

MASTER

TITLE: THE JOHNSON TRANSFORMATION SYSTEM REVISITED

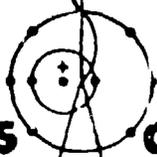
CONF. 730/103 - - 2

AUTHOR(S): Mark E. Johnson, Q-12
John S. Ramberg, University of Iowa

SUBMITTED TO:
The Eleventh Hawaii International Conference on Systems Sciences
Honolulu, Hawaii, January 5 and 6, 1978

By acceptance of this article for publication, the publisher recognizes the Government's (license) rights in any copyright and the Government and its authorized representatives have unrestricted right to reproduce in whole or in part said article under any copyright secured by the publisher.

The Los Alamos Scientific Laboratory requests that the publisher identify this article as work performed under the auspices of the USERDA.


**Los Alamos
scientific laboratory**
of the University of California
LOS ALAMOS, NEW MEXICO 87545

An Affirmative Action/Equal Opportunity Employer

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Energy Research and Development Administration, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

THE JOHNSON TRANSFORMATION SYSTEM REVISITED

by

Mark E. Johnson (1-505-667-6334)
Energy Systems and Statistics, Q-12
Los Alamos Scientific Laboratory MS 254
Los Alamos, NM 87545

John S. Ramberg (1-319-353-5264)
Systems Division
The University of Iowa
Iowa City, Iowa 52242

THE JOHNSON TRANSFORMATION SYSTEM REVISITED

Some results on the Johnson transformation system are obtained which can enhance applications of this system in multivariate Monte Carlo studies. The primary contribution is that the mean vector and the covariance matrix in the transformed population can be specified. This result is applied in a small Monte Carlo study which is devised to examine the effect of non-normality on the performance of Fisher's linear discriminant function. The observed performance conflicts in some respects with the findings of other investigators.

I. INTRODUCTION

The Johnson transformation system [3] was originally developed in a univariate setting and consisted of the normal distribution and three transformations for "normalizing" distributions. These transformations are the logarithmic, the inverse hyperbolic sine and the logit transformations. This system was later extended by Johnson [2] to a bivariate setting by applying these transformations marginally to obtain the bivariate normal distribution. In Monte Carlo robustness studies, however, the direction of application of the transformations is reversed. The inverses of these transformations are applied marginally to a multinormal variate to obtain a multivariate distribution having lognormal, inverse hyperbolic sine normal or logit normal marginals. The simulation study by Lachenbruch, Sneeringer and Revo [6], for example, use random variates obtained in this manner. In this paper attention is restricted to two distributions in the Johnson transformation system:

1. Each marginal distribution is lognormal (multivariate lognormal).
2. Each marginal distribution is inverse hyperbolic sine normal (multivariate inverse hyperbolic sine normal).

Distributions having logit-normal marginals are not considered because the expressions for the moments are intractable [3].

In section II the basic results are derived. In section III

these results are used in a small Monte Carlo study to investigate the robustness of Fisher's linear discriminant function on non-normal populations having equal covariance matrices.

II. DERIVATION OF FORMULAE

In this section the appropriate results are derived. The multivariate lognormal distribution is considered first. The goal is to determine the scale, location and correlation parameters in the multinormal population which yield upon transformation a specified mean vector and covariance matrix in the multivariate lognormal population. Let $\underline{Y} = (Y_1, Y_2, \dots, Y_n)'$ have a multivariate normal distribution with means μ_i , variances σ_i^2 , and correlations ρ_{ij} . Let \underline{X} be a random vector defined as $\underline{X} = (X_1, X_2, \dots, X_n)' = [\exp(Y_1), \exp(Y_2), \dots, \exp(Y_n)]$. The random vector \underline{X} has a multivariate lognormal distribution. The first and second order moments of \underline{X} are derived by Jones and Miller [5] as

$$E(X_i) = \exp(\mu_i + \sigma_i^2/2) \quad i=1,2,\dots,n \quad (2.1)$$

$$\text{Var}(X_i) = [\exp(2\mu_i + \sigma_i^2)] [\exp(\sigma_i^2) - 1] \quad i=1,2,\dots,n \quad (2.2)$$

$$\text{Corr}(X_i, X_j) = \frac{\exp(\rho_{ij}\sigma_i\sigma_j) - 1}{[\exp(\sigma_i^2) - 1]^{1/2} [\exp(\sigma_j^2) - 1]^{1/2}} \quad i, j=1,2,\dots,n. \quad (2.3)$$

Suppose $E(X_i)$, $\text{Var}(X_i)$ and $\text{Corr}(X_i, X_j)$ are specified as follows:

$$E(X_i) = \mu_i', \quad \text{Var}(X_i) = \sigma_i'^2 \quad i=1,2,\dots,n, \quad (2.4)$$

$$\text{Corr}(X_i, X_j) = \rho_{ij}' \quad i, j=1,2,\dots,n. \quad (2.5)$$

The parameters μ_i , σ_i^2 and ρ_{ij} determine the first and second order moments of the multivariate lognormal population. By solving for the normal population parameters in equations (2.1), (2.2) and (2.3), the following relationships can be shown to hold:

$$\mu_i = \ln[\mu_i'^2 / (\sigma_i'^2 + \mu_i'^2)^{1/2}] \quad i = 1, 2, \dots, n \quad (2.6)$$

$$\sigma_i^2 = \ln[1 + \sigma_i'^2 / \mu_i'^2] \quad i = 1, 2, \dots, n \quad (2.7)$$

$$\rho_{ij} = \frac{1}{\sigma_i \sigma_j} \ln[1 + \rho_{ij}' |\sigma_i' \sigma_j' / \mu_i' \mu_j'|] \quad i \neq j. \quad (2.8)$$

Hence, by assigning parameters in the multivariate normal distribution according to (2.6), (2.7), and (2.8), the specified moments in the transformed populations can be obtained.

Similar results are now derived for the multivariate inverse hyperbolic sine normal distribution. Let \underline{Y} be defined as before, and let \underline{X} be a vector defined as $\underline{X} = (X_1, X_2, \dots, X_n)' = [\sinh(Y_1), \sinh(Y_2), \dots, \sinh(Y_n)]'$, where $\sinh(y) = [\exp(y) - \exp(-y)]/2$.

Johnson and Kotz [4], for example, include the following results:

$$E(X_i) = \exp(\sigma_i^2/2) \cdot \sinh(\mu_i) \quad i=1, 2, \dots, n \quad (2.9)$$

$$\text{Var}(X_i) = [\exp(\sigma_i^2 - 1)] \cdot [1 + \cosh(2\mu_i) \cdot \exp(\sigma_i^2)] \quad i=1, 2, \dots, n. \quad (2.10)$$

The covariance between X_i and X_j can be derived directly using the moment-generating function of the bivariate normal to yield:

$$\begin{aligned} \text{Cov}(X_i, X_j) = \exp[(\sigma_i^2 + \sigma_j^2)/2] \cdot \{ & [\exp(\rho_{ij} \sigma_i \sigma_j) \cosh(\mu_i + \mu_j) \\ & - \exp(-\rho_{ij} \sigma_i \sigma_j) \cdot \cosh(\mu_i - \mu_j)] / 2 - \sinh(\mu_i) \cdot \sinh(\mu_j) \} \end{aligned} \quad (2.11)$$

Let μ_i' , $\sigma_i'^2$ and ρ_{ij}' denote the specified moments in the multivariate inverse hyperbolic sine distribution. The corresponding required multinormal parameters are obtained by solving equations (2.9), (2.10), and (2.11) to yield

$$\mu_i = (1/2) \cdot \cosh^{-1} \left\{ [1 + 2\mu_i' / (-\mu_i'^2 + \sqrt{\mu_i'^4 + 2\mu_i'^2 + 2\sigma_i'^2 + 1})] \right\} \quad i=1, 2, \dots, n \quad (2.12)$$

$$\sigma_i'^2 = \begin{cases} 2 \ln[\mu_i' / \sinh(\mu_i)] & \mu_i' \neq 0 \\ (1/2) \cdot \ln(2\sigma_i'^2 + 1) & \mu_i' = 0 \end{cases} \quad i=1, 2, \dots, n \quad (2.13)$$

$$\rho_{ij}' = \frac{1}{\sigma_i' \sigma_j'} \ln[(C + C^2 + 4AB)^{1/2} / 2A], \quad i \neq j \quad (2.14)$$

where

$$A = \cosh(\mu_i + \mu_j)$$

$$B = \cosh(\mu_i - \mu_j)$$

$$C = 2\rho_{ij}' \sigma_i' \sigma_j' \cdot \exp[-(\sigma_i'^2 + \sigma_j'^2)/2] + 2\sinh(\mu_i) \sinh(\mu_j)$$

Since hyperbolic cosine is an even function, two values for μ_i are generally possible in (2.12). For the variance given in (2.13) to be defined, μ_i and μ_i' must agree in sign. If the mean μ_i' is specified as zero, the mean μ_i must be set to zero.

III. APPLICATION TO A DISCRIMINANT ANALYSIS MONTE CARLO STUDY

The results obtained in the previous section are used to study the performance of Fisher's linear discriminant function (LDF) in the two population discriminant analysis problem [1]. Population one is denoted π_1 and has a known mean vector $\underline{\mu}_1$ and covariance matrix Σ_1 . Similarly, population two is denoted π_2 and has a known mean vector $\underline{\mu}_2$ and covariance matrix Σ_2 . If the populations are each governed by bivariate normal distributions with $\Sigma_1 = \Sigma_2$, Fisher's LDF, which is defined as $L(\underline{z}) = \underline{z}'\Sigma_1^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$, is the "optimal" discriminant function. It is optimal in the sense of minimizing the total probability of misclassification, which is given by

$$P[\underline{z} \text{ classified in } \pi_1 | \underline{z} \in \pi_2] + P[\underline{z} \text{ classified in } \pi_2 | \underline{z} \in \pi_1]. \quad (3.1)$$

If the normality assumption is preserved, but the equality of the covariance matrices assumption is invalid, then the optimal procedure is the quadratic discriminant function (QDF), which is defined by $Q(\underline{z}) = (\underline{z} - \underline{\mu}_2)' \Sigma_2^{-1} (\underline{z} - \underline{\mu}_2) - (\underline{z} - \underline{\mu}_1)' \Sigma_1^{-1} (\underline{z} - \underline{\mu}_1)$. Marks and Dunn [7] conclude that under the assumption of normality and unequal covariance matrices, the optimal QDF procedure performs substantially better than the LDF. If the assumption of normality were additionally violated, then the performance of the LDF could not be attributed exclusively to one of these effects. The study by Lachenbruch, Sneeringer and Revo [6], however, employs the Johnson transformation system in a fashion that the two populations are non-normal, and their covariance matrices are unequal. They compute probabilities of misclassification in each of the transformed populations using the LDF and compare them to the corresponding probabilities using the optimal procedure. The optimal

procedure is to transform back to the normal population and then to apply the LDF, since the original normal populations have equal covariance matrices. The authors observe striking imbalances in the misclassification probabilities and suboptimal performance by the LDF. They attribute this phenomena to the non-normality of the populations.

To test the authors' conclusions in a more controlled setting, a small Monte Carlo study is devised that isolates the effect of non-normality by specifying equal covariance matrices in the two non-normal populations. The results derived in the previous section lend themselves to this purpose. In Table 1 the specific population parameters used in this study are indicated. Population π_1 has a mean vector at the origin except for cases 7 and 8, which have means at $(\exp(0.5), \exp(0.5))$. This selection of parameters in cases 7 and 8 is motivated by the result that the exponential transformation of a standard normal variate yields a mean of $\exp(0.5)$. Population π_2 has a mean vector that is shifted from the population π_1 mean vector either to the right or up the diagonal $y=x$ so that the Mahalanobis distance between the two populations is $4/3$. Each component of each population has unit variance, and the correlation within a population is $1/2$.

For each of the first eight cases given in Table 1, the required parameters in the normal populations are determined by using the results in Section II. The computed parameters serve as inputs to the Monte Carlo study. For each of the populations π_1 and π_2 , 100,000 variates are generated, and the number that are misclassified into π_2 and π_1 , respectively, according to the LDF are tabulated. These results appear in Table 2.

To compute the misclassification probabilities for the optimal

procedure, some preliminary derivations are made, which lead to a more efficient Monte Carlo study. The optimal classification procedure is to apply the QDF in the pretransformed normal populations that have unequal covariance matrices. The QDF is defined as

$$Q(\underline{z}) = (\underline{z} - \underline{\mu}_2)' \Sigma_2^{-1} (\underline{z} - \underline{\mu}_2) - (\underline{z} - \underline{\mu}_1)' \Sigma_1^{-1} (\underline{z} - \underline{\mu}_1), \quad (3.2)$$

where μ_i is the mean vector and Σ_i is the covariance matrix of π_i . If

\underline{v} is defined as $\underline{v} = \Sigma_1^{-1/2} (\underline{z} - \underline{\mu}_1)$, then

$$\begin{aligned} Q(\underline{z}) &= (\underline{z} - \underline{\mu}_2)' \Sigma_2^{-1} (\underline{z} - \underline{\mu}_2) - \underline{v}' \underline{v} \\ &= (\underline{z} - \underline{\mu}_1 + \underline{\mu}_1 - \underline{\mu}_2)' \Sigma_2^{-1/2} \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} \Sigma_2^{-1/2} (\underline{z} - \underline{\mu}_1 + \underline{\mu}_1 - \underline{\mu}_2) - \underline{v}' \underline{v}. \end{aligned}$$

If $\underline{\ell}$ is defined as $\underline{\ell} = \Sigma_1^{-1/2} (\underline{\mu}_1 - \underline{\mu}_2)$, then

$$Q(\underline{z}) = (\underline{v} - \underline{\ell})' \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} (\underline{v} - \underline{\ell}) - \underline{v}' \underline{v}.$$

The matrix $\Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2}$ can be diagonalized as $P' D_\gamma P$, where D_γ is a diagonal matrix consisting of the eigenvalues of $\Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2}$, and P is an orthonormal matrix of the corresponding eigenvectors. This implies that

$$Q(\underline{z}) = (\underline{v} - \underline{\ell})' P' D_\gamma P (\underline{v} - \underline{\ell}) - \underline{v}' P' P \underline{v}.$$

Defining $\underline{\omega} = P \underline{v}$ and $\underline{k} = P \underline{\ell}$ yields

$$Q(\underline{z}) = (\underline{\omega} - \underline{k})' D_\gamma (\underline{\omega} - \underline{k}) - \underline{\omega}' \underline{\omega}.$$

By completing the square and collecting terms, the QDF can be written as

$$Q(\underline{z}) = \sum_{i=1}^2 (\gamma_i - 1) \left(\omega_i - \frac{\gamma_i k_i}{\gamma_i - 1} \right)^2 - \sum_{i=1}^2 \frac{\gamma_i k_i^2}{\gamma_i - 1}. \quad (3.3)$$

If \underline{z} belongs to π_1 , then \underline{z} has a bivariate normal distribution with mean vector $\underline{\mu}_1$ and covariance matrix Σ_1 . Thus, both \underline{v} and \underline{w} have a bivariate normal distribution with zero mean vector and identity covariance matrix. The misclassification probabilities are given as

$$\begin{aligned} P\{Q(\underline{z}) < \ln(|\Sigma_1|/|\Sigma_2|) \mid \underline{z} \in \pi_1\} \\ P\{Q(\underline{z}) > \ln(|\Sigma_1|/|\Sigma_2|) \mid \underline{z} \in \pi_2\} . \end{aligned} \quad (3.4)$$

The simulation process for estimating the probabilities in (3.4) is to generate pairs (ω_1, ω_2) of independent normal variates, to evaluate $Q(\underline{z})$ in (3.3), and to classify according to the cutoff value in (3.4). Again 100,000 pairs for each population are generated and classified, yielding the results that appear in Table 2.

Table 1

Description of the Populations

Case	π_1	π_2	Location π_2
1	S_{UU}	S_{UU}	(1.0,0.0)
2	S_{UU}	S_{UU}	(1.0,1.0)
3	S_{NN}	S_{UU}	(1.0,0.0)
4	S_{NN}	S_{UU}	(1.0,1.0)
5	S_{UU}	S_{NN}	(1.0,0.0)
6	S_{UU}	S_{NN}	(1.0,0.0)
7	S_{LL}	S_{LL}	[1+exp(0.5), exp(0.5)]
8	S_{LL}	S_{LL}	[1+exp(0.5), 1+exp(0.5)]
9	S_{NN}	S_{NN}	(1.0,0.0)
10	S_{NN}	S_{NN}	(1.0,1.0)

(S_{NN} = bivariate normal)

(S_{UU} = bivariate inverse hyperbolic sine normal)

(S_{LL} = bivariate lognormal)

Table 2

Estimates of Misclassification Probabilities

Case	LDF			Optimal Procedure		
	π_1	π_2	Average	π_1	π_2	Average
1	0.239	0.282	0.250	0.224	0.160	0.192
2	0.243	0.290	0.266	0.224	0.163	0.193
3	0.281	0.278	0.279	0.296	0.131	0.213
4	0.280	0.291	0.285	0.271	0.129	0.200
5	0.241	0.281	0.261	0.173	0.301	0.237
6	0.241	0.280	0.260	0.178	0.298	0.238
7	0.205	0.294	0.249	0.128	0.069	0.098
8	0.215	0.310	0.262	0.127	0.071	0.099
9	0.282	0.282	0.282	0.282	0.282	0.282
10	0.282	0.282	0.282	0.282	0.282	0.282

(Estimated standard deviation = 0.0015)

IV. CONCLUSIONS

As expected, the performance of Fisher's LDF (see Table 2), is sub-optimal with these populations. The imbalances in the misclassification probabilities, however, are not nearly as dramatic as those observed by Lachenbruch, Sneringer, and Revo. Certainly, Fisher's LDF is not robust against these deviations from normality (cases 1-8) despite the equal covariance matrices.

The results derived in Section II should enhance other applications of the Johnson transformation system in Monte Carlo studies.

REFERENCES

1. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, New York, 1958.
2. Higgins, J. J., "A Geometrical Method of Constructing Multivariate Densities and Some Related Inferential Procedures." Communications in Statistics, 1975, 4, 955-966.
3. Johnson, N. L., "Bivariate Distributions Based on Simple Translation Systems," Biometrika, 1949, 36, 297-304.
4. Johnson, N. L., "Systems of Frequency Curves Generated by Methods of Translation," Biometrika, 1949, 36, 149-176.
5. Johnson, N. L., and Kotz, S., Continuous Univariate Distributions, Vol. I, John Wiley & Sons, Inc., New York, 1970.
6. Lachenbruch, P. A., Sneeringer, C., and Revo, L. T., "Robustness of the Linear and Quadratic Discriminant Function to Certain Types of Non-Normality," Communications in Statistics, 1973, 1, 39-56.
7. Marks, S. and Dunn, O. J., "Discriminant Functions when Covariance Matrices are Equal," Journal of the American Statistical Association, 1974, 69, 555-559.