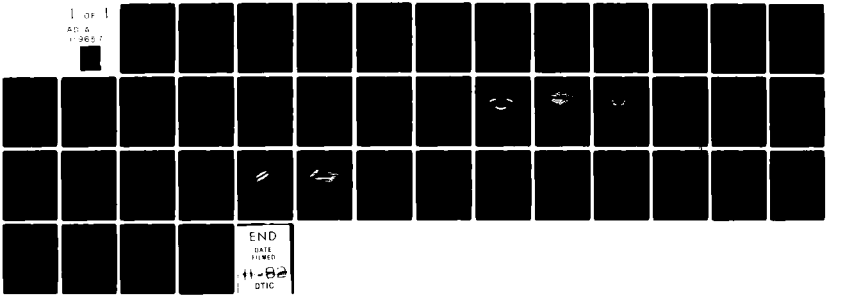
AD-A119 657 RENSSELAER POLYTECHNIC INST TROY NY F/G 12/1
SOME MODIFIED INTEGRATED SQUARED ERROR PROCEDURES FOR MULTIVARI--ETC(U)
JUN 82 A S PAULSON, C E LAWRENCE DAAG29-81-K-0110
UNCLASSIFIED A-2 ARO-18072.2-MA NL



ARO 18072.2-MA

(12)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

REPORT NUMBER: A-2	2. GOVT ACCESSION NO.: AD-A119 657
3. RECIPIENT'S CATALOG NUMBER	
TITLE (and Subtitle): SOME MODIFIED INTEGRATED SQUARED ERROR PROCEDURES FOR MULTIVARIATE NORMAL DATA	5. TYPE OF REPORT & PERIOD COVERED: Interim Technical Report
	6. PERFORMING ORG. REPORT NUMBER: A-2
AUTHOR(s): A. S. Paulson, C. E. Lawrence	8. CONTRACT OR GRANT NUMBER(s): DAA G29-81-K-0110
PERFORMING ORGANIZATION NAME AND ADDRESS: Rensselaer Polytechnic Institute, Troy, New York 12181	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
CONTROLLING OFFICE NAME AND ADDRESS: Approved for public release; distribution unlimited.	12. REPORT DATE: 1 June 1982
	13. NUMBER OF PAGES: 40
MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Department of the Navy, Office of Naval Research, 715 Broadway (5th Floor), New York, New York 10003	15. SECURITY CLASS. (of this report)
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

DTIC ELECTE SEP 28 1982 S D H

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

multivariate normal, sensitivity analysis, M-estimators, parametric density estimation, adaptive estimation, two-way cross classification

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A method of estimation for the parameters of the multivariate normal distribution based on the characteristic function (density) and its sample counterpart is given. These M-estimators are dependent on a user-specified parameter. The response of the parameter estimates and observation weights to variation of this user-specified parameter allows a sensitivity analysis of the data and the model considered as a single entity. The estimators have desirable robustness properties, are easy to compute and use, are relatively efficient at the multivariate normal and are useful in idendifying potential

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (cont'd)

outliers and problems with the statistical assumptions or the data. The method is extended to include multivariate experimental designs with attention restricted to the two-way cross classification. Several illustrations are provided.

SOME MODIFIED INTEGRATED SQUARED ERROR PROCEDURES FOR MULTIVARIATE NORMAL DATA

by

A. S. Paulson*
Rensselaer Polytechnic Institute

and

C. E. Lawrence
New York State Health Department

*Research supported in part by U.S. Army Research Office under contract DAA G29-81-K-0110

Summary

A method of estimation for the parameters of the multivariate normal distribution based on the characteristic function 'density) and its sample counterpart is given. These M-estimators are dependent on a user-specified parameter. The response of the parameter estimates and observation weights to variation of this user-specified parameter allows a sensitivity analysis of the data and the model considered as a single entity. The estimators have desirable robustness properties, are easy to compute and use, are relatively efficient at the multivariate normal and are useful in identifying potential outliers and problems with the statistical assumptions or the data. The method is extended to inc' ıde multivariate experimental designs with attention restricted to the two-way cross classification. Several illustrations are provided.

Key Words: multivariate normal, sensitivity analysis, M-estimators, parametric density estimation, adaptive estimation, two-way cross classification

1. Introduction

A large number of robust procedures are available for univariate normal data. The number of procedures for the multivariate normal distribution are fewer in number. The recent books by Huber (1981), Barnett and Lewis (1978), and Gnanadesikan (1977) provide excellent reviews and discussions of a major portion of the literature.

The weighted and integrated distance between the assumed distribution function and its empirical counterpart has recently been used by Parr and Shucany (1980) to produce attractive robust estimators in the univariate case. The extension of such procedures to the multivariate case will, however, present computational difficulties which may be insuperable. This is not the case with the estimators for location and covariance matrix that we propose. The univariate case was considered by Paulson and Nicklin (1981). Simultaneous estimators for location and the covariance matrix are determined from consideration of the sum of integrated residuals squared where the residuals are defined as the difference between a Gaussian characteristic function and its empirical counterpart . This sum is equivalent to the sum of integrated squared differences between a Gaussian density and its empirical counterpart. The approach that we propose is equivalent to one of parametric density estimation, but is quite different in spirit from the work of Parzen (1962).

The estimators we develop depend on a user chosen parameter λ (or parameters λ_{jk}). Variation of this parameter from large values (most efficient) to successively smaller values allows the user to determine the response in the parameter values to this variation. If the data

$x_1, x_2, \ldots, x_n$ and the Gaussian model are internally consistent, then a flat response surface will result. If the model and data are not internally consistent, then the response surface will not be flat and this will signify potential difficulties in the data or with the Gaussian model or both. For each observation x_j a weight $\tilde{v}_{j\lambda}$ is determined. The variation in the $\tilde{v}_{j\lambda}$ as a function of λ is useful in determining the weakest interfaces between the data and the Gaussian model. Accordingly, an examination of the response surface of the $\tilde{v}_{j\lambda}$ as a function of λ is useful in identifying potential outliers. We envision that the primary use for our procedure will be in performing sensitivity analyses. Secondarily, a robust procedure results if λ is chosen as fixed in the range $-\frac{1}{2} < \lambda < \infty$. This resultant robust procedure is of a somewhat different character from those discussed by Maronna (1976) and Devlin et al. (1975).

Detailed statistical properties of the modified integrated weighted squared error are given. An extension of the procedure to examine univariate problems and to the analysis of a two-way layout of multivariate data is also given. It is indicated that the procedure can be used in a clustering context. Several examples are provided.

2. The Estimation Procedure

The χ^2 minimum procedure consists of determining estimators of a set of parameters $\theta_1, \theta_2, \ldots, \theta_s$ by minimizing

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} \tag{2.1}$$

with respect to the θ's. The $p_i = p_i(\theta_1, \theta_2, \ldots, \theta_s)$ and ν_i represents the number of observations which fall in cell i where the k cells constitute a mutually exclusive and exhaustive partitioning of the sample

space of the random sample $x_1, x_2, \ldots, x_n$ from a density $f(x;\theta_1,\theta_2\ldots,\theta_s)$. The modified χ^2 minimum procedure consists of determining estimators of $\theta_1,\theta_2,\ldots,\theta_s$ from the system

$$\sum_{i=1}^{k} \frac{(\nu_i - np_i)}{p_i} \frac{\partial p_i}{\partial \theta_j} = 0, \tag{2.2}$$

namely the equations which result from differentiation of (2.1) with respect to θ while regarding the denominator as constant (Cramér, 1946, pp. 424-428). The method we propose for the multivariate Gaussian is entirely analogous.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from the p-dimensional Gaussian distribution with p×1 mean vector μ and p×p covariance vector V. The density is

$$f(x) = f(x;\mu,V) = \frac{|V|^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}p}} \exp(-\tfrac{1}{2}(x-\mu)^T V^{-1}(x-\mu)) \tag{2.3}$$

and corresponding characteristic function

$$\phi(u) = \phi(u;\mu,V) = \exp(i\mu^T u - \tfrac{1}{2} u^T V u) \tag{2.4}$$

where $u^T = (u_1,u_2,\ldots,u_p)$ is a 1×p vector of real numbers. We shall generally suppress the arguments μ and V of $f(x)$ and $\phi(u)$. Define

$$\Delta = \Delta_{n,\omega}(\mu,V) = \sum_{j=1}^{n} \int_{R_p} |\phi(u) - \exp iux_j|^2 |\omega(\phi(u))|^2 du \tag{2.5}$$

where $\omega(\cdot)$ is a function to be chosen. The quantity Δ represents a sum of integrated weighted squared residuals. We shall designate the solutions for μ and V determined from the system

$$2 \text{ Re} \sum_{j=1}^{n} \int_{R_p} \frac{\partial \phi(u)}{\partial \mu} (\phi(u) - \text{expiux}_j)^* |\omega(\phi(u))|^2 \, du = 0, \qquad (2.6)$$

$$2 \text{ Re} \sum_{j=1}^{n} \int_{R_p} \frac{\partial \phi(u)}{\partial V} (\phi(u) - \text{expiux}_j)^* |\omega(\phi(u))|^2 \, du = 0, \qquad (2.7)$$

as modified integrated weighted squared error estimators. The dimension of the 0's in (2.6) and (2.7) are defined from context. The notation '*' denotes complex conjugate. The equations (2.6) and (2.7) are completely analogous to the modified χ^2 minimum equations of (2.3). The estimators of μ and V determined from (2.6) and (2.7) are M-estimators and are affine invariant. If $\omega(\phi(u))$ is itself a Fourier transform with inverse $f_\omega(x)$, then (2.6) and (2.7) admit of a representation in terms of densities. By Parseval's theorem (2.6) and (2.7) have equivalent expressions, respectively

$$2(2\pi)^p \sum_{j=1}^{n} \int_{R_p} \left(\frac{\partial f(x)}{\partial \mu} \overset{x}{*} f_\omega(x)\right)\left(f(x) \overset{x}{*} f_\omega(x) - f_\omega(x-x_j)\right) dx = 0, \qquad (2.8)$$

$$2(2\pi)^p \sum_{j=1}^{n} \int_{R_p} \left(\frac{\partial f(x)}{\partial V} \overset{x}{*} f_\omega(x)\right)\left(f(x) \overset{x}{*} f_\omega(x) - f_\omega(x-x_j)\right) dx = 0, \qquad (2.9)$$

where the symbol $\overset{x}{*}$ represents convolution of the density $f(x)$ with the function $f_\omega(x)$. It is important to note that $\frac{\partial}{\partial V}(f(x) \overset{x}{*} f_\omega(x)) \neq \frac{\partial f(x)}{\partial V} \overset{x}{*} f_\omega(x)$. Expressions (2.8) and (2.9) show that $\omega(\phi(u)) = \exp(-\frac{1}{2} u^T(\lambda V)u)$ provides an attractive choice since then $f_\omega(x)$ is a Gaussian density with mean vector 0 and variance-covariance matrix λV, λ a constant scalar. The convolution of a Gaussian density with mean μ and variance-covariance matrix V with a Gaussian density with mean 0 and variance-covariance matrix λV is again Gaussian but now with mean μ and

variance-covariance matrix $(1+\lambda)V$. This choice for $\omega(\phi(u))$ leads to attractive computational properties as well as some attractive statistical properties.

Other attractive choices of $\omega(\phi(u))$ are easily found. For example, the choice $\omega(\phi(u))=(1-|\phi(u)|^2)^{-\frac{1}{2}}$ effectively makes (2.6) and (2.7) an approximate modified integrated correlated χ^2 minimum procedure since, for every fixed u, $\sum_j (\phi(u) - \exp iux_j)$ is asymptotically complex Gaussian. This choice leads to more complicated numerical algorithms but more efficiency.

The most important aspect of equations (2.6) - (2.9) is that they show that the characteristic function procedure we are suggesting is generally equivalent to a multivariate parametric density estimation procedure since $n^{-1} \sum_{j=1}^{n} f_\omega(x-x_j)$ in (2.8) and (2.9), is an unbiased kernel density estimator for $f(x) \overset{x}{*} f_\omega(x)$. Thus the estimation procedure involves a reconstruction of the smoothed-by-$f_\omega(x)$ error density. The density $f(x)$ is assumed a priori while the estimate $n^{-1} \sum f_\omega(x-x_j)$ is a kind of posterior estimate based on the kernel $f_\omega(x)$ and the data.

We shall be primarily concerned with the choice

$$\omega(\phi(u)) = \exp(-\tfrac{1}{2} u^T (\lambda V)u) \tag{2.10}$$

in this and the next section. It may be helpful to observe that differentiation of

$$\Delta_D = \sum_{j=1}^{n} \int_{R_p} |\phi(u) - \exp iux_j|^2 \exp(-u^T Du) \tag{2.11}$$

with respect to μ and V with subsequent setting of $D = \lambda V$ produces equations (2.6) and (2.7) under the choice (2.10) of $\omega(\cdot)$. The use of this observation,

although unnecessary, allows for a somewhat simpler derivation of the estimators; see Paulson and Nicklin (1981) for the univariate version. The integral Δ_D may be explicitly integrated to give

$$\Delta_D = \int_{R_p} |\phi_n(u)|^2 \exp(-u^T D u)du$$

$$- \frac{2}{n} \frac{(2\pi)^{\frac{1}{2}p}}{|V+2D|^{\frac{1}{2}}} \sum_{j=1}^{n} \exp\{- \tfrac{1}{2} (x_j-\mu)^T (V+2D)^{-1} (x_j-\mu)\}$$

$$+ \frac{\pi^{\frac{1}{2}p}}{|V+D|^{\frac{1}{2}}} .$$

We thus find (Dwyer, 1967)

$$\frac{\partial \Delta}{\partial \mu} = - \frac{2}{n} \frac{(2\pi)^{\frac{1}{2}p}}{|V+2D|^{\frac{1}{2}}} \sum_{j=1}^{n} (V+2D)^{-1} (x_j-\mu) \exp(- \tfrac{1}{2} Q_j(D)) = 0, \tag{2.12}$$

where

$$Q_j(D) = (x_j-\mu)^T (V+2D)^{-1} (x_j-\mu). \tag{2.13}$$

We use the right hand side of (2.12) to generate an estimating equation by setting $D = \lambda V$; then the estimator for the mean satisfies the implicit equation

$$\mu = \frac{\sum_{j=1}^{n} x_j \exp(- \tfrac{1}{2} Q_j(\lambda V))}{\sum_{j=1}^{n} \exp(- \tfrac{1}{2} Q_j(\lambda V))}. \tag{2.14a}$$

The simultaneous implicit matrix equation for the covariance matrix V is determined from (Dwyer, 1967)

$$\frac{\partial \Delta}{\partial V} = \frac{1}{n} \frac{(2\pi)^{\frac{1}{2}p}}{|V+2D|^{\frac{1}{2}}} (V+2D)^{-1} \sum_{j=1}^{n} \exp(-\tfrac{1}{2} Q_j(D))$$

$$- \frac{1}{n} \frac{(2\pi)^{\frac{1}{2}p}}{|V+2D|^{\frac{1}{2}}} (V+2D)^{-1} \sum_{j=1}^{n} (x_j-\mu)(x_j-\mu)^T (V+2D)^{-1} \exp(-\tfrac{1}{2} Q_j(D))$$

$$- \frac{1}{n} \frac{(2\pi)^{\frac{1}{2}p}}{|2V+2D|^{\frac{1}{2}}} \sum_{j=1}^{n} (2V+2D)^{-1} = 0. \qquad (2.15)$$

On pre- and post-multiplying (2.15) by (V+2D) and then setting $D = \lambda V$ we obtain

$$\sum_{j=1}^{n} \{V \exp(-\tfrac{1}{2} Q_j(\lambda V) - \frac{1}{1+2\lambda} (x_j-\mu)(x_j-\mu)^T \exp(-\tfrac{1}{2} Q_j(\lambda V))$$

$$- \left(\frac{1+2\lambda}{2+2\lambda}\right)^{\frac{1}{2}(p+2)} V\} = 0 \qquad (2.16)$$

whence V satisfies, in conjunction with (2.14a), the implicit equation

$$V = (1+2\lambda)^{-1} \frac{\sum_{j=1}^{n} (x_j-\mu)(x_j-\mu)^T \exp(-\tfrac{1}{2} Q_j(\lambda V))}{\sum_{j=1}^{n} \left[\exp(-\tfrac{1}{2} Q_j(\lambda V) - \left(\frac{1+2\lambda}{2+2\lambda}\right)^{\frac{1}{2}(p+2)}\right]}. \qquad (2.14b)$$

The implicit relationships given in (2.14) suggest that the estimators $\tilde{\mu}$ and $\tilde{V}$ be computed via a fixed point algorithm with $\bar{x} = n^{-1} \sum x_j$ and $S = n^{-1} \sum (x_j-\bar{x})(x_j-\bar{x})^T$ supplying the initial guesses μ_0 and V_0 respectively. Substitute these initial guesses into the right hand side of (2.14). New estimates μ_1 and V_1 are obtained from the left hand side. The new estimates μ_1 and V_1 are now substituted into the right hand side of (2.14) and the process is continued until some

pre-specified absolute or relative tolerance between successive estimates is satisfied. We have found this fixed point algorithm to be effective in a variety of practical problems and computer simulations. We have not found it necessary to use second order search methods. This is a real advantage when the dimension p is large. The question of the values of λ which will be useful in practice is addressed in the sequel.

3. Properties of the Estimators

The estimators for μ and V given by (2.10), say $\tilde{\mu}$ and $\tilde{V}$, are M-estimators as is evident from the estimating equations. They are explicitly dependent on the user-chosen parameter λ, even though this dependence on λ will generally be suppressed for convenience.

The estimators $\tilde{\mu}$ and $\tilde{V}$ are well-defined for $\lambda > -\frac{1}{2}$. Modification of the arguments of Bryant and Paulson (1979) shows that these estimators are consistent for μ and V for $\lambda > -\frac{1}{2}$ when the x_j constitute a random sample from $N_p(\mu,V)$. It is obvious from (2.14) that $\lim_{\lambda\to\infty} \tilde{\mu} = \hat{\mu} = n^{-1} \sum x_j = \bar{x}$, the usual method of moments or maximum likelihood estimator of μ. If $y_1, y_2, \ldots, y_n$ is a random sample from $N_p(\mu_y,V_y)$, and for any nonsingular matrix A, $x_j = a + A\, y_j$, $j = 1, 2, \ldots, n$, then $\tilde{\mu}_x = a + A\tilde{\mu}_y$ and $\tilde{V}_x = A\, \tilde{V}_y\, A^T$.

The asymptotic variance-covariance matrices of the estimators $\tilde{\mu}$ and $\tilde{V}$ depends on the user-specified parameter λ. From (2.12) we find that the score function for μ is

$$s_\mu = (x_j-\mu) \exp(-\tfrac{1}{2} Q(D))\Big|_{D=\lambda V} \tag{3.1}$$

where $Q(D) = (x-\mu)^T(V+2D)^{-1}(x-\mu)$. By a standard Taylor's series expansion the quality $n^{\frac{1}{2}}(\tilde{\mu}-\mu)$ is asymptotically p-variate normal $N_p(0,\sum_\mu)$ where, for easily computed expectations,

$$\begin{aligned}\operatorname{cov}(\tilde{\mu}) = \sum_\mu &= \left|E\left(\frac{\partial s_\mu}{\partial \mu^T}\right)\right|^{-1} E(s_\mu s_\mu{}^T) \left|E\left(\frac{\partial s_\mu}{\partial \mu^T}\right)\right|^{-1} \Big|_{D=\lambda V} \\ &= \{-(1+c)^{-\frac{1}{2}(p+2)} I\}^{-1}\{(1+2c)^{-\frac{1}{2}(p+2)} V\}\{-(1+c)^{-\frac{1}{2}(p+2)} I\}^{-1} \\ &= \left(\frac{1+2c+c^2}{1+2c}\right)^{\frac{1}{2}(p+2)} V = \left(\frac{4+8\lambda+4\lambda^2}{3+8\lambda+4\lambda^2}\right)^{\frac{1}{2}(p+2)} V, \qquad (3.2)\end{aligned}$$

where $c = (1+2\lambda)^{-1}$ and I is the p×p identity matrix. The asymptotic efficiency of $\tilde{\mu}$ relative to $\hat{\mu}$ is the ratio of determinants of covariance matrices, namely

$$e(\tilde{\mu}) = \left(\frac{3+8\lambda+4\lambda^2}{4+8\lambda+4\lambda^2}\right)^{\frac{1}{2}p(p+2)}. \qquad (3.3)$$

Table 1 gives a short listing of the pth root of these efficiencies.

Table 1

pth Root of Asymptotic Relative Efficiency of $\tilde{\mu}$

	λ = 0	1	2	4	∞
1	.65	.91	.96	.99	1
2	.56	.88	.95	.98	1
p 3	.49	.85	.93	.98	1
4	.42	.82	.92	.97	1
8	.24	.72	.87	.95	1

The efficiency of $\tilde{\mu}$ declines rapidly with increasing dimension for low values of λ. Efficiency of $\hat{\mu}$ increases with increasing λ as $\tilde{\mu}$ becomes increasingly like $\bar{x}$.

The asymptotic variance-covariance matrix of the $\frac{1}{2}p(p+1)$ estimators $\tilde{v}_{jk}$, $j \le k$ of $\tilde{V}$ is considerably more difficult to obtain. The covariances $\mathrm{cov}(\tilde{v}_{jk}, \tilde{v}_{\ell m})$ can be determined from the score functions determined from (2.12). For example, from (2.16) we find that the score function for v_{kk} is

$$s_{v_{kk}} = v_{kk} \exp(-\tfrac{1}{2} Q(\lambda V)) - c(x_k - \mu_k)^2 \exp(-\tfrac{1}{2} Q(\lambda V)) - \left(\frac{1+2\lambda}{2+2\lambda}\right)^{\frac{1}{2}(p+2)} v_{kk}$$

where x_k is the kth component of the vector x and $c=(1+2\lambda)^{-1}$. The asymptotic variance of $\tilde{v}_{kk}$ is

$$\left| E\left(\frac{\partial s_{v_{kk}}}{\partial v_{kk}}\right)\right|^{-2} E(s^2_{v_{kk}}).$$

Straightforward, but tedious, computations gives the asymptotic efficiency of $\tilde{v}_{kk}$ relative to the maximum likelihood estimator $\hat{v}_{kk}$ as

$$e(\tilde{v}_{kk}) = \frac{9}{2} \frac{\left(\frac{1}{1+2\lambda}\right)^2 \left(\frac{1+2\lambda}{2+2\lambda}\right)^{p+4}}{\left(\frac{1+2\lambda}{3+2\lambda}\right)^{\frac{1}{2}p} \frac{6+8\lambda+4\lambda^2}{(3+2\lambda)^2} - \left(\frac{1+2\lambda}{2+2\lambda}\right)^{p+2}}. \qquad (3.4)$$

Selected values of these efficiencies are given in Table 2. The efficiencies of $\tilde{v}_{jk}$ are about the same magnitude as those for $\tilde{v}_{kk}$. It would be desirable to have the efficiencies of the matrix estimator $\tilde{V}$ relative to $\hat{V}$, the maximum likelihood estimator, but these values would be troublesome to obtain and would not be much different from (3.4).

Table 2

Asymptotic Efficiencies of the Estimators v_{kk}

		λ				
		.5	1	2	4	∞
p	1	.78	.87	.94	.98	1
	2	.68	.77	.84	.88	.90
	3	.59	.69	.76	.80	.82
	4	.52	.62	.69	.73	.75
	5	.46	.56	.63	.67	.69

In the one dimensional case, $e(\tilde{v}_{kk})$ tends to unity with increasing λ. In the $p>1$ dimensional case the efficiencies $e(\tilde{v}_{kk})$ are bounded away from unity. The estimators $\tilde{V}$ thus have a different character in the cases $p=1$ and $p>1$. The efficiencies of $\tilde{v}_{kk}$ for fixed λ decline rapidly with increasing dimensionality. This implies that the higher the dimensionality, the larger the values of λ one should like to use if efficiency is a major consideration in the choice of λ.

The nature of $\tilde{V}$ as $\lambda\to\infty$ may be determined by an application of L'Hospital's rule. Some elementary manipulations yield the implicit equation

$$V = \frac{\sum_{j=1}^{n} (x_j-\mu)(x_j-\mu)^T}{\tfrac{1}{2}\, n(p+2) - \tfrac{1}{2} \sum_{j=1}^{n} (x_j-\mu)^T V^{-1}(x_j-\mu)} \tag{3.5}$$

that the estimator V must satisfy asymptotically in λ. When $p=1$ (3.6) may be rearranged to give the usual maximum likelihood or moment estimator for V. But for $p\geq 2$ the estimator cannot be so arranged; thus the

efficiencies in Table 2 are bounded away from unity when $p>1$ and $\lambda=\infty$. The curious estimator of V defined implicitly in (3.6) does not seem to be especially interesting.

The estimators $\tilde{v}_{11}, \tilde{v}_{12}, \ldots, \tilde{v}_{1p}, \tilde{v}_{22}, \ldots, \tilde{v}_{pp}$ are asymptotically $\frac{1}{2}\, p(p+1)$ variate Gaussian and are asympotically independent of $\tilde{\mu}$ which is asymptotically p-dimensional Gaussian.

There are a number of measures of qualitative robustness but the most important is the influence function. Most of the other measures are derived from the influence function. The influence function is simply proportional to the score function (Huber, 1981, p. 45). The influence function at the p-variate Gaussian distribution $N_p(\mu,V)$ is

$$IC(x;\tilde{\mu},N) = \left| E\left(\frac{\partial s_\mu}{\partial \mu^T}\right)\right|^{-1} s_\mu \Big|_{D=\lambda V}$$

$$= (1+c)^{\frac{1}{2}(p+2)}(x-\mu)\, \exp(-\frac{c}{2}\,(x-\mu)^T V^{-1}(x-\mu)). \qquad (3.6)$$

This function is bounded and redescending to zero for all $-\frac{1}{2} < \lambda < \infty$. It also shows that the assumed Gaussian distribution is playing an adaptive role in the estimation of μ. Furthermore, the component x_1 of the vector x plays a role in the estimation of the components $\mu_2, \mu_3, \ldots, \mu_p$ of μ as well as μ_1. Figures 1 provide contours of this influence function for $\tilde{\mu}$ for several values of λ and correlation ρ at the standard bivariate Gaussian distribution. The forms of these contours suggest that the procedure will adaptively cluster the observations assumed to follow a Gaussian parent according to those which belong to the parent and those which do not - provided λ is small enough.

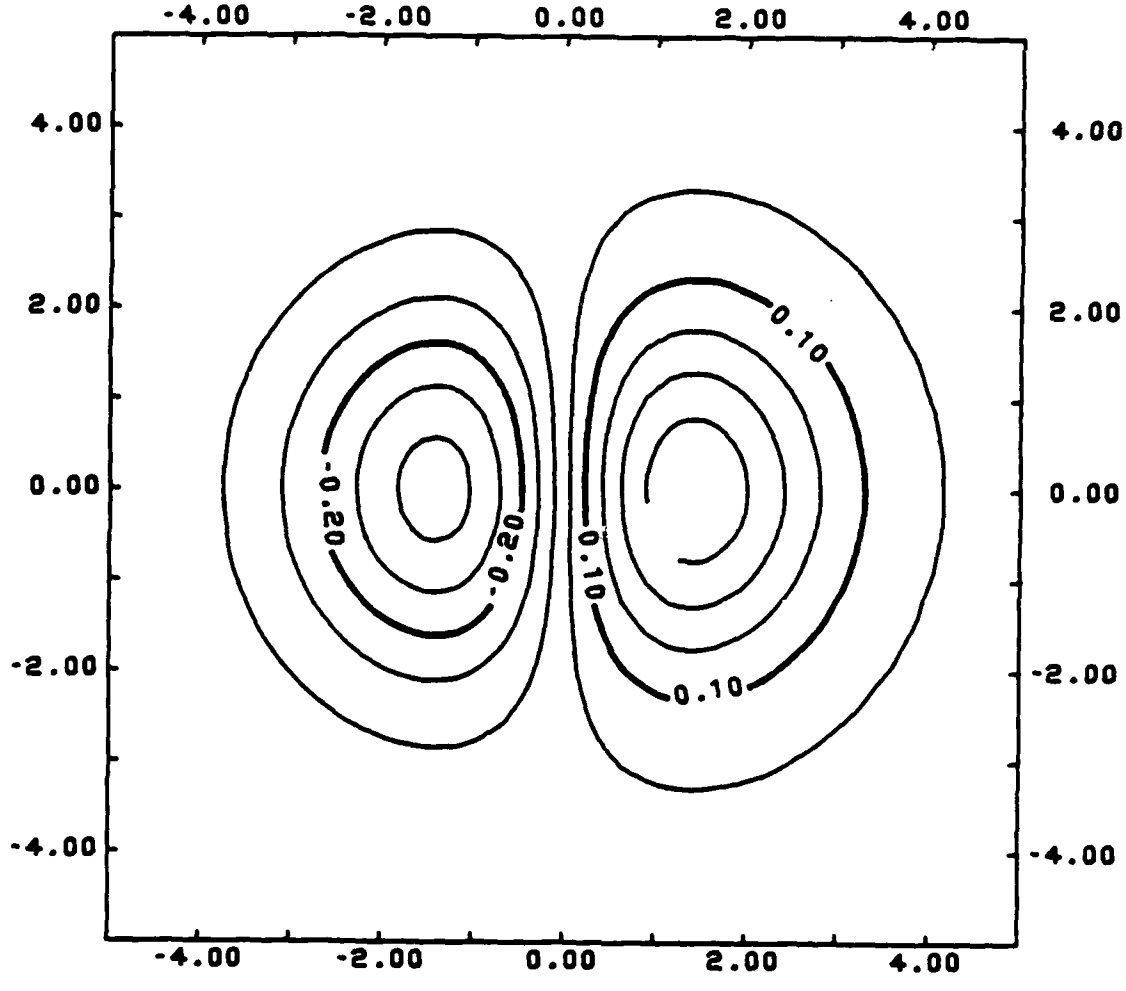


Figure 1a. Influence function contours for $\tilde{\mu}_1$ at the standard bivariate normal, $\rho=0$, $\lambda=.5$

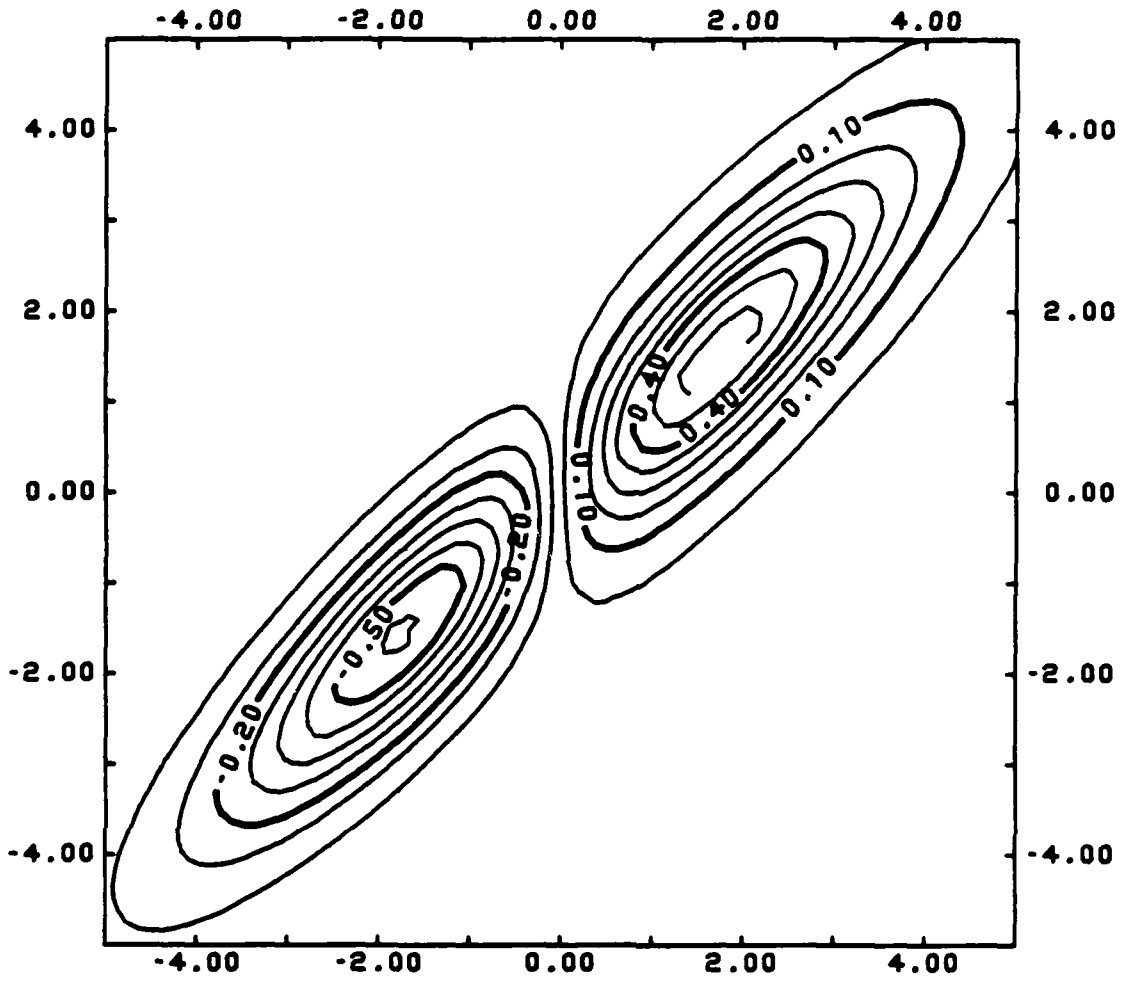


Figure 1b. Influence function contours for $\tilde{\mu}_1$ at the standard bivariate normal, $\rho=.9$, $\lambda=1$

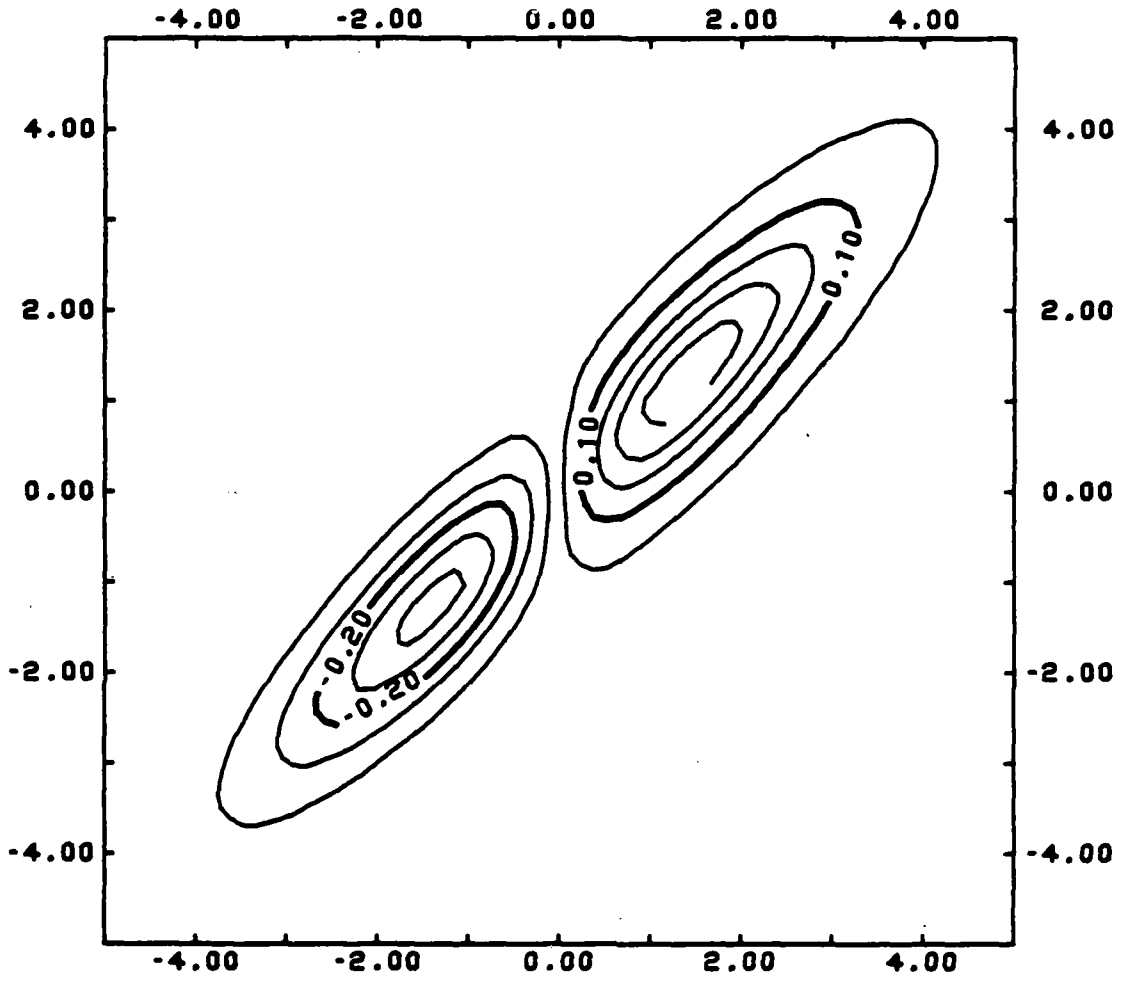


Figure 1c. Influence function contours for $\tilde{\mu}_1$ at the standard bivariate normal, $\rho=.9$, $\lambda=.5$

The influence function for $\tilde{V}$ at $N_p(\mu,V)$ is more difficult to obtain, at least computationally. For the j,k element of (2.12), we note that the partial derivative of this element with respect to V is a p×p matrix and hence differentiation of (2.12) with respect to V is facilitated by the introduction of tensors. We need not go to such lengths. The shape of the influence function is the important property and it is determined by the score function s_V. The influence function for $\tilde{V}$ at $N_p(\mu,V)$ has shape determined by

$$s_V = c(x-\mu)(x-\mu)^T \exp(-\tfrac{1}{2} Q(\lambda V)) + (1+c)^{-\frac{1}{2}(p+2)} V - V \exp(-\tfrac{1}{2} Q(\lambda V)) \tag{3.7}$$

This function is bounded and re-descending. It does not redescend to zero, however, but rather to a positive definite matrix constant. Score functions of $s_{v_{ij}}$ for λ=1 and several values of correlation ρ at the standard bivariate normal distribution are given in Figures 2. The contours are not closed for large values of the arguments x_1 and x_2 of the vector x. However, the appearance of these contours still suggest the possibility of clustering by the estimation procedure. Moreover, the procedure can also be used to determine observations which might be trimmed from the sample, if trimming or peeling were for some reason a desirable thing to do. The observations with relatively low values of the final weights $\tilde{v}_j = \exp(-\frac{c}{2}(x_j-\tilde{\mu})^T\tilde{V}^{-1}(x_j-\tilde{\mu}))$ are those which might be peeled off.

The estimators $\tilde{\mu}$ and $\tilde{V}$ are not in the class of those considered by Maronna (1976) even though they are similar in form. Devlin, Gnanadesikan, and Kettenring (1981) report favorable results obtained with the use of Maronna- and Huber- type multivariate estimators.

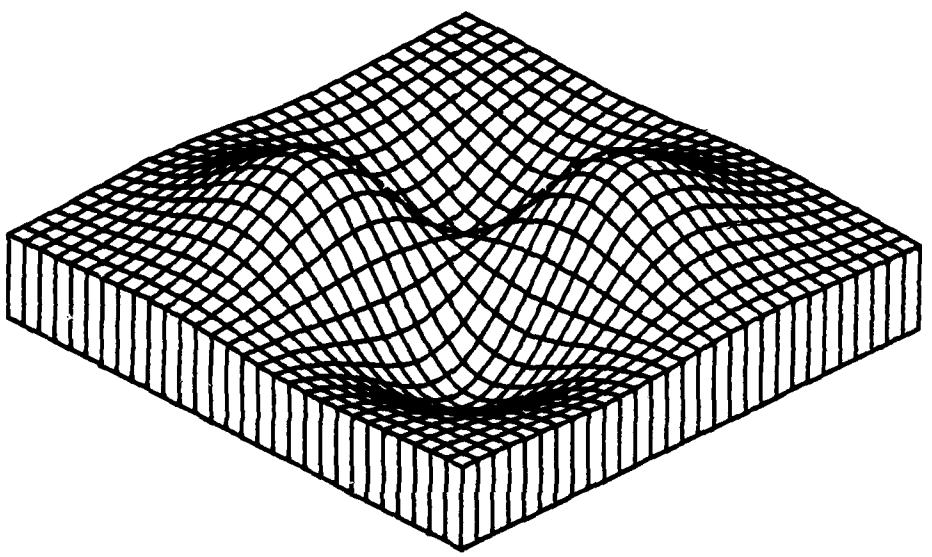


Figure 2a. Score function for $\tilde{v}_{12}$ at the standard bivariate normal, $\rho=0$, $\lambda=1$ (Azimuth = 45°, Elevation = 30°).

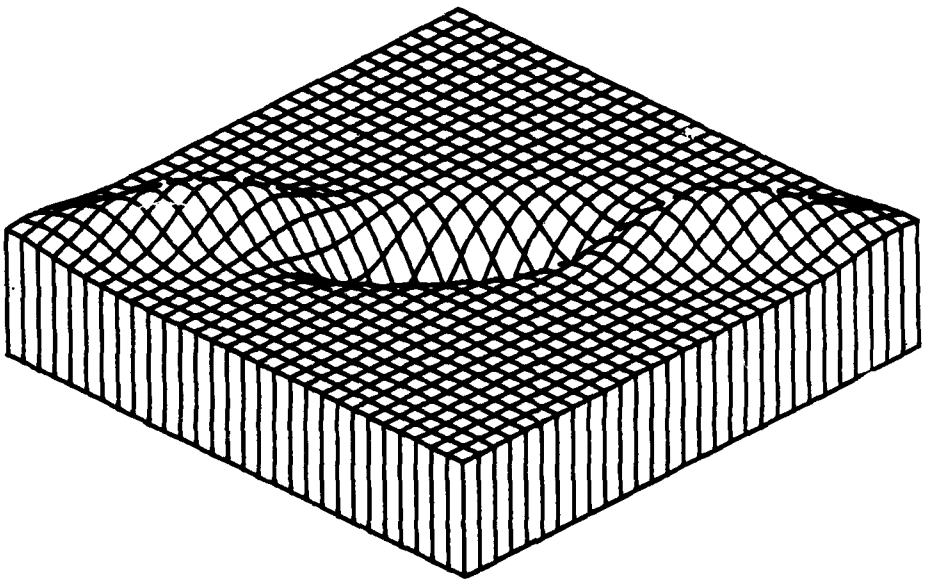


Figure 2b. Score function for $\tilde{v}_{11}$ at the standard bivariate normal, $\rho=.9$, $\lambda=1$ (Azimuth = 45^{o}, Elevation = 30^{o}).

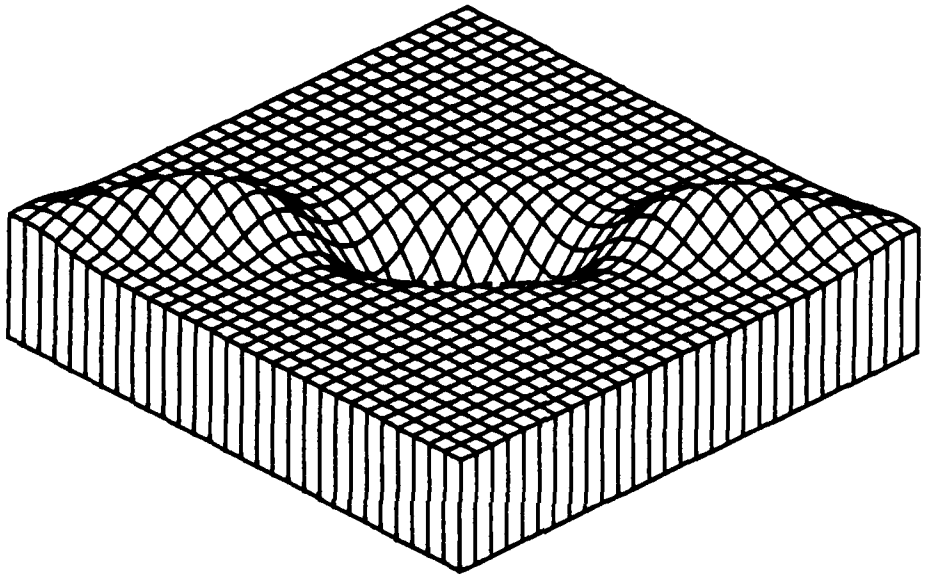


Figure 2c. Score function for $\tilde{v}_{12}$ at the standard bivariate normal, $\rho=.9$, $\lambda=1$ (Azimuth = 45°, Elevation = 30°).

4. Choice of λ

There are two possible uses for the procedure we propose here. The first is to choose a single value of λ, possibly based on efficiency considerations, and use it as a robust procedure. The choices $\lambda=1$ or $\lambda=2$ provide high efficiencies and good robustness properties. The second use is the one for which the procedure was developed and which has proved most useful in practical applications. We use the procedure to generate a sensitivity analysis. In a practical exploratory setting we first compute the maximum likelihood estimators and use these for starting values in the iterative algorithm. Next we take a value of λ, 4 or 2, and examine the behavior of the estimates. Finally, we would take $\lambda=1$ or $\frac{1}{2}$ and gradually decrease it. The response surface of the parameter values and the final weights as a function of λ are of primary interest. In the process we determine estimates and final weights $\tilde{v}_{j\lambda} = \exp(-\frac{1}{2}(1+2\lambda)^{-1}(x_j-\tilde{\mu})^T\tilde{V}^{-1}(x_j-\tilde{\mu}))$ associated with each observation. If the estimates are sensitive to this variation in λ, then there are problems associated with either the data or with the Gaussian error model or both. The particular observation(s) which is (are) the potential cause of the sensitivity are identified by low values of $\tilde{v}_{j\lambda}$ vis-a-vis the whole set of these weights. As c increases (λ decreases) observations a distance removed from $\tilde{\mu}$ receive lower weight. This discussion will be subsequently illustrated with an example.

The derivation which led to the estimators given in (2.14) did not require that λ in (2.14a) be the same as in (2.14b). We could for example use $\lambda=1$ for $\tilde{\mu}$ and $\lambda=2$ for $\tilde{V}$. Under such a choice we would then be able to fix the efficiencies that might be desired for both $\tilde{\mu}$ and $\tilde{V}$.

The asymptotic efficiencies are appropriate because some evidence is available that these estimators approach their asymptotic distributions very rapidly.

Furthermore, we need not have restricted ourselves to a scalar value of λ in order to arrive at (2.14). At the expense of greater algorithmic complexity we could have chosen values λ_{ij} corresponding to each v_{ij} in the covariance matrix V. Let the p×p matrix $L = (1+2\lambda_{ij})$ and the p×p matrix $M = (2+2\lambda_{ij})$ and let $L\times V = ((1+2\lambda_{ij})v_{ij})$ denote the Hadamard product of L with V. Then by arguments similar to those employed to arrive at (2.14) it may be shown that the more general estimators for μ and V satisfy the implicit relations

$$\mu = \frac{\sum_{j=1}^{n} x_j \exp(-\tfrac{1}{2}(x_j-\mu)^T(L\times V)^{-1}(x_j-\mu))}{\sum_{j=1}^{n} \exp(-\tfrac{1}{2}(x_j-\mu)^T(L\times V)^{-1}(x_j-\mu))}, \tag{4.1a}$$

and

$$L\times V = \left\{ \sum_{j=1}^{n} \exp(-\tfrac{1}{2}(x_j-\mu)^T(L\times V)^{-1}(x_j-\mu)) \right\}^{-1} \left\{ n \frac{|L\times V|^{\frac{1}{2}}}{|M\times V|^{\frac{1}{2}}} (L\times V)(M\times V)^{-1}(L\times V) \right.$$

$$\left. + \sum_{j=1}^{n} (x_j-\mu)(x_j-\mu)^T \exp(-\tfrac{1}{2}(x_j-\mu)^T(L\times V)^{-1}(x_j-\mu)) \right\}. \tag{4.1b}$$

The identity $M\times V = \left(\frac{2+2\lambda_{ij}}{1+2\lambda_{ij}}\right) \times (L\times V)$ is useful in computing (4.1b). The estimators of v_{ij} are computed in a component-wise fashion from the final iteration of (4.1b). These estimators would be of interest when it is desired to treat different components of the x_j differently. Such a situation arises in the bounded influence regression problem (Belsley, Kuh, and Welsch, 1980) where we may wish to be relatively critical in our

analysis of the dependent variables but less so for the independent variables since considerable information can be associated with a wide spread in the independent variables.

Specifically, let us partition the vectors x_j as $(y_j, z_{1j}, z_{2j}, \ldots, z_{qj}) = (y_j, z_j)$ where $q = p-1$. Let y_j be the dependent variable in a regression framework and let z_j represent a q-vector of independent variables. Corresponding to this partition we have $\mu^T = (\mu_1, \nu^T)$ where $E(y_j) = \mu_1$, $E(z_j) = \nu$. Further,

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} \cdots v_{1p} \\ v_{21} & v_{22} \cdots v_{2p} \\ \vdots & \vdots \\ v_{p1} & v_{p2} \cdots v_{pp} \end{pmatrix}$$

represents the corresponding partition of the covariance matrix. The regression of y_j on z_j is (Anderson, 1958, Ch. 2)

$$E(y_j | z_j) = \mu_1 + V_{12} V_{22}^{-1} (z_j - \nu). \tag{4.2}$$

We may wish to estimate μ_1 and v_{11} in a more critical fashion than V_{12} and this in turn in a more critical fashion than V_{22}. To achieve this we would take values λ_{11} associated with v_{11}, λ_{12} associated with V_{12}, and λ_{22} associated with V_{22} where $\lambda_{11} < \lambda_{12} < \lambda_{22}$. The matrix L would thus be

$$L = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{12} \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{22} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{22} \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12}\underset{\sim}{1}^T \\ \lambda_{12}\underset{\sim}{1} & \lambda_{22}\underset{\sim}{1}\underset{\sim}{1}^T \end{pmatrix}$$

where $\underset{\sim}{1} = (1,1,\ldots,1)^T$, a $q \times 1$ vector of ones. For exploratory and sensitivity analysis purposes we could consider first, say, λ_{11}, λ_{12}, λ_{22}, next $\frac{1}{2}\lambda_{11}$, $\frac{1}{2}\lambda_{12}$, $\frac{1}{2}\lambda_{22}$, and in general $k\lambda_{11}$, $k\lambda_{12}$, $k\lambda_{22}$ for some values of k. Points z_j far out and isolated in factor space or points whose response y_j give rise to large residuals will be identified by the response of the final observation weight $\tilde{v}_{jL} = \exp(-\frac{1}{2}(x_j-\tilde{\mu})^T(L\times V)^{-1}(x_j-\tilde{\mu}))$. For a fixed value of L, an estimate of the regression equation $E(y|z)$ is obtained by substituting parameter estimates $\tilde{\mu}_1$, $\tilde{V}_{12}$, $\tilde{V}_{22}$, $\tilde{\nu}$ in (4.2). Some of these ideas will be subsequently illustrated.

5. Examples

Example 5.1. The basic data for this example are taken from Anderson (1958). The first 25 points consist of the first two (of four) components of this data with five additional (outlying) observations appended. We have chosen λ=4, 2, 1, ½ for this illustration. Table 3 provides the estimates of the means and covariances as well as the maximum likelihood estimators. Table 3 also summarizes the weights $\tilde{v}_{j\lambda} = \exp(-\frac{c}{2}(x-\tilde{\mu})^T \tilde{V}^{-1} (x-\tilde{\mu}))$ associated with each point on the assumption that the data follow a single multivariate Gaussian distribution.

With $\lambda = +\infty$, all weights are the same. As λ decreases from $+\infty$, the weights become differentiated. The 5 outlying observations are rendered distinctive by their diminishing weights $\tilde{v}_{j\lambda}$ as λ decreases. This indicates that these observations are not consistent with the remainder of the observations and the assumption of a single Gaussian distribution. This is further highlighted by referral to equations (2.8) and (2.9). At convergence of the iterative estimation procedure we find for $\theta=\mu$ or $\theta=V$

$$\int_{R_p} \frac{\partial f(x)}{\partial \theta} \overset{x}{*} f_\omega(x) \left(f(x) \overset{x}{*} f_\omega(x) - n^{-1} \sum_{j=1}^{n} f_\omega(x-x_j)\right) dx = 0 \quad (5.1)$$

where $f_\omega(x)$ is the Gaussian density with mean 0 and covariance matrix λV, which implies that the density $f(x) \overset{x}{*} f_\omega(x)$ is being estimated by

$$\hat{g}_\lambda(x) = n^{-1} \sum_{j=1}^{n} |\lambda \tilde{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda} (x-x_j)^T \tilde{V}^{-1} (x-x_j)\right).$$

The density $f(x) \overset{x}{*} f_\omega(x)$ may be regarded as a prior distribution, $\hat{g}_\lambda(x)$ as a posterior density given the x_j. Figures 3a, 3b, 3c, depicts what the estimation procedure perceives as λ decreases. For large λ the density estimate $\hat{g}_\lambda(x)$ is approximately uniform. At $\lambda=2$ the density estimate contours are smooth except for some slight distortion in the area of (195,130). At $\lambda=1$ the probability surface is becoming distorted in the vicinity of the contamination but the distortion is not yet pronounced. Compare the estimates of the parameters and the observation weights $\tilde{v}_{j\lambda}$. At $\lambda=\frac{1}{2}$ the distortion has become dramatic and indeed separate "hills" for the outlying points have formed. Again compare the estimates and $\tilde{v}_{j\lambda}$ for $\lambda=\frac{1}{2}$. Since the density estimate perceives the outlying observations

Table 3a

Sensitivity of Observational Weights $\tilde{v}_{j\lambda}$(×100) to Variation in λ

	x_1	x_2	λ = 4	2	1	.5
1	179	145	36	39	45	50
2	201	152	33	36	23	16
3	185	149	37	41	47	55
4	188	149	37	40	44	49
5	171	142	34	35	39	39
6	192	152	37	39	44	48
7	190	149	37	39	41	42
8	189	152	37	40	47	53
9	197	159	35	36	39	36
10	187	151	37	41	47	55
11	186	148	37	40	45	50
12	174	147	35	37	38	36
13	185	152	37	41	46	50
14	195	157	36	38	42	43
15	187	158	35	36	30	20
16	161	130	27	23	17	8
17	183	158	34	34	22	10
18	173	148	35	36	33	27
19	182	146	37	40	45	50
20	165	137	31	30	30	25
21	185	152	37	41	46	50
22	178	147	36	39	45	49
23	176	143	36	38	42	45
24	200	158	34	35	37	36
25	187	150	37	41	47	54
26	200	130	18	6	0	0
27	200	135	23	11	0	0
28	165	160	24	13	1	0
29	195	170	28	22	6	1
30	220	170	23	18	13	6

Table 3b

Sensitivity of Parameter Estimates to Variation in λ

	λ = 4	2	1	.5
μ_1	185.5	185.2	184.9	184.9
μ_2	150.1	150.3	149.9	149.7
v_{11}	148.1	148.8	155.4	136.6
v_{22}	80.2	74.1	65.7	52.3
ρ_{12}	.52	.67	.85	.87

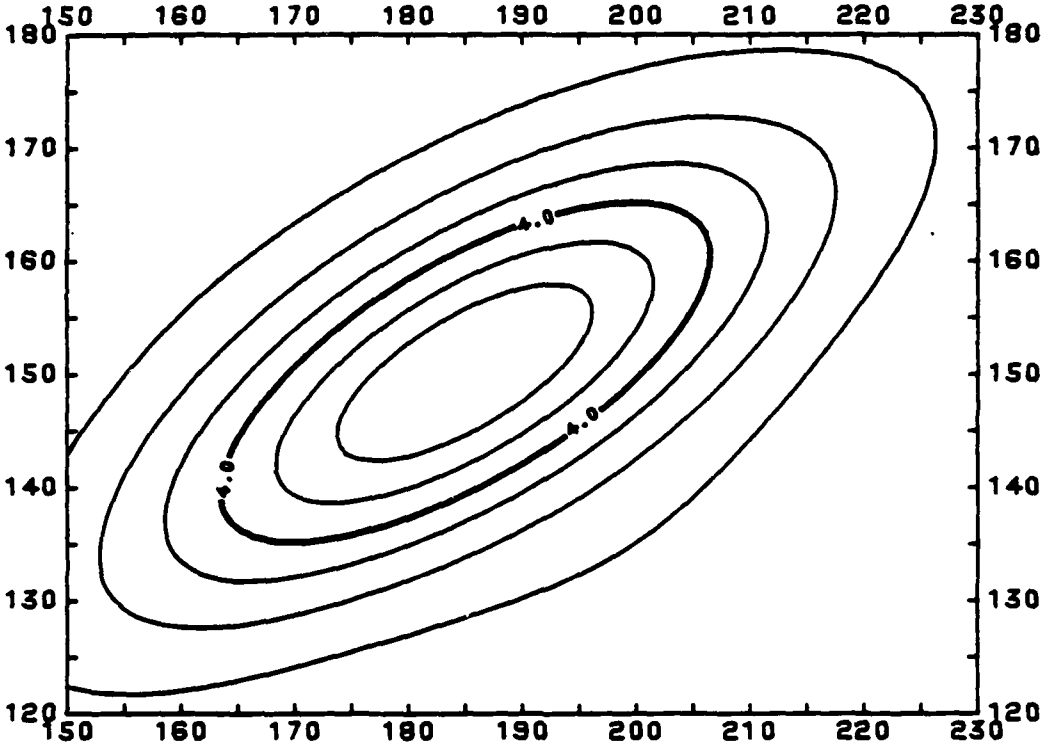


Figure 3a. Contour set of $\hat{g}_\lambda(x)$ of example 5.1, $\lambda=2$

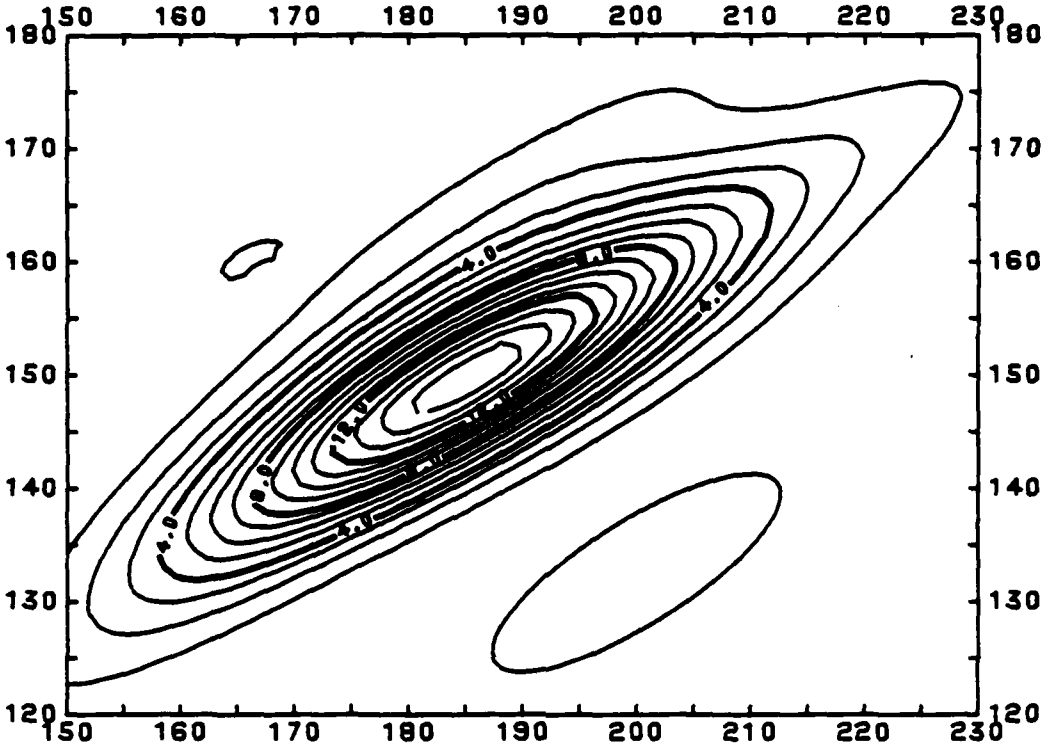


Figure 3b. Contour plot of $\hat{g}_\lambda(x)$ of example 5.1, $\lambda=1$

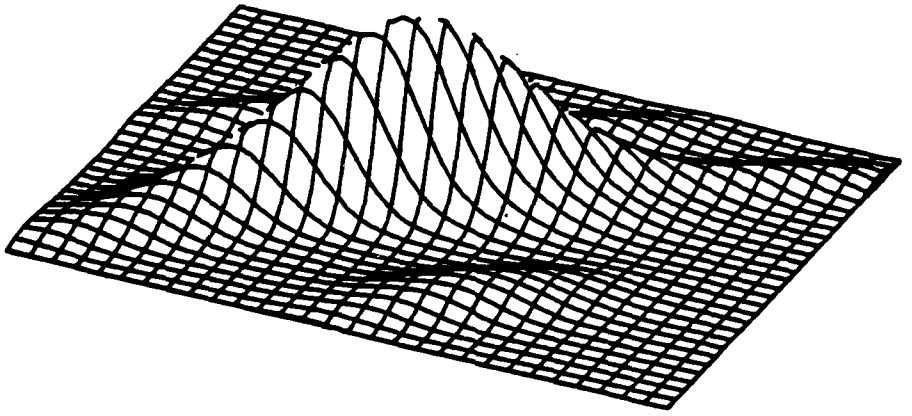


Figure 3c. Plot of $\hat{g}_\lambda(x)$ for data of example 5.1, λ=.5

as not consistent with the remainder of the data and the single Gaussian assumption, the reason for the down weighting of the outlying observations has become clear. If we let $\lambda \to 0+$, the density estimator becomes a set of Dirac delta functions located at each point.

At $\lambda = \frac{1}{2}$, the procedure has effectively clustered the data with the outlying observations excluded from the main cluster. Accordingly, as λ is varied, a dramatic change in the estimators implies the existence of clusters of observations different from the main cluster and not consistent with the prior assumption of strict Gaussianity. The reconstruction of the error density becomes increasingly dependent on the data as λ decreases and hence the procedure is increasingly critical with decreasing λ.

Example 5.2. If (y_j, z_j), $j = 1,2,\ldots,n$ represents a random sample from $N_2(\mu,V)$, $\mu = (\mu_1,\nu)^T$, then the regression of y on z is given by

$$E(y|z) = \mu_1 + V_{12} V_{22}^{-1} (z-\nu) = \beta_0 + \beta_1 z,$$

say, and the β's may be computed from the estimates of μ and V. The data for this example is taken from Andrews and Pregibon (1978) who were concerned with regression models. The data are presented in Table 4. The least squares or maximum likelihood estimates are also presented in Table 4. We shall use the L matrix version of the modified integrated squared error procedure as discussed in section 4. We wish to be most critical with respect to estimation of v_{11} and μ_1, to somewhat less critical with respect to estimation of V_{12}, and least critical with respect to estimation of V_{22} and ν. Thus we take, for any single application such as robost regression, $\lambda_{11} < \lambda_{12} < \lambda_{22}$. This choice seems to be particularly appealing since, in a regression framework, we wish to retain the high

Table 4a

Sensitivity of $\tilde{v}_{jL}$ (×100) to Variation in λ_{11}, λ_{12}, λ_{22}

			$(\lambda_{11}, \lambda_{12}, \lambda_{22})$		
	y	x	(2,4.8)	(1,2,4)	(.5,1,2)
1	95	15	100	99	97
2	71	26	72	48	20
3	83	10	85	76	65
4	91	9	96	93	88
5	106	15	88	82	74
6	87	20	96	89	72
7	93	18	99	96	87
8	100	11	98	97	97
9	104	8	95	91	84
10	94	20	96	91	78
11	113	7	81	70	54
12	96	9	99	97	95
13	83	10	85	76	65
14	84	11	88	81	71
15	102	11	97	95	93
16	100	10	98	97	96
17	105	12	92	89	84
18	57	42	40	11	0
19	121	17	50	33	19
20	86	11	92	86	79
21	100	10	98	97	96

Table 4b

Sensitivity of Estimates to Variation in λ_{11}, λ_{12}, λ_{22}*

	(2,4,8)	(1,2,4)	(.5,1,2)
μ_1	13.4	12.8	12.3
ν	94.6	95.3	95.9
v_{11}	45.6	32.7	23.6
v_{12}	-49.2(-.56)	-29.2(-.43)	-14.8(-.27)
v_{22}	168.3	143.9	125.1

*The estimate of correlation is given in parentheses along with v_{12}

efficiency associated with substantial spread in the independent variable. If on the other hand, we wish to be aware of extreme values of the independent variables, a sensitivity analysis may be more appropriate. We first take $\lambda_{11}=2$, $\lambda_{12}=4$, $\lambda_{22}=8$. The weights $\tilde{v}_{jL} = \exp(-\frac{1}{2}(x_j-\tilde{\mu})^T(L\times V)^{-1})(x_j-\mu))$ are presented in column 3 of Table 4. Next we take $\lambda_{11}=1$, $\lambda_{12}=2$, $\lambda_{22}=4$ and the results are presented in column 4. The parameter estimates are sensitive functions of L. The points which are most influential or potentially inconsistent vis-a-vis the linear model with a Gaussian error structure are determined from observational weights $\tilde{v}_{jL}$, i.e. those with low values of $\tilde{v}_{jL}$. Three points, 2, 18, 19, are especially singled out. Point 18 represents an extreme point in the z-space and is most influential on the estimate of the slope $\beta_1 = v_{12}/v_{22}$ of the regression line. Point 2 has the second-most extreme value of z. Point 19 produces an extreme residual. Analogous results obtain if all $\lambda_{ij} = \lambda$ and λ is varied in order to determine the response of the parameter estimates and the observational weights to variation in λ. In many cases, not all, taking just a single value of λ (or set of values λ_{ij}) provides sufficient information concerning the response of the parameter estimates and weights $\tilde{v}_{j\lambda}$ (or $\tilde{v}_{jL}$) to variation in λ (or L) in the sense that if λ (the λ_{ij}) were further decreased, the trend in response will be continued. In these cases a robust analysis will lead to the same conclusions as the sensitivity analysis. In some cases, a change in trends will be observed as λ decreases.

Example 5.3. It is possible to produce non-Gaussian data for which variation in λ does not lead to dramatic changes in the parameter estimates or low values of $\tilde{v}_{j\lambda}$. The three dimensional data for this illustration are taken from Gnanadesikan (1977, pp. 50-52). The 61 triads of his

example 7 were obtained by appending a standard normal deviate to each of the coordinates on the surface of a specified paraboloid. The two-dimensional scatter plots of these data are not suggestive of the data in three dimensions lying near a curved surface. We determine the response of the parameter estimates to changes in λ. The mean vector estimate at $\lambda=8$ is (-3.54, 4.72, 26.94) while at $\lambda=\frac{1}{4}$ it is (-3.53, 4.70, 26.95). The variance estimates at $\lambda=8$ are (3.81, 2.33, 2.94) while at $\lambda=\frac{1}{4}$ they are (4.43, 2.76, 3.79); the correlation estimates at $\lambda=8$ are (-.51, -.47, .20) while at $\lambda=\frac{1}{4}$ they are (-.56, -.50, .27). The estimates of the mean are remarkably stable but the estimated variances increase with a decrease in λ. These characteristics imply that if the data or the Gaussian distribution model is not appropriate, the best place to look for difficulties is at the centroid. This is confirmed by the distribution of the weights $\tilde{v}_{j\lambda}$, especially for $\lambda=\frac{1}{4}$. For example, for $\lambda=\frac{1}{4}$, the largest three weights $\tilde{v}_{j\lambda}$, .89, .84, .82, which indicate that there are no observations near the centroid. This deficiency of observations near the centroid is also determinable from $\lambda=8$ results, but it is highlighted at the smaller values of λ. A plot of $-2(1+2\lambda) \log \tilde{v}_{j\lambda}$ on χ^2 paper with 3 degrees of freedom further emphasizes the inappropriateness of the Gaussian assumption.

6. Multivariate Two-Way Cross Classification

The system of equations (2.6) and (2.7) are readily extended to include multivariate regression and design situations. We indicate how this may be done for the case of a two-way cross classified design. The arguments are similar for other designs and regression problems.

The two-way cross classified model may be written

$$E(x_{jk\ell}) = \mu + \alpha_j + \beta_k \tag{6.1}$$

$j = 1,2,\ldots,a$, $k = 1,2,\ldots,b$, $\ell = 1,2,\ldots,n_{jk}$, $n_{jk} \geq 1$. The $x_{jk\ell}$ are assumed to be p-dimensional Gaussian with covariance matrix V. The quantities μ, α_j, β_k are p×1 location vectors. Define

$$\phi_{jk}(u) = \exp\{iu^T(\mu + \alpha_j + \beta_k) - \tfrac{1}{2}\, u^T V u\}, \tag{6.2}$$

$$f_{jk}(x) = |2\pi V|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}\,(x-\mu-\alpha_j-\beta_k)^T\, V^{-1}\,(x-\mu-\alpha_j-\beta_k)\}, \tag{6.3}$$

and

$$\omega(\phi(u)) = \exp\Big(-\frac{\lambda}{2}\, u^T V u\Big), \tag{6.4}$$

$$f_\omega(x) = |2\pi\lambda V|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\, x^T(\lambda V)^{-1}x)\ . \tag{6.5}$$

The system of equations parallel to (2.6) and (2.7) are

$$\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{\ell=1}^{n_{jk}} \int_{R_p} \frac{\partial\phi_{jk}(u)}{\partial\theta}\,(\phi_{jk}(u) - \exp(iux_{jk\ell}))^*\, |\omega(\phi(u))|^2\, du = 0 \tag{6.6}$$

or, equivalently

$$(2\pi)^p \sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{\ell=1}^{n_{jk}} \int_{R_p} \frac{\partial f_{jk}(x)}{\partial\theta} \overset{x}{*} f_\omega(x)(f_{jk}(x) \overset{x}{*} f_\omega(x) - f_\omega(x-x_{jk\ell}))dx = 0, \tag{6.7}$$

with arguments $\theta=\mu$, α_j, β_k, V, $i = 1,2,\ldots,a$, $k = 1,2,\ldots,b$. Equation (6.6) may be explicitly evaluated and leads to the implicit equations

$$\sum_j \sum_k \sum_\ell (x_{jk\ell} - \mu - \alpha_j - \beta_k)\, v_{jk\ell,\lambda} = 0,$$

$$\sum_k \sum_\ell (x_{jk\ell} - \mu - \alpha_j - \beta_k)\, v_{jk\ell,\lambda} = 0,\ j = 1,2,\ldots,a$$

$$\sum_j \sum_\ell (x_{jk\ell} - \mu - \alpha_j - \beta_k)\, v_{jk\ell,\lambda} = 0,\ k = 1,2,\ldots,b.$$

The rank of this system of equations is $a+b-1$. The first of these equations suggest the constraints

$$\sum_j \sum_k \sum_\ell \alpha_j\, v_{jk\ell} = \sum_j \sum_k \sum_\ell \beta_k\, v_{jk\ell} = 0 \tag{6.8}$$

be appended to produce a full rank system. Along with (6.8) we obtain the implicit equations

$$\mu = \frac{\sum_j \sum_k \sum_\ell x_{jk\ell}\, v_{jk\ell,\lambda}}{\sum_j \sum_k \sum_\ell v_{jk\ell,\lambda}} \tag{6.9}$$

$$\alpha_j = \frac{\sum_k \sum_\ell (x_{jk\ell} - \mu - \beta_k)\, v_{jk\ell,\lambda}}{\sum_k \sum_\ell v_{jk\ell,\lambda}},\ j = 1,2,\ldots,a-1, \tag{6.10}$$

$$\beta_k = \frac{\sum_j \sum_\ell (x_{jk\ell} - \mu - \alpha_j)\, v_{jk\ell,\lambda}}{\sum_j \sum_\ell v_{jk\ell,\lambda}},\ k = 1,2,\ldots,b-1 \tag{6.11}$$

and

$$V = (1+2\lambda)^{-1} \frac{\sum_j \sum_k \sum_\ell (x_{jk\ell} - \mu - \alpha_j - \beta_k)(x_{jk\ell} - \mu - \alpha_j - \beta_k)^T v_{jk\ell,\lambda}}{\sum_j \sum_k \sum_\ell \left(v_{jk\ell,\lambda} - \left(\frac{1+2\lambda}{2+2\lambda}\right)^{\frac{1}{2}(p+2)}\right)} \tag{6.12}$$

where

$$v_{jk\ell,\lambda} = \exp(-\tfrac{1}{2}(1+2\lambda)^{-1}(x_{jk\ell} - \mu - \alpha_j - \beta_k)^T V^{-1}(x_{jk\ell} - \mu - \alpha_j - \beta_k)). \tag{6.13}$$

Observations $x_{jk\ell}$ which require special consideration are indicated, as in section 5, by low values of $\tilde{v}_{jk\ell,\lambda}$ vis-a-vis the whole set. A low value of $\tilde{v}_{jk\ell,\lambda}$ may mean that the particular observation is a potential outlier, Too many low values will imply that the model assumption of a single Gaussian parent may not be warranted or that the model is mis-specified or that there are indeed a number of potential outliers. In the latter case a goodness-of-fit test will usually declare against the Gaussian error distribution. Furthermore, if $n_{jk} > 1$ and we find that individual cells have low values $v_{jk\ell,\lambda}$ associated with them, then interaction in the table is a distinct possibility. In this case we generalize the model to

$$E(x_{jk\ell}) = \mu + \alpha_j + \beta_k + \gamma_{jk}.$$

and proceed accordingly.

This multivariate procedure can be especially useful for exploratory purposes. Determination of the sensitivity of $\tilde{v}_{jk\ell,\lambda}$ and the parameter estimates to changes in λ will serve to uncover potential problems with the data or the model considered as a unit. The procedure is computationally inexpensive and easy to use. This procedure does not apparently lend itself to hypothesis testing problems per se. However, it could be effectively used in conjunction with tests of hypotheses. If the sensitivity analysis uncovers some difficulty with the data or the model, then a test of hypothesis may be appropriate.

Example 6.1. The data (Table 5) for this example were taken from Anderson (1958, p. 218) who gives some additional background concerning these data. The first component of the observation vector is a barley yield in a given year; the second component is the same measurement made the following year.

Table 5

VARIETIES

LOCATION	1	2	3	4	5
1	81	105	120	110	98
	81	82	80	87	84
2	147	142	151	192	146
	100	116	112	148	108
3	82	77	78	131	90
	103	105	117	140	130
4	120	121	124	141	125
	99	62	96	126	76
5	99	89	69	89	104
	66	50	97	62	80
6	87	77	79	102	96
	68	67	67	92	94

We fit the model (6.1) to this data by the method of maximum likelihood and by the modified integrated squared error method for various values of λ with the objective of performing a sensitivity analysis. The results of this analysis are summarized in Tables 6 and 7. We have only given the results for $\lambda=2$ since the response of the parameter estimates and the final weights to decreases in λ continues the trend evidenced in Tables 6 and 7. Table 6 indicates that the largest change occurred in the parameter α_5 and the covariance. The correlation increased from .22 to .37. The final weights $\tilde{v}_{jk,2}$ are given in Table 7. Observation (5,3) receives an especially low weight while observations (1,3), (3,4), (5,4), and, to a lesser extent, (2,4) also receive low weights. It is likely that (5,3) is an outlier although we have not applied any tests of discordancy. We are suggesting that low weights raise the suspicion of potential outliers or other difficulties with the data and the model (Gaussianity, additivity, etc.). We are not suggesting that this procedure be used as a formal test for outliers. These observations have had the effect of reducing the correlation between the first and second year yields.

A rough rule of how improbable the weights are may be determined by the following considerations. The quadratic form $(x_{jk}-\tilde{\mu})^T\tilde{V}^{-1}(x_{jk}-\tilde{\mu})$ is approximately χ^2 on 2 degrees of freedom. The 1% point of this distribution is 9.21. Roughly, for $\lambda=2$, we would expect weights less than $\exp(-9.21/(2(1+2\lambda))) = .40$ about 1% of the time.

Additional reduction of λ say to 1.5 will produce a somewhat stronger version of basically the same results. However, when λ is decreased to unity the procedure starts to break down in the sense that the singled-out observations above receive weights near 0 and a few of the other originally

Table 6

Maximum Likelihood (ML) and Modified Integrated Squared Error (λ=2) Parameter Estimates

	μ	β_1	β_2	β_3	β_4	β_5
ML	109.1	-6.4	- 7.2	-5.6	18.4	0.8
	93.2	-7.0	-12.8	1.7	16.0	2.2
λ=2	108.8	-5.7	- 7.2	-3.5	18.6	1.0
	92.8	-6.1	-11.2	-1.6	17.9	4.2

	α_1	α_2	α_3	α_4	α_5	α_6
ML	- 6.3	46.5	-17.5	17.1	-19.1	-20.9
	-10.4	23.6	25.8	-1.4	-22.2	-15.6
λ=2	- 7.9	45.3	-19.7	16.9	-12.7	-21.4
	-10.8	22.3	25.3	- .1	-25.5	-16.1

$\hat{V}$

109.3	.22
26.7	133.9

$\tilde{V}$

101.9	.37
42.4	125.5

Table 7

Final Weights $\tilde{v}_{jk,2}$ (×100)

VARIETIES

	1	2	3	4	5
1	73	85	56	85	99
2	94	32	100	66	88
3	94	98	93	57	95
4	87	68	98	77	69
5	93	97	12	46	93
6	95	98	94	98	87

low weights have been further reduced. The first component variance is dramatically reduced. The reason for this is that some of the observations are separated in such a way that the empirical density estimator around it intersects only slightly with the empirical density estimator around other observations. Accordingly, multiple densities are perceived by the procedure and the more isolated and less massive with respect to the model (6.1) and the Gaussian assumption are basically excluded from the estimates by virtue of their low weights.

When $\lambda=0$, the density estimate $f_{\omega}(x-x_{jk})$of (6.7) around each point x_{jk} becomes a Dirac delta function. This fact does not ensure that all but a few weights will tend to zero, however. If the model (6.1) is truly appropriate and if the data $x_{jk\ell}$, $\ell>1$ replicates are truly Gaussian, then the variances in the covariance matrix will ultimately begin to increase with further decreases in λ.

It would be desirable to have an estimator of V whose influence function redescends to zero. Such an estimator is derivable from the modified integrated squared error procedure but we do not present it here. Still another approach, involving a generalization of Shannon's information or likelihood produces very similar estimators.

References

1. Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. New York: Wiley.

2. Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter. Journal of the Royal Statistical Society, B, 40, pp. 85-93.

3. Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. New York: Wiley.

4. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

5. Bryant, J., and Paulson, A.S. (1979). Some comments on characteristic function-based estimators. Sankhyā, A, pp. 109-116.

6. Cramér, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

7. Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1976). Robust estimation and outlier detection with correlation coefficients. Biometrika, 62, pp. 531-545.

8. Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981). Journal of the American Statistical Association, 76, pp. 354-362.

9. Dwyer, P.S. (1967). Some applications of matrix derivatives in multivariate analysis. Journal of the American Statistical Association, 62, pp. 607-625.

10. Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.

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

12. Maronna, R.A. (1976). Robust M-estimators of multivariate location and Scatter. Annals of Statistics, 4, pp. 51-67.

13. Parr, W.C. and Schucany, W.R. (1980). Minimum distance and robust estimation. Journal of the American Statistical Association, 75, pp. 616-624.

14. Parzen, E. (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33, pp. 1065-1076.

15. Paulson, A.S. and Nicklin, E.H. (1981). The form, and some robustness properties, of integrated distance estimators for linear models, applied to some published data sets. To appear.


ATE
LME