AN ABSTRACT OF THE THESIS OF

GARY JOE SEXTON (Name) for the degree DOCTOR OF PHILOSOPHY (Degree)

in STATISTICS (Major Department) presented on February 21, 1975 (Date)

Title: MULTIVARIATE DENSITY ESTIMATION--A BAYESIAN APPROACH

Redacted for privacy

Abstract approved: _____
Dr. H.D. Brunk

Methods of estimating a multivariate density over $\mathbb{R}^m$ are investigated from a Bayesian point of view.

The motivation for this thesis comes from the classification problem where the goal is to estimate the probability of class C_i, $i = 1, \ldots, N$ given an observed vector. Initial transformations of the data are introduced so that the components of the transformed data vector $\underset{\sim}{U}$ are uncorrelated. In this case the density of $\underset{\sim}{U}$ is close to multivariate normal. Under the assumption that both the density $p(u)$ and $\log p(u)$ have expansions in terms of an orthonormal basis, estimates are derived for the coefficients in the expansions. The approximation of the density is in terms of the best L_2 approximation.

When $p(u)$ is a multivariate normal density, estimating the leading coefficients in the expansion of $p(u)$ is equivalent to

estimating the covariance matrix. Estimating the corresponding coefficients in the expansion of $\log p(u)$ is equivalent to estimating the inverse of the covariance matrix. In this case, estimates of the covariance matrix are proposed. The "goodness" of these estimates is measured by a quadratic loss function.

Multivariate Density Estimation--A Bayesian Approach

by

Gary Joe Sexton

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1975

APPROVED:

Redacted for privacy

Professor of Statistics

in charge of major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Dean of Graduate School

Date thesis is presented February 21, 1975

Typed by Clover Redfern for Gary Joe Sexton

ACKNOWLEDGMENT

The author wishes to thank the many members of the Department of Statistics with whom he has had valuable course work and discussions. This would include Dr. Mark Lembersky and Dr. Charles Land.

The author wishes to express a special note of appreciation to Dr. Justus Seely and Dr. David Birkes for many stimulating and rewarding conversations during the time this thesis was being written.

Deepest appreciation is expressed to Professor H.D. Brunk who served as the author's major professor. Professor Brunk originally suggested the problem area to the author and gave guidance during its investigation. In addition to technical guidance, Professor Brunk gave abundantly of his time, encouragement, and above all, patience, without which this thesis would never have been completed.

The author is grateful for the financial support provided by the Department of Statistics through teaching assistantships and an N.I.H. Biometry traineeship.

Final thanks go to the author's wife, Mary Lou, and their children, Craig and Stacey, for their patience, understanding and moral support.

TABLE OF CONTENTS

MULTIVARIATE DENSITY ESTIMATION--A BAYESIAN APPROACH

I. INTRODUCTION

We initially consider the following two problems of practical importance. Simply stated they are:

(i) An investigator observes some grid, say $k \times k$ with k^2 lattice points. At each lattice point the investigator observes a number. This number could be an index number that is an index of greyness as in the data recorded by the NASA ERTS satellite and in a sense represents an aerial photograph. The investigator would like to identify the object of the photograph.

(ii) The second problem corresponds to the situation where several measurements are taken on an individual as in the case of diagnosing a disease.

In both of the above situations we are observing a random vector and the objective is to determine to which of several classes or populations the object of measurement belongs. In the first case we might be interested in differentiating urban areas from rural areas or differentiating between grain fields and grass fields for example. In the second case we might be interested in classifying a patient as to the type of disease he has.

In both of the above examples we assume that we are dealing with several classes which are denoted by $C_1, C_2, \ldots, C_N$. Having observed a random vector $X^T = (X_1, \ldots, X_k)$, the problem of fundamental interest is to determine to which of the N classes the object represented by X^T belongs. We would also like to give odds that realistically reflect the accuracy of our classification scheme. That is, we would like to know the probability of class C_i given the observed vector X which is denoted $P(C_i|X)$ $i = 1,2,\ldots,N$. If we know $P(C_i|X)$ for each i the the "posterior odds" for class C_i as opposed to class C_j will be $P(C_i|X)/P(C_j|X)$.

In the broadest sense it is the classification problem given in (i) and (ii) that motivated this thesis.

The determination of the quantity $P(C_i|X)$ may be approximated by a scalar multiple of the quantity $P(X|C_i)$. The emphasis of this thesis is the estimation of $P(X|C_i)$ where it is assumed that $P(X|C_i)$ is a multivariate continuous density for $i = 1, \ldots, N$.

Chapter II contains a discussion of certain initial transformations of the data vector that lead to convenient methods of reducing the number of components in the data vector. We also discuss the expansion of densities in terms of multidimensional Hermite polynomials, which form a complete orthonormal family on the m-dimensional Euclidean space $\mathbb{R}^m$. We assume that the unknown

density $p(u)$ has an expansion given by

$$p(u) = p_0(u) \sum_r \beta_r H_r(u)$$

where $p_0(u)$ is the multidimensional weight function for the Hermite polynomials $H_r(u)$. We also assume that $\log p(u)$ has a similar expansion so that

$$p(u) = p_0(u)\, e^{\Sigma_r \alpha_r H_r(u)} .$$

The details of these expansions are given in terms of certain model specifics.

Chapter III contains a discussion of Wishart and inverted Wishart distributions including properties that are important to the estimation problem of Chapter IV.

Chapter IV contains a discussion of estimating Fourier type coefficients of the expansions developed in Chapter II. The general approach is Bayesian in nature. We initially derive an estimator for the coefficients β_r using very general considerations. We also consider the problem of estimating the coefficients α_r in the expansion of $\log p(u)$. The problem of estimating α_r when p is multivariate normal is equivalent to estimating the precision matrix in a multivariate normal density while estimating the corresponding

β_r is equivalent to estimating the covariance matrix. In this case estimators for the covariance and precision matrices are derived.

The general problem of estimating densities by way of orthogonal expansions seems to be of fairly recent origin.

The earliest work known to the author was a paper by Čencov (1962) where he considered the Fourier type expansion of a univariate density in terms of an orthonormal basis. An unbiased estimate of the Fourier coefficents and properties of the corresponding density estimate are derived in terms of an L_2-norm.

Following the paper by Čencov, the orthogonal function approach was pursued by Schwartz (1967), Kronmal and Tarter (1968), Watson (1969) and Crain (1974). The emphasis in these papers is the estimation of univariate densities.

Schwartz addressed himself to the expansion of a univariate density in terms of Hermite functions and used an unbiased estimate of the coefficients. Under certain conditions he shows that his estimate is consistent with respect to mean squared error, $E[f(x)-f_n(x)]^2$, and mean integrated squared error, $E\int[f(x)-f_n(x)]^2dx$.

Kronmal and Tartar used standard Fourier series and derived trigonometric estimates along with their properties.

Watson used general orthogonal series to derive an estimate involving a weighting factor dependent upon sample size where the

weighting factor is chosen to minimize $E\int[f(x)-f_n(x)]^2dx$.

Crain studied extensively the estimation of univariate densities and distribution functions when both the density and the log of the density have expansions in terms of orthonormal functions over compact sets. In particular, normalized Legendre polynomials were considered.

It seems to the author that very little has been done by way of estimating multivariate densities by way of orthogonal functions from a Bayesian viewpoint.

One notable paper is that of Brunk and Pierce (1974). They develop estimators for a random vector with dichotomous components by way of an orthonormal basis. They assume that $p(x)$ and $\log p(x)$ both have expansions. The coefficients in each of these expansions is related through the notion of "near independence" of the components of X. The notion of "near independence" is represented through a prior distribution on the coefficients in the expansions and resulting estimates of these coefficients are derived.

An interesting paper on the estimation of a covariance matrix in a multivariate normal distribution using a Bayesian approach is found in Stein, Efron and Morris (1972). Stein considers estimators of the form $[aS^{-1}+(b/\text{tr } S)I]^{-1}$ where S is a matrix that is a function of the data. Stein shows that his estimate strictly dominates the maximum likelihood estimate, which is in the above class, when

the "loss" function is given by

$$L(\hat{\Sigma}, \Sigma) = \frac{\mathrm{tr}(\hat{\Sigma}^{-1} - \Sigma^{-1})S(\hat{\Sigma}^{-1} - \Sigma^{-1})}{k \, \mathrm{tr} \, \Sigma^{-1}} .$$

This "loss" function arises in a natural way from considerations found in a paper by Efron and Morris (1972).

We now introduce some of the notation used in this thesis. Other notation is defined as it is introduced.

We use the symbol := to mean the quantity on the left of the symbol is defined by the quantity on the right of the symbol. We use $\langle \cdot, \cdot \rangle$ to denote inner product. The symbol $(U|X) \sim$ will mean "U given X is distributed as" while the symbol $\dot{\sim}$ will mean "is approximately distributed as." We use the generic symbol Σ to denote a covariance matrix with σ_{ij} and σ^{ij} denoting the ij elements of Σ and Σ^{-1} respectively. We use the symbol $|\cdot|$ to denote the determinant of a matrix. T as a superscript will denote the transpose. The symbols $N_m(\mu, \Sigma)$ and $W_m(\nu, \Sigma)$ are used to denote the m-dimensional normal and Wishart distributions with mean μ and covariance Σ in the normal case and with ν denoting the "degrees of freedom" and Σ denoting the parametric matrix in the Wishart case. The term "precision" will refer to the reciprocal of the variance in the case of a univariate distribution and the inverse of the covariance matrix in the case of a multivariate distribution.

II. THE CLASSIFICATION PROBLEM--SPECIFICS

2.1 Initialization of the Data

If $X^T = (X_1, \ldots, X_n)$ represents a vector of observations then we know that the posterior probabilities for the various classes satisfy the relationship

$$P(C_i|X) \propto P(C_i)P(X|C_i) .$$

We will assume that there is sufficient information to give reliable estimates of the probabilities $P(C_i)$, $i = 1, \ldots, N$. Under this assumption we will be able to concentrate our attention on $P(X|C_i)$.

We also assume that population i has a covariance structure given by Σ_i, mean μ_i and the probability of class i occurring is p_i with $p_i > 0$.

The problem of estimating the mean has been studied extensively and many properties as well as the accuracy are known about the estimates. With this in mind, we will initially concentrate on the problem of estimating other structure acting as though the means were known.

Prior to discussing the estimation problem we will discuss representation of the data.

If the random vector $\underset{\sim}{X}$ has k components and k is large, it will be desirable in many situations to reduce the number of

components. In order to perform this reduction in some optimal manner it will be convenient to initially transform the vector $\underset{\sim}{X}$.

Let us first suppose that the data is centered in the usual way by the transformation $\underset{\sim}{Y} = \underset{\sim}{X} - \mu_i$ for each of the N classes.

The formula $P_{\underset{\sim}{X}}(X|C_i) = P_{\underset{\sim}{Y}}(Y+\mu_i|C_i)$ will allow us to concentrate on the centered data $\underset{\sim}{Y}$ and relate the results to the original problem.

The joint covariance structure for the N classes can be expressed in terms of the marginal covariance structure by way of the following formula.

$$
\begin{aligned}
K(t,s) &:= E(\underset{\sim}{Y}_s \underset{\sim}{Y}_t) - E(\underset{\sim}{Y}_s)E(\underset{\sim}{Y}_t) \\
&= E\{E(\underset{\sim}{Y}_s \underset{\sim}{Y}_t)|C_i\} - E\{E(\underset{\sim}{Y}_s|C_i)\}E\{E(\underset{\sim}{Y}_t|C_i)\} \\
&= E\{\mathrm{cov}(\underset{\sim}{Y}_s, \underset{\sim}{Y}_t|C_i) + E(\underset{\sim}{Y}_s|C_i)E(\underset{\sim}{Y}_t|C_i)\} \\
&\qquad - E\{E(\underset{\sim}{Y}_s|C_i)\}E\{E(\underset{\sim}{Y}_t|C_i)\} \\
&= E\{\mathrm{cov}(\underset{\sim}{Y}_s, \underset{\sim}{Y}_t)|C_i\} \\
&= \sum_{i=1}^{N} E(\underset{\sim}{Y}_s \underset{\sim}{Y}_t|C_i) .
\end{aligned}
\tag{2.1}
$$

Let Σ denote the joint covariance matrix of the N classes.

We can represent $\underset{\sim}{Y}$ without error by an expansion of the form $\underset{\sim}{Y} = \psi \underset{\sim}{V}$ where ψ is a deterministic $k \times k$ matrix of

rank k and $\underset{\sim}{V}$ is a random vector. From this exact expansion we may write $\underset{\sim}{V}$ as $\underset{\sim}{V} = \psi^{-1}\underset{\sim}{Y}$ so that $\underset{\sim}{V}$ is simply a transformation of $\underset{\sim}{Y}$.

For the general classification problem, it will be desirable, if we are going to transform the data, to preserve the original structure as much as possible. In particular, it is especially desirable to preserve distance. With this in mind we state the following known result from linear theory.

Lemma 2.1.1. If $\psi: \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation then $\|x\| = \|\psi x\|$ if and only if ψ is orthogonal.

Proof: See Smith (1971). ∎

In our situation the primary implication of Lemma 2.1.1 is that we shall restrict ourselves to those transformations ψ which are orthonormal. That is, those transformations ψ for which $\psi^T\psi = I$. Additionally, by restricting our attention to orthogonal transformations, the distributions are related in the following simple manner:

$$\begin{aligned} F_{\underset{\sim}{V}}(v) &= P(\underset{\sim}{V} \leq v) \\ &= P(\psi^T\underset{\sim}{Y} \leq v) \\ &= P(\underset{\sim}{Y} \leq \psi^T v) \\ &= F_{\underset{\sim}{Y}}(\psi^T v). \end{aligned}$$

Under this restriction we have $\underset{\sim}{Y}$ represented without error by the transformation $\underset{\sim}{Y} = \psi\underset{\sim}{V}$ where ψ is orthonormal and in this case $\underset{\sim}{V} = \psi^T\underset{\sim}{Y}$ and $\underset{\sim}{V}$ is simply an orthonormal transformation of $\underset{\sim}{Y}$.

It should be noted that the covariance structure of $\underset{\sim}{V}$ is given by:

$$\text{cov } \underset{\sim}{V} = \text{cov}(\psi^T\underset{\sim}{Y}) = \psi^T(\text{cov } \underset{\sim}{Y})\psi = \psi^T\Sigma\psi$$

From among the many orthonormal transformations we single out the Karhunen-Loeve transformation which will be denoted by κ. In matrix form, the columns of κ will consist of the k eigenvectors of Σ with the ith column, κ_i, being the eigenvector associated with the eigenvalue λ_i. Moreover, the eigenvalues have been ordered so that they satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

In using the transformation κ we now have that $\text{cov } \underset{\sim}{V} = \kappa^T\Sigma\kappa = D$ where D is the diagonal matrix with $D_{ii} = \lambda_i$, $i = 1, 2, \ldots, k$. By rotating (or reflecting) $\underset{\sim}{Y}$ in this manner we have arrived at a vector $\underset{\sim}{V}$ that has the property that the components of $\underset{\sim}{V}$ are uncorrelated and in this sense we say they are nearly independent.

We make special note of the Karhunen-Loève transformation in as much as it has certain optimal properties as we shall see. Moreover, it is the transformation that we will use.

In order to justify our choice of the Karhunen-Loeve transformation we return to the situation where ψ is an arbitrary orthonormal transformation so that $\underset{\sim}{Y} = \psi\underset{\sim}{V}$ and $\underset{\sim}{V} = \psi^T\underset{\sim}{Y}$.

In the language of pattern recognition, the components of $\underset{\sim}{V}$ are called features or indicants representing the vector $\underset{\sim}{Y}$. As commented earlier, if the number of features, k, is large, it will be desirable in certain situations to reduce the number of features used to represent the vector $\underset{\sim}{Y}$. That is, suppose we only determine m of the components of $\underset{\sim}{V}$. Moreover, we may assume that it is the first m components that are determined.

If we replace each of the indicants $\underset{\sim}{V}_i$ with a constant b_i, $i = m+1, \ldots, k$, we will have an approximate representation of $\underset{\sim}{Y}$. In introducing this approximation we wish to do so in a manner that will minimize the mean square error in the representation.

For notation, set $\underset{\sim}{V}_1 := (v_1, \ldots, v_m)^T$, $\underset{\sim}{V}_2 := (v_{m+1}, \ldots, v_k)^T$, $b := (b_{m+1}, \ldots, b_k)^T$, $\underset{\sim}{Y}_b = \psi\begin{pmatrix}\underset{\sim}{V}_1 \\ b\end{pmatrix} = \sum_{i=1}^{m} \psi_i \underset{\sim}{V}_i + \sum_{i=m+1}^{k} \psi_i b_i$ where ψ_i is the ith column of ψ. Also, suppose we partition ψ as $(\psi_1 \,\vdots\, \psi_2)$ where ψ_2 is the last $k-m$ columns of ψ.

We now have the following relationships among the partitioned matrices:

$$V = \begin{pmatrix}V_1 \\ V_2\end{pmatrix} = \begin{pmatrix}\psi_1^T \\ \cdots \\ \psi_2^T\end{pmatrix} Y$$

$$\operatorname{cov} \underset{\sim}{Y} = \begin{pmatrix} \psi_1^T \\ \hline \psi_2^T \end{pmatrix} \Sigma (\psi_1 \,\vdots\, \psi_2) = \left(\begin{array}{c|c} \psi_1^T \Sigma \psi_1 & \psi_1^T \Sigma \psi_2 \\ \hline \psi_2^T \Sigma \psi_1 & \psi_2^T \Sigma \psi_2 \end{array}\right)$$

$$\underset{\sim}{V}_2 = (O \,\vdots\, I_{k-m}) \begin{pmatrix} \underset{\sim}{V}_1 \\ \underset{\sim}{V}_2 \end{pmatrix} = (O \,\vdots\, I_{k-m}) \begin{pmatrix} \psi_1^T \\ \hline \psi_2^T \end{pmatrix} \underset{\sim}{Y} = \psi_2^T \underset{\sim}{Y}$$

$$\operatorname{cov} V_2 = \psi_2^T \Sigma \psi_2$$

$$\psi_2^T \psi_2 = I_{k-m}$$

The following known result gives us the proper choice of b_i, $i = m+1, \ldots, k$ to minimize the mean square error in the representation.

Lemma 2.1.2. Let ψ be an arbitrary orthogonal transformation and define $\underset{\sim}{V}$ by $\underset{\sim}{V} := \psi^T \underset{\sim}{Y}$ so that $\underset{\sim}{Y} = \psi \underset{\sim}{V}$. Set $\underset{\sim}{Y}_b = \psi \begin{pmatrix} \underset{\sim}{V}_1 \\ b \end{pmatrix}$ where b is a $(k-m) \times 1$ vector of constants so that $\underset{\sim}{Y}_b$ approximates $\underset{\sim}{Y}$. The choice of b that minimizes the mean square error in the approximation is $b = E\underset{\sim}{V}_2$.

Proof: The mean square error in the approximation is given by $E\|\underset{\sim}{Y} - \underset{\sim}{Y}_b\|^2$ where $\|\cdot\|$ denotes the usual Euclidean norm. If we set $\underset{\sim}{\varepsilon}_\psi := \underset{\sim}{Y} - \underset{\sim}{Y}_b$ we then have the following:

$$E\|\underset{\sim}{Y}-\underset{\sim}{Y}_b\|^2 = E\|\underset{\sim}{\xi}_\psi\|^2 = E\,\underset{\sim}{\xi}_\psi^T\underset{\sim}{\xi}_\psi = \operatorname{tr} E\,\underset{\sim}{\xi}_\psi\underset{\sim}{\xi}_\psi^T$$

$$= \sum_{i=m+1}^{k} E(\underset{\sim}{V}_i-b_i)^2 .$$

Each term in the sum is minimum for $b_i = E\underset{\sim}{V}_i$. ∎

We should note that in our case $E\underset{\sim}{V}_i = 0$ so we set $b_i = 0$, $i = m+1,\ldots,k$. We note that $\underset{\sim}{\xi}_\psi = \psi\begin{pmatrix}0\\ \underset{\sim}{V}_2-b\end{pmatrix}$ and in general, when $b = E\underset{\sim}{V}_2$ so that $\underset{\sim}{\xi}_\psi\begin{pmatrix}0\\ \underset{\sim}{V}_2-E\underset{\sim}{V}_2\end{pmatrix} = \psi_2(\underset{\sim}{V}_2-E\underset{\sim}{V}_2)$, we have:

$$E\|\underset{\sim}{\xi}_\psi\|^2 = \sum_{i=m+1}^{k} E(\underset{\sim}{V}_i-E\underset{\sim}{V}_i)^2 = \sum_{i=m+1}^{k} \operatorname{var}\underset{\sim}{V}_i$$

$$= \operatorname{tr} E\,\underset{\sim}{\xi}_\psi\underset{\sim}{\xi}_\psi^T$$

$$= \operatorname{tr}\,\psi_2\{E(\underset{\sim}{V}_2-E\underset{\sim}{V}_2)(\underset{\sim}{V}_2-E\underset{\sim}{V}_2)^T\}\psi_2^T$$

$$= \operatorname{tr}\,\psi_2(\operatorname{cov}\underset{\sim}{V}_2)\psi_2^T$$

$$= \operatorname{tr}\operatorname{cov}\underset{\sim}{V}_2$$

$$= \operatorname{tr}\,\psi_2^T\Sigma\psi_2 .$$

Having chosen the vector b to minimize $E\|\underset{\sim}{\xi}_\psi\|^2$ for any fixed orthogonal transformation ψ the next problem is to select an optimal transformation. The selection of an optimal transformation will rest upon the following result attributed to Poincaré.

For notation, if Q is any matrix we will denote the ith ordered eigenvalue of Q by $\lambda_i(Q)$ with $\lambda_i(Q) \geq \lambda_{i+1}(Q)$.

Lemma 2.1.3. Let A be a $k \times k$ symmetric matrix with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ its eigenvalues. Let B be a $k \times q$ matrix such that $B^T B = I_q$. It then follows that

$$\lambda_i(B^T AB) \leq \lambda_i(A), \quad i = 1, \ldots, q,$$

and

$$\lambda_{q-j}(B^T AB) \geq \lambda_{k-j}(A), \quad j = 0, 1, \ldots, q-1$$

Proof: See Rao (1965). ∎

The following known result will follow immediately from Lemma 2.1.3.

Under the hypothesis of Lemma 2.1.2, if we set $b_i = EV_i$ for $i = m+1, \ldots, k$ then the problem

$$\underset{\psi}{\text{minimize}}\ E\| \underset{\sim}{\xi}_{\psi}\|^2$$

has a solution given by $\psi = \kappa$ where κ is the Karhunen-Loève transformation as described on page 10.

Proof: As noted above, for any orthogonal transformation ψ,

$$E\| \underset{\sim}{\xi}_{\psi} \|^2 = \text{tr } \psi_2^T \hat{\Sigma} \psi_2 .$$

It, therefore, suffices to show that

$$\text{tr } \psi_2^T \hat{\Sigma} \psi_2 \geq \text{tr } \kappa_2^T \hat{\Sigma} \kappa_2 .$$

Note that ψ_2 is $k \times (k-m)$ and $\psi_2^T \psi_2 = I_{k-m}$ so that we may invoke Lemma 2.1.3 to conclude that

$$\lambda_{k-m-j}(\psi_2^T \hat{\Sigma} \psi_2) \geq \lambda_{k-j}(\hat{\Sigma}) \geq 0 \quad j = 0, 1, \ldots, k-m-1.$$

Therefore,

$$\sum_{j=0}^{k-m-1} \lambda_{k-m-j}(\psi_2^T \hat{\Sigma} \psi_2) \geq \sum_{j=0}^{k-m-1} \lambda_{k-j}(\hat{\Sigma}).$$

Now, $\lambda_{k-j}(\hat{\Sigma}) = \lambda_{k-j}$ so that

$$\sum_{j=0}^{k-m-1} \lambda_{k-j}(\hat{\Sigma}) = \sum_{j=0}^{k-m-1} \lambda_{k-j} = \sum_{i=m+1}^{k} \lambda_i = \text{tr } \kappa_2^T \hat{\Sigma} \kappa_2 .$$

That is, the sum of the eigenvalues, hence the trace, of $\psi_2^T \hat{\Sigma} \psi_2$ is bounded below by $\text{tr } \kappa_2^T \hat{\Sigma} \kappa_2$ and the result is established. ∎

Before proceeding we present here a brief summary of the initialization process. The primary function of the initialization process is to obtain a convenient coordinate system with which to

represent the data. We have chosen to use the Karhunen-Loève coordinate system. To use this system we require the knowledge of Σ, the overall covariance structure of the N populations. There are two basic viewpoints that we have considered to obtain an estimate of Σ.

In the first case we assume that we have some reliable estimate of p_i for $i = 1, 2, \ldots, N$. We also have a training set selected at random from each of the N populations. From each of these training sets we estimate μ_i in the event μ_i is unknown. We also estimate Σ_i from the sample training set corresponding to population i. We then obtain an estimate of Σ using (2.1).

In the second case we consider one sample from the combined training sets. After obtaining the overall sample we are able to identify to which population each observation belongs. We now estimate Σ from the overall sample. Moreover, we are able to directly obtain estimates of p_i by taking p_i to be the proportion of times that we observe an observation from population i.

Inasmuch as the second case allows us to perform all estimations from the one overall sample we have focused on the second viewpoint in this thesis.

Since the fundamental purpose of our estimate of Σ is to establish an initial coordinate system, the accuracy of this estimate does not seem to be critical. This was found to be the case in an

analogous situation by Martin and Bradley (1972).

With this in mind we are content to use the usual estimate of Σ. That is, we set $\overline{x_i x_j} := \frac{1}{n}\sum_{t=1}^{n} {}^t x_i \, {}^t x_j$, $\overline{x_i} := \frac{1}{n}\sum_{t=1}^{n} {}^t x_i$ and estimate σ_i by $\overline{x_i x_j} - \overline{x_i}\,\overline{x_j}$. In matrix form,

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} {}^k X \, {}^k X^T - \bar{X}\,\bar{X}^T .$$

Given that we have an estimate of Σ we find the Karhunen-Loève transformation κ whose columns will be the eigenvectors of Σ arranged in decreasing order by the magnitudes of the associated eigenvalues.

Now suppose we observe a random vector $\underset{\sim}{X}$ from the population at large. For fixed i, we center this observation and set $\underset{\sim}{Y} = \underset{\sim}{X} - \mu_i$. We now transform $\underset{\sim}{Y}$ by κ^T to obtain a transformed observation $\underset{\sim}{V}$ where $\underset{\sim}{V} = \kappa^T \underset{\sim}{Y}$. We now have an observation $\underset{\sim}{V}$ and the components $\underset{\sim}{V}$ are uncorrelated.

2.2 Expansion of Densities

From this point on, we shall assume that all of the previous initializations have been performed. We will now denote $\underset{\sim}{Y}_b$ as simply $\underset{\sim}{Y}$ and assume $\underset{\sim}{Y}$ is an $m \times 1$ random vector, κ is the $k \times m$ orthonormal transformation consisting of the eigenvectors

of Σ. $\underset{\sim}{V} = \kappa^T \underset{\sim}{Y}$ so that the feature vector $\underset{\sim}{V}$ has its components uncorrelated.

We shall concentrate on estimating the density $p(V|C_i)$ for fixed i which we will denote simply by $p(v)$. The estimation procedure will be to represent $p(v)$ by an expansion of Fourier type in terms of a complete orthonormal family and estimating the coefficients in the expansion.

We shall outline the general procedure in the univariate case and then proceed to the details of our model.

2.3 Orthonormal Families, Fourier Expansions and Hermite Polynomials

In general, let us assume that P is a probability measure on the Borel subsets $\mathcal{B}^1$ of $\mathbb{R}^1$ and P is absolutely continuous with respect to some measure ℓ so that its density is given by $p = \frac{dP}{d\ell}$. Let P_0 be any measure on $(\mathbb{R}^1, \mathcal{B}^1)$ that is absolutely continuous with respect to the measure ℓ, so that its density is $p_0 = \frac{dP_0}{d\ell}$.

Consider the Hilbert space, $L_2(\mathbb{R}^1, \mathcal{B}^1, P_0)$ with the usual inner product and norm given by $\langle f, g \rangle := \int fg dP_0$ and $\|f\|^2 = \int |f|^2 dP_0$. Let $\{T_s : s \in S\}$ be a family of functions from $\mathbb{R}^1$ to $\mathbb{R}^1$ that is orthonormal with respect to P_0 so that $\langle T_s, T_r \rangle = \delta_{r,s}$ where $\delta_{r,s}$ is the Kronecker delta. Moreover,

if it happens that $\{T_s : s \in S\}$ is complete, that is, $\langle f, T_s \rangle = 0$ for every $s \in S$ implies $f = 0$, then any function $f \in L_2(P_0)$ has an expansion of the form

$$f = \Sigma_s \langle f, T_s \rangle T_s$$

If we restrict our attention to densities p such that $\frac{p}{p_0}$ and $\log \frac{p}{p_0}$ are members of $L_2(P_0)$ then we have representations for p given by

$$p(v) = p_0(v)\Sigma_s \beta_s T_s(v)$$

and

$$p(v) = p_0(v)e^{\Sigma_s \alpha_s T_s(v)}$$

Regarding the terms β_s and α_s as parameters, the problem of estimating p is one of estimating the parameters in at least one of the above expansions.

A class of functions that we will find useful is the class of Hermite polynomials. Let h be the weighting function on $\mathbb{R}^1$ given by $h(v) = e^{-v^2/2}$ and set $g(v) = \frac{1}{\sqrt{2\pi}} h(v)$ so that $g(v)$ may be regarded as the density of a standardized normal variate.

The Hermite polynomials are defined in terms of the above functions. For $n = 0$, set $H_0(v) := 1$. For $n = 1, 2, 3, \ldots$, set $H_n(v) := \frac{(-1)^n}{\sqrt{n!}} e^{v^2/2} h^{(n)}(v)$ where $h^{(n)}(v) := \frac{d^n}{dv^n} h(v)$. $H_n(v)$ is

called the nth order normalized Hermite polynomial. We should note here that we are considering the space $L_2(\mathbb{R}^1, \mathcal{B}^1, P_0)$ where P_0 is determined by $g = \frac{dP_0}{d\ell}$ and ℓ is Lebesgue measure.

Let m and n be integers and assume $m \leq n$. We then have:

$$<H_m, H_n> = \int H_m(v)H_n(v)g(v)dv = \frac{1}{\sqrt{2\pi}} \int e^{-v^2/2} H_m(v)H_n(v)dv$$

$$= \frac{(-1)^n}{\sqrt{n!}} \int H_m(v)g^{(n)}(v)dv$$

Integrating by parts m times gives:

$$<H_m, H_n> = \frac{(-1)^{n+m}}{\sqrt{n!}} \int H_m^{(m)}(v)g^{(n-m)}(v)dv$$

If $m < n$, then $n-m > 0$ and one more integration by parts will give $<H_m, H_n> = 0$ since H_m is a polynomial of degree m and we will be taking the $(m+1)$st derivative.

If $m = n$ then $m-n = 0$ and we have

$$<H_m, H_n> = \int [H_n(v)]^2 g(v)dv$$

$$= \frac{1}{\sqrt{n!}} \int H_n^{(n)}(v)g(v)dv$$

$H_n(v)$ is a polynomial of degree n with leading coefficient being $\frac{1}{\sqrt{n!}}$. In this case $H_n^{(n)} = \frac{n!}{\sqrt{n!}} = \sqrt{n!}$ and we have

$$\langle H_m, H_n \rangle = \int g(v)dv = 1.$$

From this argument we can see that $\langle H_m, H_n \rangle = \delta_{m,n}$ for $m, n = 0, 1, 2, \ldots$ and the normalized Hermite polynomials are orthonormal on $\mathbb{R}^1$ with respect to the density $g(v) = \frac{1}{\sqrt{2\pi}} e^{-v^2/2}$. It is also true that $\{H_n : n = 0, 1, \ldots\}$ is complete with respect to g (see Burrill (1972)).

The first few normalized polynomials are:

$$H_0(v) = 1 \qquad H_3(v) = \frac{1}{\sqrt{6}}(v^3 - 3v)$$

$$H_1(v) = v \qquad H_4(v) = \frac{1}{\sqrt{24}}(v^4 - 6v^2 + 3)$$

$$H_2(v) = \frac{1}{\sqrt{2}}(v^2 - 1) \qquad H_5(v) = \frac{1}{\sqrt{120}}(v^5 - 10v^3 + 5v)$$

2.4 The Model Specifics

In constructing this model we assume that we are dealing with a random vector $\underset{\sim}{V}$ defined on $\mathbb{R}^m$ and that P is some probability measure on $(\mathbb{R}^m, \mathcal{B}^m)$ associated with $\underset{\sim}{V}$. Additionally, we assume that P is absolutely continuous with respect to some measure ℓ on $(\mathbb{R}^m, \mathcal{B}^m)$ so that the joint density function

is given by $p = \frac{dP}{d\ell}$.

Consider the normalized Hermite polynomials described previously which form a complete orthonormal family on $\mathbb{R}^1$ with respect to the density $\frac{1}{\sqrt{2\pi}} e^{-v^2/2}$ when ℓ is Lebesgue measure. The following known theorem will allow us to find a suitable orthonormal family that is complete on $\mathbb{R}^m$.

Lemma 2.3.1. Suppose $\{X_m\}$ and $\{Y_n\}$ are complete orthonormal families with respect to probability measures μ and ν. It then follows that the product set $\{X_m Y_n\}$ is a complete orthonormal family on the product space with respect to the product measure $\mu \times \nu$.

Proof: See Lancaster (1969) or Zygmund (1959). ∎

By applying this theorem and letting $\{X_m\} = \{Y_m\} = \{H_m\}$ we have, extending by induction, that all finite products consisting of m terms of Hermite polynomials forms a complete orthonormal system on $\mathbb{R}^m$ with respect to the measure P_0 whose density is given by

$$p_0(v) = \frac{dP_0}{d(\ell_1 \times \ldots \times \ell_n)} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda_i}} e^{-v_i^2/2\lambda_i}.$$

Let $B = \{r = (r_1, \ldots, r_m) : r_i = 0, 1, 2, \ldots, i = 1, \ldots, m\}$. For $r \in B$, a non zero entry in the jth position will indicate the

presence of a Hermite polynomial whose argument is the jth feature and the value of r_j is the degree of the polynomial. With this indexing scheme we can denote the complete orthonormal family consisting of m-products of Hermite polynomials. Writing $u_i = v_i/\sqrt{\lambda_i}$ we have that $\{H_r(u) : r \in B\}$ is a complete orthonormal family on $(\mathbb{R}^m, \mathcal{B}^m, P_0)$ where P_0 has a density with respect to the measure ℓ on $\mathbb{R}^m$ given by

$$p_0(u) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-u_i^2/2}$$

We assume that $\frac{p(u)}{p_0(u)}$ and $\log \frac{p(u)}{p_0(u)}$ have expansions given by:

$$p(u) \simeq p_0(u)\Sigma_r \beta_r H_r(u) \tag{2.2}$$

$$p(u) = p_0(u) e^{\Sigma_r a_r H_r(u)} \tag{2.3}$$

A statistician, when faced with a real problem, will formulate a model that he believes most accurately reflects the true situation. In attempting the analysis he may encounter many mathematical and/or computational difficulties that are not easily, if ever, overcome. He may reject his model on the grounds that it is of little practical value if he cannot carry out the analysis. He then may reformulate the problem from an entirely different viewpoint perhaps

encountering the same or similar problems. Alternatively, he may hold with his initial viewpoint but adopt simplifying assumptions in the model to allow for the analysis. In either case, his final model is a formulation of the problem that most accurately reflects the true situation subject to the condition that the analysis is possible.

When faced with the problem of estimating multivariate densities two simplifying assumptions come immediately to mind.

(i) The components of the vector $\underset{\sim}{U}$ are independent.

(ii) The vector $\underset{\sim}{U}$ is multivariate normal.

Both of these assumptions are implicit in the use of P_0 as a distribution for $\underset{\sim}{U}$ and we will later see that assumption (ii) may be reflected in (2.3) by setting certain coefficients equal to zero.

Before proceeding to the details of the expansion, a final comment is in order concerning expansions (2.2) and (2.3). Both expansions involve expressions of the form $\Sigma_r \beta_r H_r(u)$ where r is an m-component vector. There are several ways to interpret this sum.

Perhaps the most natural way is to write

$$\Sigma_r \beta_r H_r(u) := \sum_{r_1=0}^{\infty} \cdots \sum_{r_m=0}^{\infty} \beta_{(r_1,\ldots,r_m)} H_{(r_1,\ldots,r_m)}(u).$$

For our purposes it will be more convenient to interpret the

sum in the following manner. Set

$$n(r) := \sum_{i=1}^{m} r_i$$

for $r \in B$ and we will call $n(r)$ the rank of r. We will also say that the coefficients of $H_r(u)$ in (2.1) and (2.2) have rank $n(r)$. We now think of our countable collection of basis functions as being grouped by blocks according to increasing rank with an arbitrary ordering within blocks. That is, block zero contains $H_r(u)$ for which $n(r) = 0$, block one contains $H_r(u)$ for which $n(r) = 1$ with arbitrary ordering within the block and so on. We now have the functions $H_r(u)$ ordered by rank but arbitrary within rank with the result that the coefficients will be ordered similarly.

We now write:

$$\Sigma_r \beta_r H_r(u) := \sum_{\nu=0}^{\infty} \sum_{r:n(r)=\nu} \beta_r H_r(u)$$

2.5 Details of the Expansion

We develop our expansion in a manner analogous to that in Brunk and Pierce (1974).

Suppose we denote $(0, 0, \ldots, 0) \in B$ as simple 0 and then

from (2.3),

$$1 = \int p(u)du$$

$$= \int p_0(u)e^{\Sigma_r a_r H_r(u)}\, du$$

from which we can write

$$e^{-a_0} = \int p_0(u)e^{\Sigma_{r\neq 0} a_r H_r(u)}\, du. \tag{2.4}$$

Let a denote a vector whose components are the coefficients a_s for which $s \neq 0$. Where a is specified then a_0 is determined by the requirement given in (2.4). Setting $A(a) = -a_0$ we have defined a function $A(\cdot)$ whose domain is a set of vectors. We can now rewrite (2.3) as

$$p(u) = p_0(u)e^{\Sigma_{r\neq 0} a_r H_r(u) - A(a)}. \tag{2.5}$$

Recalling that $H_0(u) = 1$ and that

$$E[H_r(u)H_s(u)\,|\,p_0] = \langle H_r(u), H_s(u)\rangle$$

$$= \delta_{r,s}$$

and using the dominated convergence theorem in conjunction with (2.4) we can write

$$\frac{\partial}{\partial a_s} e^{A(a)} = \frac{\partial A(a)}{\partial a_s} e^{A(a)}$$

$$= \frac{\partial}{\partial a_s} \int p_0(u) e^{\sum_{r\neq 0} a_r H_r(u)} du$$

$$= \int \frac{\partial}{\partial a_s} p_0(u) e^{\sum_{r\neq 0} a_r H_r(u)} du$$

$$= \int p_0(u) e^{\sum_{r\neq 0} a_r H_r(u)} H_s(u) du$$

so that

$$\frac{\partial}{\partial a_s} A(a) = \int p_0(u) e^{\sum_{r\neq 0} a_r H_r(u) - A(a)} H_s(u) du \,. \tag{2.6}$$

Using (2.3),

$$\frac{\partial}{\partial a_s} A(a) = \int H_s(u) p(u) du$$

$$= E[H_s(u) | p] \,, \quad \text{provided} \quad s \neq 0. \tag{2.7}$$

By writing

$$\int H_s(u) p(u) du = \int H_s(u) \frac{p(u)}{p_0(u)} p_0$$

and using (2.2), we have

$$\frac{\partial A(a)}{\partial a_s} = \int H_s(u) \sum_r \beta_r H_r(u) p_0(u) du \tag{2.8}$$

$$= \sum_r \beta_r \int H_s(u) H_r(u) p_0(u) du \;=$$

$$= \sum_r \beta_r E[H_s(u)H_r(u)|p_0(u)]$$

$$= \sum_r \beta_r \delta_{r,s}$$

$$= \beta_s \qquad \text{provided} \quad s \neq 0.$$

Similarly, finding the second partials of $A(\alpha)$ gives

$$(2.9) \qquad e^{A(\alpha)}\left[\frac{\partial^2 A(\alpha)}{\partial \alpha_s \partial \alpha_t} + \frac{\partial A(\alpha)}{\partial \alpha_s}\frac{\partial A(\alpha)}{\partial \alpha_t}\right]$$

$$= \int p_0(u) e^{\sum_{r\neq 0} \alpha_r H_r(u)} H_s(u)H_t(u)du$$

so that

$$(2.10) \qquad \frac{\partial^2 A(\alpha)}{\partial \alpha_s \partial \alpha_t} + \frac{\partial A(\alpha)}{\partial \alpha_s}\frac{\partial A(\alpha)}{\partial \alpha_t} = E[H_s(u)H_t(u)|p]$$

and

$$(2.11) \qquad \frac{\partial^2 A(\alpha)}{\partial \alpha_s \partial \alpha_t} = \mathrm{cov}[H_s(u), H_r(u)|p] .$$

From (2.3) and (2.5),

a. If $\alpha = 0$ then $A(\alpha) = A(0) = 0$

b. If $s \neq 0$ then

$$(2.12) \qquad \frac{\partial A(0)}{\partial \alpha_s} = \int p_0(u)H_s(u)du = \int p_0(u)H_s(u)H_0(u)du$$

$$= E[H_s(u)H_0(u)|p_0]$$

$$= \delta_{s,0}$$

$$= 0$$

c. If $s \neq 0$ then

$$\frac{\partial^2 A(0)}{\partial a_s^2} = E[H_s(u)H_s(u)|p_0]$$
$$= 1$$

If $t \neq 0$ and $t \neq s$ then

$$\frac{\partial^2 A(0)}{\partial a_s \partial a_t} = E[H_s(u)H_t(u)|p_0]$$
$$= 0$$

That is,

$$\frac{\partial^2 A(0)}{\partial a_s \partial a_t} = \delta_{s,t} \quad \text{if} \quad s \neq 0, \quad t \neq 0 . \tag{2.13}$$

As an aside we note here that by (2.7) and (2.8) the coefficients of rank one in (2.2) all vanish. That is, if $n(r) = 1$ then $H_r(u) = H_1(u_i)$ for some i and

$$\beta_r = E[H_r(u)|p]$$
$$= E[H_1(u_i)|p]$$
$$= E[u_i|p]$$
$$= 0 .$$

Recall that in the general classification problem Σ represents the covariance structure of the mixture of all classes. In fact, Σ will be taken to be the usual estimate of the mixture covariance

in case the individual covariance structures are not known. Recall also that U is a transformation of our original vector so that the components are uncorrelated.

If we were initially to assume that $\underset{\sim}{Y}$ is $N(0, \Phi)$ then $P_0(u)$ would be the distribution resulting from the application of κ to $\underset{\sim}{Y}$. That is, $N(0,I)$ would be an initial estimate of the distribution of $\underset{\sim}{U}$.

With this in mind we will be interested in the situation where the components of $\underset{\sim}{U}$ are nearly independent. If we interpret the parameters a_r as interactions and assume that interactions of high order are relatively small then we may interpret near independence in terms of a_r of rank larger than one being near zero.

In this context we will consider the Maclaurin expansion of the function $A(a)$.

$$A(a) = A(0) + \sum_{r \neq 0} a_r^2 + \ldots$$

$$\doteq \sum_{r \neq 0} a_r^2 + \ldots$$

From (2.7) and (2.8) we have $\beta_s = \frac{\partial A(a)}{\partial a_s} = E[H_s(u)|p]$ and β is a function of a through p. We shall denote this fact by writing β as $\beta(a)$. Moreover,

$$\beta_0(a) = E[H_0(u)|p] = 1$$

and

$$\beta_s(0) = \frac{\partial A(0)}{\partial a_s} = 0 \quad \text{by (2.12).}$$

These two facts give $\beta_s(0) = \delta_{s,0}$.

On the other hand, from (2.8) we have

$$\frac{\partial \beta_s(a)}{\partial a_t} = \frac{\partial^2 A(a)}{\partial a_s \partial a_t} = \operatorname{cov}[H_s(u), H_t(u) \mid p]$$

and

$$\frac{\partial \beta_s(0)}{\partial a_t} = \delta_{s,t} \quad \text{for } s \neq 0,\ t \neq 0.$$

If we consider the expansion of $\beta_s(a)$ about zero,

$$\beta_s(a) = \beta_s(0) + a_s + \text{higher order terms}$$

$$= a_s + \text{higher order terms}$$

From this we can infer that when the a's are near zero then $\beta_s \doteq a_s$ if $s \neq 0$.

In view of the two basic equations (2.2) and (2.3), the estimation of the density p involves the estimation of the parameters β_r in (2.2) and/or the estimation of the parameters a_r in (2.3). In either case our aim is to derive some suitable estimate of p.

III. DISTRIBUTION THEORY

In this chapter we will give some results in distribution theory that will be useful in the parameter estimation problem of Chapter IV. For our purposes, the distributions of most interest are distributions from families of Wishart and Inverted Wishart distributions.

Generally, the results of this chapter may be found in various texts such as Anderson (1958) and Press (1972). The proofs supplied are believed by the author to be his own and are presented here for the sake of completeness.

Selected results concerning characteristic functions and the differentiation of matrices may be found in the Appendix.

3.1 The Wishart Distribution

Historically, the derivation of the Wishart distribution originates from observing a sample of size ν from the distribution of an m-component random vector $\underset{\sim}{Z}$ that is $N_m(0, \Sigma)$ and considering the random matrix

$$\underset{\sim}{V} := \sum_{a=1}^{\nu} \underset{\sim}{Z}_a \underset{\sim}{Z}_a^T .$$

The random matrix $\underset{\sim}{V}$ is symmetric and may therefore be considered as a collection of $\frac{m(m+1)}{2}$ distinct random variables

V_{ij}, $i,j = 1,\ldots,m$, $i \leq j$. In the event that $\nu > m-1$ and Σ is non singular, the joint distribution of the random variables $\underset{\sim}{V}_{ij}$, $i,j = 1,\ldots,m$, $i \leq j$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{m(m+1)/2}$ and therefore has a density. The value of the density function is given by

$$(3.1) \qquad f(V|\nu,\Sigma) := C|\Sigma|^{-\nu/2}|V|^{\nu-m-1/2}e^{-(1/2)tr(\Sigma^{-1}V)}1_{\mathcal{V}}(V)$$

where $tr(\cdot)$ denotes, as usual, the trace operator,

$$\Gamma_m(\frac{\nu}{2}) := \prod_{i=1}^{m}\Gamma(\frac{\nu+1-i}{2})$$

$$C := \{2^{\nu m/2}\Gamma^{m(m-1)/4}\Gamma_m(\frac{\nu}{2})\}^{-1}$$

and $1_{\mathcal{V}}(\cdot)$ represents the indicator function of the class of positive definite matrices $\mathcal{V}$.

In case $\underset{\sim}{V}$ has a density given as above, we say that $\underset{\sim}{V}$ has a non-singular Wishart distribution with ν degrees of freedom and parametric matrix Σ. We will denote this fact by simply writing $\underset{\sim}{V} \sim W(\nu,\Sigma)$. In many instances it is convenient to set $\tau := \Sigma^{-1}$ and in this case we call τ the precision matrix of the variate $\underset{\sim}{V}$.

For a complete discussion and derivation of the above density one can see for example Anderson (1958).

In specifying a set of values v_{ij} for the random variables $\underset{\sim}{V}_{ij}$, we may regard these values as a vector in $\mathbb{R}^{m(m+1)/2}$. Equivalently, we may think of these values as characterizing an $m \times m$ symmetric matrix V. In either case we will refer to (3.1) as the density function.

At this point we note two additional properties of the Wishart distribution that are apparent from the definition given in (3.1). The first is that when $\nu > m-1$ and when Σ is non singular then the random matrix $\underset{\sim}{V}$ with density (3.1) is positive definite with probability one. The second is that while we initially interpret the random matrix $\underset{\sim}{V}$ as a finite sum, that is, ν is an integer, the function given by (3.1) will be a density for non integer ν, provided that $\nu > m-1$ and Σ is positive definite.

3.2 Properties of the Wishart Distribution

For notation in all that follows we let i denote $\sqrt{-1}$.

Lemma 3.2.1. Suppose the random matrix $\underset{\sim}{V}$ has a Wishart distribution with ν degrees of freedom and parametric matrix Σ, then the joint characteristic function of the distinct random variables $\underset{\sim}{V}_{ij}$, $i, j = 1, 2, \ldots, m$, $i \leq j$ evaluated at the random vector t_{ij}, $1 \leq i \leq j \leq m$ is given by the formula

(3.2) $$\phi_{\underset{\sim}{V}}(t) := |\Psi^{-1}|^{\nu/2}|\Psi^{-1} - it|^{-\nu/2}$$

where

$$t = \begin{bmatrix} 2t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & 2t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & & \vdots \\ t_{m1} & t_{m2} & \cdots & 2t_{mm} \end{bmatrix}.$$

Proof: See Kshirsagar (1972). ∎

We note here that, consistent with the comment following example A.2.2, the characteristic function given in (3.2) can be viewed as the characteristic function of the random matrix $\underset{\sim}{V}$ evaluated at a symmetric matrix t as in (3.2) with $\langle t, \underset{\sim}{V} \rangle = \frac{1}{2}\operatorname{tr}(t\underset{\sim}{V})$. If we view $\underset{\sim}{V}$ as a random matrix and use the inner product $\operatorname{tr}(t\underset{\sim}{V})$ as in example A.2.2, we then have a characteristic function evaluated at any symmetric matrix t given by

(3.3) $$\phi_{\underset{\sim}{V}}(t) = |\Psi^{-1}|^{\nu/2}|\Psi^{-1} - 2it|^{-\nu/2}.$$

For this viewpoint one can see Anderson (1958).

We should comment here that in interpreting the characteristic function of a Wishart distribution, (3.2) is the joint characteristic function of the random variables $\underset{\sim}{V}_{11}, \underset{\sim}{V}_{12}, \cdots, \underset{\sim}{V}_{1m}, \underset{\sim}{V}_{22}, \underset{\sim}{V}_{23}, \cdots, \underset{\sim}{V}_{2m}, \cdots, \underset{\sim}{V}_{mm}$ while (3.3) is the joint characteristic function of the

variables $\underset{\sim}{V}_{11}, \underset{\sim}{V}_{22}, \ldots, \underset{\sim}{V}_{33}, 2\underset{\sim}{V}_{12}, 2\underset{\sim}{V}_{13}, \ldots, 2\underset{\sim}{V}_{m-1,m}$. In this thesis we have adopted the viewpoint consistent with (3.2) and use (3.2) as the characteristic function for a Wishart variate with ν degrees of freedom and precision Σ.

It should also be noted here that the proofs of this lemma that were cited initiate their discussion by assuming a sample of size ν from a $N_m(0, \Sigma)$ distribution. An examination of the proofs reveals that the lemma is true when $\nu \leq m-1$ and integer even though the density exists only when $\nu > m-1$. In case $\nu < m-1$ we say that we have a singular Wishart distribution with ν degrees of freedom. We also note that Lukacs (1969) has shown that if $\nu > m-1$ and not necessarily integer then $\underset{\sim}{V}$ has a moment generating function of the form $\frac{|\Sigma^{-1}|^{\nu/2}}{|\Sigma^{-1}-t|^{\nu/2}}$ and hence a characteristic function of the form given by the lemma. Olkin and Rubin (1962) have shown that $\frac{|\Sigma^{-1}|^{\nu/2}}{|\Sigma^{-1}-t|^{\nu/2}}$ is not the moment generating function of a probability distribution if $\nu < m-1$ and non integer.

Lemma 3.2.2. Suppose we have the transformation $\underset{\sim}{V} = C\underset{\sim}{A}C^T$ where C is some fixed non singular matrix. If $\underset{\sim}{V} \sim W_m(\nu, \Sigma)$ then $\underset{\sim}{A} \sim W_m(\nu, C^{-1}\Sigma(C^T)^{-1})$.

Proof: If $\underset{\sim}{V} = C\underset{\sim}{A}C^T$ then $\underset{\sim}{A} = C^{-1}\underset{\sim}{V}(C^T)^{-1}$. Now apply Lemma A.2.5 from the Appendix to conclude that the characteristic function of $\underset{\sim}{A}$ is given by

$$\phi_{\underset{\sim}{A}}(t) := \phi_{\underset{\sim}{V}}[((C^T)^{-1}tC^{-1})^T]$$

$$= \phi_{\underset{\sim}{V}}[(C^{-1})^T tC^{-1}]$$

$$= \frac{|\Sigma^{-1}|^{\nu/2}}{|\Sigma^{-1} - i(C^{-1})^T tC^{-1}|^{\nu/2}}$$

$$= \frac{|\Sigma^{-1}|^{\nu/2}}{|(C^{-1})^T(C^T\Sigma^{-1}C - it)C^{-1}|^{\nu/2}}$$

$$= \frac{|C^T\Sigma^{-1}C|^{\nu/2}}{|C^T\Sigma^{-1}C - it|^{\nu/2}} \quad .$$

This is the characteristic function of a Wishart variate with parameters ν and $(C^T\Sigma^{-1}C)^{-1} = C^{-1}\Sigma(C^T)^{-1}$. ∎

This lemma may be used in the following way to derive conclusions concerning a Wishart random matrix. Suppose we are interested in some property $\mathcal{P}$ regarding a Wishart variate $\underset{\sim}{V} \sim W_m(\nu, \Sigma)$. In general we may verify $\mathcal{P}$ for a variate $\underset{\sim}{A} \sim W(\nu, I)$. Next set $\Sigma = CC^T$ where C is $m \times m$ and non singular and $\underset{\sim}{A} := C^{-1}\underset{\sim}{V}(C^T)^{-1}$ so that $\underset{\sim}{V} = C\underset{\sim}{A}C^T$. Now by Lemma 3.2.2, $\underset{\sim}{V} \sim W(\nu, \Sigma)$ implies $\underset{\sim}{A} \sim W(\nu, I)$. In many instances, having verified $\mathcal{P}$ for the variate $\underset{\sim}{A}$ we will be able to invoke simple properties of transformations to conclude the validity of $\mathcal{P}$ for $\underset{\sim}{V}$. We will use this result in the proofs of several following lemmas.

We first prove several preliminary lemmas.

Lemma 3.2.3. If Σ and t are symmetric matrices with t as in (3.2) then $\frac{\partial|\bar{A}|}{\partial t_{ij}} = -2i\,|\bar{A}_{ij}|$ where $\bar{A} := \Sigma^{-1} - it$ and $|\bar{A}_{ij}|$ is the ij cofactor of $\bar{A}$.

Proof: By A.1.1 (iii) of the Appendix,

$$\frac{\partial|\bar{A}|}{\partial t_{ij}} = \mathrm{tr}[\bar{A}^{\#}\,\frac{\partial \bar{A}^{T}}{\partial t_{ij}}]\,.$$

By A.1.1 (i) and (iv),

$$\frac{\partial \bar{A}^{T}}{\partial t_{ij}} = \frac{\partial(\Sigma^{-1} - it)}{\partial t_{ij}}$$

$$= -i\,\frac{\partial t}{\partial t_{ij}}$$

$$= \begin{cases} -i\Delta_{ij}^{*} & \text{if } i \neq j \\ -2i\Delta_{ii}^{*} & \text{if } i = j \end{cases}$$

where Δ_{ij}^{*} is a matrix of zeroes except for ones as the ij and ji elements. Therefore,

$$\frac{\partial|\bar{A}|}{\partial t_{ij}} = \begin{cases} -i\,\mathrm{tr}[\bar{A}^{\#}\Delta_{ij}^{*}] & \text{if } i \neq j \\ -2i\,\mathrm{tr}[\bar{A}^{\#}\Delta_{ii}^{*}] & \text{if } i = j \end{cases}.$$

If $i \neq j$ then $\bar{A}^{\#}\Delta_{ij}^{*}$ will have as its i-th column the j-th column of $\bar{A}^{\#}$ and its j-th column will be the i-th column of

$\bar{A}^{\#}$ with all other entries being zero. Therefore,

$$tr[\bar{A}^{\#}\Delta^{*}_{ij}] = |\bar{A}_{ij}| + |\bar{A}_{ji}|$$

and

$$\frac{\partial|\bar{A}|}{\partial t_{ij}} = -i[|\bar{A}_{ij}| + |\bar{A}_{ji}|] .$$

Since $\bar{A}$ is symmetric, $|\bar{A}_{ij}| = |\bar{A}_{ji}|$ so that

$$\frac{\partial|\bar{A}|}{\partial t_{ij}} = -2i\,|\bar{A}_{ij}| .$$

If $i = j$ then $\bar{A}^{\#}\Delta^{*}_{ii}$ will have as its i-th column the i-th column of $\bar{A}^{\#}$ and

$$tr[\bar{A}^{\#}\Delta^{*}_{ii}] = |\bar{A}_{ii}| .$$

Therefore,

$$\frac{\partial|\bar{A}|}{\partial t_{ij}} = -2i\,|\bar{A}_{ii}| . \quad \blacksquare$$

Lemma 3.2.4. If $\bar{A}$ is as in the previous lemma then

$$\frac{\partial|\bar{A}_{ij}|}{\partial t_{k\ell}} = |\bar{A}|(\bar{a}^{ik}\bar{a}^{j\ell} + \bar{a}^{i\ell}\bar{a}^{jk}) - 2i\,\frac{|\bar{A}_{ij}|}{|\bar{A}|}\,|\bar{A}_{k\ell}|$$

where $\bar{a}^{gh}$ denotes the gh element of $\bar{A}^{-1}$.

Proof: We first note that for any non singular matrix $\bar{M}$,

$$\frac{|\bar{M}_{ij}|}{|\bar{M}|} = m^{ji} .$$

Now,

$$\frac{\partial |\bar{A}_{ij}|}{\partial t_{k\ell}} = \frac{\partial}{\partial t_{k\ell}} |\bar{A}| \frac{|\bar{A}_{ij}|}{|\bar{A}|} .$$

By the chain rule,

$$\frac{\partial |\bar{A}_{ij}|}{\partial t_{k\ell}} = |\bar{A}| \frac{\partial}{\partial t_{k\ell}} \frac{|\bar{A}_{ij}|}{|\bar{A}|} + \frac{|\bar{A}_{ij}|}{|\bar{A}|} \frac{\partial}{\partial t_{k\ell}} |\bar{A}| .$$

From Lemma 3.2.3,

$$\frac{\partial |\bar{A}_{ij}|}{\partial t_{k\ell}} = |\bar{A}| \frac{\partial}{\partial t_{k\ell}} \frac{|\bar{A}_{ij}|}{|\bar{A}|} - 2 i \frac{|\bar{A}_{ij}|}{|\bar{A}|} |\bar{A}_{k\ell}| .$$

From the initial remark,

$$\begin{aligned}
\frac{\partial}{\partial t_{k\ell}} \frac{|\bar{A}_{ij}|}{|\bar{A}|} &= \frac{\partial}{\partial t_{k\ell}} \bar{a}^{ji} \\
&= \sum_{r,s} \frac{\partial \bar{a}^{ji}}{\partial \bar{a}_{rs}} \frac{\partial \bar{a}_{rs}}{\partial t_{k\ell}} \\
&= \frac{\partial \bar{a}^{ji}}{\partial \bar{a}_{k\ell}} \frac{\bar{a}_{k\ell}}{\partial t_{k\ell}} \\
&= - i \frac{\partial \bar{a}^{ji}}{\partial \bar{a}_{k\ell}} .
\end{aligned}$$

The expression $\frac{\partial \bar{a}^{ji}}{\partial \bar{a}_{k\ell}}$ is simply the ji element of the matrix

$$\frac{\partial \bar{A}^{-1}}{\partial \bar{a}_{k\ell}} = -\bar{A}^{-1}\Delta^{*}_{k\ell}\bar{A}^{-1}$$

where as before $\Delta^{*}_{k\ell}$ is a matrix of zeroes except for ones as the $k\ell$ and ℓk elements. A simple multiplication gives

$$(-\bar{A}^{-1}\Delta^{*}_{k\ell}\bar{A}^{-1})_{ji} = -(\bar{a}^{ik}\bar{a}^{j\ell} + \bar{a}^{k\ell}\bar{a}^{jk}).$$

Therefore,

$$\frac{\partial}{\partial t_{k\ell}}\frac{|\bar{A}_{ij}|}{|\bar{A}|} = (\bar{a}^{ik}\bar{a}^{j\ell} + \bar{a}^{i\ell}\bar{a}^{jk})$$

and the result follows. ∎

We are now ready to find some of the moments of the Wishart distribution.

Lemma 3.2.5. If $\underset{\sim}{V} \in W(\nu, \Sigma)$ then $E\underset{\sim}{V} = \nu\Sigma$.

Proof: If we are restricting ν to be integer then the result follows immediately from the characterization

$$\underset{\sim}{V} = \sum_{\alpha=1}^{\nu} \underset{\sim}{Z}_{\alpha}\underset{\sim}{Z}_{\alpha}^{T}.$$

The result is not quite so apparent if we allow ν to be non integer.

Let us initially suppose that $\Sigma = I$. Regarding the random variables $\underset{\sim}{V}_{ij}$, $1 \leq i \leq j \leq m$ as a vector in $\mathbb{R}^{m(m+1)/2}$ we have from Lemma A.2.3,

$$E\underset{\sim}{V}_{ij} = \frac{1}{i} \frac{\partial \phi_{\underset{\sim}{V}}(t)}{\partial t_{ij}}\Big|_{t=0}$$

Now,

$$\frac{\partial \phi_{\underset{\sim}{V}}(t)}{\partial t_{ij}}\Big|_{t=0} = \frac{\partial}{\partial t_{ij}} |I|^{\nu/2} |I - it|^{-\nu/2}\Big|_{t=0}$$

$$= \{-\frac{\nu}{2} |I - it|^{-\nu/2-1} \frac{\partial}{\partial t_{ij}} |I - it|\}\Big|_{t=0} .$$

By Lemma 3.2.3 with $\Sigma = I$ we have

$$\frac{\partial \phi_{\underset{\sim}{V}}(t)}{\partial t_{ij}}\Big|_{t=0} = \{i\nu |I - t|^{-\nu/2-1} |(I - it)_{ij}|\}\Big|_{t=0}$$

$$= i\nu \delta_{ij}$$

and

$$E\underset{\sim}{V}_{ij} = \nu \delta_{ij} .$$

Therefore,

$$E\underset{\sim}{V} = \nu I.$$

For the general case, if $\underset{\sim}{V} \sim W(\nu, \Sigma)$ set $\Sigma = CC^T$ where C is $m \times m$ and non singular and set $\underset{\sim}{A} := C^{-1}\underset{\sim}{V}(C^T)^{-1}$ so that

$\underset{\sim}{V} = C\underset{\sim}{A}C^T$. By Lemma 3.2.2, $\underset{\sim}{A} \sim W(\nu, I)$. We now have

$$\begin{aligned} E\underset{\sim}{V} &= EC\underset{\sim}{A}C^T \\ &= C(E\underset{\sim}{A})C^T \\ &= C(\nu I)C^T \\ &= \nu \Sigma . \end{aligned}$$ ∎

Additional useful moments are given in the following lemma.

Lemma 3.2.6. If $\underset{\sim}{V} \in W_m(\nu, \Sigma)$ then

i) $E(\underset{\sim}{V}_{ij}) = \nu \sigma_{ij}$

ii) $E(\underset{\sim}{V}_{ij}\underset{\sim}{V}_{k\ell}) = \nu^2\sigma_{ij}\sigma_{k\ell} + \nu\sigma_{ik}\sigma_{j\ell} + \nu\sigma_{i\ell}\sigma_{jk}$

iii) $\text{cov}(\underset{\sim}{V}_{ij}, \underset{\sim}{V}_{k\ell}) = \nu(\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})$

iv) $\text{var}(\underset{\sim}{V}_{ij}) = \nu(\sigma_{ii}\sigma_{jj}+\sigma_{ij}^2)$.

Proof: Result (i) was proved in Lemma 3.2.5. (iii) follows from (i) and (ii) in the usual way by writing

$$\text{cov}(\underset{\sim}{V}_{ij}, \underset{\sim}{V}_{k\ell}) = E(\underset{\sim}{V}_{ij}\underset{\sim}{V}_{k\ell}) - (E\underset{\sim}{V}_{ij})(E\underset{\sim}{V}_{k\ell})$$

and (if) follows from (iii) by the relationship

$$\text{var}(\underset{\sim}{V}_{ij}) = \text{cov}(\underset{\sim}{V}_{ij}, \underset{\sim}{V}_{ij}) .$$

All that needs to be proved is (ii) and the proof of this is a direct computation using the characteristic function and results on matrix

differentiation found in the Appendix. We can write

$$(3.4) \qquad E\underset{\sim}{V}_{ij}\underset{\sim}{V}_{k\ell} = \frac{1}{i^2}\,\frac{\partial^2 \phi_{\underset{\sim}{V}}(t)}{\partial t_{k\ell}\,\partial t_{ij}}\Big|_{t=0}$$

For notation, as before let $\bar{A} := \Sigma^{-1} - i\,t$ and $A := \Sigma^{-1}$. Using Lemma 3.2.3 we can write

$$\frac{\partial \phi_{\underset{\sim}{V}}(t)}{\partial t_{ij}} = i\nu|\Sigma^{-1}|^{\nu/2}|\Sigma^{-1} - i\,t|^{-\nu/2-1}|\bar{A}_{ij}| \,.$$

Taking second partials with respect to $t_{k\ell}$ gives

$$\begin{aligned}
\frac{\partial^2 \phi_{\underset{\sim}{V}}(t)}{\partial t_{k\ell}\,\partial t_{ij}} &= \frac{\partial}{\partial t_{k\ell}}\left\{ i\nu|\Sigma^{-1}|^{\nu/2}|\Sigma^{-1} - it|^{-\nu/2-1}|\bar{A}_{ij}|\right\} \\
&= i\nu|\Sigma^{-1}|^{\nu/2}\left\{-(\tfrac{\nu}{2}+1)|\bar{A}_{ij}||\bar{A}|^{-\nu/2-2}\frac{\partial}{\partial t_{k\ell}}|\bar{A}|\right. \\
&\qquad \left. + |\bar{A}|^{-\nu/2-1}\frac{\partial}{\partial t_{k\ell}}|\bar{A}_{ij}|\right\} \\
&= i\nu|\Sigma^{-1}|^{\nu/2}\left\{2i(\tfrac{\nu}{2}+1)|\bar{A}_{ij}||\bar{A}|^{-\nu/2-2}|\bar{A}_{k\ell}|\right. \\
&\qquad \left. + |\bar{A}|^{-\nu/2-1}\frac{\partial}{\partial t_{k\ell}}|\bar{A}_{ij}|\right\}
\end{aligned}$$

Now,

$$\frac{\partial^2 \phi_{V}(t)}{\partial t_{k\ell}\,\partial t_{ij}}\Big|_{t=0} = i^2(\nu^2+2\nu)\left\{\frac{|\bar{A}_{ij}||A_{k\ell}|}{|\Sigma^{-1}||\Sigma^{-1}|}\right\}\Big|_{t=0} + i\nu|\Sigma|\left\{\frac{\partial}{\partial t_{k\ell}}|\bar{A}_{ij}|\right\}_{t=0}.$$

When $t = 0$, $|\bar{A}_{ij}|$ is the (i,j) cofactor of Σ^{-1} so that

$$\left.\frac{\partial^2 \phi_{\underset{\sim}{V}}(t)}{\partial t_{k\ell}\,\partial t_{ij}}\right|_{t=0} = i^2(\nu^2+2\nu)\,\frac{|\Sigma^{-1}_{ij}||\Sigma^{-1}_{k\ell}|}{|\Sigma^{-1}||\Sigma^{-1}|} + i\nu|\Sigma|\left\{\frac{\partial}{\partial t_{k\ell}}|\bar{A}_{ij}|\right\}\Bigg|_{t=0}.$$

Noting that $\dfrac{|\Sigma^{-1}_{ij}|}{|\Sigma^{-1}|}$ is the (i,j) element of Σ and using (3.4) we have

$$E\underset{\sim}{V}_{ij}\underset{\sim}{V}_{k\ell} = (\nu^2+2\nu)\sigma_{ij}\sigma_{k\ell} + \frac{1}{i}\,\nu\,|\Sigma|\left\{\frac{\partial}{\partial t_{k\ell}}|\bar{A}_{ij}|\right\}_{t=0}$$

By Lemma 3.2.4

$$\frac{\partial}{\partial t_{k\ell}}\,\frac{\partial}{\partial t_{k\ell}}\left\{|\bar{A}_{ij}|\right\}_{t=0} = i|\Sigma^{-1}|(\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk}) - 2i\,\frac{|\Sigma^{-1}_{ij}|}{|\Sigma^{-1}|}\,|\Sigma^{-1}_{k\ell}|$$

so that

$$E\underset{\sim}{V}_{ij}\underset{\sim}{V}_{k\ell} = \nu^2\sigma_{ji}\sigma_{\ell k} + 2\nu\sigma_{ij}\sigma_{k\ell} - 2\nu\frac{|\Sigma^{-1}_{ij}|}{|\Sigma^{-1}|}\,\frac{|\Sigma^{-1}_{k\ell}|}{|\Sigma^{-1}|} + \nu(\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})$$

and the result is established. ∎

3.3 The Inverted Wishart Distribution

Another distribution that will be of use is the distribution of a random matrix $\underset{\sim}{U}$ where $\underset{\sim}{U} = \underset{\sim}{V}^{-1}$ and $\underset{\sim}{V} \sim W(\nu, \Sigma)$.

Definition 3.3.1. An $m \times m$ random matrix $\underset{\sim}{U}$ is said to have an inverted Wishart distribution with parameters S and t

where S is a positive definite matrix and t is a real number in case $\underset{\sim}{U}$ has a density given by

$$f(U|S,t) = \begin{cases} \dfrac{|U|^{-t/2}|S|^{(t-m-1)/2}e^{-(1/2)tr(U^{-1}S)}}{2^{(t-m-1)(m/2)}\pi^{m(m-1)/2}\Gamma_m(\frac{t}{2})} & \text{if } U > 0 \\ 0 & \text{otherwise} \end{cases} \qquad 2m < t$$

When $\underset{\sim}{U}$ has an inverted Wishart distribution we will write $\underset{\sim}{U} \sim W_m^{-1}(t,S)$.

The inverse Wishart that will be of interest to us arises when we have $\underset{\sim}{V} \sim W_m(\nu,\Sigma)$ and set $\underset{\sim}{U} = \underset{\sim}{V}^{-1}$. A direct computation shows that in this case $\underset{\sim}{U} \sim W_m^{-1}(\nu+m+1,\Sigma^{-1})$. A discussion of this density may be found in Press (1972).

3.4 Properties of the Inverted Wishart Distribution

Let us first recall that if a random variable $\underset{\sim}{X}$ is distributed as χ^2 with K degrees of freedom then $E[\frac{1}{\underset{\sim}{X}}] = \frac{1}{K-2}$. Also, we should recall the following long known lemma.

Lemma 3.4.1. If $\underset{\sim}{V} \sim W_m(\nu,\Sigma)$ and L is a fixed vector then

$$\frac{L^T\Sigma^{-1}L}{L^T\underset{\sim}{V}^{-1}L} \sim \chi^2_{[\nu-(m-1)]}$$

Proof: See Rao (1965). ∎

We now can prove the following useful result.

Lemma 3.4.2. If $\underset{\sim}{V} \sim W(\nu, \Sigma)$ then

$$E\underset{\sim}{V}^{-1} = \frac{1}{\nu-m-1}\Sigma^{-1}$$

Proof: Let us first assume $\Sigma = I$ so that $\underset{\sim}{V} \sim W_m(\nu, I)$ and let $L := e_i$ where e_i is a standard basis vector with a one in the ith component. By Lemma 3.4.1.

$$\frac{1}{\underset{\sim}{U}_{ii}} \sim \chi^2_{[\nu-(m-1)]}$$

where $\underset{\sim}{U} = \underset{\sim}{V}^{-1}$. It now follows that each diagonal element of $\underset{\sim}{U}$ has the same expectation. That is,

$$E\underset{\sim}{U}_{ii} = \frac{1}{\nu-(m-1)-2} \qquad i = 1, 2, \ldots, m$$

$$= \frac{1}{\nu-m-1} \qquad i = 1, 2, \ldots, m.$$

Now, let $e_{ij} := e_i + e_j$ and set $L := e_{ij}$. Again, by Lemma 3.4.1,

$$\frac{L^T I L}{L^T \underset{\sim}{U} L} \sim \chi^2_{[\nu-(m-1)]}.$$

That is,

$$\frac{2}{\underset{\sim}{U}_{ii}+\underset{\sim}{U}_{jj}+2\underset{\sim}{U}_{ij}} \sim \chi^2_{(\nu-m+1)} \quad \text{for} \quad i, j = 1, 2, \ldots, m, \; i \neq j.$$

Again taking the expectation of the reciprocal of a χ^2 variate we have

$$\frac{1}{2}E(\underset{\sim}{U}_{ii}+\underset{\sim}{U}_{jj}+2\underset{\sim}{U}_{ij}) = \frac{1}{\nu-m-1}$$

or

$$E\underset{\sim}{U}_{ii} + E\underset{\sim}{U}_{jj} + 2E\underset{\sim}{U}_{ij} = \frac{2}{\nu-m-1}$$

so that

$$\frac{1}{\nu-m-1} + \frac{1}{\nu-m-1} + 2E\underset{\sim}{U}_{ij} = \frac{2}{\nu-m-1}$$

and

$$E\underset{\sim}{U}_{ij} = 0$$

To summarize, when $\Sigma = I$.

$$\begin{aligned} E\underset{\sim}{V}^{-1} &= E\underset{\sim}{U} \\ &= (E\underset{\sim}{U}_{ij})_{m \times m} \\ &= \frac{1}{\nu-m-1} I \end{aligned}$$

For the general case the result follows by setting $\Sigma = CC^T$, C non singular, $\underset{\sim}{A} = C^{-1}\underset{\sim}{V}(C^T)^{-1}$ so that $\underset{\sim}{V} = C\underset{\sim}{A}C^T$, and applying Lemma 3.2.2 to conclude that $\underset{\sim}{A} \sim W_m(\nu, I)$. Now,

$$\begin{aligned} \underset{\sim}{V}^{-1} &= (C\underset{\sim}{A}C^T)^{-1} \\ &= (C^T)^{-1}\underset{\sim}{A}^{-1}C^{-1} \end{aligned}$$

and

$$E\underset{\sim}{V}^{-1} = (C^T)^{-1}E(\underset{\sim}{A}^{-1})C^{-1}$$

$$= (C^T)^{-1}\left(\frac{1}{\nu-m-1} I\right)C^{-1}$$

$$= \frac{1}{\nu-m-1}(CC^T)^{-1}$$

$$= \frac{1}{\nu-m-1}\Sigma^{-1} . \quad \blacksquare$$

Other moments for an inverted Wishart distribution are given in the following lemma.

Lemma 3.4.3. If $\underset{\sim}{V} \sim W_m(\nu, \Sigma)$ and $\underset{\sim}{U} = \underset{\sim}{V}^{-1}$ so that $\underset{\sim}{U} \sim W_m^{-1}(\nu+m+1, \Sigma^{-1})$ then

i) $$\text{cov}(\underset{\sim}{U}_{ij}, \underset{\sim}{U}_{k\ell}) = \frac{\left[\frac{2\sigma^{ij}\sigma^{k\ell}}{\nu-m-1} + \sigma^{ik}\sigma^{j\ell} + \sigma^{i\ell}\sigma^{kj}\right]}{(\nu-m)(\nu-m-1)(\nu-m-3)}$$

ii) $$\text{var}(\underset{\sim}{U}_{ii}) = \frac{2(\sigma^{ii})^2}{(\nu-m-1)^2(\nu-m-3)}$$

iii) $$\text{var}(\underset{\sim}{U}_{ij} = \frac{\sigma^{ii}\sigma^{jj} + \frac{\nu-m+1}{\nu-m-1}(\sigma^{ij})^2}{(\nu-m)(\nu-m-1)(\nu-m-3)}$$

Proof: See Press (1972). $\blacksquare$

IV. ESTIMATION OF PARAMETERS

4.1 A Natural Conjugate Family of Priors

Our model assumes, after suitable transformations, that we are observing a random vector U defined on $(\mathbb{R}^m, \mathcal{B}^m)$ having a multivariate density given by $p(u)$ with $EU_i = 0$. We also assume that the density $p(u)$ has expansions given by (2.2) and (2.3) both of which involve parameters to be estimated. In most applications we will be interested in truncating these series at some point. In examining expansion (2.3), we see that by truncating the series at some point we have a representation of $p(u)$ by a distribution that is a member of an exponential family.

With this in mind we shall initiate our discussion of the estimation problem with a general comment concerning exponential families of distributions and certain related priors.

Suppose we have an m-dimensional random vector $\underset{\sim}{U}$ whose distribution is in some exponential family. As an example let us assume that $\underset{\sim}{U}$ is multivariate normal. Furthermore, let us assume that the mean is known to be zero and we are interested in the problem of estimating the precision matrix R. In such a problem, one might be interested in using a Bayesian approach by introducing a prior distribution for the parameters to be estimated.

In general, a useful family of priors when considering a Bayesian approach to an estimation problem is a family of priors that is a "natural conjugate" family with respect to the likelihood function. We mean by this that both the prior and posterior distributions of the parameters involved are in the same family of distributions. Having the prior and the posterior distribution both in the same family has the advantage of mathematical tractibility. A natural conjugate family also allows us, in many instances, to determine in an intuitive way the relative contributions of the prior distribution and the sample data to the posterior knowledge of the parameter. For example, suppose we consider a random sample of size n from a normal distribution with unknown mean μ_1 and known variance σ_1^2. It is well known that the family of normal distributions is a natural conjugate family for this problem and that the posterior distribution of μ_1 is normal with mean and variance given by

$$\mu_2 = \frac{\frac{n\bar{x}}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}}$$

and

$$\sigma_2^2 = \frac{1}{\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)}$$

where μ_0 and σ_0^2 are the mean and variance of the prior distribution for μ_1. From these relationships we may interpret the prior distribution for μ_1 as being equivalent to a prior sample of size $n_0 = \sigma_1^2/\sigma_0^2$ from the normal distribution with unknown mean μ_1 and known variance σ_1^2 starting with a flat prior. That is, we suppose we start with no observations and an improper uniform prior distribution for μ_1. We then suppose we have a sample of size n_0 and sample mean μ_0 from the given distribution. This "equivalent" sample produces a posterior distribution that is normal with mean μ_0 and variance σ_0^2. When this equivalent sample is combined with the actual sample the composite sample yields the posterior distribution with mean μ_2 and variance σ_2^2. For a more complete discussion of natural conjugate families the reader is referred to Raiffa and Schlaifer (1961) or Barnett (1973).

Suppose we now consider an exponential family of distributions determined by the vector θ of parameters. The density may be written as

$$f(u|\theta) = e^{\theta^T W(u) - \psi(\theta)}$$

where $W(u)$ is a vector dependent upon U and $\psi(\theta)$ is a normalizing constant.

The likelihood function for a sample of size n from the above distribution may be written as

$$(4.1) \qquad f({}^{1}u, {}^{2}u, \ldots, {}^{n}u \mid \theta) = e^{\sum_{t=1}^{n} \theta^T W({}^{t}u) - n\psi(\theta)}$$

$$= e^{n\theta^T \bar{w} - n\psi(\theta)}$$

where

$$\bar{w} = \frac{1}{n} \sum_{t=1}^{n} W({}^{t}u)$$

Now suppose that we have a prior distribution for θ that has a density given by

$$(4.2) \qquad g(\theta) \propto e^{b^*[\theta^T a^* - \psi(\theta)]}$$

where b^* is some specified constant and a^* is some prescribed vector.

The posterior distribution of θ given the observations is given by

$$(4.3) \qquad g(\theta \mid {}^{1}u, {}^{2}u, \ldots, {}^{n}u) \propto e^{\theta^T[n\bar{w}+b^*a^*] - (n+b^*)\psi(\theta)}$$

$$= e^{(n+b^*)[\theta^T(\frac{n\bar{w}+b^*a^*}{n+b^*}) - \psi(\theta)]}$$

$$= e^{b[\theta^T a - \psi(\theta)]}$$

where b is the constant $n+b^*$ and a is the vector $\frac{n\bar{w}+b^*a^*}{n+b^*}$. From Equations (4.2) and (4.3) we see that the family of distributions given by (4.2) forms a natural conjugate family of priors for the

exponential family of distributions.

One very useful result concerning families which are natural conjugate families for exponential distributions is the following result.

Lemma 4.1.1. Suppose the density of the random vector $\underset{\sim}{U}$ is given by $f(u|\theta) = e^{\theta^T W(u) - \psi(\theta)}$. If a prior distribution for θ is given by (4.2) so that the corresponding posterior distribution is given by (4.3) and we set $\psi_i(\theta) := \frac{\partial \psi(\theta)}{\partial \theta_i}$, then under mild regularity conditions $\psi_i(\theta) = E[W_i(u)|\theta]$ and $a_i = E\psi_i(\theta)$ $(a_i^* = E\psi_i(\theta))$ where the expectation is with respect to the posterior (prior) distribution of θ. Moreover, a_i (a_i^*) is the modal value of the posterior (prior) in the sense that if $\overline{\theta}$ is defined by $\psi_i(\overline{\theta}) = a_i$ then $\overline{\theta}$ is the mode of the posterior.

Proof: The first statement in the lemma may be found in Barndorff-Nielsen (1973). The result $a_i = E\psi_i(\theta)$ is due to Morgan (1969). The regularity conditions given by Morgan are given in the Appendix. ∎

In the event that $\underset{\sim}{U}$ is an m-dimensional random vector that has a normal distribution with mean zero and precision R, we may write the density as

$$(4.4) \qquad f(u|R) = (2\pi)^{-m/2}|R|^{1/2} e^{-\frac{1}{2}u^T Ru}$$

$$= (2\pi)^{-m/2}|R|^{1/2} e^{-\frac{1}{2}\Sigma_{ij} r_{ij} u_i u_j}$$

$$= (2\pi)^{-m/2} e^{\Sigma_i r_{ii} w_{ii}(u) + \Sigma_{i<j} r_{ij} w_{ij}(u) - \psi(R)}$$

where

$$w_{ii}(u) := -\frac{1}{2}u_i^2$$

$$w_{ij}(u) := -u_i u_j \quad \text{for} \quad i < j$$

$$\psi(R) := -\frac{1}{2}\log|R| \ .$$

The likelihood function for a sample of size n is given by

$$(4.5) \qquad f({}^1u, {}^2u, \ldots, {}^nu|R) \propto e^{n\Sigma_i r_{ii}\bar{w}_{ii} + n\Sigma_{i<j} r_{ij}\bar{w}_{ij} - n\psi(R)}$$

where $\bar{w}_{ij} = \frac{1}{n}\sum_{t=1}^{n} w_{ij}({}^tu)$. Now suppose that R is given a prior distribution that is Wishart with ν degrees of freedom and precision matrix τ. If $\nu > m-1$ the density of R is given by

$$f(R|\nu,\tau) \propto |R|^{\frac{\nu-m-1}{2}} e^{-\frac{1}{2}tr(\tau R)}$$

$$= e^{-\frac{1}{2}tr(\tau R) + \frac{\nu-m-1}{2}\log|R|}$$

where we are assuming R and τ are positive definite. Since τ and R are symmetric we have

$$\operatorname{tr}(\tau R) = \sum_{i=1}^{m} \sum_{j=1}^{m} \tau_{ij} r_{ij}$$

$$= \sum_{i=1}^{m} \tau_{ii} r_{ii} + 2 \sum_{i<j} \tau_{ij} r_{ij} .$$

Therefore,

$$(4.6) \qquad f(R \mid \nu, \tau) \propto e^{-\frac{1}{2}[\Sigma_{i=1}^{m} \tau_{ii} r_{ii} + 2\Sigma_{i<j} \tau_{ij} r_{ij}] + \frac{\nu-m-1}{2} \log|R|}$$

$$= e^{(\nu-m-1)[-\frac{1}{2(\nu-m-1)} \Sigma_i \tau_{ii} r_{ii} - \Sigma_{i<j} \frac{\tau_{ij} r_{ij}}{(\nu-m-1)} + \frac{1}{2} \log|R|]}$$

$$= e^{b^*[\Sigma_i r_{ii} a_{ii}^* + \Sigma_{i<j} r_{ij} a_{ij}^* - \psi(R)]}$$

where

$$b^* := \nu-m-1$$

$$a_{ii}^* := -\frac{\tau_{ii}}{2(\nu-m-1)}$$

$$a_{ij}^* := -\frac{\tau_{ii}}{\nu-m-1} .$$

The posterior distribution of $\underset{\sim}{R}$ given the observations is given by

$$(4.7)\qquad f(R \mid {}^{1}u, \ldots, {}^{n}u, \nu, \tau)$$

$$\propto e^{\Sigma_{i=1}^{m} r_{ii}[n\bar{w}_{ii}+b^*a^*_{ii}]+\Sigma_{i<j} r_{ij}[n\bar{w}_{ij}+b^*a^*_{ij}]-(n+b^*)\psi(R)}$$

$$= e^{(n+b^*)\{\Sigma_i r_{ii}[\frac{n\bar{w}_{ii}+b^*a^*_{ii}}{n+b^*}]+\Sigma_{i<j} r_{ij}[\frac{n\bar{w}_{ij}+b^*a^*_{ij}}{n+b^*}]-\psi(R)\}}$$

$$= e^{b[\Sigma_i r_{ii}a_{ii}+\Sigma_{i<j} r_{ij}a_{ij}-\psi(R)]}$$

where

$$b := n+b^*$$

$$a_{ii} := \frac{n\bar{w}_{ii}+b^*a^*_{ii}}{n+b^*}$$

$$a_{ij} := \frac{n\bar{w}_{ij}+b^*a^*_{ij}}{n+b^*}\,.$$

From (4.6) and (4.7) we note that the prior and posterior distributions for the precision matrix R are members of the same family. Therefore, we conclude that the family of Wishart distributions is a natural conjugate family for the precision matrix in a multivariate normal family with zero mean.

We note in passing that from this result and the relationship between a Wishart and an inverted Wishart distribution the family of inverted Wishart distributions is a natural conjugate family for the covariance matrix in the same problem.

4.2 An Estimator for β_r

In this section we concentrate on deriving an estimator $\hat{\beta}_r$ for the coefficient β_r in the expansion given by (2.2). We assume that we have a sample training set ${}^t\underset{\sim}{U}, t = 1, 2, \ldots, n$ from one of our classes where $\underset{\sim}{U}$ is a vector of m components. We recall that $H_r({}^t u)$ denotes the m-dimensional basis function evaluated at the observation ${}^t u$. Set

$$\overline{H}_r := \frac{1}{n} \sum_{t=1}^{n} H_r({}^t u)$$

so that if n is large, $\overline{H}_r$ is approximately multivariate normal. We now consider the likelihood function for α. (Recall that α was defined on page 26.) We also recall that if α_r is near zero for all r then $\beta_r \doteq \alpha_r$. The likelihood function is given by

$$L(\alpha) = \left\{ \prod_{t=1}^{n} p_0({}^t u) \right\} e^{n[\Sigma_{r \neq 0} \alpha_r \overline{H}_r - A(\alpha)]} \tag{4.8}$$

From the relationships

$$E[H_r(U) \mid p] = \beta_r$$

and

$$\text{var}[H_r(U)|p] = \text{cov}[H_r(U), H_r(U)|p]$$

$$= \frac{\partial^2 A(a)}{\partial a_r^2}$$

and the fact that when the a's are near zero,

$$\frac{\partial^2 A(a)}{\partial a_r^2} \doteq 1$$

we can conclude that

(4.9) $$E[\underset{\sim}{\overline{H}}_r|p] = \beta_r$$

and

(4.10) $$\text{var}[\underset{\sim}{\overline{H}}_r|p] \doteq \frac{1}{n^2}\sum_{t=1}^{n}\text{var}[H_r(\underset{\sim}{U}^t)|p]$$

$$\doteq \frac{1}{n} .$$

From (4.8) we note that the statistics $\overline{H}_r$, $r \neq 0$, are sufficient for a.

In order to derive a possible estimator we can, as Brunk and Pierce (1974) point out in an analogous situation, represent the notion of near independence among indicants by way of a prior distribution on the a_r's. As Brunk and Pierce (1974) also point out such a prior might be of the form

$$\pi(\alpha) \propto e^{-\frac{1}{2}\Sigma_{r\neq 0} c_r \alpha_r^2} \tag{4.11}$$

where c_r is the precision associated with α_r. That is, we assume a prior whereby the α_r are approximately independent and normally distributed with mean 0 and precision c_r where c_r increases with rank. As a possible estimator, suppose we consider the mode of the posterior distribution of α for such a prior.

The posterior mode will maximize $\log L(\alpha)\pi(\alpha)$ as well as $L(\alpha)\pi(\alpha)$. That is, it will maximize

$$n\left\{\sum_{r\neq 0} \alpha_r \overline{H}_r - A(\alpha)\right\} - \frac{1}{2}\sum_{r\neq 0} c_r \alpha_r^2 .$$

In the usual way, taking partials and using (2.8) we have

$$\frac{\partial}{\partial \alpha_r}\left[n\left\{\sum_{r\neq 0} \alpha_r \overline{H}_r - A(\alpha)\right\} - \frac{1}{2}\sum_{r\neq 0} c_r \alpha_r^2\right] = n[\overline{H}_r - \beta_r] - c_r \alpha_r \tag{4.12}$$

$$= 0 .$$

When $\beta_r \doteq \alpha_r$ then (4.12) reduces to

$$\hat{\beta}_r \doteq \frac{n\overline{H}_r}{n+c_r} . \tag{4.13}$$

We first note that the estimator (4.13) of β_r is a function of the

sufficient statistic $\overline{H}_r$ and the precision c_r of α_r. We next note that from (4.9),

$$E[\hat{\underset{\sim}{\beta}}_r | p] = \frac{n}{n+c_r} E[\overline{\underset{\sim}{H}}_r | p]$$

$$= \frac{n}{n+c_r} \beta_r$$

$$= (1 - \frac{c_r}{n+c_r}) \beta_r \quad .$$

That is $\hat{\beta}_r$ is a biased estimator of β_r with negative bias. Finally we note that the expected mean square error of the estimate $\hat{\beta}_r$ is given by

$$E[(\hat{\underset{\sim}{\beta}}_r - \beta_r)^2 | p] = var[\hat{\underset{\sim}{\beta}}_r | p] + \beta_r^2 (\frac{n}{n+c_r} - 1)^2$$

and

$$var[\hat{\underset{\sim}{\beta}}_r | p] = var[(1 - \frac{c_r}{n+c_r}) \overline{\underset{\sim}{H}}_r | p]$$

$$= (1 - \frac{c_r}{n+c_r})^2 var[\overline{\underset{\sim}{H}}_r | p] \quad .$$

Now, $\beta_r^2 (\frac{n}{n+c_r} - 1)^2 \to 0$ as $n \to \infty$ and since $var[\underset{\sim}{H}_r | p] < \infty$, $var[\hat{\underset{\sim}{\beta}}_r | p] \to 0$ as $n \to \infty$ so that

$$E[(\hat{\underset{\sim}{\beta}}_r - \beta_r)^2 | p] \to 0 \quad \text{as} \quad n \to \infty$$

and $\hat{\beta}_r$ is a consistent estimator for β_r.

If one is interested in using a prior such as (4.11) then it is necessary to specify the precisions c_r. Brunk and Pierce (1974) have indicated that one possible way of specifying precisions c_r for such a prior is the following. The investigator might feel that all a_r of the same rank have the same precision and precision increases with rank. For example,

$$c_r = \pi\lambda^{n(r)-1} . \tag{4.14}$$

This condition is equivalent to the condition that if $n(s) = n(r)+1$ then for $\epsilon > 0$ and $\lambda > 1$, $\Pr\{|a_s| < \epsilon\} = \Pr\{|a_r| < \epsilon\sqrt{\lambda}\}$. This statement in turn is equivalent to the condition $\Pr\{|\sqrt{c_s}\, a_s| < \epsilon\sqrt{c_s}\} = \Pr\{|\sqrt{c_r}\, a_r| < \epsilon\sqrt{(\lambda c_r)}\}$. Now, $\sqrt{c_s}\, a_s$ and $\sqrt{c_r}\, a_r$ are standardized normal variables so that this last statement is equivalent to asserting that $\epsilon\sqrt{c_s} = \epsilon\sqrt{(\lambda c_r)}$ which finally reduces to $c_s = \lambda c_r$.

With a prior as in (4.11) and the precisions as in (4.14), the investigator has reduced the problem of specifying a prior with many parameters to a problem of specifying a two parameter prior. That is, he must specify λ and π. If the investigator feels that he can do an adequate job of determining λ and π he might well consider using (4.13) with no further ado.

Suppose we now consider the situation where the determination

of λ and π is a difficult or costly process. That is, the investigator does not feel competent enough to specify λ and π. We consider the problem from a different viewpoint with the idea that we will let the data aid us in specifying the prior precisions.

Let us consider β_r as a parameter and $\overline{H}_r$ as an observation. When n is large and when β_r and α_r are near zero then, as before, the Central Limit Theorem gives us the validity of (4.9) and (4.10) so that $(\underset{\sim}{\overline{H}}_r \mid \beta_r) \dot{\sim} N(\beta_r, \frac{1}{n})$. Moreover, $\text{cov}[\underset{\sim}{\overline{H}}_r, \underset{\sim}{\overline{H}}_s \mid p]$ is approximately zero so that we may consider the collection of means $\underset{\sim}{\overline{H}}_r$, $r \in B$ as a collection of approximately independent random variables. Now consider a prior on the parameters $\underset{\sim}{\beta}_r$ whereby they are independent normal variables with zero means. Moreover, assume that all $\underset{\sim}{\beta}_r$ with the same rank have the same precision. Denote by φ_k the common variance of all $\underset{\sim}{\beta}_r$ with $n(r) = k$. If the φ_k are all specified then $\underset{\sim}{\beta}_r \mid \overline{H}_r$ is distributed as a normal variate with mean $n\overline{H}_r/(n+(1/\varphi_k))$ and precision $n + \frac{1}{\varphi_k}$. Additionally, the marginal distribution of $\underset{\sim}{\overline{H}}_r$ is normal with mean 0 and variance $\varphi_k + \frac{1}{n}$. In this case we have a "natural estimate" for β_r given by (4.13) with $\frac{1}{\varphi_k} = c_r$ whenever $n(r) = k$. We also have a "natural estimate" for $\frac{1}{n} + \varphi_k$ which is to let

$$\hat{D} = \frac{1}{\#\{r : n(r)=k\}} \sum_{r\,:\,n(r)=k} \overline{H}_r^2 \tag{4.15}$$

where $\#\{r : n(r)=k\}$ denotes the number of indices r having rank k. With this estimate for $\frac{1}{n} + \varphi_k$, we let

$$\hat{\varphi}_k = \hat{D} - \frac{1}{n} \quad . \tag{4.16}$$

At this point we have derived using Bayesian ideas and some rather general considerations our estimator for β_r under the assumption that the parameters β_r satisfy the condition that all coefficients of the same rank have the same precision. We see also that, all other considerations aside, the estimator given by (4.13) or (4.14) and (4.16) is extremely simple. We shall also see in Section 4.3 that a similar estimator may be derived without the Central Limit Theorem using a method of moments.

We have concentrated in this section on the estimation of the coefficients in (2.2). In principle, $\beta_r \doteq \alpha_r$ when α is near zero so that we can use

$$\hat{\alpha}_r = \frac{n\overline{H}_r}{n + \frac{1}{\varphi_k}} \tag{4.17}$$

as an estimate of α_s in (2.3). However, taking as an estimate of α_r the same estimate as for β_r could lead to a poor estimate for α_s. That is, if the α's are "too far" from zero then it might not be true that $\alpha_r \doteq \beta_r$. With this in mind we shall next take up the

problem of estimating a_r independently of β_r.

4.3 An Estimator for a_r

Before discussing an estimator for a_r we shall briefly consider a_r and β_r for certain specific ranks in order to determine what these coefficients are measuring.

We note here that if $n(r) = 1$ then $\beta_r = 0$. Suppose we now consider for a moment only those r's with rank two.

For notation we shall let ii denote an entry of two in the ith component of r and ij denote an entry of one in each of the i and j components of r. All other components will be zero since $n(r) = 2$.

If we consider our estimator for β_r given in (4.13) we see that it is a simple linear function of $\overline{H}_r$. If we restrict ourselves to those r's with rank two we then have

$$\overline{H}_r = \frac{1}{n} \sum_{t=1}^{n} H_r({}^t u)$$

with

$$\overline{H}_{ii}({}^t u) = H_2({}^t u_i) \tag{4.18}$$

$$= \frac{1}{\sqrt{2}} ({}^t u_i^2 - 1)$$

and

$$H_{ij}({}^{t}u) = H_i({}^{t}u_i)H_1({}^{t}u_j)$$

$$= {}^{t}u_i\,{}^{t}u_j\,.$$

The relationships in (4.18) may be rewritten as

$$\overline{H}_{ii} = \frac{1}{\sqrt{2}}\,(\sigma_i^2 - 1) \tag{4.19}$$

and

$$\overline{H}_{ij} = \sigma_{ij}$$

where

$$\sigma_i^2 := \frac{1}{n}\sum_{t=1}^{n} {}^{t}u_i^2$$

and

$$\sigma_{ij} := \frac{1}{n}\sum_{t=1}^{n} {}^{t}u_i\,{}^{t}u_j\,.$$

From these relationships we see that when the rank of r is two, $\hat{\beta}_r$ is a simple linear function of the sample covariances.

Suppose we now consider (2.3) and introduce a prior on the α's so that all α's of rank larger than two are zero. We note that in this case if $n(r) = 1$ then $\alpha_r = 0$. We now write (2.3) as

$$p(u) = p_0(u)\, e^{\sum_{r\neq 0}\alpha_r H_r(u) - A(\alpha)} \tag{4.20}$$

$$= (2\pi)^{-\frac{m}{2}}\, e^{-\frac{1}{2}\sum_1^m u_i^2}\, e^{\sum_{r:n(r)=2}\alpha_r H_r(u) - A(\alpha)}\,.$$

The exponent in (4.20) may be written as

$$-\frac{1}{2}\sum_{i=1}^{m} u_i^2 + \sum_{n(r)=2} a_r H_r(u)$$

$$= \frac{1}{2}\sum_{i=1}^{m} u_i^2 + \frac{\sqrt{2}}{2}\sum_{i=1}^{m} a_{ii}(u_i^2 - 1) + a_{12}u_1u_2 + \ldots + a_{1m}u_1u_m$$

$$+ a_{23}u_2u_3 + \ldots$$

$$= -\frac{1}{2}\sum_{i=1}^{m}(1-\sqrt{2}\,a_{ii})u_i^2 - \sum_{i=1}^{m}\frac{a_{ii}}{\sqrt{2}} + \sum_{i=1}^{m}\sum_{j=i+1}^{m} a_{ij}u_iu_j$$

$$= -\frac{1}{2}\sum_{i=1}^{m}(1-\sqrt{2}\,a_{ii})u_i^2 - 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} a_{ij}u_iu_j - \sum_{i=1}^{m}\frac{a_{ii}}{\sqrt{2}}\,.$$

Now the quadratic form within the braces is easily recognized and (4.20) may be written as

$$p(u) = (2\pi)^{-m/2} e^{-\frac{1}{2}u^T \Sigma^{-1} u - \sum_{i=1}^{m}\frac{a_{ii}}{\sqrt{2}} - A(a)}$$

where Σ^{-1} is the symmetric matrix

$$\begin{bmatrix} 1-\sqrt{2}\,a_{11} & -a_{12} & \cdots & -a_{1m} \\ & 1-\sqrt{2}\,a_{22} & & -a_{2m} \\ & & & \vdots \\ & & & 1-\sqrt{2}\,a_{mm} \end{bmatrix}$$

$A(\alpha)$ is the normalizing constant and when

$$A(\alpha) = -\frac{1}{2}\left\{ \sum_{i=1}^{m} \sqrt{2}\, \alpha_{ii} + \ln[|\Sigma^{-1}|] \right\}$$

we may write (4.20) as

$$p(u) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{m/2}} e^{-\frac{1}{2} u^T \Sigma^{-1} u} \tag{4.21}$$

We may summarize here that when the rank of r is two, β_r is a simple linear function of the sample covariances. It is also true that from the relationship $\beta_r = E[H_r(\underset{\sim}{u})|p]$ we can conclude that when $n(r) = 2$, β_r is a simple linear function of the covariance matrix Σ. If in addition $n(r) > 2$ implies $\alpha_r = 0$ then α_r with $n(r) = 2$ is a simple linear function of the precision matrix Σ^{-1}. That is, to estimate β_r when $n(r) = 2$ is equivalent to estimating Σ and to estimate α_r when $n(r) = 2$ and when $n(r) > 2$ implies $\alpha_r = 0$ is equivalent to estimating Σ^{-1}. We also note here that the condition $n(r) > 2$ implies $\alpha_r = 0$ tacitly assumes that $p(u)$ is an m-dimensional normal density.

Estimation of the Precision Matrix of a Multivariate Normal Distribution. As in Section 4.2 let us suppose that we have the random observations ${}^t\underset{\sim}{U}$, $t = 1, \ldots, n$ from one of our classes.

Furthermore we make the assumption that the distribution of ${}^{t}\underset{\sim}{U}$ is multivariate normal with a non singular covariance matrix denoted by Σ_i. We will henceforth suppress the i and simply write Σ. The joint density of the n random vectors is then given by

$$(4.22) \qquad ({}^{1}\underset{\sim}{U}, \ldots, {}^{n}\underset{\sim}{U} | R) \sim (2\pi)^{-mn/2} |R|^{n/2} e^{-\frac{1}{2}\Sigma_{t=1}^{n} {}^{t}u^{T} R {}^{t}u}$$

where $R := \Sigma^{-1}$. Let us also suppose that the precision matrix R is given a prior distribution that is a Wishart distribution with ν degrees of freedom and precision τ so that we have the density

$$(4.23) \qquad (R | \nu, \tau) \sim c |\tau|^{\nu/2} |R|^{\nu-m-1/2} e^{-\frac{1}{2} \operatorname{tr}(\tau R)}$$

where

$$c := \left\{ 2^{\nu m/2} \pi^{m(m-1)/4} \prod_{i=1}^{m} \tau\left(\frac{\nu+1-i}{2}\right) \right\}^{-1}$$

and

$$\nu > m-1 .$$

In order to estimate the a's our aim is to find estimates of the a's in terms of the sample random vectors ${}^{1}\underset{\sim}{U}, \ldots, {}^{n}\underset{\sim}{U}$. Inasmuch as estimating a_r for $n(r) = 2$ when $n(s) > 2$ implies $a_s = 0$ is equivalent to estimating R we focus our attention on R. We have assumed a prior on $\underset{\sim}{R}$ that is $W(\nu, \tau^{-1})$ so that we have from (4.6) and (4.7) the result that the posterior distribution of $\underset{\sim}{R}$

is also Wishart. In fact,

$$(\underset{\sim}{R} \mid {}^{1}U, \ldots, {}^{n}U, \nu, \tau) \sim W(\nu+n, \tau^{*-1})$$

where

$$\tau^{*} = \tau + \mathcal{U}\mathcal{U}^{T} \tag{4.24}$$

and

$$\mathcal{U}\mathcal{U}^{T} := \sum_{t=1}^{n} {}^{t}u\,{}^{t}u^{T} .$$

Since ${}^{t}u^{T}R\,{}^{t}u$ is a scalar for each $t = 1, \ldots, n$, we have, using properties of the trace operator

$$\begin{aligned} \text{tr} \sum_{t=1}^{n} {}^{t}u^{T}R\,{}^{t}u &= \sum_{t=1}^{n} \text{tr}[{}^{t}u^{T}R\,{}^{t}u] \\ &= \sum_{t=1}^{n} \text{tr}[{}^{t}u\,{}^{t}u^{T}R] \\ &= \text{tr}\left[\left(\sum_{t=1}^{n} {}^{t}u\,{}^{t}u^{T}\right) R\right] \\ &= \text{tr}[\mathcal{U}\mathcal{U}^{T}R] . \end{aligned} \tag{4.25}$$

With this relationship we note that (4.22) depends upon the observations only through $\mathcal{U}\mathcal{U}^{T}$ so that $\mathcal{U}\mathcal{U}^{T}$ is a sufficient statistic for R.

If ν and τ are specified then the posterior distribution of R will give us an estimate of R based upon the sufficient statistic $\mathcal{U}\mathcal{U}^T$. An estimate that comes immediately to mind is the mean of the posterior. That is, by Lemma 3.2.7 we have as an estimate $(\nu+n)\tau^{*-1}$. From the relationship $R = \Sigma^{-1}$ and Lemma 3.4.2 we have as another candidate

$$\hat{R} = \{E[R^{-1} \mid {}^1U, \ldots, {}^nU, \nu, \tau]\}^{-1} \tag{4.26}$$

$$= (\nu+n-m-1)\tau^{*-1} .$$

We comment here that the estimate given in (4.26) has certain intuitive appeal when τ is of a certain form which will be discussed later.

If ν and τ are specified by the investigator then we have an estimate of R available. We next address ourselves to the problem of specifying ν and τ.

We may view $p_0(u)$ as initially approximating the density $p(u)$ after the Karhunen-Loève transformation. From this viewpoint we initially estimate R^{-1} by I. Under the additional assumption that $(\underset{\sim}{R} \mid \nu, \tau) \sim W(\nu, \tau^{-1})$ we have by Lemma 3.4.2

$$E[\underset{\sim}{R}^{-1} \mid \nu, \tau] = \frac{1}{\nu-m-1}\tau .$$

Combining these two ideas we set

$$\frac{1}{\nu\text{-m-1}}\tau = I$$

so that

$$\tau = (\nu\text{-m-1})I .$$

We have used these interpretations to derive τ as a function of ν and thereby reduce the problem of specifying ν and τ to a problem of specifying only ν.

Putting aside for a moment any previous estimators for R but retaining the assumption that $\tau = (\nu\text{-m-1})I$, we could use a procedure similar to that suggested by Lindley and Smith (1971). That is, we could think of ν as a hyperparameter and give $\underset{\sim}{\nu}$ a prior distribution. We then would consider the joint posterior distribution of $\underset{\sim}{R}$ and $\underset{\sim}{\nu}$ and integrate with respect to ν to obtain the marginal posterior distribution of $\underset{\sim}{R}$. We then would consider some function of the marginal posterior moments as an estimate of R.

We choose to pursue a different option. That is, we choose to let the data furnish direct estimates of ν.

At this point let us recall the discussion involving equations (4.4) through (4.7). In the case that $\tau = (\nu\text{-m-1})I$ we have the constants in (4.6) given by

$$b^* = \nu - m - 1$$

$$a_{ii}^* = -\frac{1}{2}$$

$$a_{ij}^* = 0 \ .$$

The constants in (4.7) are given by

$$b = \nu + n - m - 1$$

$$a_{ii} = \frac{n\bar{w}_{ii} + b^* a_{ii}^*}{n + b^*}$$

$$a_{ij} = \frac{n\bar{w}_{ij}}{n + b^*} \ .$$

If we now apply Lemma 4.1.1 we have

(4.27)
$$a_{ii} = E\{E[\underset{\sim}{W}_{ii} | R] |^1 U, \ldots, {}^n U\}$$

$$= -\frac{1}{2} E\{E[\underset{\sim}{u}_i^2 | R] |^1 U, \ldots, {}^n U\}$$

$$= -\frac{1}{2} E\{\underset{\sim}{\sigma}_i^2 |^1 U, \ldots, {}^n U\}$$

and

(4.28)
$$a_{ij} = E\{E[\underset{\sim}{W}_{ij} | R] |^1 U, \ldots, {}^n U\}$$

$$= -E\{E[\underset{\sim}{u}_i \underset{\sim}{u}_j | R] |^1 U, \ldots, {}^n U\}$$

$$= -E\{\underset{\sim}{\sigma}_{ij} |^1 U, \ldots, {}^n U\}$$

What we have here is that when $\tau = (\nu-m-1)I$ and the posterior distribution is written as in (4.7) the a_{ij} terms in the posterior distribution give us the Bayes estimates for β_r when $n(r) = 2$.

With this development we shall digress for a moment and compare this estimate with the one derived in Section 4.2.

Recall from Section 4.2 that when $n(r) = 2$, the estimator takes the form

$$\hat{\beta}_{ij} = \frac{n\overline{H}_{ij}}{n+c_k}$$

where c_k is the precision common to all β_r where $n(r) = 2$ and

$$\overline{H}_{ij} = \frac{1}{n}\sum_{t=1}^{n} H_{ij}({}^{t}u) \ .$$

Also,

$$H_{ij}(u) = \begin{cases} \frac{1}{\sqrt{2}}(u_i^2-1) & \text{if } i = j \\ u_i u_j & \text{if } i \neq j \end{cases} .$$

Therefore,

$$\overline{H}_{ii} = \frac{1}{n\sqrt{2}}\sum_{t=1}^{n}({}^{t}u_i^2-1) \qquad i = j$$

$$\overline{H}_{ij} = \frac{1}{n}\sum_{t=1}^{n} {}^{t}u_i \, {}^{t}u_j \qquad i \neq j.$$

From the relationships in (4.4) and this last relationship we can write

$$\overline{w}_{ii} = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{2}t_{u_i}^2 \qquad i = j$$

$$\overline{w}_{ij} = \frac{1}{n}\sum_{t=1}^{n} -t_{u_i}t_{u_j} \qquad i \neq j$$

$$= -\frac{1}{n}\sum_{t=1}^{n} t_{u_i}t_{u_j}$$

$$= -\overline{H}_{ij}\,.$$

We can now write $\hat{\beta}_{ij}$ as

(4.29)
$$\hat{\beta}_{ij} = \frac{n\overline{H}_{ij}}{n+c_k}$$

$$= -\frac{n\overline{w}_{ij}}{n+c_k}$$

$$= \left(\frac{n+b^*}{n+c_k}\right)(-a_{ij}) \qquad \text{if} \quad i \neq j.$$

If $b^* = c_k$, we then have

$$\hat{\beta}_{ij} = -a_{ij} \qquad \text{if} \quad i \neq j.$$

That is, if $b^* = c_k$ then the estimate $\hat{\beta}_{ij}$ and the Bayes estimate are one and the same. Note, if we replace c_k by $\hat{c}_k$ where $\hat{c}_k = \frac{1}{\hat{\varphi}_k}$ in (4.29) then $\hat{\beta}_{ij}$ and a_{ij} may still both be considered as estimators of σ_{ij} when $i \neq j$. If $\hat{c}_k = b^*$ then they are the same estimators.

(4.30)
$$\hat{\beta}_{ij} = \frac{n\overline{H}_{ii}}{n+c_k}$$

$$= \frac{-\frac{2}{\sqrt{2}}\overline{w}_{ii} - \frac{n}{2}}{n+c_k} .$$

If $b^* = c_k$ then

$$\hat{\beta}_{ii} = \frac{-\frac{2}{\sqrt{2}}[(n+b^*)a_{ii} - b^*a_{ii}^*] - \frac{n}{2}}{n+b^*}$$

$$= -\frac{1}{\sqrt{2}}(2a_{ii}+1) .$$

We note that from (4.27), a_{ii} is an estimate of $-\frac{1}{2}\sigma_i^2$ so that

$$-\hat{\sigma}_i^2 = 2a_{ii}$$

and we may write

$$\hat{\beta}_{ii} = \frac{1}{\sqrt{2}}(\sigma_i^2 - 1) .$$

If we replace c_k by $\hat{c}_k$ in (4.30) then

$$\hat{\beta}_{ii} = \frac{\frac{1}{\sqrt{2}}(\hat{\sigma}_i^2 - 1)^2 - \frac{1}{\sqrt{2}}n}{(\hat{\sigma}_i^2 - 1)}$$

$$\doteq \frac{1}{\sqrt{2}}(\hat{\sigma}_i^2 - 1) .$$

Once again, with the assumption that $b^* = c_k$ or $b^* = \hat{c}_k$ and the fact that $2a_{ii} = -\hat{\sigma}_i^2$ we get the same estimate as in Section 4.2.

At this point we note that in Section 4.2 we derived an estimator for β_r under some very general assumptions. Under the assumption of a prior on $\underset{\sim}{R}$ that is $W(\nu, \tau^{-1})$ with τ being equal to $(\nu-m-1)I$ and considering the posterior mean of $\underset{\sim}{R}^{-1}$ as an estimate of the covariance structure we have just observed that our original estimate is appropriate in this special case with b^* equal to c_k or $\hat{c}_k$.

From these considerations we have

$$\hat{c}_k = \frac{1}{\hat{\varphi}_k}$$

and

$$\hat{c}_k = b^*$$

$$= \nu - m - 1 .$$

We set

(4.31) $$\hat{\nu} = m+1 + \frac{1}{\hat{\varphi}_k}$$

$$\hat{\varphi}_k + \frac{1}{n} = \left\{ \frac{2}{m(m+1)} \sum_{r:\, n(r)=2} \overline{H}_r^2 - \frac{1}{n} \right\}^{-1}$$

and take

$$\hat{\nu} = m+1 + \left\{ \frac{2}{m(m+1)} \sum_{r:\, n(r)=2} \bar{H}_r^2 - \frac{1}{n} \right\}^{-1}$$

to be an estimate of ν.

We now have an estimate for ν and have expressed τ as a function of ν. That is, we have a specification for ν and τ as was required in the estimation process for a_r when $n(r) = 2$.

We now consider one final aspect of the problem of determining ν.

We first note that again by (4.22) and (4.25), $\mathcal{U}\mathcal{U}^T$ is a sufficient statistic for ν. Conceptually we consider finding the joint distribution of $\underset{\sim}{\mathcal{U}}\underset{\sim}{\mathcal{U}}^T$ and $\underset{\sim}{R}$ given ν and τ then integrate with respect to R. This will leave us with the marginal density of $\underset{\sim}{\mathcal{U}}\underset{\sim}{\mathcal{U}}^T$ in terms of ν and τ. We then consider the first few moments of this distribution and choose some function of the lower order moments as an estimate of ν.

In point of fact, we can find some of the lower order moments without specifically finding the marginal distribution.

Lemma 4.3.1. If $\tau = (\nu-m-1)I$ then $E[\underset{\sim}{\mathcal{U}}\underset{\sim}{\mathcal{U}}^T \mid \nu, \tau] = nI$.

Proof: We first note that $(\underset{\sim}{\mathcal{U}}\underset{\sim}{\mathcal{U}}^T \mid R, \nu, \tau) \sim W(n, R^{-1})$ and by Lemma 3.2.5

$$E[\,\mathcal{U}\mathcal{U}^T \mid R, \nu, \tau] = nR^{-1} .$$

$\underset{\sim}{R}$ being Wishart implies by Lemma 3.4.2,

$$E[n\underset{\sim}{R}^{-1}] = \frac{n}{\nu-m-1}\tau .$$

The result now follows from the formula

$$E[\,\underset{\sim}{\mathcal{U}}\,\underset{\sim}{\mathcal{U}}^T \mid \nu, \tau] = E\{E[\,\underset{\sim}{\mathcal{U}}\,\underset{\sim}{\mathcal{U}}^T \mid \underset{\sim}{R}, \nu, \tau] \mid R\} . \quad \blacksquare$$

If we set $\underset{\sim}{S} := \underset{\sim}{\mathcal{U}}\,\underset{\sim}{\mathcal{U}}^T$ so that $\underset{\sim}{S}_{ij} = n\,\overline{\underset{\sim}{u}_i \underset{\sim}{u}_j}$ for $i, j = 1, \ldots, m$ then Lemma 4.3.1 gives us $E[\underset{\sim}{S}_{ij} \mid \nu, \tau] = n\delta_{ij}$. Suppose we now consider functions of the second moments of $\underset{\sim}{S}_{ij}$.

Lemma 4.3.2. If $\tau = (\nu-m-1)I$ then

$$\operatorname{var}(\overline{\underset{\sim}{u}_i \underset{\sim}{u}_j} \mid \nu, \tau) = \begin{cases} \frac{2}{n} + \frac{2(\frac{2}{n}+1)}{\nu-m-3} & \text{if } i = j \\ \frac{(\nu-m-1)(\nu+n-m-1)}{n(\nu-m)(\nu-m-3)} & \text{if } i \neq j \end{cases}$$

Proof: The proof of this lemma rests on the conditional variance formula

$$\operatorname{var}(\underset{\sim}{S}_{ij} \mid \nu, \tau) = E[\operatorname{var}(\underset{\sim}{S}_{ij} \mid R, \nu, \tau)] + \operatorname{var}[E(\underset{\sim}{S}_{ij} \mid R, \nu, \tau)].$$

Since $(\underset{\sim}{R} \mid \nu, \tau) \sim W(\nu, \tau^{*-1})$, $\underset{\sim}{R}^{-1}$ follows an inverted Wishart distribution with $\nu+m+1$ degrees of freedom and parametric matrix

τ. With $\Sigma = R^{-1}$ we have by Lemmas 3.4.2 and 3.4.3 the following.

$$E(\underset{\sim}{\sigma}_{ij}) = \frac{1}{\nu-m-1}\tau_{ij} = \delta_{ij} \tag{4.32}$$

$$E(\underset{\sim}{\sigma}_{ij}^2) = \operatorname{var}\underset{\sim}{\sigma}_{ij} + [E\underset{\sim}{\sigma}_{ij}]^2 = \operatorname{var}\underset{\sim}{\sigma}_{ij} + \delta_{ij}\,. \tag{4.33}$$

$$\operatorname{var}\underset{\sim}{\sigma}_{ij} = \begin{cases} \dfrac{2}{\nu-m-3} & \text{if } i = j \\ \dfrac{(\nu-m-1)}{(\nu-m)(\nu-m-3)} & \text{if } i \neq j\,, \end{cases} \tag{4.34}$$

Combining (4.33) and (4.34),

$$E(\underset{\sim}{\sigma}_{ij}^2) = \begin{cases} \dfrac{\nu-m-1}{\nu-m-3} & \text{if } i = j \\ \dfrac{\nu-m-1}{(\nu-m)(\nu-m-3)} & \text{if } i \neq j\,. \end{cases} \tag{4.35}$$

Again by Lemma 3.4.3

$$\operatorname{cov}(\underset{\sim}{\sigma}_{ii}, \underset{\sim}{\sigma}_{jj}) = \begin{cases} \dfrac{2}{\nu-m-3} & \text{if } i = j \\ \dfrac{2}{(\nu-m)(\nu-m-3)} & \text{if } i \neq j. \end{cases} \tag{4.36}$$

From the relationships

$$E\underset{\sim}{\sigma}_{ii}\underset{\sim}{\sigma}_{jj} = \operatorname{cov}(\underset{\sim}{\sigma}_{ii}, \underset{\sim}{\sigma}_{jj}) + E\underset{\sim}{\sigma}_{ii}E\underset{\sim}{\sigma}_{jj}\,,$$

(4.32) and (4.36) we have

$$(4.36) \qquad E\underset{\sim}{\sigma}_{ii}\underset{\sim}{\sigma}_{jj} = \begin{cases} \dfrac{\nu-m-1}{\nu-m-3} & \text{if } i = j \\ \dfrac{2}{(\nu-m)(\nu-m-3)} + 1 & \text{if } i \neq j . \end{cases}$$

By noting that $(\underset{\sim}{S}|R, \nu, \tau) \sim W(n, R^{-1})$ and using Lemma 3.2.6 we have

$$(4.38) \qquad E(\underset{\sim}{S}_{ij}|R, \nu, \tau) = n\sigma_{ij}$$

$$\text{var}(\underset{\sim}{S}_{ij}|R, \nu, \tau) = n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})$$

Therefore,

$$\text{var}(\underset{\sim}{S}_{ij}|R, \nu, \tau) = nE\underset{\sim}{\sigma}_{ij}^2 + nE\underset{\sim}{\sigma}_{ii}\underset{\sim}{\sigma}_{jj}$$

Using (4.32)-(4.38) we have

$$(4.39) \qquad \text{var}(\underset{\sim}{S}_{ij}|\nu, \tau) = \begin{cases} \dfrac{2n(\nu-m-1)}{\nu-m-3} + \dfrac{2n^2}{\nu-m-3} & \text{if } i = j \\ \dfrac{n(\nu-m+1)}{(\nu-m)(\nu-m-3)} + \dfrac{n^2(\nu-m-1)}{(\nu-m)(\nu-m-3)} + n & \text{if } i \neq j. \end{cases}$$

Noting that

$$\text{var}(\overline{\underset{\sim}{u}_i \underset{\sim}{u}_j}|\nu, \tau) = \text{var}(\frac{1}{n}\underset{\sim}{S}_{ij}|\nu, \tau) ,$$

the result follows. ∎

We note that

$$\overline{H}_{ij} = \begin{cases} \frac{1}{\sqrt{2}}\,\overline{u_i^2} - \frac{1}{\sqrt{2}} & \text{if} \quad i = j \\ \overline{u_i u_j} & \text{if} \quad i \neq j \end{cases}$$

so that

$$\operatorname{var}(\overline{H}_{ij}) = \begin{cases} \frac{1}{2}\operatorname{var}(\overline{u_i^2}) & \text{if} \quad i = j \\ \operatorname{var}(\overline{u_i u_j}) & \text{if} \quad i \neq j\,. \end{cases}$$

Moreover it follows from Lemma 4.3.1 that $E(\overline{H}_{ii}) = 0$ for all i and j.

From Lemma 4.3.2 we can estimate ν by a method of moments. For example we might choose the case $i = j$ when

$$\operatorname{var}(\overline{H}_{ii}) = \frac{1}{2}\operatorname{var}\,\overline{u_i^2}$$

$$= \frac{1}{n} + \frac{\frac{2}{n}+1}{\nu-m-3}\,.$$

Since $E(\overline{H}_{ii}) = 0$ we have a "natural" estimate for $\operatorname{var}(\overline{H}_{ii})$ given by

$$\frac{1}{m}\sum_{i=1}^{m} \overline{H}_{ii}^2\,.$$

We set

$$\frac{\frac{2}{n}+1}{\hat{\nu}-m-3} = \frac{1}{m}\sum_{i=1}^{m} \overline{H}_{ii}^2 - \frac{1}{n}$$

so that

$$\hat{\nu} = m+3 + (\frac{2}{n}+1)\left\{\frac{1}{m}\sum_{i=1}^{m}\overline{H}_{ii}^{2} - \frac{1}{n}\right\}^{-1}$$

and take $\hat{\nu}$ to be an estimate of ν.

Similarly, we could choose the case $i \neq j$ to derive an estimate of ν. The case $i \neq j$ is not as simple as the case $i = j$ inasmuch as

$$\text{var}(\underset{\sim}{\overline{H}}_{ii}) = \frac{(\nu-m-1)(\nu+n-m-1)}{n(\nu-m)(\nu-m-3)} .$$

We note here that we get results that are quite similar to the estimates derived using the assumptions of Section 4.2. That is, the estimate of ν given through (4.40) are quite similar to the estimate of ν given by (4.31). The estimate given by (4.31) has one very attractive feature--ease of computation.

In this section we have concentrated on deriving an estimator for the a_r when $n(r) = 2$ and $n(s) > 2$ implies $a_s = 0$. We have noted in the process that estimating the a's of rank two is equivalent to estimating the precision matrix R and estimating the β's of rank two is equivalent to estimating the covariance structure Φ.

4.4 Properties of Estimators

In Section 4.2 we developed the formula for $\hat{\beta}_r$ given by (4.13) from rather general considerations. We also noted two properties of $\hat{\beta}_r$:

(i) $\hat{\beta}_r$ is biased for β_r

(ii) $\hat{\beta}_r$ is consistent for β_r.

In this section we further investigate properties of estimates based on (4.13) as well as estimates based on the formula given by (4.26).

We comment here that the formula

$$\hat{\beta}_r = \frac{n\overline{H}_r}{n+c_r}$$

gives a way of estimating β_r for each r in the expansion

$$\frac{p(u)}{p_0(u)} = \sum_r \beta_r H_r(u) .$$

However, the fundamental estimation problem is not to estimate β_r for each r but rather to estimate $p(u)$.

In most practical situations we are not interested in estimating the entire series $\sum_r \beta_r H_r(u)$ but rather in considering a convenient truncation of this series and using this truncation to estimate $p(u)$.

With this in mind we recall that the notation $\sum_r \beta_r H_r(u)$ represents the series $\sum_{c=0}^{\infty} \sum_{r: n(r)=c} \beta_r H_r(u)$. We now introduce the following notation: let $\underline{H}_k(u)$ and B_k be the vectors whose components are $H_r(u)$ and β_r respectively when $n(r) = k$. We also consider the vector b defined by

$$(b_0, b_1, \ldots, b_\ell)^T := (B_0, B_1, \ldots, B_k)^T$$

where b_0 corresponds to the first element of B_0 and b_ℓ corresponds to the last element of B_k. We also denote by $H^s(u)$ that polynomial $H_r(u)$ that corresponds to b_s for $s = 1, \ldots,$.

Now let us suppose we wish to approximate $p(u)$ by a truncation of the series at $n(r) = k$. That is, we wish to use only those terms of the expansion (2.2) through rank k.

We have by (2.2),

$$\frac{p(u)}{p_0(u)} = \sum_r \beta_r H_r(u)$$

$$= \sum_{c=0}^{\infty} \sum_{r: n(r)=c} \beta_r H_r(u) .$$

We set

$$\frac{p_k(u)}{p_0(u)} := \sum_{r:\, n(r) \le k} \beta_r H_r(u) \tag{4.41}$$

$$= \sum_{c=0}^{k} \sum_{r:\, n(r)=c} \beta_r H_r(u)$$

$$= \sum_{s=0}^{\ell} b_s H^s(u) \, .$$

Now consider any other linear combination of the $H_r(u)$ where $n(r) \le k$. Let us denote this arbitrary linear combination by

$$\frac{\pi_k(u)}{p_0(u)} := \sum_{s=0}^{\ell} a_s H^s(u)$$

Since $H^s(u)$ is a complete orthonormal family in $L_2(P_0)$, Parsevals equation is valid. That is,

$$\left\| \frac{p(u)}{p_0(u)} \right\|^2 = \sum_{s=0}^{\infty} b_s^2 \, .$$

Now, it is a well known result [e.g., Natanson (1965)] that the error as measured by the $L_2(P_0)$ norm in approximating $\frac{p(u)}{p_0(u)}$ by $\frac{\pi_k(u)}{p_0(u)}$ is given by

$$(4.42) \qquad \left\| \frac{p(u)}{p_0(u)} - \frac{\pi_k(u)}{p_0(u)} \right\|^2 = \sum_{s=0}^{\infty} b_s^2 - \sum_{s=0}^{\ell} b_s^2 + \sum_{s=0}^{\ell} (a_s - b_s)^2 .$$

From this result we see that the error in terms of the $L_2(P_0)$ norm is minimum when $\pi_k(u) = p_k(u)$. That is, when $a_s = b_s$, $s = 0, 1, 2, \ldots, n$.

Having determined that the best $L_2(P_0)$ approximation to $\frac{p(u)}{p_0(u)}$ by a linear combination of the $H_r(u)$ for which $n(r) \leq k$ is given by (4.41) we next consider as an estimate of $\frac{p(u)}{p_0(u)}$

$$(4.42) \qquad \overline{\frac{p_k(u)}{p_0(u)}} := \sum_{r:\, n(r) \leq k} \hat{\beta}_r H_r(u) .$$

We now consider properties of the expected mean square error in using (4.42) to estimate $\frac{p(u)}{p_0(u)}$.

We have immediately the following results.

Lemma 4.4.1. For k fixed, the expected mean squared error of $\overline{\frac{p_k(u)}{p_0(u)}}$ is

$$\sum_{r:\, n(r)>k} \beta_r^2 + \sum_{r:\, n(r) \leq k} \{ \operatorname{var}(\hat{\beta}_r | p) + \beta_r^2 \left(\frac{n}{n+c_r} - 1 \right)^2 \}$$

Proof: The result is immediate by setting

$$\frac{\overline{p_k(u)}}{p_0(u)} = \frac{\pi_k(u)}{p_0(u)}$$

in (4.42), taking the expectation of (4.42) with respect to $p(u)$ and recalling from Section 4.2 that

$$E[(\hat{\beta}_r - \beta_r)^2 | p] = \text{var}(\hat{\beta}_r | p) + \beta_r^2 \left(\frac{n}{n+c_r} - 1\right)^2 . \quad \square$$

Lemma 4.42. For fixed k, $\lim_{n \to \infty} E \left\| \frac{p(u)}{p_0(u)} - \frac{\overline{p_k(u)}}{p_0(u)} \right\|^2 = \sum_{n(r)>k} \beta_r^2 .$

Proof: This follows immediately from Lemma 4.4.1 and the consistency of $\hat{\beta}_r$ for β_r derived in Section 4.2. $\square$

We now turn our attention to the estimators developed in Section 4.3 under the assumption that when $n(s) > 2$, $a_s = 0$.

We recall that this assumption is equivalent to assuming $p(u)$ is a multivariate normal density. Moreover, with this assumption the problem of estimating a_r when $n(r) = 2$ is identical to estimating R, the inverse of the covariance matrix. We also recall that having assumed a prior on $\underset{\sim}{R}$ that is $W(\nu, \tau^{-1})$ with $\tau = (\nu - m - 1)I$ we found that the posterior distribution of $\underset{\sim}{R}$ was $W(\nu+n, \tau^{*-1})$ where $\tau^* = \tau + \mathcal{U}\mathcal{U}^T$ and $\mathcal{U}\mathcal{U}^T = \sum_{t=1}^{n} t_u t_u^T$. It is also true that the posterior distribution of $\underset{\sim}{\Sigma} = \underset{\sim}{R}^{-1}$ is inverted Wishart with

$$E\underset{\sim}{\Sigma} = \frac{1}{\nu+n-m-1}\tau^* .$$

We consider the posterior mean as an estimate of Σ so that we set

$$\hat{\Sigma} := \frac{1}{\nu+n-m-1}\tau^* .$$

Equation (4.26) gives an estimate of R based on the inverse of the posterior mean of Σ. That is, (4.26) gives

$$\begin{aligned}\hat{R} &= (\nu+n-m-1)\tau^{*-1} \\ &= (\hat{\Sigma})^{-1} .\end{aligned}$$

We now consider some of the properties of these estimates.

We set $S := \mathcal{U}\mathcal{U}^T$ and $\Sigma_0 := \frac{1}{n}S$ so that Σ_0 is the maximum likelihood estimate of Σ when sampling from a normal distribution with mean 0 and covariance matrix Σ.

With $\tau = (\nu-m-1)I$ we set $a := \nu-m-1$ and have the result

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{a+n}[aI+S] \\ &= \frac{a}{a+n}I + \frac{1}{a+n}S \\ &= \frac{a}{a+n}I + \frac{n}{a+n}\Sigma_0 .\end{aligned} \tag{4.43}$$

We may write (4.43) in component form as

$$\hat{\Sigma}_{ij} = \begin{cases} \frac{n}{a+n}(\Sigma_0)_{ij} & \text{if } i \neq j \\ \frac{a}{a+n} + \frac{n}{a+n}(\Sigma_0)_{ij} & \text{if } i = j. \end{cases}$$

From this equation we see that for $i \neq j$, $\hat{\Sigma}$ tends to be shifted from the maximum likelihood estimate towards zero while if $i = j$, $\hat{\Sigma}$ tends to be shifted from the maximum likelihood estimate towards one.

From (4.43) we see that

$$E\hat{\underset{\sim}{\Sigma}} = \frac{a}{a+n} I + \frac{n}{a+n} E\underset{\sim}{\Sigma}_0 \tag{4.44}$$

$$= \frac{a}{a+n} I + \frac{n}{a+n} \Sigma .$$

From this equation it is immediate that $\hat{\Sigma}$ is asymptotically unbiased.

We now consider the norm on the space of all $m \times m$ symmetric matrices given by

$$\|\Sigma\|^2 = \langle \Sigma, \Sigma \rangle$$

$$= \operatorname{tr} \Sigma^2 .$$

In the classical sense, recall that the relative efficiency of an unbiased estimate $\tilde{\Sigma}$ relative to the unbiased estimate $\bar{\Sigma}$ is defined by

$$r.e. := \frac{E\|\bar{\Sigma}-\Sigma\|^2}{E\|\bar{\bar{\Sigma}}-\Sigma\|^2}$$

An estimator $\bar{\Sigma}$ is called most efficient in case $r.e. \leq 1$ for all $\bar{\bar{\Sigma}}$. It is well known that Σ_0 is most efficient in the class of unbiased estimates. We note that in the classical sense the expectations are computed with respect to the distribution of $(\underset{\sim}{U}|\Sigma)$.

Lemma 4.4.3. Under conditions of sampling from a multivariate normal distribution with mean 0 and covariance matrix Σ, if $\hat{\Sigma}$ is given by (4.43) then

$$E\|\hat{\underset{\sim}{\Sigma}}-\Sigma\|^2 = (\frac{a}{a+n})^2 tr[I-2\Sigma+\Sigma^2] + \frac{n}{(a+n)^2}[(tr\ \Sigma)^2+tr(\Sigma^2)]$$

where the expectation is with respect to the distribution of $\underset{\sim}{U}$ given Σ.

Proof: Using the fact that $(\underset{\sim}{S}|\Sigma, \nu, \tau) \sim W(n, \Sigma)$

$$E\|\hat{\underset{\sim}{\Sigma}}-\Sigma\|^2 = E\ tr[\frac{a}{a+n} I + \frac{1}{a+n}\underset{\sim}{S}-\Sigma]^2$$

$$= tr\ E[(\frac{a}{a+n})^2 I + \frac{2a}{(a+n)^2}\underset{\sim}{S} - \frac{2a}{a+n}\Sigma + \frac{1}{(a+n)^2}\underset{\sim}{S}^2 - \frac{2}{(a+n)}\underset{\sim}{S}\Sigma+\Sigma^2]$$

$$= tr[(\frac{a}{a+n})^2 I + \frac{2an}{(a+n)^2}\Sigma - \frac{2a}{a+n}\Sigma + \Sigma^2 + \frac{1}{(a+n)^2}E\underset{\sim}{S}^2 - \frac{2}{(a+n)}E\underset{\sim}{S}\Sigma] =$$

$$= \operatorname{tr}[(\frac{a}{a+n})^2 I - 2(\frac{a}{a+n})^2 \underset{\sim}{\Sigma} + \frac{a-n}{a+n} \underset{\sim}{\Sigma}^2] + \operatorname{tr} E[\frac{1}{(a+n)^2} \underset{\sim}{S}^2]$$

$$= (\frac{a}{a+n})^2 \operatorname{tr}[I - 2\underset{\sim}{\Sigma} + \underset{\sim}{\Sigma}^2] - \operatorname{tr}[(\frac{n}{a+n})^2 \underset{\sim}{\Sigma}^2] + \operatorname{tr} E[\frac{1}{(a+n)^2} \underset{\sim}{S}^2]$$

$$= (\frac{a}{a+n})^2 \operatorname{tr}[I - 2\underset{\sim}{\Sigma} + \underset{\sim}{\Sigma}^2] - \operatorname{tr}[(\frac{n}{a+n})^2 \underset{\sim}{\Sigma}^2] + \frac{1}{(a+n)^2} E \operatorname{tr} \underset{\sim}{S}^2 .$$

Now,

$$E \operatorname{tr} \underset{\sim}{S}^2 = E \sum_i \sum_j \underset{\sim}{S}_{ij}^2$$

$$= \sum_i \sum_j E \underset{\sim}{S}_{ij}^2$$

$$= \sum_i \sum_j [\operatorname{var} \underset{\sim}{S}_{ij} + (E \underset{\sim}{S}_{ij})^2] .$$

From Lemmas 3.2.6 we have

$$E \operatorname{tr} \underset{\sim}{S}^2 = \sum_i \sum_j [n(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2) + n^2 \sigma_{ij}^2]$$

$$= n(\operatorname{tr} \underset{\sim}{\Sigma})^2 + n \operatorname{tr}(\underset{\sim}{\Sigma}^2) + n^2 \operatorname{tr}(\underset{\sim}{\Sigma}^2) .$$

Therefore,

$$E \| \hat{\underset{\sim}{\Sigma}} - \underset{\sim}{\Sigma} \|^2 = (\frac{a}{a+n})^2 \operatorname{tr}[I - 2\underset{\sim}{\Sigma} + \underset{\sim}{\Sigma}^2] + \frac{n}{(a+n)^2} [(\operatorname{tr} \underset{\sim}{\Sigma})^2 + \operatorname{tr}(\underset{\sim}{\Sigma}^2)]. \quad \blacksquare$$

Lemma 4.4.4. Under the same conditions as Lemma 4.4.3 we have for $\Sigma_0 = \frac{1}{n} S$,

$$E\|\underset{\sim}{\Sigma}_0 - \Sigma\|^2 = \frac{1}{n}[(\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)] .$$

Proof:

$$E\|\underset{\sim}{\Sigma}_0 - \Sigma\|^2 = E \operatorname{tr}(\frac{1}{n}\underset{\sim}{S} - \Sigma)^2$$

$$= \sum_i \sum_j E(\frac{1}{n}\underset{\sim}{S}_{ij} - \sigma_{ij})^2$$

$$= \sum_i \sum_j \operatorname{var}(\frac{1}{n}\underset{\sim}{S}_{ij})$$

$$= \sum_i \sum_j \frac{1}{n^2} \operatorname{var} \underset{\sim}{S}_{ij}$$

$$= \sum_i \sum_j \frac{1}{n^2} n(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2)$$

$$= \frac{1}{n}[(\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)] . \quad \|$$

We note here that $E\|\hat{\underset{\sim}{\Sigma}} - \Sigma\|^2 < E\|\underset{\sim}{\Sigma}_0 - \Sigma\|^2$ if and only if

$$\frac{an}{a+2n} \operatorname{tr}(I - \Sigma)^2 < [(\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)] .$$

From this relationship we have that $E\|\hat{\underset{\sim}{\Sigma}} - \Sigma\|^2 < E\|\underset{\sim}{\Sigma}_0 - \Sigma\|^2$ when

Σ is "close" to I, the prior mean of Σ,

Lemmas 4.4.3 and 4.4.4 allow us to write

$$\lim_{n\to\infty} \frac{E\|\underset{\sim}{\Sigma}_0-\Sigma\|^2}{E\|\underset{\sim}{\hat{\Sigma}}-\Sigma\|^2} = \lim_{n\to\infty} \frac{\frac{1}{n}[(\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2)]}{(\frac{a}{a+n})^2\operatorname{tr}(I-\Sigma)^2+\frac{n}{(a+n)^2}[(\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2)]}$$

$$= 1 .$$

We may conclude from this last expression that $\hat{\Sigma}$ is asymptotically most efficient.

We also have by Lemma 4.4.3 the following lemma.

Lemma 4.4.5. $\hat{\Sigma}$ is a consistent estimator of Σ.

Proof: By Lemma 4.4.3 we write

$$E\|\underset{\sim}{\hat{\Sigma}}-\Sigma\|^2 = (\frac{a}{a+n})^2\operatorname{tr}[I-\Sigma]^2 + \frac{n}{(a+n)^2}[(\operatorname{tr}\Sigma)^2+\operatorname{tr}(\Sigma^2)]$$

from which we conclude that

$$E\|\underset{\sim}{\hat{\Sigma}}-\Sigma\|^2 \to 0 \quad \text{as} \quad n\to\infty.$$

That is, $\hat{\Sigma} \to \Sigma$ in mean of order two and therefore $\hat{\Sigma} \to \Sigma$ in probability. ‖

These results involving asymptotic results are not surprising inasmuch as we can see from (4.43) that as $n \to \infty$, $\hat{\Sigma}$ is

essentially the same as $\hat{\Sigma}_0$.

We have been considering the quadratic loss function for Σ given by $L(\hat{\Sigma}, \Sigma) = tr[\hat{\Sigma}-\Sigma]^2$. Equivalently, we have been considering the loss function for R given by

$$L(\hat{R}, R) = tr[\hat{R}^{-1}-R^{-1}]^2$$

which is quadratic in R^{-1}.

It is well known that under quadratic loss the Bayes rule for estimating Σ is the mean of the posterior. That is, $\hat{\Sigma}$ is Bayes with respect to $tr[\hat{\Sigma}-\Sigma]^2$ for estimating Σ and equivalently $\hat{R} = \hat{\Sigma}^{-1}$ is Bayes with respect to $tr[\hat{R}^{-1}-R^{-1}]^2$ for estimating R.

Lemma 4.4.6. Under the quadratic loss function $L(\hat{\Sigma}, \Sigma) = tr[\hat{\Sigma}-\Sigma]^2$ the Bayes risk, $\mathcal{R}_0$ in estimating Σ by $\hat{\Sigma}_0$ is given by

$$\mathcal{R}_0 = \frac{1}{n}\left[\frac{(m^2+m)a^2+2am}{(a+1)(a-2)}\right]$$

Proof:

$$\mathcal{R}_0 = E_{\Sigma}\{E_{u|\Sigma} L(\underset{\sim}{\Sigma}_0, \underset{\sim}{\Sigma})\}$$

$$= E_{\Sigma}\{E_{u|\Sigma} tr[\underset{\sim}{\Sigma}_0-\underset{\sim}{\Sigma}]^2\}$$

From Lemma 4.4.4,

$$E_{u|\Sigma} \operatorname{tr}[\underset{\sim}{\Sigma}_0 - \Sigma]^2 = \frac{1}{n}[(\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)]$$

$$= \sum_i \sum_j \frac{1}{n}(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2) .$$

Therefore,

$$\mathcal{R}_0 = E_{\Sigma} \sum_i \sum_j \frac{1}{n}(\underset{\sim}{\sigma}_{ii}\underset{\sim}{\sigma}_{jj} + \underset{\sim}{\sigma}_{ij}^2)$$

$$= \frac{1}{n} \sum_i \sum_j [E\underset{\sim}{\sigma}_{ii}\underset{\sim}{\sigma}_{jj} + E\underset{\sim}{\sigma}_{ij}^2]$$

$$= \frac{1}{n} \sum_i \sum_j [\operatorname{cov}(\underset{\sim}{\sigma}_{ii}, \underset{\sim}{\sigma}_{jj}) + (E\underset{\sim}{\sigma}_{ii})(E\underset{\sim}{\sigma}_{jj})]$$

$$+ \sum_i \sum_j [\operatorname{var}\underset{\sim}{\sigma}_{ij} + (E\underset{\sim}{\sigma}_{ij})^2] .$$

Now, $\underset{\sim}{R}$ has prior that is $W(\nu, \tau^{-1})$ so that the distribution of $\underset{\sim}{\Sigma}$ is inverted Wishart with parameters $\nu+m+1$ and aI so we may invoke Lemmas 3.4.2 and 3.4.3 with $\tau_{ij} = a\delta_{ij}$. Therefore,

$$\mathcal{R}_0 = \frac{1}{n}\left\{\sum_i \sum_j [\operatorname{cov}(\underset{\sim}{\sigma}_{ii}, \underset{\sim}{\sigma}_{jj})] + \sum_i \sum_j 1 + \sum_i \sum_j \operatorname{var}\underset{\sim}{\sigma}_{ij} + \sum_i \sum_j \delta_{ij}\right\}$$

$$= \frac{1}{n}\left\{\sum_i \sum_j \frac{2a+2a^2\delta_{ij}}{(\nu-m)a(\nu-m-3)} + m^2 + \sum_i \sum_j \frac{a^2+(\nu-m+1)a\delta_{ij}}{(\nu-m)a(\nu-m-3)} + m\right\} =$$

$$= \frac{1}{n}\left\{m^2+m+\sum_i\sum_j \frac{2a+2a^2\delta_{ij}}{(\nu-m)(\nu-m-3)} + m^2+\sum_i\sum_j \frac{a+(\nu-m+1)\delta_{ij}}{(\nu-m)(\nu-m-3)}\right\}$$

$$= \frac{1}{n}\left\{m^2+m+\frac{2m^2}{(\nu-m)(\nu-m-3)}+\frac{2am}{(\nu-m)(\nu-m-3)}+\frac{am^2}{(\nu-m)(\nu-m-3)} + \frac{(\nu-m+1)m}{(\nu-m)(\nu-m-3)}\right\}$$

$$= \frac{1}{n}\left\{m^2+m+\frac{2m^2+2am+am^2+(a+2)m}{(a+1)(a-2)}\right\}$$

$$= \frac{1}{n}\left\{\frac{(m^2+m)a^2+2am}{(a+1)(a-2)}\right\} . \qquad \blacksquare$$

Lemma 4.4.7. Under the conditions of Lemma 4.4.6, the Bayes risk $\mathcal{R}$ in using the Bayes rule $\hat{\Sigma}$ is given by

$$\mathcal{R} = \frac{a^3m^2+a^2(a+2)m}{(a+n)^2(a+1)(a-2)} + \frac{n^2}{(a+n)^2}\mathcal{R}_0 .$$

Proof: From Lemma 4.4.3,

$$E_{u|\Sigma}\ \mathrm{tr}\,[\underset{\sim}{\hat{\Sigma}}-\Sigma]^2 = \left(\frac{a}{a+n}\right)^2 \mathrm{tr}[I-\Sigma]^2 + \frac{n}{(a+n)^2}[(\mathrm{tr}\,\Sigma)^2+\mathrm{tr}(\Sigma^2)] .$$

Therefore,

$$\mathcal{R} = \left(\frac{a}{a+n}\right)^2 E\ \mathrm{tr}[I-\Sigma]^2 + \frac{n}{(a+n)^2}E[(\mathrm{tr}\,\Sigma)^2+\mathrm{tr}(\Sigma^2)] .$$

From Lemma 4.4.4,

$$(\mathrm{tr}\,\Sigma)^2 + \mathrm{tr}(\Sigma^2) = nE_{u|\Sigma}\,\mathrm{tr}[\underset{\sim}{\Sigma}_0-\Sigma]^2 .$$

Therefore,

$$= \left(\frac{a}{a+n}\right)^2 E\ \mathrm{tr}[I-\Sigma]^2 + \left(\frac{n}{a+n}\right)^2 \mathcal{R}_0$$

$$= \left(\frac{a}{a+n}\right)^2 \sum_i \sum_j E(\delta_{ij}-\sigma_{ij})^2 + \left(\frac{n}{a+n}\right)^2 \mathcal{R}_0$$

$$= \left(\frac{a}{a+n}\right)^2 \sum_i \sum_j \mathrm{var}\ \sigma_{ij} + \left(\frac{n}{a+n}\right)^2 \mathcal{R}_0$$

$$= \frac{a^2}{(a+n)^2} \frac{am^2+(a+2)m}{(a+1)(a-2)} + \frac{n^2}{(a+n)^2} \mathcal{R}_0 \quad \|$$

It is the content of the next lemma that the Bayes risk in using $\hat{\Sigma}$ is strictly less than the Bayes risk in using Σ_0.

Lemma 4.4.8. Under the conditions of Lemmas 4.4.6 and 4.4.7,

$$\mathcal{R} < \mathcal{R}_0 .$$

Proof: From Lemmas 4.4.6 and 4.4.7,

$$\mathcal{R}_0 - \mathcal{R} = \left[1 - \frac{n^2}{(a+n)^2}\right]\mathcal{R}_0 - \frac{a^2}{(a+n)^2}\left[\frac{am^2+(a+2)m}{(a+1)(a-2)}\right]$$

$$= \frac{a^2+2an}{(a+n)^2} \frac{(m^2+m)a^2+2am}{n(a+1)(a-2)} - \frac{a^2[am^2+(a+2)m]}{(a+n)^2(a+1)(a-2)}$$

$$= \frac{1}{(a+n)^2(a+1)(a-2)} \left\{\left(\frac{a^2+2an}{n}\right)[(m^2+m)a^2+2am] - a^2[am^2+(a+2)m]\right\} =$$

$$= \frac{1}{(a+n)^2(a+1)(a-2)} \{\frac{a^4}{n} (m^2+m)+a^3m^2+a^3m + \frac{2a^3m}{n} + 2a^2m\}$$

$$= \frac{1}{(a+n)^2(a+1)(a-2)} \{\frac{a^4}{n} (m^2+m)+a^3(m^2+m)+2a^2m(\frac{a}{n}+1)\}$$

$$= \frac{a^2}{(a+n)^2(a+1)(a-2)} \{\frac{a^2}{n} (m^2+m)+a(m^2+m)+2m(\frac{a}{n}+1)\} .$$

We note that for the required moments of the inverted Wishart distribution to exist we require that $\nu-m-3 > 0$. That is, $a-2 > 0$. With this fact we see that the right hand side of the last equation is the product of two strictly positive terms and the result is established. ∎

As a final comment in this section we consider a problem that arises in the classification problem.

In certain classification problems particularly in the area of pattern recognition the problem of insufficient data occurs. By this we mean that we are trying to classify an m-component vector based on a sample of size n where n may be much smaller than m. If we are using certain linear discriminant functions it will be required that we invert our estimate of Σ. If we consider Σ_0 then we need $n \geq m$. A possible way around this problem is to use a feature extractor to select the "best" m_1 components from X where $n \geq m_1$. With the requirement that $m_1 \leq n$, it is easy to

see that the investigator might be required to eliminate components which he feels are pertinent to the problem.

An alternative would be to consider another estimator of Σ. The estimator $\hat{\Sigma}$ given by (4.43) has one very nice property with respect to this problem. Since

$$\hat{\Sigma} = \frac{a}{a+n} I + \frac{1}{a+n} S$$

$$= \frac{a}{a+n} I + \frac{n}{a+n} \Sigma_0$$

and $\frac{a}{a+n} > 0$ we have that $\hat{\Sigma}$ is positive definite and therefore invertible for every $n \geq 1$.

BIBLIOGRAPHY

Anderson, T.W. 1958. An introduction to multivariate statistical analysis. New York, Wiley. 374 p.

Barnett, Vic. 1973. Comparative statistical inference. New York, Wiley. 287 p.

Barndorff-Nielsen, Ole. 1973. Exponential families and conditioning. New York, Wiley.

Bellman, Richard. 1960. Introduction to matrix analysis. New York, McGraw-Hill. 328 p.

Brunk, H.D. and D.A. Pierce. 1974. Estimation of discrete multivariate densities for computer aided differential diagnosis of disease. Biometrika 61:493-501.

Burrill, Claude. 1972. Measure, integration, and probability. New York, McGraw-Hill. 464 p.

Crain, Bradford R. 1974. Estimation of distributions using orthogonal expansions. The Annals of Statistics 2:454-463.

Efron, B. and Carl Morris. 1972. Empirical Bayes on vector observations: An extension of Stein's method. Biometrika 59:335-347.

Efron, B. and Carl Morris. 1973. Stein's estimation rule and its competitors--an Empirical Bayes approach. Journal of the American Statistical Association 68:117-130.

Ferguson, T.S. 1967. Mathematical Statistics a decision theoretic approach. New York, McGraw-Hill. 396 p.

Fukunaga, K. 1972. Introduction to statistical pattern recognition. New York, Academic Press. 369 p.

Graybill, F. 1969. Introduction to matrices with applications in statistics. Belmont, Wadsworth. 372 p.

Jackson, D. 1941. Fourier series and orthogonal polynomials. Carus Mathematical Monograph No. 6. Menasha, Wisc. Banta. 234 p.

Kshirsagar, A. 1972. Multivariate analysis. New York, Marcel Dekker. 534 p.

Kronmanl, R. and Michael Tarter. 1968. The estimation of probability densities and accumulatives by Fourier series methods. Journal of the American Statistical Association 63:925-952.

Lancaster, H.O. 1969. The Chi-Squared Distribution. New York, Wiley. 356 p.

Lindley, D.V. and A.F.M. Smith. 1972. Bayes estimates for the linear model. Journal of the Royal Statistical Society. 34:1-41.

Lukacs, E. and R.G. Laha. 1964. Applications of characteristic functions. London. Griffin. 202 p.

Martin, D.C. and R.A. Bradley. 1972. Probability models, estimation and classification for multivariate dichotomous populations. Biometrics 23:203-221.

Morgan, R.L. 1969. A class of conjugate prior distributions and optimal allocation. 1969. Unpublished Ph.D. thesis at Univ. of Missouri.

Natanson, I.P. 1965. Constructive function theory Vol II. New York. Ungar. 174 p.

Olkin, I. and H. Rubin. 1962. A characterization of the Wishart distribution. The Annals of Mathematical Statistics: 33. No. 4. 1272-1280.

Press, S.J. 1972. Applied Multivariate Analysis. New York. Holt, Rinehart and Winston. 521 p.

Raiffa, H. and R. Schlaifer. 1961. Applied statistical decision theory. Cambridge, Mass. M.I.T. Press. 356 p.

Rao, C.R. 1965. Linear statistical inference and its applications. New York. Wiley. 522 p.

Sansone, G. 1959. Orthogonal functions. New York. Interscience. 411 p.

Schwartz, Stuart C. 1967. Estimation of probability density by orthogonal series. The Annals of Mathematical Statistics 38:1262-1265.

Smith, K.T. 1971. Smiths primer of modern analysis. New York. Bogden and Quigley. 412 p.

Stein, C., Effron, B. and C. Morris. 1972. Improving the usual estimator of a normal covariance matrix. Stanford University Dept. of Statistics, Technical Report No. 37.

Watson, Geoffrey S. 1969. Density estimation by orthogonal series. The Annals of Mathematical Statistics 40:1496-1498.

Zygmund, A. 1959. Trigonometric series. Vol II. Cambridge Univ. Press. 354 p.

APPENDIX

This appendix contains results scattered throughout various reference works that were found to be useful in writing this thesis.

A.1 Matrix Differentiation

The results given here are formulas for differentiating matrices and functions of matrices with respect to matrices as well as scalars. Most of these results are found in Graybill (1969) as well as various books in econometrics and nonlinear programming.

Definition A.1.1. If f is a function of the real variables $x_1, x_2, \ldots, x_m$ then f is a function of the vector $(x_1, \ldots, x_m)^T$. We define the derivative of f with respect to $x = (x_1, \ldots, x_m)^T$ by the formula

$$\frac{\partial f}{\partial x} := \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_m}\right)^T$$

Definition A.1.2. If f is a function of the mn independent real variables $x_{11}, x_{12}, \ldots, x_{mn}$ then f is a function of the $m \times n$ matrix (x_{ij}) $i = 1, \ldots, m$, $j = 1, \ldots, n$. We define the derivative of f with respect to the matrix $X = (x_{ij})$ to be the $m \times n$ matrix with

$$\left(\frac{\partial f}{\partial X}\right)_{ij} := \frac{\partial f}{\partial x_{ij}} .$$

We should note here that x_{ij} denotes the element in the ith row and jth column of X. If there are any functional relationships between the elements they are disregarded in this definition. For example if X is m x m and symmetric then we consider the matrix Y where the components of Y are functions of the matrix X given by $y_{ij}(x) = y_{ji}(x) = x_{ij}$, $1 \leq i \leq j \leq m$. We now have by the chain rule,

$$\frac{\partial f}{\partial x_{ij}} = \sum_{k,\ell} \frac{\partial f}{\partial y_{k\ell}} \frac{\partial y_{k\ell}}{\partial x_{ij}}$$

$$= \frac{\partial f}{\partial y_{ij}} + \frac{\partial f}{\partial y_{ji}} .$$

Definition A.1.3. If Y is an m x m matrix and y_{ij} is some function of the real variable ξ we define the derivative of Y with respect to ξ to be the m x m matrix with

$$\left(\frac{\partial Y}{\partial \xi}\right)_{ij} := \frac{\partial y_{ij}}{\partial \xi} .$$

For notation, if X is a matrix we shall denote the ij cofactor of X by $|X_{ij}|$ and the matrix of cofactors of X by $X^{\#}$. We now state several results regarding matrix differentiation most of which

can be found in Graybill (1969).

Result A.1.1.

(i) If X and Y are matrices where x_{ij} and y_{ij} are functions of the real variable ξ then

$$\frac{\partial(X+Y)}{\partial\xi} = \frac{\partial X}{\partial\xi} + \frac{\partial Y}{\partial\xi} .$$

(ii) If $X = (x_{ij})$ is an $m \times m$ matrix of independent variables then

$$\frac{\partial|X|}{\partial X} = X^{\#}$$

(iii) If $Y = (y_{k\ell})$ is an $m \times m$ matrix and each element $y_{k\ell}$ is a function of several independent real variables including ξ then

$$\frac{\partial|Y|}{\partial\xi} = \mathrm{tr}[Y^{\#} \frac{\partial Y^{T}}{\partial\xi}] .$$

(iv) If $X = (x_{ij})$ is an $m \times m$ non singular matrix of independent variables then

$$\frac{\partial X^{-1}}{\partial x_{ij}} = -X^{-1}\Delta_{ij}X^{-1}$$

where Δ_{ij} is an $m \times m$ matrix with an entry of one as the ij element and zeroes elsewhere.

We comment here that if X is a symmetric matrix then we may view it as a matrix whose elements are functions of the $\frac{m(m+1)}{2}$ independent variables x_{ij}, $1 \leq i \leq j \leq m$. With this interpretation we state the following results.

Result A.1.2. If X is an $m \times m$ symmetric matrix then for $1 \leq i \leq j \leq m$:

(i) $\frac{\partial X}{\partial x_{ij}} = \Delta^*_{ij}$ where Δ^*_{ij} contains entries of zero as the ij and ji elements and zeroes elsewhere.

(ii) $\frac{\partial |X|}{\partial x_{ij}} = 2|X_{ij}| - \delta_{ij}|X_{ij}|$.

(iii) If in addition X is non singular, $\frac{\partial X^{-1}}{\partial x_{ij}} = -X^{-1}\Delta^*_{ij}X^{-1}$.

We note here that (i) follows directly from Definition A.1.3., (ii) follows direction from A.1.1, (iii) and A.1.2. (i), and (iii) follows from the relationship $X^{-1}X = I$ and the chain rule.

A.2 Characteristic Functions

Definition A.2.1. Let $(W, \mathcal{W}, \lambda)$ be a measure space and let $\underset{\sim}{X}$ be a random vector defined on W into the m-dimensional Hilbert space $(\mathcal{H}, <\cdot,\cdot>)$. We define the characteristic function of $\underset{\sim}{X}$ to be the function $\phi : \mathcal{H} \to \mathbb{R}^1$ defined for each $t \in \mathcal{H}$ by

$$\phi_{\underset{\sim}{X}}(t) = \int e^{i\langle t, x\rangle} dP_X(x)$$

$$= \int e^{i\langle t, X(\omega)\rangle} d\lambda(\omega) .$$

We note here that for random vectors $\underset{\sim}{X} \in \mathbb{R}^m$ we will use

$$\langle t, \underset{\sim}{X}\rangle = \sum_{i=1}^{m} t_i \underset{\sim}{X}_i .$$

We also note that when dealing with random matrices there are two Hilbert spaces that are particularly useful.

Example A.2.1. Let $\mathcal{H}_{mn} = \{A : A \text{ is an } m \times n \text{ real matrix}\}$ $(\mathcal{H}_{mn}, \langle \cdot, \cdot \rangle)$ is a Hilbert space of dimension mn where $\langle A, B\rangle = tr(A^T B)$.

Example A.2.2. Let $\mathcal{H}_{mn} = \{A : \text{ is an } m \times m \text{ real symmetric matrix}\}$ $(\mathcal{H}_{mm}, \langle \cdot, \cdot \rangle$ is a Hilbert space of dimension $\frac{m(m+1)}{2}$ where $\langle A, B\rangle = tr(AB)$.

We comment here that if we have the symmetric matrices V and t where

$$t = \begin{bmatrix} 2t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & 2t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & & \vdots \\ t_{m1} & t_{m2} & \cdots & 2t_{mm} \end{bmatrix}$$

and we use the inner product $\frac{1}{2}\mathrm{tr}(tV)$ in Example A.2.2 then numerically this is equivalent to viewing V and t as vectors V_{ij} and t_{ij}, $1 \leq i \leq j \leq m$ in $\mathbb{R}^{m(m+1)/2}$ and using the standard inner product.

Definition A.2.2. If $\underset{\sim}{X}$ is a random vector with characteristic function $\phi_{\underset{\sim}{X}}(t) = E_{\underset{\sim}{X}}\{e^{i\langle t, \underset{\sim}{X}\rangle}\}$ then a product moment of order k is defined to be $E[\underset{\sim}{X}_1^{J_1}\underset{\sim}{X}_2^{J_2}\ldots\underset{\sim}{X}_m^{J_m}]$ provided the expectation exists and where $\sum_{i=1}^{m} J_i = k$, $\{e_i : i = 1, \ldots, m\}$ is a basis and $\underset{\sim}{X}_i = \langle \underset{\sim}{X}, e_i \rangle$.

The next three lemmas concerning the characteristic functions of random vectors are found in Lukacs (1964).

Lemma A.2.1. Suppose g is a finite real valued, measurable function defined on the range of X. Set $\underset{\sim}{Y} = g(\underset{\sim}{X})$. We then have that

$$\begin{aligned}\varphi_{\underset{\sim}{Y}}(t) &= E_{\underset{\sim}{Y}}\{e^{\dot{\imath}ty}\} \\ &= E_{\underset{\sim}{X}}\{e^{\dot{\imath}tg(x)}\} \\ &= \int_{\mathbb{R}^m} e^{\dot{\imath}tg(x)} dP_X .\end{aligned}$$

Lemma A.2.2. Let g_i, $i = 1, \ldots, s$ be single valued functions and set $\underset{\sim}{Y} = (\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_s)^T$ where $\underset{\sim}{Y}_i = g_i(\underset{\sim}{X})$. The

characteristic function of $\underset{\sim}{Y}$ is then given by

$$\phi_{\underset{\sim}{Y}}(t) = E_{\underset{\sim}{Y}} e^{i t^T Y}$$

$$= \int_{R^m} e^{i t^T g(x)} dP_X .$$

We should note here that if we set $g_i(x) = x_i$ in Lemma A.2.2 we then have the result that the joint characteristic function of any s components of the random vector $\underset{\sim}{X}$ may be found by setting $t_i = 0, \quad i \neq 1, 2, \ldots, s$ in the characteristic function of $\underset{\sim}{X}$.

Lemma A.2.3. Suppose $\underset{\sim}{X}$ is a random vector with each component of $\underset{\sim}{X}$ having finite moments up to order k. It then fillows that all product moments of order k exist. Moreover,

$$\left. \frac{\partial^k \phi_{\underset{\sim}{X}}(t)}{\partial t_1^{J_1} \ldots \partial t_m^{J_m}} \right|_{t=0} = i^m E[\underset{\sim}{X}_1^{J_1} \underset{\sim}{X}_2^{J_2} \ldots \underset{\sim}{X}_m^{J_m}]$$

In the special case where each component of $\underset{\sim}{X}$ has a finite first moment then we define the expectation of $\underset{\sim}{X}$ to be the vector of expectations. That is,

$$E\underset{\sim}{X} = \begin{pmatrix} E\underset{\sim}{X}_1 \\ \vdots \\ E\underset{\sim}{X}_m \end{pmatrix} .$$

Lemma A.2.3 tells us that $E\underset{\sim}{X}_j = \left.\frac{\partial \phi_{\underset{\sim}{X}}(t)}{\partial t_j}\right|_{t=0}$ so that we may write

$$E\underset{\sim}{X} = \frac{1}{i}\left(\left.\frac{\partial \phi_{\underset{\sim}{X}}(t)}{\partial t_i}\right|_{t=0}\right) m \times 1$$

$$= \frac{1}{i}\left(\left.\frac{\partial \varphi_{\underset{\sim}{X}}(t)}{\partial t}\right)\right|_{t=0}$$

We conclude the section on characteristic functions with the following useful lemma.

Lemma A.2.5. If $\underset{\sim}{X}$ is a random $m \times n$ matrix, set $\underset{\sim}{Z} = A\underset{\sim}{X}B + C$ where A, B, C are fixed matrices, then

$$\phi_{\underset{\sim}{Z}}(U) = e^{i\,tr(U^T C)}\phi_{\underset{\sim}{X}}[A^T U B^T]$$

Proof:

$$\begin{aligned}
\phi_{\underset{\sim}{Z}}(U) &= E_{\underset{\sim}{Z}} e^{i\langle U, \underset{\sim}{Z}\rangle} \\
&= E_{\underset{\sim}{X}} e^{i\langle U, A\underset{\sim}{X}B+C\rangle} \\
&= E_{\underset{\sim}{X}} e^{i\,tr[U^T(A\underset{\sim}{X}B+C)]} \\
&= e^{i\,tr\,U^T C} E_{\underset{\sim}{X}} e^{i\langle (BU^T A)^T, \underset{\sim}{X}\rangle} \\
&= e^{i\,tr\,U^T C}\phi_{\underset{\sim}{X}}[(BU^T A)^T] \\
&= e^{i\,tr\,U^T C}\phi_{\underset{\sim}{X}}[A^T U B^T]
\end{aligned}$$

A.3 Remarks Concerning Lemma 4.1.1

We recall the situation in which Lemma 4.1.1 was applied

$$f(u|R) = (2\pi)^{-m/2}|R|^{1/2} e^{-\frac{1}{2}u^T Ru}$$

$$= (2\pi)^{-m/2} e^{\Sigma_i r_{ii} W_{ii}(u) + \Sigma_{i<j} r_{ij} W_{ij}(u) - \psi(R)}$$

where

$$W_{ii}(u) = -\frac{1}{2}u_i^2$$

$$W_{ij}(u) = -u_i u_j, \quad i < j$$

$$\psi(R) = -\frac{1}{2}\log|R|$$

Suppose the prior density is given by

$$f(R|\nu,\tau) \propto |R|^{\nu-m-1/2} e^{-\frac{1}{2}tr(\tau R)}$$

$$= e^{b^*\{\Sigma_i r_{ii} a_{ii}^* + \Sigma_{i<j} r_{ij} a_{ij}^* - \psi(R)\}}$$

where

$$b^* = \nu - m - 1$$

$$a_{ii}^* = -\frac{1}{2}$$

$$a_{ij}^* = 0$$

The posterior distribution of $\underset{\sim}{R}$ is then given by

$$f(R \mid {}^{1}U, \ldots, {}^{n}U, \nu, \tau) \propto e^{b\{\Sigma_i r_{ii} a_{ii} + \Sigma_{i<j} r_{ij} a_{ij} - \psi(R)\}}$$

where

$$b = n + b^*$$

$$a_{ii} = \frac{n\bar{w}_{ii} + b^* a^*_{ii}}{n + b^*}$$

$$a_{ij} = \frac{n\bar{w}_{ij} + b^* a^*_{ij}}{n + b^*}$$

$$\bar{w}_{ij} = \frac{1}{n} \sum_{t=1}^{n} w_{ij}({}^{t}u)$$

We note that the conditional, prior and posterior distributions have a natural parametrization. The condition necessary to write

$$\psi_i(R) = E[W_i(\underset{\sim}{u}) \mid R]$$

is that we must pass a derivative under an integral sign which can be done [Barndorff-Nielsen (1973)] if the natural parameter space is open.

If we let $\mathcal{S}$:= {all symmetric m x m matrices} then $\mathcal{S}$ may be viewed as $\mathbb{R}^{m(m+1)/2}$. We let $\mathcal{V}$:= {all positive definite matrices}. $\mathcal{V}$ is a subset of $\mathcal{S}$ so that we view $\mathcal{V}$ as being a subset of $\mathbb{R}^{m(m+1)/2}$ and we need to show that $\mathcal{V}$ is open in $\mathbb{R}^{m(m+1)/2}$.

Consider the following characterization of positive definiteness for matrices found in Bellman (1960).

Lemma A.3.1. Let A be an $m \times m$ symmetric matrix (a_{ij}). Let $D_k := |(a_{ij})| \quad i,j = 1,\ldots,k$. A is positive definite if and only if $D_k > 0,\ k = 1,\ldots,m$.

We note that $D_k : \mathcal{S} \to R^1$ is a continuous function for $k = 1,\ldots,m$ and that we may characterize $\mathcal{V}$ as that subset of $\mathcal{S}$ for which $D_k > 0,\ k = 1,\ldots,m$. That is,

$$\begin{aligned}\mathcal{V} &= \{R \in \mathcal{S} : D_1(R) > 0, D_2(R) > 0, \ldots, D_m(R) > 0\} \\ &= \{R \in \mathcal{S} : D_1(R) > 0\} \cap \{R \in \mathcal{S} : D_2(R) > 0\} \cap \ldots \\ &\qquad \cap \{R \in \mathcal{S} : D_m(R) > 0\} \\ &= \{D_1^{-1}(0,\infty)\} \cap \{D_2^{-1}(0,\infty)\} \cap \ldots \cap \{D_m^{-1}(0,\infty)\}\end{aligned}$$

Now D_k being continuous for $k = 1,\ldots,m$ implies that each of the sets $\{D_k^{-1}(0,\infty)\}$ is open and thus $\mathcal{V}$ is the finite intersection of open sets and is therefore open.

The regularity condition given by Morgan (1969) sufficient for the validity of the statement $a_{ij} = E[\psi_{ij}(R)]$ that requires comment involves upper and lower limits of integration. The condition arises in the proof of his theorem where the integration by parts formula

$$\int udv = uv - \int vdu$$

is used. Roughly speaking the regularity condition requires that the uv term be zero when evaluated at the upper and lower limits of integration.

To be more precise in our situation, we consider a point R in $\mathcal{V}$ as a vector r in $\mathbb{R}^{m(m+1)/2}$ with components r_{ij}, $1 \le i \le j \le m$. We define the vector $\mathcal{r}_{ij}$ by $\mathcal{r}_{ij} = r$ with the ij component deleted and view $\mathcal{r}_{ij}$ as a vector in $\mathbb{R}^{m(m+1)/2 - 1}$. We define the functions corresponding to upper and lower limits by

$$u_{ij}(s) = \sup\{r_{ij} : R \in \mathcal{V}, \mathcal{r}_{ij} = s\}$$

$$\ell_{ij}(s) = \inf\{r_{ij} : R \in \mathcal{V}, \mathcal{r}_{ij} = s\}$$

We note that the domain of $u_{ij}(\cdot)$ and $\ell_{ij}(\cdot)$ is $\mathbb{R}^{\frac{m(m+1)}{2} - 1}$.

The required regularity condition is:

$$\lim_{r_{ij} \to \ell_{ij}(s)} e^{b^*\{-\frac{1}{2}\Sigma_i r_{ii} + \frac{1}{2}\log|R|\}} = \lim_{r_{ij} \to u_{ij}(s)} e^{b^*\{-\frac{1}{2}\Sigma_i r_{ii} + \frac{1}{2}\log|R|\}}$$

for $1 \le i \le j \le m$.

We now show that this condition is satisfied. Our verification rests on Lemma A.3.1. For notation let $D_{i:j}$ be the determinant of the matrix obtained from R by the deletion of the ith row

and jth column and $D_{ij:j\ell}$ the determinant obtained from R by the deletion of the ith and kth rows and the jth and ℓth columns.

Consider

$$u_{11}(s) = \sup\{r_{11}: R \in \quad , \quad _{11} = s\}$$

Since $R \in \mathcal{V}$, $|R| > 0$. If we expand $|R|$ along the first column of R we have

$$r_{11}D_{1:1} - r_{12}D_{1:2} + r_{13}D_{1:3} - \ldots + (-1)^{m+1}r_{1m}D_{1:m} = |R| > 0$$

We now note that r_{11} is involved in only the first term of the above sum. That is $r_{11}D_{1:1}$ is the only term involving r_{11}. Since $D_{1:1} > 0$, it now follows that with r_{ij} fixed for $(i, j) \neq (1, 1)$, we have that $u_{11}(s) = \infty$. We now note that this same argument is valid for any diagonal element so that

$$u_{ii}(s) = \infty, \quad i = 1, \ldots, m$$

Now consider

$$\ell_{mm}(s) = \inf\{r_{mm}: R \in \mathcal{V}, \mu_{mm} = s\}.$$

We note that the constraints D_k given by Lemma A.3.1 do not involve r_{mm} for $k = 1, \ldots, m-1$. The only constraint on r_{mm} is the constraint $D_m = |R| > 0$. Therefore,

$$\inf\{r_{mm} : R \in \mathcal{V}, \mu_{mm} = s\} = \bar{r}_{mm}$$

where $\bar{r}_{mm}$ is that value of r_{mm} for which $|R| = 0$ and we have

$$\ell_{mm}(s) = \bar{r}_{mm} .$$

By an even number of row and column interchanges any diagonal element can be thought of as the (m, m) element. Therefore,

$$\ell_{ii}(s) = \bar{r}_{ii}, \quad i = 1, \ldots, m.$$

We now consider the "off-diagonal" terms. That is, consider

$$u_{ij}(s) = \sup\{r_{ij} : R \in \mathcal{V}, \mu_{ij} = s\}$$

and

$$\ell_{ij}(s) = \inf\{r_{ij} : R \in \mathcal{V}, \mu_{ij} = s\} .$$

We expand R by column j :

$$\text{(A. 1)} \qquad |R| = (-1)^{1+j} r_{ij} D_{1:j} + (-1)^{2+j} r_{2j} D_{2:j} + \ldots + (-1)^{i+j} r_{ij} D_{i:j}$$

$$+ \ldots + (-1)^{m+j} r_{mj} D_{m:j} > 0 .$$

We note that each matrix corresponding to $D_{t:j}$ for $t \neq i$ contains the component r_{ij} at most once. Therefore, the terms $r_{ij} D_{t:j}$ are linear or constant in r_{ij} for $t \neq i$ and quadratic in r_{ij} for $t = i$.

We now consider the quadratic term $(-1)^{i+j} r_{ij} D_{i:j} \cdot D_{i:j}$ is itself a determinant. The matrix corresponding to this determinant contains the element r_{ji}. We now expand $D_{i:j}$ along the column containing r_{ji} and consider only that term involving r_{ji}. Namely, $(-1)^{i+j-1} r_{ji} D_{ij:ji}$. The quadratic term in (A. 1) now may be written as

$$(-1)^{i+j} r_{ij} (-1)^{i+j-1} r_{ji} D_{ij:ji} = -r_{ij}^2 D_{ij:ij}$$

Now, $D_{ij:ij}$ is the determinant of the matrix R with the i and j rows and columns deleted and is therefore positive.

We now have that the expansion given in (A. 1) is quadratic in r_{ij} with the coefficient of the quadratic term being negative. That is the quadratic is concave. Therefore,

$$\inf\{r_{ij} : R \in \mathcal{V}, \mathcal{R}_{ij} = s\} = \underline{r}_{ij}$$

and

$$\sup\{r_{ij} : R \in \mathcal{V}, \mathcal{R}_{ij} = s\} = \overline{r}_{ij}$$

where $\underline{r}_{ij}$ and $\overline{r}_{ij}$ are respectively the smallest and largest values of r_{ij} for which $|R| = 0$ when the remaining elements in R are fixed.

From the relationship,

$$e^{b^*\{-\frac{1}{2}\Sigma_i r_{ii}+\frac{1}{2}\log|R|\}} = |R|^{b^*/2} e^{b^*\{-\frac{1}{2}\Sigma_i r_{ii}\}}$$

and the values of $u_{ij}(s)$ and $\ell_{ij}(s)$ as defined above, the regularity condition follows. ∎