# ASYMPTOTIC THEORY FOR PRINCIPAL COMPONENT ANALYSIS[1]

By T. W. Anderson

*Columbia University*

**0. Summary.** The asymptotic distribution of the characteristic roots and (normalized) vectors of a sample covariance matrix is given when the observations are from a multivariate normal distribution whose covariance matrix has characteristic roots of arbitrary multiplicity. The elements of each characteristic vector are the coefficients of a principal component (with sum of squares of coefficients being unity), and the corresponding characteristic root is the variance of the principal component. Tests of hypotheses of equality of population roots are treated, and confidence intervals for assumed equal roots are given; these are useful in assessing the importance of principal components. A similar study for correlation matrices is considered.

**1. Introduction.** Let $x$ be a $p$-component random vector[2] with mean vector $\mathcal{E}x = \mu$ and covariance matrix $\mathcal{E}(x - \mu)(x - \mu)' = \Sigma$ (where the prime denotes the transpose of the vector). The variance of a linear combination $\gamma' x$ is

$$\mathcal{E}(\gamma' x - \mathcal{E}\gamma' x)^2 = \mathcal{E}[\gamma'(x - \mu)]^2 = \mathcal{E}\gamma'(x - \mu)(x - \mu)'\gamma = \gamma'\Sigma\gamma. \tag{1.1}$$

The linear combination normalized by $\gamma'\gamma = 1$ which has maximum variance may be called the first principal component of $x$. The linear combination uncorrelated with the first principal component and similarly normalized which has maximum variance may be called the second principal component. The other $p - 2$ principal components are similarly defined. (See [2], Chapter 11, for more detail; Hotelling [9] developed much of the theory.)

To give these linear combinations precisely we use the characteristic roots and vectors of $\Sigma$. Let $\delta_1 \geqq \cdots \geqq \delta_p > 0$ be the $p$ characteristic roots of $\Sigma$ (assumed to be positive definite). They are the roots of

$$|\Sigma - \delta I| = 0. \tag{1.2}$$

Let $\gamma_1, \cdots, \gamma_p$ be the corresponding normalized characteristic vectors; they satisfy

$$\Sigma\gamma_i = \delta_i\gamma_i, \tag{1.3}$$

$$\gamma_i'\gamma_i = 1. \tag{1.4}$$

Received June 26, 1961; revised July 19, 1962.

[1] This research was sponsored by the Office of Naval Research under Contract Number Nonr 3279(00) and Nonr 266(33). Reproduction in whole or in part is permitted for any purpose of the United States Government. Some of the research in this paper was conducted at Stanford University.

[2] Latin letters denote random variables and running variables, and Greek letters denote parameters. Scalars and vectors are indicated by lower case letters, and matrices are indicated by capital letters. To the extent of convenience the estimate of a parameter is indicated by a Latin letter corresponding to the Greek letter for the parameter.