Multivariate Analysis

Dietrich von Rosen

Swedish University of Agricultural Sciences, Uppsala, Sweden
Linköping University, Sweden

Abstract

Multivariate statistical analysis has a long history, but most of us probably do not have a clear picture of when it really started, what it was in the past and what it is today. In the present introduction we give a few personal reflections about some areas which are connected to the analysis based on the dispersion matrix or the multivariate normal distribution, omitting a discussion of many "multivariate areas" such as factor analysis, structural equations modelling, multivariate scaling, principal components analysis, multivariate calibration, cluster analysis, path analysis, canonical correlation analysis, non-parametric multivariate analysis, graphical models, multivariate distribution theory, and Bayesian multivariate analysis, to mention a few.

To begin with, it is of interest to cite a reply made by T.W. Anderson, concerning a discussion of the 2nd edition of his book on multivariate analysis "For a confident and thorough understanding, the mathematical theory is necessary" (Schervish, 1987). Although these words were written more than 25 years ago, they make even more sense today.

The multivariate normal (Gaussian) distribution was first applied about 200 years ago. Today one possesses substantial knowledge of the distribution: the characteristic function, moments, density, derivatives of the density, characterizations, and marginal distributions, among other topics. Closely connected to the distribution are the Wishart and the inverse Wishart distributions and different types of multivariate beta distributions. When extending the multivariate normal distribution the class of elliptical distributions is sometimes used since it includes the normal distribution. Other types of multivariate normal distributions which share many basic properties with the classical "vector-normal" distribution are the matrix normal, the bilinear normal and the multilinear normal distributions. To some extent they are all special cases of the multivariate normal distribution (classical vector-valued distribution), but in view of the possible applications, there are some advantages to be gained from studying all these different cases.

It is interesting to observe that it is still a relatively open question how to decide if data follows a multivariate normal distribution. The existing tests may be classified either as goodness-of-fit tests or as tests based on characterizations. However, most of the tests are connected

with some asymptotic result and the size of the samples needed to make testing interesting is not obvious. Too large samples will usually lead to the test statistics becoming asymptotically normally distributed, even if the original data is not normal, whereas small samples will mean that there is no power when testing for normality. Here one can envisage computer-intensive methods to becoming beneficial, since they can speed up convergency.

Concerning modelling there has been a tendency to create more and more complicated models: i.e. the parametrization has tended to become more advanced and the distributions have tended to deviate more from the normal distribution. An interesting class to study is skew-symmetric distributions, which include a skew-normal distribution. One natural field of application of skewed distributions is cases when there exist certain detection limits. However, one should not forget that a small change in the parametrization may have drastic inferential consequences, for example, when extending the MANOVA model

$$\mathbf{X} = \mathbf{BC} + \mathbf{E}, \qquad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}),$$

where $\mathbf{B}$ and $\mathbf{\Sigma}$ are unknown parameters, to the Growth Curve model

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E}, \qquad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}),$$

where $\mathbf{B}$ and $\mathbf{\Sigma}$ are unknown parameters, as in MANOVA, and $\mathbf{A}$ and $\mathbf{C}$ are known design matrices. With the Growth Curve model we actually move from the exponential family to the curved exponential family with significant consequences, e.g. for the Growth Curve model the MLEs of $\mathbf{B}$ are non-linear and the estimators are not independent of the unique MLE of $\mathbf{\Sigma}$. A further generalization is a spatial-temporal setting

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E}, \qquad \mathbf{E} \sim N_{p,nk}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I} \otimes \mathbf{\Psi}),$$

where $\mathbf{\Sigma}$ models the dependency over time and $\mathbf{\Psi}$ is connected to spatial dependency. In summary, in MANOVA most things work as in the corresponding univariate case, i.e. easily interpretable mean and dispersion estimators are obtained, while in the Growth Curve model explicit estimators are also obtained, but the mean estimators are non-linear and more difficult to interpret. For the spatial-temporal model, no explicit MLEs are available but one has algorithms which deliver unique estimators. Concerning the future we will probably see more articles where for $\mathbf{X} \in N(\mu, \mathbf{\Sigma})$ there are models which state that $\mu \in \mathcal{C}(\mathbf{C}_1) \otimes \mathcal{C}(\mathbf{C}_2) \otimes \ldots \otimes \mathcal{C}(\mathbf{C}_m)$, i.e. a tensor product of $\mathcal{C}(\mathbf{C}_i)$, where $\mathcal{C}(\mathbf{C}_i)$ stands for the space generated by the columns of $\mathbf{C}_i$, and $\mathbf{\Sigma} = \mathbf{\Sigma}_1 \otimes \mathbf{\Sigma}_2 \otimes \ldots \otimes \mathbf{\Sigma}_m$. Another type of generalization which has been taking place for decades is the assumption of different types of dispersion structures, e.g. structures connected to factor analysis, structures connected to spatial relationships, and structures connected to time series, structures connected to random effects models, structures connected to graphical normal models, structures connected to the complex normal and quaternion normal distributions.

High-dimensional statistical analysis is, with today's huge amount of available data, of the utmost interest. Indeed various different high-dimensional approaches are natural extensions of classical multivariate methods. A general characterization of high-dimensional analysis is that in the multivariate setting there are more dependent variables than independent observations. It is driven by theoretical challenges as well as numerous applications such as applications within signal processing, finance, bioinformatics, environmetrics, chemometrics, etc. The area comprises, but is not limited to, random matrices, Gaussian and Wishart matrices with sizes which turn to infinity, free probability, the R-transform, free convolution, analysis of large data sets, various types of p, n-asymptotics including the Kolmogorov asymptotic approach, functional data analysis, smoothing methods (splines); regularization methods (Ridge regression, partial least squares (PLS), principal components regression (PCR), variable selection, blocking); and estimation and testing with more variables than observations.

If one considers the asymptotics with p indicating the number of dependent variables and n the number of independent observations, there are a number of different cases: $p/n \to c$, where c is a known constant, and both p and n go to infinity without any relationship between p and n. The latter case, however, has to be treated very carefully in order to obtain interpretable results. For example, one has to distinguish if first p and then n goes to infinity or vice versa, or $\min(p, n) \to \infty$. When studying proofs of different situations in the literature, it is not obvious which situation is considered and many results can only be viewed as approximations and not as strict asymptotic results, at least on the basis of the presented proofs.

One of the main problems in multivariate statistical analysis as well as high-dimensional analysis occurs when the inverse dispersion matrix, $\mathbf{\Sigma}^{-1}$, has to be estimated. If $\mathbf{\Sigma}$ is known, it often follows from univariate analysis that the statistic of interest is a function of $\mathbf{\Sigma}^{-1}$. Then one tries to replace $\mathbf{\Sigma}^{-1}$ with an estimator. If $\mathbf{S}$ is an estimator of $\mathbf{\Sigma}$, the problem is that $\mathbf{S}^{-1}$ may not exist or may perform poorly due to multicollinearity, for example. If $\mathbf{S}$ is singular, then $\mathbf{S}^{+}$ has been used. Moreover, "ridge type" estimators of the form $(\mathbf{S} + \lambda\mathbf{I})^{-1}$ are in use (Tikhonov regularization). Sometimes a shrinking takes place through a reduction of the eigenspace by removing the part which corresponds to small eigenvalues. A different idea is to use the Cayley-Hamilton theorem and utilize the fact that

$$\mathbf{\Sigma}^{-1} = \sum_{i=1}^{p} c_i \mathbf{\Sigma}^{i-1},$$

where $\mathbf{\Sigma}$ is of the size $p \times p$ and since $\mathbf{\Sigma}$ is unknown the constants c_i are also unknown. Then an approximation of $\mathbf{\Sigma}^{-1}$ is given by

$$\mathbf{\Sigma}^{-1} \approx \sum_{i=1}^{a} c_i \mathbf{\Sigma}^{i-1}, \qquad a \leq p,$$

and an estimator is found via $\widehat{\mathbf{\Sigma}}^{-1} \approx \sum_{i=1}^{a} \widehat{c}_i \mathbf{S}^{i-1}$. When determining c_i a Krylov space method, partial least squares (PLS), is used.

Needless to say, there are many interesting research questions to work on. Computers are nowadays important tools but much more important are ideas which can challenge some fundamental problems. For example in high-dimensional analysis we have parameter spaces which are infinitely large and it is really unclear how to handle and interpret this situation. Hopefully the discussions in this conference will deal with some of the challenging multivariate statistical problems.

References

[1] Schervish, M.J. (1987). A review of multivariate analysis. With discussion and a reply by the author. *Statist. Sci.* **2**, 396–433.