Invariance, Identifiability, and Morphometrics

Subhash R. Lele and Charles E. McCulloch

The form of an object is that characteristic that remains invariant under a group of transformations comprising translation, rotation, and possibly reflection. Group invariance thus naturally plays an important role in the statistical analysis of forms. We examine the existing methods for the statistical analysis of form from the invariance perspective. We begin with a review of the important basic ideas behind invariance and derive a maximal invariant under the group of transformations consisting of rotation, reflection, and translation and its distribution under the Gaussian and elliptically symmetric perturbation models. We first consider the single-sample case and discuss the issue of identifiability of the parameters. We show that method-of-moments estimators based on the distances between landmarks and maximum likelihood estimators (MLEs) based on the size and shape coordinates are invariant and estimate identifiable parameters. However, a number of commonly used methods do not. We compare the statistical and computational efficiencies of the method-of-moments estimator and MLE and show that the method of moments substantially simplifies the estimation procedure computationally with only a small loss of statistical efficiency. We then extend the discussion of invariance to the comparison of two forms. We discuss the relationship between identifiability and invariance in the two-sample case and again show that many commonly used methods base inference on nonidentifiable parameters and discuss the scientific implications of basing inferences on nonidentifiable parameters using a biological example. We provide a brief summary of a method for shape analysis that is invariant.

KEY WORDS: Maximal invariants; Nuisance parameters; Shape analysis.

1. INTRODUCTION

The analysis of the form (or shape and size) of an object has a long history in the biological sciences (e.g., Thompson 1917; Huxley 1932). One of the reasons for the importance of the study of form is the belief that the form is directly related to the function of the organism (Thompson 1917). Using such relationships is useful not only in the study of evolution of organisms, but also in medical statistics; for example, in plastic surgery, where prediction of the effect of the surgical procedure on the growth of a child is necessary to decide on the best course of action (Dufresne and Richtsmeier 1995; Fang, Srinivasan, Raghavan, and Richtsmeier 2000). Morphometrics, or the quantitative analysis of form, is thus an important methodology in a number of fields, including evolutionary biology (Gould 1977), surgery, brain mapping (Richtsmeier 1993), genetics, and protein science (Godzik 1996).

The analysis of form has also attracted substantial attention in the statistical literature since the publication of review articles by Bookstein (1986) and Kendall (1989). Kendall’s approach was recently reviewed by Small (1996). The main emphasis in Kendall’s work is development of the probabilistic and mathematical framework for studying shapes using the Kendall shape space. In contrast, Bookstein and colleagues have emphasized the statistical and inferential aspects of shape analysis. Our article deals mainly with these latter aspects of shape analysis and hence concentrates on the Bookstein school of shape analysis. Intuitively, the definition of the “form” of a geometrical object should be invariant under the group of transformations consisting of translation, rotation, and possibly reflection. Bookstein et al. (1985) argued that whenever an arbitrary choice is made, “its impact upon subsequent statistical analysis [should be] explicitly investigated.” This article takes a careful look at the relationship between this innate invariance and various methods developed for the statistical analysis of forms. Invoking invariance leads to questions of identifiability of parameters. Unfortunately, the conclusion is that in many commonly used and/or intuitively appealing methods, inferences are not based on identifiable parameters.

The article is organized as follows. Section 2 gives a brief discussion of the nature of landmark coordinate data and specifies the basic stochastic model that we use. Section 3 discusses the single population case and considers parameter identifiability and estimation. Section 4 studies the problem of comparing two forms using deformation and superimposition of mean forms, as well as comparison in terms of distances between landmarks. Section 5 illustrates the practical importance of invariance and identifiability by analyzing a biological dataset. We conclude by summarizing the implications of invariance to the statistical analysis of landmark coordinate data.

2. A BRIEF INTRODUCTION TO LANDMARK COORDINATE DATA AND THE PERTURBATION MODEL

Practical analysis of biological form is hindered by the complexity of quantifying an entire form. One approach to reducing this complexity is to consider a few biologically important points on the form under consideration and assume that the configuration of these points approximates the underlying form adequately for the problem at hand. Such points are called “landmarks.” These landmarks not only summarize the form of the object in the geometrical sense, but also define a correspondence of biological structures between the objects under study. Hence they are called “homologous landmarks.” Various authors have discussed the biological basis of “homology and landmarks” (see, e.g., Roth 1988; Van Valen 1982; Hall 1994; and Valeri, Cole, Lele, and Richtsmeier 1997).

Suppose that the choice of the landmarks to be collected is decided on. They can be easily quantified using current

© 2002 American Statistical Association
Journal of the American Statistical Association
DOI 10.1198/016214502388618609

796
imaging technology. For example, we could fix a coordinate system arbitrarily and locate the object in this system. Next, we calculate the coordinates of each landmark with respect to this arbitrarily chosen coordinate system using a digitizer or other device (Corner, Lele, and Richtsmeier 1992; Richtsmeier, Paik, Elfert, Cole, and Dahlman 1995). If there are \( K \) landmarks on a \( D \)-dimensional object, then we can construct a \( K \times D \) matrix whose \( j \)th row consists of the \( D \) coordinates of the \( j \)th landmark. In typical applications, \( D \) is usually either 2 or 3 and \( K \) is assumed to be larger than \( D \) to avoid degeneracy. We use \( X_i \) to denote the \( K \times D \) matrix of coordinates for the \( i \)th individual.

### 2.1 Perturbation Model

We model the interindividual variability by the Gaussian perturbation model described by, for example, Goodall (1991) and Lele (1993). The perturbation model may be thought of as representing the following process. To generate a random geometrical object or, equivalently, a \( K \) point configuration in \( D \)-dimensional Euclidean space, nature first chooses a mean form (represented by matrix \( M \)) and perturbs the elements of this matrix by adding noise to this mean form according to a matrix-valued Gaussian distribution. The \( K \) point configuration so obtained is then rotated and/or reflected by an unknown angle and translated by an unknown amount. Such perturbed, translated, rotated, and/or reflected \( K \) point configurations constitute our data.

The foregoing description can be put in a mathematical form as follows. Let \( M \) be a \( K \times D \) matrix corresponding to the mean form. Let \( E_j \) be a \( K \times D \) matrix-valued Gaussian random variable (Arnold 1981, pp. 309–323) representing the error for the \( j \)th individual. We assume that \( E_j \) is Gaussian with mean matrix 0 and covariance matrices given by \( \Sigma_k \) and \( \Sigma_D \). The matrix \( \Sigma_k \) describes the covariances between elements within the same column of \( E \) and \( \Sigma_D \) describes the covariances within the rows of \( E \). Notationally, we have \( E_j \sim N(0, \Sigma_k, \Sigma_D) \). More precisely, if we stack the matrix, \( E_j \), into a vector, \( \text{vec}(E_j) \), then we have \( \text{var}(\text{vec}(E_j)) = \Sigma_k \otimes \Sigma_D \). Let \( R \) be an orthogonal matrix corresponding to the rotation and possibly reflection of the \( i \)th individual, and let \( t_i \) be a \( D \times 1 \) matrix corresponding to the translation of the \( i \)th individual. Then the landmark coordinate matrix corresponding to the \( i \)th individual may be represented as

\[
X_i = (M + E_i)R_i + t_i^T,
\]

where \( 1 \) is a \( K \times 1 \) matrix of 1s. The random matrices, \( X_i \), thus follow:

\[
X_i \sim N(MR_i + 1t_i^T, \Sigma_k, R_i^T \Sigma_D R_i).
\]

An important thing to notice is that the matrices \( R_i \) and \( t_i \) are unknown and unknowable. These are nuisance parameters, whereas the parameter of interest is \( (M, \Sigma_k \otimes \Sigma_D) \), which we notationally write as \( (M, \Sigma_K, \Sigma_D) \). In fact, there are more unknown parameters than there are number of observations, and the number of parameters grows with the sample size. This problem thus falls in the class of problems described by Neyman and Scott (1948).

### 2.2 Invariance and Elimination of Nuisance Parameters

There are a number of methods for eliminating nuisance parameters so as to deal with Neyman-Scott-type problems (Reid 1995; Liang and Zeger 1995). In particular, Liang and Zeger (1995, p. 165) discussed the use of transforming the data in such a fashion that the distribution of the transformed data is independent of the nuisance parameters. One of their examples is restricted maximum likelihood estimation of variance components, in which invariance to the mean parameters is desired. Although the traditional use of invariance has been in a decision theoretic setting, we instead use invariance to find a transformation of the data such that the distribution of the transformed data does not involve nuisance parameters. The marginal distribution of this transformed data is then used to conduct statistical inferences. Given the strong scientific justification for the use of invariance in size and shape analysis, this approach is natural. Thus the basic idea is to eliminate the nuisance parameters by invoking invariance under the group of transformations induced by rotation, translation, and possibly reflection. Invariance, by reducing the data to an invariant statistic, also typically shrinks the parameter space. To avoid overreduction of the data to whatever extent possible leads us naturally to consideration of the maximal invariant statistic. We thus state the following definition, in which the set \( G \) denotes a group of invertible functions from a set \( C \) to itself.

**Definition 1.** Let \( G \) be a group of invertible functions from a set \( C \) to itself. A statistic \( T(X) \) is called a **maximal invariant statistic** if \( T(\cdot) \) satisfies the following conditions:

a. \( T(g(c)) = T(c) \) for all \( g \in G \) and \( c \in C \).

b. If \( T(c_1) = T(c_2) \), then there exists \( g \in G \) such that \( c_2 = g(c_1) \).

Maximal invariants are not unique. However, all maximal invariants are equivalent in the sense that their sets of constancy, or orbits, coincide. A maximal invariant maps the complete orbit of equivalent distributions to a unique single point. Thus there is a one-to-one-relationship between the orbits defined by the group and points in the maximal invariant space. A consequence of this is that only the orbits are identifiable and individual points on these orbits are not. We clarify this point in the next section.

### 2.3 Invariance and Identifiability

As noted earlier, the use of a maximal invariant statistic, \( T \), typically also shrinks the parameter space. This can create a lack of identifiability of the parameters. We use the following main tool to investigate this. Suppose that \( X \) is distributed according to a probability distribution \( P_\theta, \theta \in \Theta \). We say that \( \theta_1 \) and \( \theta_2 \) are nonidentifiable if and only if for every set \( A, P_{\theta_1}(X \in A) = P_{\theta_2}(X \in A) \). Let \( T(X) \) be a maximal invariant whose distribution depends on \( \nu(\theta) \). It then follows that if \( \theta_1 \) and \( \theta_2 \in \Theta \) are such that \( \nu(\theta_1) = \nu(\theta_2) \), then \( \theta_1 \) and \( \theta_2 \) are not identifiable based on \( T(X) \). Though obvious, two points should be emphasized: (1) any two parameter values for which \( \nu(\theta_1) = \nu(\theta_2) \) will have the same value of the likelihood based on \( T(X) \), and hence (2) cannot be distinguished based on the data, \( T(X) \). From this line of argument, it is
clear that we actually use only the fact that the distribution of $T(X)$ is invariant.

### 2.4 Implications for Landmark Data

We now consider the landmark coordinate data problem. The landmark coordinate matrix is denoted by $X$. Recall that the group of transformations involved in the analysis of landmark coordinates consists of translation, rotation, and possibly reflection. Thus $gX = XR + y^T$. It is easy to see that if $X = N(M, \Sigma_k, \Sigma_D)$, then $gX = N(MR + y^T, \Sigma_k, R^T \Sigma_D R)$, where $=$ means equal in distribution.

We now consider one particular maximal invariant under this group and derive the parameters obtained under it, $T(\cdot)$ and $\nu(\theta)$ in the foregoing notation. Because all of the maximal invariants are equivalent, the identifiability issues that we discuss are the same for all of them.

Let $L$ be a $(K - 1) \times K$ matrix whose first column consists of $-1$s and the rest of the matrix is an identity matrix of dimension $(K - 1) \times (K - 1)$. Now define $T(X) = LXX^T L^T$. Because $Ll = 0$ and $R$ is an orthogonal matrix (i.e., $RR^T = I$), it is easy to see that $T(X) = T(XR + y^T)$ and hence is invariant. To show that it is a maximal invariant, we need to show that, given $T(X)$, it can be mapped back to a unique orbit in the original space. This can be proved using the fact that $T(X)$ is a centered inner product matrix and that there exists a unique (up to rotation, translation, and reflection) mapping from the centered inner product matrix to a coordinate matrix (see Lele 1991, 1993). Next, we derive the distribution of this maximal invariant to understand how the parameter space has been shrunk and to clarify what is identifiable based on the maximal invariant.

We derive the distribution of the maximal invariant where the error matrix $E$, follows an elliptically symmetric distribution with mean 0 and variance-covariance matrix given by $\text{var}(\text{vec}(E^T)) = \Sigma_k \otimes \Sigma_D$.

### 3. ESTIMATION OF PARAMETERS IN THE SINGLE-SAMPLE CASE

Having established the identifiability of certain parameters, the next natural question is whether these parameters can be estimated in a practically suitable and statistically desirable fashion. Various methods have been suggested in the literature. The most commonly used method is based on the generalized Procrustes analysis. The review article by Goodall (1991) provides a description of this method. Alternatively, maximum likelihood estimation can be conducted using the exact shape distributions derived by Mardia and colleagues, or estimators.
can be constructed via the method of moments (Stoyan 1990; Lele 1993).

Method-of-moments estimators are based on the moments of the distribution of the maximal invariant described in the previous section. Lele (1993) showed that under the assumption that \( \Sigma_p = I \), the estimating functions based on the method of moments have a unique, analytical solution. Moreover, Lele (1993) also showed that the estimator of the mean form \( M \) can be estimated up to translation, rotation, and reflection consistently, and the covariance parameter \( \Sigma^T_k = L \Sigma_k L^T \) is also estimable consistently. However, assuming that \( \Sigma_p = I \) imposes restrictions on the applicability of this model, we next extend the method to the situation where \( \Sigma_p \neq I \).

### 3.1 Method-of-Moments Estimators

For notational simplicity, let \( Y = T(X) \). Also let \( Y = [Y_{im}] \), where \( l = 1, 2, \ldots, K \); \( m = 1, 2, \ldots, K \) denote the individual elements of the matrix \( Y \). Similarly, let \( L \Sigma_k L^T = [\sigma_{lm}] \) and \( \Delta_j = [\delta_{jm}] \). Notice that the matrices \( \Delta_j \)'s are symmetric and \( \delta_{jm} = \sqrt{\sigma_{jm} \sigma_{mn}} \). This follows easily from the definition of these matrices.

From the previous section, we know that \( Y \) is distributed as a linear combination of non-central Wishart random variables. The following expressions provide the first two moments of the random matrix \( Y \) (Alam and Mitra 1990):

\[
E(Y_{im}) = \sum_{j=1}^{p} \lambda_j \delta_{jm} + \sigma_{im},
\]

\[
\text{Var}(Y_{im}) = \sum_{j=1}^{p} \lambda_j^2 \delta_{jm}^2 + \sigma_{jm}^2 + 2 \sigma_{jm} \delta_{jm} + 2 \sigma_{im} \delta_{jm} + 2 \sigma_{jm} \delta_{im},
\]

and

\[
\text{Cov}(Y_{im}, Y_{ip}) = \sum_{j=1}^{p} \lambda_j^2 \delta_{jm} \delta_{ip} + \sigma_{jm} \sigma_{ip} + \sigma_{jm} \delta_{ip} + \sigma_{ip} \delta_{jm} + \sigma_{jm} \delta_{ip} + \sigma_{ip} \delta_{jm}.
\]

Equating the sample moments with the population moments, one can numerically obtain the estimates of the parameters \( \lambda_j L \Sigma_k L^T, \ldots, \lambda_j L \Sigma_k L^T \) and \( \lambda_j, \lambda_k, \lambda_k, \ldots, \lambda_k \). These estimators are computationally simple, are especially convenient with large numbers of landmarks, and easily extend from two- to three-dimensional data.

### 3.2 Maximum Likelihood Estimators

Dryden and Mardia (1991, 1992) derived the exact distributions of the form coordinates and the shape coordinates. The likelihood function can be maximized to obtain maximum likelihood estimators (MLEs) of the mean form as well as the covariance parameters. It should be noted again that these parameters are identifiable only within the orbit defined by the group structure. A significant drawback of this approach is that the exact distributions of form and shape coordinates are mathematically complex for large numbers of landmarks and for three-dimensional objects. It is not uncommon in medical problems to need at least 10, 15, or more landmarks to represent an object reasonably. In such a situation, the exact shape distribution is extremely complicated (see eq. 2.4 of Dryden and Mardia 1991 or eq. 2.7 of Dryden and Mardia 1992). It involves telescoping sums increasing geometrically with the number of landmarks.

All of the maximal invariants are equivalent in the sense that MLEs based on any of them give estimates on the same orbit. The use of maximal invariants to eliminate nuisance parameters usually leads to some loss of efficiency as compared to the situation where the nuisance parameters are known. As shown earlier, the method-of-moments estimators based on the maximal invariant \( T(X) \) are easy to obtain. We next study the loss of efficiency of method of moments estimators as compared to MLEs.

Table 1 gives the result of a small simulation study comparing the method-of-moments method based on the interlandmark distances, the MLE based on the Dryden and Mardia distribution (Dryden and Mardia 1992), and the MLE based on the (unobserved) data before translation or rotation. Thus the "unobservable MLE" represents an idealized situation for comparison, because the nuisance parameters, \( R \) and \( t \), are assumed known.

In estimating the interlandmark distances or the variance covariance parameters, none of the methods exhibited any bias; hence Table 1 concentrates on the relative root mean squared error, which primarily reflects variability. Because performance of the methods was similar within the mean parameters, within the variance parameters, and within the covariance parameters, the results are averaged across the (three in each case) parameters. We can see that for estimating the mean form, both the method-of-moments estimator and the MLE do as well as the unobservable MLE. For estimating the variance parameters, the method-of-moments estimator and the MLE do worse than the unobservable MLE, and the method-of-moments estimator does slightly worse than the MLE. For the covariance parameters, some loss of efficiency results from using the method-of-moments estimator compared to the MLE, but not a dramatic amount.

Another point is worth mentioning in comparing the method-of-moments estimator and the MLE. Even for a simple case of three landmarks in two dimensions, we found it

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Parameters</th>
<th>EDMA</th>
<th>MLE</th>
<th>Unobserved MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (M_1, \Sigma_k, 1) )</td>
<td>Mean distances</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Variances</td>
<td>23.2</td>
<td>19.5</td>
<td>16.3</td>
<td></td>
</tr>
<tr>
<td>Covariances</td>
<td>45.6</td>
<td>37.7</td>
<td>28.4</td>
<td></td>
</tr>
<tr>
<td>( (M_1, \Sigma_k, 2) )</td>
<td>Mean distances</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Variances</td>
<td>21.4</td>
<td>18.7</td>
<td>14.6</td>
<td></td>
</tr>
<tr>
<td>Covariances</td>
<td>36.4</td>
<td>31.0</td>
<td>24.7</td>
<td></td>
</tr>
<tr>
<td>( (M_2, \Sigma_k, 1) )</td>
<td>Mean distances</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>Variances</td>
<td>24.2</td>
<td>19.4</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>Covariances</td>
<td>48.4</td>
<td>41.3</td>
<td>30.1</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Shown are the percent relative root mean squared errors for the two methods based on 100 simulations. Samples of size 30 were generated under two different mean forms and two different covariance structures (see Table 2). The column "Unobserved MLE" corresponds to the relative root mean squared error for the MLEs based on assuming that the nuisance parameters are known. Hence these values represent best achievable results.

* Because of nonconvergence of the maximum likelihood routine, these are based on only 98 out of 100 simulations.
Table 2. The Mean Form Coordinates (centered) and the Covariance Matrices Used in the Simulation Study

<table>
<thead>
<tr>
<th>$\Sigma_{k,1}$</th>
<th>$\Sigma_{k,2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.78 .44 .28</td>
<td>.78 0 0</td>
</tr>
<tr>
<td>.44 .66 .40</td>
<td>0 .66 0</td>
</tr>
<tr>
<td>.28 .40 .53</td>
<td>0 0 .53</td>
</tr>
</tbody>
</table>

$$M_1 = \begin{pmatrix} 0 & 4.72 \\ 7.07 & -2.36 \end{pmatrix}$$
$$M_2 = \begin{pmatrix} 1 & 2.33 \\ 0 & -0.67 \end{pmatrix}$$

very difficult to maximize the likelihood numerically using the Dryden and Mardia (1992) distribution. Despite estimating the variance-covariance matrix using a Cholesky decomposition to guarantee positive semidefiniteness, many datasets required new starting values and/or a preliminary sizing step for the variance-covariance parameters to achieve convergence. Despite repeated attempts, there were still 2 datasets (out of 100) in the first parameter configuration for which we could not obtain convergence. More details of the simulation are given in the Appendix.

3.3 Superimposition-Based Estimators

As a practical alternative to maximum likelihood estimation, Goodall and Bose (1987) and Goodall (1991) discussed superimposition-based estimators. (See also Bookstein 1986 for a similar estimator based on shape coordinates.) The basic idea is to first superimpose the objects as closely as possible and then use least squares to estimate the parameters. Intuitively, this attempts to remove the unknown translations and rotations so that the data can be treated as multivariate normal with common mean and variance. If the unknowns could be removed, then estimation of the parameters would be straightforward. But this cannot be done. As a result, these methods lead to inconsistent estimates in the general case where $\Sigma_k \neq I$ and $\Sigma_D \neq I$ (Lele 1993; Kent and Mardia 1997). A further and an important implication of this result is that the covariance estimator is also inconsistent. Lele (1993) illustrated this fact by simulations and provided an intuitive explanation for it.

4. COMPARISON OF FORMS: THE TWO-SAMPLE CASE

We now proceed to apply the principle of invariance to the important case of comparison between two or more populations. Along the lines developed in the first part of this article, it can be shown that the mean forms are identifiable only up to translation, rotation, and reflection. Identifiability of the covariance parameters is also similar to that in the one-sample case. Now suppose that, using two samples, we have estimated the mean forms for each population using either the method-of-moments estimator or the MLE based on the maximal invariants. The question of scientific interest is that of studying the difference in the two mean forms. Several different measures of form difference are used in practice; we describe some of the prominent ones here.

4.1 Deformation Approach to Form Difference

Let $M_1$ and $M_2$ be two mean forms under consideration. A typical way to represent the relationship between $M_1$ and $M_2$ is to consider $M_1$ being deformed into $M_2$ by a function $h$. Typically, a parametric family of functions represents this deformation. Thus we have $M_2 = h(M_1, \phi)$, and the value of the parameter $\phi$ helps us understand the underlying biological mechanisms associated with this deformation.

To make the basic ideas understandable, we consider the simplest case of two-dimensional objects with three landmarks each. We assume that the only possible deformation is of the simplest form, an affine deformation. Under this model, $M_2 = M_1 A + 1^T t$, where $A$ is any $2 \times 2$ matrix and $t$ is a translation vector. Thus $A$ is the parameter of interest and $t$ is the nuisance parameter. It is easy to see that $A = (LM_1)^{-1}(LM_2)$, where $L$ is the centering matrix defined earlier. A physical interpretation can be given to this matrix $A$ by writing its singular value decomposition (Goodall and Green 1986) as

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} \begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}.$$

The angle $\theta$ corresponds to the initial rotation to the principal axes of $M_1$, the singular values $p$ and $q$ correspond to the stretching along the principal axes, and the angle $\varphi$ corresponds to the rotation of the deformed $M_1$ to match exactly with $M_2$. From the scientific viewpoint, the parameters of interest are $(\theta, p, q)$ because they contribute to the deformation, whereas the parameter $\varphi$ does not.

The affine deformation is the simplest type of deformation function. In practice, more complex forms of deformation functions are required. For example, Bookstein (1989, 1991, 1996) suggested using thin-plate splines as a deformation function. An advantage of thin-plate splines is that they depict the deformation geometrically and allow the total deformation to be decomposed into several orthogonal components, which in turn can be studied to localize the form differences. Such localization presumably helps us understand the biological processes that might be responsible for the form change. (See Bookstein 1989, 1991, 1996 for more details and applications.) We would also like to point out that deformation functions other than the thin-plate splines could also be used. Next, we provide a brief description of this method and its variants.

Assume that the objects under consideration are two-dimensional. Let $d_{ij}$ denote the Euclidean distance between landmarks $i$ and $j$ in $M_1$ and let $U(d) = d^2 \log d^2$. Let $U$ be a $K \times K$ matrix given by $U = [U(d_{ij})]_{i,j=1,2,\ldots,K}$. According to the thin-plate splines model, the relationship between the two forms is given by $M_2 = 1^T r + M_1 A + U/N$, where $r$ is the translation parameter, $A$ is any $2 \times 2$ matrix corresponding to the affine deformation, and $N$ is a $K \times 2$ matrix corresponding to the nonaffine deformation parameters and the parameter $N$ is such that $[1^T] N = 0$. The resultant thin-plate spline function is termed the "minimum bending energy" spline.

A variant of this deformation was considered by Rohlf (1996). Let $V = [V_{ij}]$ and $W$ be any $K \times K$ symmetric matrix. Instead of specifying the parameter $N$ to satisfy the above
constraint, this variant specifies that the parameters corresponding to translation and affine deformation are such that 
\[(VWV^T)^{-1}V1W_{M_2} = [t_1']\]. Depending on the choice of the weight matrix, different solutions are obtained.

### 4.2 Superimposition Methods for Comparison of Forms

We now examine the superimposition method for comparing two forms. (For details on this method, see Siegel and Benson 1982; Rohlf and Slice 1990; Lele 1991.) This method of form difference has been used in morphometrics since the beginning of the 20th century (Cole 1996) and was recently suggested by Goodall (1991) and Rohlf and Slice (1990). Dryden and Mardia (1998) provided a book-length treatment of this approach. Small (1996, p. 35) gave a succinct description of this approach: “An attempt to discover the shape differences between sets will typically involve a matching of the sets to determine how differences in the coordinates of corresponding points can be explained through similarity transformations. Any residual differences that cannot be explained through similarity transformations can be understood to be due to differences in shape.”

The model used in the superimposition approach thus may be described in mathematical terms as follows. Suppose that we are interested in describing the difference in two forms \(M_1\) and \(M_2\). The superimposition approach postulates that the two forms relate to each other according to the model 
\[M_2 = 1r + M_1R + D_F\], where \(r\) is a translation parameter, \(R\) is an orthogonal matrix, and the matrix \(D_F\) describes the form difference. If we are interested in shape differences, then the model is given by 
\[M_2 = 1r + \alpha M_1R + D_S\], where \(\alpha > 0\) is a real number taking the scaling of an object into consideration, and the residual matrix \(D_S\) describes the shape difference. The model component \(1r + \alpha M_1R\) is known as the similarity transformation because the resultant objects, after such a transformation, are geometrically similar to the original object \(M_1\). According to the superimposition approach, what remains after the similarity transformation, namely \(D_S\), describes the shape difference.

Consider the model describing the form difference, 
\[M_2 = 1r + M_1R + D_F\]. The Procrustes superimposition approach (Siegel and Benson 1982), also known as the least squares fitting criterion, assumes that the rotation and translation parameters are such that 
\[tr(M_2 - M_1R - 1r) = tr(D_F)\] is minimum. Once the translation and rotation parameters are fixed in this fashion, the form difference is well defined. Other sets of constraints are also used in practice, for example, the edge superimposition scheme commonly used in biology (Kreiberg 1981). This particular method of studying form difference translates and rotates \(M_1\) in such a manner that a landmark and the direction of a chosen edge match with the corresponding landmark and edge in \(M_2\). This in effect imposes the constraint that one particular row and one other element of \(D_F\) is 0. Alternatively, we may be interested in shape differences and hence additionally scale the two objects so that the length of the matching edge is of unit length. This introduces the additional constraint that in the row where one element is 0, the other element is equal to 1.

In practice, of course, the mean forms are unknown, and the estimated mean forms, rotated and translated in some chosen orientation, are used to estimate the various parameters. Dryden and Mardia (1998) have discussed such applications.

### 4.3 Matrix Transformations, Invariance, and Identifiability Issues

The problem of studying form differences can be stated in terms of studying transformation of one \(K \times D\) matrix into another \(K \times D\) matrix. The invariance ideas described in the first part of the article can then be easily extended to the two-sample problem.

Let \(\Omega\) denote the space of all \(K \times D\) matrices. Let \(H: \Omega \rightarrow \Omega\) be the collection of all transformations that map any given \(K \times D\) matrix to another \(K \times D\) matrix. For example, the affine transformation belongs to this class, as do the thin-plate splines and superimposition transformations.

In morphometrics, two \(K \times D\) matrices are considered equivalent to each other if they are a rotation, reflection, and/or translation of each other. Thus we can partition the space \(\Omega\) into equivallency classes, each class consisting of equivalent matrices. Let \(O(M)\) denote the class of matrices equivalent to a given matrix \(M\).

Let \(h_1\) and \(h_2\) be two members of \(H\) such that 
\[h_1(MR_1 + 1r_1') = h_2(MR_2 + 1r_2')R + 1r_2'.\] That is, the resultant matrices from both transformations acting on equivalent matrices are also equivalent to each other. Let us denote by \(H_{M_1,M_2}\), the set of transformations that map an element of the orbit of \(M_1\) to an element of the orbit of \(M_2\).

In our particular case, the key result that follows is that for 
\[h_1, h_2 \in H_{M_1,M_2} \text{ likelihood}(h_1 | \text{ data}) = \text{ likelihood}(h_2 | \text{ data}),\] that is, data cannot help us distinguish between the transformations that map an element of the orbit of \(M_1\) to an element of the orbit of \(M_2\). Data help us determine only the set \(H_{M_1,M_2}\), but not a particular transformation of this set. It seems, though, that the different approaches to studying form difference described in the previous section actually provide a particular transformation. How is this possible? The superimposition method chooses a particular transformation from the class \(H_{M_1,M_2}\) using an external criterion such as least squares. Similarly, the thin-plate splines method chooses a particular transformation from the class \(H_{M_1,M_2}\) that minimizes the bending energy or some such external criterion.

To argue that this may be problematic, we consider a simpler setting. Suppose that we are studying two treatments for reducing blood pressure in which each subject is randomized to one of two treatments and reductions in both diastolic blood pressure and systolic blood pressure are recorded. Let \(X_1, X_2, \ldots, X_n\) denote the bivariate observations under the first treatment and let \(Y_1, Y_2, \ldots, Y_m\) denote the bivariate observations under the second treatment. Suppose that we model these data as \(X_i \sim N(\mu + 1_{ij}, 1)\) and \(Y_j \sim N(\mu + \Delta + 1_{ij}, 1)\), where \(\mu = (\mu')^T\) is the mean reduction in diastolic and systolic blood pressure due to treatment 1, \(\Delta = (\Delta')^T\) is the effect of treatment 2 over and above treatment 1, and the \(i_j's\) correspond to the subject effects.

This problem belongs to the class of problems considered by Neyman and Scott (1948). Suppose that we use invariance to eliminate the nuisance parameters, \(i_j's\), corresponding to
the subjects. Notice that $X_{i,2} - X_{i,1}$ is a maximal invariant under the group of translations and that $X_{i,2} - X_{i,1} \sim N(\mu_2 - \mu_1, 1)$. Similarly, $Y_{i,2} - Y_{i,1} \sim N((\mu_2 - \mu_1) + (\Delta_2 - \Delta_1), 1)$.

Let us denote $\mu_2 - \mu_1 = d_\mu$, $(\Delta_2 - \Delta_1) = d_\Delta$. Arguing as before, $d_\mu$ and $d_\Delta$ are identifiable, and any values in the original parameter space with equal values of $d_\mu$ and $d_\Delta$ will give the same value of the likelihood based on the maximal invariants. An immediate consequence of this is that $\Delta_1$ and $\Delta_2$, the changes in diastolic and systolic blood pressure, are not identifiable. We can only find out if the change in diastolic blood pressure is larger or smaller than the change in systolic blood pressure and, if so, by how much. So the question of how much treatment 2 reduces systolic blood pressure or diastolic blood pressure is unanswered in this situation.

Suppose that instead of accepting this limitation of the data, we decide to follow the superimposition logic and conduct the following analysis:

1. Least squares approach. Because there are nuisance parameters related to the translation group, we first superimpose (in this case, translate) the observations in sample 1 in such a manner that $\sum_{i=1}^{n} \sum_{j=1}^{n} tr((X_i - X_j)(X_i - X_j)^T)$ is minimized. This corresponds to transforming the observations $X_i = (x_{i,1}, \ldots, x_{i,n})$ to $X_i = (x_{i,1} - x_{i,n}/2)$. Similarly, we translate the observations in sample 2 to $Y_i = (y_{i,1} - y_{i,n}/2)$. Now we average these transformed observations to obtain $\overline{X}$ and $\overline{Y}$. It is easy to see that

$$\overline{X} \sim N\left(\frac{(\mu_1 - \mu_2)/2}{(\mu_2 - \mu_1)/2}, \frac{1}{n} \begin{pmatrix} .5 & - .5 \\ - .5 & .5 \end{pmatrix}\right)$$

and

$$\overline{Y} \sim N\left(\frac{(\mu_1 - \mu_2)/2}{(\mu_2 - \mu_1)/2} + \frac{(\Delta_1 - \Delta_2)/2}{(\Delta_2 - \Delta_1)/2}, \frac{1}{m} \begin{pmatrix} .5 & - .5 \\ - .5 & .5 \end{pmatrix}\right).$$

To study the difference between the two populations, now we superimpose (i.e., translate) $\overline{X}$ so that it matches $\overline{Y}$ in the sense that $tr((\overline{X} - \overline{Y})(\overline{X} - \overline{Y})^T)$ is minimized. It can be seen that the best translation in this case is no translation at all, and hence the difference between two populations is given by $\overline{Y} - \overline{X}$. As the sample sizes converge to infinity, this quantity converges to

$$\frac{(\Delta_1 - \Delta_2)/2}{(\Delta_2 - \Delta_1)/2} = \frac{-d_\Delta}{d_\Delta/2}.$$

2. Edge superimposition-type approach. We can conduct similar calculations where instead of translating the observations so that least squares criterion is satisfied, we translate them so that the first component (landmark 1) is matched. Such calculations lead to the difference in two populations, which converges to $\frac{0}{(\Delta_2 - \Delta_1)} = \frac{0}{d_\Delta}$. It leads us to the conclusion that treatment 2 is effective only in reducing systolic blood pressure and has no effect on diastolic blood pressure. Had we decided to match the second component (landmark 2), then we would have come to the conclusion that treatment 2 affects only systolic blood pressure, not diastolic blood pressure.

The choice of a particular side condition thus leads to the choice of a particular solution. Another context where this is familiar is that of factor analysis where loadings are not identifiable, but imposing varimax or quartimax conditions can yield unique values for the factor loadings. In the foregoing example, once a particular side condition is chosen, inferences thereafter are invariant to the translation group. However, it is obvious in this example that the choice of a particular side condition has a significant impact on the scientific conclusions.

Unfortunately, as was pointed out earlier, any values in the original parameter space with equal values of $d_\mu$ and $d_\Delta$ will give the same value of the likelihood, and hence even an infinite amount of data cannot tell us which side condition is “correct.” One may argue that the choice of the side condition is similar to the choice of the “normal” distribution model. However, the difference between choosing a particular model, such as the normal distribution, in the foregoing example and choosing a side condition is that data potentially can refute or support the choice of the normal model, but there is no potential for refuting or supporting a particular choice of the side condition.

The situation in morphometrics is the same as described before, except that the group structure is more complicated, involving rotations and reflections. Once a side condition, such as minimum bending energy or least squares fitting, is chosen, the inferences thereafter are invariant (strictly speaking, equivariant) to this group. However, similar to the simple situation discussed earlier, because the choice of the side condition is not even potentially refutable and can have a significant affect on the scientific inferences, we suggest that only those conclusions that remain stable to the choice of these side conditions are valid conclusions. Immediate questions that come to mind are what quantities remain invariant to the choice of the side conditions, and are they of any use in scientific enquiries. We address the first question in Section 4.4 and the second issue, of scientific relevance in morphometrics, in Section 5.

4.4 Form Comparisons Based on Distances

In the classical approaches to the problem of size and shape (Huxley 1932; Mosimann 1970), the forms and shapes are compared in terms of the distances between landmarks. This approach was extended by Lele and colleagues (Lele 1991, 1993; Lele and Richtsmeier 1991, 1992, 1995, 2000; Richtsmeier and Lele 1992; Lele and Cole 1996). Do these methods satisfy the invariance principle?

Consider a mean form coordinate matrix $M$. Define a new matrix, denoted by $F(M)$, where $F_{ij}(M) = \text{Euclidean distance between landmarks } i \text{ and } j$. This is a symmetric, $K \times K$ matrix with diagonal elements 0, called a Euclidean distance matrix corresponding to the form $M$. Lele (1991, 1993) proved that this is also a maximal invariant under the group of transformations comprising translation, rotation, and reflection. Notice that, unlike the maximal invariants described in the first part of the article, the Euclidean distance matrix $F(M)$, which is also a maximal invariant, is independent of the choice of the translation matrix $L$. Euclidean distance matrix analysis (EDMA) involves the statistical procedures based on analysis of these matrices.
Because $F_y(M)$'s are continuous functions, it is clear that $F(\hat{M}_1)$ and $F(\hat{M}_2)$ are consistent estimators of the corresponding Euclidean distance matrices for the mean form of the two populations. Thus these quantities can be estimated. Moreover, any continuous function of $F(M_1)$ and $F(M_2)$ also can be estimated. Thus, any procedure based only on the entries of these matrices can be estimated and is invariant to the choice of the translation matrix $L$ and to rotation and reflection. For example, the testing procedures and confidence intervals proposed by Lele and Richatsemer (1991, 1995) satisfy the principle of invariance. Similarly, procedures suggested by Mosimann (1970) and by Rao and Suryawanshi (1996) are also invariant, because they are based on distances between the landmarks.

4.5 Comparison of Growth Patterns and s-Sample Comparisons

One of the major problems confronting modern biology is to understand how complex morphologic structures arise during development and how they are altered during evolution (Atchley and Hall 1991). An important component of these processes is the way in which component structures rearrange relative to one another with an overall increase in size. A growth pattern is defined as the composite of geometric changes in structure occurring over time (Richatsmeier and Lele 1993). Gould (1977) discussed the relationships between growth patterns and their implications in the evolutionary context. Richatsmeier and colleagues have studied the implications of growth patterns in clinical situations involving genetic disorders (Dufresne and Richatsmeier 1995). A theoretical framework for studying growth using EDMA was developed by Richatsmeier and Lele (1993).

5. NONINVARIANT DESCRIPTORS AND SCIENTIFIC CONSEQUENCES

In the foregoing discussion, we argued that inferences affected by the choice of the side condition are scientifically invalid. However, a strict application of this principle may be unnecessarily stringent if under all possible choices of the external constraints, the scientific conclusions drawn are essentially identical or at least similar enough to make no practical difference. Nonetheless, if the scientific conclusions vary substantially depending on the choice of the external constraint used to describe the form change, then one should be extremely cautious. Are the descriptors of form change based on superimposition and deformation "scientifically stable?" As shown in the blood pressure example (conceptually) and later, "scientific invariance" does not necessarily hold in real biological situations.

The following is a paraphrased description of an example discussed by Richatsmeier (1987, p. 474) in which the craniofacial form of a normal child is compared with that of a child with Apert syndrome, known to be caused by specific genetic mutations. The purpose behind such studies is to propose possible surgical procedures, determine the timing of these procedures, or infer about the underlying biological processes responsible for the obvious differences in craniofacial form. Morphometric studies of these syndromes routinely use cephalometric radiographs. A cephalometric radiograph (roentgenogram) is an X-ray produced under controlled conditions that allow for correction of distortion and enlargement that occur during exposure of the film by an X-ray beam. The data analyzed consist of two-dimensional coordinates of landmarks from cephalometric radiographs. (More details of the genetic disorder and the justification of the two-dimensional approximation are available in Richatsmeier 1987 and references therein.) Eleven landmarks were collected from a normal individual and an individual with Apert syndrome. The landmarks are listed in Table 3 and the form represented by the landmarks, with the outline of the skull added, is shown in Figure 1.

We first illustrate the effect of the choice of the constraint on the method of superimposition. The solid dots are the landmarks corresponding to the normal skull, whereas the open circles correspond to the malformed skull. Figure 2(a) shows one possible superimposition of the two forms. This superimposition corresponds to matching landmarks 1 (nasion) and 7 (tuberculum sella) on the two forms. These landmarks for superimposition are commonly chosen in biological studies (Broadbent, Broadbent, and Golden 1975; Kreiberg 1981).

Table 3. Abbreviation and Description of Landmarks Used in the Comparison of the Craniofacial Form of a Normal Individual With an Individual With Apert Syndrome

<table>
<thead>
<tr>
<th>Landmark index</th>
<th>Landmark abbreviation</th>
<th>Landmark description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NAS</td>
<td>Nasion</td>
</tr>
<tr>
<td>2</td>
<td>NSL</td>
<td>Nasale</td>
</tr>
<tr>
<td>3</td>
<td>ANS</td>
<td>Anterior nasal spine</td>
</tr>
<tr>
<td>4</td>
<td>IDS</td>
<td>Intradentale superior</td>
</tr>
<tr>
<td>5</td>
<td>IGW</td>
<td>Intersection of the greater wings of the sphenoid with planum sphenoidale and/or the cribiform plate.</td>
</tr>
<tr>
<td>6</td>
<td>PNS</td>
<td>Posterior nasal spine</td>
</tr>
<tr>
<td>7</td>
<td>TBS</td>
<td>Tuberculum sella</td>
</tr>
<tr>
<td>8</td>
<td>SEF</td>
<td>Sella floor</td>
</tr>
<tr>
<td>9</td>
<td>PSL</td>
<td>Posterior sella</td>
</tr>
<tr>
<td>10</td>
<td>BAS</td>
<td>Basion</td>
</tr>
<tr>
<td>11</td>
<td>CRU</td>
<td>Cruciate eminence</td>
</tr>
</tbody>
</table>

Figure 1. Landmark Positions on a Human Skull. See Table 3 for the names and the description of the landmarks.
Using superimposition 2(b), the surgeon would plan to lower the palate. If the surgeon used Figure 2(a) or 2(c), however, he would plan his surgical procedure quite differently, moving the face in different directions and to different magnitudes. It should be clear that superimposition approach provides very different scientific inferences depending on the choice of the external constraint.

Next, we consider the form difference using the thin-plate spline methodology. One attraction of the thin-plate splines methodology is its ability to present the form change graphically. This is depicted by showing how a square grid would have changed under the estimated deformation. Figure 3(a) and (b) shows the deformation grid when the constraint is based on the minimum bending energy of the thin-plate spline and when the constraint is based on $W = E^{-1}$, where $E$ is the matrix of squared distances between all pairs of landmarks in the normal child. It is clear that the two grids are quite different, leading to different scientific conclusions about where and how much the two forms differ. For example, looking at the inverse-squared distance weighted grid, we see that the form has been markedly compressed on the posterior inferior corner. This means that the posterior aspect of the neurocranium (landmarks 10 and 11) and the distance from the back of the palate (landmark 6) to the anterior aspect of the foramen magnum (landmark 10) have been compressed. A less obvious constriction is noted on the anterosuperior corner, whereas the anteroinferior edge is stretched. Within the grid, some antero-posterior expansion or stretching is shown local to the pituitary fossa region (landmarks 7, 8, and 9), some superoin-