

A hierarchical eigenmodel for pooled covariance estimation

Peter D. Hoff ¹

Working Paper no. 86
Center for Statistics and the Social Sciences
University of Washington
Seattle, WA 98195-4322.

March 31, 2008

¹Departments of Statistics and Biostatistics and the Center for Statistics and the Social Sciences. University of Washington, Seattle, Washington 98195-4322. Web: <http://www.stat.washington.edu/hoff/>. This research was partially supported by NSF grant SES-0631531. The author thanks Michael Perlman for helpful discussions.

Abstract

While a set of covariance matrices corresponding to different populations are unlikely to be exactly equal they can still exhibit a high degree of similarity. For example, some pairs of variables may be positively correlated across most groups, while the correlation between other pairs may be consistently negative. In such cases much of the similarity across covariance matrices can be described by similarities in their principal axes, the axes defined by the eigenvectors of the covariance matrices. Estimating the degree of across-population eigenvector heterogeneity can be helpful for a variety of estimation tasks. Eigenvector matrices can be pooled to form a central set of principal axes, and to the extent that the axes are similar, covariance estimates for populations having small sample sizes can be stabilized by shrinking their principal axes towards the across-population center. To this end, this article develops a hierarchical model and estimation procedure for pooling principal axes across several populations. The model for the across-group heterogeneity is based on a matrix-valued antipodally symmetric Bingham distribution that can flexibly describe notions of “center” and “spread” for a population of orthonormal matrices.

Some key words: Bayesian inference, copula, Markov chain Monte Carlo, principal components, random matrix, Stiefel manifold.

1 Introduction

Principal component analysis is a well-established procedure for describing the features of a covariance matrix. Letting $\mathbf{U}\Lambda\mathbf{U}^T$ be the eigenvalue decomposition of the covariance matrix of a p -dimensional random vector \mathbf{y} , the principal components of \mathbf{y} are the elements of the transformed mean-zero vector $\mathbf{U}^T(\mathbf{y} - \mathbf{E}[\mathbf{y}])$. From the orthonormality of \mathbf{U} it follows that the elements of the principal component vector are uncorrelated, with variances equal to the diagonal of Λ . Perhaps more importantly, the matrix \mathbf{U} provides a natural coordinate system for describing the orientation of the multivariate density of \mathbf{y} : Letting \mathbf{u}_j denote the j th column of \mathbf{U} , \mathbf{y} can be expressed as $\mathbf{y} - \mathbf{E}[\mathbf{y}] = z_1\mathbf{u}_1 + \cdots + z_p\mathbf{u}_p$, where $(z_1, \dots, z_p)^T$ is a vector of uncorrelated mean-zero random variables with diagonal covariance matrix Λ .

Often the same set of variables are measured in multiple populations. Even if the covariance matrices differ across populations, it is natural to expect that they share some common structure, such as the correlations between some pairs of variables having common signs across the populations. With this situation in mind, Flury [1984] developed estimation and testing procedures for the ‘‘common principal components’’ model, in which a set of covariance matrices $\{\Sigma_1, \dots, \Sigma_K\}$ have common eigenvectors, so that $\Sigma_j = \mathbf{U}\Lambda_j\mathbf{U}^T$ for each $j \in \{1, \dots, K\}$. A number of variations of this model have since appeared: Flury [1987] and Schott [1991, 1999] consider cases in which only certain columns or subspaces of \mathbf{U} are shared across populations, and Boik [2002] describes a very general model in which eigenspaces can be shared between all or some of the populations.

These approaches all assume that certain eigenspaces are either exactly equal or completely distinct across a collection of covariances matrices. In many cases these two alternatives are too extreme, and it may be desirable to recognize situations in which eigenvectors are similar but not exactly equal. To this end, this article develops a hierarchical model to assess heterogeneity of principal axes across a set of populations. This is accomplished with the aid of a probability distribution over the orthogonal group \mathcal{O}_p which can be used in a hierarchical model for sample covariance matrices, allowing for pooling of covariance information and a description of similarities and differences across populations. Specifically, this article develops a sampling model for across-population covariance heterogeneity in which

$$\begin{aligned} p(\mathbf{U}|\mathbf{A}, \mathbf{B}, \mathbf{V}) &= c(\mathbf{A}, \mathbf{B})\text{etr}(\mathbf{B}\mathbf{U}^T\mathbf{V}\mathbf{A}\mathbf{V}^T\mathbf{U}) \\ \mathbf{U}_1, \dots, \mathbf{U}_K &\sim \text{i.i.d. } p(\mathbf{U}|\mathbf{A}, \mathbf{B}, \mathbf{V}) \\ \Sigma_k &= \mathbf{U}_k\Lambda_k\mathbf{U}_k^T, \end{aligned} \tag{1}$$

where \mathbf{A} and \mathbf{B} are diagonal matrices and $\mathbf{V} \in \mathcal{O}_p$. The above distribution is a type of generalized Bingham distribution [Khatri and Mardia, 1977, Gupta and Nagar, 2000] that is appropriate for modeling principal component axes. Section 2 of this article describes some features of this distribution, in particular how \mathbf{A} and \mathbf{B} represent the variability of $\{\mathbf{U}_1, \dots, \mathbf{U}_K\}$ and how \mathbf{V}

represents the mode. Parameter estimation is discussed in Section 3, in which a Markov chain Monte Carlo algorithm is developed which allows for the joint estimation of $\{\mathbf{A}, \mathbf{B}, \mathbf{V}\}$ as well as $\{(\mathbf{U}_k, \Lambda_k), k = 1, \dots, K\}$. The estimation scheme is illustrated with two example data analyses in Sections 4 and 5. The first dataset, previously analyzed by Flury [1984] and Boik [2002] among others, involves skull measurement data on four populations of voles. Model diagnostics and comparisons indicate that the proposed hierarchical model represents certain features of the observed covariance matrices better than do less flexible models. The second example involves survey data from different states across the U.S.. The number of observations per state varies a great deal, and many states have only a few observations. The example shows how the hierarchical model shrinks correlation estimates towards the across-group center when within-group data is limited. Section 6 provides a discussion of the hierarchical model and a few of extensions of the approach.

2 A generalized Bingham distribution

The eigenvalue decomposition of a positive definite covariance matrix Σ is given by $\Sigma = \mathbf{U}\Lambda\mathbf{U}^T$, where Λ is a diagonal matrix of positive numbers $(\lambda_1, \dots, \lambda_p)$ and \mathbf{U} is an orthonormal matrix, so that $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$. Writing $\mathbf{U}\Lambda\mathbf{U}^T = \sum_{j=1}^p \lambda_j \mathbf{u}_j \mathbf{u}_j^T$, we see that multiplication of a column of \mathbf{U} by -1 does not change the value of the covariance matrix, highlighting the fact that the columns of \mathbf{U} represent not directions of variation, but axes. As such, any probability model representing variability across a set of principal axes should be antipodally symmetric in the columns of \mathbf{U} , meaning that \mathbf{U} is equal in distribution to $\mathbf{U}\mathbf{S}$ for any diagonal matrix \mathbf{S} having diagonal elements equal to plus or minus one, since \mathbf{U} and $\mathbf{U}\mathbf{S}$ represent the same axes.

Bingham [1974] described a probability distribution having a density proportional to $\exp\{\mathbf{u}^T \mathbf{G} \mathbf{u}\}$ for normal vectors $\{\mathbf{u} : \mathbf{u}^T \mathbf{u} = 1\}$. This density has antipodal symmetry, making it a candidate model for a random axis. Khatri and Mardia [1977] and Gupta and Nagar [2000] discuss a matrix-variate version of the Bingham distribution,

$$p(\mathbf{U}|\mathbf{G}, \mathbf{H}) \propto \text{etr}(\mathbf{H}\mathbf{U}^T \mathbf{G} \mathbf{U}) \quad (2)$$

where \mathbf{G} and \mathbf{H} are $p \times p$ symmetric matrices. Using the eigenvalue decompositions $\mathbf{G} = \mathbf{V}\mathbf{A}\mathbf{V}^T$ and $\mathbf{H} = \mathbf{W}\mathbf{B}\mathbf{W}^T$, this density can be rewritten as $p(\mathbf{U}|\mathbf{A}, \mathbf{B}, \mathbf{V}, \mathbf{W}) \propto \text{etr}(\mathbf{B}[\mathbf{W}^T \mathbf{U}^T \mathbf{V}] \mathbf{A} [\mathbf{V}^T \mathbf{U} \mathbf{W}])$. A well-known feature of this density is that it depends on \mathbf{A} and \mathbf{B} only through the differences among their diagonal elements. This is because for any orthonormal matrix \mathbf{X} (such as $\mathbf{V}^T \mathbf{U} \mathbf{W}$), we have

$$\begin{aligned} \text{tr}([\mathbf{B} + d\mathbf{I}]\mathbf{X}^T[\mathbf{A} + c\mathbf{I}]\mathbf{X}) &= \text{tr}(\mathbf{B}\mathbf{X}^T \mathbf{A} \mathbf{X}) + d \times \text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X}) + c \times \text{tr}(\mathbf{B}\mathbf{X}^T \mathbf{X}) + cd \times \text{tr}(\mathbf{X}^T \mathbf{X}) \\ &= \text{tr}(\mathbf{B}\mathbf{X}^T \mathbf{A} \mathbf{X}) + d \times \text{tr}(\mathbf{A}) + c \times \text{tr}(\mathbf{B}) + cdp \end{aligned}$$

and so the probability densities $p(\mathbf{U}|\mathbf{A}, \mathbf{B})$ and $p(\mathbf{U}|\mathbf{A} + c\mathbf{I}, \mathbf{B} + d\mathbf{I})$ are proportional as functions of \mathbf{U} and therefore equal. By convention \mathbf{A} and \mathbf{B} are usually taken to be non-negative. In what follows we will set the smallest eigenvalues a_p and b_p to be equal to zero.

Although a flexible class of distributions for orthonormal matrices, densities of the form (2) are not necessarily antipodally symmetric. However, we can identify conditions on \mathbf{G} and \mathbf{H} which give the desired symmetry. If either \mathbf{G} or \mathbf{H} have only one unique eigenvalue, then $\text{tr}(\mathbf{H}\mathbf{U}^T\mathbf{G}\mathbf{U})$ is constant in \mathbf{U} and therefore trivially antipodally symmetric in the columns of \mathbf{U} . Otherwise, we have the following result:

Proposition 1. *If \mathbf{G} and \mathbf{H} both have more than one unique eigenvalue, then a necessary and sufficient condition for $\text{tr}(\mathbf{H}\mathbf{U}^T\mathbf{G}\mathbf{U})$ to be antipodally symmetric in the columns of \mathbf{U} is that \mathbf{H} be a diagonal matrix.*

Proof. The symmetry condition requires that $\text{tr}(\mathbf{H}\mathbf{U}^T\mathbf{G}\mathbf{U}) = \text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}\mathbf{U}^T\mathbf{G}\mathbf{U})$ for all $\mathbf{U} \in \mathcal{O}_p$ and diagonal sign matrices \mathbf{S} . If \mathbf{H} is diagonal then $\mathbf{S}\mathbf{H}\mathbf{S} = \mathbf{H}$ and so diagonality is a sufficient condition for antipodal symmetry. To show that it is also a necessary condition, first let \mathbf{V} and \mathbf{A} be the eigenvector and eigenvalue matrices of \mathbf{G} . Then for any orthonormal \mathbf{X} we have $\mathbf{X} = \mathbf{U}^T\mathbf{V}$ if $\mathbf{U} = \mathbf{V}\mathbf{X}^T$. Therefore the symmetry condition is that

$$\text{tr}([\mathbf{H} - \mathbf{S}\mathbf{H}\mathbf{S}]\mathbf{X}\mathbf{A}\mathbf{X}^T) = 0 \quad \text{for all } \mathbf{X} \in \mathcal{O}_p \text{ and } \mathbf{S} \in \{\text{diag}(\mathbf{s}), \mathbf{s} \in \{\pm 1\}^p\}.$$

If we let $\text{diag}(\mathbf{S}) = (-1, 1, 1, \dots, 1)$ then $\mathbf{D} = \mathbf{H} - \mathbf{S}\mathbf{H}\mathbf{S}$ is zero except for possibly $\mathbf{D}_{[1,-1]}$ and $\mathbf{D}_{[-1,1]}$, the first row and column of \mathbf{D} absent $d_{1,1}$. Additionally, if not all entries are zero then \mathbf{D} is of rank two with eigenvectors \mathbf{e} and $\mathbf{S}\mathbf{e}$, where $\mathbf{e} = (|\mathbf{H}_{[-1,1]}|, h_{2,1}, h_{3,1}, \dots, h_{p,1}) / (\sqrt{2}|\mathbf{H}_{[-1,1]}|)$, and corresponding eigenvalues $\pm d$, where $d = 2|\mathbf{H}_{[-1,1]}|$. Writing the symmetry condition in terms of the eigenvalues and vectors of \mathbf{D} gives

$$0 = \text{tr}(\mathbf{D}\mathbf{X}\mathbf{A}\mathbf{X}^T) = d \sum_{i=1}^p a_i \mathbf{x}_i^T (\mathbf{e}\mathbf{e}^T - \mathbf{S}\mathbf{e}\mathbf{e}^T\mathbf{S}) \mathbf{x}_i.$$

The symmetry condition requires that this hold for all orthonormal \mathbf{X} . Now if k and l are the indices of any two unequal eigenvalues of \mathbf{G} , we can let $\mathbf{x}_k = \mathbf{e}$ and $\mathbf{x}_l = -\mathbf{S}\mathbf{e}$, giving

$$\begin{aligned} 0 = \text{tr}(\mathbf{D}\mathbf{X}\mathbf{A}\mathbf{X}^T) &= d[a_k(1 - \mathbf{e}^T\mathbf{S}\mathbf{e}\mathbf{e}^T\mathbf{S}\mathbf{e}) + a_l(\mathbf{e}^T\mathbf{S}\mathbf{e}\mathbf{e}^T\mathbf{S}\mathbf{e} - 1)] \\ &= d(1 - [\mathbf{e}^T\mathbf{S}\mathbf{e}]^2)(a_k - a_l). \end{aligned}$$

Since $a_k \neq a_l$ by assumption, this means that either $d = 0$ or $(\mathbf{e}^T\mathbf{S}\mathbf{e})^2 = 1$. Neither of these conditions are met unless all entries of \mathbf{D} are zero, implying that the off-diagonal elements in the first row and column of \mathbf{H} must be zero. Repeating this argument with the diagonal of \mathbf{S} ranging over all p -vectors consisting of one negative-one and $p - 1$ positive ones shows that all off-diagonal elements of \mathbf{H} must be zero. \square

Based on this results we fix the eigenvector matrix of \mathbf{H} to be \mathbf{I} and our column-wise antipodally symmetric model for $\mathbf{U} \in \mathcal{O}_p$ is

$$p_B(\mathbf{U}|\mathbf{A}, \mathbf{B}, \mathbf{V}) = c(\mathbf{A}, \mathbf{B}) \text{etr}(\mathbf{B}\mathbf{U}^T \mathbf{V} \mathbf{A} \mathbf{V}^T \mathbf{U}) \quad (3)$$

where \mathbf{A} and \mathbf{B} are diagonal matrices with $a_1 \geq a_2 \geq \dots \geq a_p = 0$, $b_1 \geq b_2 \geq \dots \geq b_p = 0$ and $\mathbf{V} \in \mathcal{O}_p$. Interpreting these parameters is made easier by writing $\mathbf{X} = \mathbf{V}^T \mathbf{U}$ and expanding out the exponent of p_B as

$$\text{tr}(\mathbf{B}\mathbf{U}^T \mathbf{V} \mathbf{A} \mathbf{V}^T \mathbf{U}) = \sum_{i=1}^p \sum_{j=1}^p a_i b_j (\mathbf{v}_i^T \mathbf{u}_j)^2 = \sum_{i=1}^p \sum_{j=1}^p a_i b_j x_{i,j}^2 = \mathbf{a}^T (\mathbf{X} \circ \mathbf{X}) \mathbf{b}, \quad (4)$$

where “ \circ ” is the Hadamard product denoting element-wise multiplication. The value of $x_{i,j}^2$ describes how close column i of \mathbf{V} is to column j of \mathbf{U} . Since both a and b are in decreasing order, $a_1 b_1$ is the largest term and the density will be large when $x_{1,1}^2$ is large. However, due to the orthonormality of \mathbf{X} , a large $x_{1,1}^2$ restricts $x_{1,2}^2$ and $x_{2,1}^2$ to be small, which then allows $x_{2,2}^2$ to be large. Continuing on this way suggests that the density is maximized if $\mathbf{X} \circ \mathbf{X}$ is the identity, i.e. $\mathbf{U} = \mathbf{V}\mathbf{S}$ for some diagonal sign matrix \mathbf{S} .

Proposition 2. *The modes of p_B include \mathbf{V} and $\{\mathbf{V}\mathbf{S} : \mathbf{S} = \text{diag}(\mathbf{s}), \mathbf{s} \in \{\pm 1\}^p\}$. If the diagonal elements of \mathbf{A} and \mathbf{B} are distinct then these are the only modes.*

Proof. The matrix $\mathbf{X} \circ \mathbf{X}$ is an element of the set of orthostochastic matrices, a subset of the doubly stochastic matrices. The set of doubly stochastic matrices is a compact convex set whose extreme points are the permutation matrices. Since every element of this compact convex set can be written as a convex combination of the extreme points, we have

$$\max_{\mathbf{X} \in \mathcal{O}_p} \mathbf{a}^T (\mathbf{X} \circ \mathbf{X}) \mathbf{b} \leq \max_{\theta} \mathbf{a}^T \left(\sum \theta_k \mathbf{P}_k \right) \mathbf{b}$$

for probability distributions θ over the finite set of permutation matrices. If the elements of \mathbf{a} and \mathbf{b} are distinct and ordered it is easy to show that $\mathbf{a}^T \mathbf{P} \mathbf{b}$ is uniquely maximized over permutation matrices \mathbf{P} by $\mathbf{P} = \mathbf{I}$, and so the right-hand side is maximized when θ is the point-mass measure on \mathbf{I} . Since \mathbf{I} is orthostochastic, the maximum on the left-hand side is achieved at $(\mathbf{X} \circ \mathbf{X}) = \mathbf{I}$. \square

If two adjacent eigenvalues are equal then the density has additional maxima. For example, if $b_1 = b_2$ then $\mathbf{a}^T \mathbf{Z} \mathbf{b}$ is maximized over doubly stochastic matrices \mathbf{Z} by any convex combination of the matrices corresponding to the permutations $\{1, 2, 3, \dots, p\}$ and $\{2, 1, 3, \dots, p\}$. In terms of \mathbf{U} , this would mean that modes are such that $\mathbf{u}_i^T \mathbf{v}_i = \pm 1$ for $i > 2$, with \mathbf{u}_1 and \mathbf{u}_2 being any orthonormal vectors in the null space of $\{\mathbf{v}_3, \dots, \mathbf{v}_p\}$. More generally, how the parameters (\mathbf{A}, \mathbf{B})

control the variability of \mathbf{U} around \mathbf{V} can be seen by rewriting the exponent of (3) a few different ways. For example, equation (4), can be expressed as

$$\begin{aligned} \text{tr}(\mathbf{B}\mathbf{U}^T\mathbf{V}\mathbf{A}\mathbf{V}^T\mathbf{U}) &= \sum_{i=1}^p \sum_{j=1}^p a_i b_j (\mathbf{v}_i^T \mathbf{u}_j)^2 \\ &= \sum_{j=1}^p b_j \mathbf{u}_j^T (\mathbf{V}\mathbf{A}\mathbf{V}^T) \mathbf{u}_j \end{aligned} \quad (5)$$

From equations (4) and (5) it is clear that $b_j = b_{j+1}$ implies that $\mathbf{u}_j \stackrel{d}{=} \mathbf{u}_{j+1}$ and $a_i = a_{i+1}$ implies $\mathbf{v}_i^T \mathbf{u}_j \stackrel{d}{=} \mathbf{v}_{i+1}^T \mathbf{u}_j$. In this way the model can represent *eigenspaces* of high probability, not only eigenvectors, providing a probabilistic analog to the common space models of Flury [1987]. To illustrate this further, Figure 1 shows the expectations of the squared elements of $\mathbf{X} = \mathbf{V}^T\mathbf{U}$ for two different values of (\mathbf{A}, \mathbf{B}) (calculations were based on a Monte Carlo approximation scheme described in Hoff [2007b]). The plot in the first panel is based on the generalized Bingham distribution in which $\text{diag}(\mathbf{A}) = \text{diag}(\mathbf{B}) = (7, 5, 3, 0, 0, 0)$. For this distribution, \mathbf{u}_1 , \mathbf{u}_2 and \mathbf{u}_3 are highly concentrated around \mathbf{v}_1 , \mathbf{v}_2 and \mathbf{v}_3 respectively. Since $b_4 = b_5 = b_6$, the vectors \mathbf{u}_4 , \mathbf{u}_5 and \mathbf{u}_6 are equal in distribution and close to being uniformly distributed on the null space of $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$. These particular values of (\mathbf{A}, \mathbf{B}) could represent a situation in which the first three eigenvectors are conserved across populations but the others are not. The second panel of Figure 1 represents a more complex situation in which $\text{diag}(\mathbf{A}) = (7, 5, 3, 0, 0, 0)$ and $\text{diag}(\mathbf{B}) = (7, 7, 0, 0, 0, 0)$. For these parameter values the following components of $\mathbf{X} = \mathbf{V}^T\mathbf{U}$ are equal in distribution: columns 1 and 2; columns 3, 4, 5 and 6; rows 2 and 3; rows 4, 5 and 6. Such a distribution might represent a situation in which a vector \mathbf{v}_1 is shared across populations, but it is equally likely to be represented within a population by either \mathbf{u}_1 or \mathbf{u}_2 .

3 Pooled estimation of covariance eigenstructure

In the case of normally distributed data the sampling model for $\mathbf{S}_k = \mathbf{Y}_k^T (\mathbf{I} - \frac{1}{n_k} \mathbf{1}\mathbf{1}^T) \mathbf{Y}_k$, the observed sum of squares matrix in population k , is a Wishart distribution. Combining this with the model for principal axes developed in the last section gives the following hierarchical model for covariance structure:

$$\begin{aligned} \mathbf{U}_1, \dots, \mathbf{U}_K &\sim \text{i.i.d. } p(\mathbf{U} | \mathbf{A}, \mathbf{B}, \mathbf{V}) && \text{(across-population variability)} \\ \mathbf{S}_k &\sim \text{Wishart}(\mathbf{U}_k \Lambda_k \mathbf{U}_k^T, n_k - 1) && \text{(within-population variability)} \end{aligned}$$

The unknown parameters to estimate include $\{\mathbf{A}, \mathbf{B}, \mathbf{V}\}$ as well as the within-population covariance matrices, parameterized as $\{(\mathbf{U}_1, \Lambda_1), \dots, (\mathbf{U}_K, \Lambda_K)\}$. In this section we describe a Markov chain Monte Carlo algorithm that generates approximate samples from the posterior distribution for these

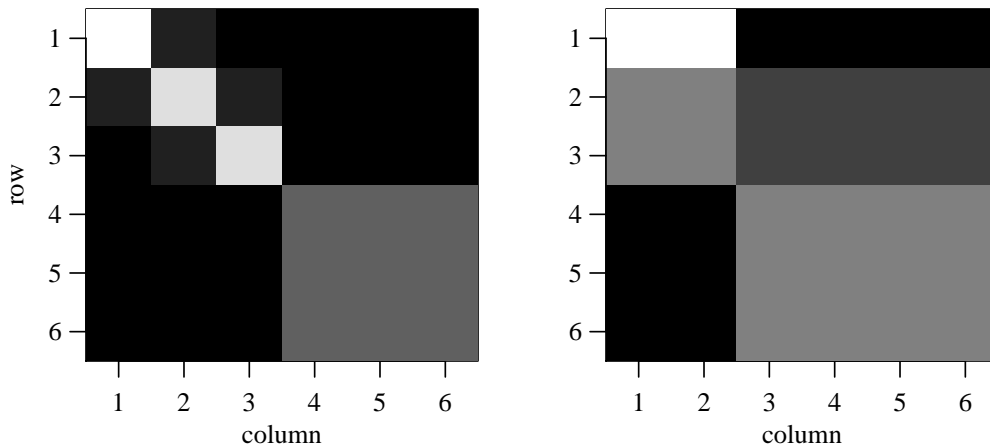


Figure 1: Expected values of the squared entries of $\mathbf{X} = \mathbf{V}^T \mathbf{U}$ under two different Bingham distributions. Light shading indicates high values.

parameters, allowing for estimation and inference. The Markov chain is constructed with Gibbs sampling, in which each parameter is iteratively resampled from its full conditional distributions. We first describe conditional updates of the across-population parameters $\{\mathbf{A}, \mathbf{B}, \mathbf{V}\}$, then describe pooled estimation of each \mathbf{U}_k , which combines the population-specific information \mathbf{S}_k with the across-population information in $\{\mathbf{A}, \mathbf{B}, \mathbf{V}\}$.

3.1 Estimation of across-population parameters

Letting the prior distribution for \mathbf{V} be the uniform (invariant) measure on \mathcal{O}_p , we have

$$\begin{aligned}
 p(\mathbf{V} | \mathbf{A}, \mathbf{B}, \mathbf{U}_1, \dots, \mathbf{U}_K) &\propto p(\mathbf{V}) \prod_{k=1}^K p(\mathbf{U}_k | \mathbf{A}, \mathbf{B}, \mathbf{V}) \\
 &\propto \text{etr} \left(\sum_{k=1}^K \mathbf{B} \mathbf{U}_k^T \mathbf{V} \mathbf{A} \mathbf{V}^T \mathbf{U}_k \right) \\
 &= \text{etr} \left(\sum_{k=1}^K \mathbf{A} \mathbf{V}^T \mathbf{U}_k \mathbf{B} \mathbf{U}_k^T \mathbf{V} \right) = \text{etr} \left(\mathbf{A} \mathbf{V}^T \left[\sum_{k=1}^K \mathbf{U}_k \mathbf{B} \mathbf{U}_k^T \right] \mathbf{V} \right),
 \end{aligned}$$

and so the full conditional distribution of \mathbf{V} is a generalized Bingham distribution, of the same form as described in the previous section. Conditional on the values $\{\mathbf{A}, \mathbf{B}, \mathbf{U}_1, \dots, \mathbf{U}_K\}$, pairs of columns of \mathbf{V} can be sampled from their full conditional distributions using a method described in Hoff [2007b].

Obtaining full conditional distributions for \mathbf{A} and \mathbf{B} is more complicated. The joint density of

$\{\mathbf{U}_1, \dots, \mathbf{U}_K\}$ is $c(\mathbf{A}, \mathbf{B})^K \text{etr}(\sum \mathbf{A} \mathbf{V}^T \mathbf{U}_k \mathbf{B} \mathbf{U}_k^T \mathbf{V})$. From the terms in the exponent we have

$$\text{tr}(\sum_{k=1}^K \mathbf{B} \mathbf{U}_k^T \mathbf{V}^T \mathbf{A} \mathbf{V}^T \mathbf{U}_k) = \sum_{k=1}^K \text{tr}(\mathbf{B} \mathbf{U}_k^T \mathbf{V} \mathbf{A} \mathbf{V}^T \mathbf{U}_k) = \sum_{k=1}^K \sum_{i=1}^p \sum_{j=1}^p a_i b_j (\mathbf{v}_i^T \mathbf{u}_{j,k})^2 = \mathbf{a}^T \mathbf{M} \mathbf{b}$$

where \mathbf{M} is the matrix $\mathbf{M} = \sum_{k=1}^K (\mathbf{V}^T \mathbf{U}_k) \circ (\mathbf{V}^T \mathbf{U}_k)$. The normalizing constant $c(\mathbf{A}, \mathbf{B})$ is equal to ${}_0F_0(\mathbf{A}, \mathbf{B})^{-1}$, where ${}_0F_0(\mathbf{A}, \mathbf{B})$ is a type of hypergeometric function with matrix arguments [Herz, 1955]. Exact calculation of this quantity is problematic, although approximations have been discussed in Anderson [1965], Constantine and Muirhead [1976] and Muirhead [1978]. The first-order term in these approximations is

$$c(\mathbf{A}, \mathbf{B}) \approx \tilde{c}(\mathbf{A}, \mathbf{B}) = 2^{-p} \pi^{-\binom{p}{2}} e^{-\mathbf{a}^T \mathbf{b}} \prod_{i < j} (a_i - a_j)^{1/2} (b_i - b_j)^{1/2}.$$

This gives the following approximation to the likelihood for \mathbf{A}, \mathbf{B} :

$$\begin{aligned} p(\mathbf{U}_1, \dots, \mathbf{U}_K | \mathbf{A}, \mathbf{B}, \mathbf{V}) &\approx \tilde{c}(\mathbf{A}, \mathbf{B})^K \text{etr}(\mathbf{a}^T \mathbf{M} \mathbf{b}) \\ &\propto \exp\{-\mathbf{a}^T (\mathbf{K} \mathbf{I} - \mathbf{M}) \mathbf{b}\} \prod_{i < j} (a_i - a_j)^{K/2} (b_i - b_j)^{K/2} \end{aligned} \quad (6)$$

However, there is an identifiability issue with this likelihood: As seen above, since \mathbf{A} and \mathbf{B} are diagonal, $\text{tr}(\mathbf{B} \mathbf{X}^T \mathbf{A} \mathbf{X})$ simplifies to $\mathbf{a}^T (\mathbf{X} \circ \mathbf{X}) \mathbf{b} = \sum_i \sum_j a_i b_j x_{i,j}^2$. This means that for any $c > 0$, $p(\mathbf{U} | \mathbf{A}, \mathbf{B}, \mathbf{V}) = p(\mathbf{U} | c\mathbf{A}, c^{-1}\mathbf{B}, \mathbf{V})$, and so the scale of \mathbf{A} and \mathbf{B} are not separately identifiable. To account for this, we parameterize \mathbf{A} and \mathbf{B} as follows:

$$\begin{aligned} \text{diag}(\mathbf{A}) &= (a_1, \dots, a_p) = \sqrt{w}(\alpha_1, \dots, \alpha_p) \\ \text{diag}(\mathbf{B}) &= (b_1, \dots, b_p) = \sqrt{w}(\beta_1, \dots, \beta_p) \end{aligned}$$

where $w > 0$, $1 = \alpha_1 > \alpha_2 > \dots > \alpha_{p-1} > \alpha_p = 0$ and $1 = \beta_1 > \beta_2 > \dots > \beta_{p-1} > \beta_p = 0$. Rewriting (6) in terms of these parameters and multiplying by a prior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, w)$ gives

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, w | \mathbf{M}) \propto p(\boldsymbol{\alpha}, \boldsymbol{\beta}, w) \times \exp\{-w \boldsymbol{\alpha}^T (\mathbf{K} \mathbf{I} - \mathbf{M}) \boldsymbol{\beta}\} w^{\binom{p}{2} K/2} \prod_{i < j} (\alpha_i - \alpha_j)^{K/2} (\beta_i - \beta_j)^{K/2}. \quad (7)$$

In what follows we take the prior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, w)$ such that $1 > \alpha_2 > \dots > \alpha_{p-1} > 0$ and $1 > \beta_2 > \dots > \beta_{p-1} > 0$ are two independent sets of order statistics of uniform random variables on $[0, 1]$, and w has a gamma distribution. With these priors, the values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be sampled from their full conditional distributions on a grid of $[0, 1]$, and the full conditional distribution of w is a gamma distribution. For example, if $w \sim \text{gamma}(\eta_0/2, \tau_0^2/2)$ *a priori*, then $p(w | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{M})$ is $\text{gamma}(\eta_0/2 + \binom{p}{2} K/2, \eta_0 \tau_0^2/2 + \boldsymbol{\alpha}^T (\mathbf{K} \mathbf{I} - \mathbf{M}) \boldsymbol{\beta})$.

We should keep in mind that this full conditional distribution is based on an approximation to the normalizing constant $c(\mathbf{A}, \mathbf{B})$ (although it is a “bona fide” full conditional distribution under

the prior $\tilde{p}(\mathbf{A}, \mathbf{B}) = p(\mathbf{A}, \mathbf{B})[\tilde{c}(\mathbf{A}, \mathbf{B})/c(\mathbf{A}, \mathbf{B})]^K$. Results from Anderson [1965] show that in terms of the parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, w\}$,

$$\frac{c(\mathbf{A}, \mathbf{B})}{\tilde{c}(\mathbf{A}, \mathbf{B})} \approx 1 + \frac{1}{4w} \sum_{i < j} [(\alpha_i - \alpha_j)(\beta_i - \beta_j)]^{-1} + O\left(\frac{1}{w^2}\right) \quad (8)$$

Further terms in the expansion are available in Anderson [1965]. In problems where w is large then the approximation is likely to be a good one. In cases where the differences between consecutive α_i 's or β_j 's is small compared to $1/w$ then it may be desirable to correct for the approximation using a few additional terms from (8) via a Metropolis-Hastings procedure. For example, letting $h(\boldsymbol{\alpha}, \boldsymbol{\beta}, w) = 1 + \sum_{i < j} [4w(\alpha_i - \alpha_j)(\beta_i - \beta_j)]^{-1}$, if w_s is the current value of w in the Markov chain and \tilde{w} is sampled from the approximate full conditional distribution of w based on (7), then the correction can be implemented as follows:

1. sample a proposal \tilde{w} from the full conditional of w based on (7);
2. sample $u \sim \text{uniform}(0,1)$ and compute $r = [h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \tilde{w})/h(\boldsymbol{\alpha}, \boldsymbol{\beta}, w_s)]^K$;
3. if $u < r$ then set $w_{s+1} = \tilde{w}$, otherwise set $w_{s+1} = w_s$.

A similar procedure using one more order in the expansion is used in the example data analyses of the next section.

Finally, we note that there has been some recent progress in computing ${}_0F_0(\mathbf{A}, \mathbf{B})$ exactly. Koev and Edelman [2006] provide an algorithm that is fast enough to be used in MCMC algorithms for problems in which p roughly 5 or less and the values of \mathbf{A} and \mathbf{B} are not too large. For other problems the approximations based on (7) and (8) still seem necessary.

3.2 Estimation of population-specific principal axes

As described above, the within-population sampling model for the sample sum-of-squares matrix \mathbf{S}_k is $\text{Wishart}(\mathbf{U}_k \Lambda_k \mathbf{U}_k^T, n_k - 1)$, so that as a function of \mathbf{S}_k , \mathbf{U}_k and Λ_k ,

$$p(\mathbf{S}_k | \mathbf{U}_k, \Lambda_k) \propto |\Lambda_k|^{-(n_k-1)/2} |\mathbf{S}_k|^{(n_k-p-2)/2} \times \text{etr}\left(-\frac{1}{2} \Lambda_k^{-1} \mathbf{U}_k^T \mathbf{S}_k \mathbf{U}_k\right). \quad (9)$$

In the absence of information from other populations, a uniform prior distribution on \mathbf{U}_k would yield a generalized Bingham full conditional distribution for \mathbf{U}_k . However, combining (9) with the across-population information $p(\mathbf{U}_k | \mathbf{A}, \mathbf{B}, \mathbf{V})$ gives a non-standard full conditional distribution for \mathbf{U}_k :

$$\begin{aligned} p(\mathbf{U}_k | \mathbf{S}_k, \Lambda_k, \mathbf{A}, \mathbf{B}, \mathbf{V}) &\propto p(\mathbf{S}_k | \mathbf{U}_k, \Lambda_k) \times p(\mathbf{U}_k | \mathbf{A}, \mathbf{B}, \mathbf{V}) \\ &\propto \text{etr}\left(-\frac{1}{2} \Lambda_k^{-1} \mathbf{U}_k^T \mathbf{S}_k \mathbf{U}_k\right) \times \text{etr}(\mathbf{B} \mathbf{U}_k^T \mathbf{V} \mathbf{A} \mathbf{V}^T \mathbf{U}_k). \end{aligned}$$

The terms in the exponents are difficult to combine as they are both quadratic in \mathbf{U}_k . Writing out the expression in terms of the columns $\{\mathbf{u}_{1,k}, \dots, \mathbf{u}_{p,k}\}$ yields some insight:

$$\text{tr}(\mathbf{B}\mathbf{U}_k^T\mathbf{V}\mathbf{A}\mathbf{V}^T\mathbf{U}_k - \frac{1}{2}\Lambda_k^{-1}\mathbf{U}_k^T\mathbf{S}_k\mathbf{U}_k) = \sum_{j=1}^p \mathbf{u}_{j,k}^T \left(b_j\mathbf{V}\mathbf{A}\mathbf{V}^T - \frac{1}{2}\lambda_{j,k}^{-1}\mathbf{S}_k \right) \mathbf{u}_{j,k}$$

This suggests that the full conditional distribution of the j th column vector of \mathbf{U}_k is a vector-valued Bingham distribution. This is true in a very limited sense: Since \mathbf{U}_k is an orthonormal matrix the full conditional distribution of $\mathbf{u}_{j,k}$ given the other columns of \mathbf{U}_k must have support only on $\pm\tilde{\mathbf{u}}$, where $\tilde{\mathbf{u}}$ represents the null space of the vectors $\{\mathbf{u}_{1,k}, \dots, \mathbf{u}_{j-1,k}, \mathbf{u}_{j+1,k}, \dots, \mathbf{u}_{p,k}\}$. Iteratively sampling the columns of \mathbf{U}_k from their full conditional distributions would therefore produce a reducible Markov chain which would not converge to the target posterior distribution. One remedy to this situation, used by Hoff [2007b] in the context of sampling from the Bingham distribution, is to sample from the full conditional distribution of columns taken two at a time. Conditional on $\{\mathbf{u}_{3,k}, \dots, \mathbf{u}_{p,k}\}$, the vectors $\{\mathbf{u}_{1,k}, \mathbf{u}_{2,k}\}$ are equal in distribution to $\mathbf{N}\mathbf{Z}$, where \mathbf{N} is any $p \times 2$ dimensional orthonormal basis for the null space of $\{\mathbf{u}_{3,k}, \dots, \mathbf{u}_{p,k}\}$ and \mathbf{Z} is a random 2×2 orthonormal matrix whose density with respect to the uniform measure is proportional to

$$p(\mathbf{Z}) \propto \text{etr}(\mathbf{z}_1^T \mathbf{G} \mathbf{z}_1 + \mathbf{z}_2^T \mathbf{H} \mathbf{z}_2),$$

where $\mathbf{G} = \mathbf{N}^T(b_1\mathbf{V}\mathbf{A}\mathbf{V}^T - \lambda_{1,k}^{-1}\mathbf{S}_k)\mathbf{N}$, $\mathbf{H} = \mathbf{N}^T(b_2\mathbf{V}\mathbf{A}\mathbf{V}^T - \lambda_{2,k}^{-1}\mathbf{S}_k)\mathbf{N}$ and \mathbf{z}_1 and \mathbf{z}_2 are the columns of \mathbf{Z} . Since \mathbf{Z} is orthogonal, we can parameterize it as

$$\mathbf{Z} = \begin{pmatrix} \cos \phi & s \sin \phi \\ \sin \phi & -s \cos \phi \end{pmatrix}$$

for some $\phi \in (0, 2\pi)$ and $s = \pm 1$. The uniform density on the circle is constant in ϕ , so the joint density of (ϕ, s) is simply $p(\mathbf{Z}(\phi, s))$. Sampling from this distribution can be accomplished by first sampling $\phi \in (0, 2\pi)$ from a density proportional to

$$p(\phi) \propto \exp([g_{1,1} + h_{2,2}] \cos^2 \phi + [h_{1,1} + g_{2,2}] \sin^2 \phi + [g_{1,2} + g_{2,1} - h_{1,2} - h_{2,1}] \cos \phi \sin \phi),$$

and then sampling s uniformly from $\{-1, +1\}$.

3.3 Estimation of eigenvalues

From (9) we see that the conditional distribution of Λ_k given \mathbf{U}_k and \mathbf{S}_k has the following form:

$$\begin{aligned} p(\Lambda_k | \mathbf{U}_k, \mathbf{S}_k) &\propto p(\Lambda_k) |\Lambda_k|^{-(n_k-1)/2} \text{etr}(-\frac{1}{2}\Lambda_k^{-1}\mathbf{U}_k^T\mathbf{S}_k\mathbf{U}_k) \\ &= p(\Lambda_k) \prod_{j=1}^p \lambda_{j,k}^{-(n_k-1)/2} \exp\{-\frac{1}{2}\sum_{j=1}^p \lambda_{j,k}^{-1} \mathbf{u}_{j,k}^T \mathbf{S}_k \mathbf{u}_{j,k}\} \end{aligned}$$

The part not involving the prior distribution has the form of an inverse-gamma density, and indeed, if $p(\Lambda_k)$ were the product of inverse-gamma densities with parameters $(\nu_0/2, \nu_0\sigma_0^2/2)$ then the full conditional distribution of $\lambda_{j,k}$ would be inverse-gamma $[(\nu_0 + n - 1)/2, (\nu_0\sigma_0^2 + \mathbf{u}_{j,k}^T \mathbf{S}_k \mathbf{u}_{j,k})/2]$. However, it may be desirable to add more structure to the estimation of the eigenvalues. In usual one-sample principal component analysis the eigenvalues are labeled in order of decreasing magnitude and attention is focused on the “first few” eigenvectors, i.e. those corresponding to the largest eigenvalues. In terms of making comparisons of eigenvectors across groups, restricting the eigenvalues to be ordered means that the ordered columns of \mathbf{V} refer to the ordered columns of \mathbf{U} . One concern about such a restriction would be how it might affect inference in the case of a shared eigenvector that is the first principal axes in some groups, and possibly the second or third in other groups. This sort of heterogeneity can in fact be represented with the generalized Bingham distribution even if the eigenvalues are order-restricted. For example, the distribution with $\text{diag}(\mathbf{A}) = (a, 0, 0, \dots)$ and $\text{diag}(\mathbf{B}) = (b, b, 0, \dots)$ represents a population in which, with equal frequency, one of the first two columns of \mathbf{U} is near the first column of \mathbf{V} . Because of this flexibility, in what follows we estimate the eigenvalues in each group as being ordered. A convenient prior distribution is that $p(\Lambda_k)$ is the product of inverse-gamma densities described above, but restricted to the space $\lambda_{1,k} > \lambda_{2,k} > \dots > \lambda_{p,k}$. The full conditional distribution of $\lambda_{j,k}$ is then inverse-gamma $[(\nu_0 + n - 1)/2, (\nu_0\sigma_0^2 + \mathbf{u}_{j,k}^T \mathbf{S}_k \mathbf{u}_{j,k})/2]$ but restricted to the interval $(\lambda_{j+1,k}, \lambda_{j-1,k})$.

3.4 Summary of MCMC algorithm

The unknown parameters in the hierarchical model are the group-specific eigenvectors and values $\{\mathbf{U}_1, \Lambda_1\}, \dots, \{\mathbf{U}_K, \Lambda_K\}$ and the parameters $\{\mathbf{A}, \mathbf{B}, \mathbf{V}\}$ describing the across-group heterogeneity of eigenvector matrices. The diagonal matrices \mathbf{A} and \mathbf{B} are parameterized as

$$\begin{aligned} \text{diag}(\mathbf{A}) &= (a_1, \dots, a_p) = \sqrt{w}(\alpha_1, \dots, \alpha_p) \\ \text{diag}(\mathbf{B}) &= (b_1, \dots, b_p) = \sqrt{w}(\beta_1, \dots, \beta_p) \end{aligned}$$

with $1 = \alpha_1 > \dots > \alpha_p = 0$ and $1 = \beta_1 > \dots > \beta_p = 0$. Convenient prior distributions are $\mathbf{V} \sim$ uniform \mathcal{O}_p , $(\alpha_2, \dots, \alpha_{p-1})$ and $(\beta_2, \dots, \beta_{p-1})$ are uniform on $[0, 1]$ subject to the ordering restriction, $w \sim$ gamma $(\eta_0/2, \tau_0^2/2)$ and $(1/\lambda_{1,k}, \dots, 1/\lambda_{p,k})$ are the order statistics of a sample from a gamma $(\nu_0/2, \sigma_0^2/2)$ distribution. With these prior distributions, a Markov chain in the unknown parameters that converges to the posterior distribution $p(\{\mathbf{U}_1, \Lambda_1\}, \dots, \{\mathbf{U}_K, \Lambda_K\}, \mathbf{A}, \mathbf{B}, \mathbf{V} | \mathbf{Y}_1, \dots, \mathbf{Y}_K)$ can be constructed by iteration of the following sampling scheme:

1. Update the within-group parameters:
 - (a) Update $\{\mathbf{U}_1, \dots, \mathbf{U}_K\}$: For each k and a randomly selected pair $\{j_1, j_2\} \subset \{1, \dots, p\}$;
 - i. let \mathbf{N} be the null space of the columns $\{\mathbf{u}_{j,k} : j \notin \{j_1, j_2\}\}$;

- ii. compute $\mathbf{G} = \mathbf{N}^T (b_{j_1} \mathbf{V} \mathbf{A} \mathbf{V}^T - \lambda_{j_1, k}^{-1} \mathbf{S}_k) \mathbf{N}$, $\mathbf{H} = \mathbf{N}^T (b_{j_2} \mathbf{V} \mathbf{A} \mathbf{V}^T - \lambda_{j_2, k}^{-1} \mathbf{S}_k) \mathbf{N}$;
 - iii. sample $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2) \in \mathcal{O}_2$ from the density proportional to $\exp(\mathbf{z}_1^T \mathbf{G} \mathbf{z}_1 + \mathbf{z}_2^T \mathbf{H} \mathbf{z}_2)$
 - iv. set $\mathbf{u}_{j_1, k}$ to be the first column of $\mathbf{N} \mathbf{Z}$ and $\mathbf{u}_{j_2, k}$ to be the second.
- (b) Update $\{\Lambda_1, \dots, \Lambda_K\}$: Iteratively for each $j \in \{1, \dots, p\}$ and $k \in \{1, \dots, K\}$, sample $\lambda_{j, k} \sim \text{inverse-gamma}[(\nu_0 + n_k - 1)/2, (\nu_0 \sigma_0^2 + \mathbf{u}_{j, k}^T \mathbf{S}_k \mathbf{u}_{j, k})/2]$, but constrained to be in $(\lambda_{j-1, k}, \lambda_{j+1, k})$.

2. Update the across-group parameters:

- (a) Update \mathbf{V} : Sample \mathbf{V} from the Bingham density proportional to $\text{etr}(\mathbf{A} \mathbf{V}^T [\sum \mathbf{U}_k \mathbf{B} \mathbf{U}_k^T] \mathbf{V})$.
- (b) Update w , $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$: Compute $\mathbf{M} = \sum_{k=1}^K (\mathbf{V}^T \mathbf{U}_k) \circ (\mathbf{V}^T \mathbf{U}_k)$ and
 - i. sample $w \sim \text{gamma}(\eta_0/2 + \binom{p}{2} K/2, \eta_0 \tau_0^2/2 + \boldsymbol{\alpha}^T [K \mathbf{I} - \mathbf{M}] \boldsymbol{\beta})$
 - ii. for each $i \in \{2, \dots, p-1\}$ sample $\alpha_i \in (\alpha_{i-1}, \alpha_{i+1})$ from the density proportional to $\exp\{-\alpha_i (w \boldsymbol{\beta}^T \mathbf{M}_{[i, \cdot]})\} \prod_{j: j \neq i} |a_i - a_j|^{K/2}$.
 - iii. for each $j \in \{2, \dots, p-1\}$ sample $\beta_j \in (\beta_{j-1}, \beta_{j+1})$ from the density proportional to $\exp\{-\beta_j (w \mathbf{M}_{[\cdot, j]}^T \boldsymbol{\alpha})\} \prod_{i: i \neq j} |\beta_j - \beta_i|^{K/2}$.

As discussed in section 3.1, it may be desirable to make Metropolis-Hastings adjustments to the steps in 2(b) to account for the approximation to the normalizing constant $c(\mathbf{A}, \mathbf{B})$. Functions and example code for this algorithm, written in the R programming environment, are available at my website, <http://www.stat.washington.edu/hoff/>.

4 Example: Vole measurements

Flury [1987] describes an analysis of skull measurements on four different groups of voles. The four groups, defined by species and sex, are male and female *Microtus californicus* and male and female *Microtus ochrogaster*, having sample sizes of 82, 70, 58 and 54 respectively. Flury provides the sample covariance matrices of four log-transformed measurements corresponding to skull length, toothrow length, cheekbone width and interorbital width. The eigenvectors of the four empirical covariance matrices are given in Table 1. The first eigenvector in each group can roughly be interpreted as measuring overall size variation, and its values seem fairly similar across groups. The remaining eigenvectors also display a high degree of similarity across groups. By performing statistical tests of various hypotheses regarding the four population covariance matrices, Flury concludes that although the sample covariance matrices appear similar, there is enough evidence to reject exact equality of the population matrices. Furthermore, Flury rejects a hypotheses that the population covariances are proportional to each other, and then suggests a model in which the the covariance matrices share a single eigenvector (interpreted as corresponding to size), with

| <i>M. californicus</i> | | | | | | | |
|------------------------|-------|-------|-------|---------|-------|-------|-------|
| males | | | | females | | | |
| 36.31 | 27.01 | 8.05 | 2.78 | 52.44 | 21.14 | 3.75 | 3.17 |
| 0.49 | -0.31 | -0.19 | 0.79 | 0.53 | -0.31 | -0.29 | 0.74 |
| 0.60 | -0.10 | 0.76 | -0.24 | 0.56 | -0.06 | 0.82 | -0.11 |
| 0.55 | -0.12 | -0.63 | -0.54 | 0.57 | -0.14 | -0.48 | -0.65 |
| 0.30 | 0.94 | -0.06 | 0.17 | 0.30 | 0.94 | -0.11 | 0.14 |
| <i>M. ochrogaster</i> | | | | | | | |
| males | | | | females | | | |
| 36.30 | 9.67 | 7.97 | 2.80 | 35.61 | 12.35 | 8.32 | 3.38 |
| 0.58 | 0.04 | -0.38 | 0.71 | 0.56 | 0.06 | 0.05 | 0.82 |
| 0.45 | 0.72 | 0.51 | -0.13 | 0.47 | 0.02 | 0.80 | -0.38 |
| 0.51 | -0.08 | -0.51 | -0.69 | 0.66 | -0.25 | -0.58 | -0.40 |
| 0.45 | -0.68 | 0.58 | -0.02 | 0.13 | 0.97 | -0.17 | -0.14 |

Table 1: Eigenvalues and eigenvectors of empirical covariance matrices. Eigenvalues are given in the first row for each group.

the remaining eigenvectors and all of the eigenvalues being distinct across groups. In this section we reanalyze these data using the hierarchical eigenmodel discussed above, and compare it to the model in Flury [1987] and a few others. In particular, we show that allowing information-sharing across the groups where appropriate, but not forcing any of the eigenvectors to be exactly equal, results in a model that better represents features of the observed dataset.

Using the sample sizes and sample covariance matrices provided in Flury [1987], centered versions of $\mathbf{Y}_k^T \mathbf{Y}_k$ for each group $k \in \{1, \dots, 4\}$ were reconstructed and used as data for the model described in Section 3. The prior distribution of w was taken to be a diffuse exponential with a mean of 1000, and the prior distribution for the inverse-eigenvalues was exponential with a mean of 1. A Markov chain consisting of 10,000 iterations was constructed, for which parameter values were saved every 10th iteration giving a total of 1000 posterior samples for each parameter. Mixing of the Markov chains was monitored via a variety of parameter summaries computed at each saved iteration. For example, for each saved value of \mathbf{A} and \mathbf{B} the average and standard deviation of the logs of the $(p-1) \times (p-1) = 9$ non-zero values of $\mathbf{A} \circ \mathbf{B}$ were obtained and plotted sequentially in the first panel of Figure 2.

A posterior point estimate of \mathbf{V} can be obtained from the eigenvector matrix of the posterior mean of \mathbf{VAV}^T , obtained by averaging across samples of the Markov chain. This produces the

| hierarchical model estimate | | | | empirical estimate | | | |
|-----------------------------|-------|-------|-------|--------------------|-------|-------|-------|
| 0.54 | -0.27 | -0.19 | 0.77 | 0.55 | -0.25 | -0.17 | 0.78 |
| 0.54 | -0.10 | 0.80 | -0.22 | 0.53 | -0.10 | 0.81 | -0.23 |
| 0.56 | -0.15 | -0.56 | -0.59 | 0.57 | -0.16 | -0.56 | -0.57 |
| 0.30 | 0.95 | -0.06 | 0.10 | 0.30 | 0.95 | -0.05 | 0.08 |

Table 2: Model-based and empirical estimates of \mathbf{V} .

matrix in Table 2, which is nearly identical to the eigenvector matrix of the pooled covariance matrix $\sum \mathbf{Y}_k^T \mathbf{Y}_k / (n_k - 1)$, which is also given in the table. Posterior mean estimates of the eigenvalue matrices $\{\Lambda_1, \dots, \Lambda_K\}$ were all within 1.0 of their corresponding values based on the the empirical covariance matrices.

Table 1 suggests that the first and fourth eigenvectors are the most preserved across groups, whereas the other two are less well-preserved. Letting $\hat{\mathbf{U}}_k$ be the eigenvector matrix of $\mathbf{Y}_k^T \mathbf{Y}_k$ and $\hat{\mathbf{V}}$ the eigenvector matrix of $\sum \mathbf{Y}_k^T \mathbf{Y}_k / (n_k - 1)$, the differential heterogeneity of the eigenvectors can be described numerically by computing the value of $\text{diag}(\hat{\mathbf{V}}^T \hat{\mathbf{U}}_k)^2$ and averaging each of the p entries of this vector across the K groups. This p -dimensional function of the observed data gives $t(\mathbf{Y}_1^T \mathbf{Y}_1, \dots, \mathbf{Y}_4^T \mathbf{Y}_4) = (0.98, 0.85, 0.86, 0.96)$, indicating that by this metric the first and last eigenvectors are most preserved across groups. To examine how well the model represents this observed heterogeneity, the value of $t(\mathbf{Y}_1^T \mathbf{Y}_1, \dots, \mathbf{Y}_4^T \mathbf{Y}_4)$ can be compared to its posterior predictive distribution under the model. This was done by generating simulated values $\tilde{\mathbf{Y}}_1^T \tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_4^T \tilde{\mathbf{Y}}_4$ every 10th iteration of the Markov chain and computing the statistic $t()$ described above, resulting in 1000 samples of $t(\tilde{\mathbf{Y}}_1^T \tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_4^T \tilde{\mathbf{Y}}_4)$ from the posterior predictive distribution. For simplicity we present below only the minimum and maximum values of this statistic, which for our observed data are (0.85, 0.98).

The top of the second panel of Figure 2 shows the posterior predictive distributions of this statistic on a logit scale under the hierarchical model which pools information across eigenvector matrices. The observed values are well within the predicted range, indicating that the model is able to represent the differential amounts of eigenvector preservation among the observed covariance matrices. In contrast, the lower part of the plot shows a posterior predictive distribution generated under a one-shared-eigenvector model similar to the one in Flury [1987], obtained obtained by fixing $w = 1000$, $\alpha_1 = \beta_1 = 1$ and $\alpha_j = \beta_j = 0$ for $j > 1$. This model accurately predicts the highest degree of preservation across eigenvectors, but underestimates the preservation among other eigenvectors. This is not surprising, as this model shares information only across a single eigenvector.

Lastly we fit two other related models: a “no pooling” model in which no information was shared

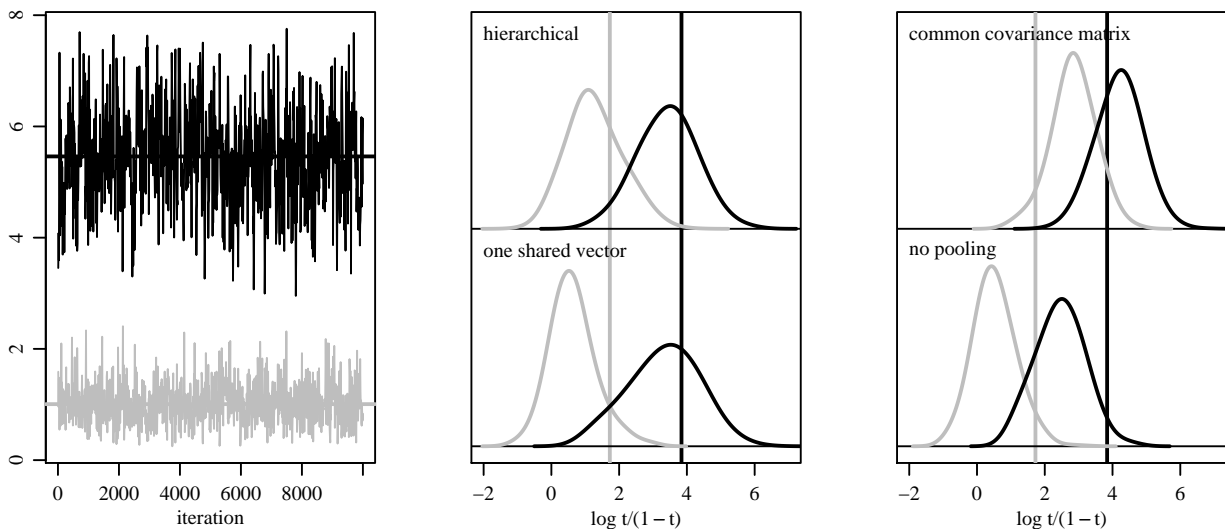


Figure 2: MCMC and model diagnostics for the Vole example. The first panel shows averages (black) and standard deviations (gray) of the log entries of $\mathbf{A} \circ \mathbf{B}$ at every 10th scan of the Markov chain. The second and third panels show posterior predictive distributions of the minimum and maximum similarity statistics under a variety of models, with the observed value of each statistic represented by a vertical line.

across groups and a common covariance matrix model in which it is assumed that the covariances are exactly identical across groups. Not surprisingly these two models under- and over-represent the similarity across eigenvectors of observed covariance matrices, as shown in the third panel of Figure 2. Taken together, these results indicate that assuming complete equality, or completely ignoring similarity, can misrepresent the variability of covariance structure across groups.

5 Example: National Health Communication Study

The 2005 Annenberg National Health Communication Survey (anhcs.asc.upenn.edu) gathered self-reported health and lifestyle data from 2,989 members of the adult U.S. population under the age of 65. Among the variables recorded were the following:

state: state of residency (including the District of Columbia)
fruitveg: typical number of servings of fruit and vegetables per day
exercise: typical weekly frequency of exercise
bmi: body mass index
alcohol: number of days in the month consuming five or more alcoholic drinks
smoke: typical number of cigarettes smoked per day
age: in years
female: indicator of being female
income: household income (19 ordered categories)
edu: education level (no degree, high school, some college, Bachelor’s degree or higher)

In this section we estimate state-specific correlation matrices in a Gaussian copula model for the $p = 9$ ordinal variables above. More specifically, we model the observed data vectors $\mathbf{y}_1, \dots, \mathbf{y}_n$ within a particular state as monotone functions of latent Gaussian random variables, so that

$$\begin{aligned}
 \mathbf{z}_1, \dots, \mathbf{z}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \Sigma) \\
 y_{i,j} &= g_j(z_{i,j}).
 \end{aligned}$$

The non-decreasing functions $\{g_1, \dots, g_p\}$ are state-specific as is the covariance matrix Σ . We compare two models for $\Sigma_1, \dots, \Sigma_K$, the first being one in which no information is shared and $\mathbf{U}_1, \dots, \mathbf{U}_K$ are *a priori* independent and uniformly distributed on \mathcal{O}_p . The second is the hierarchical eigenmodel in which $\mathbf{U}_1, \dots, \mathbf{U}_K \sim \text{i.i.d. } p_B(\mathbf{U}|\mathbf{A}, \mathbf{B}, \mathbf{V})$, with the parameters $\{\mathbf{A}, \mathbf{B}, \mathbf{V}\}$ unknown and estimated from the data, using the same prior distributions as in the previous section. For both models, the prior distribution on the eigenvalues in each group is such that $\{1/\lambda_1 < \dots < 1/\lambda_p\}$ are the order statistics of p independent exponential(1) random variables. We note that both of these models ignore the possibility that heterogeneity in correlation matrices might be associated with state-specific characteristics such as population size or geographic location (although some ad-hoc exploratory analyses suggest these effects are small).

Parameter estimation for this hierarchical copula model can be accomplished by iterative sampling of the parameters from their full conditional distributions as in Section 3 with the latent \mathbf{z} ’s taking the roles of the observed \mathbf{y} ’s, along with iterative sampling of the \mathbf{z} ’s from their full conditional distributions (which are constrained normal distributions). This latter step is described for a one-group discrete-data copula model in Hoff [2007a]. We note that this is a type of parameter-expanded estimation scheme [Gelman et al., 2008] in that the scale of each variable j can be represented by both g_j and $\Sigma_{j,j}$, and so these quantities are not separately identifiable. However, the posterior distribution of $\{\Sigma_1, \dots, \Sigma_K\}$ induces a posterior distribution over state-specific correlation matrices $\{\mathbf{C}_1, \dots, \mathbf{C}_K\}$, which are the parameters of primary interest in copula models. In the posterior analysis that follows we focus mostly on comparing the hierarchical and

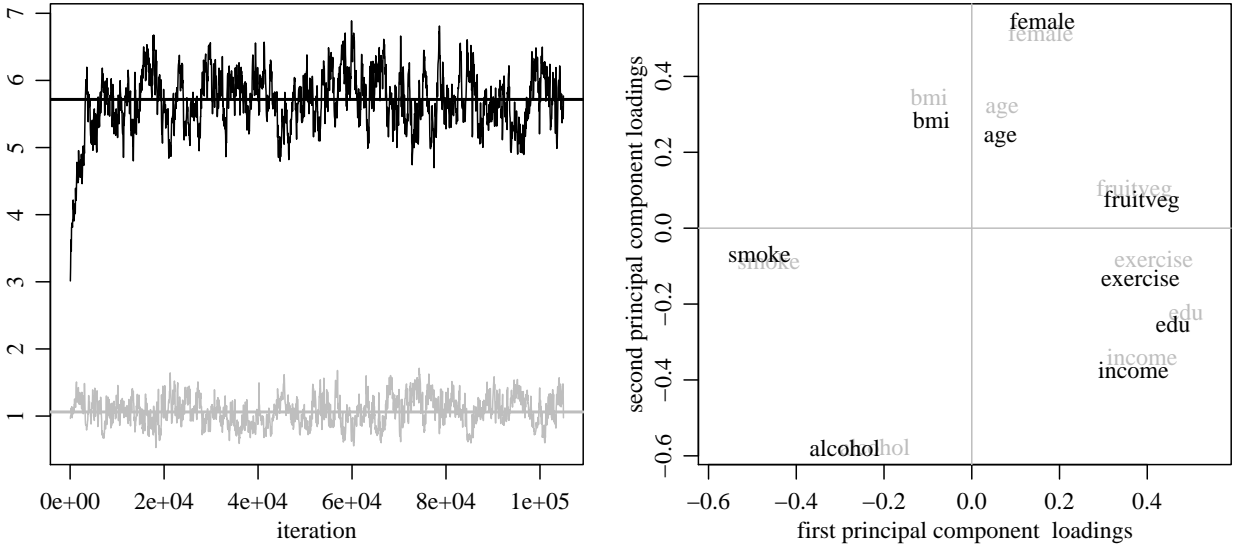


Figure 3: The first panel shows averages (black) and standard deviations (gray) of the log entries of $\mathbf{A} \circ \mathbf{B}$. The second panel shows the first two principal component loadings for the hierarchical (black) and non-hierarchical (gray) models.

non-hierarchical posterior mean estimates of the state-specific correlation matrices.

Markov chains consisting of 105,000 iterations were constructed for each of the two models, with results from the first 5000 iterations being discarded to allow for burn-in. Parameter values from the remaining iterations were saved every 50th iteration, leaving 2000 Monte Carlo samples for approximating the posterior distributions. The correlation parameters mixed reasonably well: In the hierarchical model, the effective sample sizes for 90% of the parameters was greater than 500, and for 50% it was greater than 1200 (effective sample size is an estimate of the number of independent samples required to estimate the mean to the same precision as with a given auto-correlated sample). Mixing of the hierarchical parameters was slower: The first panel of Figure 5 plots the mean and standard deviation of the logs of the 64 non-zero values of the matrix $\mathbf{A} \circ \mathbf{B}$ at every 50th scan of the Markov chain for the hierarchical model. The effective sample sizes for these two functions of the parameters were both just over 100.

The second panel of Figure 5 plots the first two eigenvectors of the posterior mean of the state-averaged correlation matrix $\sum_{k=1}^K \mathbf{C}_k / K$ for each of the two models. The results are quite similar, indicating that the main correlations across states are described by smoking and drinking behavior being negatively correlated with education level, income, fruit and vegetable intake and exercise. In terms of state-specific correlation matrices however, the two models produce quite different results: The top and bottom plots of Figure 5 give posterior mean estimates of state-specific correlations from the non-hierarchical and hierarchical models, respectively. For each pair of variables, a boxplot

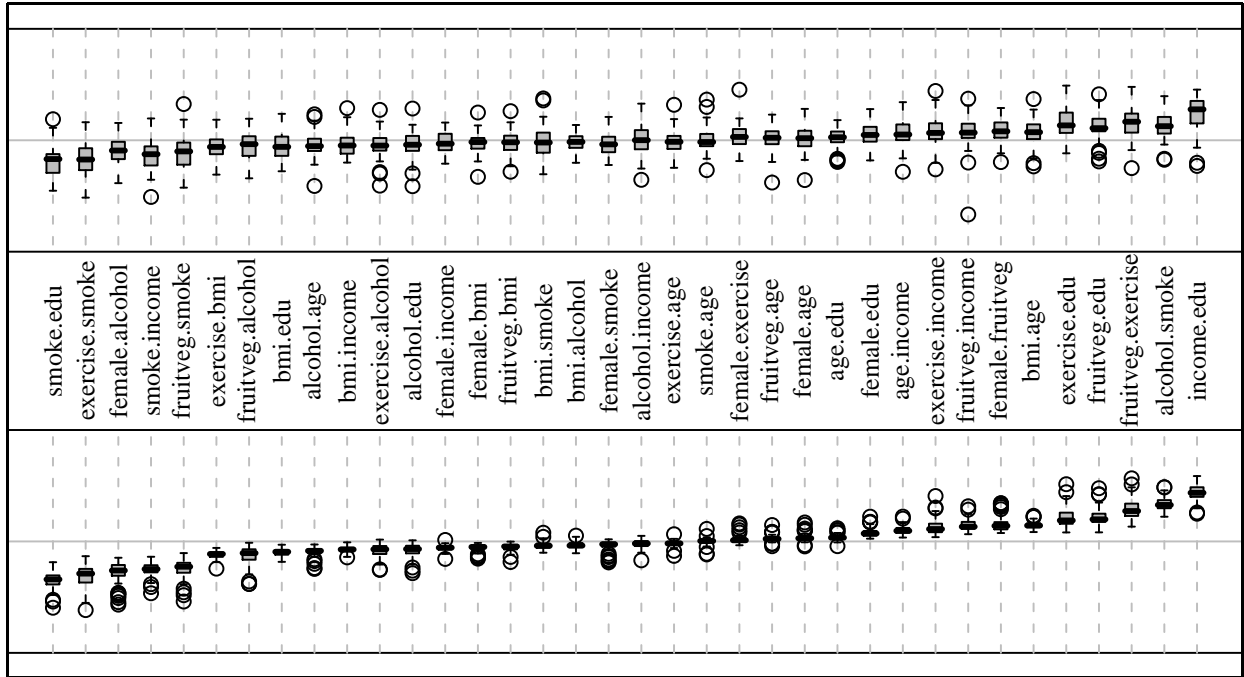


Figure 4: Posterior mean estimates of state-specific correlations. The top row are from the non-hierarchical model, the bottom from the hierarchical eigenmodel.

indicates the across-state heterogeneity among the $K = 51$ parameter estimates for each correlation coefficient. The number of observations per state varies widely, with North Dakota and the District of Columbia having 3 respondents each, whereas Texas and California have 225 and 312 respondents respectively. Generally speaking, the estimates for states with lower sample sizes appear at the extremes of the boxplots, which is not surprising as these estimates have a higher degree of sampling variability. Also of note is the fact that there is no coefficient that is estimated as either positive across all states or negative across all states. For example, among the highest correlations is that between income and education, with an across-state median estimate of 0.28 based on the non-hierarchical model. However, there were four states (VT, AK, WY, NE) which were estimated as having a negative correlation between these variables. The sample sizes from these states were 5, 9, 5 and 15 respectively, suggesting that these low correlations may be due to unrepresentative samples. In contrast, the hierarchical model recognizes that much of the across-state heterogeneity in correlation estimates may be due to sampling variability, and shrinks estimates from low-sample size states towards the across-state center. For example, the hierarchical model gives positive point estimates for the correlation between income and education for all of the states, including VT, AK, WY and NE. As shown in the lower half of Figure 5, across-state heterogeneity among the other correlation coefficients is similarly reduced, with nearly two-thirds (23 of 36) of the correlation coefficients having sign-consistent estimates across all 51 states.

The effects of hierarchical estimation are explored further in Figure 5. We have two estimates of the eigenvectors for each of the k state-specific correlation matrices: $\hat{\mathbf{U}}_k$ from the hierarchical model and $\check{\mathbf{U}}_k$ from the non-hierarchical model. We can compute a similarity between these two matrices as the average of the p values of $\text{diag}(\check{\mathbf{U}}_k^T \hat{\mathbf{U}}_k)^2$. The first panel of Figure 5 shows that the relationship between the similarity and the within-state sample size is positive as expected: Covariance matrices for states with large sample sizes are well-estimated based on within-state data alone, and their eigenvector estimates are largely unaffected by hierarchical estimation. In contrast, the amount of information from states with low sample sizes is small, and so the estimates for the hierarchical model are pulled towards the population mode and away from $\check{\mathbf{U}}_k$. The effects of this shrinkage on the principal axes of the correlation matrices are shown graphically in the second and third panels of Figure 5. The second panel plots the projections of the first two columns of each $\check{\mathbf{U}}_k$ onto the first two columns of the eigenvector matrix of the pooled correlation matrix. Although heterogeneous, the vectors are generally in the same direction, and further inspection shows that outliers tend to be states with low sample sizes. The third panel of the figure shows the same plot for the projections of the columns of each $\hat{\mathbf{U}}_k$ from the hierarchical model. The heterogeneity here represents the estimated across-state variability in eigenvectors after accounting for the within-state sampling variability. The axis in this plot that is furthest from the center is that representing Wisconsin, which has relatively high sample size of 69 but some extreme correlations: For example, among

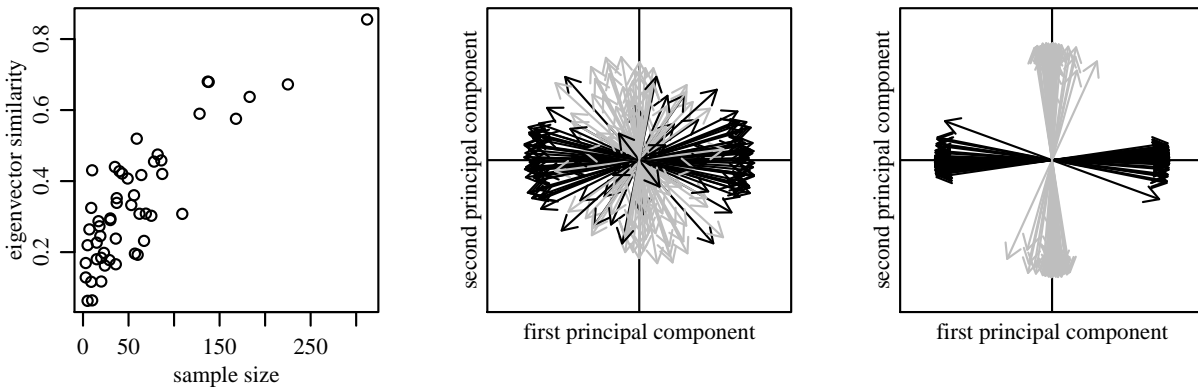


Figure 5: Effects of shrinkage on the estimated principal axes. The first panel shows the similarity between non-hierarchical and hierarchical estimates as a function of sample size. The second and third show heterogeneity across the first two principal axes in the non-hierarchical and hierarchical models, respectively.

states with sample sizes greater than 20, Wisconsin has the lowest non-hierarchical estimate of the correlation for (`income`, `education`) and (`female`, `bmi`), and the highest non-hierarchical estimate of the correlation for (`income`, `alcohol`). These correlations make Wisconsin somewhat of an outlier in terms of the correlations represented by the first two principal components. The relatively large sample size for Wisconsin suggests these extreme correlations cannot be solely attributed to within-state sampling variability, and this is reflected in the state-specific estimated correlation matrix from the hierarchical model.

6 Discussion

As an alternative to the Bingham distribution, a simpler model for across-group covariance heterogeneity would be that $\Sigma_1, \dots, \Sigma_K$ are i.i.d. samples from an inverse-Wishart distribution. For many applications however, such a model may be too simple: The inverse-Wishart distribution has only one parameter to represent heterogeneity around the mean covariance matrix, and cannot represent differential amounts of eigenvector heterogeneity as the Bingham distribution can. Additionally, the inverse-Wishart cannot distinguish between across-group eigenvector heterogeneity and across-group eigenvalue heterogeneity, as these quantities are modeled simultaneously.

As we are pooling eigenvector information across groups it is natural to consider pooling eigenvalues as well. This would entail modeling $\{\Lambda_1, \dots, \Lambda_K\}$ as being samples from a common population, and estimating the parameters of this population using the data from all K groups. One simple approach to doing this would be to estimate the parameters (ν_0, σ_0^2) in the prior distribu-

tion for the eigenvalues, thus treating the distribution as a sampling model. As with the other unknown parameters, this can be done by iteratively updating these parameters based on their full conditional distributions. Straightforward calculations show that a gamma prior distribution for σ_0^2 results in a gamma full conditional distribution. The full conditional distribution for ν_0 is non-standard, but if ν_0 is restricted to the integers then its full conditional distribution can easily be sampled from.

Another possible model extension is to situations where the number of variables is larger than any of the within-group sample sizes. In these cases, full-rank covariance estimation can become unstable and computationally intractable. A remedy to this problem is to use a factor analysis model, in which a covariance matrix Σ_k is assumed equal to $\mathbf{U}_k \mathbf{D}_k \mathbf{U}_k^T + \sigma_k^2 \mathbf{I}$, where \mathbf{D}_k is a positive diagonal matrix and \mathbf{U}_k is a $p \times r$ orthonormal matrix with $r < p$, an element of the Stiefel manifold $\mathcal{S}_{r,p}$. As before, heterogeneity across covariance matrices can be expressed by heterogeneity in these matrix components, and a Bingham model on $\mathcal{S}_{r,p}$, similar to the one used in this paper, can be used to express heterogeneity among $\mathbf{U}_1, \dots, \mathbf{U}_K$.

Computer code and data for the examples in this article are available at

<http://www.stat.washington.edu/hoff/>.

References

- George A. Anderson. An asymptotic expansion for the distribution of the latent roots of the estimated covariance matrix. *Ann. Math. Statist.*, 36:1153–1173, 1965. ISSN 0003-4851.
- Christopher Bingham. An antipodally symmetric distribution on the sphere. *Ann. Statist.*, 2:1201–1225, 1974. ISSN 0090-5364.
- Robert J. Boik. Spectral models for covariance matrices. *Biometrika*, 89(1):159–182, 2002. ISSN 0006-3444.
- A. G. Constantine and R. J. Muirhead. Asymptotic expansions for distributions of latent roots in multivariate analysis. *J. Multivariate Anal.*, 6(3):369–391, 1976. ISSN 0047-259x.
- Bernhard K. Flury. Two generalizations of the common principal component model. *Biometrika*, 74(1):59–69, 1987. ISSN 0006-3444.
- Bernhard N. Flury. Common principal components in k groups. *J. Amer. Statist. Assoc.*, 79(388):892–898, 1984. ISSN 0162-1459.
- Andrew Gelman, David A. van Dyk, Zaiying Huang, and John W. Boscardin. Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1):95–122, March 2008.

- A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*. Chapman & Hall/CRC, Boca Raton, FL, 2000. ISBN 1-58488-046-5.
- Carl S. Herz. Bessel functions of matrix argument. *Ann. of Math. (2)*, 61:474–523, 1955. ISSN 0003-486X.
- Peter D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Statist.*, 1(1):265–283, 2007a.
- Peter D. Hoff. Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. Technical report, University of Washington, 2007b.
- C. G. Khatri and K. V. Mardia. The von Mises-Fisher matrix distribution in orientation statistics. *J. Roy. Statist. Soc. Ser. B*, 39(1):95–106, 1977. ISSN 0035-9246.
- Plamen Koev and Alan Edelman. The efficient evaluation of the hypergeometric function of a matrix argument. *Math. Comp.*, 75(254):833–846 (electronic), 2006. ISSN 0025-5718.
- Robb J. Muirhead. Latent roots and matrix variates: a review of some asymptotic results. *Ann. Statist.*, 6(1):5–33, 1978. ISSN 0090-5364.
- James R. Schott. Some tests for common principal component subspaces in several groups. *Biometrika*, 78(4):771–777, 1991. ISSN 0006-3444.
- James R. Schott. Partial common principal component subspaces. *Biometrika*, 86(4):899–908, 1999. ISSN 0006-3444.