Quantification, Variable Selection and Sensitivity Analysis in Multivariate Methods

Yutaka Tanaka

1. Introduction

Variable selection and sensitivity analysis in multivariate methods are two major topics which I have been studying with my colleagues in Okayama for many years. We studied methodologies and developed two kinds of statistical software. One is VASMM (VAriable Selection in Multivariate Methods) developed by Y. Mori, M. Iizuka, T. Tarumi and Y. Tanaka (see, e.g., Iizuka et al., 2002) and the other is SAMMIF (Sensitivity Analysis in Multivariate Methods based on Influence Function) developed by Y. Mori, S. Watadani, Y. Yamamoto, Y. Odaka, T. Tarumi and Y. Tanaka (see, e.g., Mori et al., 1998). In this presentation I will review methods and software putting emphasis on motivations and applications.

If I have time I will briefly mention, as an additional topic, an interesting result of our recent study on the performance of classification of Hayashi’s second method of quantification.

2. Variable selection in multivariate methods

When I was working for a pharmaceutical company, I encountered a problem of variable selection in factor analysis (FA). I joined in a statistical analysis group of a project with experts of neurosurgery to make a rating scale to measure the improvement of mild disturbance of consciousness (MDOC) (Sano, K. et al., 1977). The questionnaire consisted of 23 items or variables selected carefully for representing all the important symptoms of MDOC, and as the result of analysis a two-factor model fitted the data well. The obtained two factors were named as verbal factor and performance factor. The questionnaire was suitable for the research purposes, but it was time-consuming to be used in daily examinations in ordinary hospitals and therefore a simplified version was required. Then we studied methods of variable selection in factor analysis (Tanaka & Kodake, 1981; Tanaka, 1983). In these studies we tried to select variables so that the difference was as small as possible between the configurations of the factor scores based on all the variables and those based on selected variables.

Similar situations exist for variable selection in principal component analysis (PCA). Consider a situation where we wish to select variables so as to delete redundant variables or to make a small dimensional, possibly one dimensional, rating scale by using PCA. Validity requires all related variables to be included, however, practical applications require the number of variables to be as small as possible. For example, consider a case where there are a large number of sensors (checkpoints) in a plant which are used to measure some quantity at each checkpoint and evaluate the performance of the entire plant. Exact evaluation requires to be made on the basis of the data measured at all points, but the number of points may be too large.

---

1Professor Emeritus, Okayama University
to obtain the result within limited time for temporary evaluation. It may be helpful if we can reduce the number of points to be used for temporary analysis without losing the degree of accuracy. To be used in such situations we proposed a modified PCA (Tanaka & Mori, 1997). As variable selection procedures in VASMM we can choose any procedure in stepwise procedures such as forward, backward, forward-backward and backward-forward procedures as well as the procedure of selecting the best one among all possible subsets.

3. Sensitivity analysis in multivariate methods

There are two major tools for sensitivity or influence analysis in statistical methods. One is Hampel’s influence function (Hampel, 1974) and the other is Cook’s local influence (Cook, 1986). Various methods of sensitivity analysis have been studied in multivariate methods such as PCA, FA, canonical correlation analysis, correspondence analysis, Hayashi’s methods of quantification and covariance structure analysis (simultaneous equation modeling) by many authors using either of these two tools in recent three decades. In the influence function approach a case weight is introduced to each observation and the influence of a small change of the weight is evaluated on the results of analysis so that we can detect “influential” observations. On the other hand in the local influence approach a weight vector is introduced to all observations simultaneously and the influential direction of the weight vector is searched for by maximizing the likelihood displacement defined by the difference between maximized likelihood functions for perturbed and unperturbed likelihoods. We studied the relationships between these two approaches and proposed a general procedure for sensitivity analysis in statistical modeling. It consists of 1) the computation of influence functions for the parameters, 2) the evaluation of the covariance matrix $V$ of the estimated parameters, and 3) PCA of the influence function vectors with the (generalized) inverse of $V$ as its metric. It can be verified that the $k$-th principal component scores provide the $k$-th influential direction in the sense of Cook’s local influence. It can detect not only singly but also jointly influential observations (Tanaka, 1994; Tanaka & Zhang, 1999; Tanaka, Zhang & Mori, 2003). Cook’s local influence has provided a general framework for sensitivity analysis in statistical models, but our general procedure has advantages that it can be applied to situations where the likelihood function is not available and that we can easily modify the procedure so that it becomes robust.

Acknowledgment

I would like to express my sincere thanks to the members of VASMM and SAMMIF projects and coauthors of related papers for their active contributions to the research and development of the methodologies and software. I am also grateful to Dr. Yuichi Mori and Dr. Masaya Iizuka for their kindness of providing me the materials to be used for the presentation.
References


