \bar{X} , which is a nuisance parameter that can be fixed arbitrarily in the analysis.

5.3 Comparison of 3-step model and joint model

While the motivations and specifications of the 3-step model and joint model are very similar, by linking the inference of all parameters, the joint model may provide improved estimation of the relationship between morphological and gene expression variation in the larvae. While the joint model provides coherence in the model, it also carries additional computational challenges.

By explicitly separating inference of morphological (form) variation from the gene expression variation, the 3-step model only tests for gene expression correlates of the largest sources of variation in larval morphology. In the joint model, information from the gene expression measurements is used directly during the inference of the factors underlying variation in larval form. Thus, the joint analysis can identify gene expression correlates of form factors more directly related to gene expression. Whether this behavior is desirable depends on the the research question. The 3-step model may be sufficient to identify major determinants of larval morphology, because the first two steps will retain all the major axes of variation in form. However, the joint model should provide a more comprehensive description of the relationship between gene expression and morphology. By conditioning on the gene expression in (5.7), axes of morphological variation that are not correlated with any genes will

be less influential, allowing for better inference of more relevant gene expression - morphological correlations.

The 3-step model is analogous to multivariate principle components regression where principle components of the dependent variables are calculated first, and then each principle component axis is treated as a unique trait and regressed on the full set of independent variables. However, we instead of fixing these axes, we perform the regression with the equivalence class of factors defined up to a rotation. This is necessary because the factors themselves are only estimated up to this equivalence class. The joint model proposed in this chapter is analogous to a multivariate factor regression model. In this model, the covariance of the dependent variables and the independent variables are considered at once, and the factors in the dependent variables are estimated conditionally given the gene expression, rather than marginally. This is similar in principle to the two-block partial least-squares analysis introduced for genetic analysis of shape traits by Rohlf and Corti (2000). However, in both the 3-step and the joint model I use the group Lasso prior to induce sparsity in gene expression regression coefficients. This sparsity provides variable selection, which is useful because experiments to validate the gene expression - morphology relationship will require perturbations to a low number of genes.

However, the joint analysis will be computationally demanding. The iteration time and mixing of the Gibbs sampler derived for characterizing form variation alone in Chapter 3 were slow. Adding draws of rotation, translation, residual form (5.1), and regression (5.7) parameters will slow each iteration of the sampler, and may make the mixing even slower. Therefore, it may not be feasible to exhaustively search the space of model dimensions in (5.5), as I did in Chapter 3. Other techniques that automatically select the number of factors in a factor model (eq. Carvalho et al 2008, Bhattacharya and Dunson 2011) may be more efficient. I have not attempted to fit this model yet. Designing efficient algorithms to fit this joint model will make

the methods developed in this thesis more useful to a broad set of problems in the genetic analysis of morphological shape and evolutionary developmental biology.

Bibliography

- Aguilar, O. and West, M. (2000), “Bayesian dynamic factor models and variance matrix discounting for portfolio allocation,” *Journal of Business & Economic Statistics*.
- Alberts, B. (2008), *Molecular biology of the cell*, New York : Garland Science.
- Armagan, A., Dunson, D. B., and Clyde, M. (2011), “Generalized Beta Mixtures of Gaussians,” *Neural Information Processing Systems*.
- Bookstein, F. L. (1986), “Size and Shape Spaces for Landmark Data in Two Dimensions,” *Statistical Science*, 1, 181–222.
- Dryden, I. L. and Mardia, K. V. (1998), *Statistical Shape Analysis*, John Wiley & Sons, New York, NY, 1 edn.
- Goodall, C. R. and Mardia, K. V. (1991), “A Geometrical Derivation of the Shape Density,” *Advances in Applied Probability*, 23, 496–514.
- Hart, M. W. and Strathmann, R. R. (1994), “Functional Consequences of Phenotypic Plasticity in Echinoid Larvae,” *The Biological bulletin*, 186, 291.
- Hoff, P. D. (2009), “Simulation of the Matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data,” *Journal of Computational and Graphical Statistics*.
- Kendall, D. G. (1984), “Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces,” *Bulletin of the London Mathematical Society*, 16, 81–121.
- Klingenberg, C. P. (2010), “Evolution and development of shape: integrating quantitative approaches.” *Nature Reviews Genetics*, 11, 623–635.
- Klingenberg, C. P., Barluenga, M., and Meyer, A. (2002), “Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry.” *Evolution; international journal of organic evolution*, 56, 1909–1920.
- Mardia, K. V., Bookstein, F. L., and Moreton, I. J. (2000), “Statistical assessment of bilateral symmetry of shapes,” *Biometrika*, 87, 285–300.

- McEdward, L. and Herrera, J. (1999), “Body form and skeletal morphometrics during larval development of the sea urchin *Lytechinus variegatus* Lamarck,” *Journal Of Experimental Marine Biology And Ecology*, 232, 151–176.
- Oliveri, P., Tu, Q., and Davidson, E. (2008), “Global regulatory logic for specification of an embryonic cell lineage,” *Proceedings of the National Academy of Sciences USA*, 105, 5955–5962.
- Peter, I. S. and Davidson, E. H. (2010), “The endoderm gene regulatory network in sea urchin embryos up to mid-blastula stage,” *Developmental biology*, 340, 188–199.
- Peter, I. S. and Davidson, E. H. (2011), “A gene regulatory network controlling the embryonic specification of endoderm.” *Nature*, 474, 635–639.
- Raff, R. A. (1996), *The shape of life*, genes, development, and the evolution of animal form, University Of Chicago Press.
- Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., and Roth, V. (2009), “The Bayesian Group-Lasso for Analyzing Contingency Tables,” in *Proceedings of the 26th International Conference on Machine Learning*, pp. 1–8, Montreal, Canada, ACM Press.
- Rohlf, F. J. and Corti, M. (2000), “Use of two-block partial least-squares to study covariation in shape,” *Systematic Biology*, 49, 740–753.
- Sharma, T. and Ettensohn, C. A. (2011), “Regulative deployment of the skeletogenic gene regulatory network during sea urchin development.” *Development*, 138, 2581–2590.
- Strathmann, R. R. (2006), “Good eaters, poor swimmers: compromises in larval form,” *Integrative and Comparative Biology*, 46, 312–322.
- Su, Y.-H., Li, E., Geiss, G. K., Longabaugh, W. J. R., Krämer, A., and Davidson, E. H. (2009), “A perturbation model of the gene regulatory network for oral and aboral ectoderm specification in the sea urchin embryo,” *Developmental biology*, 329, 410–421.
- Viklands, T. (2006), “Algorithms for the Weighted Orthogonal Procrustes Problem and other Least Squares Problems,” Ph.D. thesis, Umeå University, Umeå, Sweden.
- West, M. (2003), “Bayesian Factor Regression Models in the “Large p, Small n” Paradigm,” *Bayesian Statistics*.
- Yuan, M. and Lin, Y. (2006), “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 68, 49–67.