Proc. Indian Acad. Sci. (Math. Sci.) (2020) 130:46
https://doi.org/10.1007/s12044-020-00586-4

C R Rao and Mahalanobis' distance

PROBAL CHAUDHURI

Indian Statistical Institute, 203, B. T. Road, Kolkata 700 108, India
E-mail: probalchaudhuri@gmail.com

Abstract. C R Rao has made seminal contributions in many areas in statistics. A review of some of his fundamental contributions towards the development of the theory and application of Mahalanobis distance in classification problems is presented here. The review is based on information provided in some of Rao's famous research and autobiographical papers.

Keywords. Bayes risk optimality; discriminant analysis; Gaussian populations; linear discriminant function.

Mathematics Subject Classification. 62-03, 01A32, 01A6, 62H30.

1. Introduction

Prasanta Chandra Mahalanobis met Nelson Annandale at the Indian Science Congress held in Nagpur in the year 1920. During their conversation, Annandale asked Mahalanobis for help in analyzing certain anthropometric measurements of Anglo-Indians living in Calcutta. While doing that analysis, Mahalanobis invented his famous D^2 statistic, which is now well-known as the Mahalanobis distance. For two multivariate populations with means μ_1 and μ_2 and a common dispersion matrix Σ, D^2 is defined as $D^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$ (see [2]). It so happened that C R Rao was born in the same year in Huvvina Hadagalli, a small town in Karnataka. So, in a sense, 2020 is the birth centenary year for C R Rao as well as the D^2 statistic.

Rao graduated with a B.A.(Hons) degree in mathematics from Andhra University in 1940 but failed to obtain a scholarship to do research in mathematics in the same university because his application was late. He then came to Kolkata to appear for an interview for the position of a mathematician at the army survey unit, and was not selected for the job either. Later, when Calcutta University started a Master's degree programme in statistics in 1941, he joined that as a student. He passed the M.A. degree examination in 1943 with a first class and securing the first rank. Rao was one of the first five students, who graduated from the Calcutta University with Master's degree in statistics. Rao [5] gave an entertaining description of that post-graduate degree programme in statistics in Calcutta University that is worth quoting here. "None of the teachers had any experience in teaching statistics, and further they were as ignorant as the students in some areas of statistics. As there were no text books on statistics, the teachers had to learn by reading original papers and then teach. The courses given in the first two years benefitted the faculty as well as the students."

This article is part of the "Special Issue in Honour of Professor C R Rao on His Birth Centenary"

Published online: 25 July 2020

In his final year, Rao submitted a thesis in lieu of two practical papers. In that thesis, he developed and studied, what he called the "perimeter test", which is a test for comparing the means of several multivariate populations, when their dispersions are known or can be estimated from very large samples drawn from each of these populations. For comparing the means of two multivariate populations, the standard test is the Hotelling's T^2 test, which is essentially based on the D^2 statistic. Rao's perimeter test is based on a weighted sum of all D^2 statistics for pairs of populations, and it has a chi-square distribution with appropriate degrees of freedom. This was Rao's first involvement with Mahalanobis' distance as a young statistician, and it continued for many years leading to many fundamental discoveries related to D^2.

Mahalanobis offered jobs to all of the first batch of five students from the Calcutta University, who graduated with the Master's degree in staistics. Rao joined Indian Statistical Institute as a statistical apprentice in 1943 and decided to settle down in the city of Kolkata. He started teaching soon after joining the Institute, and it was in the course of his interactions with the students at the time, when he invented his famous Cramér–Rao Lower Bound and Rao–Blackwell theorem (see [5]).

2. C R Rao's Ph.D. thesis

In Indian Statistical Institute, one of Rao's several responsibilities was the analysis of anthropometric data on different castes and tribes collected during the decennial census of Indian population in 1941. In 1946, J. C. Trevor, a Cambridge University anthropologist, requested Mahalanobis to depute a scholar from Indian Statistical Institute, who would analyze some anthropometric data collected by Cambridge University Anthropological Museum using Mahalanobis' distance. Mahalanobis deputed Rao along with another person, and they arrived in England in the August of 1946. Rao [5] has mentioned that they sailed for England by a steamship named *Andes* that was taking mostly Italian prisoners of the Second World War from Mumbai to Naples. The Italian soldiers were captured in Africa and detained in India.

Soon after arriving in Cambridge, Rao took admission as a research scholar in King's College, where Mahalanobis also studied as a Tripos student. This was in addition to Rao's assignment at the Anthropological Museum at Duckworth Laboratory at Cambridge. Fisher was the Balfour Professor of Genetics at Cambridge at that time. Rao requested Fisher, whom he met in 1944, when Fisher visited India, to accept him as a PhD student. Fisher agreed and suggested that he should work in the genetics laboratory to gain experience with breeding of mice for linkage studies. On the other hand, Rao's assignment at the Anthropological Museum involved analysis of measurements taken on skeletal materials excavated by a British expedition in the ancient graves in Jebel Moya in North Africa. Rao [5] wrote: "I kept myself busy in Cambridge dividing my time between bones and stones at the University Anthropological Museum and mice at Whittingham Lodge, the official residence of the Balfour Professor of Genetics converted into a genetics laboratory."

In his seminal paper "The Use of Multiple Measurements in Taxonomic Problems" published by Fisher in 1936 in *Annals of Eugenics*, he developed the concept of linear discriminant analysis for two multivariate populations. Fisher [1] constructed his linear discriminant function by maximizing the ratio of the 'between population variation' to the 'within population variation' over different choices of the linear functions of multivariate measurements. Clearly, Fisher [1] was motivated by ideas from analysis of variance

and thought that the best linear discriminator should have the highest F-ratio among all possible linear functions of the variables. Fisher's linear discriminant function relates to Mahalanobis' distance in an interesting way. Suppose that we define the Mahalanobis' distance between a multivariate observation x from a multivariate population with mean μ and dispersion Σ using $D^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$. Then it is straight-forward to verify that for two multivariate populations that have a common dispersion matrix and differ only in their means, Fisher's linear discriminant function classifies the observation x into the population, which is closer to x in Mahalanobis' distance.

The general theme of Rao's PhD research was discriminant analysis using multivariate data, and it was deeply connected with Fisher's work on discriminant analysis and Mahalanobis' distance. The research was motivated by his work at the Cambridge Anthropological Museum. An illustrative example mentioned by Rao [3] is worth mentioning here. "In August, 1939, a relatively complete male human skeleton was recovered from the ditch of an Iron Age camp on Highdown Hill, Goring by Sea, in the course of the excavations conducted under the auspices of the Worthing Archaeological Society. Fragments associated with the bones suggested that burial could not have taken place later than the Iron Age in Sussex, about 500 B.C. The camp went out of use no later than 250 B.C., and the remains themselves can be assigned to a 500 B.C. "invasion" horizon. It is doubtful, however, whether their owner was a Bronze Age "defender" or an Iron Age "invader". The principal question to be considered, in the present context, is whether the Highdown skull is more likely to have belonged to the Bronze Age or to an Iron Age population."

Towards the end of his two-year long stay at Cambridge, Rao met H E Daniels, who asked him to submit a discussion paper to the Royal Statistical Society based on his PhD research. The paper that was submitted by Rao had the title "The Utilization of Multiple Measurements in Problems of Biological Classification", which was very similar to the title of the 1936 *Annals of Eugenics* paper by Fisher, mentioned above. Rao's paper was accepted for a discussion at the Royal Statistical Society, and it was read before the Research Section of the society in 1948. It was also published in the *Journal of the Royal Statistical Society, Series B (Methodological)* in the same year. Rao [3] extended Fisher's discriminant analysis for more than two populations. He also established the Bayes risk optimality of Fisher's linear discriminant function for multivariate Gaussian populations that have same dispersion matrix and differ only in their means. Another siginificant part of Rao's paper is the development of a graphical tool for visual representation of multivariate populations by projecting them into some appropriate lower dimensional spaces so that observations from different populations cluster into groups of similar populations. Both of Fisher's 1936 paper and Rao's 1948 paper are fundamental contributions in discriminant analysis that have influenced several generations of research in multivariate statistics.

While discussing his PhD research, Rao [5] wrote: "For my PhD thesis, I wanted a theoretical topic. I asked Fisher to suggest some problem on which I could work. He said: "Problem must be yours and I shall help you if I can". I wrote my PhD thesis on classification problems extending Fisher's work on the discriminant function to more than two populations, which arose in my work at the museum. I did not take much help from Fisher. Finally when I showed him the thesis, he said : "The problem was worth investigating". I have been telling my research students what Fisher told me that they should choose their own problems for research. I could succeed only in a very few cases. $\cdots$ Fisher's students generally worked on problems in genetics. I was, perhaps, the only one, who worked and wrote a thesis in statistics under his guidance for the PhD degree of Cambridge University."

3. Discrimination in infinite dimension

After returning to India from Cambridge, Rao joined the Indian Statistical Institute as a professor, when he was only 28 years old. While working on statistical analysis of data collected in several anthropometric surveys, Rao became interested in discrimination problems involving very large number of variables and relatively fewer cases. He collaborated on this with the famous probabilist V. S. Varadarajan, who was a research scholar at Indian Statistical Institute in the early sixties and was specializing in stochastic processes. Rao and Varadarajan [4] considered the problem of discriminant analysis of Gaussian processes in infinite dimensional Hilbert spaces.

It is a well-known fact that any two Gaussian distributions with positive definite covariance matrices in finite dimensional spaces are mutually absolutely continuous. It follows from Rao [3] that the optimal Bayes classifier to discriminate between two such Gaussian distributions is Fisher's linear or quadratic discriminant function depending on whether the two Gaussian distributions have the same or different covariance matrices, respectively. However, in infinite dimensional Hilbert spaces, two Gaussian probabilities with positive definite covariance operators will be either mutually orthogonal or mutually absolutely continuous. From the point of view of discriminating between two Gaussian distributions, the case when they are mutually orthogonal is trivial and uninteresting. Rao and Varadarajan [4] derived the necessary and sufficient condition for two Gaussian probabilities in Hilbert spaces to be mutually absolutely continuous. For instance, if we have two Gaussian probabilities with means μ_1 and μ_2 and a common positive definite covariance operator Σ, they will be mutually absolutely continuous if and only if $\mu_1 - \mu_2$ lies in the range space of $\Sigma^{1/2}$. If this condition holds, there is a unique δ such that $\Sigma^{1/2}\delta = \mu_1 - \mu_2$, and we can write $\delta = \Sigma^{-1/2}(\mu_1 - \mu_2)$. Then one can define Mahalanobis distance between the two Gaussian distributions as $\|\Sigma^{-1/2}(\mu_1 - \mu_2)\|$. It follows from Rao and Varadarajan [4] that in this case, Mahalanobis distance becomes a monotonically increasing function of the well-known Hellinger distance between the two Gaussian probabilities. Further, the optimal Bayes classifier to discriminate between two such Gaussian distributions, which is the likelihood ratio classifier based on the Radon–Nikodym derivative of one Gaussian distribution with respect to the other one, is the natural extension of Fisher's linear discriminant function for such infinite dimensional problems.

4. The birth centenary year

C R Rao was born in 1920, when the infamous Spanish Flu pandemic had just subsided, and the First World War was coming to an end. 2020, his birth cenetenary year, happens to be a year that will be remembered in the history of mankind for several landmark events. Coronavirus pandemic, which is currently ravaging the world, infecting a few millon people and killing thousands of them, might top the list of events for which we shall remember 2020 forever. We shall all be living in a socially distanced world and struggling to get used to new normals for weeks and months to come. The world is probably never going to be the same even after this pandemic subsides at some point of time in the future.

While scientists are at a loss and looking helplessly at the rising death toll due to coronavirus infection all over the world, the world is being shaken by riots and protests arising from the killing of George Floyd by some police officers in Minnesota in May 2020. Rao [5] wrote the following while reminiscing about his days in Cambridge during 1946–48: "J Wishart was working in the School of Agriculture, which was next door

to the museum, and I had the opportunity to meet him frequently and discuss statistical problems. I remember one occasion when he told me that Professor H Hotelling was thinking of offering me a position at the University of North Carolina and wrote to him to check up on my skin colour, as there was difficulty in hiring dark coloured persons at the University of North Carolina."

References

[1] Fisher R A, The use of multiple measurements in taxonomic problems, *Ann. Eugenics* **19** (1936) 179–188

[2] Mahalanobis P C, On the generalized distance in statistics, *Proc. Natl Inst. Sci., India* **12** (1936) 49–55

[3] Rao C R, The utilization of multiple measurements in problems of biological sciences, *J. R. Stat. Soc., Series B* **10** (1948) 159–203

[4] Rao C R and Varadarajan V S, Discrimination of Gaussian Processes, *Sankhya* **25** (1963) 303–330

[5] Rao C R, Statistics as a last resort, Glimpses of India's Statistical Heritage, edited by J K Ghosh, S K Mitra and K R Parthasarathy (1993) (Wiley Eastern Limited)