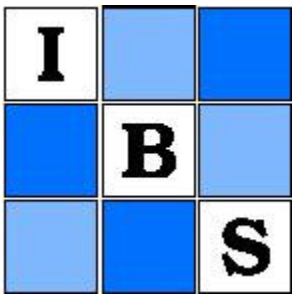


BIOMETRICS *27*, 825–40
December 1971

TRANSFORMATIONS OF MULTIVARIATE DATA

D. F. ANDREWS, R. GNANADESIKAN, AND J. L. WARNER

Bell Telephone Labs., Murray Hill, New Jersey 07974, U.S.A.

SUMMARY

Methods, which are extensions of the techniques of Box and Cox [1964], are proposed for obtaining data-based transformations of multivariate observations to enhance the normality of their distribution and also possibly to simplify the model (e.g. improve additivity, homoscedasticity, etc.). Specifically, power transformations of the original variables are estimated to effect both marginal and joint normality. A method for improving directional normality is also described. Examples are included to illustrate some properties of the methods.

1. INTRODUCTION

A well-recognized limitation of multivariate theory is the paucity of models which have been proposed and investigated as the focus for multivariate statistical methods and data analysis. The classical theory has been based largely on the multivariate normal distribution, and knowledge is sparse on the robustness of the methods to such a distributional assumption.

One way of handling this limitation has been to develop distribution-free methods for specifically posed problems such as tests of hypotheses. Although such methods may be useful for the specific purpose for which they are designed, the statistics employed by them are often of little value for summarizing the structure in a body of data. On the other hand, the serendipitous value of many of the classical methods (e.g. calculation of the mean or the covariance matrix, performing an analysis of variance, etc.) lies in their usefulness for summarizing the structure underlying data. Hence it seems reasonable and appropriate to inquire about ways of transforming the data so as to enable the use of more familiar statistical techniques that are based implicitly or explicitly on normal distributional theory. The choice of a transformation, of course, would depend on the nature of the objectives of the data analysis, and transforming to obtain more nearly normally distributed data is only one of several possible and reasonable motivations.

A transformation may be based on theoretical considerations or be bootstrapped from the particular body of data which is to be analyzed. Univariate examples of the former type are the logistic transformation for binary data (Cox [1970]) and the well-known variance stabilizing transformations for the binomial, the Poisson, and the correlation coefficient (cf. Anscombe [1948], Bartlett [1947]). Techniques for developing data-based transformations

of univariate observations have been proposed by several authors including Andrews [1971], Box and Cox [1964], Dolby [1963], Draper and Hunter [1969], Fraser [1967], Kruskal [1965], Moore and Tukey [1954], and Tukey [1949; 1957].

A central issue in the problem of transforming data is the nature of the class of transformations under consideration. Flexibility is very desirable, but conceptual and computational simplicity are not unimportant, especially in the multivariate case where one has a virtually unlimited number of transformations to consider. Interpreting analyses of multivariate data is a complex task anyway and hence, for ease of understanding as well as cumputing, it is even more crucial to remain with simple classes of transformations if at all possible.

The present paper is a preliminary report on some ongoing work concerned with statistical methods for developing relatively simple data-based transformations of multivariate observations to enhance normality of their distribution. For simplicity of exposition the detailed development is confined to the bivariate case although issues of extensions are considered and discussed in the final section. Section 2 contains a statement of the problems considered. Section 3 discusses methods appropriate for *marginal transformations* (i.e. transformations involving only one of the jointly observed variates at a time), while section 4 is concerned with a method for *joint transformations* (i.e. transformations involving more than one variate at a time). Section 5 presents some numerical examples of the application of the proposed methods. Section 6 consists of concluding discussion.

2. STATEMENT OF PROBLEMS

If $\mathbf{Y}' = (Y_1, Y_2)$ denotes the bivariate set of response variables, the general problem may be formulated as follows: to determine the vector of transformation parameters $\boldsymbol{\lambda}$ such that the transformed variables $g_1(\mathbf{Y}'; \boldsymbol{\lambda})$ and $g_2(\mathbf{Y}'; \boldsymbol{\lambda})$ are 'more nearly' bivariate normal, $N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, than Y_1 and Y_2. The elements of $\boldsymbol{\lambda}$ are unknown as are those of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Provided one can obtain an appropriate estimate, $\hat{\boldsymbol{\lambda}}$, of $\boldsymbol{\lambda}$ (as well as of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) from the data, the original bivariate observations, $\mathbf{y}_i'$, can be transformed one at a time to yield new observations, $\{g_1(\mathbf{y}_i'; \hat{\boldsymbol{\lambda}}), g_2(\mathbf{y}_i'; \hat{\boldsymbol{\lambda}})\}$, which may then be considered as more nearly conforming to a simple bivariate normal model than the original observations.

The preliminary results reported here are all concerned with transformation functions, g_i, which are direct extensions of the power transformation of a single non-negative variate X to $X^{(\lambda)}$ considered by Moore and Tukey [1954] and by Box and Cox [1964], where

$$X^{(\lambda)} = \begin{cases} (X^\lambda - 1)/\lambda & \lambda \neq 0, \\ \ln X & \lambda = 0. \end{cases}$$

Although more extensive computations would be involved in considering analogues of the more general class of shifted power transformations (i.e.

with $X + \xi$ in place of X in the above), nothing essentially new in principle is needed to extend the present approach to this more general class.

In the univariate case, after limiting oneself to the class of power transformations, various methods have been suggested for estimating the transformation parameter λ utilizing the observations on X (cf. Andrews [1971], Box and Cox [1964], Draper and Hunter [1969], Fraser [1967], and Moore and Tukey [1954]). The likelihood method, used by Box and Cox [1964] for the univariate problem, is the one adopted here for the bivariate case. (See discussion in section 6, however.)

Specifically, three approaches are discussed for the bivariate situation. Although each approach uses the likelihood method, yet the three have different objectives and properties. In each approach, both maximum likelihood (ML) estimates and the associated approximate confidence regions are obtained for the transformation parameters involved. The confidence regions may be employed for assessing the performance of the estimated transformation as well as inferences regarding the range of values of the transformation parameters which would be in concordance with the data. (Cf. discussion of Examples 1 and 2 in section 5.)

3. MARGINAL TRANSFORMATIONS

For ease of interpretation it is desirable to seek marginal transformations, i.e. transformations which operate on each of the original variables separately. In this section, the simple family of transformations defined by

$$g_i(\mathbf{Y}'; \boldsymbol{\lambda}) = Y_i^{(\lambda_i)} = \begin{cases} (Y_i^{\lambda_i} - 1)/\lambda_i & \lambda_i \neq 0 \\ \ln Y_i & \lambda_i = 0 \end{cases} \qquad i = 1, 2, \tag{1}$$

will be considered.

Two criteria are considered in the choice of the elements λ_1 and λ_2 of the vector $\boldsymbol{\lambda}$. The properties of marginal and joint normality are used in sections 3.1 and 3.2, respectively.

3.1. *Marginal normality*

A natural starting point is to choose λ_i so as to improve the marginal normality of $Y_i^{(\lambda_i)}$ for $i = 1, 2$. While it is recognized that marginal normality does not imply joint normality, it is hoped that the choice of transformations to achieve marginal normality may in many cases yield data more amenable to standard analyses.

The procedure here would be to apply the likelihood method proposed by Box and Cox [1964] to each variable separately. Two univariate computations are involved, and the theory and techniques for each are identical with those of Box and Cox [1964].

3.2. *Joint normality*

In this section, the marginal transformations in (1) are chosen with the

aim of achieving *joint* normality for the transformed data. Once again a likelihood approach will be adopted.

Consider the $n \times 2$ matrix $\mathbf{Y} = (y_{ij})$, $i = 1, \cdots, n$; $j = 1, 2$, whose rows $\mathbf{y}_i'$ are the bivariate observations, and assume that after some marginal transformation of the form in (1), the transformed data $\mathbf{Y}^{(\lambda)}$ may be described by a bivariate normal density function with mean $\boldsymbol{\mu}'$ and covariance matrix $\boldsymbol{\Sigma}$.

Let $\boldsymbol{\Xi} = E(\mathbf{Y}^{(\lambda)}) = \mathbf{1} \cdot \boldsymbol{\mu}'$. (Note: The exposition here will be restricted to the case of an unstructured sample involving n observations with a constant mean. The modification needed for extending the treatment here and in the following section to the general linear model is to change the definition of $\boldsymbol{\Xi}$ to reflect the more general case, i.e. $\boldsymbol{\Xi} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is a $n \times k$ matrix whose rows are observed or specified values of k regressor or design variables.) If $\boldsymbol{\lambda}$ is the set of marginal transformations yielding joint normality, the density function of the data $\mathbf{Y}$ is

$$f(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}n}(2\pi)^{-n} \exp\left[-\tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1}(\mathbf{Y}^{(\lambda)} - \boldsymbol{\Xi})'(\mathbf{Y}^{(\lambda)} - \boldsymbol{\Xi})\right] J,$$

where J, the Jacobian of the transformation from $\mathbf{Y}^{(\lambda)}$ to $\mathbf{Y}$, is

$$\prod_{j=1}^{2} \prod_{i=1}^{n} y_{ij}^{\lambda_j - 1}.$$

Thus the log likelihood of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\lambda}$ is given (aside from an additive constant) by

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y}) = -\tfrac{1}{2}n \ln |\boldsymbol{\Sigma}| - \tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1}(\mathbf{Y}^{(\lambda)} - \boldsymbol{\Xi})'(\mathbf{Y}^{(\lambda)} - \boldsymbol{\Xi}) + \sum_{j=1}^{2} \left[(\lambda_j - 1) \sum_{i=1}^{n} \ln y_{ij} \right]. \qquad (2)$$

For specified λ_1 and λ_2, the ML estimates of $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}$ are given, respectively, by

$$\hat{\boldsymbol{\mu}}' = \frac{1}{n} \mathbf{1}'\mathbf{Y}^{(\lambda)},$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} (\mathbf{Y}^{(\lambda)} - \hat{\boldsymbol{\Xi}})'(\mathbf{Y}^{(\lambda)} - \hat{\boldsymbol{\Xi}}),$$

where $\hat{\boldsymbol{\Xi}} = \mathbf{1} \cdot \hat{\boldsymbol{\mu}}'$. If these estimates are substituted in (2), the maximized log likelihood function (up to an additive constant) is

$$\mathcal{L}_{\max}(\lambda_1, \lambda_2) = -\tfrac{1}{2}n \ln |\hat{\boldsymbol{\Sigma}}| + \sum_{j=1}^{2} \left[(\lambda_j - 1) \sum_{i=1}^{n} \ln y_{ij} \right], \qquad (3)$$

a function of two variables which may be computed and studied. The ML estimates $\hat{\lambda}_1$, $\hat{\lambda}_2$ may be obtained by numerically maximizing (3). Also, confidence sets for λ_1, λ_2 may be obtained. One such set based on asymptotic

considerations would be those (λ_1 , λ_2) which satisfy

$$\mathfrak{L}_{\max} (\hat{\lambda}_1 , \hat{\lambda}_2) - \mathfrak{L}_{\max} (\lambda_1 , \lambda_2) \leq \tfrac{1}{2}\chi_2^2(\alpha), \tag{4}$$

where $\chi_2^2(\alpha)$ denotes the upper $100\alpha\%$ point of a χ^2 distribution with 2 D.F.

The class of transformations considered is the same in sections 3.1 and 3.2; however, the likelihood criterion used in the latter section specifies joint rather than marginal normality as the goal of the transformation. Some properties of the two methods when applied to the same body of data are illustrated in section 5.

4. JOINT TRANSFORMATIONS TO DIRECTIONAL NORMALITY

In some situations, data may exhibit non-normality only in some, and not all, directions in the space of the original variables (cf. Barnard [1962]). The method of this section is to identify such directions and to estimate a power transformation of the projections of the original observations onto these directions so as to improve conformity to bivariate normality. The specification of a direction would in general depend on both coordinates and hence the method no longer involves only marginal transformations in the sense of section 3.

Since the property of distributional normality is affine invariant, it would be desirable that statistical methods for detecting non-normality not depend on location and dispersion characteristics. One such method for selecting the directions of greatest non-normality is proposed here.

As in section 3.2, let $\mathbf{Y}$ denote the matrix whose rows $\mathbf{y}_i'$, for $i = 1, \cdots, n$, are the n bivariate observations. Let $\bar{\mathbf{y}}'$ denote the sample mean vector and $\mathbf{S} = (\mathbf{Y} - \hat{\mathbf{\Xi}})' (\mathbf{Y} - \hat{\mathbf{\Xi}})$, where $\hat{\mathbf{\Xi}} = \mathbf{1} \cdot \bar{\mathbf{y}}'$, denote the sample sum-of-products matrix. If $\mathbf{S}^{\frac{1}{2}}$ is the symmetric square root of $\mathbf{S}$, then one can obtain the set of scaled residual vectors

$$\mathbf{z}_i = [\mathbf{S}^{\frac{1}{2}}]^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}) \quad i = 1, \cdots, n. \tag{5}$$

Any non-normal characteristics of the $\mathbf{y}_i$ will be reflected in corresponding characteristics of the $\mathbf{z}_i$ which are spherically symmetric in distribution, independently of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, provided that the original distribution is normal. The direction of any non-normal clustering of points, if present, may be identified by studying a weighted sum of the $\mathbf{z}_i$,

$$\mathbf{d} = \sum_{i=1}^{n} \mathbf{z}_i \, ||\mathbf{z}_i||^{\alpha}, \tag{6}$$

where $||\mathbf{z}_i|| = (\mathbf{z}_i'\mathbf{z}_i)^{\frac{1}{2}}$, and α is a constant to be chosen.

The vector $\mathbf{d}$ provides a parametrization of directions in the z-space (and hence y-space) in terms of the single parameter α. If $\alpha = -1$, $\mathbf{d}$ is a function only of the orientation of the $\mathbf{z}$'s and gives the direction of any clustering. If $\alpha = 1$, $\mathbf{d}$ becomes sensitive primarily to those observations far from the mean. If the z's are skew in one direction, $\mathbf{d}$ will tend to point in that direction.

For a specified α, the vector $\mathbf{d}$, the direction of some non-normal characteristic of the $\mathbf{z}$'s, corresponds to the vector $\mathbf{S}^{\frac{1}{2}}\mathbf{d}$, the direction of some non-normal characteristic of the $\mathbf{y}$'s. The projection of the original observation $\mathbf{y}_i$ onto the unidimensional space specified by the direction $\mathbf{S}^{\frac{1}{2}}\mathbf{d}$ is

$$\mathbf{y}_{i\pi} = \left[\mathbf{y}_i' \cdot \frac{\mathbf{S}^{\frac{1}{2}}\mathbf{d}}{\mathbf{d}'\mathbf{S}\mathbf{d}}\right]\mathbf{S}^{\frac{1}{2}}\mathbf{d}, \tag{7}$$

for $i = 1, \cdots, n$. One can now estimate a power transformation to improve the normality of the distribution of the projections $\mathbf{y}_{i\pi}$. Since the projections constitute a univariate sample along the direction $\mathbf{S}^{\frac{1}{2}}\mathbf{d}$, the procedure is simply to apply the methods of Box and Cox [1964] to this one-dimensional sample. The effect of the transformation is to alter the data only in the direction $\mathbf{S}^{\frac{1}{2}}\mathbf{d}$.

The advantage of this method is that the relatively small class of power transformations may be applied to very complex data. The procedure may be applied iteratively, using a different value of α at each stage so as to transform along a different direction. The computations for estimating the transformations along each direction are univariate and so the extension to higher dimensions presents no crucial difficulty. With high-dimensional multivariate data, however, the chances of uncovering spurious directions of non-normality may be greater (cf. Day [1969]), and it may be argued that the method would be performing a superfluous task.

5. NUMERICAL EXAMPLES

Some examples of application of the techniques discussed in sections 3 and 4 are considered next. Both computer-generated and real data are used, the former type being useful primarily for general assessments of the statistical performance of the procedures.

For convenience in referring to the three methods, the method of Box and Cox [1964] applied to each variable separately is called Method I while the procedures described in sections 3.2 and 4 are called, respectively, Methods II and III.

Example 1

The first example consists of 50 ($= n$) sets of bivariate normal samples generated on a computer. Pairs of random standard normal deviates, (x_{1i}, x_{2i}), were transformed using the relationships

$$\begin{aligned} y_{1i} &= x_{1i} \\ y_{2i} &= \rho x_{1i} + \sqrt{(1-\rho^2)}x_{2i} \end{aligned} \qquad i = 1, 2, \cdots, 50, \tag{8}$$

to obtain the 50 samples, (y_{1i}, y_{2i}), from

$$N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].$$

To avoid negative values, the mean vector was shifted sufficiently away from the origin by adding a constant vector (c, c) to each of the observations. A range of values for ρ was employed to provide a basis for comparing the techniques.

Table 1 shows the estimates of the transformation parameters obtained by the three approaches described earlier. The actual outputs of the analyses consist not only of the ML estimates involved in each case but also, for Methods I and III, plots of the log likelihood functions involved together with the associated approximate confidence intervals, and for Method II a contour plot of the log likelihood surface displayed with the approximate confidence sets for this case. To minimize the number of displays, only a few sample plots are included here.

Over the range of 10 values of ρ shown in Table 1, it can be seen that the estimates of λ_1 and λ_2 obtained by Method I vary between 0.596 and 1.085. (Note: Because of the scheme (8) used in generating the data, the estimate of λ_1 obtained by Method I remains the same for all ρ.) Moreover, every 95% confidence interval included not only the 'true' value of $\lambda = 1$ (since the original distributions are all normal) but also every other estimate of λ. Figure 1 shows a plot of the log likelihood function of λ_2 when $\rho = 0.95$, the case in which Method I yielded the smallest (and farthest from 1) estimate of the transformation parameter.

Method II yields estimates of λ_1 and λ_2 which range between 0.878 and 1.035 and Figure 2 shows a contour plot of the log likelihood surface, defining approximate confidence regions for λ_1 and λ_2 , for the case when $\rho = 0.95$. The 95% confidence sets for all values of ρ included the point (1, 1). A very interesting feature of the results in Table 1 is the greater stability exhibited by the estimates obtained by Method II as compared to the ones yielded by Method I. The stability is particularly noticeable as ρ increases although

TABLE 1

MONTE CARLO NORMAL DATA ($p = 2$, $n = 50$)

	I		II		III†	
ρ	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}$	d'
0	0.896	0.957	0.878	0.984	0.688	−.7, .7
0.1	0.896	1.021	0.892	1.009	0.811	−.6, .8
0.3	0.896	1.085	0.890	1.035	0.714	−.9, .4
0.5	0.896	1.041	0.882	1.011	0.728	−1., .2
0.75	0.896	0.810	0.887	0.887	0.725	−.6, .8
0.8	0.896	0.745	0.886	0.852	0.729	−.5, .8
0.9	0.896	0.614	0.884	0.782	0.735	−.3, .9
0.95	0.896	0.596	0.884	0.769	0.719	−.3, .9
0.975	0.896	0.642	0.883	0.784	0.737	−.3, .9
0.999	0.896	0.833	0.884	0.859	0.715	−.6, .8

† The value of α used with Method III was 1 in all the examples in this section.

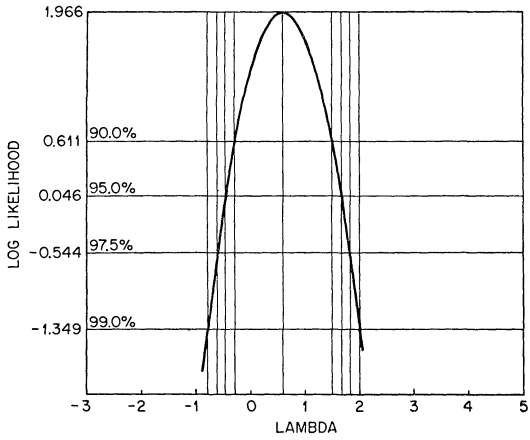


FIGURE 1

LOG LIKELIHOOD FUNCTION (METHOD I); MAXIMIZING VALUE = 0.596

it is evident even for small values of ρ. It is always legitimate to ask whether one gains anything significant by using a multivariate approach. In the present case, it seems that a multivariate approach may be able to exploit the intercorrelations among the variables to advantage and lead to more stable estimates.

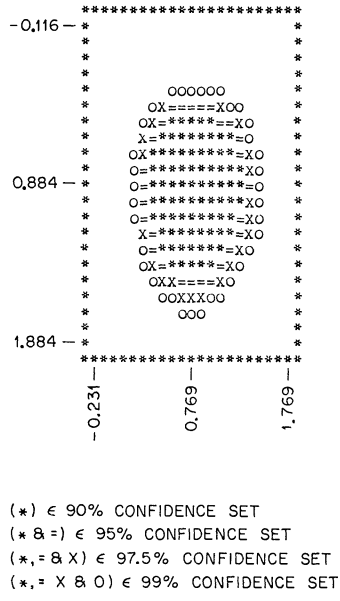


FIGURE 2

CONTOUR PLOT OF LOG LIKELIHOOD SURFACE WITH ASSOCIATED CONF. REGIONS

The results of applying Method III are also included in Table 1. The estimate of λ as well as the direction of non-normality, $\mathbf{d}'$, are shown. In this Monte Carlo 'null' example, the method appears to identify an arbitrary direction and, as seen by the $\hat{\lambda}$ values and from the fact that all the 95% confidence intervals included the value 1, the transformation has not altered the data very much.

Example 2

To try the methods next on a 'nonnull' example, 50 bivariate lognormal deviates were generated, once again employing a range of values for the underlying correlation coefficient. The results of using Methods I and II in this example are shown in Table 2. The estimates obtained in both cases are again reasonably cohesive and statistically close to the 'true' value of $\lambda_1 = \lambda_2 = 0$. The values $\lambda_1 = \lambda_2 = 1$ were far outside even the 99% confidence regions in all cases for both the methods, thus strongly indicating that the untransformed data is not bivariate normal. Furthermore, if one were to repeat the analysis with the transformed data, the point (1, 1) would be in the middle of the confidence sets obtained, thus indicating that assuming normality for the transformed data is not unreasonable.

The two methods appear to yield equally stable estimates in this example. This is not surprising since the data conform to the usual definition of a joint lognormal distribution in which the separate logarithmic transformation of each coordinate leads to variates whose joint distribution is multivariate normal. Thus the marginal transformations should yield not only marginal normality but joint normality as well.

Example 3

This illustrates the use of Method III in a 'nonnull' case, and consists of bivariate computer-generated data for which the first coordinate is dis-

TABLE 2

MONTE CARLO LOGNORMAL DATA ($p = 2$, $n = 50$)

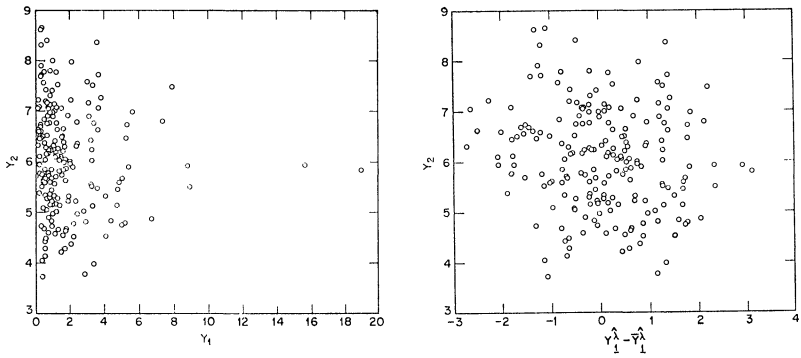
ρ	I		II	
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
0	−.004	−.017	−.008	−.015
0.1	−.004	−.005	−.001	−.004
0.3	−.004	.007	0	.008
0.5	−.004	.0002	−.005	0
0.75	−.004	−.042	−.003	−.042
0.8	−.004	−.055	−.004	−.055
0.9	−.004	−.080	−.005	−.079
0.95	−.004	−.080	−.005	−.081
0.975	−.004	−.069	−.005	−.069
0.999	−.004	−.020	−.005	−.020

tributed lognormally while the second is distributed normally independent of the first. Using Method III leads to identifying the direction of non-normality as $\mathbf{d}' = (1, 0)$, as it should in this example, and $\hat{\lambda} = -0.03$ which again is sufficiently close to zero, the value for the logarithmic transformation one would expect. Figures 3a and b show scatter plots of the data before and after transformation and the achievements of the transformation are clear.

Example 4

The data taken from Daniel and Riblett [1954] are from a 2^{8-3} fractional factorial experiment concerned with evaluating the effects of 8 factors involved in manufacturing a catalyst. Two responses called *activity* and *selectivity* were observed. It was hoped that a simple model, involving only 8 main effects, for the transformed responses would suffice. For each transformation λ, provision was made for fitting only the 8 main effects in addition to the general mean; the matrix $\hat{\mathbf{B}}$ in sections 3 and 4 consisted of the fitted values thence obtained.

The log likelihood functions involved in each of the methods were all extremely flat and the associated confidence regions were very large. This was especially true in the estimation of λ_2 by Methods I and II. Thus, in this example, the estimates of the transformation parameters in all three approaches are not sharply determined. This is perhaps not surprising since the transformations are operating on data which is 'far removed' from the origin. Specifically, the results of Draper and Cox [1969] show that the sharpness of the likelihood functions in Method I depends critically on the coefficient of variation of the data, being flat for small values and sharper



(a) (b)

FIGURE 3a
UNTRANSFORMED DATA (EX. 3)

FIGURE 3b
DATA TRANSFORMED BY METHOD III (EX. 3)

TABLE 3

DANIEL-RIBLETT [1954] DATA
(EXAMPLE 4)

I		II		III	
$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}$	$\mathbf{d}'$
−3.2	8.2	−3.1	15.9	1.62	−.8, .7

for larger values. A similar property, yet to be established formally, may be expected to hold for Method II also.

Scatter plots of the residuals (after accounting for the main effects) before and after transformations of the observations may be compared for studying the effects of the transformations. With Methods I and II one could also evaluate the accomplishments of the transformation by means of normal probability plots of the sets of residuals. Figure 4a shows a plot of 32 residuals on the untransformed scale of the activity response. The two 'smallest' points on this plot appear to depart from the configuration of the 'larger' points. Figures 4b and c are corresponding plots for the residuals on the transformed scales of this response as determined by Methods I and II, respectively. These two configurations are suggestive of more homogeneous groupings and the transformation appears to have adjusted the possibly aberrant points exhibited in Figure 4a. If the aberrant points corresponded to maverick observations, then the indication has been suppressed by the analysis on the transformed scale. Figure 4d which is for residuals in Method III, however, suggests that the transformation to directional normality seems not to deemphasize outliers in this example.

Example 5

The last example is based on data from a 2^{7-2} experiment concerned with 7 factors which affect the operation of a detergent manufacturing process. (See Roy *et al.* [1971] for more details.) The original study involved measurements on 7 responses but, for present purposes, only a bivariate subset of the original responses, viz. *rate* (bins/hr.) and *stickiness*, was considered. Once again, in applying the techniques of sections 3 and 4, allowance was made for the 7 main effects in this example. The estimates obtained by the three approaches are shown in Table 4. An indirect method of assessing the utility of the transformations obtained by Methods I and II is to study the effects of the transformations on the outputs of statistical analyses, such as analyses of variance, performed before and after the transformations. In the present example, for instance, one can obtain 31 estimated treatment effects (or contrasts) of interest for each response. A useful graphical internal comparisons technique for simultaneous assessment of the effects is a half-normal probability plot (cf. Daniel [1959]) of the absolute values of the estimated effects, or equivalently a χ_1^2 probability plot of the squared values.

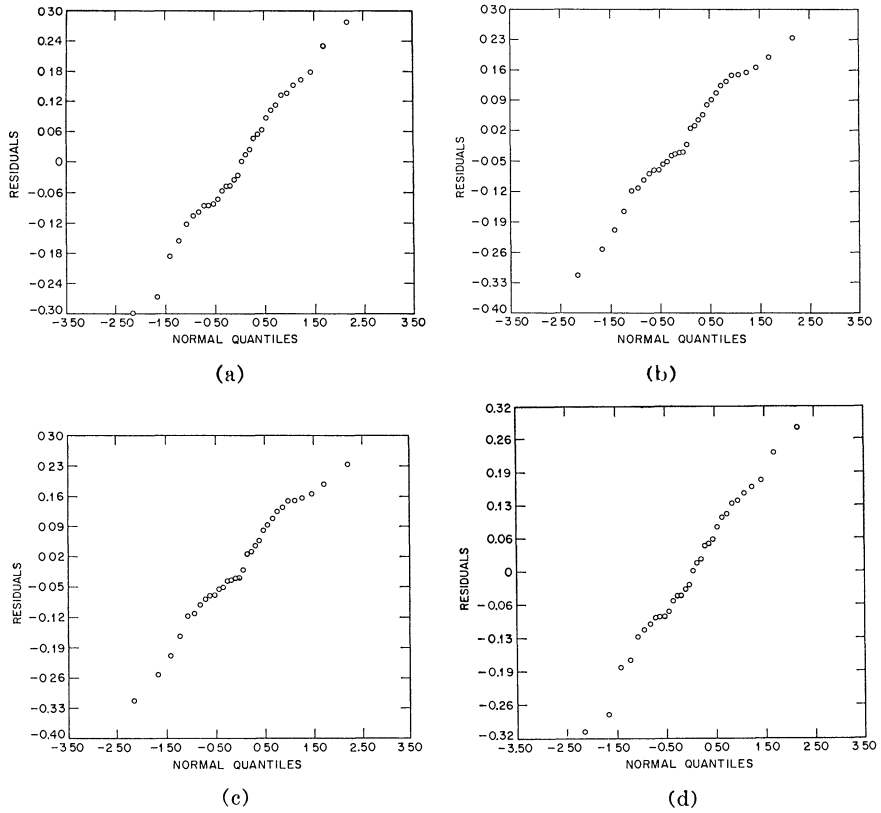


FIGURE 4

NORMAL PROBABILITY PLOT OF RESIDUALS (DANIEL-RIBLETT 'ACTIVITY' DATA (a) UNTRANSFORMED; (b) TRANSFORMED BY METHOD I; (c) TRANSFORMED BY METHOD II; (d) TRANSFORMED BY METHOD III

One can do this for effects estimated on both the untransformed and the transformed scales of the responses and compare the resulting configurations. It is perhaps reasonable to expect that, because of the averaging involved in obtaining the estimated treatment effects, except for bad non-normality of the original observations the estimated effects would be adequately normal

TABLE 4

DETERGENT MANUFACTURE DATA
(EXAMPLE 5)

I		II		III	
$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}$	d′
8.88	2.06	7.22	1.88	0.451	−0.7, −0.7

in distribution. However, in the present example, Figures 5a and b, which are χ_1^2 probability plots of the squared effects on each of the two original response scales, appear to indicate presence of considerable distributional peculiarities. The improvements achieved by using the transformations determined by Method I are evident in Figures 5c and d which show the χ_1^2 probability plots for squared effects on the transformed scales of the two variables involved. The smoother configurations of these two plots, especially at the lower end, suggest not only possible improvement of underlying normality but also the delineation of a more homogeneous grouping of smallish effects from which one can hopefully derive a 'cleaner' estimate of error variance.

A similar evaluation of Method II can be made by comparing gamma probability plots of certain squared distances associated with the bivariate effects on the untransformed and transformed scales of the two responses. (See Wilk and Gnanadesikan [1964] for a description and discussion of the

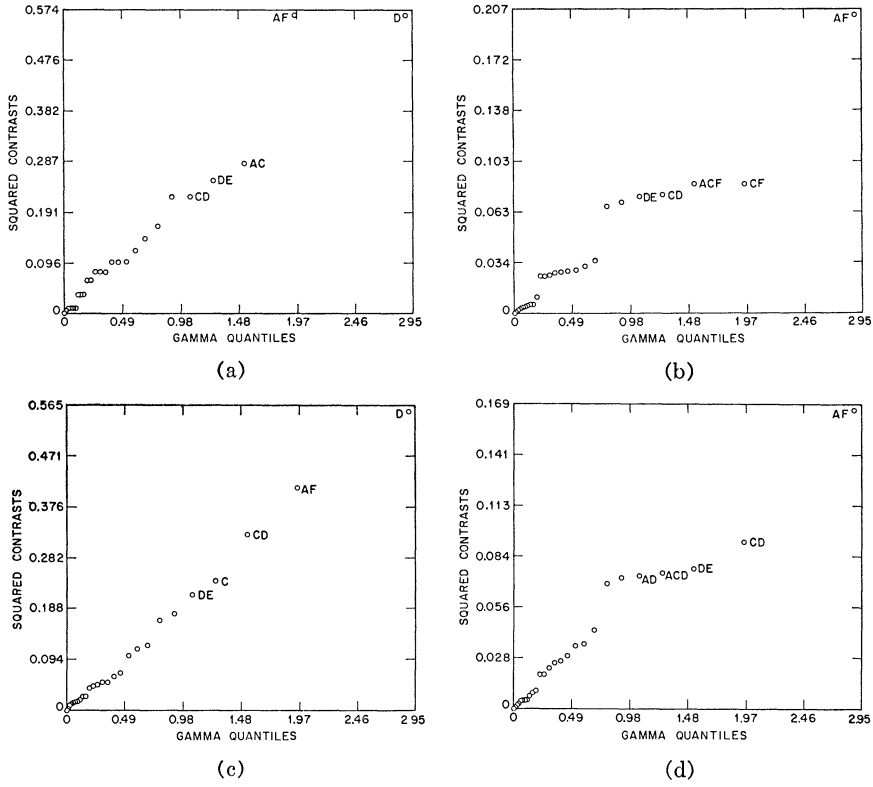


FIGURE 5

GAMMA PROBABILITY PLOT (SHAPE PARAMETER = $\frac{1}{2}$) OF SQUARED CONTRASTS (a) 'RATE' DATA UNTRANSFORMED; (b) 'STICKINESS' DATA UNTRANSFORMED; (c) 'RATE' DATA TRANSFORMED BY METHOD I; (d) 'STICKINESS' DATA TRANSFORMED BY METHOD II

method.) Figures 6a and b show the gamma plots for Example 5, the former derived from the untransformed observations while the latter is for observations transformed by using $\hat{\lambda}_1$ and $\hat{\lambda}_2$ shown for Method II in Table 4. (Note: the choice of the compounding matrix involved in the squared distances, as well as the specification of certain quantities needed for estimating the shape parameter of the evaluating gamma distribution, were such as to make the two plots quite comparable on those scores.) The null configuration of the 'smaller' points in Figure 6b is not only smoother, but also the delineation of the departures at the 'large' end is clearer. Method II appears to have improved the sensitivity of the statistical analysis. It is also worth noting that additivity has been improved by the transformations in the sense that the main effects are emphasized relative to the interactions on the transformed scales (cf. especially Figures 5c and 6b).

6. CONCLUDING REMARKS

The results presented in this paper constitute a preliminary report of some promising work in the area of transformations of multivariate data. Many other methods still need to be explored, and more experience has to be accumulated even with the methods considered in this paper. Although the present paper has emphasized the likelihood method, multivariate analogues of other univariate approaches, [e.g. Bayesian (Box and Cox [1964]), model-simplification (Draper and Hunter [1969]; Moore and Tukey [1954]), exact confidence sets (Andrews [1971])] seem to be feasible and direct.

As the dimension (p) of the data increases new problems and possibilities arise. In particular when $p > 3$, while the extension of the method of section

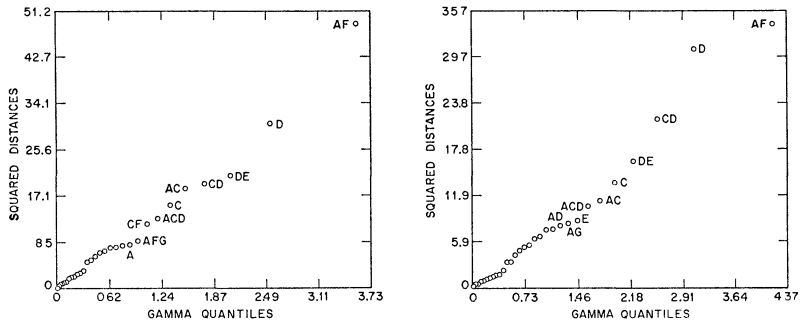


FIGURE 6a

BIVARIATE GAMMA PROBABILITY PLOT (ESTD. SHAPE PAREMETER = 0.75) OF SQD. DISTANCES (EX. 5, UNTRANSFORMED DATA)

FIGURE 6b

BIVARIATE GAMMA PROBABILITY PLOT (ESTD. SHAPE PARAMETER = 1.01) OF SQD. DISTANCES (EX. 5, DATA TRANSFORMED BY METHOD II)

3.2 involves nothing new in principle, questions of graphical representation and numerical optimization of a function of several variables become critical. A simpler, though not equivalent or sufficient, alternative approach may be to develop ways of 'combining' the estimates of transformation parameters obtained from analyzing the bivariate subsets of the variables by the method of section 3.2.

With increasing dimensionality, one may want to extend the problem in other directions. For example, the class of transformations may be extended without discarding flexibility and interpretability.

Additional criteria for choosing the particular transformations may also be desirable. For instance, a criterion based on some moderate percentage points analogous to the proposal by Tukey ([1957] p. 608) may be useful in some cases for developing methods that are less sensitive to outliers. In other cases, independence may be desirable to achieve. From a data analytic standpoint, transformations which accomplish one objective extremely well to the exclusion of all other goals are likely to be of limited value. The methods discussed in this paper have emphasized normality. However, model simplification (e.g. improving additivity, homoscedasticity, etc.) is often a more important goal in practice, and the methods here (especially the one in section 3.2) can be extended, by analogy with the suggestions of Box and Cox [1964], to aid in this process. In fact, even the present methods appear to have secondary accomplishments that may be useful in some applications (cf. Examples 4 and 5 of section 5).

The methods discussed here are not appropriate for extremely discrete (0, 1) multivariate data. In this case, theoretical considerations might suggest the use of the multivariate analogue of the logistic transformation (cf. Cox [1969]).

Finally, an issue closely related to the concerns of this paper is the one of assessing the multivariate normality of a set of observations. For this purpose, some ways of using the approximate confidence sets associated with the present methods have been mentioned. Additional statistical tests for joint normality are currently being developed.

ACKNOWLEDGMENT

The authors wish to thank Professor D. R. Cox for his helpful comments on an earlier version of this paper.

TRANSFORMATIONS DE DONNEES MULTIVARIATES

RESUME

Les auteurs proposent des méthodes, extensions des techniques de Box et Cox [1964], pour obtenir des transformations des données de base d'observations multivariates pour augmenter la normalité de leur distribution et aussi, si possible, pour simplifier le modèle (c'est à dire améliorer l'additivité, homoscédasticité, etc.). Plus précisément, des transformations à la puissance des variables originales sont estimées pour affecter à la fois la normalité marginale et liée. Les auteurs décrivent aussi une méthode pour améliorer la normalité dans une direction. Ils donnent des exemples pour illustrer quelques propriétés des méthodes.

REFERENCES

Andrews, D. F. [1971]. A note on the selection of data transformations. To appear in *Biometrika 58*.

Anscombe, F. J. [1948]. The tranformation of Poisson, binomial and negative-binomial data. *Biometrika 35*, 246–54.

Barnard, G. A. [1962]. Letter in *The Statistician 12*, 323.

Bartlett, M. S. [1947]. The use of transformations. *Biometrics 3*, 39–52.

Box, G. E. P. and Cox, D. R. [1964]. An analysis of transformations. *J. R. Statist. Soc. B 26*, 211–52.

Cox, D. R. [1969]. Discussion of paper by P. A. P. Moran, *J. R. Statist. Soc. A 132*, 521–2.

Cox, D. R. [1970]. *The Analysis of Binary Data*. Methuen, London.

Daniel, C. [1959]. The use of half-normal plots in interpreting factorial two-level experiments. *Technometrics 1*, 311–41.

Daniel, C. and Riblett, E. W. [1954]. A multifactor experiment. *Industrial and Engineering Chemistry 46*, 1465–8.

Day, N. E. [1969]. Estimating the components of a mixture of normal distributions. *Biometrika 56*, 463–74.

Dolby, J. L. [1963]. A quick method for choosing a transformation. *Technometrics 5*, 317–25.

Draper, N. R. and Cox, D. R. [1969]. On distributions and their transformation to normality. *J. R. Statist. Soc. B 31*, 472–6.

Draper, N. R. and Hunter, W. G. [1969]. Transformations: some examples revisited. *Technometrics 11*, 23–40.

Fraser, D. A. S. [1967]. Data transformations and the linear model. *Ann. Math. Statist. 38*, 1456–65.

Kruskal, J. B. [1965]. Analysis of factorial experiments by estimating monotone transformations of the data. *J. R. Statist. Soc. B 27*, 251–63.

Moore, P. G. and Tukey, J. W. [1954]. Answer to query 112. *Biometrics 10*, 562–8.

Roy, S. N., Gnanadesikan, R., and Srivastava, J. N. [1971]. *Analysis and Design of Certain Quantitative Multiresponse Experiments*. Pergamon Press, Oxford.

Tukey, J. W. [1949]. Dyadic anova, an analysis of variance for vectors. *Hum. Biol. 21*, 65–110.

Tukey, J. W. [1957]. On the comparative anatomy of transformations. *Ann. Math. Statist. 28*, 602–32.

Wilk, M. B. and Gnanadesikan, R. [1964]. Graphical methods for internal comparisons in multiresponse experiments. *Ann. Math. Statist. 35*, 613–31.

Received July 1970, *Revised February* 1971