On S.N. Roy's Legacy to Multivariate Analysis*

Kanti V. Mardia,
Department of Statistics,
University of Leeds,
Leeds, LS2 9JT, U.K.
k.v.mardia@leeds.ac.uk

January 23, 2008

Abstract

I describe here my connection with two of the major contributions of S.N. Roy: namely the Jacobians of complicated transformations for various exact distributions, rectangular coordinates and the Bartlett decomposition. Their applications have appeared in directional statistics, shape analysis and now in statistical bioinformatics.

Keywords: Bartlett decomposition; Bioinformatics; Haar measure; Jacobian under constraints; Matrix Fisher distribution; Shape analysis.

*Presented as a plenary talk to S. N. Roy Multivariate Conference, Kolkata (December 2006)

1. INTRODUCTION

I came to know of S.N. Roy while doing my M.Sc. in Statistics (University of Bombay) in 1953-1955. Indeed, I saw him first time in 1954 when he visited the Department of Statistics, Bombay University. There was a conference of the International Institute of Statistics at the Indian Statistical Institute, Calcutta, that year, and we had various visitors who visited the Department of Statistics, University of Bombay while en route to Calcutta. What I remember vividly is that he was all dressed impressively in Bengali Dhoti and Kurta. I do not recall of his work in our M.Sc. course directly but while writing Multivariate Analysis book (Mardia, Kent and Bibby 1979), I saw his contributions more clearly and at least four topics have stuck in my mind: the invariant tests, the Jacobians of complicated transformations for various exact distributions, rectangular coordinates and the Bartlett decomposition, and inverting pattern matrices (Roy and Sarhan 1956).

I will comment here on Jacobian under constraints and the Bartlett decomposition. His contribution on the Jacobians became more clear while working with Chunni Khatri in 1975. Its impact was seen clearly in calculating the Jacobians for "parametrized" rotation matrices in terms of the generalized Eulerian angles (Khatri and Mardia 1997). Surprisingly, these have reappeared recently in a matching problem in Bioinformatics (Green and Mardia 2006). We will give some details in Section 2. Another key contribution is the use of rectangular coordinates starting from Mahalanobis, Bose and Roy (1937). These are closely related to the Bartlett decomposition and have appeared in Shape Analysis (Goodall and Mardia 1992). We will give some comments on the Bartlett decomposition in Section 3. The paper ends with discussion.

2. JACOBIAN UNDER CONSTRAINTS

Suppose in p dimensions, the orientation of the object is specified by n directions $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ with

$$XX^T = I_n, \tag{1}$$

where $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$ are n rows of X and I_n is the $n \times n$ identity matrix, $n \leq p$ (for simplicity). The Riemann space whose elements are X is called the Stiefel manifold, and we shall denote it by $O(x, p)$. For $n = p$, the Stiefel manifold becomes the orthogonal space, $O(p)$. A Haar measure of unit mass on $O(n, p)$ will be written as

$$[dX], X \, \epsilon \, O \, (n, p)$$

One of the most common distributions on X is the matrix Fisher distribution (Downs 1972; Khatri and Mardia 1977), defined as

$$a(F)\exp(\mathrm{tr}FX^T)[dX], X \,\epsilon\, O\,(n,p),$$

where F is an $n \times p$ parameter matrix.

This distribution is becoming increasingly important with its new applications in Bioinformatics (see Green and Mardia 2006). The focus has shifted on how to obtain a suitable parametrization of X. An Euler-angle representation of X turns out to be effective for calculating the Haar measure $[dX]$ as well as simulating the matrix Fisher distribution. We need to obtain the Jacobian of the Euler-angle transformation under the constraints (1). The simplification follows basically from a powerful theorem of Roy (1957, Thm A.5.5., p.165) for Jacobians under constraints; further details are given in Mardia (2006).

2. THE BARTLETT DECOMPOSITION

Let $\boldsymbol{x}_i, \quad i = 1, ..., n$ be a random sample from $N_p(\mathbf{0}, I)$ and let

$$A = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = TT^T$$

where $t_{ij} = 0, i < j, t_{ii} > 0$. Then $t_{11}, t_{21}, ..., t_{pp}$ are distributed independently as $N(0,1)$ and $t_{ii}^2, \quad i = 1, ..., p$ are independent with t_{ii}^2 has a χ^2 distribution with $n - i + 1$ degrees of freedom. This is called the Bartlett Decomposition. I have used effectively Bartlett decomposition for shape analysis with Colin Goodall (Mardia and Goodall 1991; Goodall and Mardia 1992). Bartlett had the decomposition in Wishart and Bartlett (1933) and Bartlett (1939); Mahalanobis, Bose and Roy (1937) have similar ideas using what is called rectangle coordinates or QR decomposition. We in fact used the QR decomposition!

The early history of Bartlett decomposition/Wishart distribution has been summarised by John Wishart himself in 1948 since various derivations appeared for the exact form of the distribution. The following comment of Wishart (1948, p.55) is worth bearing in mind since it emphasises the parallel work by various pioneers of that time. The problem of course fascinated several of the workers:

"At the end of 1933 Prof. Mahalanobis sent the author a somewhat fuller proof on the same lines, which was published some years later (Mahalanobis et al. 1937) as part of a long study of the normalization of statistical variates and the use of rectangular co-ordinates in the theory of sampling distributions. A proof on entirely different lines had been published before the

communication referred to from Prof. Mahalanobis was received (Wishart and Bartlett 1933)."

John Gower and I organized a conference on "Multivariate Analysis and its Applications" at Hull University in 1973. A report was subsequently published in the Journal of the Royal Statistical Society Series C (Gower and Mardia 1973). Professor M.S. Bartlett gave the first paper entitled "Some historical remarks and recollections in multivariate analysis" giving some comments on the early development of distributional results in multivariate analysis. This paper was subsequently published in Sankhyā (Bartlett, 1974). As was appropriate from someone who himself was directly involved in the early development of distributional results in multivariate analysis. In 1928, Wishart had developed the now what called Wishart distribution of the samples covariance matrix from a multivariate normal distribution (with zero correlations). To quote from Gower and Mardia (1973, p.61):

"This work was ultimately derived from Fisher's treatment of the bivariate case in 1918 when he was obtaining the distribution of the correlation coefficient in the null case, and so takes us back to the Biometric school's pre-occupation with the correlation coefficient. During the early 1930's, Hotelling and Wilks were concerned with multivariate generalizations of the t and F-distributions and this was very closely linked with Mahalanobis's development of D^2. In 1933, Professor Bartlett was concerned with the relationship of these new tests with the linear models of regression theory (Bartlett 1933). Fisher's development of discriminant analysis with Hotelling's canonical correlation analysis (which includes discrimination when one set of variables are dummies) extended the class of models for which distributional results were required, and brought to light problems concerning the distribution of the latent roots of a covariance matrix. This culminated in 1939 with the publication of four papers all giving the simultaneous distribution of the latent roots of a covariance matrix in the null case (Fisher, Hsu, Roy, Girshick)."

Note that S.N. Roy was one of the pioneers in these parallel developments. Then Bartlett goes on to comment on Rao (1972) paper related to Fisher's contributions to multivariate analysis (Bartlett 1974, p.108):

"Rao (1972) has rightly drawn attention to Fisher's vital role in these first developments, though he would no doubt agree with the linking of the names of Hotelling and Mahalanobis with that of Fisher in the key developments in the 30's."

In my opinion, this comment must include the development in the paper of 1937 by Mahalanobis, Bose and Roy.

3. DISCUSSION

Bartlett (1974, p.174) starts with optimistic future for multivariate analysis.

"The outburst of activity in multivariate analysis in the last decade has been obviously influenced to a large extent by the development of computers. In particular, the exploratory numerical investigations classifiable under the general title of cluster analysis are closely linked with the availability of computer algorithms. Such algorithms have clearly been developed in response to wide demand, and the empirical and *ad hoc* elements present in their design are being gradually reduced, or at least are becoming more appreciated."

Cluster analysis, graphical models, projection pursuit, data mining are some of the areas of multivariate analysis which are either new or where there is a tremendous activity (see Mardia 2004). However, the debate of model based statistics vs "algorithmic" statistics, a hybrid statistics or a holistic statistics is very much with us (Mardia and Gilks 2005). This debate is clearly seen in Bioinformatics where we are endowed with large scale data. Bioinformatics applications have revived the subject of Directional Statistics; the previous innovations came mainly from Earth Science applications. Directional Statistics has also appeared in the new field of Shape Analysis (see Dryden and Mardia 1998). When I gave a talk on the Bayesian alignment methods in Stanford University last year, Ted Anderson aptly remarked that "this is a new Multivariate Analysis"!

Mardia and Gilks (2005) have identified three themes for statistics in the 21st century. First, statistics should be viewed in the broadest way for scientific explanation or prediction of any phenomenon. Second, the future of statistics lies in a holistic approach to interdisciplinary research. Third, a change of attitude is required by statisticians - a paradigm shift - for the subject to go forward.

REFERENCES

Bartlett, M. S. (1933), "On the theory of statistical regression," *Proceedings of the Royal Society of Edinburgh*, **53**, 260-283.

——— (1939), "A note on tests of significance in multivariate analysis," *Proceedings of the Cambridge Philosophical Society*, **35**, 180-185.

——— (1974), "Some historical remarks and recollections on multivariate analysis," *Sankhyā*, **36**, 107-114.

Downs, T. D. (1972), "Orientation statistics," *Biometrika*, **59**, 665-676.

Dryden, I. L. and Mardia, K. V. (1998), *Statistical Shape Analysis*, Chichester: Wiley.

Goodall, C. and Mardia, K. V. (1992), "The noncentral Bartlett decomposition and shape densities," *Journal of Multivariate Analysis*. **40**, 94-108.

Gower, J. C. and Mardia, K. V. (1973), "Multivariate analysis and its applications: A report on the Hull Conference, 1973," *Journal of the Royal Statistical Society Series C*, **23**, 60-66.

Green, P. J. and Mardia, K. V. (2006), "Bayesian alignment using hierarchical models, with applications in protein bioinformatics," *Biometrika*, **93**, 235-254.

Khatri, C. G. and Mardia, K. V. (1977), "The von Mises-Fisher Matrix distribution in orientation statistics," *Journal of the Royal Statistical Society Series B*, **39**, 95-106.

Mahalanobis, P. C., Bose, R. C. and Roy, S. N. (1937), "Normalisation of statistical variates and the use of rectangular coordinates in the theory of sampling distributions," *Sankhyā*, **3**, 1-40.

Mardia, K. V. (2004), "Past revolutions and future prospects in science and statistics," in *Proceedings of the International Sri Lankan Statistical Conference: Visions of Futuristic Methodologies*, eds. B. M. de Silva, and N. Mukhopadhyay, 17-34, Melbourne: Sri Lanka and RMIT University.

——— (2006), "Jacobians under constraints and statistical bioinformatics," To appear.

Mardia, K. V. and Gilks, W. (2005), "Meeting the statistical needs of 21st-century science," *Significance*, **2**, 162-165.

Mardia, K. V. and Goodall, C. (1991), "A geometrical derivation of the shape density," *Advances in Applied Probability*, **23**, 496-514.

Mardia, K. V. and Jupp, P. E. (2000), *Directional Statistics*. Wiley, Chichester.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

Rao, C. R. (1972), "Recent trends of research work in multivariate analysis," *Biometrics*, **28**, 3-22.

Roy, S. N. (1957), *Some Aspects of Multivariate Analysis*, New York: Wiley.

Roy, S. N. and Sarhan, A. E. (1956), "On inverting a class of patterned matrices," *Biometrika*, **43**, 227-231.

Wishart, J. (1948), "Proofs of the distribution law of the second order moment statistics," *Biometrika*, **35**, 55-57.

Wishart, J. and Bartlett, M. S. (1933), "The distribution of second order moment statistics in a normal system," *Proceedings*

of the Cambridge Philosophical Society, **28**, 455-459.