31. Image Registration and Unknown Coordinate Systems

This chapter deals with statistical problems involving unknown coordinate systems, either in Euclidean 3-space $\mathbb{R}^3$ or on the unit sphere Ω_3 in $\mathbb{R}^3$. We also consider the simpler cases of Euclidean 2-space $\mathbb{R}^2$ and the unit circle Ω_2. The chapter has five major sections.

Although other problems of unknown coordinate systems have arisen, a very important problem of this class is the problem of *image registration* from landmark data. In this problem we have two images of the same object (such as satellite images taken at different times) or an image of a prototypical object and an actual object. It is desired to find the rotation, translation, and possibly scale change, which will best align the two images. Whereas many problems of this type are two-dimensional, it should be noted that medical imaging is often three-dimensional.

After introducing some mathematical preliminaries we introduce the concept of M-*estimators*, a generalization of least squares estimation. In least squares estimation, the registration that minimizes the sum of squares of the lengths of the deviations is chosen; in M estimation, the sum of squares of the lengths of the deviations is replaced by some other objective function. An important case is L_1 estimation, which minimizes the sum of the lengths of the deviations; L_1 estimation is often used when the possibility of outliers in the data is suspected.

The second section of this chapter deals with the calculation of least squares estimates. Then, in the third section, we introduce an iterative modification of the least squares algorithm to calculate other M-estimates. Note that minimization usually involves some form of differentiation and hence this section starts with a short introduction to the geometry of the group of rotations and differentiation in the rotation group. Many statistical techniques are based upon approximation by derivatives and hence a little understanding of geometry is necessary to understand the later statistical sections.

The fourth section discusses the statistical properties of M-estimates. A great deal of emphasis is placed upon the relationship between the geometric configuration of the landmarks and the statistical errors in the image registration. It is shown that these statistical errors are determined, up to a constant, by the geometry of the landmarks. The constant of proportionality depends upon the objective function and the distribution of the errors in the data.

General statistical theory indicates that, if the data error distribution is (anisotropic) multivariate normal, least squares estimation is optimal. An important result of this section is that, even in this case when least squares estimation is theoretically the most efficient, the use of L_1 estimation can guard against outliers with a very modest cost in efficiency. Here optimality and efficiency refer to the expected size of the statistical errors. In practice, data is often long-tailed and L_1 estimation yields *smaller* statistical errors than least squares estimation. This will be the case with the three-dimensional image registration example given here.

Finally, in the fifth section, we discuss diagnostics that can be used to determine which data points are most influential upon the registration. Thus, if the registration is unsatisfactory, these diagnostics can be used to determine which data points are most responsible and should be reexamined.

31.1 Unknown Coordinate Systems and Their Estimation

31.1.1 Problems of Unknown Coordinate Systems

Wahba [31.1] posed the following question. Suppose we have the directions of certain stars with respect to the unknown coordinate system of a satellite. How can we estimate the orientation of the satellite? Let $\mathbf{A}$ be the unknown 3×3 matrix whose rows represent the axes of the satellite's coordinate system with respect to a fixed and known (Earth) coordinate system. Furthermore let $\boldsymbol{u}_i$ be the directions of the stars with respect to the known coordinate systems, where each $\boldsymbol{u}_i$ is written as a three-dimensional column vector with unit length. Similarly let $\boldsymbol{v}_i$ be the directions of the stars with respect to the satellite's coordinate system. Then

$$\boldsymbol{v}_i = \mathbf{A}\boldsymbol{u}_i + \text{error}\,. \tag{31.1}$$

In essence the question was to estimate $\mathbf{A}$. Wahba gave the least squares solution.

Chapman et al. [31.2] posed the same question in the following form. Suppose we have an object defined by a computer-aided design (CAD) program and a prototype is measured using a coordinate measuring machine (CMM). The orientations of lines on the object can be defined by unit vectors parallel to the lines and the orientations of planes can be defined by unit vectors normal to the planes. So we have unit vectors $\boldsymbol{u}_i$ defined by the CAD program and the corresponding unit vectors $\boldsymbol{v}_i$ as measured by the CMM. If $\mathbf{A}$ is the coordinate system of the CMM relative to the CAD program, then (31.1) holds.

Chapman et al. again used a least squares estimate $\widehat{\mathbf{A}}$ of $\mathbf{A}$. The main question of interest, that is the geometric integrity of the prototype, was then answered by analyzing the residuals of $\boldsymbol{v}_i$ from $\widehat{\mathbf{A}}\boldsymbol{u}_i$.

Since the $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are of unit length, these two problems involve spherical data.

31.1.2 Image Registration

If we enlarge the inquiry to Euclidean space data, we arrive at the widely used *image registration* problem. Suppose $\boldsymbol{u}_i \in \mathbb{R}^p$ represent the locations of some landmarks in one image, and $\boldsymbol{v}_i \in \mathbb{R}^p$ the locations of corresponding landmarks in a second image of the same

object. The usual applications occur with $p = 2.3$. Under certain conditions, it might be reasonable to suppose that

$$\boldsymbol{v}_i = \mathbf{B}\boldsymbol{u}_i + \boldsymbol{b} + \text{error} \tag{31.2}$$

for an unknown $p \times p$ matrix $\mathbf{B}$ and an unknown p-dimensional column vector $\boldsymbol{b}$. The matrix $\mathbf{B}$ represents a coordinate change and the vector $\boldsymbol{b}$ represents a translation of coordinates. The image registration problem is to estimate $\mathbf{B}$ and $\boldsymbol{b}$.

The model (31.2) also arises in a slightly different context. Suppose we have landmarks $\boldsymbol{u}_i$ on a prototypical face. For example the $\boldsymbol{u}_i$ might represent the locations of the nose, the two eyes, the base of the chin, etc. For the purpose of automated processing of a large number of facial images of different subjects, we might want to bring each facial image into alignment with the prototypical image using a transformation of the form (31.2) where the $\boldsymbol{v}_i$ represent the same locations (nose, two eyes, base of chin, etc.) on the subject facial image.

In the absence of measurement error, one does not expect the landmarks on two faces to be related using a transformation of the form

$$\boldsymbol{v}_i = \mathbf{B}\boldsymbol{u}_i + \boldsymbol{b} \,. \tag{31.3}$$

The reader might be puzzled why a transformation of this form is under consideration. Statistical error, however, is not limited to measurement error. Statistical error incorporates all effects not included in the systematic portion of the model. In building a model of the form (31.2), we hope to separate out the most important relationship (31.3) between the landmarks $\boldsymbol{u}_i$ on one object and the corresponding landmarks $\boldsymbol{v}_i$ on the other object; the rest is placed in the statistical error.

Unlike the Wahba problem, the unknown $(\mathbf{B}, \boldsymbol{b})$ of the image registration problem, or the unknown $\mathbf{A}$ in the *Chapman* et al. problem, are not of primary interest. Rather, they must be estimated as a preliminary step to more interesting problems. We will discuss herein the properties of various methods of estimating these unknowns. These properties will hopefully help the interested reader to choose a good estimation technique which will hopefully yield better results after this preliminary step is completed.

31.1.3 The Orthogonal and Special Orthogonal Matrices

Consider, for example the data set in Table 31.1 from *Chang* and *Ko* [31.3], which we will analyze repeatedly in what follows. This data consists of the digitized locations of 12 pairs of landmarks on the left and right hands of one of the authors. This is a $p = 3$ three-dimensional image registration problem. We might decide that, apart from the statistical error term, the shape of the two hands is the same; that is the distance between two points on one hand is the same as the distance between the corresponding two points on the other hand.

This condition translates mathematically to the equation $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_p$, the $p \times p$-dimensional identity matrix. We outline a derivation of this well-known mathematical fact for the primary purpose of introducing the reader to the mathematical style of the remainder of this chapter. The distance between two p-dimensional column vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is

$$||\boldsymbol{v}_2 - \boldsymbol{v}_1|| = \sqrt{(\boldsymbol{v}_2 - \boldsymbol{v}_1)^{\mathrm{T}}(\boldsymbol{v}_2 - \boldsymbol{v}_1)} \,, \tag{31.4}$$

where the operations on the right-hand side of (31.4) are matrix multiplication and transposition. If the $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$ are related by (31.3),

$$\begin{aligned}\left(\boldsymbol{v}_j - \boldsymbol{v}_i\right)^{\mathrm{T}} \left(\boldsymbol{v}_j - \boldsymbol{v}_i\right) &= \left[\mathbf{B}(\boldsymbol{u}_j - \boldsymbol{u}_i)\right]^{\mathrm{T}} \times \left[\mathbf{B}(\boldsymbol{u}_j - \boldsymbol{u}_i)\right] \\ &= \left[(\boldsymbol{u}_j - \boldsymbol{u}_i)\right] \mathbf{B}^{\mathrm{T}}\mathbf{B} \left[(\boldsymbol{u}_j - \boldsymbol{u}_i)\right] \,.\end{aligned}$$

Thus if $||\boldsymbol{v}_j - \boldsymbol{v}_i|| = ||\boldsymbol{u}_j - \boldsymbol{u}_i||$ for all i and j, and if the $\boldsymbol{u}_i$ do not all lie in a $(p-1)$-dimensional hyperplane of $\mathbb{R}^p$,

$$\mathbf{I}_p = \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{B}\mathbf{B}^{\mathrm{T}} \,. \tag{31.5}$$

Table 31.1 12 digitized locations on the left and right hand

	Left hand $\boldsymbol{u}_i$			Right hand $\boldsymbol{v}_i$		
A	5.17	11.30	16.18	5.91	11.16	16.55
B	7.40	12.36	17.50	8.63	10.62	18.33
C	8.56	12.59	17.87	10.09	10.60	18.64
D	9.75	13.62	17.01	10.89	10.95	17.90
E	11.46	14.55	12.96	12.97	10.13	13.88
F	7.10	13.12	12.56	8.79	11.21	13.17
G	8.85	13.82	12.60	10.70	11.10	13.42
H	6.77	13.07	10.32	8.47	11.09	11.35
I	6.26	11.62	13.34	7.28	12.52	14.04
J	6.83	12.00	13.83	8.05	12.42	14.56
K	7.94	12.29	13.84	9.07	12.39	14.86
L	8.68	12.71	13.67	10.15	12.17	14.44

A: Top of little finger; B: Top of ring finger; C: Top of middle finger; D: Top of forefinger; E: Top of thumb; F: Gap between thumb and forefinger; G: Center of palm; H: Base of palm; I: Little finger knuckle; J: Ring finger knuckle; K: Middle finger knuckle; L: Forefinger knuckle

Note that the first equality of (31.5) implies that $\mathbf{B}^{-1} = \mathbf{B}^{\mathrm{T}}$ and hence the second equality follows. Matrices which satisfy condition (31.5) are said to be *orthogonal*.

On the other hand, we might want to hypothesize that the two hands (again apart from statistical error) have the same shape except that one hand might be larger than the other. In this case we are hypothesizing

$$\mathbf{B} = \gamma \mathbf{A} \tag{31.6}$$

where $\mathbf{A}$ is orthogonal and γ is a positive real number.

In the Wahba and Chapman et al. problems, the rows of $\mathbf{A}$ are known to be an orthonormal basis of $\mathbb{R}^3$. Since the (i, j) entry of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is the dot product of the i-th and j-th rows of $\mathbf{A}$, it follows that $\mathbf{A}$ is orthogonal. However, more is known. Since the unknown coordinate system is known to be right-handed,

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_p, \quad \det(\mathbf{A}) = 1\,, \tag{31.7}$$

where $\det(\mathbf{A})$ is the determinant of the matrix $\mathbf{A}$. Such matrices are said to be *special orthogonal*.

In the hands data of Table 31.1, if we use the model (31.2) with condition (31.6), then $\mathbf{A}$ will not be special orthogonal. This is because the left and right hands have different orientations. However, it is common in image registration problems to assume that condition (31.6) is true with $\mathbf{A}$ assumed to be special orthogonal.

Following standard mathematical notation, we will use $\mathcal{O}(p)$ to denote the $p \times p$ orthogonal matrices [that is the set of all matrices which satisfy (31.5) and $\mathcal{SO}(p)$ to denote the subset of $\mathcal{O}(p)$ of special orthogonal matrices [that is the set of all matrices which satisfy (31.7)].

31.1.4 The Procrustes and Spherical Regression Models

In this chapter, we will be concerned with statistical methods which apply to the model (31.2) for Euclidean space data $\boldsymbol{u}_i, \boldsymbol{v}_i \in \mathbb{R}^p$, for arbitrary p, where $\mathbf{B}$ satisfies the condition (31.6) with $\mathbf{A}$ constrained to be either orthogonal or special orthogonal. Following Goodall [31.4], we will call this model the *Procrustes* model.

We will also consider models of the form (31.1), where the p-vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are constrained to be of unit length, that is

$$\boldsymbol{u}_i, \boldsymbol{v}_i \in \Omega_p = S^{p-1} = \left(\boldsymbol{x} \in \mathbb{R}^p \quad \middle| \quad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1\right)$$

and $\mathbf{A}$ is constrained to be either orthogonal or special orthogonal. Following *Chang* [31.5], we will call this model the *spherical regression* model.

The statistical methodology for these two models can easily be described in parallel. In general, we will focus on the Procrustes model, while giving the modifications that apply to the spherical regression model.

31.1.5 Least Squares, L_1, and M Estimation

In Sect. 31.2, we will derive the least squares estimate of $\mathbf{A}$, γ, $\boldsymbol{b}$ for the Procrustes model. This estimate minimizes

$$\rho_2(\mathbf{A}, \gamma, \boldsymbol{b}) = \sum_i ||\boldsymbol{v}_i - \gamma \mathbf{A}\boldsymbol{u}_i - \boldsymbol{b}||^2 \tag{31.8}$$

over all $\mathbf{A}$ in either $\mathcal{O}(p)$ or $\mathcal{SO}(p)$, constants $\gamma > 0$, and p-vectors $\boldsymbol{b} \in \mathbb{R}^p$. For the spherical regression model, the least squares estimate minimizes

$$\rho_2(\mathbf{A}) = \sum_i ||\boldsymbol{v}_i - \mathbf{A}\boldsymbol{u}_i||^2 \tag{31.9}$$

$$= 2n - 2\sum_i \boldsymbol{v}_i^{\mathrm{T}}\mathbf{A}\boldsymbol{u}_i \tag{31.10}$$

over all $\mathbf{A}$ in either $\mathcal{O}(p)$ or $\mathcal{SO}(p)$. For the second equality in (31.9), we have used that if $1 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}$, then

$$\begin{aligned} ||\boldsymbol{v} - \mathbf{A}\boldsymbol{u}||^2 &= (\boldsymbol{v} - \mathbf{A}\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{v} - \mathbf{A}\boldsymbol{u}) \\ &= \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} - \boldsymbol{v}^{\mathrm{T}}\mathbf{A}\boldsymbol{u} - (\mathbf{A}\boldsymbol{u})^{\mathrm{T}}\boldsymbol{v} + \boldsymbol{u}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{u} \\ &= 2 - 2\boldsymbol{v}^{\mathrm{T}}\mathbf{A}\boldsymbol{u}\,. \end{aligned}$$

Least squares estimates have the advantage that an explicit closed-form solution for them is available. They have the disadvantage that they are very sensitive to *outliers*, that is points $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ for which the error term in (31.2) is unusually large. In the image registration problem, an outlier can arise in several contexts. It can be the result of a measurement error, or it can be the result of a misidentified landmark. Perhaps the image is not very clear, or the landmark (e.g. 'point of the nose') cannot be very precisely determined, or the landmark is obscured (by clouds or shrubs, etc.). Or perhaps there are places in the image where the image is not really rigid, that is the ideal match (31.3) does not apply very well. It is easy to conceive of a myriad of situations which might give rise to outliers.

L_1 estimators are often used to ameliorate the effects of outliers. These estimators minimize

$$\rho_1(\mathbf{A}, \gamma, \boldsymbol{b}) = \sum_i ||\boldsymbol{v}_i - \gamma \mathbf{A}\boldsymbol{u}_i - \boldsymbol{b}||\,, \tag{31.11}$$

for the Procrustes model, or the sum of the distances along the surface of the sphere

$$\rho_1(\mathbf{A}) = \sum_i \arccos(\boldsymbol{v}_i^{\mathrm{T}} \mathbf{A} \boldsymbol{u}_i) \tag{31.12}$$

for the spherical regression model. Unfortunately, an explicit closed-form solution for the L_1 estimate is not available and it must be calculated by numerical minimization. We will offer a few suggestions on approaches for numerical minimization in Sect. 31.3.5.

The least squares and L_1 estimators are special cases of the so-called *M estimators*. These estimators minimize an objective function of the form

$$\rho(\mathbf{A}, \gamma, \boldsymbol{b}) = \sum_i \rho_0(s_i) \,, \tag{31.13}$$

where

$$s_i = ||\boldsymbol{v}_i - \gamma \mathbf{A} \boldsymbol{u}_i - \boldsymbol{b}||$$

and ρ_0 is some increasing function. Intermediate between the least squares and L_1 estimate is the *Huber* estimate for which

$$\rho_0(s) = \begin{cases} (s/b)^2 & s < b \\ s/b & s \geq b \end{cases}$$

for some preset constant b. Or we can *Windsorize* the estimate

$$\rho_0(s) = \begin{cases} (s/b)^2 & s < b \\ 1 & s \geq b \end{cases} .$$

In the linear regression context, these and other objective functions are discussed in *Huber* [31.6].

For the spherical regression model, an M-estimator minimizes an objective function of the form

$$\rho(\mathbf{A}) = \sum_i \rho_0(t_i) \,, \tag{31.14}$$

where

$$t_i = \boldsymbol{v}_i^{\mathrm{T}} \mathbf{A} \boldsymbol{u}_i \,.$$

Notice that, as $\boldsymbol{v}$ moves away from $\mathbf{A}\boldsymbol{u}$ towards the antipodal point $-\mathbf{A}\boldsymbol{u}$, $t = \boldsymbol{v}^{\mathrm{T}} \mathbf{A} \boldsymbol{u}$ *decreases* from 1 to -1. Thus, for the spherical case, $\rho_0(t)$ is chosen to be a decreasing function of t.

In Sect. 31.4 we will discuss the statistical properties of M-estimates. We will see how the geometry of the data translates into the error structure of the estimate. In the image registration problem, this information can be used, for example, to help select landmarks. General statistical theory indicates under certain conditions ("normal distribution") the least squares solution is optimal. However, if we were to use a L_1 estimate to guard against outliers, we would suffer a penalty of 13% for image registrations in two dimensions and only 8% for image registrations in three dimensions, even when least squares is theoretically optimal. We will make more precise in Sect. 31.4 how this *penalty* is defined. The important point to realize is that, especially for three-dimensional image registrations, L_1 estimators offer important protections against outliers in the data at very modest cost in the statistical efficiency of the estimator.

In Sect. 31.5, we will discuss diagnostics for the Procrustes and spherical regression models. If the image registration is not satisfactory, this section will give tools to determine which of the landmarks is causing the unsatisfactory registration. It will follow, for example, that landmarks which greatly influence $\mathbf{A}$ will have negligible influence on γ and vice versa.

31.2 Least Squares Estimation

31.2.1 Group Properties of $\mathcal{O}(p)$ and $\mathcal{SO}(p)$

It is important to note that $\mathcal{O}(p)$ and $\mathcal{SO}(p)$ are groups in the mathematical sense. That is, if $\mathbf{A}, \mathbf{B} \in \mathcal{O}(p)$, then

$$(\mathbf{AB})^{\mathrm{T}}(\mathbf{AB}) = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AB} = \mathbf{B}^{\mathrm{T}}\mathbf{I}_p\mathbf{B} = \mathbf{I}_p$$

since both $\mathbf{A}$ and $\mathbf{B}$ satisfy (31.5). Thus $\mathbf{AB} \in \mathcal{O}(p)$. Similarly if $\mathbf{A} \in \mathcal{O}(p)$, then (31.5) implies that $\mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}} \in \mathcal{O}(p)$. This implies that $\mathcal{O}(p)$ is a group. Furthermore, if $\det(\mathbf{A}) = \det(\mathbf{B}) = 1$, then $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B}) = 1$ and $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A}) = 1$. In summary we have

$$\begin{aligned} &\text{If } \mathbf{A}, \mathbf{B} \in \mathcal{O}(p), \text{ then } \mathbf{AB} \in \mathcal{O}(p) \\ &\text{and } \mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}} \in \mathcal{O}(p) \\ &\text{If } \mathbf{A}, \mathbf{B} \in \mathcal{SO}(p), \text{ then } \mathbf{AB} \in \mathcal{SO}(p) \\ &\text{and } \mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}} \in \mathcal{SO}(p) \,. \end{aligned} \tag{31.15}$$

Notice also that, if $\mathbf{A}$ satisfies (31.5), then

$$1 = \det(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = [\det(\mathbf{A})]^2$$

so that $\det(\mathbf{A}) = 1, -1$.

31.2.2 Singular Value Decomposition

Given a $p \times q$ matrix $\mathbf{X}$ its *singular value decomposition* is

$$\mathbf{X} = \mathbf{O}_1 \Lambda \mathbf{O}_2^{\mathrm{T}} , \tag{31.16}$$

where $\mathbf{O}_1 \in \mathcal{O}(p)$, $\mathbf{O}_2 \in \mathcal{O}(q)$ and $\mathbf{\Lambda}$ is $p \times q$. If $p \le q$, $\mathbf{\Lambda}$ has block form

$$\mathbf{\Lambda} = \left[\operatorname{diag}(\lambda_1, \cdots, \lambda_p) \;\; \mathbf{0}_{(p,q-p)} \right]$$

Here $\operatorname{diag}(\lambda_1, \cdots, \lambda_p)$ is a diagonal matrix with entries $\lambda_1 \ge \cdots \ge \lambda_p$ and $\mathbf{0}_{(p,q-p)}$ is a $p \times (q-p)$ matrix with all zeros. If $q \le p$

$$\mathbf{\Lambda} = \begin{pmatrix} \operatorname{diag}(\lambda_1, \cdots, \lambda_q) \\ \mathbf{0}_{(p-q,q)} \end{pmatrix} .$$

Most mathematical software packages now include the singular value decomposition. However, it can be computed using a package which only computes eigen-decompositions of symmetric matrices. Suppose temporarily $p \le q$. Since $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is a symmetric nonnegative definite matrix, its eigen-decomposition has the form

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{O}_1 \mathbf{\Lambda}_1 \mathbf{O}_1^{\mathrm{T}} ,$$

where $\mathbf{O}_1 \in \mathcal{O}(p)$ and $\mathbf{\Lambda}_1 = \operatorname{diag}(\lambda_1^2, \cdots, \lambda_p^2)$ with $\lambda_1 \ge \cdots \ge \lambda_p \ge 0$. The columns of $\mathbf{O}_1$ are the eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\lambda_1^2, \cdots, \lambda_p^2$ are the corresponding eigenvalues. Suppose $\lambda_p > 0$ and let $\widetilde{\mathbf{O}}_2 = \mathbf{X}^{\mathrm{T}} \mathbf{O}_1 \mathbf{\Lambda}_1^{-1/2}$. $\widetilde{\mathbf{O}}_2$ is $q \times p$, but

$$\begin{aligned} \widetilde{\mathbf{O}}_2^{\mathrm{T}} \widetilde{\mathbf{O}}_2 &= \mathbf{\Lambda}_1^{-1/2} \mathbf{O}_1^{\mathrm{T}} \mathbf{X}\mathbf{X}^{\mathrm{T}} \mathbf{O}_1 \mathbf{\Lambda}_1^{-1/2} \\ &= \mathbf{\Lambda}_1^{-1/2} \mathbf{O}_1^{\mathrm{T}} \mathbf{O}_1 \mathbf{\Lambda}_1 \mathbf{O}_1^{\mathrm{T}} \mathbf{O}_1 \mathbf{\Lambda}_1^{-1/2} \\ &= \mathbf{\Lambda}_1^{-1/2} \mathbf{\Lambda}_1 \mathbf{\Lambda}_1^{-1/2} = \mathbf{I}_p , \end{aligned}$$

so that the columns of $\widetilde{\mathbf{O}}_2$ are orthonormal. Furthermore

$$\mathbf{O}_1 \mathbf{\Lambda}_1^{1/2} \widetilde{\mathbf{O}}_2^{\mathrm{T}} = \mathbf{O}_1 \mathbf{\Lambda}_1^{1/2} \mathbf{\Lambda}_1^{-1/2} \mathbf{O}_1^{\mathrm{T}} \mathbf{X} = \mathbf{X} .$$

Filling $\mathbf{\Lambda}_1^{1/2}$ with $q - p$ columns of zeros, and completing the columns of $\widetilde{\mathbf{O}}_2$ to an orthonormal basis of $\mathbb{R}^q$ yields the decomposition (31.16).

Extensions to the cases when $\lambda_p = 0$ or when $q \le p$ will not be difficult for the careful reader.

31.2.3 Least Squares Estimation in the Procrustes Model

The least squares estimation of the Procrustes model (31.2) has long been known (see, for example, Goodall [31.4]). Let $\overline{\boldsymbol{u}} = n^{-1} \sum_i \boldsymbol{u}_i$, where n is the number of pairs $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ and let $\overline{\boldsymbol{v}}$ be similarly defined. Define the $p \times p$ matrix $\mathbf{X}$ by

$$\mathbf{X} = \sum_i (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{v}_i - \overline{\boldsymbol{v}})^{\mathrm{T}} .$$

Then

$$\begin{aligned} \rho_2(\mathbf{A}, \gamma, \boldsymbol{b}) &= \sum_i ||\boldsymbol{v}_i - \gamma \mathbf{A}\boldsymbol{u}_i - \boldsymbol{b}||^2 \\ &= \sum_i ||\boldsymbol{v}_i - \overline{\boldsymbol{v}} - \gamma \mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) \\ &\quad - [\boldsymbol{b} - (\overline{\boldsymbol{v}} - \gamma \mathbf{A}\overline{\boldsymbol{u}})]\,||^2 \\ &= \sum_i ||\boldsymbol{v}_i - \overline{\boldsymbol{v}}||^2 \\ &\quad - \gamma \sum_i (\boldsymbol{u}_i - \overline{\boldsymbol{u}})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} (\boldsymbol{v}_i - \overline{\boldsymbol{v}}) \\ &\quad - \gamma \sum_i (\boldsymbol{v}_i - \overline{\boldsymbol{v}})^{\mathrm{T}} \mathbf{A} (\boldsymbol{u}_i - \overline{\boldsymbol{u}}) \\ &\quad + \gamma^2 \sum_i ||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||^2 \\ &\quad + n||\boldsymbol{b} - (\overline{\boldsymbol{v}} - \gamma \mathbf{A}\overline{\boldsymbol{u}})||^2 . \end{aligned}$$

All the other cross-product terms sum to zero. Now

$$\begin{aligned} &\sum_i (\boldsymbol{v}_i - \overline{\boldsymbol{v}})^{\mathrm{T}} \mathbf{A} (\boldsymbol{u}_i - \overline{\boldsymbol{u}}) \\ &= \sum_i \operatorname{Tr}\left[(\boldsymbol{v}_i - \overline{\boldsymbol{v}})^{\mathrm{T}} \mathbf{A} (\boldsymbol{u}_i - \overline{\boldsymbol{u}}) \right] \\ &= \sum_i \operatorname{Tr}\left[\mathbf{A} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{v}_i - \overline{\boldsymbol{v}})^{\mathrm{T}} \right] = \operatorname{Tr}(\mathbf{A}\mathbf{X}) \end{aligned}$$

and

$$\begin{aligned} &\sum_i (\boldsymbol{u}_i - \overline{\boldsymbol{u}})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} (\boldsymbol{v}_i - \overline{\boldsymbol{v}}) \\ &= \sum_i (\boldsymbol{v}_i - \overline{\boldsymbol{v}})^{\mathrm{T}} \mathbf{A} (\boldsymbol{u}_i - \overline{\boldsymbol{u}}) = \operatorname{Tr}(\mathbf{A}\mathbf{X}) . \end{aligned}$$

Therefore

$$\begin{aligned} \rho_2(\mathbf{A}, \gamma, \boldsymbol{b}) &= \sum_i ||\boldsymbol{v}_i - \overline{\boldsymbol{v}}||^2 - 2\gamma \operatorname{Tr}(\mathbf{A}\mathbf{X}) \\ &\quad + \gamma^2 \sum_i ||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||^2 \\ &\quad + n||\boldsymbol{b} - (\overline{\boldsymbol{v}} - \gamma \mathbf{A}\overline{\boldsymbol{u}})||^2 . \end{aligned} \tag{31.17}$$

Substituting (31.16),

$$\begin{aligned} \operatorname{Tr}(\mathbf{A}\mathbf{X}) &= \operatorname{Tr}\left(\mathbf{A}\mathbf{O}_1 \mathbf{\Lambda} \mathbf{O}_2^{\mathrm{T}} \right) = \operatorname{Tr}\left(\mathbf{O}_2^{\mathrm{T}} \mathbf{A}\mathbf{O}_1 \mathbf{\Lambda} \right) \\ &= \sum_i \lambda_i e_{ii} , \end{aligned}$$

where e_{ii} are the diagonal entries of $\mathbf{O}_2^{\mathrm{T}}\mathbf{A}\mathbf{O}_1 \in \mathcal{O}(p)$. Now $|e_{ii}| \le 1$ and hence $\mathrm{Tr}\,(\mathbf{A}\mathbf{X})$ is maximized when $e_{ii} = 1$ or, equivalently, when $\mathbf{O}_2^{\mathrm{T}}\mathbf{A}\mathbf{O}_1 = \mathbf{I}_p$. This implies $\mathbf{A} = \mathbf{O}_2\mathbf{O}_1^{\mathrm{T}}$.

Thus if $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right)$ minimizes (31.17),

$$\begin{aligned}
\widehat{\mathbf{A}} &= \mathbf{O}_2\mathbf{O}_1^{\mathrm{T}}\,, \\
\widehat{\gamma} &= \left(\sum_i ||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||^2\right)^{-1} \mathrm{Tr}\left(\widehat{\mathbf{A}}\mathbf{X}\right) \\
&= \left(\sum_i ||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||^2\right)^{-1} \sum_i \lambda_i\,, \qquad (31.18) \\
\widehat{\boldsymbol{b}} &= \overline{\boldsymbol{v}} - \widehat{\gamma}\widehat{\mathbf{A}}\overline{\boldsymbol{u}}\,.
\end{aligned}$$

If $\mathbf{A}$ is constrained to lie in $\mathcal{SO}(p)$, we use a modified singular value decomposition. Let $\mathbf{X} = \widetilde{\mathbf{O}}_1\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{O}}_2^{\mathrm{T}}$ be the (usual) singular value decomposition of $\mathbf{X}$ and let

$$\mathbf{E} = \mathrm{diag}(1, \cdots, 1, -1) \qquad (31.19)$$

be the identity matrix with its last entry changed to -1. Let $\mathbf{O}_1 = \widetilde{\mathbf{O}}_1\mathbf{E}^{\delta_1}$ where $\delta_1 = 0$ if $\widetilde{\mathbf{O}}_1 \in \mathcal{SO}(p)$ and $\delta_1 = 1$ otherwise.

Similarly define δ_2 and $\mathbf{O}_2$. Finally write $\mathbf{\Lambda} = \widetilde{\mathbf{\Lambda}}\mathbf{E}^{\delta_1+\delta_2}$. Then (31.16) is valid with $\mathbf{O}_1, \mathbf{O}_2 \in \mathcal{SO}(p)$ and $\lambda_1 \ge \cdots \ge \lambda_{p-1} \ge |\lambda_p|$.

This is the modified singular value decomposition.

The least squares estimates, subject to the contraint $\widehat{\mathbf{A}} \in \mathcal{SO}(p)$, is still given by (31.18) when a modified singular value decomposition is used for $\mathbf{X}$.

31.2.4 Example: Least Squares Estimates for the Hands Data

Consider, for example, the hands data in Table 31.1. For this data

$$\begin{aligned}
\overline{\boldsymbol{u}} &= \begin{pmatrix} 7.8975 \\ 12.7542 \\ 14.3067 \end{pmatrix}\,, \quad \overline{\boldsymbol{v}} = \begin{pmatrix} 9.2500 \\ 11.3633 \\ 15.0950 \end{pmatrix}\,, \\
\mathbf{X} &= \left[\begin{pmatrix} 5.17 \\ 11.30 \\ 16.18 \end{pmatrix} - \overline{\boldsymbol{u}}\right]\left[\begin{pmatrix} 5.91 \\ 11.16 \\ 16.55 \end{pmatrix} - \overline{\boldsymbol{v}}\right]^{\mathrm{T}} + \\
&\quad \cdots + \left[\begin{pmatrix} 8.68 \\ 12.71 \\ 13.67 \end{pmatrix} - \overline{\boldsymbol{u}}\right]\left[\begin{pmatrix} 10.15 \\ 12.17 \\ 14.44 \end{pmatrix} - \overline{\boldsymbol{v}}\right]^{\mathrm{T}} \\
&= \begin{pmatrix} 34.0963 & -6.9083 & 3.5769 \\ 17.3778 & -4.9028 & -5.6605 \\ -2.3940 & -5.7387 & 57.8598 \end{pmatrix}\,.
\end{aligned}$$

The singular value decomposition $\mathbf{X} = \mathbf{O}_1\mathbf{\Lambda}\mathbf{O}_2^{\mathrm{T}}$ is given by

$$\mathbf{O}_1 = \begin{pmatrix} 0.0465 & -0.8896 & -0.4544 \\ -0.1012 & -0.4567 & 0.8838 \\ 0.9938 & -0.0048 & 0.1112 \end{pmatrix}\,,$$

$$\mathbf{O}_2 = \begin{pmatrix} -0.0436 & -0.9764 & -0.2114 \\ -0.0944 & 0.2147 & -0.9721 \\ 0.9946 & -0.0224 & -0.1015 \end{pmatrix}$$

$$\mathbf{\Lambda} = \mathrm{diag}(58.5564, 39.1810, 1.8855)\,.$$

Hence (31.18) yields

$$\widehat{\mathbf{A}} = \begin{pmatrix} 0.9627 & 0.2635 & -0.0621 \\ 0.2463 & -0.9477 & -0.2030 \\ 0.1123 & -0.1801 & 0.9772 \end{pmatrix}\,,$$

$$\widehat{\gamma} = 0.9925\,, \qquad (31.20)$$

$$\widehat{\boldsymbol{b}} = \begin{pmatrix} -0.7488 & 24.3115 & 2.6196 \end{pmatrix}^{\mathrm{T}}\,.$$

Notice that $\det(\widehat{\mathbf{A}}) = -1$ so $\widehat{\mathbf{A}} \notin \mathcal{SO}(3)$. We expect this result since, as previously remarked, the left and right hands have different orientations. The value of $\widehat{\gamma}$ is somewhat puzzling since the subject is right-handed and one would expect, therefore, $\gamma > 1$. Although, as we will see in Sect. 31.4, the difference between $\widehat{\gamma}$ and 1 is not significant, a better estimate would have been achieved if the L_1 objective function (31.11) were numerically minimized instead. In this case $\widehat{\gamma} = 1.0086$. Our analysis will show that the hands data set has an outlier and we see here an example of the superior resistance of L_1 estimates to outliers.

31.2.5 Least Squares Estimation in the Spherical Regression Model

Least squares estimation for the spherical regression model is similar to least squares estimation in the Procrustes model. Let $\mathbf{X} = \sum_i \boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}}$ and define $\mathbf{O}_1, \mathbf{O}_2 \in \mathcal{O}(p)$ using a singular value decomposition of $\mathbf{X}$. Then $\widehat{\mathbf{A}} = \mathbf{O}_2\mathbf{O}_1^{\mathrm{T}}$. If, on the other hand, it is desired to constrain $\widehat{\mathbf{A}}$ to $\mathcal{SO}(p)$, one defines $\mathbf{O}_1, \mathbf{O}_2 \in \mathcal{SO}(p)$ using a modified singular value decomposition and, again, $\widehat{\mathbf{A}} = \mathbf{O}_2\mathbf{O}_1^{\mathrm{T}}$.

31.3 Geometry of $\mathcal{O}(p)$ and $\mathcal{SO}(p)$

$\mathcal{O}(p)$ and $\mathcal{SO}(p)$ arise because they give distance-preserving transformations of $\mathbb{R}^p$, and to formulate properly the statistical properties of $\widehat{\mathbf{A}}$ defined by (31.18), it is important to understand the geometry of these two groups.

31.3.1 $\mathcal{SO}(p)$ for $p=2$

For $p=2$,

$$\mathcal{SO}(2)=\left\{\Phi_2(h)=\begin{pmatrix}\cos(h) & -\sin(h)\\ \sin(h) & \cos(h)\end{pmatrix}\middle| h\in\mathbb{R}^1\right\}. \tag{31.21}$$

Physically $\Phi_2(h)$ represents a rotation of $\mathbb{R}^2$ by an angle of h radians. Since $\Phi_2(h)=\Phi_2(h+2\pi)$, $\mathcal{SO}(2)$ is geometrically a circle.

Since each element of $\mathcal{SO}(2)$ has four entries, it is tempting to think of $\mathcal{SO}(2)$ as four-dimensional. However as (31.21) makes clear, $\mathcal{SO}(2)$ can be described by one parameter $h\in\mathbb{R}^1$. Thus $\mathcal{SO}(2)$ is really one-dimensional. Suppose we were constrained to live on a circle Ω_2 (instead of the sphere Ω_3). At each point on Ω_2 we can only travel to our left or to our right, and, if our travels were limited, it would appear as if we only had one-dimensional travel. Mathematicians describe this situation by saying that $\mathcal{SO}(2)$ is a one-dimensional *manifold*.

Notice also $\Phi_2(0)=\mathbf{I}_2$ and that, if h is small, then $\Phi_2(h)$ is close to $\mathbf{I}_2$. Thus, if h is small, $\Phi_2(h)\boldsymbol{x}$ is close to $\boldsymbol{x}$ for all $\boldsymbol{x}\in\mathbb{R}^2$. As we shall see, this simple observation is key to understanding our approach to the statistical properties of $\widehat{\mathbf{A}}$.

31.3.2 $\mathcal{SO}(p)$ for $p=3$

$\mathcal{SO}(3)$ can be described as the collection of all rotations in $\mathbb{R}^3$. That is

$$\mathcal{SO}(3)=\left\{\Phi_3(\boldsymbol{h})\mid \boldsymbol{h}\in\mathbb{R}^3\right\}, \tag{31.22}$$

where $\Phi_3(\boldsymbol{h})$ is right-hand rule rotation of $||\boldsymbol{h}||$ radians around the axis $||\boldsymbol{h}||^{-1}\boldsymbol{h}$. Writing $\theta=||\boldsymbol{h}||$ and $\xi=||\boldsymbol{h}||^{-1}\boldsymbol{h}$, so that ξ is a unit-length three-dimensional vector, it can be shown that

$$\begin{aligned}&\Phi_3(\boldsymbol{h})\\ &=\Phi_3(\theta\xi)\\ &=\cos(\theta)\mathbf{I}_3+\sin(\theta)M_3(\xi)+[1-\cos(\theta)]\,\xi\xi^{\mathrm{T}},\end{aligned} \tag{31.23}$$

where

$$M_3(\xi)=M_3\begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\end{pmatrix}=\begin{pmatrix}0 & -\xi_3 & \xi_2\\ \xi_3 & 0 & -\xi_1\\ -\xi_2 & \xi_1 & 0\end{pmatrix}.$$

Thus, although each $\mathbf{A}\in\mathcal{SO}(3)$ has nine entries, $\mathcal{SO}(3)$ is actually a three-dimensional manifold.

Again we notice that $\Phi_3(\mathbf{0})=\mathbf{I}_3$ and that if $||\boldsymbol{h}||$ is small then $\Phi_3(\boldsymbol{h})\boldsymbol{x}$ is close to $\boldsymbol{x}$ for all $\boldsymbol{x}\in\mathbb{R}^3$.

For future use, we note that if $\mathbf{C}\in\mathcal{SO}(3)$, then the axis ξ of the rotation represented by $\mathbf{C}$ satisfies $\mathbf{C}\xi=\xi$. Thus ξ is the eigenvector associated to the eigenvalue 1 of $\mathbf{C}$. By re-representing $\mathbf{C}$ in an orthonormal basis which includes ξ, one can show that the angle of rotation θ of the rotation represented by $\mathbf{C}$ satisfies $1+2\cos(\theta)=\mathrm{Tr}(\mathbf{C})$. Thus, if ξ and θ are calculated in this way, $\Phi_3(\theta\xi)=\mathbf{C}$.

31.3.3 $\mathcal{SO}(p)$ and $\mathcal{O}(p)$, for General p, and the Matrix Exponential Map

For general p, let $\mathbf{H}$ be a $p\times p$ *skew-symmetric* matrix; that is

$$\mathbf{H}^{\mathrm{T}}=-\mathbf{H}.$$

We define the matrix exponential map by

$$\exp(\mathbf{H})=\sum_{k=0}^{k=\infty}\frac{\mathbf{H}^k}{k!}.$$

It can be shown that the skew-symmetry condition implies that $\exp(\mathbf{H})\left[\exp(\mathbf{H})\right]^{\mathrm{T}}=\mathbf{I}_p$ and indeed

$$\mathcal{SO}(p)=\left[\exp(\mathbf{H})\mid \mathbf{H}\text{ is skew-symmetric}\right]. \tag{31.24}$$

A skew-symmetric matrix must have zeros on its main diagonal and its entries below the main diagonal are determined by its entries above the main diagonal. Thus the skew-symmetric $p\times p$ matrices have $p(p-1)/2$ independent entries and hence $\mathcal{SO}(p)$ is a manifold with dimension $p(p-1)/2$.

Let $\mathbf{0}_{(p,p)}$ be a $p\times p$ matrix of zeros. Then

$$\exp(\mathbf{0}_{(p,p)})=\mathbf{I}_p. \tag{31.25}$$

Thus, if the entries of $\mathbf{H}$ are small (in absolute value), then $\exp(\mathbf{H})$ will be close to the identity matrix.

For $p=3$, it can be shown, by using (31.23), that $\Phi_3(\boldsymbol{h})=\exp\left[M_3(\boldsymbol{h})\right]$ for $\boldsymbol{h}\in\mathbb{R}^3$. Similarly we define for

$h \in \mathbb{R}^1$ the skew-symmetric matrix

$$M_2(h) = \begin{pmatrix} 0 & -h \\ h & 0 \end{pmatrix}$$

and it follows that $\Phi_2(h) = \exp[M_2(h)]$. Thus (31.21) and (31.22) are indeed special cases of (31.24).

$\mathcal{O}(p)$ has two connected components; one is $\mathcal{SO}(p)$ and the other is

$$\mathcal{SO}(p)\boldsymbol{E} = \{\mathbf{A}\boldsymbol{E} \mid \mathbf{A} \in \mathcal{SO}(p)\},$$

where $\boldsymbol{E}$ has been previously defined in (31.19). Notice that $\boldsymbol{E}$ is a reflection of $\mathbb{R}^p$ through the $(p-1)$-dimensional hyperplane perpendicular to the last coordinate vector. Indeed all reflections of $\mathbb{R}^p$ are in $\mathcal{O}(p)$.

31.3.4 Geometry and the Distribution of *M*-Estimates

So, heuristically speaking, suppose we have estimates $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right)$ which minimize an objective function of the form (31.13). What values of the unknown parameters $(\mathbf{A}, \gamma, \boldsymbol{b})$ should we consider as reasonable given the data? The obvious answer, which is fully consistent with the usual practices of statistics, is those $(\mathbf{A}, \gamma, \boldsymbol{b})$ which do not excessively degrade the fit of the best-fit parameters $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right)$; that is those $(\mathbf{A}, \gamma, \boldsymbol{b})$ for which

$$\begin{aligned}\rho(\mathbf{A}, \gamma, \boldsymbol{b}) - \rho\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right) \\ = \sum_i \left[\rho_0(||\boldsymbol{v}_i - \gamma\mathbf{A}\boldsymbol{u}_i - \boldsymbol{b}||) \right. \\ \left. - \rho_0(||\boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}\boldsymbol{u}_i - \widehat{\boldsymbol{b}}||)\right]\end{aligned}$$

is not too large.

Recall that, for $p = 3$, if $\boldsymbol{h}$ is small, then $\Phi_3(\boldsymbol{h})\boldsymbol{u}_i$ will be close to $\boldsymbol{u}_i$. This suggests writing

$$\widehat{\mathbf{A}} = \mathbf{A}\Phi_3(\widehat{\boldsymbol{h}}), \tag{31.26}$$

where $\widehat{\boldsymbol{h}} \in \mathbb{R}^3$. Then $\mathbf{A}\boldsymbol{u}_i = \widehat{\mathbf{A}}\Phi_3(-\widehat{\boldsymbol{h}})\boldsymbol{u}_i$ will be close to $\widehat{\mathbf{A}}\boldsymbol{u}_i$ when $\widehat{\boldsymbol{h}}$ is small. Rather than focus on the distribution of $\widehat{\mathbf{A}}$, we will focus on the distribution of the *deviation* of $\widehat{\mathbf{A}}$ from $\mathbf{A}$ as measured by the (hopefully) small vector $\widehat{\boldsymbol{h}}$.

Similarly, for $p = 2$, we will write

$$\widehat{\mathbf{A}} = \mathbf{A}\Phi_2(\widehat{h}), \tag{31.27}$$

where $\widehat{h} \in \mathbb{R}^1$. For general p, one writes

$$\widehat{\mathbf{A}} = \mathbf{A}\exp(\widehat{\mathbf{H}}), \tag{31.28}$$

where $\widehat{\mathbf{H}}$ is $p \times p$ skew-symmetric.

The most elementary procedures in statistics are based upon the fact

> *If $X_1, \cdots, X_n$ are independent and each X_i is distributed $N(\mu, \sigma^2)$, then $\overline{X}$ is distributed $N(\mu, \sigma^2/n)$.*

An equivalent result is

> *If $X_1, \cdots, X_n$ are independent and each X_i is distributed $N(\mu, \sigma^2)$, then $\overline{X} - \mu$ is distributed $N(0, \sigma^2/n)$.*

In the latter form, we have an estimator (in this case $\overline{X}$) and the distribution of the *deviation* $\widehat{h} = \overline{X} - \mu$ of the estimator from the unknown parameter μ. This is sufficient for both confidence intervals and hypothesis testing and is analogous to what we propose to do in Sect. 31.4.

We note that $\Phi_3(\boldsymbol{h}) = \mathbf{I}_3$ whenever $||\boldsymbol{h}|| = 2\pi$. This implies that Φ_3 will have a singularity as $||\boldsymbol{h}|| \to 2\pi$. However, Φ_3 behaves very well for small $\boldsymbol{h}$ and hence (31.26) is a good way to parameterize $\widehat{\mathbf{A}}$ close to $\mathbf{A}$.

All parameterizations of $\mathcal{SO}(3)$ have singularities somewhere. By using parameterizations such as (31.26), (31.27), or (31.28), we put those singularities far away from the region of interest, that is far away from $\mathbf{A}$. As we will see in Sects.31.4 and 31.5, the result is very clean mathematics. However, some formulations of Euler angles [31.7] have a singularity at $\boldsymbol{h} = \mathbf{0}$. This means that Euler angles are an especially poor parameterization of small rotations in $\mathcal{SO}(3)$ (that is, for $\mathbf{A}$ close to $\mathbf{I}_3$) and that, if we were to repeat the calculations of Sect. 31.4 and 31.5 using Euler angles, the results would be much messier.

31.3.5 Numerical Calculation of *M*-Estimates for the Procrustes Model

We use here the geometric insights into $\mathcal{SO}(p)$ to propose a method of minimizing the objective function (31.13) for the Procrustes model. The simplifications necessary to minimize the objective function (31.14) for the spherical regression model should be reasonably clear.

In what follows, it will be convenient to rewrite the Procrustes model

$$\boldsymbol{v}_i = \gamma\mathbf{A}\boldsymbol{u}_i + \boldsymbol{b} + \text{error}$$

in the equivalent form

$$\boldsymbol{v}_i = \gamma\mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) + \boldsymbol{\beta} + \text{error}, \tag{31.29}$$

where $\boldsymbol{\beta} = \gamma\mathbf{A}\overline{\boldsymbol{u}} + \boldsymbol{b}$.

Let $\psi(s) = \rho_0'(s)$. Differentiating (31.13) with respect to γ and $\boldsymbol{\beta}$ we get that the M-estimates $(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{\beta}})$ must satisfy.

$$0 = \sum_i \psi(s_i) s_i^{-1} \left[\boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \widehat{\boldsymbol{\beta}} \right]^{\mathrm{T}} \times \widehat{\mathbf{A}}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) \,, \tag{31.30}$$

$$0 = \sum_i \psi(s_i) s_i^{-1} \left[\boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \widehat{\boldsymbol{\beta}} \right] \,, \tag{31.31}$$

where $s_i = \left|\left| \boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \widehat{\boldsymbol{\beta}} \right|\right|$.

To differentiate (31.13) with respect to $\mathbf{A}$, we note that, if $\mathbf{H}$ is any skew-symmetric matrix, and using (31.25),

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \left\{ \sum_i \rho_0 \left[\left|\left| \boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}\exp(t\mathbf{H}) \right.\right.\right. \\ &\quad \left.\left.\left. \times (\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \widehat{\boldsymbol{\beta}} \right|\right| \right] \right\} \\ &= -\gamma \sum_i \psi(s_i) s_i^{-1} \left[\boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}\,(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \widehat{\boldsymbol{\beta}} \right]^{\mathrm{T}} \\ &\quad \times \widehat{\mathbf{A}}\mathbf{H}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) \\ &= -\gamma \mathrm{Tr}(\widetilde{\mathbf{X}}\mathbf{H}), \end{aligned}$$

where

$$\widetilde{\mathbf{X}} = \sum_i \psi(s_i) s_i^{-1} \, (\boldsymbol{u}_i - \overline{\boldsymbol{u}}) \left[\boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \widehat{\boldsymbol{\beta}} \right]^{\mathrm{T}} \widehat{\mathbf{A}} \,.$$

Since $\mathbf{H}$ is any skew-symmetric matrix, $\widetilde{\mathbf{X}}$ is symmetric. Equivalently

$$\mathbf{X} = \sum_i \psi(s_i) s_i^{-1} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{v}_i - \widehat{\boldsymbol{\beta}})^{\mathrm{T}} \widehat{\mathbf{A}} \text{ is symmetric} \,. \tag{31.32}$$

Equations (31.32), (31.30), and (31.31) lead to the following iterative minimization algorithm. Start with the least squares solution given in Sect. 31.2.3 and use these estimates to calculate s_i. Using these s_i and the current guess for $\widehat{\mathbf{A}}$, solve (31.30) and (31.31) to update the guesses for $\widehat{\gamma}$ and $\widehat{\beta}$. Now writing $\mathbf{X} = \mathbf{O}_1 \mathbf{\Lambda} \mathbf{O}_2^{\mathrm{T}}$ for the singular value decomposition of $\mathbf{X}$, the next guess for $\widehat{\mathbf{A}}$ is $\mathbf{O}_2 \mathbf{O}_1^{\mathrm{T}}$. This yields a minimum in $\mathcal{O}(p)$. If minimization in $\mathcal{SO}(p)$ is desired, a modified singular value decomposition is used for $\mathbf{X}$ instead. Having updated the guesses for $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{\beta}}\right)$, we now iterate.

For example consider the hands data of Table 31.1. We calculate the L_1 estimate for which $\psi(s) = 1$. Starting with the least squares estimates in (31.20), we convert $\widehat{\boldsymbol{b}}$ to

$$\widehat{\beta} = \widehat{\gamma}\widehat{\mathbf{A}}\overline{\boldsymbol{u}} + \widehat{\boldsymbol{b}} = \begin{pmatrix} 9.2500 & 11.3633 & 15.0950 \end{pmatrix}^{\mathrm{T}} . \tag{31.33}$$

We use these least squares estimate as an initial guess; a single iteration of the minimization algorithm yields the updated guess

$$\begin{aligned} \widehat{\mathbf{A}} &= \begin{pmatrix} 0.9569 & 0.2823 & -0.0690 \\ 0.2614 & -0.9399 & -0.2199 \\ 0.1269 & -0.1924 & 0.9731 \end{pmatrix} , \\ \widehat{\gamma} &= 1.0015 \,, \\ \widehat{\beta} &= \begin{pmatrix} 9.2835 & 11.4092 & 15.0851 \end{pmatrix}^{\mathrm{T}} . \end{aligned}$$

Convergence is achieved after around a dozen iterations. We arrive at the L_1 estimates

$$\begin{aligned} \widehat{\mathbf{A}} &= \begin{pmatrix} 0.9418 & 0.3274 & -0.0760 \\ 0.3045 & -0.9268 & -0.2200 \\ 0.1425 & -0.1840 & 0.9725 \end{pmatrix} , \\ \widehat{\gamma} &= 1.0086 \,, \\ \widehat{\beta} &= \begin{pmatrix} 9.2850 & 11.4255 & 15.0883 \end{pmatrix}^{\mathrm{T}} . \end{aligned} \tag{31.34}$$

31.4 Statistical Properties of *M*-Estimates

31.4.1 The Σ Matrix and the Geometry of the u_i

Let $\mathbf{\Sigma}$ be the $p \times p$ matrix

$$\mathbf{\Sigma} = n^{-1} \sum_i (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^{\mathrm{T}}$$

$\mathbf{\Sigma}$ is nonnegative definite symmetric and hence its eigenvalues are real and its eigenvectors form an orthonormal basis of $\mathbb{R}^p$. We can use this eigen-decomposition of $\mathbf{\Sigma}$ to summarize the geometry of the point $\boldsymbol{u}_i$. More specifically, let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of $\mathbf{\Sigma}$ with corresponding eigenvectors $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p$. Then $\boldsymbol{e}_1$ points in the direction of the greatest variation in the $\boldsymbol{u}_i$, and $\boldsymbol{e}_p$ in the direction of the least variation.

31.4.2 Example: Σ for the Hands Data

For example, for the data of Table 31.1,

$$\overline{\boldsymbol{u}} = \begin{pmatrix} 7.8975 \\ 12.7542 \\ 14.3067 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \frac{1}{12}\left\{\left[\left(\begin{matrix} 5.17 \\ 11.30 \\ 16.18 \end{matrix}\right) - \overline{\boldsymbol{u}}\right]\left[\left(\begin{matrix} 5.17 \\ 11.30 \\ 16.18 \end{matrix}\right) - \overline{\boldsymbol{u}}\right]^{\mathrm{T}} + \cdots + \left[\left(\begin{matrix} 8.68 \\ 12.71 \\ 13.67 \end{matrix}\right) - \overline{\boldsymbol{u}}\right]\left[\left(\begin{matrix} 8.68 \\ 12.71 \\ 13.67 \end{matrix}\right) - \overline{\boldsymbol{u}}\right]^{\mathrm{T}}\right\}$$

$$= \begin{pmatrix} 2.6249 & 1.2525 & 0.1424 \\ 1.2525 & 0.8095 & -0.5552 \\ 0.1424 & -0.5552 & 4.9306 \end{pmatrix},$$

$$\lambda_1 = 5.004\,, \qquad \lambda_2 = 3.255\,, \qquad \lambda_3 = 0.1054\,,$$

$$\boldsymbol{e}_1 = \begin{pmatrix} -0.0115 \\ -0.1346 \\ 0.9908 \end{pmatrix}, \quad \boldsymbol{e}_2 = \begin{pmatrix} -0.8942 \\ -0.4420 \\ -0.0704 \end{pmatrix},$$

$$\boldsymbol{e}_3 = \begin{pmatrix} -0.4474 \\ 0.8869 \\ 0.1152 \end{pmatrix}.$$

Examining the data of Table 31.1, one sees that $\overline{\boldsymbol{u}}$ is close to point G, the center of the left palm. Examining the displacement of G to C, top of the middle finger, it is evident that left hand was close to vertically oriented. This is the direction $\boldsymbol{e}_1$. Examining the displacement of G to E, the top of the thumb, it appears that the left thumb was pointed in roughly the direction of the x-axis. This is the direction of $-\boldsymbol{e}_2$. Thus the left hand was roughly parallel to the x–z plane. The normal vector to the plane of the left hand is thus approximately parallel to the y-axis. This is the direction of $\boldsymbol{e}_3$. Notice that λ_3 is much smaller than λ_1 or λ_2, indicating that the thickness of the hand is much smaller than its length or breadth.

31.4.3 Statistical Assumptions for the Procrustes Model

Before giving the statistical properties of $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right)$ it is necessary to make explicit the statistical assumptions of the Procrustes model (31.2). These assumptions are:

- $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n \in \mathbb{R}^p$ are fixed (non-random) vectors.
- $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n \in \mathbb{R}^p$ are independent random vectors.
- The distribution of $\boldsymbol{v}_i$ is of the form $f_0(s_i)$, where $s_i = ||\boldsymbol{v}_i - \gamma\mathbf{A}\boldsymbol{u}_i - \boldsymbol{b}||$. Here $(\mathbf{A}, \gamma, \boldsymbol{b})$ are unknown, $\mathbf{A} \in \mathcal{SO}(p)$ or $\mathcal{O}(p)$, γ is a positive real constant, and $\boldsymbol{b} \in \mathbb{R}^p$.

The most obvious example of a suitable distribution f_0 is

$$f_0(s) = (2\pi\sigma^2)^{-p/2}\,\mathrm{e}^{-\frac{s^2}{2\sigma^2}} \tag{31.35}$$

for a fixed constant σ^2. In what follows, we will not need to know the value of σ^2. In fact, we will not even need to know the form of f_0, only that the distribution of $\boldsymbol{v}_i$ depends only upon its distance s_i from $\gamma\mathbf{A}\boldsymbol{u}_i + \boldsymbol{b}$.

The distribution (31.35) is a multivariate normal distribution with mean vector $\gamma\mathbf{A}\boldsymbol{u}_i + \boldsymbol{b}$ and covariance matrix $\sigma^2\mathbf{I}_p$. Equivalently, the p components of $\boldsymbol{v}_i$ are independent and each has variance σ^2. If the components of $\boldsymbol{v}_i$ were to have different variances, then the distribution of $\boldsymbol{v}_i$ would not satisfy the Procrustes model assumptions.

In essence we assume that $\boldsymbol{v}_i$ is isotropically (i.e., that all directions are the same) distributed around its mean vector.

31.4.4 Theorem (Distribution of $(\widehat{\mathbf{A}},\widehat{\gamma},\widehat{\boldsymbol{b}})$ for the Procrustes Model)

Suppose $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right)$ minimize an objective function of the form (31.13). Let $\boldsymbol{\beta} = \gamma\mathbf{A}\overline{\boldsymbol{u}} + \boldsymbol{b}$ and $\widehat{\boldsymbol{\beta}} = \widehat{\gamma}\widehat{\mathbf{A}}\overline{\boldsymbol{u}} + \widehat{\boldsymbol{b}}$. Then

- $\widehat{\mathbf{A}}$, $\widehat{\gamma}$, and $\widehat{\boldsymbol{\beta}}$ are independent;
- $\widehat{\boldsymbol{\beta}}$ is distributed multivariate normal with mean $\boldsymbol{\beta}$ and covariance matrix $\frac{k}{n}\mathbf{I}_p$;
- If $p = 2$, write $\widehat{\mathbf{A}} = \mathbf{A}\Psi_2(\widehat{h})$, for $\widehat{h} \in \mathbb{R}^1$. Then $\widehat{h}$ is normally distributed with mean 0 and variance $\frac{k}{n\mathrm{Tr}(\Sigma)}$;
- If $p = 3$, write $\widehat{\mathbf{A}} = \mathbf{A}\Psi_3(\widehat{\boldsymbol{h}})$, for $\widehat{\boldsymbol{h}} \in \mathbb{R}^3$. Let $\Sigma = \lambda_1\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}} + \lambda_2\boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}} + \lambda_3\boldsymbol{e}_3\boldsymbol{e}_3^{\mathrm{T}}$ be the spectral decomposition of $\boldsymbol{\Sigma}$. Then $\widehat{\boldsymbol{h}}$ is distributed trivariate normal with mean $\mathbf{0}$ and covariance matrix

$$\frac{k}{n}\Big[(\lambda_2+\lambda_3)^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}} + (\lambda_3+\lambda_1)^{-1}\boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}} + (\lambda_1+\lambda_2)^{-1}\boldsymbol{e}_3\boldsymbol{e}_3^{\mathrm{T}}\Big].$$

- For general p, write $\widehat{\mathbf{A}} = \mathbf{A}\exp(\widehat{\mathbf{H}})$, where $\widehat{\mathbf{H}}$ is $p \times p$ skew-symmetric. Then $\widehat{\mathbf{H}}$ has a multivariate normal density proportional to $\exp\left[-\frac{n}{2k}\mathrm{Tr}(\widehat{\mathbf{H}}^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\mathbf{H}})\right]$;
- $\widehat{\gamma}$ is normally distributed with mean γ and variance $\frac{k}{n\mathrm{Tr}(\boldsymbol{\Sigma})}$.

These results are asymptotic, that is they are large-sample approximate distributions.

The constant k is defined to be

$$k = \frac{pE\left[\psi(s)^2\right]}{E^2\left[\psi'(s)+(p-1)\psi(s)s^{-1}\right]}, \tag{31.36}$$

where $\psi(s) = \rho_0'(s)$. Thus k can be estimated from the sample by

$$\widehat{k} = \frac{np\sum_i \psi(s_i)^2}{\left\{\sum_i\left[\psi'(s_i)+(p-1)\psi(s_i)s_i^{-1}\right]\right\}^2}, \tag{31.37}$$

where $s_i = \left|\left|\boldsymbol{v}_i - \widehat{\gamma}\widehat{\mathbf{A}}\boldsymbol{u}_i - \widehat{\boldsymbol{b}}\right|\right|$.

Theorem 31.4.4 is proven in *Chang* and *Ko* [31.3]. (In [31.3], s is defined to be $s = ||\boldsymbol{v} - \widehat{\gamma}\widehat{\mathbf{A}}\boldsymbol{u} - \widehat{\boldsymbol{b}}||^2$ and this causes the formulas (31.36) and (31.37) to be somewhat different there.)

31.4.5 Example: A Test of $\gamma = 1$

For the hands data, the least squares estimates were given in Example 31.2.4. Table 31.2 gives the calculation of the s_i. Substituting $p = 3$, $\rho_0(s) = s^2$, $\psi(s) = 2s$ into (31.37), $\widehat{k} = (3n)^{-1}\sum_i s_i^2 = 0.0860$.

To test if the two hands are the same size, we test $\gamma = 1$. Using Example 31.4.2, $\mathrm{Tr}(\mathbf{\Sigma}) = 8.365$. Hence the variance of $\widehat{\gamma}$ is 0.000 860 and hence its standard error is 0.0293. Since $\widehat{\gamma} = 0.9925$, we see that $\widehat{\gamma}$ is not significantly different from 1.

The L_1 estimate of γ is 1.0086. To calculate the standard error of this estimate, we use $\rho_0(s) = s$ and $\psi(s) = 1$. Hence for the L_1 estimate, (31.37) yields $\widehat{k} = 0.75\left(n^{-1}\sum_i s_i^{-1}\right)^{-2}$. After recomputing the s_i using L_1 estimates of $(\mathbf{A}, \gamma, \boldsymbol{b})$, we obtain $\widehat{k} = 0.023$. Thus the L_1 estimate of γ has a standard error of 0.0150 and this estimate is also not significantly different from 1.

Apparently, the two hands have the same size.

General statistical theory implies that if the $\boldsymbol{v}_i$ were really normally distributed, the least squares estimates would be the most efficient. In other words, least squares estimates should have the smallest standard errors. Evidently this is not true for the hands data and it appears that this data is not, in fact, normally distributed.

31.4.6 Example: A Test on A

As discussed in 31.4.2, the eigenvector $\boldsymbol{e}_3$ of $\mathbf{\Sigma}$ is perpendicular to the plane of the left palm. It might be of interest to test if the two hands have the same orientation; that is, after reflecting the left hand in the plane perpendicular to $\boldsymbol{e}_3$, do the fingers and thumb of the two hands point in the same directions. We formulation this hypothesis as $H_0 : \mathbf{A} = \mathbf{R}_{\boldsymbol{e}_3}$ where $\mathbf{R}_{\boldsymbol{e}_3}$ is the matrix of the reflection in plane perpendicular to $\boldsymbol{e}_3$.

$$\begin{aligned}\mathbf{R}_{\boldsymbol{e}_3} &= \mathbf{I}_3 - 2\boldsymbol{e}_3\boldsymbol{e}_3^{\mathrm{T}} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &\quad -2\begin{pmatrix} -0.4474 \\ 0.8869 \\ 0.1152 \end{pmatrix}\begin{pmatrix} -0.4474 \\ 0.8869 \\ 0.1152 \end{pmatrix}^{\mathrm{T}} \\ &= \begin{pmatrix} 0.5996 & 0.7936 & 0.1031 \\ 0.7936 & -0.5731 & -0.2044 \\ 0.1031 & -0.2044 & 0.9734 \end{pmatrix},\end{aligned}$$

$\widehat{\boldsymbol{h}}$ is defined by

$$\begin{aligned}\Phi_3(\widehat{\boldsymbol{h}}) &= \mathbf{R}_{\boldsymbol{e}_3}^{\mathrm{T}}\widehat{\mathbf{A}} \\ &= \begin{pmatrix} 0.7843 & -0.6127 & -0.0976 \\ 0.5999 & 0.7890 & -0.1327 \\ 0.1583 & 0.0455 & 0.9863 \end{pmatrix},\end{aligned} \tag{31.38}$$

where $\widehat{\mathbf{A}}$ was calculated in 31.2.4.

To solve for $\widehat{\boldsymbol{h}}$ we use the results at the end of Sect. 31.3.2. The matrix of (31.38) has an eigenvector of $\xi = (0.1395 \quad -0.2003 \quad 0.9494)^{\mathrm{T}}$ corresponding to the eigenvalue of 1. Its angle of rotation is given by

$$\theta = \arccos\left[0.5\mathrm{Tr}\left(\mathbf{R}_{\boldsymbol{e}_3}^{\mathrm{T}}\widehat{\mathbf{A}}\right) - 0.5\right] = 0.6764\,.$$

Thus $\widehat{\boldsymbol{h}} = \theta\boldsymbol{\xi} = (0.0944 \quad -0.1355 \quad 0.6422)^{\mathrm{T}}$.

By Theorem 31.4.4, if H_0 is true, $\widehat{\boldsymbol{h}}$ is trivariate normally distributed with mean $\mathbf{0}$ and covariance matrix

$$\begin{aligned}&\frac{k}{n}\Big[(\lambda_2+\lambda_3)^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}} + (\lambda_3+\lambda_1)^{-1}\boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}} \\ &+(\lambda_1+\lambda_2)^{-1}\boldsymbol{e}_3\boldsymbol{e}_3^{\mathrm{T}}\Big].\end{aligned}$$

The constant k was estimated in 31.4.5 and the λ_i and $\boldsymbol{e}_i$ were calculated in 31.4.2. Using these calculations, the covariance matrix of $\widehat{\boldsymbol{h}}$ is estimated to be

$$\begin{aligned}&\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{h}}) \\ &= \begin{pmatrix} 0.001\,296 & 0.000\,2134 & 0.000\,019\,23 \\ 0.000\,2134 & 0.000\,9951 & -0.000\,1520 \\ 0.000\,019\,23 & -0.000\,1520 & 0.002\,112 \end{pmatrix}.\end{aligned}$$

Table 31.2 Calculation of residual lengths for data from Table 31.1

	Predicted $\widehat{v}_i$ $\widehat{\gamma}\widehat{\mathbf{A}}u_i+\widehat{b}$			Residual $v_i-\widehat{v}_i$			s_i $\|\|v_i-\widehat{v}_i\|\|$
A	6.148	11.687	16.868	−0.238	−0.527	−0.318	0.660
B	8.475	10.969	18.207	0.155	−0.349	0.123	0.401
C	9.620	10.962	18.654	0.470	−0.362	−0.014	0.593
D	11.080	10.457	17.769	−0.190	0.493	0.131	0.544
E	13.206	10.816	13.865	−0.236	−0.686	0.015	0.726
F	8.691	11.176	13.247	0.099	0.034	−0.077	0.129
G	10.544	10.938	13.355	0.156	0.162	0.065	0.234
H	8.501	11.594	11.046	−0.031	−0.504	0.304	0.589
I	7.449	12.225	14.178	−0.169	0.295	−0.138	0.367
J	8.062	11.908	14.649	−0.012	0.512	−0.089	0.520
K	9.198	11.904	14.730	−0.128	0.486	0.130	0.519
L	10.026	11.724	14.573	0.125	0.446	−0.133	0.481

Under the null hypothesis

$$\chi^2 = \widehat{\boldsymbol{h}}^{\mathrm{T}}\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{h}})^{-1}\widehat{\boldsymbol{h}} = 213$$

has an approximate χ^2 distribution with three degrees of freedom.

We emphatically conclude that, after reflecting the left hand, the orientations of the two hands are not the same.

31.4.7 Asymptotic Relative Efficiency of Least Squares and L_1 Estimates

Examining Theorem 31.4.4, we see that the covariance of the M-estimate $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right)$ is determined, up to a constant k, by the geometry of the $\boldsymbol{u}_i$, as summarized by the matrix $\boldsymbol{\Sigma}$. Only the constant k, see (31.36), depends upon the probability distribution of the $\boldsymbol{v}_i$ and the objective function (31.13) that $\left(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}}\right)$ minimize. Furthermore, a sample estimate of k, see (31.37) is available which does not require knowledge of the distribution of the $\boldsymbol{v}_i$.

Let $k(f_0, L_2)$ denote the constant k as defined in (31.36) when the underlying density is of the form f_0 and least squares (L_2) estimation is used, and $k(f_0, L_1)$ the corresponding value when L_1 estimation is used. The ratio $\mathrm{ARE}(L_1, L_2; f_0) = k(f_0, L_2)/k(f_0, L_1)$ is called the *asymptotic relative efficiency* of the L_1 to the least squares estimators at the density f_0.

We see that

$$\mathrm{ARE}(L_1, L_2; f_0) = \frac{\text{variance of least squares estimator}}{\text{variance of } L_1 \text{estimator}}, \tag{31.39}$$

where we recognize that both variances are matrices, but the two variance matrices are multiples of each other.

If f_0 is a p-dimensional normal density (31.35), it can be shown from (31.36) that

$$\mathrm{ARE}(L_1, L_2; N_p) = \frac{2\Gamma^2\left[(p+1)/2\right]}{p\Gamma^2(p/2)}. \tag{31.40}$$

We have used N_p in (31.40) to denote the p-dimensional normal density function.

The Γ function in (31.40) has the properties

$$\Gamma(1) = 1 \quad \Gamma(0.5) = \sqrt{\pi}$$
$$\Gamma(q+1) = q\Gamma(q).$$

Thus when $p = 2.3$

$$\mathrm{ARE}(L_1, L_2; N_2) = \frac{\pi}{4} = 0.785,$$
$$\mathrm{ARE}(L_1, L_2; N_3) = \frac{8}{3\pi} = 0.849. \tag{31.41}$$

$\mathrm{ARE}(L_1, L_2; N_p)$ increases to 1 as $p \to \infty$.

When the underlying distribution is normal, statistical theory indicates that least squares procedures are optimal, that is, they have the smallest variance. Using (31.39) and (31.41), we see that, even when the data is normal, the use of L_1 methods results in only an 8% penalty in standard error. And L_1 methods offer superior resistance to outliers.

Indeed, as we saw in Example 31.4.5, the standard error of the L_1 estimator was *smaller* than the standard error of the least squares estimator. Evidently the hands data set is long-tailed, that is it has more outliers than would be expected with normal data.

31.4.8 The Geometry of the Landmarks and the Errors in $\widehat{\mathbf{A}}$

In this section we will constrain our discussion to the case $p = 3$.

Suppose we write the estimate $\widehat{\mathbf{A}}$ in the form

$$\widehat{\mathbf{A}} = \mathbf{A}\Phi_3(\widehat{\boldsymbol{h}}) \,. \tag{31.42}$$

$\Phi_3(\widehat{\boldsymbol{h}})$ is a (hopefully) small rotation which expresses the deviation of the estimate $\widehat{\mathbf{A}}$ from the true value $\mathbf{A}$.

Recall that $\Phi_3(\widehat{\boldsymbol{h}})$ is a rotation of $||\widehat{\boldsymbol{h}}||$ radians around the axis $||\widehat{\boldsymbol{h}}||^{-1}\widehat{\boldsymbol{h}}$.

In particular $\Phi_3(\widehat{\boldsymbol{h}})^{-1} = \Phi_3(-\widehat{\boldsymbol{h}})$ and

$$\mathbf{A} = \widehat{\mathbf{A}}\Phi_3(-\widehat{\boldsymbol{h}}) \,.$$

According to Theorem 31.4.4, the covariance matrix of $\widehat{\boldsymbol{h}}$ has the form

$$\begin{aligned}\mathrm{Cov}(\widehat{\boldsymbol{h}}) = \frac{k}{n}\Big[&(\lambda_2+\lambda_3)^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}} + (\lambda_3+\lambda_1)^{-1}\boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}} \\ &+ (\lambda_1+\lambda_2)^{-1}\boldsymbol{e}_3\boldsymbol{e}_3^{\mathrm{T}}\Big] \,,\end{aligned} \tag{31.43}$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are the eigenvalues of Σ with corresponding eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$. Since $\widehat{\boldsymbol{h}}$ is normally distributed

$$\chi^2 = \widehat{\boldsymbol{h}}^{\mathrm{T}}\left[\mathrm{Cov}(\widehat{\boldsymbol{h}})\right]^{-1}\widehat{\boldsymbol{h}}$$

is distributed χ^2 with three degrees of freedom.

Thus a confidence region for $\mathbf{A}$ is of the form

$$\left\{\widehat{\mathbf{A}}\Phi_3(-\widehat{\boldsymbol{h}}) \,\middle|\, \widehat{\boldsymbol{h}}^{\mathrm{T}}\left[\mathrm{Cov}(\widehat{\boldsymbol{h}})\right]^{-1}\widehat{\boldsymbol{h}} < \chi^2_{3,\alpha}\right\} \,, \tag{31.44}$$

where $\chi^2_{3,\alpha}$ is the appropriate critical point of a χ^2_3 distribution.

Let $\theta = ||\widehat{\boldsymbol{h}}||$ and $\xi = -||\widehat{\boldsymbol{h}}||^{-1}\widehat{\boldsymbol{h}}$ so that $\widehat{\boldsymbol{h}} = -\theta\xi$.

Thus $\Phi_3(-\widehat{\boldsymbol{h}})$ is a rotation of θ radians around the axis ξ.

Substituting (31.43) into the confidence region (31.44), we can re-express this confidence region as

$$\begin{aligned}\Big\{\widehat{\mathbf{A}}\Phi_3(\theta\xi) \mid \theta^2\frac{n}{k}\Big[&(\lambda_2+\lambda_3)(\xi^{\mathrm{T}}\boldsymbol{e}_1)^2 + (\lambda_3+\lambda_1) \\ &\times(\xi^{\mathrm{T}}\boldsymbol{e}_2)^2 + (\lambda_1+\lambda_2)(\xi^{\mathrm{T}}\boldsymbol{e}_3)^2\Big] < \chi^2_{3,\alpha}\Big\} \,.\end{aligned} \tag{31.45}$$

Now

$$\lambda_2+\lambda_3 \leq \lambda_3+\lambda_1 \leq \lambda_1+\lambda_2 \,.$$

Thus the confidence region (31.45) constrains θ the most (that is the limits on θ are the smallest) when ξ points in the direction $\boldsymbol{e}_3$. It bounds θ the least when ξ points in the direction $\boldsymbol{e}_1$.

Recall also that $\boldsymbol{e}_1$ is the direction of the greatest variation in the $\boldsymbol{u}_i$ and $\boldsymbol{e}_3$ the direction of the least variation.

For the hands data of Table 31.1, $\boldsymbol{e}_1$ points in the direction of the length of the left hand and $\boldsymbol{e}_3$ in the normal direction to the palm.

Thus the angle θ of the small rotation $\Phi_3(\theta\xi)$ is the most constrained when its axis ξ points in the direction of the least variation in the $\boldsymbol{u}_i$. θ is least constrained when ξ points in the direction of the greatest variation of the $\boldsymbol{u}_i$.

For the hands data, if $\widehat{\boldsymbol{h}}$ is in the direction of $\boldsymbol{e}_1$, the length of the hand, it represents a small rotation at the elbow with the wrist held rigid. The variance of the deviation rotation $\widehat{\boldsymbol{h}}$ in the direction $\boldsymbol{e}_1$ is $(\lambda_2+\lambda_3)^{-1} = 0.298$. If $\widehat{\boldsymbol{h}}$ points in the direction of $\boldsymbol{e}_2$, the width of the hand, it represents a forwards and backwards rotation at the wrist; the variance of $\widehat{\boldsymbol{h}}$ in this direction is $(\lambda_2+\lambda_3)^{-1} = 0.196$. Finally if $\widehat{\boldsymbol{h}}$ points in the direction of $\boldsymbol{e}_3$, the normal vector to the hand, it represents a somewhat awkward sideways rotation at the wrist (this rotation is represented in Fig. 31.1b; the variance of $\widehat{\boldsymbol{h}}$ in this direction is $(\lambda_1+\lambda_2)^{-1} = 0.121$. If the variability of the component of $\widehat{\boldsymbol{h}}$ in the direction of a rotation at the elbow

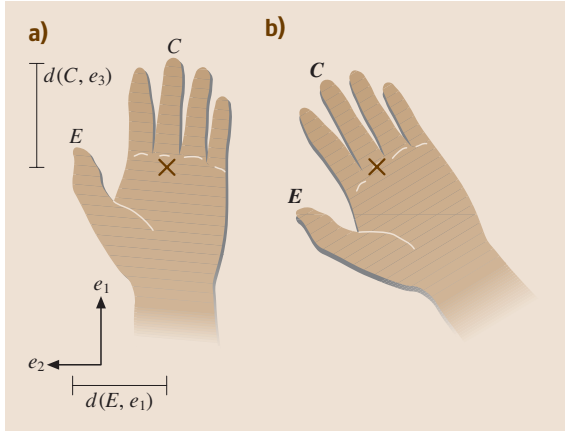

Fig. 31.1 **(a)**A hand with axes $\boldsymbol{e}_1, \boldsymbol{e}_2$; axis $\boldsymbol{e}_3$ points out of paper. X marks the center point $\overline{\boldsymbol{u}}$. The distances $d(C, \boldsymbol{e}_3)$ and $d(E, \boldsymbol{e}_1)$ are the lengths of the indicated line segments. **(b)**The effect of a rotation of angle θ around the axis $\boldsymbol{e}_3$. The point C moves a distance of approximately $d(C, \boldsymbol{e}_3)\theta$. Under a rotation of θ around $\boldsymbol{e}_1$ (not shown), the point E moves a distance of approximately $d(E, \boldsymbol{e}_1)\theta$. Notice that $d(E, \boldsymbol{e}_1) < d(C, \boldsymbol{e}_3)$, and, indeed, the landmarks $\boldsymbol{u}_i$ tend to be closer to $\boldsymbol{e}_1$ than to $\boldsymbol{e}_3$. It follows that a rotation of θ around $\boldsymbol{e}_3$ will move the figure more than a rotation of θ around $\boldsymbol{e}_1$

is unacceptably large, we need to increase λ_3; in effect to create, if possible, landmarks which effectively thicken the palm.

A heuristic derivation of this result is due to *Stock* and *Molnar* [31.8, 9]. It appeared in the geophysical literature and is considered a major development in our understanding of the uncertainties in tectonic plate reconstructions. We will present their argument below, suitably modified for the image registration context.

It is convenient to rewrite the model, as in Theorem 31.4.4, in the form (31.29). If we substitute $\mathbf{A} = \widehat{\mathbf{A}}\Phi_3(\theta\xi)$, we see that $\mathbf{A}$ first perturbs the $\boldsymbol{u}_i - \overline{\boldsymbol{u}}$ by the small rotation $\Phi_3(\theta\xi)$ and then applies the best fitting orthogonal matrix $\widehat{\mathbf{A}}$.

Let $d(\boldsymbol{u}_i, \xi)$ be the distance of the landmark $\boldsymbol{u}_i$ to the line through the center point $\overline{\boldsymbol{u}}$ and in the direction of the axis ξ. Refer to Fig. 31.1. Since the landmarks vary most in the direction $\boldsymbol{e}_1$ and least in the direction $\boldsymbol{e}_3$, the distances $d(\boldsymbol{u}_i, \boldsymbol{e}_3)$ will tend to be biggest and the distances $d(\boldsymbol{u}_i, \boldsymbol{e}_1)$ smallest.

A point $\boldsymbol{x}$ will move a distance of approximately $d(\boldsymbol{x}, \xi)\theta$ under a rotation of angle θ around the axis ξ. It follows that a rotation of angle θ will most move the landmarks $\boldsymbol{u}_i$ if the axis is $\boldsymbol{e}_3$. It will move the landmarks $\boldsymbol{u}_i$ least if the axis is $\boldsymbol{e}_1$. In other words, for a fixed θ, the small rotation $\Phi_3(\theta\xi)$ will most degrade the best fit, provided by $\widehat{\mathbf{A}}$, if $\xi = \boldsymbol{e}_3$; it will least degrade the best fit if $\xi = \boldsymbol{e}_1$.

An orthogonal transformation $\mathbf{A} = \widehat{\mathbf{A}}\Phi_3(\theta\xi)$ is considered a possible transformation if it does not degrade the best fit by too much. It follows that θ is most constrained if $\xi = \boldsymbol{e}_3$, the direction of the least variation in the landmarks $\boldsymbol{u}_i$, and is least constrained if $\xi = \boldsymbol{e}_1$, the direction of greatest variation in the landmarks $\boldsymbol{u}_i$.

Suppose instead we were to write the estimate $\widehat{\mathbf{A}}$ in the form

$$\widehat{\mathbf{A}} = \Phi_3(\widehat{\boldsymbol{h}}_v)\mathbf{A}\,, \qquad (31.46)$$
$$\mathbf{A} = \Phi_3(-\widehat{\boldsymbol{h}}_v)\widehat{\mathbf{A}}\,.$$

Then (31.43) is replaced by

$$\begin{aligned}\mathrm{Cov}(\widehat{\boldsymbol{h}}_v) = \frac{k}{n}\Big[&(\lambda_2+\lambda_3)^{-1}(\mathbf{A}\boldsymbol{e}_1)(\mathbf{A}\boldsymbol{e}_1)^{\mathrm{T}}\\ &+(\lambda_3+\lambda_1)^{-1}(\mathbf{A}\boldsymbol{e}_2)(\mathbf{A}\boldsymbol{e}_2)^{\mathrm{T}}\\ &+(\lambda_1+\lambda_2)^{-1}(\mathbf{A}\boldsymbol{e}_3)(\mathbf{A}\boldsymbol{e}_3)^{\mathrm{T}}\Big]\,.\end{aligned}$$

The same reasoning then expresses the errors of $\widehat{\boldsymbol{h}}_v$, and hence of $\widehat{\mathbf{A}}$, in terms of the geometry of the landmarks $\boldsymbol{v}_i$. In other words, for the hands data, using the definition (31.46) expresses the errors of $\widehat{\mathbf{A}}$ in terms of the orientation of the right hand.

31.4.9 Statistical Properties of *M*-Estimates for Spherical Regressions

The statistical assumptions of the spherical regression model (31.1) are:

- $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n \in \Omega_p$ are fixed (non-random) vectors.
- $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n \in \Omega_p$ are independent random vectors.
- The distribution of $\boldsymbol{v}_i$ is of the form $f_0(t_i)$ where $t_i = \boldsymbol{v}_i^{\mathrm{T}}\mathbf{A}\boldsymbol{u}_i$. Here $\mathbf{A} \in \mathcal{SO}(p)$ or $\mathcal{O}(p)$ is unknown.

A commonly used distribution for spherical data $\boldsymbol{x} \in \Omega_p$ is the distribution whose density (with respect to surface measure, or uniform measure, on Ω_p) is

$$f(\boldsymbol{x};\theta) = c(\kappa)\exp(\kappa\boldsymbol{x}^{\mathrm{T}}\theta)\,. \qquad (31.47)$$

This distribution has two parameters: a positive real constant κ which is commonly called the *concentration parameter* and $\theta \in \Omega_p$. It is easily seen that $f(\boldsymbol{x})$ is maximized over $\boldsymbol{x} \in \Omega_p$ at θ and hence θ is usually refered to as the *modal vector*; $c(\kappa)$ is a normalizing constant.

If $\kappa = 0$, (31.47) is a uniform density on Ω_p. On the other hand as $\kappa \to \infty$, the density (31.47) approaches that of a multivariate normal distribution in $p-1$ dimensions with a covariance matrix of $\kappa^{-1}\mathbf{I}_{p-1}$. Thus intuitively we can think of κ as σ^{-2}, that is think of κ as the inverse variance. As $\kappa \to \infty$, (31.47) approaches a singular multivariate normal distribution supported on the $(p-1)$-dimensional subspace $\theta^{\perp} \subset \mathbb{R}^p$. As a singular multivariate normal distribution in $\mathbb{R}^p$ its covariance matrix is $\kappa^{-1}(\mathbf{I}_p - \theta\theta^{\mathrm{T}})$.

For the circle Ω_1, (31.47) is due to von Mises. For general Ω_p, it is due (independently) to Fisher and to Langevin. More properties of the Fisher–von Mises–Langevin distribution can be found in *Watson* [31.10] or in *Fisher* et al. [31.11].

The distribution of an M-estimator $\widehat{\mathbf{A}}$ which minimizes an objective function of the form (31.14) is similar to the distribution given in Theorem 31.4.4:

- If $p = 2$, write $\widehat{\mathbf{A}} = \mathbf{A}\Psi_2(\widehat{h})$, for $\widehat{h} \in \mathbb{R}^1$. Then $\widehat{h}$ is normally distributed with mean 0 and variance $\frac{k}{n}$.
- If $p = 3$, write $\widehat{\mathbf{A}} = \mathbf{A}\Psi_3(\widehat{\boldsymbol{h}})$, for $\widehat{\boldsymbol{h}} \in \mathbb{R}^3$. Let $\boldsymbol{\Sigma} = \lambda_1\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}} + \lambda_2\boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}} + \lambda_3\boldsymbol{e}_3\boldsymbol{e}_3^{\mathrm{T}}$ be the spectral decomposition of $\boldsymbol{\Sigma}$. Then $\widehat{\boldsymbol{h}}$ is distributed trivariate normal with mean $\mathbf{0}$ and covariance matrix

$$\begin{aligned}\frac{k}{n}\Big[&(\lambda_2+\lambda_3)^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}} + (\lambda_3+\lambda_1)^{-1}\boldsymbol{e}_2\boldsymbol{e}_2^{\mathrm{T}}\\ &+(\lambda_1+\lambda_2)^{-1}\boldsymbol{e}_3\boldsymbol{e}_3^{\mathrm{T}}\Big]\,.\end{aligned}$$

- For general p, write $\widehat{\mathbf{A}} = \mathbf{A}\exp(\widehat{\mathbf{H}})$, where $\widehat{\mathbf{H}}$ is $p \times p$ skew-symmetric. Then $\widehat{\mathbf{H}}$ has a multivariate normal density proportional to $\exp\left[-\frac{n}{2k}\mathrm{Tr}(\widehat{\mathbf{H}}^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\mathbf{H}})\right]$.

Let $\psi(t) = -\rho_0'(t)$. (The sign of ψ has been chosen to make $\psi(t)$ nonnegative, since ρ_0 is a decreasing function of t.) The constant k and its sample estimate $\widehat{k}$ are given by

$$k = \frac{(p-1)E[\psi(t)^2(1-t^2)]}{E^2[(p-1)\psi(t)t - \psi'(t)(1-t^2)]}\,,$$

$$\widehat{k} = \frac{n(p-1)\sum_i \psi(t_i)^2(1-t_i^2)}{\left\{\sum_i[(p-1)\psi(t_i)t_i - \psi'(t_i)(1-t_i^2)]\right\}^2}\,. \quad (31.48)$$

For the spherical case, the matrix $\boldsymbol{\Sigma} = \sum_i \boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}$. Its dominant eigenvector $\boldsymbol{e}_1$ points in the direction of the center of the $\boldsymbol{u}_i$. The $\boldsymbol{e}_2$ is the vector perpendicular to $\boldsymbol{e}_1$ so that the two-dimensional plane spanned by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ (and the origin) best fits the $\boldsymbol{u}_i$. This continues until $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{p-1}$ is the $(p-1)$-dimensional hyperplane, among the collection of all $(p-1)$-dimensional hyperplanes that bests fits the data. This latter hyperplane is, of course, the hyperplane perpendicular to $\boldsymbol{e}_p$. Except for this slight reinterpretation of the geometric meaning of the $\boldsymbol{e}_i$, our previous comments about the relationship of the uncertainties in $\widehat{\boldsymbol{h}}$ to the geometry of the $\boldsymbol{u}$-points, as summarized by the eigen-decomposition of $\boldsymbol{\Sigma}$, remain valid. Indeed the original Stock and Molnar insights about the uncertainties of tectonic plate reconstructions were actually in the spherical data context.

Thus, as before, the uncertainties in $\widehat{\mathbf{A}}$ are determined up to the constant k by the geometry of the $\boldsymbol{u}$-points. Only the constant k depends upon the underlying data distribution f_0 or upon the objective function ρ. We can define the asymptotic relative efficiency as in Sect. 31.4.7 without change. Its interpretation (31.39) also remains valid.

Equation (31.48) implies that we can, as before, define the asymptotic efficiency of the L_1 estimator relative to the least squares estimator, at the density f_0, as $\mathrm{ARE}(L_1, L_2; f_0) = k(f_0, L_2)/k(f_0, L_1)$. The interpretation (31.39) remains valid. The constants $k(f_0, L_2)$ and $k(f_0, L_1)$ come from (31.48) using the underlying density f_0 under consideration and $\rho_0(t) = 2 - 2t$ [refer to (31.9)], $\psi(t) = 2$, for the least squares case, or $\rho_0(t) = \arccos(t)$, $\psi(t) = (1-t^2)^{\frac{1}{2}}$, for the L_1 case. If f_0 is the Fisher–von Mises–Langevin density (31.47) on Ω_p (which we will denote by $F_{\kappa,p}$ in the following)

$$\mathrm{ARE}(L_1, L_2; F_{\kappa,p}) = \frac{\left[\int_{-1}^{1} \mathrm{e}^{\kappa t}(1-t^2)^{(p-2)/2}\,\mathrm{d}t\right]^2}{\left[\int_{-1}^{1} \mathrm{e}^{\kappa t}(1-t^2)^{(p-1)/2}\,\mathrm{d}t\right]} \times \frac{1}{\left[\int_{-1}^{1} \mathrm{e}^{\kappa t}(1-t^2)^{(p-3)/2}\,\mathrm{d}t\right]}\,. \quad (31.49)$$

As $\kappa \to \infty$, the limit of (31.49) is

$$\lim_{\kappa\to\infty}\mathrm{ARE}(L_1, L_2; F_{\kappa,p}) = \frac{2\Gamma^2(p/2)}{(p-1)\Gamma^2[(p-1)/2]}\,. \quad (31.50)$$

Comparing (31.40) with (31.50), we see that (31.50) is the same as (31.40) with p replaced by $p-1$. This is as expected because, as noted above, for large κ the Fisher–von Mises–Langevin distribution approaches a $(p-1)$-dimensional multivariate normal distribution. Figure 31.2 gives a graph of $\mathrm{ARE}(L_1, L_2; F_{\kappa,p})$ for $p = 2, 3$.

In particular for $p = 3$, $\mathrm{ARE}(L_1, L_2; F_\kappa, 3) \to \pi/4$. For the Fisher–von-Mises–Langevin distribution, least squares methods are optimal. Nevertheless, in standard error terms, the penalty for using L_1 methods is at most 13%.

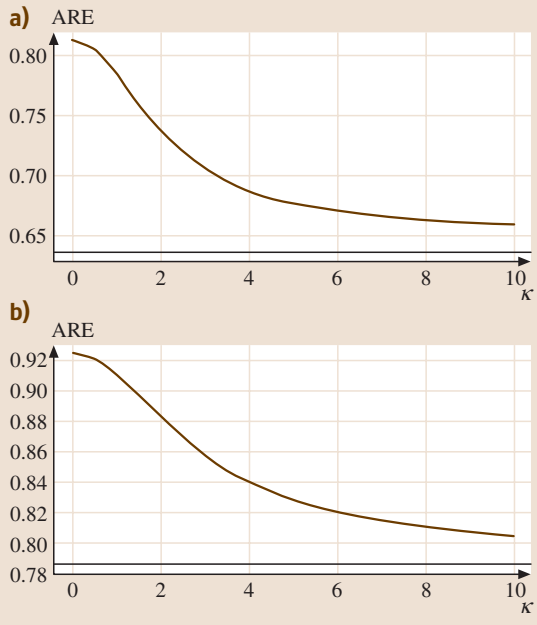

Fig. 31.2a,b Asymptotic efficiency of L_1 estimators relative to least squares estimators for Fisher–von Mises–Langevin distributions on Ω_p as a function of κ for **(a)** $p = 2$ and **(b)** $p = 3$. Horizontal lines are asymptotic limits as $\kappa \to \infty$.

31.5 Diagnostics

We discuss in this section influence function diagnostics for the Procrustes model. Suppose the registration provided by the estimates $(\widehat{\mathbf{A}}, \widehat{\gamma}, \widehat{\boldsymbol{b}})$ is unsatisfactory. These diagnostics will determine which points are influential for the estimated orthogonal matrix $\widehat{\mathbf{A}}$, which points are influential for the estimated scale change $\widehat{\gamma}$, and which are influential for the estimated translation $\widehat{\boldsymbol{b}}$.

31.5.1 Influence Diagnostics in Simple Linear Regression

As background discussion, we consider first the simple linear regression model

$$y_i = \alpha + \beta x_i + \text{error}\,, \tag{31.51}$$

where $x_i, y_i \in \mathbb{R}^1$. For simplicity, we will assume $\sum_i x_i = 0$. This can be accomplished by a centering transformation similar to that used in (31.29).

For the model (31.51), the least squares estimates are

$$\widehat{\alpha} = \overline{y}\,,$$
$$\widehat{\beta} = \left(\sum_i x_i^2\right)^{-1} \left(\sum_i x_i y_i\right)\,. \tag{31.52}$$

Suppose we delete the i-th observation (x_i, y_i) and recompute the estimates (31.52). The resulting estimates would be [see *Cook* and *Weisberg* [31.12], (3.4.6)]

$$\widehat{\alpha}_{(i)} = \widehat{\alpha} - (1 - v_{ii})^{-1} \frac{e_i}{n}\,,$$
$$\widehat{\beta}_{(i)} = \widehat{\beta} - (1 - v_{ii})^{-1} \frac{x_i e_i}{\sum_k x_k^2}\,, \tag{31.53}$$

where

$$e_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i$$

is the residual, and

$$v_{ii} = \frac{1}{n} + \frac{x_i^2}{\sum_k x_k^2}$$

is the i-th diagonal entry of the so-called *hat* matrix. It can be shown that

$$0 \le v_{ii} \le 1\,,$$
$$\sum_i v_{ii} = 2\,. \tag{31.54}$$

If $|x_i|$ is big, $1 - v_{ii}$ can be close to zero, although because of (31.54), if n is large, this will usually not be the case. Ignoring the factor of $(1 - v_{ii})^{-1}$, it follows from (31.53) that deletion of (x_i, y_i) will be influential for $\widehat{\alpha}$ when the magnitude of the residual $|e_i|$ is big. Deletion of (x_i, y_i) will be influential for $\widehat{\beta}$ when both $|x_i|$ and $|e_i|$ are big. Points with large values of $|x_i|$ [typically, due to (31.54), $|x_i| > \frac{4}{n}$] are called *high-leverage* points, whereas points with large values of $|e_i|$ are called *outliers*. (Recall we have centered the data so that $\overline{x} = 0$.)

Thus influence on $\widehat{\alpha}$ and on $\widehat{\beta}$ are different. Outliers are influential for $\widehat{\alpha}$, whereas influence for $\widehat{\beta}$ is a combination of being an outlier and having high leverage. For the model (31.51) with the least squares estimators, the *influence function* works out to be

$$\mathrm{IF}\left[\widehat{\alpha}; (x_i, y_i)\right] = \frac{y_i - \alpha - \beta x_i}{n}\,, \tag{31.55}$$

$$\mathrm{IF}\left[\widehat{\beta}; (x_i, y_i)\right] = \frac{x_i(y_i - \alpha - \beta x_i)}{\sum_k x_k^2}\,, \tag{31.56}$$

where α and β are the 'true' population values in the model (31.51). We will not give a formal definition of the influence function here, but refer the reader to *Cook* and *Weisberg* [31.12] for a more comprehensive discussion of the influence function in the regression model.

It should be noted tthat to actually calculate (31.55) and (31.56) from a sample, it is necessary to estimate α and β. Thus, even though in the left-hand sides of (31.55) and (31.56), $\widehat{\alpha}$ and $\widehat{\beta}$ are least squares estimates, we should substitute in the right-hand sides of (31.55) and (31.56) better estimates, if available, of α and β. There is no contradiction in using L_1 estimates to estimate the influence function of the least squares estimators.

31.5.2 Influence Diagnostics for the Procrustes Model

Chang and *Ko* [31.3] calculated the *standardized influence functions* (SIF) for M-estimates (31.13) in the Procrustes model (31.29). (The influence functions of the estimates $\widehat{\mathbf{A}}$ and $\widehat{\beta}$ are vectors; the standardization calculates their square length in some metric.) Using their notation

$$||\mathrm{SIF}\left[\widehat{\beta}; (\boldsymbol{u}_i, \boldsymbol{v}_i)\right]||^2 = k_I \psi(s_i)^2\,, \tag{31.57}$$

where $s_i = ||\boldsymbol{v}_i - \gamma \mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \beta||$ and $\psi(s) = \rho_0'(s)$. Therefore

- the influence of $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ on the estimate $\widehat{\beta}$ of the translation parameter depends only upon the length s_i of the residual.

This behavior is similar to that of simple linear regression (31.55). The constant k_I is given by

$$k_I = \frac{pE[g'(s)^2]}{E^2[\psi'(s)+(p-1)\psi(s)s^{-1}]} ,$$

where $g(s) = \log f_0(s)$ and $f_0(s)$ is defined in Sect. 31.4.3.

For the scale parameter γ, let $\mathbf{\Sigma} = n^{-1}\sum_i (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^{\mathrm{T}}$. Then

$$||\mathrm{SIF}[\widehat{\gamma}; (\boldsymbol{u}_i, \boldsymbol{v}_i)]||^2 = \frac{k_I \psi(s_i)^2}{\mathrm{Tr}(\mathbf{\Sigma})} \left[\boldsymbol{w}_i^{\mathrm{T}} \mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})\right]^2 . \quad (31.58)$$

Here

$$\boldsymbol{w}_i = [\boldsymbol{v}_i - \gamma \mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \beta]/s_i .$$

Notice that $\boldsymbol{v}_i - \gamma\mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \beta$ is the residual of the i-th data point and s_i is its length. Thus $\boldsymbol{w}_i$ is a unit-length vector in the direction of the i-th data point. We conclude

- For a given length s_i of residual, a point $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ will be influential for the estimate $\widehat{\gamma}$ of the scale parameter if $\boldsymbol{u}_i$ is far from the center $\overline{\boldsymbol{u}}$ of the data and if its residual is parallel to $\mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})$.

For simplicity, we restrict the formulas of influence on the estimate of the orthogonal matrix $\mathbf{A}$ to the cases $p = 2, 3$. For $p = 2$,

$$||\mathrm{SIF}\left[\widehat{\mathbf{A}}; (\boldsymbol{u}_i, \boldsymbol{v}_i)\right]||^2 = \frac{k_I \psi(s_i)^2}{\mathrm{Tr}(\mathbf{\Sigma})} ||\boldsymbol{w}_i \times [\mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})]||^2 . \quad (31.59)$$

The product on the right-hand side of (31.59) is the vector 'cross' product. Therefore

- For $p = 2$, for a given length s_i of residual, a point $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ will be influential for the estimate $\widehat{\mathbf{A}}$ of the orthogonal matrix if $\boldsymbol{u}_i$ is far from the center $\overline{\boldsymbol{u}}$ of the data and if its residual is perpendicular to $\mathbf{A}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})$. Thus points which are influential for $\widehat{\mathbf{A}}$ will not be influential for $\widehat{\gamma}$, and vice versa. Indeed

$$||\mathrm{SIF}[\widehat{\gamma}; (\boldsymbol{u}_i, \boldsymbol{v}_i)]||^2 + ||\mathrm{SIF}\left[\widehat{\mathbf{A}}; (\boldsymbol{u}_i, \boldsymbol{v}_i)\right]||^2 = \frac{k_I \psi(s_i)^2}{\mathrm{Tr}(\mathbf{\Sigma})} ||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||^2 .$$

For $p = 3$, let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of Σ and let $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ be the corresponding eigenvectors. Write

$$\boldsymbol{w}_i \times \left(\mathbf{A}\frac{\boldsymbol{u}_i - \overline{\boldsymbol{u}}}{||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||}\right) = x_1 \mathbf{A}\boldsymbol{e}_1 + x_2 \mathbf{A}\boldsymbol{e}_2 + x_3 \mathbf{A}\boldsymbol{e}_3 .$$

Then

$$\mathrm{SIF}\left[\widehat{\mathbf{A}}; (\boldsymbol{u}_i, \boldsymbol{v}_i)\right] = k_I \psi(s_i)^2 ||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||^2 \times \left(\frac{x_1^2}{\lambda_2 + \lambda_3} + \frac{x_2^2}{\lambda_3 + \lambda_1} + \frac{x_3^2}{\lambda_1 + \lambda_2}\right) . \quad (31.60)$$

It follows

- For $p = 3$, for a given length s_i of residual and distance $||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||$ of $\boldsymbol{u}_i$ from the center of the data, a point $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ will be maximally influential for the estimate $\widehat{\mathbf{A}}$ of the orthogonal matrix if both $\boldsymbol{u}_i - \overline{\boldsymbol{u}}$ is perpendicular to the dominant eigenvector $\boldsymbol{e}_1$ of $\mathbf{\Sigma}$ and the residual

$$\boldsymbol{w}_i = \pm\mathbf{A}\left(\frac{\boldsymbol{u}_i - \overline{\boldsymbol{u}}}{||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||} \times \boldsymbol{e}_1\right) .$$

- The influence of $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ on $\widehat{\mathbf{A}}$ will be zero if

$$\boldsymbol{w}_i = \pm\mathbf{A}\left(\frac{\boldsymbol{u}_i - \overline{\boldsymbol{u}}}{||\boldsymbol{u}_i - \overline{\boldsymbol{u}}||}\right) .$$

- The maximum influence of the data on the estimate $\widehat{\mathbf{A}}$ of the orthogonal matrix can be minimized for fixed $\mathrm{Tr}(\mathbf{\Sigma})$ by making $\lambda_1 = \lambda_2 = \lambda_3$. Thus the optimal choice of landmarks would make the landmarks spherically symmetric around the center point $\overline{\boldsymbol{u}}$.

31.5.3 Example: Influence for the Hands Data

For the Procrustes model (31.29) and the hands data, we compare here the influence statistics for the least squares estimates $(\widehat{\mathbf{A}}_2, \widehat{\gamma}_2, \widehat{\boldsymbol{\beta}}_2)$ [given in (31.20) and (31.33)] to those for the the L_1 estimates $(\widehat{\mathbf{A}}_1, \widehat{\gamma}_1, \widehat{\boldsymbol{\beta}}_1)$ [in (31.34)]. These estimates correspond to $\psi_2(s) = s$ and $\psi_1(s) = 1$ respectively. In the right-hand sides of (31.57), (31.58), and (31.60), we substituted $s_i = ||\boldsymbol{v}_i - \widehat{\gamma}_1 \widehat{\mathbf{A}}_1(\boldsymbol{u}_i - \overline{\boldsymbol{u}}) - \widehat{\boldsymbol{\beta}}_1||$ to calculate the influence functions for both the L_1 and least squares estimates. Similarly the $\boldsymbol{w}_i$ were calculated using the L_1 estimates. Furthermore when (31.57), (31.58), and (31.60) were calculated for the i-th observation $(\boldsymbol{u}_i, \boldsymbol{v}_i)$, $\boldsymbol{u}_i$ was not used to calculate $\mathbf{\Sigma}$.

Using (31.57),

$$||\mathrm{SIF}[\widehat{\boldsymbol{\beta}}_2; (\boldsymbol{u}_i, \boldsymbol{v}_i)]||^2 \propto s_i^2 ,$$
$$||\mathrm{SIF}[\widehat{\boldsymbol{\beta}}_1; (\boldsymbol{u}_i, \boldsymbol{v}_i)]||^2 \propto 1 ,$$

so that E (top of thumb), followed by H (base of palm), are the most influential for $\widehat{\beta}_2$. All points are equally influential for $\widehat{\beta}_1$.

In what follows we will be interested in determining which data points are most influential for which estimates. In other words we will be interested in the relative values of $||\mathrm{SIF}||^2$. Thus, for each estimator, we renormalized the values of $||\mathrm{SIF}||^2$ so that their sum (over the 12 data points) equals 1. The results, together with the values of s_i, are shown in Fig. 31.3.

We have from (31.58) and (31.60)

$$||\mathrm{SIF}\,[\widehat{\gamma}_2;(\boldsymbol{u}_i,\boldsymbol{v}_i)]\,||^2 \propto s_i^2\left[\boldsymbol{w}_i^{\mathrm{T}}\mathbf{A}(\boldsymbol{u}_i-\overline{\boldsymbol{u}})\right]^2 ,$$

$$||\mathrm{SIF}\,[\widehat{\gamma}_1;(\boldsymbol{u}_i,\boldsymbol{v}_i)]\,||^2 \propto \left[\boldsymbol{w}_i^{\mathrm{T}}\mathbf{A}(\boldsymbol{u}_i-\overline{\boldsymbol{u}})\right]^2 ,$$

$$\left|\left|\mathrm{SIF}\left[\widehat{\mathbf{A}}_2;(\boldsymbol{u}_i,\boldsymbol{v}_i)\right]\right|\right|^2 \propto s_i^2\left|\left|\boldsymbol{u}_i-\overline{\boldsymbol{u}}\right|\right|^2 \left(\frac{x_1^2}{\lambda_2+\lambda_3}+\frac{x_2^2}{\lambda_3+\lambda_1}+\frac{x_3^2}{\lambda_1+\lambda_2}\right) ,$$

$$\left|\left|\mathrm{SIF}\left[\widehat{\mathbf{A}}_1;(\boldsymbol{u}_i,\boldsymbol{v}_i)\right]\right|\right|^2 \propto \left|\left|\boldsymbol{u}_i-\overline{\boldsymbol{u}}\right|\right|^2 \left(\frac{x_1^2}{\lambda_2+\lambda_3}+\frac{x_2^2}{\lambda_3+\lambda_1}+\frac{x_3^2}{\lambda_1+\lambda_2}\right) ,$$

$$\boldsymbol{w}_i\times\left[\widehat{\mathbf{A}}_1\frac{\boldsymbol{u}_i-\overline{\boldsymbol{u}}}{||\boldsymbol{u}_i-\overline{\boldsymbol{u}}||}\right] = x_1\widehat{\mathbf{A}}_1\boldsymbol{e}_1+x_2\widehat{\mathbf{A}}_1\boldsymbol{e}_2+x_3\widehat{\mathbf{A}}_1\boldsymbol{e}_3 .$$

Examining Fig. 31.3, we see that point E is by far the most influential point for $\widehat{\mathbf{A}}$. Its relative influence however can be somewhat diminished by using L_1 estimates. The value of $||\boldsymbol{u}_{\mathrm{E}}-\overline{\boldsymbol{u}}||$ is also the largest of the $||\boldsymbol{u}_i-\overline{\boldsymbol{u}}||$. It turns out that $\boldsymbol{u}_{\mathrm{E}}-\overline{\boldsymbol{u}}$ makes an angle of 13° with $\boldsymbol{e}_2$ and that the unit length $\boldsymbol{w}_{\mathrm{E}}$ makes an angle of 12° with $\widehat{\mathbf{A}}_1\boldsymbol{e}_3$. Thus x_1 will be relatively big and x_2, x_3 relatively small. This accounts for the strong influence of point E on both estimates of $\mathbf{A}$. Notice that s_{E} and $||\boldsymbol{u}_{\mathrm{E}}-\overline{\boldsymbol{u}}||$ are sufficiently big that, despite the directions of $\boldsymbol{w}_{\mathrm{E}}$ and $\widehat{A}_1(\boldsymbol{u}_{\mathrm{E}}-\overline{\boldsymbol{u}})$, E is still fairly influential for $\widehat{\gamma}_2$. However, its influence on $\widehat{\gamma}_1$, which does not depend upon s_{E}, is quite small.

The point H (base of the palm) is the most influential point for $\widehat{\gamma}$. H is perhaps the least well-defined point so that it is not surprising that its residual length s_{H} is relatively big. It also defines the length of the hand, so that its influence on $\widehat{\gamma}$ is not surprising. Indeed if H were completely deleted, $\widehat{\gamma}_2$ would change from 0.9925 to 1.0110 and $\widehat{\gamma}_1$ changes from 1.0086 to 1.0262.

One might think that C (top of the middle finger) would also be influential for $\widehat{\gamma}$. In a coordinate system of the eigenvectors of $\boldsymbol{\Sigma}$,

$$\boldsymbol{u}_{\mathrm{H}}-\overline{\boldsymbol{u}} = \begin{bmatrix} -3.98 & 1.15 & 0.33 \end{bmatrix}^{\mathrm{T}}$$

$$\boldsymbol{u}_{\mathrm{C}}-\overline{\boldsymbol{u}} = \begin{bmatrix} 3.55 & -0.77 & -0.03 \end{bmatrix}^{\mathrm{T}}$$

so that $\boldsymbol{u}_{\mathrm{H}}-\overline{\boldsymbol{u}} \approx -(\boldsymbol{u}_{\mathrm{C}}-\overline{\boldsymbol{u}})$. It is useful here to remember that $\boldsymbol{e}_1$ is approximately in the direction of the length of the left hand. Furthermore s_{C} and s_{H} are reasonably close.

However Fig. 31.3 indicates that C has negligible influence on both estimates of γ. Indeed if C were completely deleted, $\widehat{\gamma}_2$ would only change from 0.9925 to 0.9895 and $\widehat{\gamma}_1$ change from 1.0086 to 1.0047. These changes are much smaller than those caused by the deletion of H.

The difference is that $\widehat{\mathbf{A}}_1(\boldsymbol{u}_{\mathrm{C}}-\overline{\boldsymbol{u}})$ makes an angle of 88° with $\boldsymbol{w}_{\mathrm{C}}$. In other words $\widehat{\mathbf{A}}_1(\boldsymbol{u}_{\mathrm{C}}-\overline{\boldsymbol{u}})$ and $\boldsymbol{w}_{\mathrm{C}}$ are very close to perpendicular (Perhaps the close to perpendicularity of the residual at C to $\widehat{\mathbf{A}}_1(\boldsymbol{u}_{\mathrm{C}}-\overline{\boldsymbol{u}})$ is to be expected. The uncertainty in locating C is roughly tangential to the middle finger tip.) Hence the influence of C on $\widehat{\gamma}$ is negligible.

On the other hand, $\widehat{\mathbf{A}}_1(\boldsymbol{u}_{\mathrm{H}}-\overline{\boldsymbol{u}})$ makes an angle of 124° with $\boldsymbol{w}_{\mathrm{H}}$. This accounts for the greater influence of H.

Thus if the registration between the two hands is unsatisfactory in either the translation or rotation parameters, point E should be inspected. If it is unsatisfactory in the scale change, point H should be checked.

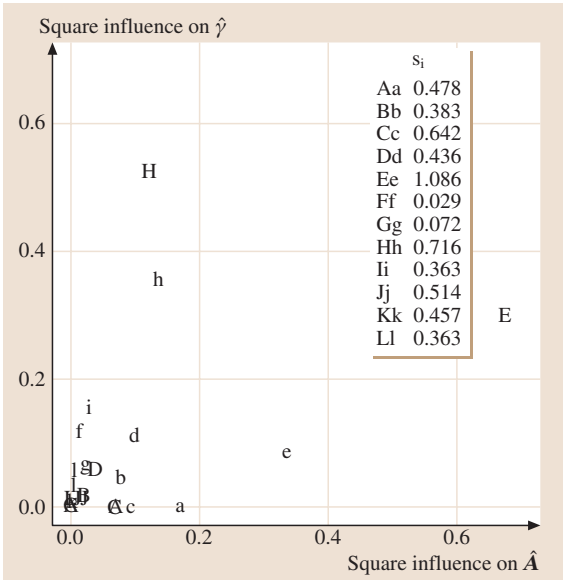

Fig. 31.3 Relative influence of hands data points on least squares (upper case) and L_1 estimates (lower case) of γ and $\mathbf{A}$

References

31.1 G. Wahba: Section on problems and solutions: A least squares estimate of satellite attitude, SIAM Rev. **8**, 384–385 (1966)

31.2 G.R. Chapman, G. Chen, P.T. Kim: Assessing geometric integrity through spherical regression techniques, Stat. Sin. **5**, 173–220 (1995)

31.3 T. Chang, D. Ko: *M*-estimates of rigid body motion on the sphere and in Euclidean space, Ann. Stat. **23**, 1823–1847 (1995)

31.4 Colin, Goodall: Procrustes methods in the statistical analysis of shape, J. R. Stat. Soc. B **53**, 285–339 (1991)

31.5 T. Chang: Spherical regression, Ann. Stat. **14**, 907–924 (1986)

31.6 P.J. Huber: *Robust Statistics* (Wiley, New York 1981)

31.7 H. Goldstein: *Classical Mechanics* (Addison–Wesley, Reading 1950)

31.8 J. Stock, P. Molnar: Some geometrical aspects of uncertainties in combined plate reconstructions, Geology **11**, 697–701 (1983)

31.9 P. Molnar, J. Stock: A method for bounding uncertainties in combined plate reconstructions, J. Geophys. Res. **90**, 12537–12544 (1985)

31.10 G.S. Watson: *Statistics on Spheres* (Wiley Interscience, New York 1983)

31.11 N.I. Fisher, T. Lewis, B.J.J. Embleton: *Statistical Analysis of Spherical Data* (Cambridge Univ. Press, Cambridge 1987)

31.12 R.D. Cook, S. Weisberg: *Residuals and Influence in Regression* (Chapman Hall, New York 1982)