Working Paper 93-39
Statistics and Econometrics Series 25
November, 1993

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

DIAGNOSTICS AND ROBUST ESTIMATION IN MULTIVARIATE DATA TRANSFORMATIONS

Santiago Velilla*

Abstract

This paper presents a method for detecting multivariate outliers which might be distorting the estimation of a transformation to normality. A robust estimator of the transformation parameter is also proposed.

Key Words
Likelihood displacement, multivariate Box-Cox transformation, S estimators.

*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

DIAGNOSTICS AND ROBUST ESTIMATION IN MULTIVARIATE DATA TRANSFORMATIONS

Santiago Velilla, *Dpto. de Estadística y Econometría, Universidad Carlos III de Madrid*, 28903-*Getafe* (*Madrid*), *Spain*.

Abstract: This paper presents a method for detecting multivariate outliers which might be distorting the estimation of a transformation to normality. A robust estimator of the transformation parameter is also proposed.

Keywords and phrases: Likelihood displacement; Multivariate Box-Cox transformation; S estimators.

1. Introduction

Consider a p-variate random vector $\mathbf{X}=(X_1, \ldots, X_p)'$ such that all its components take positive values. If $\mathbf{X}$ is not multivariate normal, Andrews et al. (1971) propose the following transformation method to normality. Defining for a>0 and scalar λ the family of transformations

$$a^{(\lambda)}=\begin{cases} \dfrac{a^{\lambda}-1}{\lambda}, & \lambda\neq 0; \\ \log a, & \lambda=0, \end{cases} \tag{1.1}$$

they consider a vector $\Lambda=(\lambda_1, \ldots, \lambda_p)'$ of transformation parameters, one for each dimension, such that when transforming each X_j in the form $X_j^{(\lambda_j)}$, the following model holds, at least approximately,

$$X^{(\Lambda)}=(X_1^{(\lambda_1)}, \ldots, X_p^{(\lambda_p)})' \sim N_p(\mu,\Sigma), \tag{1.2}$$

where $\mu=(\mu_1, \ldots, \mu_p)'$ and $\Sigma=(\sigma_{ij})_{p\times p}$. Model (1.2) is a multivariate generalization of the univariate transformation model for a random variable X of Box and Cox (1964), namely, $X^{(\lambda)}\sim N(m,\sigma^2)$.

While extensively studied both in the one sample case and in the multiple linear regression case (see e.g. Atkinson (1985)), the application of the Box-Cox transformation to multivariate data has received little attention in the literature. Since model (1.2) implies the p marginal models $X_j^{(\lambda_j)}\sim N(\mu_j,\sigma_{jj})$, j=1, ...,p, it is usually recommended, on the basis of numerical simplicity, to estimate Λ by $\hat{\Lambda}_M=(\hat{\lambda}_{1M}, \ldots, \hat{\lambda}_{pM})'$, where $\hat{\lambda}_{jM}$ is the maximum likelihood estimator (MLE) of λ_j computed under the jth marginal model. It is argued that, in general, these marginal estimators will not differ from the MLE $\hat{\Lambda}$ of Λ computed from the joint model (1.2) and, as a consequence, there is a common belief that the problems encountered in dealing with multivariate data transformations can be handled by just using routine extensions of univariate techniques. On the other hand, it is well known that the MLE estimator to normality is very sensitive to outlying observations and as, a consequence, there is the need of developping both diagnostic techniques and robust estimation procedures for the multivariate transformation parameter Λ. In agreement with the ideas above, both univariate diagnostic techniques and robust estimation applied separately to each parameter λ_j would provide, in principle, a reasonably satisfactory joint methodology for detecting and/or accommodating anomalous observations.

The aim of this paper is to propose *specific* multivariate diagnostic methods and robust estimation procedures which are shown to be, either theoretically or by example, superior to simultaneous application of existing univariate techniques. Section 2 presents some background and motivation. Section 3 is devoted to diagnostics while section 4 is devoted to robustness. Section 5 contains some final comments.

2. Background and motivation

Let $\mathcal{X}=(x_{ij})=(x_1, \ldots, x_p)$ be a nxp data matrix from a random vector X with unknown distribution F. If, according to the model (1.2), the rows of the transformed data matrix $\mathcal{X}^{(\Lambda)}$, namely, $x_i^{(\Lambda)}=(x_{i1}^{(\lambda_1)}, \ldots, x_{ip}^{(\lambda_p)})$, i=1, ..., n, are i.i.d. $N_p(\mu,\Sigma)$, it can be shown that the concentrated log-likelihood $L_{max}(\Lambda)$ for Λ is (up to an additive constant)

$$L_{max}(\Lambda)=-\frac{n}{2}\log[|Z^{(\Lambda)'}AZ^{(\Lambda)}|], \qquad (2.1)$$

where $Z^{(\Lambda)}=(z_{ij}^{(\lambda_j)})=(Z_1^{(\lambda_1)}, \ldots, Z_p^{(\lambda_p)})$ is the nxp matrix of normalized variables of generic element $z_{ij}^{(\lambda_j)}=J_{\lambda_j}^{-1/n}(x_j)x_{ij}^{(\lambda_j)}$, i=1, ..., n, j=1, ..., p, and $J_{\lambda_j}(x_j)=(\prod_{i=1}^{n} x_{ij})^{\lambda_j-1}$, j=1, ..., p, are the jacobian terms. The matrix A in (2.1) is the nxn projection matrix $A=I_n-1_n1_n'/n$. The MLE $\hat{\Lambda}=(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ minimizes then the determinant

$$|Z^{(\Lambda)'}AZ^{(\Lambda)}|. \qquad (2.2)$$

However, as mentioned in the introduction, the customary practice is to estimate Λ by the estimator $\hat{\Lambda}_M=(\hat{\lambda}_{1M}, \ldots, \hat{\lambda}_{pM})'$. It can be shown by induction (see appendix A) the identity

$$|Z^{(\Lambda)'}AZ^{(\Lambda)}|=[\prod_{k=1}^{p} Z_k^{(\lambda_k)'}AZ_k^{(\lambda_k)}]\prod_{k=1}^{p}(1-r_k^2), \qquad (2.3)$$

where $r_1^2=0$ and, for k≥2, $r_k^2=r_k^2(\lambda_1, \ldots, \lambda_k)$ is the multiple correlation coefficient of $Z_k^{(\lambda_k)}$ with $(Z_1^{(\lambda_1)}, Z_{k-1}^{(\lambda_{k-1})})$. (2.3) decomposes the objective function for $\hat{\Lambda}$ in the product of the p marginal objective functions plus a factor which depends on the sequence $\{r_k^2\}$. In applications, $\{r_k^2\}$ is, in general, quite stable for all the values of the transformation parameter Λ in a neighborhood of the optimum $\hat{\Lambda}$ and therefore, the relevant information in the determinant criterion $|Z^{(\Lambda)'}AZ^{(\Lambda)}|$ comes from the marginal criterions $Z_k^{(\lambda_k)'}AZ_k^{(\lambda_k)}$, k=1, ...,p. This explains the observed closeness between $\hat{\Lambda}$ and $\hat{\Lambda}_M$. This empirical phenomenon is further illustrated in examples 2.1 and 2.2 below. In experience of the author, the function $-L_{max}(\Lambda)$ is typically convex, a convenient feature for numerical optimization using a canned routine. Exact expressions for the partial derivatives in the gradient vector $\partial L_{max}(\Lambda)/\partial\Lambda$ and the Hessian matrix $\partial(\partial L_{max}(\Lambda)/\partial\Lambda)'/\partial\Lambda$ of $L_{max}(\Lambda)$ are in appendix B.

EXAMPLE 2.1. A random sample of size n=50 is generated trough the model

$$(\log X,\log Y)' \sim N_2[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1. & .95\\ .95 & 1.\end{pmatrix}].$$

By applying a standard Newton-Raphson algorithm in each coordinate, we get $\hat{\Lambda}_M$=(-.0441,-.1474)'. Using $\hat{\Lambda}_M$ as the initial point in the corresponding bivariate optimization, we find $\hat{\Lambda}$=(-.0594,-.1812)'. The function $1-r_2^2(\lambda_1,\lambda_2)$ varies between .1159 and .1214 in the rectangle [-.1,0.]x[-.2,-.1].

EXAMPLE 2.2. A random sample of size n=50 is generated trough the trivariate lognormal model

$$(\log X,\log Y,\log Z)' \sim N_3[\begin{pmatrix}0\\0\\0\end{pmatrix},\begin{pmatrix}1. & .75 & .6\\ .75 & 1. & .5\\ .6 & .5 & 1.\end{pmatrix}].$$

Using the same methodology as above, we find $\hat{\Lambda}_M$=(.0354, -.1354,-.1740)' and $\hat{\Lambda}$=(.0443,-.1467,-.1543)'. In the rectangle [0.,.1]x[-.2,-.1]x[-.2,-.1], the function $1-r_2^2(\lambda_1,\lambda_2)$ is bounded between .431 and .436 and $1-r_2^2(\lambda_1,\lambda_2,\lambda_3)$ between .642 and .662.

3. Diagnostics

We will be interested in constructing a diagnostic for the effect of single case deletion on $\hat{\Lambda}$. The notation $\hat{\Lambda}_{(i)}$ will be used for the estimator computed after deletion of the ith row of the data matrix. For the univariate case, case deletion diagnostics have been proposed by Cook and

Wang (1983), Hinkley and Wang (1988) and Tsai and Wu (1990). Specifically, Cook and Wang (1983) proposed taking the *likelihood displacement*

$$LD_i=2[L_{max}(\hat{\lambda})-L_{max}(\hat{\lambda}_{(i)})], \tag{3.1}$$

as a scalar measure of influence. Cook and Wang (1983) suggested also a numerical approximation for the perturbed estimators $\hat{\lambda}_{(i)}$, i=1, ...,n, which is improved by Tsai and Wu (1990). The following example shows how, for p>1, an outlier can remain undetected when applying the univariate diagnostic techniques separately in each one of the coordinates.

EXAMPLE 3.1. An outlier $(x_{51},y_{51})'$ is added to the data set in example 2.1 in such a way that $(\log x_{51},\log y_{51})'=(2.1,-1.9)'$. Plotting the data in logs, figure 3.1, it is easily seen that $(\log x_{51},\log y_{51})'$ alters the structure of correlation of the cloud defined by the bulk of the remaining 50 data points. Now we get $\hat{\Lambda}_M=(-.0612,,-.0928)'$ and $\hat{\Lambda}=(-.1393,.0252)'$. Figure 3.2 a) and figure 3.2 b) show, for j=1,2, the marginal diagnostics

$$\hat{\lambda}^1_{jM(i)}$$

and

$$LD^j_i=2[L^j_{max}(\hat{\lambda}_{jM})-L_{max}(\hat{\lambda}^1_{jM(i)})],$$

respectively, where $\hat{\lambda}^1_{jM}$ is the corresponding marginal one-step approximation of Tsai and Wu (1990). Figure 3.2 c) displays the combined marginal diagnostic $LD^1_i+LD^2_i$. Notice that case 51 remains unnoticed.

Figure 3.1

Figure 3.2 a) Figure 3.2 b) Figure 3.2 c)

In the example above, marginal diagnostics fail because they don't take into account the correlation structure of the transformed data. In principle, a suitable multivariate diagnostic measure would be

$$LD_i=2[L_{max}(\hat{\Lambda})-L_{max}(\hat{\Lambda}_{(i)})] \tag{3.2}$$

By (2.1), the measure (3.2) has a cumbersome expression and is hard to work with. By a standard first order Taylor expansion and using that $\partial L_{max}(\Lambda)/\partial\Lambda\big|_{\Lambda=\hat{\Lambda}}=0$, we can approximate

$$LD_i\cong\tilde{L}D_i=(\hat{\Lambda}-\hat{\Lambda}_{(i)})'H(\hat{\Lambda})(\hat{\Lambda}-\hat{\Lambda}_{(i)}), \tag{3.3}$$

where $H(\hat{\Lambda})=-\frac{\partial}{\partial\Lambda}\left(\frac{\partial}{\partial\Lambda}L_{max}(\Lambda)\right)'\Bigg|_{\Lambda=\hat{\Lambda}}$ is the pxp observed Fisher information matrix. Velilla (1993) proposes the ellipsoid

$$(\Lambda-\hat{\Lambda})'H(\hat{\Lambda})(\Lambda-\hat{\Lambda})\leq\chi^2_{p,\alpha}, \tag{3.4}$$

as an alternative asymptotic $(1-\alpha)\times100\%$ confidence region for the transformation parameter. Therefore, calibration of the influence measure $\tilde{L}D_i$ on the right hand side of (3.3) can be made with reference to the percentage points of the χ^2_p distribution.

From the computational point of view, $\tilde{L}D_i$ depends on: (*i*) The MLE $\hat{\Lambda}$; (*ii*) The matrix $H(\hat{\Lambda})$; and (*iii*) The estimators $\hat{\Lambda}_{(i)}$, i=1, ...,n. The MLE is obtained by maximizing $L_{max}(\Lambda)$ and the elements of the matrix $H(\hat{\Lambda})$ can be computed from the corresponding expressions in theorem B.1 of appendix B, with the additional simplification obtained from $\partial L_{max}(\Lambda)/\partial\lambda_r\big|_{\Lambda=\hat{\Lambda}}=0$. Computation of $\hat{\Lambda}_{(i)}$ requires iteration and an approximation is developed as follows. By (2.2) the MLE $\hat{\Lambda}$ minimizes $|M(\Lambda)|$, where $M(\Lambda)=|Z^{(\Lambda)'}AZ^{(\Lambda)}|$, and, therefore, $\hat{\Lambda}_{(i)}$ minimizes, in obvious notation, $|M_{(i)}(\Lambda)|$. Let $\nabla_{(i)}(\Lambda) = \partial|M_{(i)}(\Lambda)|/\partial\Lambda$ and $H_{(i)}(\Lambda)=\partial^2|M_{(i)}(\Lambda)|/\partial\Lambda\partial\Lambda'$ be, respectively, the gradient vector and the hessian matrix of $|M_{(i)}(\Lambda)|$. In the first-order Taylor expansion around an initial guess Λ_0 for $\hat{\Lambda}_{(i)}$, $\nabla_{(i)}(\Lambda)\cong\nabla_{(i)}(\Lambda_0)+H_{(i)}(\Lambda_0)(\Lambda-\Lambda_0)$, we can use the fact that $\nabla_{(i)}(\hat{\Lambda}_{(i)})=0$ to obtain a one-step approximation for $\hat{\Lambda}_{(i)}$,

$$\hat{\Lambda}^1_{(i)}=\Lambda_0-[H_{(i)}(\Lambda_0)]^{-1}\nabla_{(i)}(\Lambda_0). \tag{3.5}$$

Equation (3.5) is a multivariate extension of equation (15) in Tsai and Wu (1990). Typically $\Lambda_0=\hat{\Lambda}$ and the computational problem is finished whenever we are able to compute the explicit expressions for the elements in $\nabla_{(i)}(\hat{\Lambda})$ and $H_{(i)}(\hat{\Lambda})$. These can be seen in appendix B.

EXAMPLE 3.1 (cont.). Figure 3.3 a) is an index plot of LD_i where case 51 is clearly pinpointed. Figure 3.3 b) is the corresponding plot for $\tilde{L}D_i$. Deletion of case 51 moves the estimator to the boundary of a 99.25% confidence ellipsoid.

Figure 3.3 a) **Figure 3.3 b)**

Finally, it is important to remark that in both (3.3) and (3.4) it is crucial to use the multivariate estimators $\hat{\Lambda}$ and $\hat{\Lambda}_{(i)}$. By replacing, in (3.3), $\hat{\Lambda}$ and $\hat{\Lambda}_{(i)}$ by their marginal counterparts $\hat{\Lambda}_M$ and $\hat{\Lambda}_{M(i)}$ the "marginal" version of (3.3), namely $LDM_i=2[L_{max}(\hat{\Lambda}_M)-L_{max}(\hat{\Lambda}_{M(i)})]$ often takes negative values. On the other hand, $\tilde{L}DM_i=(\hat{\Lambda}_M-\hat{\Lambda}_{M(i)})'H(\hat{\Lambda}_M)(\hat{\Lambda}_M-\hat{\Lambda}_{M(i)})$ does not indicate the outlier because of the lack of consideration of the correlation structure. See figures 3.4 a) and b). Moreover, the general claim of closeness between $\hat{\Lambda}$ and $\hat{\Lambda}_M$ is not true in the presence of outliers. If, for a given pxl vector u, $\|u\|^2_M=u'Mu$ denotes the square of the norm of u with respect the inner product pxp matrix M, we have, in the case of example 3.1, $\|\hat{\Lambda}-\hat{\Lambda}_M\|^2_H=2.06$ and $\|\hat{\Lambda}_{(51)}-\hat{\Lambda}_{M(51)}\|^2_H=.08$.

Figure 3.4 a) Figure 3.4 b)

4. A robust estimator of the transformation parameter

As a consequence of section 3, a multivariate outlier can remain undetected if only univariate diagnostic techniques are used. Therefore, if a robust estimator of the transformation parameter $\Lambda=(\lambda_1, \ldots,\lambda_p)'$ is to be constructed, it should not be based in robustifying separately in each dimension. This section presents a robust estimator $\hat{\Lambda}_R$ of Λ which takes into account the correlation structure among the transformed variables.

Given $x_1, \ldots,x_n$ i.i.d. p-variate observations with unknown distribution F that can be approximately modelled with the transformation model (1.1), the corresponding log-likelihood is (up to an additive constant)

$$L(\mu,\Sigma,\Lambda)=-(n/2)\log|\Sigma|-(1/2)\sum_{i=1}^{n}(x_i^{(\Lambda)}-\mu)'\Sigma^{-1}(x_i^{(\Lambda)}-\mu)+\log[J_\Lambda(\mathcal{X})], \tag{4.1}$$

where $J_\Lambda(\mathcal{X})=\prod_{j=1}^{p}J_{\lambda_j}(x_j)$, $J_{\lambda_j}(x_j)=(\prod_{i=1}^{n}x_{ij})^{\lambda_j-1}$, is the jacobian term. The MLE estimator $(\hat{\mu},\hat{\Sigma},\hat{\Lambda})$ is obtained maximizing in Λ the profile log-likelihood $L_{max}(\Lambda)=L(\hat{\mu}(\Lambda),\hat{\Sigma}(\Lambda),\Lambda)$, with $\hat{\mu}(\Lambda)=n^{-1}\sum_{i=1}^{n}x_i^{(\Lambda)}$, $\hat{\Sigma}(\Lambda)=n^{-1}\mathcal{X}^{(\Lambda)'}A\mathcal{X}^{(\Lambda)}$, which can be written in the form

$$-(n/2)\log[|\hat{\Sigma}(\Lambda)|]+\log[J_\Lambda(\mathcal{X})]. \tag{4.2}$$

From (4.2), a possible method for robustifying the estimation of Λ is to replace the objective function (4.2) by:

$$L_{max,R}(\Lambda)=-(n/2)\log[|\hat{\Sigma}_R(\Lambda)|]+\log[J_\Lambda(\mathcal{X})]. \tag{4.3}$$

where $\hat{\Sigma}_R(\Lambda)$ is a robust estimator of **Σ** computed from the data matrix $\mathcal{X}^{(\Lambda)}$. Our choice for $\hat{\Sigma}_R(\Lambda)$ is the S-estimator of dispersion as described in Lopuhaä (1989). For the univariate case, Carroll (1980) has proposed a robust estimation approach by replacing, in the normal log-likelihood for (μ,σ,λ), $d_i^2/2=(x_i^{(\lambda)}-\mu)^2/2\sigma^2$ by $\rho(d_i)$. For every fixed λ, he suggests estimating μ and σ for M-estimators $\mu_M(\lambda)$ and $\sigma_M(\lambda)$ and to estimate λ by maximizing the function $-(n/2)\log[\sigma_M^2(\lambda)]-\sum_{i=1}^{n}\rho[(x_i^{(\lambda)}-\mu_M(\lambda))/\sigma_M(\lambda)]$ + $(\lambda-1)\sum_{i=1}^{n}\log(x_i)$. For p=1, our proposal maximizes $-(n/2)\log[\sigma_S^2(\lambda)]$ − $\sum_{i=1}^{n}\rho[(x_i^{(\lambda)}-\mu_S(\lambda))/\sigma_S(\lambda)]$ + $(\lambda-1)$ + $\sum_{i=1}^{n}\log(x_i)$. For the univariate case, other work in robustness in transformations can be seen in Carroll and Ruppert (1988).

Next we study the properties and discuss the computation of $\hat{\Lambda}_R$. We also propose a measure of its robustness under contamination by analyzing the expression of its influence function. For the most part, the following discussion is heuristic and concentrates mainly on the ideas. More technical details can be found in appendix C.

4.1 Existence of solutions, consistency and asymptotic normality

Appendix C contains a brief description on conditions which imply the existence, for all n large enough, of a sequence of solutions $\{\hat{\Lambda}_R\}$ of (4.3) such that

$$\hat{\Lambda}_R \longrightarrow \Lambda_0, \text{ a.e,}$$

being $\Lambda_0=\Lambda_0(F)$ the *unique* global maximum over $\Lambda\in\mathbb{K}$ of

$$-(1/2)\log[\,|\Sigma(\Lambda,F)|\,]+\Lambda'u, \tag{4.4}$$

where $\mathbb{K}$ is a compact set of $\mathbb{R}^p$, $u=(E_F[\,|\log X_1|\,], \ldots, E_F[\,|\log X_p|\,])$ and $\Sigma(\Lambda,F)$ is the corresponding block of the solution $\theta(\Lambda,F)=(\mu(\Lambda,F),\Sigma(\Lambda,F))$ of the S-estimation problem $\min|\Sigma|$ over $\theta=(\mu,\Sigma)$ $(\Sigma>0)$ restricted to $\int\rho[\{(x^{(\Lambda)}-\mu)'\Sigma^{-1}(x^{(\Lambda)}-\mu)\}^{1/2}]F(dx)=k_p$. In agreement with this notation, $(\hat{\mu}_R(\Lambda),\hat{\Sigma}_R(\Lambda))=\theta(\Lambda,F_n)$. We delete sometimes dependence on F and simply write $\theta(\Lambda,F)=\theta(\Lambda)$.

Next we assume that Λ_0 is an interior point of $\mathbb{K}$ and write $\theta_0=(\mu_0,\Sigma_0)=\theta(\Lambda_0,F)$. By expression (2.7) in Lopuhaä (1989), for each Λ in $\mathbb{K}$, $\theta(\Lambda)$ solves

$$H(\Lambda,\theta)=\int\psi(x,\Lambda,\theta)F(dx)=0, \tag{4.5}$$

where the function $\psi=(\psi_1',\psi_2')'$ has components:

$$\psi_{1,r}(x,\Lambda,\theta)=u[d](x_r^{(\lambda_r)}-\mu_r),\ r=1,\ \ldots,p; \tag{4.6}$$

$\psi_{2,rs}(x,\Lambda,\theta)=pu[d](x_r^{(\lambda_r)}-\mu_r)(x_s^{(\lambda_s)}-\mu_s)-v[d]\sigma_{rs}$, $\quad r,s=1, \quad \ldots,p, \quad (r\le s)$,

where $d=d(x,\Lambda,\theta)=[(x^{(\Lambda)}-\mu)'\Sigma^{-1}(x^{(\Lambda)}-\mu)]^{1/2}$, $u[d]=\varphi[d]/d$, and $v[d]=t\varphi[d]-\rho[d]+k_p$, being φ the derivative ρ' of ρ. By assuming an appropriate form for $\rho[.]$ and suitable conditions on the moments of F, it can be shown that $H(\Lambda,\theta)$ has continuous second partial derivatives in a neighborhood of (Λ_0,θ_0) which can be obtained differentiating under the integral sign in (4.5). Since $H(\Lambda_0,\theta_0)=0$, if $\partial H(\Lambda,\theta)/\partial\theta\Big|_{(\Lambda,\theta)=(\Lambda_0,\theta_0)}$ is nonsingular, by the implicit function theorem (see, e.g. Fleming (1977, p.148)), there exists, locally around Λ_0, a function $\theta(\Lambda)$ with continuous second partial derivatives such that $H(\Lambda,\theta(\Lambda))=0$. We introduce now the set of artificial parameters $W=\{\{\omega_{r,j}\}, \{\omega_{rs,j}\}, \{\omega_{r,jk}\}, \{\omega_{rs,jk}\}: r,s=1, \ldots,p, (r\le s), j,k=1, \ldots,p\}$, which have the meaning $\omega_{r,j}=\partial\mu_r/\partial\lambda_j$, $\omega_{rs,j}=\partial\sigma_{rs}/\partial\lambda_j$, $\omega_{r,jk}=\partial^2\mu_r/\partial\lambda_j\partial\lambda_k$, and $\omega_{rs,jk}=\partial^2\sigma_{rs}/\partial\lambda_j\partial\lambda_k$. The following equation holds

$$\int\dot\psi(x,\Lambda,\theta,W)F(dx)=0, \tag{4.7}$$

where $\dot\psi$ is the array $\partial\psi/\partial\Lambda$. Moreover, Λ_0 satisfies

$$\int\Gamma(x,\Lambda,\theta)F(dx)=0, \tag{4.8}$$

where $\Gamma=\{\Gamma_j\}$ is the px1 gradient vector of the function in (4.4). Explicit expressions for the components of $\dot\psi$ and Γ are in appendix D. In summary, if $\Omega=(\Lambda,\theta,W)$, Λ_0 is the corresponding component of the solution Ω_0 of the equation

$$\int\Phi(x,\Omega)F(dx)=0, \tag{4.9}$$

where $\Phi(x,\Lambda,\theta)=(\Gamma,\psi,\dot\psi)$. By replacing, in (4.9), F by the empirical F_n, we get the estimating equation

$$n^{-1}\sum_{i=1}^{n}\Phi(x_i,\Omega)F(dx)=0, \tag{4.10}$$

whose solution $\hat\Omega_n=(\hat\Lambda_R,\hat\theta_n,\hat\omega_n)$ determines $\hat\Lambda_R$. Observe that Ω and Φ are both of dimension $q=(p/2)[2+(p+1)(p+3)]$. Equations (4.9) and (4.10) allow for an obtention of the asymptotic normality of $\hat\Lambda_R$ in the framework of M-estimation. By application of the conditions Huber (1967), if the function $\lambda_F[\Omega]=E_F[\Phi(x,\Omega)]$ has a nonsingular qxq derivative D at Ω_0, then $n^{1/2}(\hat\Omega_n-\Omega_0)\xrightarrow{D}N_q[0,D^{-1}M(D^{-1})']$, where $M=\mathrm{cov}_F[\Phi(x,\Omega)]$, and as a consequence,

$$n^{1/2}(\hat\Lambda_R-\Lambda_0)\xrightarrow{D}N_q[0,HR],$$

where HR is the proper pxp submatrix of $D^{-1}M(D^{-1})'$. A consistent estimate HR_n of HR is given by the associated pxp submatrix of $D_n^{-1}M_n(D_n^{-1})'$, where

$D_n=n^{-1}\sum_{i=1}^{n}\partial\Phi(x_i,\Omega)/\partial\Omega|_{\Omega=\hat{\Omega}_n}$ and $M_n=n^{-1}\sum_{i=1}^{n}\Phi(x_i,\hat{\Omega}_n)\Phi(x_i,\hat{\Omega}_n)'$.

4.2 Computation of $\hat{\Lambda}_R$

The numerical problem for computing $\hat{\Lambda}_R$ is to minimize in $\Lambda=(\lambda_1, \ldots,\lambda_p)'$ the function

$$h[\Lambda]=(1/2)\log[|\hat{\Sigma}_R(\Lambda)|]-\sum_{i=1}^{n}\lambda_j p_{n,j} \tag{4.11}$$

where $p_{n,j}=n^{-1}\sum_{i=1}^{n}\log(x_{ij})$. The function used in the determination of the robust estimator $\hat{\Sigma}_R(\Lambda)$ is the function

$$\rho[t]=\begin{cases}(t^6/6c)-(t^8/4c^3)+(t^{10}/10c^5), & |t|\le c;\\ c^5/60 \quad , & |t|>c,\end{cases} \tag{4.12}$$

for a suitable positive constant c. This function satisfies the regularity conditions of appendix C. Its derivative $\varphi[t]=(t^5/c)[1-(t/c)^2]^2$, $|t|\le c$, (0 for $|t|>c$) is a multiple of Tukey's biweight function. The first partial derivatives of $h[\Lambda]$ are

$$\partial h/\partial\lambda_j=(1/2)\sum_{\substack{r,s=1\\ r\le s}}^{p}[2\hat{\sigma}^{rs}-\delta_{rs}\hat{\sigma}^{rs}]\hat{\omega}_{rs,j}-p_{n,j}, \tag{4.13}$$

$j=1, \ldots,p$. For every fixed j, the array $\{\hat{\omega}_{rs,j}\}$ is determined from the equation (4.7) or, more specifically, from the linear system

$$A\hat{W}_j=\hat{b}_j, \tag{4.14}$$

where $\hat{W}_j=\{\{\hat{\omega}_{r,j}\}, \{\hat{\omega}_{rs,j}\}\}$, and the $[p(p+3)/2]\times1$ vector $\hat{b}_j=\{\hat{b}_{r,j};\hat{b}_{rs,j}\}$

and $[p(p+3)/2]\times[p(p+3)/2]$ matrix

$$A=\left(\begin{array}{c|c}\hat{\alpha}_{r,a} & \hat{\alpha}_{r,ab}\\ \hline \hat{\alpha}_{rs,a} & \hat{\alpha}_{rs,ab}\end{array}\right), \tag{4.15}$$

have elements

$$\hat{b}_{r,j}=-n^{-1}\sum_{i=1}^{n}\partial\psi_{1,r}/\partial\lambda_j,\ \hat{b}_{rs,j}=-n^{-1}\sum_{i=1}^{n}\partial\psi_{2,rs}/\partial\lambda_j,$$

$$\hat{\alpha}_{r,a}=n^{-1}\sum_{i=1}^{n}\partial\psi_{1,r}/\partial\mu_a,\ \hat{\alpha}_{r,ab}=n^{-1}\sum_{i=1}^{n}\partial\psi_{1,r}/\partial\sigma_{ab},$$

$$\hat{\alpha}_{rs,a}=n^{-1}\sum_{i=1}^{n}\partial\psi_{2,rs}/\partial\mu_a \text{ and } \hat{\alpha}_{rs,ab}=n^{-1}\sum_{i=1}^{n}\partial\psi_{1,rs}/\partial\sigma_{ab},$$

for $r,s,a,b=1, \ldots,p$ ($r\le s$, $a\le b$). All the derivatives are evaluated at

$(\Lambda,\hat{\theta}_R(\Lambda))$, where $\hat{\theta}_R(\Lambda)=(\hat{\mu}_R(\Lambda),\hat{\Sigma}_R(\Lambda))$. Exact expressions can be found in appendix D. Recall that the matrix A above is the same for every j. Second partial derivatives for h theoretically exist but have untractable expressions. An algorithm for computing $\hat{\Lambda}_R$ based on a Newton-Raphson iteration which uses both the gradient $G=\partial h/\partial\Lambda$ and the Hessian matrix $H=\partial(\partial h/\partial\Lambda)'/\partial\Lambda$ of h is thus not recommended. We suggest instead the following algorithm for computing $\hat{\Lambda}_R$:

(*i*) Start with an initial value of Λ, Λ_0 say, and compute the robust S-estimators $\hat{\theta}_R(\Lambda_0)=(\hat{\mu}_R(\Lambda_0),\hat{\Sigma}_R(\Lambda_0))$ from the transformed data matrix $\mathcal{X}^{(\Lambda_0)}$. If there is some previous diagnostic information, Λ_0 could be the approximation $\hat{\Lambda}^1_{(1)}$ computed deleting a dubious case i. A suitable algorithm for computing $\hat{\theta}_R(\Lambda_0)$ is given in Ruppert (1992).

(*ii*) Solve the collection of systems $A_0\hat{W}_{j0}=\hat{b}_{j0}$, j=1, ...,p, obtain the gradient G_0 of h at $\Lambda=\Lambda_0$ and choose an initial guess for H_0, typically $H_0=I_p$.

(*iii*) Update (Λ_0,G_0,H_0) to (Λ_1,G_1,H_1) as in a Quasi-Newton algorithm with the BFGS formula (see, e.g. Seber and Wild (1989, pp. 605-609 for a description).

(*iv*) Iterate (*i*)-(*ii*)-(*iii*) until convergence.

4.3 Influence function of $\hat{\Lambda}_R$

The influence function of the functional $\Lambda_0=\Lambda(F)$ evaluated at a point $x\in\mathbb{R}^p$ and at underlying distribution F is defined pointwise as the limit

$$\lim_{h\to 0^+}\frac{\Lambda[(1-h)F+h\delta_x]-\Lambda[F]}{h},$$

if the limit exists. If the conditions for existence and consistency hold and $\lambda_F[\Omega]=E_F[\Phi(x,\Omega)]$ has a nonsingular qxq derivative D at Ω_0, then for the "larger" functional $\Omega_0=\Omega(F)$, the influence function $IF(x;\Omega;F)$ is the qx1 vector $-D^{-1}\Phi[x;\Omega(F)]$. We have the sampling approximation $IF(x;\Omega;F)\cong\hat{IF}_n=-D_n^{-1}\Phi[x;\hat{\Omega}_n]$. The influence function $IF(x;\Lambda;F)$ is given by the first p coordinates of $IF(x;\Omega;F)$ with sampling approximation provided by the first p coordinates of $\hat{IF}_n$.

4.4 Examples

EXAMPLE 4.1. We apply the algorithm described in 4.2 to the data set of example 3.1. We choose the constant c=3 in the definition of ρ in (4.12). By starting with $\Lambda_0=\hat{\Lambda}^1_{(51)}=(.0547,-.2024)'$ and $H_0=I_2$, we get $\hat{\Lambda}_R=(-.1057,-.2320)'$

after three iterations. To assess the distance between $\hat{\Lambda}$ and $\hat{\Lambda}_R$. we see that $\|\hat{\Lambda}-\hat{\Lambda}_R\|_M^2=6.3876$, that is $\hat{\Lambda}_R$ lies *outside* the 95% confidence region (3.4). By comparison $\|\hat{\Lambda}_{(51)}-\hat{\Lambda}_R\|_M^2=.2919$.

EXAMPLE 4.2. Let $\hat{u}_R(x)$ the pxl vector formed by the first p coordinates of the sampling approximation $IF(x;\Omega;F)\cong\hat{IF}_n=-D_n^{-1}\Phi[x;\hat{\Omega}_n]$. A suitable norming matrix for $\hat{u}_R(x)$ is the pxp matrix HR_n introduced in 4.1

$$IR(x)=[\hat{u}_R(x)'(HR_n/n)^{-1}\hat{u}_R(x)]^{1/2}.$$

The MLE estimator is also a particular case of M-estimator by choosing u[d]=v[d]=1 in (4.7). The influence function of the MLE can be found and, accordingly,

$$IL(x)=[\hat{u}_L(x)'(HL_n/n)^{-1}\hat{u}_L(x)]^{1/2},$$

can be computed. For the particular case of the data set in example 3.1, the vector $x=(x_1,x_2)'$ has two coordinates, so we can transform to polar coordinates $x_1=r\cos\theta$, $x_2=r\sin\theta$ $(r>0,\ 0<\theta<\pi/2)$ and compare the qualitative behaviour of IR(x) and IL(x) with the two auxiliary curves $AR(r)=\sup_{\|x\|=r} IR(x)$ $AL(r)=\sup_{\|x\|=r} IL(x)$, where $\|x\|=[x_1^2+x_2^2]^{1/2}$. These appear in figure 4.1, where superiority of $\hat{\Lambda}_R$ over $\hat{\Lambda}$ is apparent.

Figure 4.1

5. Final comments

This paper presents methods of diagnostic and robustness for the transformation parameter Λ with multivariate data. Techniques presented are shown to be better than simultaneous application of previously suggested univariate methods. Computational issues are discussed. The ideas on robustness presented could be, in principle, adapted easily to the multiple regression case. The results of this paper would be an improvement over previous robust methods in regression because they present a procedure to compute a sampling approximation to the influence curve of the estimator.

APPENDIX

A *Expression for the determinant* (2.3). The result obviously holds for p=2. Assuming it true for p, (2.3) follows from the well-known formula of the determinant of a partitioned squared matrix and induction hypothesis.

B *Gradient and Hessian of* $L_{max}(\Lambda)$. Theorem B.1. Introduce the notation: (*i*) $M(\Lambda)=Z^{(\Lambda)'}AZ^{(\Lambda)}$; (*ii*) $W_j^{(\lambda_j)}=\partial Z_j^{(\lambda_j)}/\partial\lambda_j$; (*iii*) $U_j^{(\lambda_j)}=\partial^2 Z_j^{(\lambda_j)}/\partial\lambda_j^2$, j=1, ...,p; and (*iv*) For a nxp matrix $H=(h_1, \ldots, h_p)$, $H_j(u)$ is H with its jth column replaced by the nx1 vector u and $H_{jk}(u,v)$ is H with its jth and kth column replaced by, respectively, u and v. We have:

a) $\partial L_{max}(\Lambda)/\partial\lambda_r=-n|M(\Lambda)|^{-1}|Z_r^{(\Lambda)'}[W_r^{(\lambda_r)}]AZ^{(\Lambda)}|$, r=1, ...p;

b) $\partial^2 L_{max}(\Lambda)/\partial\lambda_r^2=$

$$-n|M(\Lambda)|^{-2}[|M(\Lambda)|(|Z_r^{(\Lambda)'}[U_r^{(\lambda_r)}]AZ^{(\Lambda)}|+|Z_r^{(\Lambda)'}[W_r^{(\lambda_r)}]AZ_r^{(\Lambda)}[W_r^{(\lambda_r)}]|) + |Z_r^{(\Lambda)'}[W_r^{(\lambda_r)}]AZ^{(\Lambda)}|^2], \quad r=1, \ldots,p;$$

c) $\partial^2 L_{max}(\Lambda)/\partial\lambda_s\partial\lambda_r=$

$$-n|M(\Lambda)|^{-2}[|M(\Lambda)|(|Z_{sr}^{(\Lambda)'}[W_s^{(\lambda_s)},W_r^{(\lambda_r)}]AZ^{(\Lambda)}|+|Z_s^{(\Lambda)'}[W_s^{(\lambda_s)}]AZ_r^{(\Lambda)}[W_r^{(\lambda_r)}]|) + |Z_s^{(\Lambda)'}[W_s^{(\lambda_s)}]AZ^{(\Lambda)}||Z_r^{(\Lambda)'}[W_r^{(\lambda_r)}]AZ^{(\Lambda)}|],$$

for r,s=1, ...,p (r≠s).

B.1 follows Velilla (1993). Now define, for i=1, ...,n, j=1, ...,p: (*i*) The array of constants $a_{ij}=\alpha_{ij}^{1/n}(\prod_{k\neq i}^{n}\alpha_{kj})^{-1/n(n-1)}$ and the functions $q_{ij}(\lambda_j)=a_{ij}^{\lambda_j-1}$; (*ii*) The nxp matrix $Z_i^{(\Lambda)}=(Z_{i1}^{(\lambda_1)}, \ldots, Z_{ip}^{(\lambda_p)})$ of jth column $Z_{ij}^{(\lambda_j)}=q_{ij}(\lambda_j)Z_j^{(\lambda_j)}$; and (*iii*) The nxn matrix $A_i=AB_iA$, where $B_i=I_n-[1-(1/n)]^{-1}e_ie_i'$ and e_i is the ith canonical vector of $\mathbb{R}^n$.

Theorem B.2. Define, for i=1, ...,n and j=1, ...,p, the nx1 vectors $W_{ij}^{(\lambda_j)}=\partial Z_{ij}^{(\lambda_j)}/\partial\lambda_j$ and $U_{ij}^{(\lambda_j)}=\partial^2 Z_{ij}^{(\lambda_j)}/\partial\lambda_j^2$. Put $\mathbf{Z}_i=\mathbf{Z}_i^{(\hat{\Lambda})}$ and write $Z_{ij}=Z_{ij}^{(\hat{\lambda}_j)}$, $W_{ij}=W_{ij}^{(\hat{\lambda}_j)}$ and $U_{ij}=U_{ij}^{(\hat{\lambda}_j)}$. The derivatives at $\Lambda=\hat{\Lambda}$ of $|M_{(i)}(\Lambda)|$ are:

a) $\partial|M_{(i)}(\Lambda)|/\partial\lambda_r=2|\mathbf{Z}_{ir}'(W_{ir})A_i\mathbf{Z}_i|$;

b) $\partial^2|M_{(i)}(\Lambda)|/\partial\lambda_r^2=2[|\mathbf{Z}_{ir}'(U_{ir})A_i\mathbf{Z}_i|+|\mathbf{Z}_{ir}'(W_{ir})A_i\mathbf{Z}_{ir}(W_{ir})|]$;

c) $\partial^2|M_{(i)}(\Lambda)|/\partial\lambda_s\partial\lambda_r=2[|\mathbf{Z}_{isr}'(W_{is},W_{ir})A_i\mathbf{Z}_i|+|\mathbf{Z}_{ir}'(W_{ir})A_i\mathbf{Z}_{is}(W_{is})|]$ (r≠s).

B.2 is based on the following proposition which is stated without proof.

Proposition B.3. For every i=1, ...,n, we can write

$$M_{(i)}(\Lambda)=\mathbf{Z}_i^{(\Lambda)'}A_i\mathbf{Z}_i^{(\Lambda)}.$$

C *Asymptotic properties of* $\hat{\Lambda}_R$. If $E_F[|\log X_j|]<\infty$, and $\theta(\Lambda)$ solves $\min|\Sigma|$ over $\theta=(\mu,\Sigma)$ $(\Sigma>0)$ restricted to $\int\rho[\{(x^{(\Lambda)}-\mu)'\Sigma^{-1}(x^{(\Lambda)}-\mu)\}^{1/2}]F(dx)=k_p$, it can be shown that if $\{G_k\}$ converges weakly to F, there exists for all k large enough, a sequence $\{\theta(\Lambda,G_k)\}$, equicontinuous on $\Lambda\in K$, which converge pointwise to $\theta(\Lambda)$. Convergence is then uniform. Application to the sequence of empirical cdf's $\{F_n\}$ yields strong consistency of $\{\hat{\Lambda}_R\}$. Differentiation under the integral sign holds if $\rho[.]$ is such that the second derivatives of the array ψ exist with respect to the parameters and, for every j and k, the moment $E_F[X_k^{-8b}|\log X_k|^8 X_j^{-8b}|\log X_j|^8]<\infty$, where b is a positive constant such that $-b\le\lambda_j\le b$, for $j=1, \ldots,p$. This condition also suffices for the influence function to exist and for the Huber's (1967) conditions to hold.

D *Computational aspects*. Array $\dot{\psi}$ has components

$$\dot{\psi}_{1,r,j}=\partial\psi_{1,r}/\partial\lambda_j+\sum_{a=1}^{p}(\partial\psi_{1,r}/\partial\mu_a)\omega_{a,j}+\sum_{\substack{a,b=1\\a\le b}}^{p}(\partial\psi_{1,r}/\partial\sigma_{ab})\omega_{ab,j},$$

$$\dot{\psi}_{2,rs,j}=\partial\psi_{2,rs}/\partial\lambda_j+\sum_{a=1}^{p}(\partial\psi_{2,rs}/\partial\mu_a)\omega_{a,j}+\sum_{\substack{a,b=1\\a\le b}}^{p}(\partial\psi_{2,rs}/\partial\sigma_{ab})\omega_{ab,j},$$

where for example,

$$\partial\psi_{1,r}/\partial\lambda_j=u'[d](\partial d/\partial\lambda_j)(x_r^{(\lambda_r)}-\mu_r)+u[d]\delta_{jr}(\partial x_r^{(\lambda_r)}/\partial\lambda_r),$$

$$\partial\psi_{1,r}/\partial\mu_a=u'[d](\partial d/\partial\mu_a)(x_r^{(\lambda_r)}-\mu_r)-u[d]\delta_{ra},$$

$$\partial\psi_{1,r}/\partial\sigma_{ab}=u'[d](\partial d/\sigma_{ab})\ (x_r^{(\lambda_r)}-\mu_r).$$

Also: $\partial d/\partial\lambda_j=(1/d)(\partial x_j^{(\lambda_j)}/\partial\lambda_j)[e_j'\Sigma^{-1}(x^{(\Lambda)}-\mu)]$, $\partial d/\partial\mu_a=(1/d)[e_a'\Sigma^{-1}(\mu-x^{(\Lambda)})]$, $\partial d/\partial\sigma_{ab}=(1/2d)(x^{(\Lambda)}-\mu)'(\partial\Sigma^{-1}/\partial\sigma_{ab})(x^{(\Lambda)}-\mu)$, for selected canonical vectors e_j and e_a. On the other hand, $\Gamma_j=(1/2)\sum_{\substack{r,s=1\\r\le s}}^{p}[2\sigma^{rs}-\delta_{rs}\sigma^{rs}]\omega_{rs,j}-\log x_j$. A symbolic differentiation code is useful when handling this expressions.

ACKNOWLEDGEMENTS

Research partially supported by Grant PB90-0266 (Spain). The author is grateful to P. Delicado for helpful computing assistance and estimulating conversations.

REFERENCES

ANDREWS, D.F., GNANADESIKAN, R. & WARNER, J.L. (1971). Transformations of multivariate data. *Biometrics*, **27**, 825-840.

ATKINSON, A. C. (1985). *Plots, Transformations, and Regression*, New York, Oxford University Press.

BOX, G.E.P. & COX, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Ser.* B, 26, 211-252.

CARROLL, R.J. (1980). A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society, Ser.* B, **42**, 71-78.

CARROLL, R. & RUPPERT, D. (1988). *Transformations and weighting in regression*. New York, Chapman and Hall.

COOK, R.D. & WANG, P.C. (1983). Transformations and influential cases in regression. *Technometrics*, **25**, 337-345.

FLEMING, W. (1977). *Functions of several variables*. New York, Springer.

HINKLEY, D.V. & WANG, S. (1988). More About transformations and influential cases in regression. *Technometrics*, 30, 435-440.

HUBER, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Stat. Probab.* 1, 221-233. Univ. California Press.

LOPUHAÄ, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics*, **17**, 1662-1683.

RUPPERT, D. (1992). Computing S estimators for Regression and Multivariate Location/Dispersion. *Journal of Computational and Graphical Statistics*, 3, 253-270.

SEBER, G.A.F. & WILD, C.J. (1989). *Nonlinear Regression*. New York: J. Wiley.

TSAI, C.L & WU, X. (1990). Diagnostics in transformation and weighted regression. *Technometrics*, 32, 315-322.

VELILLA, S. (1993). A note on the multivariate Box-Cox transformation to normality. *Statistics and Probability Letters*, **17**, 259-263.

CAPTIONS FOR FIGURES

Figure 3.1 Data of example 3.1 in logs

Figure 3.2 a) Marginal approximations of Tsai and Wu (1990) $\hat{\lambda}_{1M(i)}$ (cont. line) and $\hat{\lambda}_{2M(i)}$(dashed line); b) Marginal likelihood displacements LD_i^1 (cont. line) and LD_i^2 (dashed line); c) Combined marginal likelihood displacement $LD_i^1+LD_i^2$

Figure 3.3 a) LD_i; b) $\tilde{L}D_i$

Figure 3.4 a) LDM_i; b) $\tilde{L}DM_i$

Figure 4.1 Curves AR(r) (continuous line) and AL(r) (dashed line)

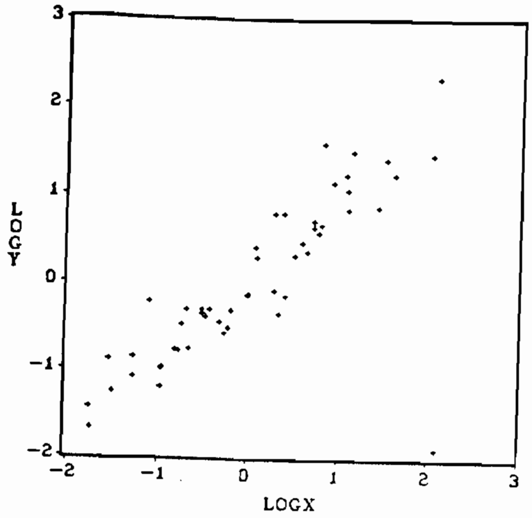
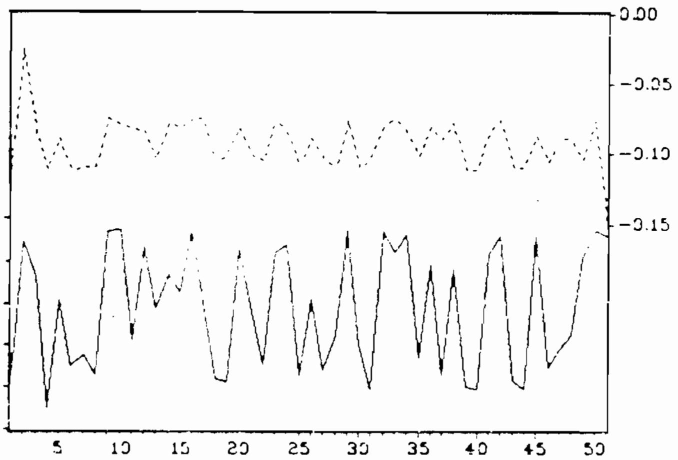
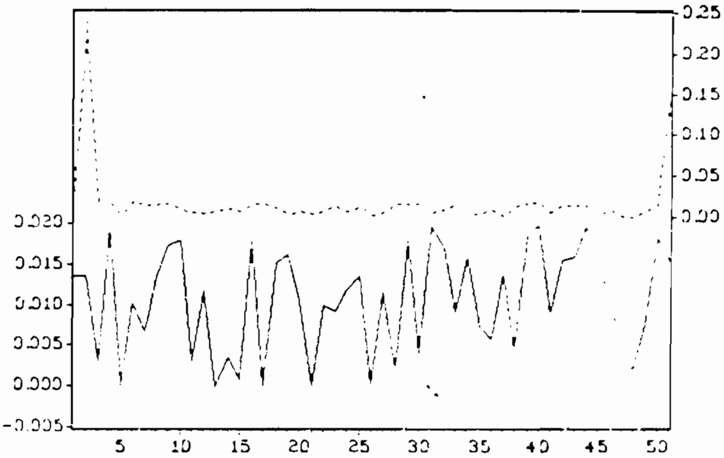


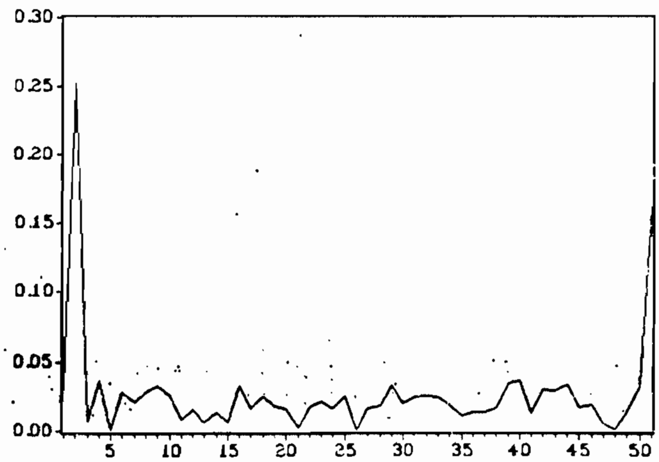
Figure 3.1



a)



b)



c)

Figure 3.2

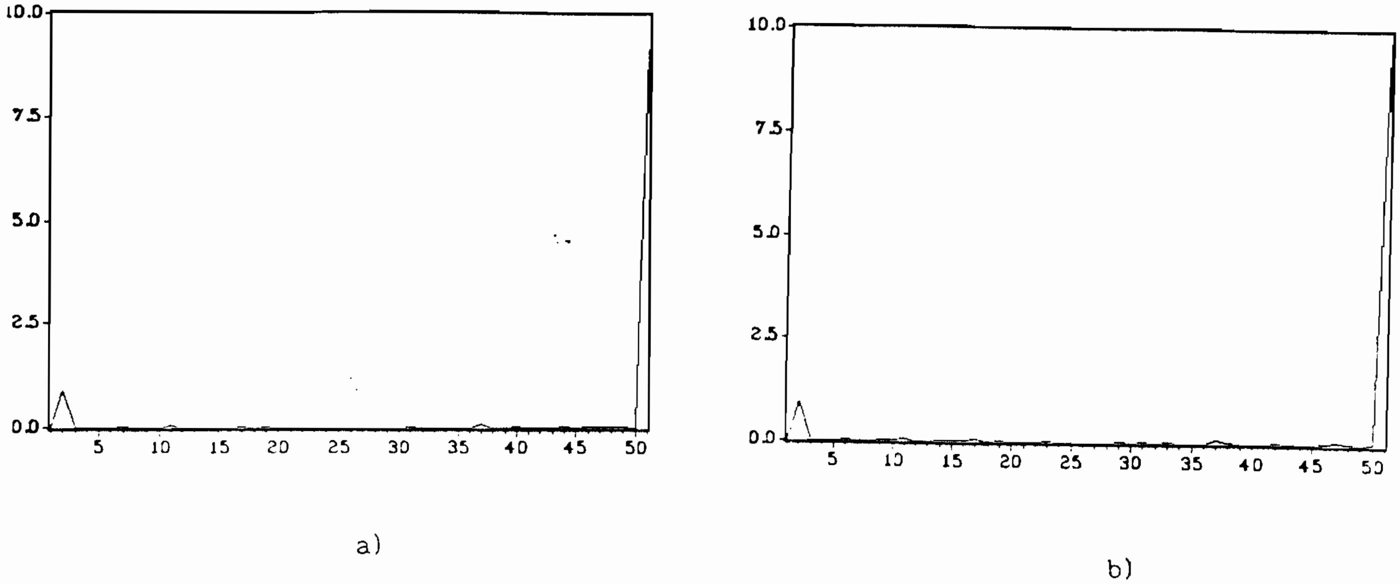


Figure 3.3

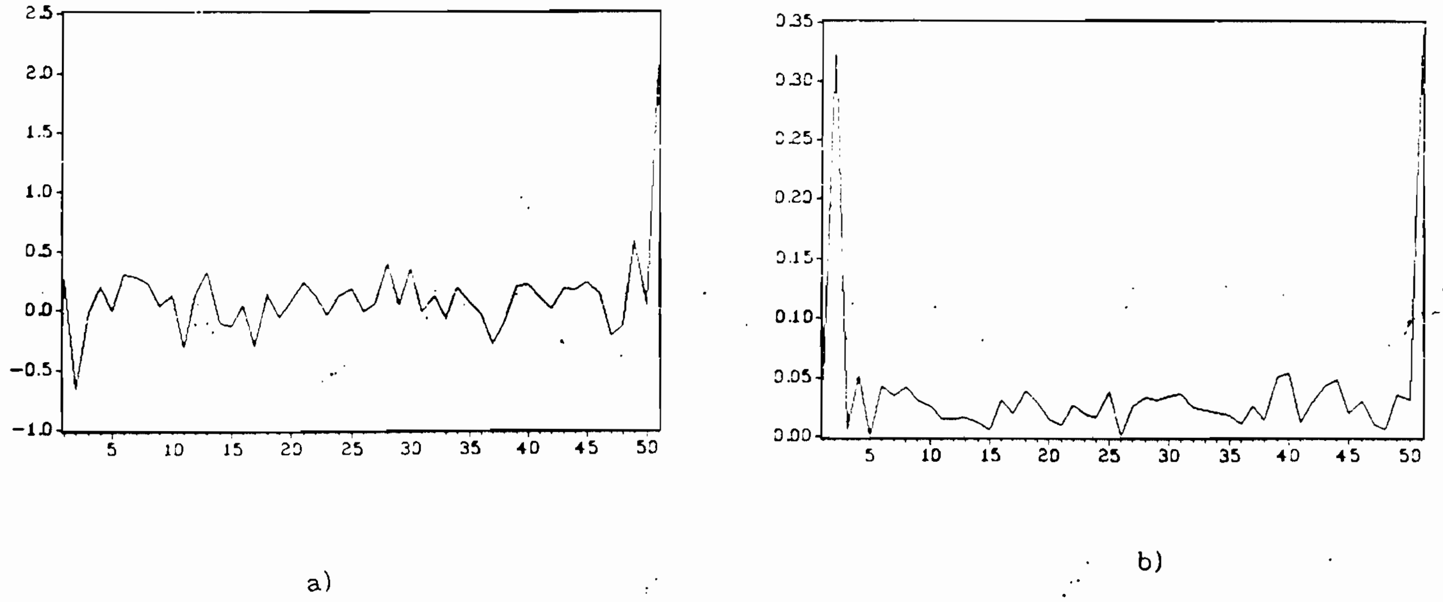


Figure 3.4

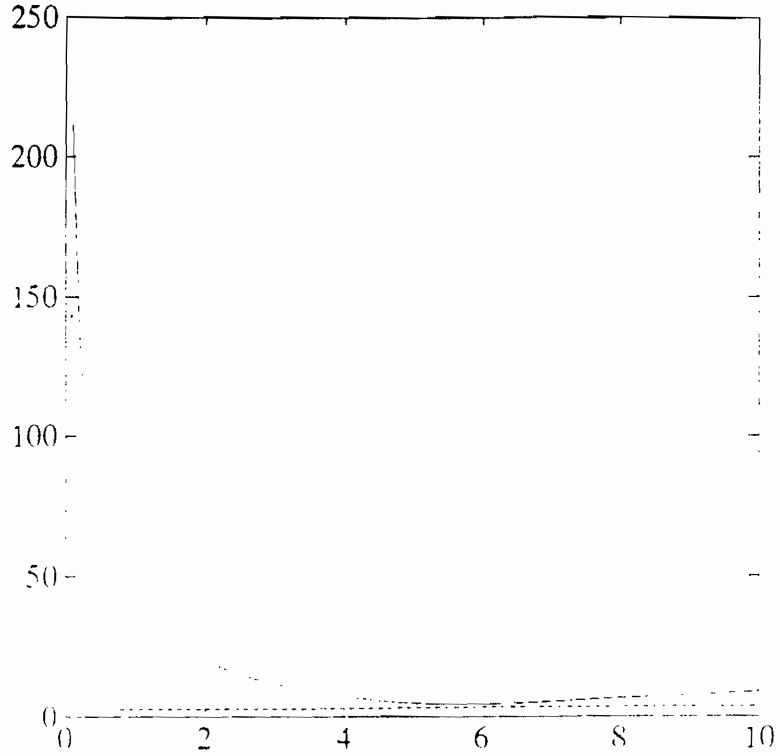


Figure 4.1