RC 6945 (#29783) 1/17/78
Mathematics 9 pages

Research Report

LEAST SQUARES AND MAXIMUM LIKELIHOOD ESTIMATES OF RIGID MOTION

by

Arthur Nádas[1]

Scientific Computation Department
IBM System Products Division
Hopewell Junction, New York 12533

Typed by Barbara Corti on CMC (Ad.2059)

IBM Research Division
San Jose · Yorktown · Zurich

RC 6945 (#29783) 1/17/78
Mathematics 9 pages

LEAST SQUARES AND MAXIMUM LIKELIHOOD ESTIMATES OF RIGID MOTION

by

Arthur Nádas[1]

Scientific Computation Department
IBM System Products Division
Hopewell Junction, New York 12533

Typed by Barbara Corti on CMC (Ad.2059)

ABSTRACT

Let $z_1,...,z_n$ denote vectors of coordinates of n points that describe some object in p-dimensional Euclidean space, $2 \leq p \leq n$. Suppose the object undergoes rigid motion $T=(a,B)$ where a is a p-vector and B is a $p \times p$ orthonormal matrix. The points are measured to be at coordinate vectors $\chi_1,...,\chi_n$: $\chi_\alpha = a+Bz_\alpha+\epsilon_\alpha$ where ϵ_α is an error vector, $\alpha=1,..,n$. We give a simple formula for the least-squares estimate of T. For independent $\{\epsilon_\alpha\}$ identically distributed as p-variate normal $N(0,\Sigma)$, Σ positive definite, we obtain maximum likelihood estimators of T (Σ known) and of (T,Σ) (Σunknown). The distribution of the estimate of T and of prediction error are discussed and in the case $p=2$ asymptotic (large n) distributions are calculated. This problem arose in a production line manufacturing ceramic substrates for silicone chips.

[1]Also at Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

LIMITED DISTRIBUTION NOTICE

This report has been submitted for publication elsewhere and has been issued as a Research Report for early dissemination of its contents. As a courtesy to the intended publisher, it should not be widely distributed until after the date of outside publication.

Copies may be requested from:
IBM Thomas J. Watson Research Center
Post Office Box 218
Yorktown Heights, New York 10598

1. Introduction and Summary

Let $z_1,...,z_n$ denote column vectors of coordinates of n points which describe some object in p-dimensional Euclidean space and define a $p \times n$ matrix Z by $Z=(z_1,...,z_n)$. Suppose that the object undergoes rigid motion and the new coordinates of the points $\chi_1,...,\chi_n$ include errors of measurement $\epsilon_1,...,\epsilon_n$. Let $X=(\chi_1,...,\chi_n)$, $E=(\epsilon_1,...,\epsilon_n)$ be $p \times n$ matrices and let $T=(a,B)$ denote the rigid motion

$$T: \ \chi_\alpha = a + Bz_\alpha + \epsilon_\alpha \quad (\alpha=1,...,n)$$

where a is the p-vector of shifts and B is an orthonormal $p \times p$ matrix. The least squares problem is to find (a,B) which minimizes the Euclidean norm of E $||E|| = ||X-A-BZ|| = (\sum_{\alpha=1}^{n} ||\chi_\alpha - a - Bz_\alpha||^2)^{1/2}$ where $A = (a,...,a)$ is $p \times n$, subject to the constraint that B is orthonormal. The solution $(\hat{a},\hat{B})$ to this problem is

$$(1) \qquad \hat{a} = \bar{\chi} - \hat{B}\bar{z}$$

where $\bar{\chi},\bar{z}$ are the aritmetic averages of $\chi_1,...,\chi_n$ and of $z_1,...,z_n$ respectively. $\hat{B}$ is any solution of

$$(2) \qquad B(C^TC)^{1/2} = C$$

where C is the $p \times p$ matrix (T denotes transpose) of inner products $C=X^*Z^{*T}$, $X^* = X-\bar{X}$, $Z^* = Z-\bar{Z}$, with $\bar{X}=(\bar{\chi},...,\bar{\chi})$ and $\bar{Z}=(\bar{z},...,\bar{z})$ $p \times n$ matrices. For any nonnegative definite M, $M^{1/2}$ denotes the unique nonnegative definite square root. If C has full rank p then

$$(2') \qquad \hat{B} = C(C^TC)^{-1/2} .$$

Suppose now that the ϵ_α are independent and identically distributed random vectors with the common p-dimensional normal distribution $N(0,\Sigma)$ having mean zero and positive definite covariance matrix Σ. If Σ is known then the maximum likelihood problem is to find (a,B) which maximizes the joint density function of the random vectors $\chi_1,\ldots,\chi_n$ (evaluated at the observed values) subject to B orthonormal. The solution $(\hat{a},\hat{B})$ to this problem is

$$\hat{a} = \bar{\chi} - \hat{B}\bar{z} \tag{3}$$

$$\hat{B} = \Sigma^{-1}C(C^T\Sigma^{-2}C)^{-1/2} \,. \tag{4}$$

If Σ is unknown then the problem is to choose (a, B, Σ) to maximize the joint density as above. Solutions $(\hat{a}, \hat{B}, \hat{\Sigma})$ to this problem satisfy the system of equations

$$\hat{a} = \bar{\chi} - \hat{B}\bar{z} \tag{5}$$

$$\hat{B} = \hat{\Sigma}^{-1}C(C^T\hat{\Sigma}^{-2}C)^{-1/2} \tag{6}$$

$$N\hat{\Sigma} = (X^* - \hat{B}Z^*)(X^* - \hat{B}Z^*)^T \,. \tag{7}$$

In Section 2 we obtain these estimators and remark on their relation to the unconstrained linear regression problem. In Section 3 we mention the problem of determining the distribution of the estimators and for $p=2$ we find the asymptotic (large n) distribution of the estimators and of the error in predicting the location of an $(n+1)^{st}$ point on the object. In connection with the latter, we obtain a general formula of some independent interest ([2],[4]) for the distribution of the length of a p-dimensional normal random vector.

2. The Estimators

2.1 The constrained estimator.

It suffices to obtain (6) since (2),(4) follow from (6) and (1),(3),(5),(7) are standard; in particular (7) follows from Lemma 3.2.2 of [1]. The logarithm of the likelihood function is -1/2 times

$$(8) \qquad n\log\det\Sigma + \sum_{\alpha=1}^{n} (\chi_\alpha - a - Bz_\alpha)^T \Sigma^{-1} (\chi_\alpha - a - Bz_\alpha).$$

Forming partial derivatives with respect to the components of a we find (5). Substituting (5) into (8) we see that the problem is to minimize

$$(9) \qquad \Psi(B,\Sigma) = n\log\det \Sigma + \sum_{\alpha=1}^{n} (\chi_\alpha^* - Bz_\alpha^*)^T \quad \Sigma^{-1} \, (\chi_\alpha^* - Bz_\alpha^*),$$

subject to

$$(10) \qquad \sum_{t=1}^{p} b_{ti}\, b_{tj} = \delta_{ij} \qquad i,j=1,\dots,p$$

where $B=(b_{ij})$ and δ_{ij} is the Kronecker delta. Let $\lambda_{ij}=\lambda_{ji}$ be Lagrange multipliers and form $\Lambda = (\lambda_{ij})$. The Lagrangian is then

$$L = \Psi + \sum_{i=1}^{P} \sum_{j=1}^{P} \quad \lambda_{ij}\Big(\sum_{t=1}^{P} b_{ti}b_{tj} - \delta_{ij}\Big).$$

The system of equations $\{\frac{\partial L}{\partial b_{ij}} = 0;\ i,j=1,\dots,p\}$ can be conveniently summarized as

$$(11) \qquad \Sigma^{-1}\,(X^* - BZ^*)\, Z^{*T} + B\Lambda = 0$$

so that

$$(12) \qquad B\Sigma^{-1}X^*Z^{*T} = Z^*Z^{*T} - \Lambda\,.$$

Left multiplying (12) by its transpose we get

$$(13) \qquad Z^*X^{*T}\,\Sigma^{-2}\,X^*Z^{*T} = (Z^*Z^{*T} - \Lambda)^2\,.$$

Taking square roots in (13) and substituting into (12) we obtain (6). Note that the assumptions insure that Prob{C is singular}=0.

2.2 Remarks

It is interesting to compare (2') with the usual (see e.g. Anderson [1]) unconstrained least squares estimate of regression coefficients, to wit: $B_o = CS^{-1}$ where $S = Z^*Z^{*T}$ is assumed nonsingular. Hoffman [3] observed that if among orthonormal matrices B maximizes $\mathrm{tr}(BC^T)$ then and only then B minimizes both $||X^*-BZ^*||$ and $||C-B||$. A simple argument, similar to the one used to derive (2), shows that if M is any nonsingular matrix then $||M-B||$ is minimized among orthonormal B by $B=M(M^TM)^{-1/2}$. This displays, with M=C, the role of C in Hoffman's observation. With $M=B_o$ we see that the best orthonormal approximation $\hat{B}_o$ to B_o is $CS^{-1}\,(S^{-1}C^TCS^{-1})^{-1/2}$, thus the two-stage minimization: $\min||X^*-BZ^*|| = ||X^*-B_oZ^*||$ followed by $\min||B_o-B|| = ||B_o-\hat{B}_o||$ will not generally produce the least-squares solution.

3. Distributions of Estimators and Prediction Error.

3.1 The general case

When exact distributions are difficult to find the usual (large sample) approach for approximating the distribution of maximum likelihood estimators is to check that certain regularity conditions are satisfied and then to use the normal approximations available in the regular case (see e.g. Zacks [5]). One of these conditions is that the space of the k-dimensional parameter contain a k-dimensional rectangle. This condition is not met in our case because of the orthonormality constraint on the matrix B which has only $r=\binom{p}{2}$

$<p^2 = k$ degrees of freedom (orthogonalizing and normalizing column by column we count $(p-1)+(p-2)+\ldots+1+0)$. A possible approach is to reparametize B with an r dimensional parameter, for example by

$$(14) \qquad B = \prod_{t=1}^{r} R_t\,(\Theta_t)$$

where $R_t(\chi)$ is the rotation matrix which rotates in the t^{th} of the r planes available. The space of $(\Theta_1,\ldots,\Theta_r)$ will be the rectangle $[0,2\pi]^r$. We have not carried out this program in general but we did so in the case $p=2$ with circular normal error which follows.

3.2 Distribution of estimates when p=2.

The model is $\chi_\alpha = a + Bz_\alpha + \epsilon_\alpha$ i.e.

$$(15) \qquad \begin{pmatrix} \chi_{1\alpha} \\ \chi_{2\alpha} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} \cos\Theta & \sin\Theta \\ -\sin\Theta & \cos\Theta \end{pmatrix} \begin{pmatrix} z_{1\alpha} \\ z_{2\alpha} \end{pmatrix} + \begin{pmatrix} \epsilon_{1\alpha} \\ \epsilon_{2\alpha} \end{pmatrix}$$

for $\alpha=1,\ldots,n$. For simplicity we assume $\Sigma = \sigma^2 I$, I is the identity, $\sigma>0$, ϵ_α are i.i.d. $N(0,\Sigma)$. A little manipulation of the logarithmic derivatives of the likelihood function yield the explicit estimates

$$\hat{a}_1 = \overline{\chi}_1 - \overline{z}_1 \cos\hat{\Theta} - \overline{z}_z \sin\hat{\Theta}$$

$$\hat{a}_2 = \overline{\chi}_2 + \overline{z}_1 \sin\hat{\Theta} - \overline{z}_2 \cos\hat{\Theta}$$

$$\tan\hat{\Theta} = \frac{\overline{\chi_1^* z_2^* - x_2^* z_1^*}}{\overline{\chi_1^* z_1^* + \chi_2^* z_2^*}}$$

$$2\hat{\sigma}^2 = \overline{(\chi_1^* - z_1^*\cos\hat{\Theta} - z_2^* \sin\hat{\Theta})^2} + \overline{(\chi_2^* + z_1^* \sin\hat{\Theta} - z_2^*\cos\hat{\Theta})^2}$$

where the quadrant of $\hat{\Theta}$ is determined by sgn $\cos\hat{\Theta}$ = sgn $\overline{\chi_1^* z_1^* + \chi_2^* z_2^*}$

(For any symbol u, $\overline{u} = n^{-1} \sum_{\alpha=1}^{n} u_\alpha$.)

The Fisher information matrix M of the vector parameter $(\beta_1,\beta_2,\beta_3,\beta_4) \equiv (a_1,a_2,\Theta,\sigma)$ is obtained as the matrix of the negatives of the expected values of the mixed second partial logarithmic derivatives of the likelihood function evaluated the random argument. Inverting this matrix we obtain the covariance matrix of the asymptotic normal approximation as

$$(16) \qquad \sigma_\beta = M^{-1} = \frac{\sigma^2}{ns^2} \begin{array}{cccc} \overline{||z||^2} - d^2 & cd & -c & 0 \\ cd & \overline{||z||^2} - c^2 & -d & 0 \\ -c & -d & 1 & 0 \\ 0 & 0 & 0 & \frac{s^2}{4} \end{array}$$

where $s^2 = \overline{||z-\overline{z}||^2}$, $C = \overline{z}_2 \cos\Theta - \overline{z}_1 \sin\Theta$ and $-d = \overline{z}_1\cos\Theta + \overline{z}_2 \sin\Theta$. In practice one estimates Σ_β by using $(\hat{\sigma},\hat{\Theta})$ in place of (σ,Θ).

3.3 The length of a normal vector

The following form of the distribution of the length of a normal vector, to be used to find the distribution of prediction errors, is of some independent interest in light of the difficulty of the general problem ([2],[4]), even though it is a trivial generalization of commonly known ([1]) results. Let X be p-variate normal $N(\mu,\Sigma)$ with Σ positive definite. Then

$$\text{Prob}\{\,||X|| \le \xi\,\} = F_1 * \ldots * F_p\,(\xi^2) \tag{17}$$

where * denotes convolution and for $i=1,\ldots,p$

$$F_i(t) = \text{Prob}\{\chi^2_{1,\eta_i^2} \le t/\lambda_i\}.$$

The constants λ_i are the characteristic values of Σ and $\eta = A\Sigma^{-1/2}\mu$ ($\eta^T = (\eta_1,\ldots,\eta_p)$ where the columns of A^T are the corresponding normalized characteristic vectors of Σ. χ^2_{1,η_i^2} is noncentral Chi-squared with one degree of freedom. The result is immediate from the representation

$$X = \mu + \Sigma^{1/2} Z$$

where Z is N(0,I), and from the spectral decomposition ($AA^T = I$)

$$\Sigma = A^T \operatorname{diag}(\lambda_1,\ldots,\lambda_p)A.$$

One writes

$$\begin{aligned} ||X||^2 \equiv X^TX = Y^T\Sigma Y &= Y^{*T} \operatorname{diag}(\lambda_1,\ldots,\lambda_p)Y^* \\ &= \sum_{i=1}^{p} \lambda_i\, Y_i^{*2} \end{aligned} \tag{18}$$

where $Y^* = AY$ is distributed as $N(A\Sigma^{-1/2}\mu, AIA^T) = N(\eta;I)$.

3.4 Prediction error when p=2.

Suppose that given $z_1,\ldots,z_n$ and $\chi_1,\ldots,\chi_n$ we have obtained the estimators as in Section 3.2. Given any $n+1^{st}$ point z_o we predict the new position $\chi_o = a+Bz_o$ of the point by $\hat{\chi}_o = \hat{a}+\hat{B}z_o$. To find the asymptotic distribution of the error $||e|| = ||\hat{\chi}_o-\chi_o||$ we may approximate e as (with $B' = (\frac{d}{d\Theta}B_{ij})$)

$$\hat{a}-a + (\hat{B}-B)z_o \doteq \hat{a}-a - (\hat{\Theta}-\Theta)\, B'\, z_o \tag{19}$$

on account of the continuity of e in Θ and of the consistency of $\hat{\Theta}$. This yields the following asymptotic covariance matrix for e

$$\Sigma_e = n^{-1}\, \sigma^2\, (I + \delta\delta^T) \tag{20}$$

where $\delta = B'(z_o-\bar{z})$ and I is the identity. In practice one uses $(\hat{\sigma},\hat{\Theta})$ in place of (σ,Θ) to evaluate (20). The characteristic values of $\frac{n}{\sigma^2}\Sigma_e$ are 1 and $\lambda = ||\delta||^2 + 1$ thus, (using (18)) we have

$$||e||^2 = \frac{\sigma^2}{n}(Z_1^2 + \lambda\, Z_2^2) \tag{21}$$

where the Z_i are i.i.d. standard normal variables. The distribution function of (21) is easy to compute; in particular the mean squared error of prediction is

$$E||e||^2 = n^{-1}\sigma^2\,(1+\lambda) = n^{-1}\sigma^2\,(2+ \frac{||z_o-\bar{z}||^2}{s^2})$$

where, as in section 3.2, $s^2 = \overline{||z-\bar{z}||^2}$.

References

[1] Anderson, T. W. (1958), Introduction to Multivariate Statistical Analysis, J. Wiley, New York.

[2] Blumenson, L. E. and Miller, K. S. (1963), Properties of generalized Raleigh distributions, Ann. Math. Stat., 34, 903-910.

[3] Hoffman, A. J. (1977), Private communication.

[4] Miller, K. S. (1964), Multidimensional Gaussian Distributions, J. Wiley, New York.

[5] Zacks, S. (1971), The Theory of Statistical Inference, J. Wiley, New York.