Covariance Reducing Models: An Alternative to Spectral Modeling of Covariance Matrices

By R. DENNIS COOK

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

dennis@stat.umn.edu

and LILIANA FORZANI

Facultad de Ingeniería Química, Universidad Nacional del Litoral and Instituto Matemática Aplicada Litoral, CONICET, Santa Fe, Argentina.

liliana.forzani@gmail.com

Summary

We introduce covariance reducing models for studying the sample covariance matrices of a random vector observed in different populations. The models are based on reducing the sample covariance matrices to an informational core that is sufficient to characterize the variance heterogeneity among the populations. They possess useful equivariance properties and provide a clear alternative to spectral models for covariance matrices.

Some key words: Central subspace, Dimension reduction, Envelopes, Grassmann manifolds, Reducing subspaces.

1. Introduction

We consider the problem of characterizing the behavior of positive definite covariance matrices $\Sigma_g = \text{cov}(X|g)$, $g = 1, \ldots, h$, of a random vector $X \in \mathbb{R}^p$ observed in each of h

populations identified by the index g. Testing for equality or proportionality (Muirhead, 1982, Ch. 8; Flury, 1988, Ch. 5; Jensen & Madsen, 2004) may be useful first steps, but lacking such a relatively simple characterization there arises a need for more flexible methodology. Perhaps the most well-known methods for studying covariance matrices stem from Flury's (1987) spectral model of partial common principal components,

$$\Sigma_g = \Gamma \Lambda_{1,g} \Gamma^T + \Gamma_g \Lambda_{2,g} \Gamma_g^T, \tag{1}$$

where $\Lambda_{1,g} > 0$ and $\Lambda_{2,g} > 0$ are diagonal matrices and (Γ, Γ_g) is an orthogonal matrix with $\Gamma \in \mathbb{R}^{p \times q}$, $q \leq p-1$, $g = 1, \ldots, h$. The linear combinations $\Gamma^T X$ are then the q principal components that are common to all populations. This model reduces to Flury's (1984) common principal component model when $q = p - 1$.

Situations can arise where the Σ_g's have no common eigenvectors, but have cardinality equal sets of eigenvectors that span the same subspace. This possibility is covered by subspaces models. Flury's (1987) common space models do not require the eigenvector sets to have the largest eigenvalues, while the common principal component subspace models studied by Schott (1991) do have this requirement. Schott's rationale was to find a method for reducing dimensionality while preserving variability in each of the h populations. Schott (1999, 2003) developed an extension to partial common principal component subspaces that targets the sum of the subspaces spanned by the first few eigenvector of the Σ_g's. Boik (2002) proposed a comprehensive spectral model for covariance matrices that allows the Σ_g's to share multiple eigenspaces without sharing eigenvectors and permits sets of homogeneous eigenvalues.

Houle, Mezey & Galpern (2002; see also Mezey & Houle, 2003) considered the suitability

of spectral methods for studying covariance matrices that arise in evolutionary biology. They concluded that Flury's principal component models perform as might be expected from a statistical perspective, but they were not encouraging about their merits as an aid to evolutionary studies. Judging from their simulations, their misgivings may stem in part from the fact that spectral methods are not generally invariant or equivariant: For a nonsingular matrix $A \in \mathbb{R}^{p\times p}$, the transformation $\Sigma_g \to A\Sigma_g A^T$ can result in new spectral decompositions that are not usefully linked to the original decompositions. For example, common principal components may not be the same or of the same cardinality after transformation.

We propose in §1 a class of new covariance reducing models as an alternative to spectral models for studying a collection of covariance matrices. Their relationship with some spectral models is discussed in §2·3. Estimation is considered in §3. Inference methods for an underlying dimension and for contributing variables are considered in §§5 and 6. §7 contains illustrations of how the proposed methodology might be employed in practice. Proofs of key results are given in the appendices.

The following notation will be used in our exposition. For positive integers p and q, $\mathbb{R}^{p\times q}$ stands for the class of real matrices of dimension $p \times q$, and $\mathbb{S}^{p\times p}$ denotes the class of symmetric $p \times p$ positive definite matrices. For $A \in \mathbb{R}^{p\times p}$ and a vector subspace $\mathcal{S} \subseteq \mathbb{R}^p$, $A\mathcal{S} \equiv \{A\mathbf{x} : \mathbf{x} \in \mathcal{S}\}$. A basis matrix for a subspace $\mathcal{S}$ is any semi-orthogonal matrix whose columns are a basis for $\mathcal{S}$. For a semi-orthogonal matrix $A \in \mathbb{R}^{p\times q}$, $q \le p$, the matrix A_0 denotes any completion of A so that $(A, A_0) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix. For $B \in \mathbb{R}^{p\times q}$, $\mathcal{S}_B \equiv \text{span}(B)$ denotes the subspace of $\mathbb{R}^p$ spanned by the columns of B. If $B \in \mathbb{R}^{p\times q}$ with rank q and $\Sigma \in \mathbb{S}^{p\times p}$, then the projection onto $\mathcal{S}_B$ relative to Σ has

the matrix representation $P_{B(\Sigma)} \equiv B(B^T\Sigma B)^{-1}B^T\Sigma$. $P_{\mathcal{S}}$ indicates the projection onto the subspace $\mathcal{S}$ in the usual inner product. The orthogonal complement $\mathcal{S}^{\perp}$ of a subspace $\mathcal{S}$ is constructed with respect to the usual inner product, unless indicated otherwise. To describe the distribution of a normal matrix $Z \in \mathbb{R}^{p\times q}$, we follow Muirhead (1982, p. 79) and use the notation $Z \sim N(M, V)$ to mean $\text{vec}(Z^T) \sim N\{\text{vec}(M^T), V\}$, where "vec" is the operator that maps a matrix to a vector by stacking its columns. The product of the non-zero eigenvalues of a positive semi-definite symmetric matrix A is indicated by $|A|_0$.

2. Population results

2·1. *Covariance reductions*

For samples of size $n_g + 1$ with $n_g \geq p$, let $\widetilde{\Sigma}_g$ denote the sample covariance matrix from population g computed with divisor n_g and let $S_g = n_g\widetilde{\Sigma}_g$, $g = 1, \ldots, h$. Random sampling may or may not be stratified by population, but in either case we condition on the observed sample sizes. Our general goal is to find a semi-orthogonal matrix $\alpha \in \mathbb{R}^{p\times q}$, $q < p$, with the property that for any two populations j and k

$$S_j|(\alpha^T S_j\alpha = B, n_j = m) \sim S_k|(\alpha^T S_k\alpha = B, n_k = m). \tag{2}$$

In other words, given $\alpha^T S_g\alpha$ and n_g, the conditional distribution of $S_g|(\alpha^T S_g\alpha, n_g)$ must not depend on g. In this way we may reasonably say that, apart from differences due to sample size, the quadratic reduction $R(S) = \alpha^T S\alpha : \mathbb{S}^{p\times p} \to \mathbb{S}^{q\times q}$ is sufficient to account for the heterogeneity among the population covariance matrices. Recalling that α_0 denotes a completion of α, (2) does not require $\alpha_0^T S_g\alpha_0$ to be constant stochastically, but this must be so conditionally given the sample size and $\alpha^T S_g\alpha$. The matrix α is not identified since,

for any full rank $A \in \mathbb{R}^{q\times q}$, (2) holds for α if and only if it holds for αA. Consequently, (2) is a requirement on the subspace $\mathcal{S}_\alpha$ rather than on its basis α. Our restriction to orthonormal bases is for convenience only. For any α satisfying (2) we will call $\mathcal{S}_\alpha$ a dimension reduction subspace for the sample covariance matrices $\widehat{\Sigma}_g$, $g = 1, \ldots, h$. The smallest dimension reduction subspace can be identified and estimated, as discussed in §§2·2 and 3. This formulation does not appeal to variability preservation or spectral decompositions for its motivation. Since it requires the conditional distribution of $S_g|(\alpha^T S_g \alpha, n_g)$ to be independent of g, it seems more demanding than approaches like (1) that model just the population covariance matrices Σ_g.

To make (2) operational we assume that the S_g's are independently distributed as Wishart random matrices, $S_g \sim W(\Sigma_g, p, n_g)$, which is a common assumption in spectral modeling (see, for example, Flury, 1987; Boik, 2002). The sum of squares matrices S_g can then be characterized as $S_g = Z_g^T Z_g$, with $Z_g \in \mathbb{R}^{n_g\times p}$ and $Z_g \sim N(0, I_{n_g} \otimes \Sigma_g)$. Therefore we have the following two results: For $g = 1, \ldots, h$,

$$Z_g|(Z_g\alpha, n_g) \sim N[Z_g P_{\alpha(\Sigma_g)}, I_{n_g} \otimes \Sigma_g\{I_p - P_{\alpha(\Sigma_g)}\}] \tag{3}$$

$$S_g|(Z_g\alpha, n_g) \sim W[\Sigma_g\{I_p - P_{\alpha(\Sigma_g)}\}, p, n_g;\, P_{\alpha(\Sigma_g)}^T Z_g^T Z_g P_{\alpha(\Sigma_g)}], \tag{4}$$

where W with four arguments describes a non-central Wishart distribution (Eaton, 1983, p. 316). From (4) we see that the distribution of $S_g|(Z_g\alpha, n_g)$ depends on $Z_g\alpha$ only through $\alpha^T Z_g^T Z_g \alpha = \alpha^T S_g \alpha$. It follows that the conditional distribution of $S_g|(\alpha^T S_g\alpha, n_g)$ is as given in (4), and thus $\mathcal{S}_\alpha$ is a dimension reduction subspace if and only if, in addition to n_g,

$$\text{(a) } P_{\alpha(\Sigma_g)} \text{ and (b) } \Sigma_g\{I_p - P_{\alpha(\Sigma_g)}\} \tag{5}$$

are constant in g. With normal populations, $\mathrm{cov}(X|\alpha^T X, g) = \Sigma_g\{I_p - P_{\alpha(\Sigma_g)}\}$ (Cook, 1998, p. 131). Thus, condition (5b) requires that $\mathrm{cov}(X|\alpha^T X, g)$ be nonrandom and constant in g. The conditional means $\mathrm{E}(X|\alpha^T X, g) = \mathrm{E}(X|g) + P^T_{\alpha(\Sigma_g)}\{X - \mathrm{E}(X|g)\}$ need not be constant in g, but condition (5a) says that the centered means $\mathrm{E}(X|\alpha^T X, g) - \mathrm{E}(X|g)$ must all lie in the same subspace $\mathcal{S}_{\Sigma\alpha}$.

The following proposition, which does not require Wishart distributions, gives conditions on $\mathcal{S}_\alpha$ that are equivalent to (5). Let $\Sigma = \sum_{g=1}^h f_g \Sigma_g$, where $f_g = n_g/n$ and $n = \sum_{g=1}^h n_g$.

PROPOSITION 1. Let $\alpha \in \mathbb{R}^{p\times q}$, $q \le p$, be any basis matrix for $\mathcal{S} \subseteq \mathbb{R}^p$. *Condition (5) and the following four statements are equivalent. For* $g = 1, \ldots, h$,

(i). $\Sigma_g^{-1}\alpha_0 = \Sigma^{-1}\alpha_0$,

(ii). the following two conditions hold

$$P_{\alpha(\Sigma_g)} = P_{\alpha(\Sigma)} \tag{6}$$

$$\Sigma_g\{I_p - P_{\alpha(\Sigma_g)}\} = \Sigma\{I_p - P_{\alpha(\Sigma)}\}, \tag{7}$$

(iii). $\Sigma_g = \Sigma + P^T_{\alpha(\Sigma)}(\Sigma_g - \Sigma)P_{\alpha(\Sigma)}$,

(iv). $\Sigma_g^{-1} = \Sigma^{-1} + \alpha\{(\alpha^T\Sigma_g\alpha)^{-1} - (\alpha^T\Sigma\alpha)^{-1}\}\alpha^T$.

Proposition 1 characterizes subspaces rather than particular bases since it holds for α if and only if it holds for any basis matrix for $\mathcal{S}_\alpha$. Its first conclusion implies that $\Sigma^{-1/2}\mathcal{S}_\alpha^\perp$ is an eigenspace with eigenvalue 1 of each of the standardized covariance matrices $\Sigma^{-1/2}\Sigma_g\Sigma^{-1/2}$. This provides a connection with Flury's models of common principal components, but the link is in term of the standardized variables $\Sigma^{-1/2}X$ rather than the original variables X.

When $h = 2$ conclusion (i) is equivalent to $\Sigma_2\Sigma_1^{-1}\alpha_0 = \alpha_0$, which is related to Flury's (1983) proposal to use the eigenvectors of $\Sigma_1^{-1}\Sigma_2$ to study the differences between two covariance matrices. A broader relationship with Flury's models in the scale of X is provided in §2·3. The second conclusion gives the constant values of the matrices in condition (5) and the final two conclusions give representations for Σ_g^{-1} and Σ_g.

2·2. *Central subspaces*

There may be many dimension reduction subspaces and one with minimal dimension is of special interest. When the intersection of all dimension reduction subspaces is itself a dimension reduction subspace we call it the central subspace (Cook, 1994, 1998) and denoted it by $\mathcal{C}$ with $d = \dim(\mathcal{C})$. If the S_g's are independent Wishart matrices then $\mathcal{S}_\alpha$ is a dimension reduction subspace if and only if it satisfies Proposition 1. This equivalence together with the next proposition implies the existence of $\mathcal{C}$ when the S_g's are Wishart.

PROPOSITION 2. *If $\mathcal{S}$ and $\mathcal{T}$ are subspaces that satisfy Proposition 1, then $\mathcal{S} \cap \mathcal{T}$ also satisfies Proposition 1.*

The central subspace serves to characterize the minimal reduction. It is equivariant under linear transformations: If $\mathcal{C}$ is the central subspace for $\widetilde{\Sigma}_g$ then $A^{-T}\mathcal{C}$ is the central subspace for $A\widetilde{\Sigma}_g A^T$, where $A \in \mathbb{R}^{p\times p}$ is nonsingular. This distinguishes the proposed approach from spectral methods, which do not have a similar property. The parameter space for $\mathcal{C}$ is a d dimensional Grassmann manifold $\mathcal{G}_{(d,p)}$ in $\mathbb{R}^p$; a single subspace in $\mathcal{G}_{(d,p)}$ is uniquely determined by choosing $d(p-d)$ real numbers (Chikuse, 2003).

We will refer to models characterized by the conditions of Proposition 1 as covariance reducing models. Part (iii) of Proposition 1 shows that Σ_g depends only on Σ, $\mathcal{C}$ and the coordinate matrices $\alpha^T\Sigma_g\alpha$ for $g = 1, \ldots, h-1$, with parameter space being the Cartesian

product of $\mathbb{S}^{p\times p}$, $\mathcal{G}_{d,p}$ and $h-1$ repeats of $\mathbb{S}^{d\times d}$. Consequently the total number of reals needed to fully specify an instance of the model is $p(p+1)/2+d(p-d)+(h-1)d(d+1)/2$. This count will be used later when determining degrees of freedom for likelihood-based inference.

2·3. *Relationships with spectral models*

Let $\Gamma_* \in \mathbb{R}^{p\times(p-q)}$ be a basis matrix for $\text{span}(\Gamma_g)$ in model (1). Then $\Gamma^T X$ and $\Gamma_*^T X$ are independent within each population, but the conditional covariance $\text{cov}(\Gamma^T X|\Gamma_*^T X, g) = \text{cov}(\Gamma^T X|g) = \Gamma^T\Sigma_g\Gamma$ need not be constant in g. In the covariance reducing model, $\alpha_0^T X$ and $\alpha^T X$ may be dependent but the conditional covariance $\text{cov}(\alpha_0^T X|\alpha^T X, g)$ must be constant in g. Because of this fundamental difference in structure it seems difficult to find direct connections between the methods. However, a relationship can be found by using the reducing subspaces of Σ. Since $\Sigma \in \mathbb{S}^{p\times p}$, a subspace $\mathcal{S}$ of $\mathbb{R}^p$ is a reducing subspace of Σ if and only if $\Sigma\mathcal{S} = \mathcal{S}$ (see, for example, Conway, 1990, p. 36). For example, the subspace spanned by any set of eigenvectors of Σ is a reducing subspace of Σ.

Let $\mathcal{E}_\Sigma(\mathcal{C})$ denote the intersection of all reducing subspaces of Σ that contain $\mathcal{C}$ and let $u = \dim\{\mathcal{E}_\Sigma(\mathcal{C})\}$, $p \geq u \geq d$. The subspace $\mathcal{E}_\Sigma(\mathcal{C})$, which is called the Σ-envelope of $\mathcal{C}$ (Cook, Li & Chiaromonte, 2007), provides a unique upper bound on $\mathcal{C}$ based on the reducing subspaces of Σ. Since $\mathcal{E}_\Sigma(\mathcal{C})$ is itself a reducing subspace of Σ we have the general form $\Sigma = \gamma_0 V_0 \gamma_0^T + \gamma V \gamma^T$, where $V_0 \in \mathbb{S}^{(p-u)\times(p-u)}$, $V \in \mathbb{S}^{u\times u}$ and $\gamma \in \mathbb{R}^{p\times u}$ is a basis matrix for $\mathcal{E}_\Sigma(\mathcal{C})$. Substituting this relationship into identity (iii) of Proposition 1 and simplifying we find that Σ_g can be parameterized in terms of the envelope $\mathcal{E}_\Sigma(\mathcal{C})$ as

$$\Sigma_g = \gamma_0 M_0 \gamma_0^T + \gamma M_g \gamma^T, \tag{8}$$

for some $M_0 \in \mathbb{S}^{(p-u)\times(p-u)}$ and $M_g \in \mathbb{S}^{u\times u}$, $g = 1, \ldots, h$. The spectral properties of this envelope model (8) can be represented explicitly by using the spectral decompositions $M_0 = v_0 D_0 v_0^T$ and $M_g = v_g D_g v_g^T$, where v_0 and v_g are orthogonal matrices, and D_0 and D_g are diagonal matrices. Let $\eta_0 = \gamma_0 v_0$ and $\eta_g = \gamma_g v_g$. Then (η_0, η_g) is an orthogonal matrix and

$$\Sigma_g = \eta_0 D_0 \eta_0^T + \eta_g D_g \eta_g^T. \tag{9}$$

This relationship shows that all eigenvectors of Σ_g can be constructed to be in either $\mathcal{E}_\Sigma(\mathcal{C})$ or $\mathcal{E}_\Sigma^\perp(\mathcal{C})$. The envelope model (8) is parameterized in terms of $\mathcal{E}_\Sigma(\mathcal{C}) \in \mathcal{G}_{(u,p)}$, and it uses a total of $u(p-u) + (p-u)(p-u+1)/2 + u(u+1)h/2$ real parameters. Representation (9) is a reparameterization in terms of the eigenvectors of Σ_g and their parameter space is a Steifel manifold. More importantly, (9) can be seen as an instance of Flury's (1987) partial common principal components model (1), while (8) is an instance of his common space model. The full versions of Flury's models allow M_0 and D_0 to depend on g, while the present formulation does not because of the sufficiency requirement (2). Additionally, (9) requires no relationship between D_0 and D_g so the common components $\eta_0^T X$ can be associated with the largest or smallest eigenvalues of Σ_g. This discussion leads to the conclusion that spectral models can be structured to provide an upper bound on $\mathcal{C}$.

For example, consider the structure

$$\Sigma_g = I_p + \sigma_g^2 \alpha\alpha^T, \;\; g = 1, \ldots, h, \tag{10}$$

where $\alpha \in \mathbb{R}^p$, $\alpha^T\alpha = 1$, the σ_g's are distinct, and $\mathcal{E}_\Sigma(\mathcal{C}) = \mathcal{C} = \mathcal{S}_\alpha$. This setting can also be described by Flury's common principal component model, or his common space model. If

the σ_g^2's are sufficiently large then $\alpha^T X$ may serve as a variance preserving reduction in the sense of Schott (1991). If we perform a nonsingular transform $A \in \mathbb{R}^{p\times p}$ and work in the scale of $\Sigma_g^* = A\Sigma_g A^T$, then the corresponding central subspace is $\mathcal{C}^* = A^{-T}\mathcal{C}$, which is still one dimensional. However, depending on the choice of A, the $\Sigma^* = \sum_{g=1}^h f_g \Sigma_g^*$ envelope of $\mathcal{C}^*$ may be $\mathbb{R}^p$, and the Σ_g^*'s may share no eigenspaces other than $\mathbb{R}^p$.

If we modify (10) to obtain $\Sigma_g^* = A + \sigma_g^2 \alpha\alpha^T$, where $A \in \mathbb{S}^{p\times p}$, then $\mathcal{C}^* = A^{-1}\mathcal{C}$ is still one-dimensional, but again the Σ_g^*'s may share no eigenspaces other than $\mathbb{R}^p$, depending on A. In short, covariance reducing models and the various spectral approaches can target the same or very different population quantities.

3. Estimation of $\mathcal{C}$ with d specified

The following proposition summarizes maximum likelihood estimation when the S_g's are Wishart matrices and $d = \dim(\mathcal{C})$ is specified. The choice of d is considered in §5.

Proposition 3. *The maximum likelihood estimator of Σ is its sample version $\widehat{\Sigma} = \sum_{g=1}^h f_g \widetilde{\Sigma}_g$. The maximum likelihood estimator $\widehat{\mathcal{C}}$ of $\mathcal{C}$ maximizes over $\mathcal{S} \in \mathcal{G}_{(d,p)}$ the log likelihood function*

$$L_d(\mathcal{S}) = c - \frac{n}{2}\log|\widehat{\Sigma}| + \frac{n}{2}\log|P_{\mathcal{S}}\widehat{\Sigma}P_{\mathcal{S}}|_0 - \sum_{g=1}^h \frac{n_g}{2}\log|P_{\mathcal{S}}\widetilde{\Sigma}_g P_{\mathcal{S}}|_0, \tag{11}$$

where c is a constant depending only on p, n_g and $\widetilde{\Sigma}_g$, $g = 1, \ldots, h$. The maximum likelihood estimator $\widehat{\Sigma}_g$ of Σ_g is constructed by substituting a basis matrix $\widehat{\alpha}$ for $\widehat{\mathcal{C}}$, $\widetilde{\Sigma}_g$ and $\widehat{\Sigma}$ for the corresponding quantities on the right of the equation in part (iii) of Proposition 1.

If $\mathcal{C} = \mathbb{R}^p$ ($d = p$) then the log likelihood (11) reduces to the usual log likelihood for fitting separate covariance matrices to the h populations. If $\mathcal{C}$ is equal to the origin

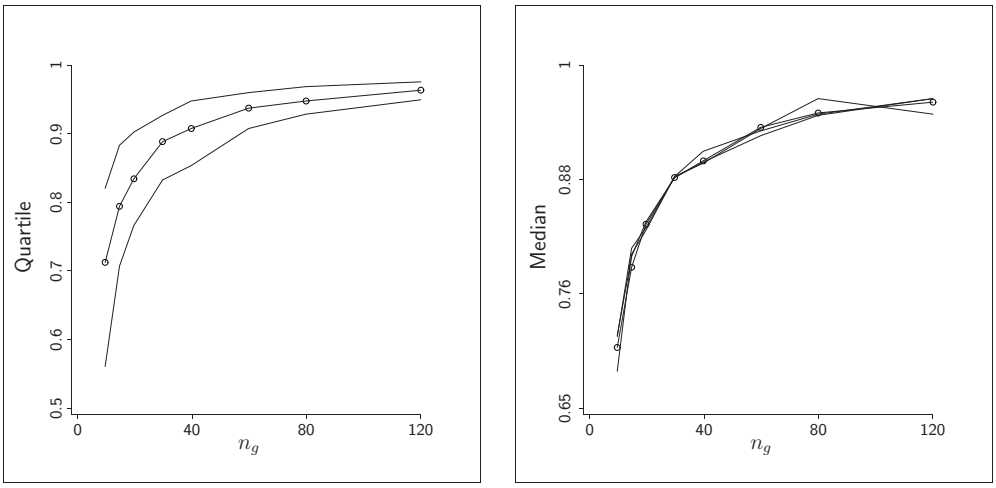
($d = 0$) then (11) becomes the log likelihood for fitting a common covariance matrix to all populations. This corresponds to deleting the two terms of (11) that depend on $\mathcal{S}$. The following corollary confirms the invariance of the estimated reduction $\widehat{R}$ under full rank quadratic transformations.

COROLLARY 1. *If* $A \in \mathbb{R}^{p\times p}$ *is full rank and* $\widehat{R}(S) = \widehat{\alpha}^T S\widehat{\alpha}$, *then* $\widehat{R}(ASA^T) = \widehat{\gamma}^T ASA^T\widehat{\gamma}$, *with* $\mathcal{S}_{\widehat{\gamma}} = A^{-T}\mathcal{S}_{\widehat{\alpha}}$.

To illustrate basic properties of estimation we simulated observations from model (10) with $p = 6$, $\alpha = (0, \ldots, 0, 1)^T$, $h = 3$, $\sigma_1 = 1$, $\sigma_2 = 4$ and $\sigma_3 = 8$. The use of the identity matrix I_p in the construction of Σ_g was for convenience only since the results are invariant under full rank transformations, as indicated in Corollary 1. The $\widetilde{\Sigma}_g$'s were constructed using observed vectors $X = \varepsilon + \sigma_g \alpha\epsilon$ generated from independent vectors $(\varepsilon^T, \epsilon)$ of independent standard normal variates, with $\varepsilon \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^1$. The ε term in X represents the component that is stochastically the same in all populations and the other term represents the population-specific component. Maximization of the log likelihood (11) was carried out using computer code developed from Liu, et al. (2004). Figure 1a shows the sample quartiles from 400 replications of the cosine of the angle between $\widehat{\mathcal{C}}$ and $\mathcal{C}$ for several sample sizes and normal errors. The method seems to respond reasonably to increasing sample size.

4. CENTRAL MEAN SUBSPACES

As represented in Proposition 1, the assumption of Wishart distributions for the S_g's implies informative and rather elegant equivariant characterizations of the covariance matrices Σ_g in terms of a basis matrix α for $\mathcal{C}$. While a straightforward connection between $\mathcal{C}$ and Proposition 1 may be problematic without Wishart distributions, its equivalences



a. Normal Error b. Other Errors

Figure 1: Quartiles (a) and median (b) of the cosine of the angle between $\widehat{\alpha}$ and $\mathcal{C}$ versus sample size.

can be used without distributional assumptions as a model for the population covariance matrices Σ_g, just as spectral decompositions like (1) have been used. Let $\mathcal{M}$ denote the intersection of all subspaces that satisfy Proposition 1. It follows from Proposition 2 that $\mathcal{M}$ also satisfies Proposition 1 and consequently it is a well-defined parameter that can be used as an inferential target. We refer to $\mathcal{M}$ as the central mean subspace since its role is to characterize the structure of the conditional means $\mathrm{E}(\widetilde{\Sigma}_g) = \Sigma_g$. If the S_g's are Wishart matrices then $\mathcal{C} = \mathcal{M}$.

The following proposition shows that without Wishart distributions the likelihood (11) still provides a Fisher consistent estimator of $\mathcal{M}$. Consequently, (11) can be used as a distribution-free objective function with the goal of modeling Σ_g in terms of the equivalences of Proposition 1.

PROPOSITION 4. *Let $d = \dim(\mathcal{M})$. Then for $\mathcal{S} \in \mathcal{G}_{(d,p)}$, $L_d(\mathcal{S})/n$ converges to*

$$K_d(\mathcal{S}) = c + (1/2)\log|P_{\mathcal{S}}\Sigma P_{\mathcal{S}}|_0 - \sum_{g=1}^{h}(f_g/2)\log|P_{\mathcal{S}}\Sigma_g P_{\mathcal{S}}|_0 \tag{12}$$

and $\mathcal{M} = \arg\max K_b(\mathcal{S})$, where c is a constant not depending on $\mathcal{S}$.

Figure 1b shows the median over 400 replication of the cosine of the angle between $\widehat{\mathcal{M}}$ and $\mathcal{M} = \mathcal{S}_\alpha$ for normal, t_5, χ^2_5 and uniform $(0,1)$ error $(\varepsilon^T, \epsilon)$ distributions with simulation model (10) and parameter values stated in §3. The results in Figure 1b match so well that the individual curves were not marked. This along with other unreported simulations suggest that a normal error distribution is not essential for the likelihood-based objective function (11) to give good results for the estimation of $\mathcal{M}$.

In §§5 and 6 we consider methods for inference about d and tests for active predictors. These methods are developed assuming Wishart distributions, and then applied in simulations with non-Wishart distributions to gain insights into their behaviour in such cases.

5. CHOICE OF d

In this section we consider ways in which $d = \dim(\mathcal{C})$ can be chosen in practice, distinguishing the true value d from value d_0 used in fitting.

The hypothesis $d = d_0$ can be tested by using the likelihood ratio statistic $\Lambda(d_0) = 2\{\hat{L}_p - \hat{L}_{d_0}\}$, where $\hat{L}_p$ denotes the value of the maximized log likelihood for the full model with $d_0 = p$ and $\hat{L}_{d_0}$ is the maximum value of the log likelihood (11). Following standard likelihood theory, under the null hypothesis $\Lambda(d_0)$ is distributed asymptotically as a chi-squared random variable with degrees of freedom $(p-d)\{(h-1)(p+1) + (h-3)d\}/2$, for

$h \geq 2$ and $d < p$. The statistic $\Lambda(d_0)$ can be used in a sequential testing scheme to choose d: Using a common test level and starting with $d_0 = 0$, choose the estimate $\widehat{d}$ of d as the first hypothesized value that is not rejected. The test for $d = 0$ is the same as Bartlett's test for equality of the Σ_g's, but without his proportional correction of $\Lambda(0)$ (Muirhead, 1982, Ch. 8). This method for dimension selection is common in dimension reduction literature (see Cook, 1998, p. 205 for background).

A second approach is to use, for instance, the Akaike or Bayes information criterion. The Bayes information criterion is consistent while Akaike's is minimax-rate optimal (Burnham & Anderson, 2002). In this approach $\widehat{d}$ is selected to minimize over d_0 the information criterion $IC(d_0) = -2\,\hat{L}_{d_0} + h(n)g(d_0)$, where $g(d_0)$ is the number of parameters to be estimated, and $h(n)$ is equal to $\log n$ for the Bayes criterion and 2 for Akaike's.

We use the sequential testing method to illustrate that useful inference for d is possible, without recommending a particular method. There are many methods that could be used to select d and a comprehensive comparison is outside the scope of this report. Table 1 gives the empirical distribution of $\widehat{d}$ from 200 replications from the simulation model described in §3. The first column labeled "Law" gives the distribution of the error $(\varepsilon^T, \epsilon)$. For normal distributions $d = \dim(\mathcal{C}) = \dim(\mathcal{M})$, while for the non-normal distributions $d = \dim(\mathcal{M})$. The second column gives the common intra-population sample size. All tests were performed with constant nominal level 0·01. The relatively poor showing at $n_g = 15$ with normal errors seems due to the power of Barlett's test at this small sample size. The method responded well to increasing sample size and the expected asymptotic results were observed at $n_g = 40$ with normal errors (N). Uniform errors (U) did not have a notable impact on the results, but skewed and heavy tailed errors resulted in more overestimation than expected with normal

errors. On balance, we regard the sequential method as useful, although the development of robust methods for $\dim(\mathcal{M})$ might mitigate overestimation due to skewed and heavy tailed errors.

Table 1: Empirical distribution of $\widehat{d}$ in percent.

		$\widehat{d}$				
Law	n_g	0	1	2	3	4
N	15	13·0	75·5	8·0	3·0	0·5
N	20	2·5	94·0	3·0	0	0
N	30	0·5	95·0	2·0	1·5	0·5
N	40	0	99·0	1·0	0	0
U	40	0	100	0	0	0
χ^2_5	40	0	88·5	9·5	2	0
t_{10}	40	0	94·0	5·5	0·5	0
t_7	40	0	82·0	15·0	2·5	0·5

N, standard normal; U, uniform $(0, 1)$.

6. Testing variates

With d specified a priori or after estimation, it may of interest in some applications to test an hypothesis that a selected subspace $\mathcal{H}$ of dimension $k \leq p - d$ is orthogonal to $\mathcal{C}$ in the usual inner product. The restriction on k is to insure that the dimension of $\mathcal{C}$ is still d under the hypothesis. The hypothesis $P_{\mathcal{H}}\mathcal{C} = 0$ can be tested by using a standard likelihood test. The test statistic is $\Lambda_d(\mathcal{H}) = 2(\hat{L}_d - \hat{L}_{d,\mathcal{H}})$, where $\hat{L}_d$ is the maximum value of the log likelihood (11), and $\hat{L}_{d,\mathcal{H}}$ is the maximum value of (11) with $\mathcal{C}$ constrained by the hypothesis. Under the hypothesis $P_{\mathcal{H}}\mathcal{C} = 0$ the statistic $\Lambda_d(\mathcal{H})$ is distributed asymptotically as a chi-squared random variable with dk degrees of freedom.

The maximized log likelihood $\hat{L}_{d,\mathcal{H}}$ can be obtained by maximizing over $\mathcal{S} \in \mathcal{G}_{(d,p-k)}$ the constrained log likelihood

$$L_d(\mathcal{S}) = c - \frac{n}{2}\log|\widehat{\Sigma}| + \frac{n}{2}\log|P_{\mathcal{S}}H_1^T\widehat{\Sigma}H_1P_{\mathcal{S}}|_0 - \sum_{g=1}^{h}\frac{n_g}{2}\log|P_{\mathcal{S}}H_1^T\widetilde{\Sigma}_gH_1P_{\mathcal{S}}|_0, \qquad (13)$$

where $H_1 \in \mathbb{R}^{p\times(p-k)}$ is a basis matrix for $\mathcal{H}^{\perp}$. When testing that a specific subset of k variables is not directly involved in the reduction, the role of H_1 in (13) is to select the parts of $\widehat{\Sigma}$ and $\widetilde{\Sigma}_g$ that correspond to the other variables.

Shown in Table 2 are the empirical levels based on 1000 simulations of nominal 1, 5 and 10 percent tests of the hypothesis that the first variate does not contribute directly to the reduction in model (10) with $\alpha = (0,\ldots,0,1)^T$, $\mathcal{H} = \text{span}\{(1,0,\ldots,0)^T\}$. For the three non-normal distributions the hypothesis tested is $P_{\mathcal{H}}\mathcal{M} = 0$. The agreement seems quite good for large samples, but otherwise the results indicate a clear tendency for the actual level to be larger than the nominal, a tendency that is made worse by skewness or heavy tails. Use of this test may be problematic when the sample size is not large and very accurate test levels are required. However, in some settings it may be sufficient to have the actual level be between 1 and 5 percent, and our results indicate that this can be achieved by testing at the nominal 1 percent level.

7. Garter snakes

Phillips & Arnold (1999) used Flury's hierarchy of principal component models to study genetic covariance matrices for six traits of female garter snakes in costal and inland populations of northern California. We illustrate aspects of the proposed methodology using the same covariance matrices. The sample sizes for the costal and inland populations are 90 and

Table 2: Simulation results on the level of the variate test using the likelihood ratio statistic $\Lambda_d(\mathcal{H})$.

Law	p	n_g	1%	5%	10%
N	6	20	3·0	8·6	15·2
N	6	40	1·5	5·8	10·9
N	10	50	2·2	8·9	14·6
N	10	70	1·4	5·4	11·2
N	15	80	1·3	6·5	13·6
N	15	120	1·2	5·6	10·7
U	10	70	1·6	5·9	13·0
χ^2_5	10	70	1·6	6·8	12·3
t_7	10	70	1·8	7·3	13·0

N, standard normal; U, uniform $(0, 1)$.

139, so we expect the large-sample methods proposed here to be reasonable. Conclusion (iii) of Proposition 1 implies that the difference $\Sigma_g - \Sigma$ will be of rank d, $g = 1, \ldots, h$. The eigenvalues of $\widetilde{\Sigma}_g - \widehat{\Sigma}$ for the inland population are (0·69, 0·14, 0·09, 0·041, −0·10, −0·82). The magnitude of these values suggests that $d = 2$ is plausible. The tests of $d = 0$, $d = 1$ and $d = 2$ resulted in the nominal p-values $4{\cdot}3 \times 10^{-9}$, 0·007 and 0·12, yielding the sequential estimate $\widehat{d} = 2$. The estimates based on the Bayes and Akaike information criteria were $\widehat{d} = 1$ and $\widehat{d} = 3$. The estimate $\widehat{d} = 2$ is also be reasonable under Akaike's criterion since the values of its objective function for $d_0 = 2$ and $d_0 = 3$ were quite close.

Phillips & Arnold (1999) concluded that the partial common principal component model (1) with $q = 4$ common components is likely the best. We use the envelope (8) to contrast this finding with that based on the covariance reducing model. Using the notation of Proposition 3 and adapting the derivation of (11), it can be shown that the maximum

likelihood estimator $\widehat{\mathcal{E}}$ of $\mathcal{E}_\Sigma(\mathcal{C})$ maximizes over $\mathcal{S} \in \mathcal{G}_{(u,p)}$ the log likelihood function

$$L_u(\mathcal{S}) = c - \frac{n}{2}\log|\widehat{\Sigma}| - \frac{n}{2}\log|P_{\mathcal{S}}\widehat{\Sigma}^{-1}P_{\mathcal{S}}|_0 - \sum_{g=1}^{h}\frac{n_g}{2}\log|P_{\mathcal{S}}\widetilde{\Sigma}_g P_{\mathcal{S}}|_0, \tag{14}$$

where $u = \dim\{\mathcal{E}_\Sigma(\mathcal{C})\}$. The maximum likelihood estimators of M_0 and M_g are $\widehat{M} = \widehat{\gamma}_0^T\widehat{\Sigma}\widehat{\gamma}_0$ and $\widehat{M}_g = \widehat{\gamma}^T\widetilde{\Sigma}_g\widehat{\gamma}$, where $\widehat{\gamma}$ is a basis matrix for $\widehat{\mathcal{E}}$, $g = 1, \ldots, h$. The tests of $u = 1$, $u = 2$ and $u = 3$ based on (14) gave the nominal p-values 0·0088, 0·03 and 0·17. Accordingly, it seems reasonable to conclude that u is either 2 or 3. At $u = 2$ Flury's spectral model (1), the covariance reducing model and the envelope model (8) can all agree with $u = d = p - q = 2$, $\text{span}(\Gamma) = \text{span}(\alpha_0) = \text{span}(\gamma_0)$ and $\Lambda_{1,g}$ a constant in g. At $u = 3$ and $d = 2$ the models can no longer agree since the envelope model requires that we condition on an additional linear combination.

To emphasize the potential differences due to invariance properties, we re-estimated the dimensions d and u after transforming each sample covariance matrix as $\Sigma_g \to A\Sigma_g A^T$, where $A \in \mathbb{R}^{6\times 6}$ was generated as a matrix of standard normal variates. As the theory predicted, the transformation had no effect on the estimated dimension $\widehat{d}$ of the covariance reducing model, but the estimated dimension of the envelope model was $\widehat{u} = 6$.

We continue this illustration using the covariance reducing model with $d = 2$, so two linear combinations of the traits are needed to explain differences in variation. In units of the observed trait standard deviations, the estimated direction vectors that span $\widehat{\mathcal{C}}$ are $\widehat{\alpha}_1 = (0{\cdot}13,\ 0{\cdot}31,\ -0{\cdot}17,\ -0{\cdot}91,\ 0{\cdot}04,\ 0{\cdot}17)^T$ and $\widehat{\alpha}_2 = (0{\cdot}07,\ -0{\cdot}13,\ -0{\cdot}86,\ 0{\cdot}33,\ -0{\cdot}13,\ 0{\cdot}34)^T$, where the trait order is as given by Phillips & Arnold (1999, Table 1). These results suggest that the third and fourth traits are largely responsible for the differences in the covariance matrices. The variate test of §6 applied to each trait individually resulted in the p-values

$(0{\cdot}39,\ 0{\cdot}24,\ 1{\cdot}6 \times 10^{-6},\ 2{\cdot}4 \times 10^{-9},\ 0{\cdot}52,\ 0{\cdot}01)$, which agrees with the qualitative impression from the standardized spanning vectors. Testing the joint hypothesis that only the third and fourth traits are involved in the conditioning resulted in a p-value of $0{\cdot}04$. These and other unreported results indicate that the third and fourth traits are largely responsible for the differences between the genetic covariance matrices at the two locations. The sixth trait may also contribute to the differences but its relevance is not as clear. The overall indication then is that only the third and fourth rows of α are non-zero. We have illustrated estimation of a basis matrix α for $\mathcal{C}$ since that is needed prior to determining all other parameter estimates, as shown in Proposition 3. The analysis could now continue in a variety of ways.

8. Discussion

We proposed a new point of view for the study of covariance matrices, gave first Wishart methodology and included some results on the behaviour of that methodology in non-Wishart settings. Our most ambitious goal is to reduce the sample covariance matrices to an informational core $\alpha^T \widetilde{\Sigma}_g \alpha$ that is sufficient to characterize the variance heterogeneity among the populations. The invariant and equivariant properties of covariance reducing models seem particularly appealing. Nevertheless, if substantive questions in application directly involve the spectral structures of the covariance matrices, then spectral modeling would of course be appropriate. On the other hand, if such questions are not spectral-specific then covariance reducing models may be a useful alternative. Both approaches could be helpful in exploratory analyses.

There are many open questions and directions for future study. Of immediate interest is the development of methodology for estimating $\mathcal{C}$ that does not require Wishart distributions, but perhaps constrains some of the conditional moments of $S_g|(\alpha^T S_g \alpha, n_g)$. Standard

errors of identified functions of the parameters can be determined from the limiting distributions of the estimates. With $b \in \mathbb{R}^p$ and $P_{\alpha(\Sigma)}b = 0$, quantities of the form $\Sigma_g b$ are constant in g and may be of interest in some studies. In such cases it might be worthwhile to consider inferences conditional on $\alpha^T S_g \alpha = B$, $g = 1, \ldots, g$. There may be new ideas and methodology for the study of correlation matrices that parallels those expressed here for covariance matrices. For instance, the equivalences of Proposition 1 still hold if we reinterpret Σ_g as a correlation matrix and the likelihood function (11) will still give a Fisher consistent estimator of the central mean subspace for correlation matrices.

Acknowledgments

Research for this article was supported in part by a grant from the U.S. National Science Foundation, and by Fellowships from the Isaac Newton Institute for Mathematical Sciences, Cambridge, U.K. The authors are grateful to Patrick Phillips for providing the covariance matrices for the garter snake illustration, and to the referees for their helpful comments.

Appendix

Proofs

The first of the following two preliminary propositions was given by Rao (1973, p. 77).

Proposition A1. *Let $B \in \mathbb{S}^{p\times p}$ and let $\alpha \in \mathbb{R}^{p\times d}$ be a semi-orthogonal matrix. Then*

$$\alpha(\alpha^T B\alpha)^{-1}\alpha^T + B^{-1}\alpha_0(\alpha_0^T B^{-1}\alpha_0)^{-1}\alpha_0^T B^{-1} = B^{-1}. \tag{A1}$$

As a consequence we have

$$
\begin{aligned}
(\alpha_0^T B^{-1}\alpha_0)^{-1} &= \alpha_0^T B\alpha_0 - \alpha_0^T B\alpha(\alpha^T B\alpha)^{-1}\alpha^T B\alpha_0 && \text{(A2)}\\
I_p - P^T_{\alpha(B)} &= P_{\alpha_0(B^{-1})}, && \text{(A3)}\\
-(\alpha_0^T B^{-1}\alpha_0)^{-1}(\alpha_0^T B^{-1}\alpha) &= (\alpha_0^T B\alpha)(\alpha^T B\alpha)^{-1}. && \text{(A4)}
\end{aligned}
$$

PROPOSITION A2. *Suppose that $B \in \mathbb{S}^{p\times p}$ and $\alpha \in \mathbb{R}^{p\times d}$ is a semi-orthogonal matrix. Then* $|\alpha_0^T B\alpha_0| = |B||\alpha^T B^{-1}\alpha|$.

Proof of Proposition A2. Let $K \in \mathbb{R}^{p\times p}$ with first block of rows $(I_d, \alpha^T B\alpha_0)$ and second block of rows $(0, \alpha_0^T B\alpha_0)$. Since (α, α_0) is an orthogonal matrix,

$$
\begin{aligned}
|\alpha_0^T B\alpha_0| &= |(\alpha,\alpha_0)K(\alpha,\alpha_0)^T| = |\alpha\alpha^T + \alpha\alpha^T B\alpha_0\alpha_0^T + \alpha_0\alpha_0^T B\alpha_0\alpha_0^T|\\
&= |B - (B - I_p)\alpha\alpha^T| = |B||I_d - \alpha^T(I_p - B^{-1})\alpha| = |B||\alpha^T B^{-1}\alpha|.
\end{aligned}
$$

Proof of Proposition 1. We will show that $(i) \Rightarrow (5) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$. We begin by showing that condition $(i) \Rightarrow (5)$. By applying (A3) with $B = \Sigma_g$:

$$
\begin{aligned}
I_p - P^T_{\alpha(\Sigma_g)} &= \alpha_0(\alpha_0^T\Sigma_g^{-1}\alpha_0)^{-1}\alpha_0^T\Sigma_g^{-1} = C_1 && \text{(A5)}\\
\{I_p - P^T_{\alpha(\Sigma_g)}\}\Sigma_g &= \alpha_0(\alpha_0^T\Sigma_g^{-1}\alpha_0)^{-1}\alpha_0^T = C_2, && \text{(A6)}
\end{aligned}
$$

where C_1 and C_2 are constant matrices since $\alpha_0^T\Sigma_g^{-1}$ is constant by hypothesis (i).

If (5) is true then (A5) and (A6) must hold. This implies that $\alpha_0^T\Sigma_g^{-1}$ is constant and thus equal to $\alpha_0^T\Sigma^{-1}$. Conclusion (ii) follows from (5) by application of (A3) with $B = \Sigma$.

(iii) follows from (ii) by replacing $P_{\alpha(\Sigma_g)}$ with $P_{\alpha(\Sigma)}$ in the second condition of (ii) and

rearranging terms: $\Sigma_g - \Sigma = (\Sigma_g - \Sigma)P_{\alpha(\Sigma)} = P^T_{\alpha(\Sigma)}(\Sigma_g - \Sigma)P_{\alpha(\Sigma)}$.

Conclusion (iv) follows from (iii) by direct multiplication. Finally, multiplying (iv) on the right by α_0 immediately gives condition (i).

Proof of Proposition 2. Let α and β be two semi-orthogonal matrices that satisfy (5). Then $\alpha_0^T\Sigma_g^{-1}$ and $\beta_0^T\Sigma_g^{-1}$ are constant, and consequently $(\alpha_0, \beta_0)^T\Sigma_g^{-1}$ is constant. This implies that $(\mathcal{S}_\alpha^\perp + \mathcal{S}_\beta^\perp)^\perp$ is a dimension reduction subspace. The conclusion follows since $\mathcal{S}_\alpha^\perp + \mathcal{S}_\beta^\perp = (\mathcal{S}_\alpha \cap \mathcal{S}_\beta)^\perp$ (Greub, 1981, page 74).

The following characterization facilitates finding the maximum likelihood estimators for the parameters when α satisfies (5a) and (5b).

PROPOSITION A3. *$R(S) = \alpha^T S\alpha$ is a sufficient reduction if and only if the following three conditions are satisfied for $g = 1, \ldots, h$:*

1. $(\alpha_0^T S_g^{-1}\alpha_0)^{-1} \sim W\{(\alpha_0^T\Sigma^{-1}\alpha_0)^{-1}, p-d, n_g - d\}$
2. $\alpha^T S_g\alpha_0|\alpha^T S_g\alpha \sim N\{-\alpha^T S_g\alpha(\alpha^T\Sigma^{-1}\alpha_0)(\alpha_0^T\Sigma^{-1}\alpha_0)^{-1}, \alpha^T S_g\alpha \otimes (\alpha_0\Sigma^{-1}\alpha_0)^{-1}\}$
3. $\alpha^T S_g\alpha \sim W(\alpha^T\Sigma_g\alpha, d, n_g)$

and $(\alpha_0^T S_g^{-1}\alpha_0)^{-1}$ and $(\alpha^T S_g\alpha_0, \alpha^T S_g\alpha)$ are stochastically independent.

Proof of Proposition A3. Using (A2) it follows that (Eaton, 1983; prop. 8.1 and 8.7),

$$\begin{aligned}
(\alpha_0^T S_g^{-1}\alpha_0)^{-1} &\sim W\{(\alpha_0^T\Sigma_g^{-1}\alpha_0)^{-1}, p-d, n_g-d\} \\
\alpha^T S_g\alpha_0|\alpha^T S_g\alpha &\sim N\{\alpha^T S_g P_{\alpha(\Sigma_g)}\alpha_0, \alpha^T S_g\alpha \otimes (\alpha_0\Sigma_g^{-1}\alpha_0)^{-1}\} \\
\alpha^T S_g\alpha &\sim W(\alpha^T\Sigma_g\alpha, d, n_g),
\end{aligned}$$

and that $(\alpha_0^T S_g^{-1}\alpha_0)^{-1}$ and $(\alpha^T S_g\alpha_0, \alpha^T S_g\alpha)$ are stochastically independent. From Propo-

sition 1, $\alpha_0^T\Sigma_g^{-1} = \alpha_0^T\Sigma^{-1}$ and $P_{\alpha(\Sigma_g)} = P_{\alpha(\Sigma)}$. The conditions of the proposition follow by using (A4) to re-express $P_{\alpha(\Sigma)}\alpha_0$.

Proof of Proposition 3. Transforming S_g to $(\alpha, \alpha_0)^T S_g(\alpha, \alpha_0)$, we have from Proposition A3 that the log likelihood is the sum of the log likelihoods arising from the densities of $(\alpha_0^T S_g^{-1}\alpha_0)^{-1}$, $\alpha^T S_g\alpha_0|\alpha^T S_g\alpha$ and $\alpha^T S_g\alpha$. Let $D = (\alpha_0^T\Sigma^{-1}\alpha_0)^{-1}$ and $H = D(\alpha_0^T\Sigma^{-1}\alpha)$.

For any semi-orthogonal matrix $\alpha \in \mathbb{R}^{p\times d}$, the transformation of $\Sigma \in \mathbb{S}^{p\times p}$ to $(\alpha, \alpha_0)^T\Sigma(\alpha, \alpha_0)$ is a one to one and onto. The transformation from $\mathbb{S}^{p\times p}$ to $\mathbb{S}^{d\times d} \times \mathbb{S}^{p-d\times p-d} \times \mathbb{R}^{(p-d)\times d}$ given by $\alpha^T\Sigma\alpha$, $D = \alpha_0^T\Sigma\alpha_0 - \alpha_0\Sigma\alpha(\alpha^T\Sigma\alpha)^{-1}\alpha^T\Sigma\alpha_0$ and $H = -(\alpha_0^T\Sigma\alpha)(\alpha^T\Sigma\alpha)^{-1}$ is also one to one and onto (Eaton, 1983, prop. 5.8). Proposition 1, (A2) and (A4) imply that fixing α for each g the dimension reduction subspace model places no constraints on D, H or $\alpha^T\Sigma_g\alpha$, which are the parameters we used for the likelihood.

The likelihood L_g for population g can be expressed prior to notable simplification as

$$\begin{aligned} L_g &= c_g - \frac{n_g - d}{2}\log|D| - \frac{n_g - p - 1}{2}\log|\alpha_0^T S_g^{-1}\alpha_0| - \frac{1}{2}\operatorname{tr}\{D^{-1}(\alpha_0^T S_g^{-1}\alpha_0)^{-1}\} \\ &\quad -\frac{n_g}{2}\log|\alpha^T\Sigma_g\alpha| + \frac{n_g - d - 1}{2}\log|\alpha^T S_g\alpha| - \frac{1}{2}\operatorname{tr}\{(\alpha^T\Sigma_g\alpha)^{-1}(\alpha^T S_g\alpha)\} \\ &\quad -\frac{p-d}{2}\log|\alpha^T S_g\alpha| - \frac{d}{2}\log|D| \\ &\quad -\frac{1}{2}\operatorname{tr}\{(\alpha^T S_g\alpha)^{-1}(\alpha^T S_g\alpha_0 + \alpha^T S_g\alpha H^T)D^{-1}(\alpha_0^T S_g\alpha + H\alpha^T S_g\alpha)\}. \end{aligned}$$

where c_g is a constant depending only on n_g and p. Using (A2) and Proposition A2, simplifying and absorbing the term $(n_g - p - 1)/2\log|S_g|$ into c_g we have

$$\begin{aligned} L_g &= c_g - \frac{n_g}{2}\log|D| - \frac{n_g}{2}\log|\alpha^T\Sigma_g\alpha| - \frac{n_g}{2}\operatorname{tr}\{(\alpha^T\Sigma_g\alpha)^{-1}(\alpha^T\widetilde{\Sigma}_g\alpha)\} \\ &\quad -\frac{1}{2}\operatorname{tr}(D^{-1}\alpha_0^T S_g\alpha_0) - \operatorname{tr}(\alpha^T S_g\alpha_0 D^{-1}H) - \frac{1}{2}\operatorname{tr}(\alpha^T S_g\alpha H^T D^{-1}H). \end{aligned}$$

With α fixed, L_g is maximized over $\alpha^T\Sigma_g\alpha$ by $\alpha^T\widetilde{\Sigma}_g\alpha$. Plugging this into L_g we get the partially maximized form

$$\begin{aligned} L_g^{(1)} &= c_g - \frac{n_g}{2}d - \frac{n_g}{2}\log|\alpha^T\widetilde{\Sigma}_g\alpha| - \frac{n_g}{2}\log|D| \\ &\quad -\frac{1}{2}\operatorname{tr}(D^{-1}\alpha_0^T S_g\alpha_0) - \operatorname{tr}(\alpha^T S_g\alpha_0 D^{-1}H) - \frac{1}{2}\operatorname{tr}(\alpha^T S_g\alpha H^T D^{-1}H). \end{aligned}$$

Let $L^{(1)} = \sum_{g=1}^h L_g^{(1)}$. Then

$$\frac{\partial L^{(1)}}{\partial H} = -\sum_{g=1}^h n_g D^{-1}\alpha_0^T\widetilde{\Sigma}_g\alpha - \sum_{g=1}^h n_g D^{-1}H\alpha^T\widetilde{\Sigma}_g\alpha$$

giving the maximum at $\widehat{H} = -\alpha_0^T\widehat{\Sigma}\alpha(\alpha^T\widehat{\Sigma}\alpha)^{-1}$, where $\widehat{\Sigma} = \sum_{g=1}^h f_g\widetilde{\Sigma}_g$. Substituting this into $L^{(1)}$ we obtain a second partially maximized log likelihood

$$\begin{aligned} L^{(2)} &= \sum_{g=1}^h c_g - \frac{n}{2}d - \sum_{g=1}^h \frac{n_g}{2}\log|\alpha^T\widetilde{\Sigma}_g\alpha| \\ &\quad -\frac{n}{2}\log|D| - \frac{n}{2}\operatorname{tr}\left\{\left(\alpha_0^T\widehat{\Sigma}\alpha_0 + 2\alpha_0^T\widehat{\Sigma}\alpha\widehat{H}^T + \widehat{H}\alpha^T\widehat{\Sigma}\alpha\widehat{H}^T\right)D^{-1}\right\}. \end{aligned}$$

This is maximized over D at $\widehat{D} = \left(\alpha_0^T\widehat{\Sigma}\alpha_0 + 2\alpha_0^T\widehat{\Sigma}\alpha\widehat{H}^T + \widehat{H}\alpha^T\widehat{\Sigma}\alpha\widehat{H}^T\right) = (\alpha_0^T\widehat{\Sigma}^{-1}\alpha_0)^{-1}$, where the second equality follows from the definition of $\widehat{H}$ and Proposition A1. Using Proposition A2, the log likelihood maximized over all parameters except α can now be written as

$$L^{(3)} = c - (n/2)\log|\widehat{\Sigma}| + \frac{n}{2}\log|\alpha^T\widehat{\Sigma}\alpha| - \sum_{g=1}^h \frac{n_g}{2}\log|\alpha^T\widetilde{\Sigma}_g\alpha|,$$

where $c = \sum_{g=1}^h c_g - np/2$. The partially maximized log likelihood (11) now follows since $|P_{\mathcal{S}_\alpha}\widehat{\Sigma}P_{\mathcal{S}_\alpha}|_0 = |\alpha^T\widehat{\Sigma}\alpha|$. Finally, since α, $\alpha^T\Sigma\alpha$, H and D uniquely determine Σ, it follows

that the maximum likelihood estimator of Σ is $\widehat{\Sigma}$

Proof of Corollary 1. Let L_A denote the log likelihood that depends on covariance matrices matrices $A\widetilde{\Sigma}_g A^T$. Then

$$\begin{aligned}\arg\max_{\mathcal{S}_\alpha} L_A(\mathcal{S}_\alpha) &= \arg\max_{\mathcal{S}_\alpha}\left\{-\sum_{g=1}^{h}\frac{n_g}{2}\log|\alpha^T A\widetilde{\Sigma}_g A^T\alpha| + \frac{n}{2}\log|\alpha^T A\widehat{\Sigma}A^T\alpha|\right\}\\ &= \arg\max_{A^{-T}\mathcal{S}_\beta} L(\mathcal{S}_\beta).\end{aligned}$$

And therefore $\arg\max L_A(\mathcal{S}_\alpha) = A^{-T}\arg\max L(\mathcal{S}_\alpha)$.

Proof of Proposition 4. Equation (12) is immediate. To show the second conclusion – $\mathcal{M} = \arg\max K_d(\mathcal{S})$ – let B_0 be a basis matrix for $\mathcal{M}^\perp$ and use Proposition A2 to write

$$\begin{aligned}K_b(\mathcal{S}) &= c + \frac{1}{2}\log|B_0^T\Sigma^{-1}B_0| - \sum_{g=1}^{h}\frac{f_g}{2}\log|B_0^T\Sigma_g^{-1}B_0| - \frac{1}{2}\log|\Sigma| + \sum_{g=1}^{h}\frac{f_g}{2}\log|\Sigma_g|\\ &\leq c - \frac{1}{2}\log|\Sigma| + \sum_{g=1}^{h}\frac{f_g}{2}\log|\Sigma_g|,\end{aligned}$$

where the inequality follows since $\log|B_0^T\Sigma^{-1}B_0|$ is a convex function of Σ. Using Proposition 1(i), we see that the upper bound is attained when B_0 is a basis matrix for $\mathcal{M}^\perp$.

References

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika* **89**, 159–182.

Burnham, K. & Anderson, D. (2002). *Model Selection and Multimodel Inference.* New York: Wiley.

Chikuse, Y. (2003). *Statistics on Special Manifolds.* New York: Springer.

Conway, J. B. (1990). *A Course in Functional Analysis, Second Edition.* New York: Springer.

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences*, pp. 18-25. Alexandria, VA: American Statistical Association.

Cook, R. D. (1998). *Regression Graphics.* New York: Wiley.

Cook. R.D., Li. B. & Chiaromonte, F. (2007). Dimension reduction without matrix inversion. *Biometrika* **94**, 569–584.

Eaton, M. (1983), *Multivariate Statistics.* New York: Wiley.

Flury, B. (1983). Some relations between the comparison of covariance matrices and principal component analysis. *Computational Statistics and Data Analysis* **1**, 97–109.

Flury, B. (1984). Common principal components in K groups. *J. Am. Statist. Assoc.* **79**, 892–898.

Flury, B. (1987). Two generalizations of the common principal component model. *Biometrika* **74**, 59–69.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models.* New York: Wiley.

Greub, W. (1981). *Linear Algebra.* New York: Springer.

Houle, D. Mezey, J. & Galpern, P. (2002). Interpretation of the results of common principal component analysis. *Evolution* **56**, 433–440.

Jensen, S. T. & Madsen, J. (2004). Estimation of proportional covariance matrices. *Ann. Statist.*, **32**, 219–232.

Liu, X., Srivastava, A. & Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 662-666.

Mezey, J. G. & Houle, D. (2003). Comparing G matrices: Are common principal components informative? *Genetics* **165**, 411–425.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* New York: Wiley.

Phillps, P. C. & Arnold, S. J. (1999). Hierarchical comparison of genetic variance-covariance matrices I. Using the Flury hierarchy. *Evolution* **53**, 1506-1515.

Rao, C. R. (1973) *Linear Statistical Inference and its Applications*, second ed. New York: Wiley.

Schott, J. R. (1991). Some tests for common principal component subspaces. *Biometrika* **75**, 229–236.

Schott, J. R. (1999). Partial common principal component subspaces. *Biometrika* **86**, 899–908.

Schott, J. R. (2003). Weighted chi-squared test for partial common principal component subspaces. *Biometrika* **90**, 411–421.