FINITE DEFINETTI THEOREMS IN
LINEAR MODELS AND MULTIVARIATE ANALYSIS

BY

PERSI DIACONIS, MORRIS EATON AND STEFFEN LAURITZEN

TECHNICAL REPORT NO. 343
May 1990

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS89–05874

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Finite deFinetti Theorems in
Linear Models and Multivariate Analysis

By

Persi Diaconis, Morris Eaton† and Steffen Lauritzen

Technical Report No. 343
May 1990

Prepared Under the Auspices
Of
National Science Foundation Grant DMS89–05874

† Also supported by National Science Foundation Grants DMS83-19924 and DMS88-00284.

Department of Statistics
Stanford University
Stanford, California
Finite deFinetti Theorems in
Linear Models and Multivariate Analysis

by

Persi Diaconis\(^1\), Morris Eaton\(^2\) and Steffen Lauritzen
Harvard University, University of Minnesota and Aalborg University

\(^1\)Research partially supported by National Science Foundation Grant DMS 86-00235

\(^2\)Research partially supported by National Science Foundation Grants
DMS 83-19924 and DMS 88-00284.

The authors also acknowledge the generous support of the Danish Research Council for portions of this research work.
Abstract

Let $X_1, \ldots, X_k$ be a sequence of random vectors. We give symmetry conditions on the joint distribution which imply that it is well approximated by a mixture of normal distributions. Examples include linear regression models, fixed and random effects analysis of variance models, and models with patterned covariance matrices. The main technical tool shows that a uniformly distributed $n \times n$ orthogonal matrix has any $n^{1/3} \times n^{1/3}$ block well approximated (in total variation) by independent normal random variables.
1. Introduction

definetti style theorems characterize models in terms of invariance. The idea is to begin with observables, postulate symmetries or summary statistics, and then find a simple description of all the models with the given symmetries.

This paper develops this program for the usual general linear model, analysis of variance and some standard covariance models in multivariate analysis. Postulated symmetries are expressed in terms of distributional invariance under the action of various groups of orthogonal transformations.

Our approach is to prove finite versions of the theorems. That is, if $X_1, \ldots, X_k$ can be extended to $X_1, \ldots, X_n$ which has the same type of symmetries, then the law of $X_1, \ldots, X_k$ is "almost" a mixture of a standard parametric family. "Almost" is expressed by an explicit error bound on the total variation distance between the law of $X_1, \ldots, X_k$ and an approximating mixture of normals.

A review of some previous work is given at the end of this section. The following example illustrates our results.

Example 1.1. Simple linear regression can be formulated as the model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, k$$

where $x_1, \ldots, x_k$ are known constants, $\alpha$ and $\beta$ are unknown parameters, and $\epsilon_1, \ldots, \epsilon_k$ are iid $N(0, \sigma^2)$ random variables. If $\alpha, \beta, \sigma$ are given the prior distribution $\mu(d\alpha, d\beta, d\sigma)$, the joint distribution of $Y_1, \ldots, Y_k$ is

$$(1.1) \quad P(Y_1 \leq y_1, \ldots, Y_k \leq y_k) = \int \prod_{i=1}^k \Phi\left( \frac{y_i - \alpha - \beta x_i}{\sigma} \right) \mu(d\alpha, d\beta, d\sigma).$$

Here $\Phi$ is the standard normal distribution.

Our object is to characterize such a mixture by symmetry conditions on the $Y_i$'s. Let $M_k$ be the two dimensional subspace of $R^k$ which is spanned by
\[ e = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad x^{(k)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}. \]

Let \( O_k(M_k) \) be all the \( k \times k \) orthogonal transformations of \( \mathbb{R}^k \) which fix each element of \( M_k \). The law of \( Y_1, \ldots, Y_k \) is invariant under each element of \( O_k(M_k) \), but this is not enough to characterize the normal mixture model. It is necessary to suppose that the experiment can be continued in the following sense.

For \( n > k \), let \( Q \) be a probability measure on \( \mathbb{R}^n \) and write \( Q_k \) for the law of the first \( k \)-coordinates induced by \( Q \). Let

\[ x^{(n)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \]

be an extension of \( x^{(k)} \). Suppose that \( Q \) is invariant under \( O_n(M_n) \). We will say that \( Q \) extends \( Q_k \). Our theorems show that \( Q_k \) is almost a mixture of normal models.

**Theorem 1.1:** Fix \( x^{(n)} \in \mathbb{R}^n \). Suppose \( O \) is a probability on \( \mathbb{R}^n \) which is \( O_n(M_n) \) invariant. Let \( Q_k \) be the marginal law of the first \( k \) coordinates. Then there exists a probability measure \( \mu(d\alpha, d\beta, d\sigma) \) such that

\[ ||Q_k - P_{\mu,k}|| \leq B(k,n,x^{(n)}). \]

Here, ||·|| denotes variation distance and \( P_{\mu,k} \) is the \( \mu \) mixture of normal distributions defined at (1.1). The bound \( B(k,n,x^{(n)}) \) has the form

\[ B(k,n,x^{(n)}) = c \frac{k}{n} + d(x^{(n)}) \]
where $c$ is a fixed constant and

$$d(x^{(n)}) = 2 \left[ \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-k) \sum_{k+1}^{n} (x_i - \bar{x})^2} \right]^{1/2} - 1.$$

As usual,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and

$$\bar{x} = \frac{1}{n-k} \sum_{k+1}^{n} x_i.$$

The interpretation of this result is that if $k/n$ and $d(x^{(n)})$ are both small, then $Q_k$ is close to some model of the form (1.1). It is argued later in this paper that $d(x^{(n)}) \to 0$ as $n \to \infty$ iff the parameter vector $\theta$ has a consistent estimator. This connects the Bayesian interpretation to frequentist notions.

In a Bayesian setting, one usually begins with $Q_k$. If $Q_k$ can be extended to an invariant $Q$ on $\mathbb{R}^n$, then $Q_k$ is almost a mixture of normal models.

Here is an outline of the paper. An abstract argument which pervades the examples is given in Section 2. A finite version of a result due to Dawid (1977) illustrates the argument.

Univariate linear models are treated in Section 3. Results for the linear regression model as well as results for both fixed and random effects two way layouts illustrate the general methods for linear models.

A special multivariate linear model is discussed in Section 4. The main ideas here extend the univariate results of Smith (1981) in two directions - our model is multivariate and ours is a finite (as opposed to infinite) form of a deFinetti theorem.

Multivariate models with structured covariance matrices are discussed in
Section 5. In particular models with diagonal covariances are treated. These examples lead to an interesting conjecture.

The main technical tool in this paper concerns random orthogonal matrices. In Appendix A, it is shown that a $k \times k$ block of a uniformly distributed $n \times n$ orthogonal matrix, has approximately iid normal entries. Deviations are measured in terms of variation distance which is shown to be small provided $k \ll n^{1/3}$. Appendix B contains some technical results regarding maximal invariants which are used in Section 5.

1.2 Historical Notes

deFinetti (1931) began the subject by characterizing exchangeable probabilities on infinite binary sequences as mixtures of coin tossing. Diaconis and Freedman (1980) give a finite version of the theorem: an exchangeable probability on binary sequences of length $n$ has its $k$ dimensional marginals within $2k/n$ of a mixture of coin tossing in total variation distance.

deFinetti (1938) found the appropriate theorem for exchangeable sequences with values in Euclidean spaces. Hewitt and Savage (1955) extended the result to sequences with values in general spaces. They give a thorough review of the early history.

deFinetti (1938, 1959) and Freedman (1962, 1963) began work on more general types of symmetry. They treated multivariate discrete data, Markov chains, and mixtures of standard univariate exponential families. A typical result from Freedman (1963) is: A real valued orthogonally invariant process is a scale mixture of independent mean 0 normal variables.

The work up to this time was carried out in a Bayesian context. Per Martin-Lof (1970, 1974) began work with the following motivation: suppose you are presented a data
set and only certain summary statistics are deemed relevant. What is the set of all probability models such that the given summaries are sufficient statistics? This work has been extended and summarized in Lauritzen (1982, 1988) where many further models are treated.

Following this early work, a host of individual models were analyzed. Dawid (1978) treated the mean 0 p-dimensional covariance mixture of normals and Smith (1981) characterized location-scale mixtures of univariate normals. A survey of work on special cases is given in Diaconis (1988). As far as we know, the multidimensional extension of Smith's theorem and the finite version of Dawid's theorem appear here for the first time.

Work on finite versions appears in Diaconis and Freedman (1981, 1987) and in Zaman (1986). It is fair to say that previous arguments depended heavily on the family of measures under study. We hope that the present unified treatment of many finite theorems has some appeal. It is often an easy matter to use the finite theorems with a passage to the limit to obtain the corresponding infinite theorems.

An abstract formalism for extreme point models was pioneered by Lauritzen (1982, 1988) in the language of projective systems. Diaconis and Freedman (1984) develop similar results in Bayesian language. Dynkin (1978) and Dobrusyn-Lanford-Ruelle (Ruelle (1978) is a convenient reference) developed essentially the same theory for statistical mechanics. The theory yields integral representations indexed by points in a "tail set". These are rather indirect descriptions and further work is required to put them into digestible form.

This brief review has not mentioned a variety of applications of deFinetti's theorem in probability. These are splendidly treated in Aldous (1985). Further
significant work appears in Ressel's (1985) analytic treatment and Kallenberg's extended work on exchangeability for general processes. Kallenberg (1976, 1981) are good references leading to others.
2. A General Argument

Here we describe a general argument which underlies all of the examples in this paper. Before giving the argument, it is useful to look at an important example which yields the finite version of Dawid's theorem alluded to earlier.

**Example 2.1**: Let $X_1, X_2, \ldots$ be a sequence of $p$-dimensional column random vectors. For each $n$, let

$$X^{(n)} = \begin{bmatrix}
X_1' \\
X_2' \\
\vdots \\
X_n'
\end{bmatrix}$$

be the $n \times p$ matrix whose rows are $X_1', X_2', \ldots, X_n'$. The group of $n \times n$ orthogonal matrices, $O_n$, acts on $X^{(n)}$ by left multiplication. The sequence $X_1, X_2, \ldots$ is **orthogonally invariant** (that is, the probability law of the sequence is orthogonally invariant) if for each $n$, the law of $X^{(n)}$, say $P_n'$, is $O_n$ invariant.

As an example, if the $X_i$'s are iid multivariate normal with mean $0$ and covariance matrix $\Sigma$, then the distribution of $X^{(n)}$, denoted by $N(0, I_n \otimes \Sigma)$, is $O_n$ invariant. Thus for any probability measure $\mu$ defined on the set of $p \times p$ non-negative definite covariance matrices, the mixture

$$Q_{n\mu} = \int N(0, I_n \otimes \Sigma) \, \mu(d\Sigma) \quad (2.1)$$

is also $O_n$ invariant.

Now, let $P$ be the law of the sequence $X_1, X_2 \ldots$ and let $Q_{\Sigma}$ denote the law of $X_1, X_2, \ldots$ when the $X_i$'s are iid $N(0, \Sigma)$. Thus the law of $X^{(n)}$, $P_n'$, is obtained from $P$ by projection, and $Q_{n\Sigma}$, the projection of $Q_{\Sigma}$, is the $N(0, I_n \otimes \Sigma)$ distribution.
Theorem 2.1 (David (1977)): If $X_1, X_2, \ldots$ is orthogonally invariant, then there exists a probability measure $\mu$ defined on the Borel sets of all $p \times p$ non-negative definite matrices such that

\begin{equation}
\text{P} = \int \text{Q}_\Sigma \mu(d\Sigma) = \text{Q}_{\mu}
\end{equation}

In particular, for each $k$,

\begin{equation}
P_k = \int \text{Q}_{k\Sigma} \mu(d\Sigma) = \text{Q}_{k\mu}
\end{equation}

The converse also holds.

Theorem 2.1 is false if orthogonal invariance is just assumed for a fixed $n$. That is, if $X^{(n)}$ is orthogonally invariant, it is not necessarily true that any $P_k$ has the representation (2.3). However, it is true that if $k$ is much smaller than $n$, then (2.3) is approximately true. This is the issue with which our paper deals.

Fix $n$ and assume $X^{(n)}$ is $Q_{\mu}$ invariant. Let $P_n$ be the law of $X^{(n)}$ so the projections of $P_n$, say $P_k$, are the laws of $X^{(k)}$, $k=1,2,\ldots,n$. For a fixed $k<n$, how close is $P_k$ to some covariance mixture of normals? That is, how small is

\begin{equation}
d_{kn} = \inf_{\mu} \left\| P_k - Q_{k\mu} \right\|
\end{equation}

Here, $\|\cdot\|$ is variation distance. An upper bound on $d_{kn}$ is obtained in two steps, the first of which is rather general. At this point we leave Example 2.1 to present this general argument.

Consider a Polish space $X_2$ which is acted on continuously by a compact group $G$. Let $\mathcal{C}$ be all the $G$ invariant probability measures defined on the Borel sets of $X_2$. The unique invariant probability measure on $G$ is $\xi$. Given $x \in X_2$, define the probability $\nu_x$ by

$$
\nu_x(B) = \xi(g | gx \in B).
$$

It is clear that $\nu_x$ is the unique $G$ invariant probability defined on the orbit
of \( x \). [The orbit of \( x \) is compact and \( G \) acts transitively on the orbit. The uniqueness of \( \nu_x \) is a standard result - see Nachbin (1965).] Further, every \( P \in \mathcal{C} \) is a mixture of the \( \nu_x \)'s - that is, each \( P \) can be written

\[
P = \int \nu_x H(dx)
\]

where \( H \) is a probability on the Borel sets of \( X_2 \). A proof of this well known result can be found in Eaton (1989).

Now, let \( X_1 \) be another Polish space and suppose \( \pi : X_2 \to X_1 \) is a measurable mapping. Let \( \mathcal{G}_{12} = \{ \pi P \mid P \in \mathcal{G} \} \) be all the "projected" invariant measures. Suppose that for each \( x \in X_2 \), \( Q_x \) is a probability measure which is an "approximation" to \( \nu_x \). The following result provides an upper bound on the variation distance between

(i) \( \pi P \in \mathcal{G}_{12} \)

(ii) the closest approximation to \( \pi P \) which is a projection of a mixture of the \( Q_x \)'s.

**Proposition 2.1** : Given \( \pi P \in \mathcal{G}_{12} \),

\[
\Delta = \inf_{\mu} \| \pi P - \pi(\int \nu_x \mu(dx)) \| \leq d
\]

where

\[
d = \sup_x \| \nu_x - Q_x \|.
\]

**Proof** : From (2.5),

\[
\pi P = \int \nu_x H(dx)
\]

for some \( H \). Choosing \( \mu = H \) in (2.6) and using the convexity of variation distance yields

\[
\Delta \leq \| \int \nu_x H(dx) - \int Q_x H(dx) \| = \\
\| \int (\nu_x - Q_x) H(dx) \| \leq \\
\int \| \nu_x - Q_x \| H(dx) \leq d.
\]
This completes the proof.

The proof of Proposition 2.1 shows that, with $\mu = H$ in (2.6), we have the inequality

$$\Delta \leq \| \pi P - \pi(\int Q_x H(dx)) \| \leq d$$

where $d$ depends only on $\pi$ and $X_1$. Observe that the same $H$ works for all choices of $\pi$ and $X_1$. This fact is used in many of our theorems below - for example, see Theorem 2.2.

**Example 2.1 continued**: To apply Proposition 2.2, take $X_2$ to be the set of all $n \times p$ real matrices and let $G = \mathbb{O}_n$. Also take $X_1$ to be all $k \times p$ real matrices and define $\pi$ by the matrix

$$\pi = (I_k, 0) : n \times p$$

which maps $X_2$ onto $X_1$ linearly. Given an orthogonally invariant $P_n$ on $X_2$, obviously $P_k = \pi^* P_n \pi$.

For $x \in X_2$, the probability $\nu_x$ can be described as follows. Let $U : n \times n$ be a random orthogonal matrix ($U$ has invariant measure on $O_n$ as its law). Thus $\nu_x$ is just the law of the random matrix $UX$ so that $\pi \nu_x$ is the law of $\pi UX$. It is not hard to show (see Appendix A) that $UX$ has mean 0 and covariance given by

$$\text{Cov}(UX) = I_n \otimes (n^{-1} x'x)$$

where $\otimes$ denotes Kronecker product (the notation here is consistent with Eaton (1983)). As an "approximation" to $\nu_x$, take $Q_x$ to be $N(0, I_n \otimes (n^{-1} x'x))$ so the mean and covariance of $\nu_x$ and $Q_x$ match. Then

$$\pi Q_x = N(0, I_k \otimes (n^{-1} x'x))$$

Now, Proposition 2.1 shows that $d_{kn}$ in (2.4) is bounded above by

$$d = \sup_{x} || \mathcal{L}(\pi UX) - \pi Q_x ||$$

$$= \sup_{x} || \mathcal{L}(\pi UX) - N(0, I_k \otimes (n^{-1} x'x)) ||$$

where $\mathcal{L}(\cdot)$ denotes "the law of \cdot". To bound $d$, write the $n \times p$ matrix $x$ as

$$x = \Gamma \left( \begin{array}{c} I_p \\ 0 \end{array} \right) s$$

where $\Gamma$ is in $O_n$ and $s : p \times p$ is non-negative definite with $s^2 = x'x$ (see Eaton (1983, Chapter 5) for this well known decomposition). Also let $Z : k \times p$ be $N(0, I_k \otimes I_p)$ so

$$\mathcal{L}(n^{-1/2} Z s) = N(0, I_k \otimes (n^{-1} x'x)).$$

12
Because \( U \) is uniform on \( O_n \), we have,

\[
d = \sup_s \left| \mathcal{L}(\pi U \left[ \begin{array}{c} I_p \\ 0 \end{array} \right] s) - \mathcal{L}(n^{-1/2}Zs) \right|
\]

\[
= \sup_s \left| \mathcal{L}(\pi U \left[ \begin{array}{c} I_p \\ 0 \end{array} \right] s) - \mathcal{L}(n^{-1/2}Zs) \right|
\]

\[
= \sup_s \left| \mathcal{L}(n^{1/2}U_{11}s) - \mathcal{L}(Zs) \right|
\]

\[
= \mathcal{L}(n^{1/2}U_{11}) - \mathcal{L}(Z)
\]

where

\[
(2.8) \quad U_{11} = \pi U \left[ \begin{array}{c} I_p \\ 0 \end{array} \right]
\]

is the upper left hand \( k \times p \) block of \( U \). Thus our problem has been reduced to bounding the variation distance between \( n^{1/2}U_{11} \) and a matrix \( Z \) of iid \( N(0,1) \)'s.

This calculation is carried out in Appendix A to show that \( d_{kn} \) in (2.4) is bounded above by \( B(k,p;n) \) given in (A.32) in Appendix A.

Our conclusions are summarized in the following theorem.

**Theorem 2.2:** Let \( X_1, \ldots, X_n \) be \( p \)-dimensional random vectors, and let \( X^{(n)} \) be the \( n \times p \) matrix with rows \( X_1', \ldots, X_n' \). Suppose \( P_n \), the law of \( X^{(n)} \), is \( O_n \)-invariant. Then there exists a measure \( \mu \) on the set of \( p \times p \) non-negative definite covariance matrices such that for all \( k \) satisfying \( p + k + 2 < n \), the \( \mu \)-mixture of normals, \( Q_{k\mu} \), defined at (2.1) satisfies

\[
\|L(X^{(k)}) - Q_{k\mu}\| \leq 2 \left\{ \left[ 1 - \frac{p + k + 2}{n} \right]^{1/2} - 1 \right\}
\]

where \( c = \frac{1}{2} \min(k,p) \).

Here are a few remarks concerning Proposition 2.1 and its application.

**Remark 2.1:** It is clear that the measure \( \nu_x \) in Proposition 2.1 is a \( G \)-invariant function of \( x \). That is

\[
\nu_x = \nu_{gx} \quad \text{for} \quad x \in X_2, \quad g \in G.
\]

In all the applications that we know, \( Q_x \) is also a \( G \)-invariant function of \( x \).

Thus the sup in (2.7) is actually a sup over the orbits in \( X_2 \) rather than a sup over \( X_2 \). In Example 2.1, the orbits are indexed by \( s = (x'x)^{1/2} \) which explains the appearance of \( s \) in this example.

**Remark 2.2:** It is not uncommon that the sup in (2.7) is actually achieved at a point \( x_0 \in X_2 \) which one can identify. This happens in most of the examples in later sections.
Remark 2.3: In most examples, $Q_x$ is a very bad "approximation" to $\nu_x$, but $\pi Q_x$ is a very good approximation to $\nu_x$. In Example 2.1, $Q_x$ and $\nu_x$ are mutually singular, but $\pi Q_x$ and $\nu_x$ are very close when $p$ and $k$ are small compared to $n$.

Proposition 2.1 can be extended to cases where it is more convenient to approximate $\nu_x$ directly rather than first guess at $Q_x$ and then project. This method, described formally in the following proposition is used in Example 3.3. Let the probability measure $R_x$ be an "approximation" to $\nu_x$ for each $x \in X_2$.

**Proposition 2.2:** Given $\pi p \in \mathcal{P}_{12}$, let

$$\Delta = \inf_{\mu} \|\pi p - \int R x \mu(dx)\|$$

where the inf ranges over all probability measures on $X_2$. Then $\Delta \leq d$ where

$$d = \sup_x \|\nu_x - R_x\|$$

**Proof:** The proof is essentially the same as the proof of Proposition 2.2.

**Remark 2.4:** Because $x \to \nu_x$ is an invariant function of $x$, one can always take the approximators $R_x$ to be invariant functions of $x$ without increasing $d$ in (2.10). In all the examples that we know, the natural choice for $R_x$ is invariant.
3. The General Univariate Linear Model

In this section, we investigate the implications of symmetry assumptions on the probability models typically associated with univariate linear models. The regression model introduced in Section 1 serves as motivation for the more general treatment here. The main result of this section shows that certain extendability and invariance assumptions imply that the probability structure associated with a univariate linear model is approximately that obtained from a mean vector-scale mixture of normals. The examples here include linear regression, the two way fixed effect model and the two-way layout random effects model.

In a finite dimensional inner product space \((V,(.,.))\), a univariate normal linear model is often specified by

\[
Y = \mu + \epsilon
\]

(3.1)

where \(\mu\), the mean vector of \(Y\), is unknown but is assumed to be in a known subspace \(M \subseteq V\), and the error vector \(\epsilon\) is assumed to be \(N(0,\sigma^2I)\) on \(V\), with \(\sigma^2 > 0\) but unknown. Let \(O_V\) be the group of orthogonal transformations of \((V,(.,.))\) and let \(O(M)\) be the subgroup of \(O_V\) given by

\[
O(M) = \{ g \mid g \in O_V, gx = x \text{ for all } x \in M \}.
\]

Because \(L(Y) = N(\mu, \sigma^2I)\), it is obvious that

\[
L(gY) = L(Y) \text{ for } g \in O(M).
\]

(3.2)

There are certainly non-normal distributions for \(Y\) for which (3.2) holds - in particular, any \(O_V\) invariant distribution for \(\epsilon\) yields (3.2). The purpose here is to study the implications of the invariance in (3.2), not for the distribution of \(Y\) but for the distribution of a "projection" of \(Y\). The regression example of Section 1 is an example where the "extended" experiment corresponds to the model (3.) and the actual experiment corresponds to the
projection of Y.

To introduce the projection formally, let \((V_1, (\cdot, \cdot)_1)\) be a second inner product space of dimension less than \(V\). Consider a linear map \(\pi\) from \(V\) to \(V_1\) which satisfies

\[
\pi \pi' = I_1
\]

where \(I_1\) is the identity transformation on \(V_1\) and \(\pi'\) is the transpose of the "projection" \(\pi\). Then the projected linear model is

\[
Y_1 = \pi(Y) = \pi(\mu) + \pi(\epsilon) = \mu_1 + \epsilon_1.
\]

The subspace

\[
M_1 = \pi(M) \subseteq V_1
\]

is the mean space for \(Y_1\).

To state our main result, consider probability measures \(S_\zeta\) of the form

\[
S_\zeta = \int_M \int_0^\infty N(\mu, \sigma^2 I) \zeta(d\mu, d\sigma)
\]

where \(\zeta\) is any probability on \(M \times [0, \infty)\). Thus, \(S_\zeta\) is a mean-variance mixture of a normal model on \(V\) (with mean in \(M\) and covariance \(\sigma^2 I\)).

Note that

\[
\pi S_\zeta = \int_M \int_0^\infty N(\pi\mu, \sigma^2 I_1) \zeta(d\mu, d\sigma)
\]

is a mixture of normal models (with mean \(\pi\mu \in M\), and covariance \(\sigma^2 I_1\)) on \(V_1\). Let \(n = \dim V, m = \dim M, k = \dim V_1\), and let \(Q\) be the orthogonal projection onto the orthogonal complement of \(M\) in \(V\).

Theorem 3.1: Assume that \(P = L(Y)\) satisfies the invariance condition (3.2), and assume \(k + 3 \leq n - m\). There exists a probability \(\zeta_0\) on \(M \times [0, \infty)\) such that for all \((V_1, (\cdot, \cdot)_1)\) and all projections \(\pi\) with \(\pi'Q\pi\) nonsingular,

\[
\|\pi P - \pi S_{\zeta_0}\| \leq B(k, 1; n - m) + 2 \left(\det \pi'Q\pi\right)^{-1/2} - 1
\]

where \(B(k, 1; n - m)\) is given in (A.32).

Proof: We apply the general method described in Section 2 and obtain an upper bound on \(d\) in (2.7). Let \(U\) be uniform on \(Q(M)\). For \(y \in V\), write \(y = y_1 + y_2\)
where \( y_1 \in M \) and \( y_2 \in M \). Since \( U \in O(M) \),
\[
Uy = y_1 + Uy_2,
\]
so
\[
\pi Uy = \pi y_1 + \pi Uy_2.
\]
Fix a point \( y_0 \in M \) with \( |y_0| = (y_0, y_0)^{1/2} - 1 \). The transitivity of \( O(M) \) on
\[
(y \mid (y, y) = 1, y \in M)
\]
is implied that
\[
\mathcal{I}(\pi Uy_2) = \mathcal{I}(|y_2| \pi Uy_0).
\]
Since \( \nu_y = \mathcal{I}(Uy) \) (see (2.5)),
\[
(3.5) \quad \pi \nu_y = \mathcal{I}(\pi Uy) = \mathcal{I}(\pi y_1 + |y_2| \pi Uy_0).
\]
For our approximator, we choose
\[
(3.6) \quad Q_y = N(y_1, (n-m)^{-1} |y_2|^2 I_1)
\]
so that
\[
(3.7) \quad \pi Q_y = N(\pi y_1, (n-m)^{-1} |y_2|^2 I_1).
\]
To bound
\[
d = \sup_y || \pi \nu_y - \pi Q_y ||,
\]
use the translation and scale invariance of variation distance plus the triangle inequality to obtain
\[
d = \sup_y || \mathcal{I}(\pi y_1 + |y_2| \pi Uy_0) - N(\pi y_1, (n-m)^{-1} |y_2|^2 I_1) ||
\]
\[
= \left| \left| \mathcal{I}(\pi Uy_0) - N(0, (n-m)^{-1} I_1) \right| \right|
\]
\[
\leq \left| \left| \mathcal{I}(\pi Uy_0) - N(0, (n-m)^{-1} \pi Q_{\pi'}) \right| \right|
\]
\[
+ \left| \left| N(0, (n-m)^{-1} I_1) - N(0, (n-m)^{-1} \pi Q_{\pi'}) \right| \right|.
\]
An application of Proposition A.2 (with \( \alpha = \pi \) and \( \beta = y_0 \)) together with the non-singularity of \( \pi Q_{\pi'} \) shows that
\[
\left| \left| \mathcal{I}(\pi Uy_0) - N(0, (n-m)^{-1} \pi Q_{\pi'}) \right| \right|
\]
\[ ||\mathcal{L}(\Delta) - N(0,(n-m)^{-1} I)|| \]

where \( \Delta: k \times 1 \) is the first \( k \)-coordinates of a random vector which is uniform on \( \{x \mid x \in \mathbb{R}^{n-m}, ||x|| = 1\} \). Proposition A.4 yields \( B(k,1;n-m) \) as a bound.

Proposition A.5 yields the bound

\[ 2[(\det \pi Q \pi')^{-1/2} - 1] \]

for the second term. This completes the proof. \( \square \)

**Example 3.1**: The classical regression model on \( \mathbb{R}^k \) is typically written

\[ Y_1 = X_1 \beta + e_1 \]

where \( X_1 \) is a \( k \times p \) known matrix of rank \( p \) and \( \beta \in \mathbb{R}^p \) is a vector of unknown regression parameters. Assume this model is invariantly extendable to

\[ Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = X \beta + \epsilon \]

on \( \mathbb{R}^n \) where \( M \) is column space of \( X \) and the distribution of \( Y \) is \( \mathcal{Q}(M) \) invariant. With

\[ \pi = (I_k, 0) : k \times n \]

and \( p = \text{dim } M = \text{rank } X = \text{rank } X_1 \), Theorem 3.1 shows that the distribution of \( Y_1 \) is close to a mixture of normals with the variation distance being bounded by (3.4). For this example,

\[ Q = I - X(X'X)^{-1}X' \]

\[ \pi Q \pi' = I_k - X_1 (X'X)^{-1}X'_1 \]

and

\[ \det \pi Q \pi' = \frac{\det X_2 X_2'}{\det(X_1X_1 + X_2X_2')} \]

With \( k \) and \( p \) fixed and \( n \to \infty \), it is natural to ask for conditions under which (3.8)

\[ \lim_{n \to \infty} [(\det \pi Q \pi')^{-1/2} - 1] = 0 \]
when $X_1$ (our original design matrix) is fixed. Since the eigenvalues of $\pi'Q\pi$
are less than or equal to one, (3.8) holds iff
\[ (3.9) \quad \lim_{\mathcal{N} \to \infty} (X'X)^{-1} = 0. \]
However, in the regression problem with $n$ observations, the least squares
estimator of $\beta$ is $\hat{\beta}_n$ whose covariance matrix is
\[ \text{Cov}(\hat{\beta}_n) = \sigma^2 (X'X)^{-1} \]
where $\sigma^2$ is the error variance. Thus (3.9) holds iff $\hat{\beta}_n$ is a consistent
estimator of $\beta$.

The above discussion includes the simple linear regression example given in
Section 2. In particular, it is known (Lauritzen 1989, p.246-7) that (3.9)
holds iff
\[ (3.10) \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 \to +\infty \]
in the notation of Example 1.1. \hfill \Box

**Example 3.2: Two way layout**

Theorem 3.1 is applied to the standard two way layout (without interactions) in
this example. Take $V$ to be the linear space of all $K \times L$ real matrices so
dim $V = KL$. The regression subspace for this example is
\[ M = \left\{ x \in V \left| \begin{array}{c}
   x_{ij} = \alpha_i + \beta_j, \quad \alpha_i \text{ and } \beta_j \text{ in } \mathbb{R}, \\
   i = 1, \ldots, K, \ j = 1, \ldots, L
\end{array} \right. \right\}. \]
The subspace $M$ is the set of vectors that are additively composed of row effects
plus column effects. The observation vector $Y \in V$ is assumed to have a
distribution, say $P$, which is $\omega(M)$ invariant.

Let $V_1$ be the vector space of $k \times l$ real matrices with $k < K$ and $l < L$, and
let $\pi$ be the linear map from $V$ to $V_1$ which picks out the upper left $k \times l$ block
of an element of $V$. It is easy to check that (3.3) holds. Clearly elements of $M_1 - \pi(M)$ consist of $k \times \ell$ matrices which have an additive representation. Now, Theorem 3.1 applies directly to this example to give a bound on the variation distance between $\mathcal{I}(\pi Y)$ and the closest translation-scale mixture of normals with mean vectors in $M_1$. This bound is

\begin{equation}
B(k \ell, 1; (K-1)(L-1)) + 2 \left[ (\text{det } \pi Q \pi')^{-1/2} - 1 \right],
\end{equation}

where $Q$ is the orthogonal projection onto $M_1$ in $V$. Thus, the calculation of

\begin{equation}
b = \text{det } \pi Q \pi'
\end{equation}

remains. Standard results in the analysis of variance show that

\[ \pi' Q \pi = \left( I_k - K^{-1} E_k \right) \otimes \left( I_\ell - L^{-1} E_\ell \right) \]

where $I_m$ is an $m \times m$ identity matrix and $E_m$ is an $m \times m$ matrix of all ones.

Hence

\[ b = \left[ \text{det} (I_k - K^{-1} E_k) \right]^\ell \left[ \text{det} (I_\ell - L^{-1} E_\ell) \right]^k = (1 - k/K)^\ell (1 - \ell/L)^k \]

For $k$ and $\ell$ fixed, with $K, L \to \infty$, (3.11) converges to zero. This can be used to establish an infinite version of this example - a result established by other means in Lauritzen (1982, 1989).

The following example, a two way layout with one effect random, cannot be treated directly by Theorem 3.1. However, the general method described in Section 2 is applicable. This example provides an interesting comparative contrast to Example 3.2 and illustrates the applicability of the general argument.

**Example 3.3 : Two way layout - one effect random**

The symbols $K, L, V, k, \ell, V_1$ and $\pi$ denote what they did in Example 3.2. The two subspaces of $V$

\[ M_R = \{ x \mid x_{ij} = \alpha_i \text{ for } \alpha_i \in \mathbb{R}, i = 1, \ldots, K, j = 1, \ldots, L \} \]
and

\[ M_C = \{ x | x_{ij} - \beta_j \text{ for } \beta_j \in \mathbb{R}, i = 1, \ldots, K, j = 1, \ldots, L \} \]

are the row and column subspaces. The subspace \( M \) of Example 3.2 is just \( M_R + M_C \) but \( M_R \cap M_C = M_0 \) is the one dimensional subspace of the constant matrices.

The normal model for this situation for \( Y \in V \), is

\begin{equation}
Y_{ij} = \alpha_i + \beta_j + \epsilon_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, L
\end{equation}

(3.13)

where the \( \alpha_i \) are real numbers, the \( \beta_j \) are iid \( N(0, \omega^2) \), and \( \epsilon_{ij} \) are iid \( N(0, \sigma^2) \) with the \( \beta \)'s and \( \epsilon \)'s independent. Thus the normal model for the projected \( Y \), say \( Y_1 = \pi Y \), is just (3.13) with \( i = 1, \ldots, k \) and \( j = 1, \ldots, l \). The model (3.13) is invariant under the group \( G \) defined by

\[
G = \left\{ g \in O_V \bigg| gx = x \text{ for } x \in M_R, \quad gx \in M_C \text{ for } x \in M_C \right\}
\]

Thus \( G \) is the identity on the row space \( M_R \) and leaves the column space \( M_C \) invariant. Hence \( M \) also remains invariant under \( G \).

Now, let \( P \) be a distribution on \( V \) which is \( G \)-invariant and let \( P_1 = \pi P \) be the projection of \( P \) down to \( V_1 \). We want to show that \( P_1 \) is close to an average (over \( \alpha_1, \ldots, \alpha_k, \omega^2, \sigma^2 \) of the distribution of \( Y_1 = \pi Y \) when (3.13) holds. More precisely, let \( P_{\alpha, \omega, \sigma} \) denote the normal distribution of \( Y_1 \) when (3.13) holds for \( Y \). To describe this normal distribution, decompose \( V^* = \pi(V) \) as

\[
V^* = (W_R + W_C) \oplus W_3
\]

where \( W_R \) is the row space of \( k \times l \) matrices, \( W_C \) is the column space of \( k \times l \) matrices and \( W_3 \) is the orthogonal complement of \( W_R + W_C \). Let \( A_C \) denote the orthogonal projection onto \( W_C \). A routine calculation then shows that when (3.13) holds

\begin{equation}
P_{\alpha, \omega, \sigma} = N(e e', \omega^2 A_C + \sigma^2 I^*)
\end{equation}

(3.14)

where \( e \) is the vector of ones in \( \mathbb{R}^l \), \( I^* \) is the identity on \( V^* \), and \( \alpha \) is a
k-vector with coordinates \( \alpha_1, \ldots, \alpha_k \). Given any distribution \( \mu \) on \((\alpha, \omega, \sigma)\), let

\[
S_\mu = \int P_{\alpha, \omega, \sigma} \mu(d\alpha, d\omega, d\sigma).
\]

The argument below establishes an upper bound on

\[
\inf_{\mu} \left\| P_1 - S_\mu \right\|.
\]  
(3.15)

Following our general method, let \( U \) be uniform on \( G \). For \( y \in V \) write

\[
y = y_1 + y_2 + y_3
\]
where \( y_1 \in M_R, \ y_2 \in (M_C - M_0) \) and \( y_3 \in M_1 \). Because \( U \in G \),

\[
Uy = y_1 + Uy_2 + Uy_3.
\]  
(3.16)

The arguments in Section 2 show that

\[
d = \sup \left\| \mathcal{Z}(\pi Uy) - P_{\alpha, \omega, \sigma} \right\|
\]  
(3.17)

is an upper bound on (3.15) where \( \alpha, \omega, \sigma \) are all allowed to depend on \( y \). This dependence is specified explicitly later in the argument.

We now proceed to bound (3.17). Let \( E_m \) denote the \( m \times m \) matrix of all ones. Then

\[
T_2 = K^{-1}E_K \otimes (I_L - L^{-1}E_L)
\]
is the orthogonal projection onto \( M_C - M_0 \) and with

\[
T_1 = I_K \otimes (L^{-1}E_L),
\]

\[
T_3 = I - T_1 - T_2 \text{ is the orthogonal projection onto } M_1.
\]

Because \( M_C \) and \( M \) are \( G \)-invariant subspaces, it follows that

(i) \( Uy_2 \) and \( Uy_3 \) are independent

(ii) \( Uy_2 \) has the same distribution as \( |y_2| Uv_2 \) where \( v_2 \) is a fixed unit vector in \( M_C \)

(iii) \( Uy_3 \) has the same distribution as \( |y_3| Uv_3 \) where \( v_3 \) is a fixed unit vector in \( M_1 \).

The independence and results in Appendix A imply
\[
\text{Cov}(\pi U y) = \text{Cov}(y_2|\pi UV_2) + \text{Cov}(y_3|\pi UV_3) = \\
(3.18) \\
|y_2|^2(L-1)^{-1} \pi T_2 \pi' + |y_3|^2[(K-1)(L-1)]^{-1} \pi T_3 \pi' - \Sigma(y).
\]

Let \(\nu_2\) and \(\nu_3\) be normal distributions on \(V^*\) with mean 0 and covariances given by the summands above. With

\[\mu_i = \mathcal{L}(y_i|\pi U y_i), \ i = 2, 3,\]

and \(\sigma e' = \pi y_1\), we have

\[
||\mathcal{L}(\pi U y) - P_\omega, \omega, \sigma|| = ||\mathcal{L}(\pi U y_2 + \pi U y_3) - P_0, \omega, \sigma|| \leq \\
(3.19) \\
||\mu_2^* \mu_3 - \nu_2^* \nu_3|| + ||\nu_2^* \nu_3 - P_0, \omega, \sigma||
\]

However, using the elementary inequality

\[
||\mu_2^* \mu_3 - \nu_2^* \nu_3|| \leq ||\mu_2 - \nu_2|| + ||\mu_3 - \nu_3||
\]

and the results in Appendix A, it follows that for all \(y\),

\[
||\mu_2^* \mu_3 - \nu_2^* \nu_3|| \leq B(k, l; K-L) + B(kl, 1; (K-L)(L-1)) \\
(3.20)
\]

Because \(\nu_3^* \nu_2\) and \(P_0, \omega, \sigma\) are both normal, Proposition A.5 yields

\[
||\nu_2^* \nu_3 - P_0, \omega, \sigma|| \leq 2 \left[\left(\frac{\det (\omega^2 A_C + \sigma^2 I^*)}{\det \Sigma(y)}\right)^{1/2} - 1\right] \\
(3.21)
\]

as long as

\[
\Sigma(y) \leq \omega^2 A_C + \sigma^2 I^*. \\
(3.22)
\]

By choosing

\[
\sigma^2 = [(K-1)(L-1)]^{-1} |y_3|^2
\]

and

23
\[ \omega^2 = \max \left\{ 0, \frac{k}{K} \left( (K-1)^{-1} |y_2|^2 - \sigma^2 \right) \right\} \]

(3.22) is satisfied. Further a routine, but very tedious calculation shows that

\[ \frac{\det (\omega A_C + \sigma^2 J^*)}{\det \Sigma(y)} \leq \left( \left[ 1 - \frac{k}{K} \right] \left[ 1 - \frac{l}{L} \right] \right)^{-\frac{1}{2}}. \]

(3.23)

Combining (3.20) and (3.23) yields a final upper bound of

\[ B(k,1;K-1) + B(kl,1;(K-1)(L-1)) + \]

\[ 2 \left[ \left\{ \left[ 1 - \frac{k}{K} \right] \left[ 1 - \frac{l}{L} \right] \right\}^{-\frac{1}{2}} - 1 \right] \]

on the variation distance between \( \mathcal{I}(Y_1) \) and the closest mixture of normal models given by (3.14). For fixed \( k \) and \( l \), this bound converges to zero as \( K,L \to +\infty \).

Summary of Example 3.3:
Consider the normal “random effects” model (3.13) whose probability model is denoted by \( Q_{\alpha,\omega,\sigma} \) where \( \alpha \in \mathbb{R}^k, \omega \in [0,\infty) \) and \( \sigma \in [0,\infty) \). A group \( G \) which preserves this model consists of all orthogonal transformations on \( V \), say \( g \), which satisfy

(i) \( gx = x \) for \( x \in M_R \)

(ii) \( g(M_C) = M_C \).

That is, \( gQ_{\alpha,\omega,\sigma} = Q_{\alpha,\omega,\sigma} \) for \( g \in G \).

Theorem 3.2: Let \( P \) be any probability on \( V \), the space of \( K \times L \) real matrices. Also, let \( \pi \) be the “projection” of \( V \) to \( V^* \), the space of \( k \times l \) real matrices, which picks out the \( k \times l \) upper left block of elements of \( V \). If \( P \) is \( G \)-invariant, then there exists a probability measure \( \xi \) on \( \mathbb{R}^k \times [0,\infty) \times [0,\infty) \) such that, for all \( k,l \) satisfying \( k + 4 < K, l < L \), and \( kl + 3 < (K-1)(L-1) \), the variation distance between \( \pi P \) and

\[ \pi \left( \int Q_{\alpha,\omega,\sigma,\xi}(d\alpha, d\omega, d\sigma) \right) \]

is bounded above by

\[ 2 \left[ \left( 1 - \frac{k+3}{K-1} \right)^{-1/2} - 1 \right] + 2 \left[ \left( 1 - \frac{kl+3}{(K-1)(L-1)} \right)^{-1/2} - 1 \right] + \]

\[ 2 \left[ \left\{ \left( 1 - \frac{k}{K} \right) \left( 1 - \frac{l}{L} \right) \right\}^{1/2} - 1 \right]. \]
§4: A Simple Multivariate Linear Model

Theorem 3.1 provides a finite version of a result due to Smith (1981). Smith showed that a sequence $Y_1, Y_2, \ldots$ of random variables is invariant under the group of orthogonal transformations which fix the vector of ones iff the sequence is a translation-scale mixture of iid normals. The finite version of the result follows from Example 3.1 with

(4.1) \[ X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n, \quad X_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^k \]

so $\text{rank}(X) = \text{dim} M = 1$. In this case

(4.2) \[ \text{det} \pi Q \pi' = \frac{n-k}{n} \]

so the variation distance bound is

(4.3) \[ B(k, 1; n-1) + 2 \left[ \left( \frac{n}{n-k} \right)^{1/2} - 1 \right] \]

This yields

**Theorem 4.1:** Let $e$ be the vector of ones in $\mathbb{R}^n$ and set $O_n(e) = \{ g \in O_n | ge = e \}$. Consider random variables $Y_1, \ldots, Y_n$ with distribution $P^{(n)}$ on $\mathbb{R}^n$. Also, let $Z_1, \ldots, Z_n$ be iid $N(\beta, \sigma^2)$ with joint distribution $Q^{(n)}_{\beta, \sigma}$ on $\mathbb{R}^n$. Given $k \leq n$, $P^{(k)}(Q^{(k)}_{\beta, \sigma})$ denotes the joint distribution of $Y_1, \ldots, Y_k(Z_1, \ldots, Z_k)$. If $P^{(n)}$ is $O_n(e)$ invariant, then there exists a probability measure $\xi$ on $\mathbb{R}^1 \times [0, \infty)$ such that for each $k \in \{1, \ldots, n-5\}$, the variation distance between $P^{(k)}$ and

\[ \int_{-\infty}^{\infty} \int_{0}^{\infty} Q^{(k)}_{\beta, \sigma} \xi(d\beta, d\sigma) \]

is bounded above by

\[ 2 \left[ \left( 1 - \frac{k+3}{n-1} \right)^{-1/2} - 1 \right] + 2 \left[ \left( \frac{n}{n-k} \right)^{1/2} - 1 \right]. \]

The main result of this section is a multivariate version of Theorem 4.1. To describe this result, let $Y_1, \ldots, Y_n$ be $p$-dimensional random vectors and form the $n \times p$ matrix
\( Y(n) = \begin{bmatrix} Y_1' \\
\vdots \\
Y_n' \end{bmatrix} \)

This random matrix takes values in the vector space \( V \) of all \( n \times p \) real matrices. The law of \( Y(n) \), say \( P(n) = I(Y(n)) \), is assumed to be \( O_n(e) \) invariant in the sense that

\[(4.4) \quad I(gY(n)) = I(Y(n)), \quad g \in O_n(e). \]

For example, if

\[(4.5) \quad I(Y(n)) = N(e\xi', I_n \otimes \Sigma), \]

then the rows of \( Y(n) \) are iid p-dimensional \( N(\xi, \Sigma) \) random vectors and (4.4) holds.

Now, consider the "projection" \( \pi \) given by

\[ \pi = (I_k, 0) : k \times n \]

so

\[ Y(k) = \pi Y(n) \]

consists of the first \( k \) rows of \( Y(n) \). Our goal is to approximate \( P(k) = I(Y(k)) \) by probabilities of the form

\[(4.6) \quad \xi^{(k)}_\mu \int N(e^{(k)}_\xi', I_k \otimes \Sigma) \mu(\xi, d\Sigma) \]

where \( \mu \) is a probability measure on \( R^p \times (p \times p \) non-negative definite matrices).

**Theorem 4.2:** Suppose (4.4) holds. Then there exists a probability measure \( \mu \) such that for all \( k \) with \( p + k + 2 < n - 1 \), the variation distance between \( P^{(k)} \) and \( S^{(k)}_\mu \) is bounded above by

\[ 2 \left[ \left( 1 - \frac{k + p + 2}{n - 1} \right)^c - 1 \right] + 2 \left[ \left( \frac{n}{n - k} \right)^{p/2} - 1 \right] \]

where \( c = t^2/2 \) and \( t = \min \{k, p\} \).

**Proof:** Let \( U \) be uniform on \( O_n(e) \) and note that

\[ P = \frac{1}{n} ee' \]

is the orthogonal projection onto the one dimensional subspace \( M = \text{span}(e) \subseteq R^n \).

The plan is to compute an upper bound on \( d \) given in (2.7). Set \( Q = I - P \) and fix \( y \in V \). Then Proposition A.1 implies that

\[ 26 \]
(4.8) \[ I(\pi U_y) = I(A^{1/2} \Delta B^{1/2} + \pi P_y) \]

where

\[
(i) \quad A = \pi Q' \quad : \quad k \times k \\
(ii) \quad B = y' Q y \quad : \quad p \times p \\
(iii) \quad \Delta : k \times p \text{ has the distribution of the } k \times p \text{ upper left block of an } (n-1) \times (n-1) \text{ orthogonal matrix uniform on } \mathbb{O}_{n-1}.
\]

Next, pick \( Q_y \) to be

\[ Q_y = N(e \xi', I_n \otimes ((n-1)^{-1} B)) \]

where

\[ \xi = n^{-1} y' e. \]

Then

\[ \pi Q_y = N(\pi^{(k)} \xi', I_k \otimes ((n-1)^{-1} B)) = N(\pi P_y, I_k \otimes ((n-1)^{-1} B)). \]

Therefore

\[
\sup_y || I(\pi U_y) - \pi Q_y || = \\
||I(A^{1/2} \Delta) - N(0, (n-1)^{-1} I_k \otimes I_p)|| + \\
||I(A^{1/2} \Delta) - N(0, (n-1)^{-1} A \otimes I_p)|| + \\
||N(0, (n-1)^{-1} A \otimes I_p) - N(0, (n-1)^{-1} I_k \otimes I_p)||.
\]

Proposition (A.4) shows the first term in the final expression is bounded by \( B(k,p;n-1) \). Applying Proposition A.5, the second term is bounded by

\[ 2 \left[ \det (A \otimes I_p)^{-1/2} - 1 \right] = \\
2 \left[ \left( \frac{n}{n-k} \right)^{p/2} - 1 \right]. \]

\[ \square \]

It should be fairly clear that the methods of Section 3 and 4 can be combined to yield results for multivariate linear models. For example, only minor modifications are necessary to obtain a multivariate version of Example 3.1. The details are omitted.
§5: Symmetry Models

Example 2.1 shows that if a \( k \times p \) random matrix is extendable to an \( n \times p \) random matrix which has an \( O_n \) invariant distribution, then the distribution of the \( k \times p \) matrix is approximable in variation distance by a covariance mixture of \( N(0, I_k \otimes \Sigma) \) distributions. The bound on the variation distance given in (2.8) makes this statement precise. In general, the support of the mixing measure is the set of all \( p \times p \) non-negative definite matrices, \( S^+_p \). However, under further symmetry assumptions, it is possible that the support of the mixing measure can be a "small" subset of \( S^+_p \). This issue is treated in the present section.

Before giving an example, it is useful to introduce symmetry models (see Perlman (1987) for a discussion). Let \( G \) be a closed subgroup of \( O_p \). The set of matrices in \( S^+_p \) which satisfy

\[
g \Sigma g' = \Sigma \quad \text{for all } g \in G
\]

is denoted by \( \phi_p(G) \). The set \( \phi_p(G) \subseteq S^+_p \) is called the symmetry model of \( G \). Obviously \( (cI_p \mid c \geq 0) \) is contained in \( \phi_p(G) \). Classical examples of symmetry models include

\[
\left\{ \begin{array}{l}
G = D_p \quad \text{the group of all coordinate sign changes} \\
\phi_p(D_p) \quad \text{all diagonal covariances}
\end{array} \right.
\]

and

\[
\left\{ \begin{array}{l}
G = \mathcal{O}_p \quad \text{all permutation matrices} \\
\phi_p(\mathcal{O}_p) \quad \text{all intraclass covariance matrices}
\end{array} \right.
\]
If a random matrix $X: n \times p$ has a $N(0, I_n \otimes \Sigma)$ distribution, then

$$I(hX) = I(X)$$

for all $h \in O_n$. Because the rows of $X$, say $X_1', \ldots, X_n'$, are iid $N(0, \Sigma)$, it is also clear that

$$I(g_iX_i) = I(X_i), \quad g_i \in G$$

when $\Sigma \in \phi_p(G)$, $i = 1, \ldots, n$. Thus the distribution of $X$ is invariant under left multiplication by an element of $O_n$ and under multiplication of the rows of $X$ by elements of $G$ (possibly different elements of $G$ for different rows) as long as $\Sigma \in \phi_p(G)$. All such transformations are elements of the orthogonal group $O_V$ of $V$, the space of $n \times p$ matrices with the usual inner product. The smallest closed subgroup of $O_V$ which contains all these transformations is denoted by

$$O_n \cup G.$$  

This symbol was chosen because of the similarly of the group action to that of the wreath product of permutation groups (see Suzuki (1982)). It is not clear to us how to describe the size or structure of $O_n \cup G$. Fortunately, such information is not needed in what follows.

Here is the basic question of this section. Suppose a random matrix $X:n \times p$ has an $O_n \cup G$ invariant distribution and consider $X^{(1)}$, the $k \times p$ upper block of $X$. (Thus, by definition, the distribution of $X^{(1)}$ is extendable to an $O_n \cup G$ invariant $X$.) Is it true that the distribution of $X^{(1)}$ is approximable by mixtures of $N(0, I_k \otimes \Sigma)$ distributions where the support of the mixing measure is contained in $\phi_p(G)$? When $G = (I_p)$, Example 2.1 gives an affirmative answer. Further, when $G = O_p$, $\phi_p(G) = \{cI_p \mid c \geq 0\}$ and a simple application of the method described in Section 2 provides an affirmative answer. Affirmative answers in other cases follow, but we do not know the answer in general. The general question is discussed again later in this section.

Because we cannot describe properties of the uniform distribution on
$O_n \cup \mathbb{C}$, the methods used here are slightly different than those of previous examples. The argument in the following example is typical.

**Example 5.1**: Assume $X: n \times p$ has an $O_n \cup \mathbb{C}$ invariant distribution. The symmetry model, $\phi_p(D_p)$, consists of all $p \times p$ diagonal covariance matrices. With

$$\pi = (I_k, 0): k \times n,$$

$X^{(k)} = \pi X$ is the $k \times p$ upper block of $X$. Let $S$ be all distributions of the form

$$S = \int N(0, I_k \otimes \Sigma) \mu(d\Sigma)$$

where the support of the mixing measure $\mu$ is $\phi_p(D_p)$. In this example, we will show that when $k + 3 < n$,

$$\inf S \| X^{(k)} - S \| \leq p B(k, l; n) \quad (5.5)$$

where $B$ is given in (A.32).

The argument runs as follows. Decompose the vector space $V$ of all $n \times p$ real matrices into the orthogonal direct sum

$$V = V_1 + \ldots + V_p$$

where $V_i$ is the set of $x$'s with $i$th column arbitrary and all remaining elements zero. Each $V_i$ is invariant under the action of $O_n \cup \mathbb{C}$. Consider the mapping

$$\tau(x) = \begin{pmatrix} |x_1| & 0 & \ldots & 0 \\ 0 & |x_2| & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & |x_p| \\ 0 \end{pmatrix} \in V \quad (5.6)$$

where $x_1, \ldots, x_p$ are the columns of $x$. The arguments in Appendix B show that $\tau$ is a maximal invariant mapping (that is; $\tau(x) = \tau(gx)$ for $x \in V$ and $g \in O_n \cup \mathbb{C}$, and $\tau(x) = \tau(y)$ implies $x = gy$ for some $g$). Thus $\tau(x)$ is an
orbit index. Let $U$ be uniform on $O_{-n} \omega_{-p}$, and for $y \in V$, let

$$Q_y = N(0, I_n \otimes (r(y))' r(y)).$$

To establish (5.5), it must be verified that

(5.7) \[ \sup_{y} || \mathcal{L}(\pi U y) - \pi Q_y || \leq k B(k, l; n). \]

Write $y = y_1 + \ldots + y_p$ with $y_i \in V_i$, $i = 1, \ldots, p$, so

$$\pi U y = \pi U y_1 + \ldots + \pi U y_p.$$  

Because $O_{-n} \omega_{-p}$ acts transitively on the unit sphere in each $V_i$, it follows that

(i) $U y_1, \ldots, U y_p$ are independent

(ii) $U y_1$ has the same distribution as $|y_1| U_1$ where $U_1 \in V_1$ is uniformly distributed on the unit sphere in $V_1$.

(iii) if $W_i \in V_i$ is $N(0, |y_i|^2 n^{-1} I_n \otimes R_i)$ where $R_i$ is the orthogonal projection onto $V_i$, then

(5.8) \[ || \mathcal{L}(\pi U y_1) - \mathcal{L}(\pi W_i) || \leq B(k, l; n). \]

Now, (i) implies that the distribution of $\pi U y$ is a $p$-fold convolution, say

$$\mu_1 * \ldots * \mu_p,$$

where

$$\mu_i = \mathcal{L}(\pi U y_i).$$

Further, an easy calculation shows that

$$\pi Q_y = \mathcal{L}(\pi W_1 + \ldots + \pi W_p)$$

where $W_1, \ldots, W_p$ are independent as given in (iii). Hence $\pi Q_y$ is a $p$-fold convolution, say $\nu_1 * \ldots * \nu_p$, with

$$\nu_i = \mathcal{L}(\pi W_i), i = 1, \ldots, p.$$  

An induction argument and the triangle inequality yields the general fact that

(5.9) \[ || \mu_1 * \ldots * \mu_p - \nu_1 * \ldots * \nu_p || \leq P || \mu_i - \nu_i ||. \]

This together with (5.8) yields (5.5)

In summary, we have the following theorem.
Theorem 5.1: Suppose $X_1, \ldots, X_n$ are random vectors in $\mathbb{R}^p$. For each $k = 1, 2, \ldots, n$, form the $k \times p$ matrix $X^{(k)}$ with rows $X'_1, \ldots, X'_k$, and let $P^{(k)}$ denote the law of $X^{(k)}$. Assume that the distribution of $X^{(n)}, P^{(n)}$, is invariant under

(i) left multiplication of $X^{(n)}$ by elements of $O_n$

(ii) transformations of the type $X'_i \rightarrow X'_i g_i, i = 1, \ldots, n$ where each $g_i$ is a $p \times p$ diagonal matrix with $\pm 1$'s on the diagonal.

Then there exists a mixing distribution $\mu$ on $p \times p$ diagonal covariance matrices $\Sigma$ such that for each $k$ with $k + 3 < n$, the variation distance between the mixture of normals

$$\int N(0, I_k \otimes \Sigma) \mu(d\Sigma)$$

and $P^{(k)}$ is bounded above by

$$2p \left[ \left( 1 - \frac{k + 3}{n} \right)^{-1/2} - 1 \right]. \quad \square$$

Example 5.2: We use the notation of the previous example. Let $B_p \cdot \Phi_p$ denote the group $B_p$ generated by coordinate sign changes and permutations. It is well known that

$$\phi(B_p) = \{ cI_p \mid c \geq 0 \}.$$

Thus, we are trying to approximate the distribution of $X^{(k)}$ by distributions of the form

$$S = \int N(0, cI_k \otimes I_p) \mu(dc).$$

The arguments in Appendix B show that $O_n \cup B_p$ acts transitively on the unit sphere in $V$. Thus, for $y \in V$ and $U$ uniform on $O_n \cup B_p$, $Uy$ is uniform on the sphere of radius $|y|$ in $V$. A straightforward application of Proposition A.4 shows that the variation distance between $\mathcal{L}(X^{(k)})$ and the closest $S$ of the form (5.10) is bounded above by $B(kp, 1; np)$ given in (A.32). This yields

Theorem 5.2: In the notation of Theorem 5.1, assume that the distribution of $X^{(n)}, P^{(n)}$, is invariant under

(i) left multiplication of $X^{(n)}$ by elements of $O_n$

(ii) transformations of the type

$$X'_i \rightarrow X'_i g_i, i = 1, \ldots, n$$
where each \( g_i \) is a \( p \times p \) permutation matrix or is a \( p \times p \) diagonal matrix with \( \pm 1 \)'s on the diagonal.

Then there exists a mixing distribution \( \mu \) on \([0, \infty)\) such that for each \( k \) with \( kp + 3 < np \), the variation distance between the mixture of normals
\[
\int N(0, cI_k \otimes I_p) \mu(dc)
\]
and \( P^{(k)} \) is bounded above by
\[
2 \left[ \left( 1 - \frac{kp + 3}{np} \right)^{1/2} - 1 \right].
\]

**Example 5.3:** Here we extend Example 5.1. Write \( \mathbb{R}^p = M_1 + \ldots + M_r \) into an \( r \)-component direct sum orthogonal decomposition and let \( R_i \) be the orthogonal projection onto \( M_i \), \( i = 1, \ldots, r \). Consider the group \( G \subseteq O_p \) given by
\[
G = \{ g \in O_p \mid g(M_i) \subseteq M_j, i = 1, \ldots, r \}.
\]
It is not hard to show that
\[
\phi_p(G) = \{ \Sigma \in S^+ \mid \Sigma = \sum_{i=1}^{r} \alpha_i R_i, \quad \alpha_i \in [0, \infty) \}.
\]
Consider distributions of the form
\[
S = \int N(0, I_k \otimes \Sigma) \mu(d\Sigma)
\]
where \( \mu \) is supported on \( \phi_p(G) \). Thus we can think of \( \mu \) as a measure on the \( r \)-fold product \([0, \infty) \times \ldots \times [0, \infty)\). With \( p_i = \dim M_i \), arguments analogous to those in Example 5.1 show that the variation distance between \( \mathcal{Z}(X^{(k)}) \) and the closest \( S \) of the form (5.11) is bounded above by
\[
\sum_{i=1}^{r} B(kp_i, 1; np_i).
\]
Models of this sort appear in the analysis of designed experiments as developed by Bailey, Nelder, Speed and their coworkers. Speed (1987) and Bailey (1990) are recent survey articles. This completes example 5.3. \( \square \)

There are a few other natural examples where the methods of the three examples above yield positive results. For example, we can establish versions of Example 2.1 for normal models with complex or quaternionic covariance.
structures (see Andersson (1975) for a description of such models). But these
are only minor variations on a theme and are not included here. An obvious example
where we do not have an honest theorem is the case when \( G \) is the permutation
group \( \mathcal{S}_p \), \( p \geq 3 \). (The case of \( p = 2 \) is essentially the same as Example 5.1 with
\( p = 2 \).) Basically the problem is to figure out what a maximal invariant is for
this case. It seems natural to conjecture that a maximal invariant is two
dimensional (corresponding to an intraclass covariance matrix), but our attempts
to prove this have failed, even for the case \( p = 3 \).

Returning to the general case, the group \( O_n \) \( \sim \) \( G \) acts on \( n \times p \) matrices \( x \)
in the manner specified above. Let \( \nu \) be Haar measure on \( G \) (\( \nu(G) = 1 \)), and
consider the map

\[
\rho(x) = \int_G (gx'xg') \, \nu(dg).
\]

In all of the examples we can do, \( \rho(x) \) turns out to be a maximal invariant under
the action of \( O_n \) \( \sim \) \( G \). It is not hard to show \( \rho(x) \in \mathcal{S}_p(G) \) and \( n^{-1} \rho(x) \) is
the maximum likelihood estimator of \( \Sigma \) for the model

\[
\begin{align*}
\mathcal{I}(X) &= N(0, I_n \otimes \Sigma) ; \\
\Sigma &\text{ positive definite, } \Sigma \in \mathcal{S}_p(G) .
\end{align*}
\]

see Andersson, Brosn and Jensen (unpublished). Showing that \( \rho(x) \) is a maximal
invariant in general would be a big step in trying to settle our original
question concerning the support of the mixing measure. The results of Andersson
(1975) on the structure of symmetry models may be relevant.
Appendix A

Several technical results are collected in this appendix. These include distributional results about random orthogonal matrices and variation distance bounds.

A.1: Random Orthogonal Transformations

On $\mathbb{R}^n$, $O_n$ denotes the compact group of $n \times n$ orthogonal matrices. Given a subspace $M$ of $\mathbb{R}^n$, the compact subgroup $O_n(M)$ is defined by

$$O_n(M) = \{ g \in O_n \mid gx = x \text{ for } x \in M \}.$$  

Because $O_n(M)$ is compact, it has a unique invariant probability measure, say $\nu$. If $U$ is a random element of $O_n(M)$ whose law is $\nu$, we say $U$ is uniform on $O_n(M)$.

Elements of $O_n$ are also elements of the $n^2$ dimensional vector space $\mathcal{L}_{n,n}$ of all $n \times n$ real matrices equipped with the usual inner product. Covariances of random elements of $O_n(M)$ are given relative to this inner product on $\mathcal{L}_{n,n}$. In what follows, the notation and use of the Kronecker product is that given in Eaton (1983, Chapter 2).

The easily verified relation

$$O_n(\Gamma M) = \Gamma O_n(M) \Gamma', \quad \Gamma \in O_n$$

is used to simplify some calculation below.

**Proposition:** Let $m = \text{dim } M$, $0 \leq m < n$ and let $P$ be the orthogonal projection onto $M$. If $U$ is uniform on $O_n(M)$, the mean and covariance of $U$ are

$$E U = P$$

$$\text{Cov}(U) = (n-m)^{-1} Q \otimes Q$$

where $Q = I - P$ is the orthogonal projection onto $M^\perp$.

**Proof:** Because of (A.2), it suffices to verify (A.3) for
\[ M_0 = \left\{ x \mid x \in \mathbb{R}^n, x = \begin{bmatrix} \dot{x} \\ 0 \end{bmatrix}, \dot{x} \in \mathbb{R}^m \right\}. \]

For \( M_0 \), it is clear that

\[ \pi(M_0) = \left\{ g \in O_n \mid g = \begin{bmatrix} I_m & 0 \\ 0 & g^* \end{bmatrix}, g^* \in O_{n-m} \right\}. \]

Hence if \( U^*: (n-m) \times (n-m) \) is uniform on \( O_{n-m} \), then

\[ U_o = \begin{bmatrix} I_m & 0 \\ 0 & U^* \end{bmatrix} \]

is uniform on \( O(M_0) \). Because \( \pi(U^*) = \pi(-U^*) \), \( EU^* = 0 \) so

\[ E(U_o) = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} = P_o \]

where \( P_o \) is the orthogonal projection onto \( M_0 \). Two applications of Proposition 2.19 in Eaton (1983) show that

\[ \text{Cov}(U^*) = c I_{n-m} \otimes I_{n-m} \]

where \( c \) is the variance of any element of \( U^* \). That \( c = (n-m)^{-1} \) is an easy calculation. A routine argument shows (A.5) implies (A.3). \( \square \)

Again let \( U \) be uniform on \( O(M) \), let \( P \) be the orthogonal projection onto \( M \) and set \( Q = I - P \). For two given matrices \( \alpha: r \times n \) and \( \beta: s \times n \), the results below describe the distribution of

\[ V = (\alpha \otimes \beta)U - \alpha U \beta' \]

With \( m = \text{dim} M < n \), it is assumed that \( r \) and \( s \) are no larger than \( n - m \). The two non-negative definite matrices

\[ A = \alpha Q \alpha', \quad B = \beta Q \beta' \]

appear below as do their symmetric square roots denoted by \( A^{1/2} \) and \( B^{1/2} \).

**Proposition A.2:** Let \( U^* \) be uniform on \( O_{n-m} \) and let \( \Delta: r \times s \) be the upper left
corner block of $U^\ast$. Then
\begin{equation}
I(V) = I(A^{1/2}A^\ast B^{1/2} + aPb') .
\end{equation}

Further, if $r + s \leq n - m$, then $\Delta$ has a density (with respect to Lebesgue measure) concentrated on the set $(\Delta'\Delta \leq I_s)$. When $s \leq r$, the density of $\Delta$ is
\begin{equation}
f(\Delta; r, s) = (2\pi)^{\frac{-rs}{2}} \frac{w(n-m-r,s)}{w(n-m,s)} |I_s - \Delta'\Delta|^{(n-m-r-s-1)/2}
\end{equation}
where $w(\cdot, \cdot)$ is the Wishart constant defined by
\begin{equation}
[w(t,p)]^{-1} = \pi^{p(p-1)/4} 2^{tp/2} \prod_{j=1}^{p} \Gamma \left( \frac{t-j+1}{2} \right).
\end{equation}
Here $p$ is a positive integer and $t$ is a real number, $t > p-1$. When $r \leq s$, the density of $\Delta$ is obtained by interchanging $r$ and $s$ in the Wishart constants in (A.9).

Proof: As in Proposition A.1, it suffices to establish the proposition when $M = M_0$. In this case
\begin{equation}
P_o = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}
\end{equation}
is the projection onto $M_0$ and
\begin{equation}
Q_o = (I - P_o) = C_o C_o'
\end{equation}
where
\begin{equation}
C_o = \begin{bmatrix} 0 \\ I_{n-m} \end{bmatrix} : n \times (n-m).
\end{equation}
When $M = M_0$
\begin{equation}
V = \alpha U_o \beta' = \alpha(P_o + Q_o) U_o (P_o + Q_o) \beta' = \alpha Q_o U_o Q_o \beta' + aP_o \beta'.
\end{equation}
The third equality follows from $P_o Q_o = 0$ and $P_o U_o = U_o P_o = P_o$.

As in the proof of Proposition A.1,
\[ U_0 = \begin{pmatrix} I_m & 0 \\ 0 & U^* \end{pmatrix} \]

where \( U^* \) is uniform on \( O_{n-m} \). The use of (A.11) now yields

(A.13) \[ V = \alpha C_{O} C' U_{O} C C' \beta' + \alpha P_{O} \beta' = \gamma U^* \delta' + \alpha P_{O} \beta' \]

where

(A.14) \[ \gamma = \alpha C_{O} \text{ and } \delta = \beta C_{O} . \]

With

\[ A_{O} = \gamma \gamma' = \alpha Q_{O} \alpha', B_{O} = \delta \delta' = \beta Q_{O} \beta' , \]

\( \gamma \) and \( \delta \) can be written

\[ \gamma = A_{O}^{1/2} (I_{rO}) \psi_{1}, \delta = B_{O}^{1/2} (I_{sO}) \psi_{2} \]

where \( \psi_{i} \in O_{n-m} , i = 1,2 \). This well-known representation (sometimes called the polar decomposition) follows easily from Vinograd’s Theorem (see Eaton, (1983), Example 1.11, p. 37). Because \( U^* \) is uniform on \( O_{n-m} \)

\[ \mathcal{I}(U^*) = \mathcal{I}(\psi_{1} U^* \psi_{2}). \]

Substitution into (A.13) yields

\[ \mathcal{I}(V) = \mathcal{I}(A_{O}^{1/2} (I_{rO}) \psi_{1} U^* \psi_{2} \begin{pmatrix} I_{s} \\ 0 \end{pmatrix} B_{O}^{1/2} + \alpha P_{O} \beta') = \]

\[ \mathcal{I}(A_{O}^{1/2} (I_{rO}) U^* \begin{pmatrix} I_{s} \\ 0 \end{pmatrix} B_{O}^{1/2} + \alpha P_{O} \beta') = \]

\[ \mathcal{I}(A_{O}^{1/2} \Delta B_{O}^{1/2} + \alpha P_{O} \beta') \]

where

\[ \Delta = (I_{rO}) U^* \begin{pmatrix} I_{s} \\ 0 \end{pmatrix} \]

is the \( r \times s \) upper left block of \( U^* \). Thus (A.8) holds when \( M = M_{O} \).

The second assertion concerning the density of \( \Delta \), originally due to Khatri (1970), is proved via invariance methods in Eaton (1985).

**Corollary A.3**: If \( r + s \leq n-m \) and if \( A \) and \( B \) have full rank, then \( V \) has a
density given by
\begin{equation}
(A.15) \quad f(V) = |A|^{-s/2} |B|^{-r/2} f(A^{-1/2}(V-aP'b')B^{-1/2};r,s)
\end{equation}
where \( f(\cdot; r,s) \) is given by (A.9).

**Proof:** Using (A.8) and computing a Jacobian, the result follows immediately. \( \square \)

Note that (A.3) implies that \( V \) given by (A.6) satisfies
\begin{equation}
(A.16) \quad \text{Cov}(V) = (n-m)^{-1} A \otimes B
\end{equation}
with \( A \) and \( B \) given in (A.7).

**A.2: Variation Distance Bounds**

Now, suppose \( \Delta : r \times s \) has the density (A.9) with \( m = 0 \), and \( r + s \leq n \).

Thus \( \Delta \) is the \( r \times s \) left upper block of a random matrix \( U \) on \( \mathbb{O}_n \). Clearly \( E\Delta = 0 \) and
\begin{equation}
\text{Cov}(\Delta) = \frac{1}{n} I_r \otimes I_s.
\end{equation}
Also let \( X : r \times s \) have a multivariate normal distribution with the same mean and covariance as \( \Delta \). With \( \mathcal{I}(X) = P_1 \) and \( \mathcal{I}(\Delta) = P_2 \), the results below give an upper bound on the variation distance
\begin{equation}
\delta_{r,n} = \|P_1 - P_2\| = 2 \sup_B |P_1(B) - P_2(B)|
\end{equation}
between \( P_1 \) and \( P_2 \). Here \( B \) ranges over all Borel sets. In what follows, the case of \( r \geq s \) is treated. In this case the density of \( X \) is
\begin{equation}
(A.17) \quad f_1(x) = (\sqrt{2\pi n}^{rs})^{rs/2} \exp \left[ -\frac{1}{2} n \text{tr} x'x \right]
\end{equation}
where \( x \) is a real matrix.

The density of \( \Delta \) is
\begin{equation}
(A.18) \quad f_2(x) = (\sqrt{2\pi})^{-rs} \frac{w(n-r,s)}{w(n,s)} |I_s - x'x|^{(n-r-s-1)/2} I_o(x'x)
\end{equation}
where \( I_o \) is given by
\begin{equation}
I_o(x'x) = \begin{cases} 
1 & \text{if } 0 < x'x \leq I_s \\
0 & \text{otherwise}
\end{cases}
\end{equation}
Because \( f_1 \) and \( f_2 \) are both functions of \( x'x \), the variation distance \( \delta_{r,n} \) is equal to the variation distance between the distribution of \( x'x \) and of \( \Delta'\Delta \). The density function of \( x'x \) is

\[
g_1(v) = \frac{w(r,s)n^{r s/2}}{\sqrt{\frac{1}{\mathbf{v}}} (r-s-1)/2} \exp[-1/2n \mathbf{v}] \mathbf{J}_0(\mathbf{v})
\]

where \( \mathbf{v} \) is a \( s \times s \) symmetric matrix and

\[
\mathbf{J}_0(\mathbf{v}) = \begin{cases} 1 & \text{if } \mathbf{v} > 0 \\ 0 & \text{otherwise} \end{cases}
\]

The density function of \( \Delta'\Delta \) is

\[
g_2(v) = \frac{w(r,s)w(n-r,s)}{w(n,s)} |\mathbf{v}|^{(r-s-1)/2} |\mathbf{I}-\mathbf{v}|^{(n-r-s-1)/2} \mathbf{I}_0(\mathbf{v}).
\]

This multivariate beta density can be found a number of places for example, see Olkin and Rubin (1964). The variation distance \( \delta_{r,n} \) is thus

\[
\delta_{r,n} = \int |g_2(v) - g_1(v)| dv = 2 \int_{\mathbf{E}} \left( \frac{g_2(v)}{g_1(v)} - 1 \right) g_1(v) dv
\]

where \( \mathbf{E} \) is the set of \( s \times s \) positive definite matrices such that \( g_2(v) > g_1(v) \).

Hence

\[
\frac{1}{2} \delta_{r,n} \leq \sup_{v \in \mathbf{E}} \left[ \frac{g_2(v)}{g_1(v)} - 1 \right] = M_{r,n}.
\]

Differentiation shows that the sup in (A.21) is achieved uniquely for \( \mathbf{v} \) equal to

\[
\hat{\mathbf{v}} = (r+s+1)n^{-1} \mathbf{I}_s.
\]

After some algebra this yields

\[
M_{r,n} = \frac{1}{2} \sum_{j=1}^{s} \left[ \frac{\Gamma\left( \frac{n-j+1}{2} \right)}{\Gamma\left( \frac{n-r-j+1}{2} \right) \Gamma\left( \frac{n}{2} \right)} \exp \left\{ \frac{n-r-s-1}{2} \log \left( 1 - \frac{r+s+1}{n} \right) + \frac{r+s+1}{2} \right\} \right]
\]

We now proceed to bound \( M_{r,n} + 1 \) when \( r \) is an even integer. The case of \( r \) odd is treated later. For notational convenience, let
\[
\begin{aligned}
\begin{cases}
m = n/2 \\
p = \frac{r}{2} + 1 \\
t = (r + s + 1)/2 \\
m_j = (n - j + 1)/2
\end{cases}
\end{aligned}
\]

Then

\[
M_{r,n+1} = \prod_{j=1}^{s} \exp \left\{ \log \frac{\Gamma(m_j)}{\Gamma(m_j-p+1)m_j^{p-1}} + m \int_0^{t/m} -\log(1-x)dx \right\}
\]

Because \( r \) is even, \( p \) is an integer, so

\[
M_{r,n+1} = \prod_{j=1}^{s} A_j
\]

where

\[
A_j = \sum_{i=1}^{p-1} \log \left( 1 - \frac{i + (j-1)/2}{m} \right) + m \int_0^{t/m} -\log(1-x)dx
\]

Because \( A_j \leq A_1 \), \( j = 1, \ldots, s \),

\[
M_{r,n+1} \leq e^{sA_1}
\]

We now proceed to bound

\[
A_1 = \sum_{i=1}^{p-1} \log(1 - i/m) + m \int_0^{t/m} -\log(1-x)dx
\]

Because \( x \to -\log(1-x) \) is an increasing convex function on \([0,1]\), for

\( 0 \leq a \leq b < 1 \) we have

\[
\int_a^b -\log(1-x)dx \leq \frac{b-a}{2} \left[ -\log(1-b) - \log(1-a) \right].
\]

Applying (A.29) to (A.28) yields
\[ A_1 \leq \sum_{i=1}^{p-1} \frac{1}{2} \left[ \log \left( 1 - \frac{i}{m} \right) - \log \left( 1 - \frac{i-1}{m} \right) \right] + \]
\[ \sum_{i=1}^{p-1} \frac{m}{i-1} \int_0^{m} \log(1-x)dx + m \int_0^{m} \frac{t}{m} \log(1-x)dx \]
\[ = \frac{1}{2} \log(1 - \frac{p-1}{m}) + m \int_0^{m} \frac{t}{p-1} \log(1-x)dx. \]

Using (A.29) on the final integral in (A.30) then gives

\[ A_1 \leq \frac{1}{2} \log \left( 1 - \frac{p-1}{m} \right) + m \left[ \frac{t-p+1}{2m} \left[ - \log \left( 1 - \frac{t}{m} \right) - \log \left( 1 - \frac{p-1}{m} \right) \right] \right] \]
\[ = \frac{t-p+1}{2} \left[ - \log \left( 1 - \frac{t}{m} \right) \right] + \frac{t-p}{2} \left[ - \log \left( 1 - \frac{p-1}{m} \right) \right] \]
\[ \leq \frac{t-p+1}{2} \left[ - \log \left( 1 - \frac{t}{m} \right) \right] + \frac{t-p}{2} \left[ - \log \left( 1 - \frac{t}{m} \right) \right] \]
\[ = \frac{s}{2} \left[ - \log \left( 1 - \frac{r+s+1}{n} \right) \right] \]

where the second inequality follows from \( p-1 \leq t \).

**Proposition A.4:** For \( r+s+2< n \), the variation between \( \mathcal{I}(A) \) and \( \mathcal{I}(X) \) is bounded above by \( B(r,s;n) \) where

\[ B(r,s;n) = 2 \left[ \left( 1 - \frac{r+s+2}{n} \right)^{-c} - 1 \right] \]

with \( c = \frac{t^2}{2} \) and \( t = \min(r,s) \).

**Proof:** When \( s \leq r \) and \( r \) is even, (A.21), (A.25), (A.27) and (A.31) combine to give

\[ \delta_{r,n} \leq 2 \left[ \left( 1 - \frac{r+s+1}{n} \right)^{-c} - 1 \right] \]

which is bounded above by \( B(r,s;n) \). When \( s \leq r \) and \( r \) is odd, the argument given in Diaconis and Freedman (1987) shows that \( \delta_{r,n} \leq \delta_{r+1,n} \). By (A.33),
\[
\delta_{n+1,n} \leq 2 \left[ \left( 1 - \frac{r+s+2}{n} \right)^c - 1 \right] = B(r,s;n).
\]

When \( s > r \), just reverse the roles of \( r \) and \( s \). This completes the proof. ||

Remarks A.1: It is possible to obtain a slightly better bound on \( \delta_{r,n} \) by writing (A.25) as

\[
(A.34) \quad \log(M_{r,n+1}) = sA_1 + \sum_{j=2}^{s} (A_j - A_1)
\]

where

\[
(A.35) \quad A_j - A_1 = \sum_{i=1}^{p-1} \log \left( 1 - \frac{i}{m} - \frac{(j-1)}{2m} \right) - \log \left( 1 - \frac{i}{m} \right)
\]

and then carefully bounding (A.35). However, this extra effort yields only a minor improvement on the bound in (A.32).

Remark A.2: In most applications, \( r \) and \( s \) are small compared to \( n \). When \( (r+s+2) < \gamma n \) where \( \gamma \) is fixed, \( 0 < \gamma < 1 \), an easy convexity argument yields

\[
(A.36) \quad B(r,s;n) \leq a \frac{r+s+2}{n}
\]

where

\[
a = 2 \frac{(1 - \gamma)^c - 1}{\gamma}
\]

This bound in (A.36) is qualitatively similar to bounds in Diaconis and Freedman (1980, 1987).

The following provides a useful upper bound on the variation distance between certain normal distributions. The setting for this result is in a finite dimensional inner product space because this generality is needed in our applications. Let \((V, (\cdot, \cdot))\) be a finite dimension inner product space so that the density of a \( N(\mu, \Sigma) \) distribution on \((V, (\cdot, \cdot))\) is
\[ h(x) = (\sqrt{2\pi})^{-n/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} (x-\mu, \Sigma^{-1}(x-\mu)) \right] \]

where \( n = \text{dim } V \) and \(|\Sigma|\) denotes the determinant of \( \Sigma \). The dominating measure is Lebesgue measure on \( V \), and of course \( \Sigma \) is assumed to be non-singular.

**Proposition A.5:** Suppose \( X \) is \( N(0, \Sigma_1) \) and \( Y \) is \( N(0, \Sigma_2) \) on \((V, (\cdot, \cdot))\). If \( \Sigma_1 \leq \Sigma_2 \), and \( \Sigma_1 \) is non-singular, then the variation distance between \( \mathcal{L}(X) \) and \( \mathcal{L}(Y) \) is bounded above by

\[
2 \left[ \left( \frac{|\Sigma_2|}{|\Sigma_1|} \right)^{1/2} - 1 \right]
\]

**(A.37)**

**Proof:** Without loss of generality \( \Sigma_2 = I \). Let \( h_1 \) be the density for \( \mathcal{L}(X) \) and \( h_2 \) be the density for \( \mathcal{L}(Y) \). Then

\[
||\mathcal{L}(X) - \mathcal{L}(Y)|| = 2 \int_E \left( \frac{h_2(x)}{h_1(x)} - 1 \right) h_1(x) dx
\]

where

\[
E = \{ x | h_2(x) > h_1(x) \}.
\]

Hence

\[
||\mathcal{L}(X) - \mathcal{L}(Y)|| \leq 2 \sup_x \left[ \frac{h_2(x)}{h_1(x)} - 1 \right].
\]

But, for each \( x \),

\[
\frac{h_2(x)}{h_1(x)} \leq |\Sigma_1|^{-1/2}
\]

because \( \Sigma_1 \leq I \). This completes the proof. \( \Box \)

**Appendix B: Some maximal invariants**

Here we identify certain functions on \( n \times p \) matrices as maximal invariants under the action of \( O_n \circ \mathbb{G} \) introduced in Section 5. To show that a given
function \( r(x) \) is maximal invariant, it is necessary and sufficient that

(i) \( r(x) \) be invariant

(ii) for each \( r_o \) in the range of \( r \), \( 0 \leq G \) act transitively on

the level set \( \{ x | r(x) = r_o \} \).

Before proceeding to specific cases, we introduce some basic operations:

For \( \theta \in \mathbb{R}^1 \) and integers \( i, j \) \((i \neq j, i, j = 1, \ldots, n)\) let \( R_{ij}(\theta) \) be the \( n \times n \) orthogonal matrix which rotates (counter clockwise) the \((i, j)\) coordinates of an \( n\)-vector through an angle \( \theta \). Further, let \( C_{jk} \) denote the linear transformation on \( n \times p \) matrices which changes sign of the \((j,k)\) element.

By an elementary wiggle of angle \( \theta \), we understand an operation on \( n \times p \) matrices \( x \) of the type

\[
W_{ij};k(\theta) (x) = R_{ij}(\theta) [C_{jk}(R_{ij}(-\theta)x)] .
\]

Thus, a wiggle is composed of a rotation of angle \(-\theta\), then a sign change and finally a rotation back again. Clearly an elementary wiggle only affects two elements of the \( n \times p \) matrix - namely \( x_{ik} \) and \( x_{jk} \). Moreover, if

\[
\tilde{v} = (\tilde{x}_{ik}, \tilde{x}_{jk})
\]

is the image of \((x_{ik}, x_{jk})\) after wiggling, we have

**Lemma B.1:** Given \( x_{ik} \) and \( x_{jk} \), there is an elementary wiggle such that \( \tilde{x}_{ik} = 0 \).

**Proof:** Write

\[
(x_{ik}, x_{jk}) = \rho (\cos \eta, \sin \eta)
\]

with \( \rho \geq 0 \). Wiggling with an angle \( \theta \) produces

\[
\rho (\cos \tilde{\eta}, \sin \tilde{\eta})
\]

where

\[
\tilde{\eta} = 2\theta - \eta .
\]

Choosing \( \theta = 1/2\eta + 1/4\pi \) gives \( \tilde{x}_{ik} = 0 \).
Case B.1: $G = D_p$.

Here, it is shown that $\tau(x)$ given in (5.6) is a maximal invariant under the action of $O \omega D_p$ on $n \times p$ real matrices, when $n \geq p$. $\tau(x)$ is invariant because it is invariant under the generators of $O \omega D_p$. Given $x: n \times p$, we first claim that there is an $h \in O_n \omega D_p$ such that

(B.1) \hspace{1cm} h(x) = \tau(x).

The group element $h$ is constructed in steps. First, there is a $\psi \in O_n$ such that $\psi x = y$ is upper triangular. For $i=1$, $j=2$ and $k=1, \ldots, p$, wiggle $y$ to give $\tilde{y}$ with $\tilde{y}_{1,k} = 0$ for $k = 2, \ldots, p$. Each of these wiggles is in $O_n \omega D_p$. Now, continue this wiggling process row by row to obtain a matrix with zeroes everywhere except possibly on the main diagonal. Finally, change the sign of the diagonals to be positive. This process produces an $h$ such that (B.1) holds.

Thus, if $\tau(x) = \tau(y)$, then we can find $h_1$ and $h_2$ such that

$$h_2(x) = \tau(x) - \tau(y) = h_2(y)$$

so that

$$x = h_1^{-1} h_2(y).$$

Hence $O \omega D_p$ acts transitively on the level sets of $\tau$ and $\tau$ is a maximal invariant.

Case B.2: $G = B_p$ (Example 5.2). The claim is that a maximal invariant is

$$\tau(x) = |x| x_0$$

where $|x|$ is the norm of the matrix $x$ and $x_0$ is the fixed matrix whose $(1,1)$ element is one and all other elements are zero. To see this, let $h \in O_n \omega B_p$ be chosen as in Case B.1 so that $h(x)$ has the form (5.6). Then, permute the elements of $h(x)$ one row at a time to put all the non-zero elements of $h(x)$ into the first column. Finally, use an element of $O_n$ to transform this into $|x| x_0$. 

46
This construction shows there is an $h_1 \in \omega \omega_B$ so that

$$h_1(x) = |x|_x.$$ 

Now repeat the argument of Case B.1 to conclude that $\tau(x)$ is a maximal invariant. \hfill \Box
References


