Hochdimensionale Statistik – Mathematik und Computer als Werkzeuge

Jürgen Läuter
Otto-von-Guericke-Universität Magdeburg
Juergen.Laeuter@medizin.uni-magdeburg.de

Keywords: multivariate Analyse, hohe Dimension, stabile statistische Verfahren, exakter Test.

Seit es sie gibt, ist die multivariate statistische Analyse eine Herausforderung für die Mathematiker. Einerseits lassen sich mit Mengen von Variablen mehr Informationen über die betrachteten Individuen als mit Einzelvariablen darstellen. Andererseits entstehen durch die Vielfalt von Variablen auch neue Probleme. Gerade in den letzten Jahren hat die multivariate Statistik durch Aufgaben aus der Bioinformatik wieder erhöhte Aktualität gewonnen.

Die klassische multivariate Analyse (siehe z.B. Anderson, 1984) kommt an ihre Grenzen, wenn die Dimension p nur wenig kleiner oder sogar größer als der Stichprobenumfang n ist. Eine statistische Inferenz ist dann kaum mehr bzw. überhaupt nicht möglich. Aber die praktischen Anforderungen entwickelten sich in der Richtung, dass die Zahl p der Variablen infolge der verbesserten Messgeräte stark zunahm, ohne dass die Zahl n der unabhängigen Individuen mitwuchs.

In dieser Situation bieten sich neuartige Verfahren einer sog. stabilen multivariaten Statistik zur Problemlösung an (siehe z.B. Läuter, Glimm und Kropf, 1996). Die klassische multivariate Analyse und auch die später entstandene Mixed-Model-Theorie (siehe z.B. Diggle, Liang und Zeger, 1994) streben eine möglichst weit gehende explizite Parametrisierung der auftretenden Verteilungen an, und demgemäß sind im ersten Schritt die unbekannten Parameter zu schätzen. Die erwähnte stabile Statistik dagegen verzichtet auf die durchgehende explizite Parametrisierung und benutzt stattdessen Linearkombinationen der gegebenen p Variablen (sog. Scores), für die nur schwache Restriktionen bestehen und die daher auf der Grundlage von Erfahrungen und Vermutungen passend festgelegt werden können.

Die klassische, auf der Methode der kleinsten Quadrate und der Maximum-Likelihood-Methode basierende Strategie erfordert eine Eingrenzung bzw. Verminderung der Variablenzahl p. Im Gegensatz dazu kann in der stabilen Statistik Redundanz der Variablen sinnvoll benutzt werden. Es kommen verstärkt Methoden der Glättung und des gegenseitigen Ausgleichs zur Anwendung.

Trotz der engen Verflechtung exploratorischer und konfirmatorischer Verfahrensschritte arbeitet die stabile Statistik nach strengen mathematischen Kriterien. Eine entscheidende Grundlage ist durch die Theorie der sphärischen Matrixverteilungen (Fang und Zhang, 1990) gegeben. In den Anwendungen der letzten Zeit konnten z.B. exakte Mittelwertsvergleiche an Genexpressionsmustern der Dimension $p = 12625$ durchgeführt werden. Auch mathematisch exakte multiple Testprozeduren zur Erkennung relevanter Variablenteilmengen wurden bereitgestellt.

Die Entwicklung der stabilen bzw. sphärischen Analyse ist das Ergebnis jahrelanger Bemühungen um effektive multivariate Auswertungsmethoden. Diese Entwicklung ist ein gutes Beispiel dafür, dass in der Statistik eine Symbiose theoretischer Überlegungen mit Computertechniken sinnvoll, wenn nicht sogar dringend erforderlich ist.

Literatur

1. Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.
2. Diggle, P.J., Liang, K.-Y. und Zeger, S.L. (1994). *Analysis of Longitudinal Data.* Clarendon Press, Oxford.
3. Fang, K.-T. und Zhang, Y.-T. (1990). *Generalized Multivariate Analysis.* Springer, Berlin.
4. Läuter, J., Glimm, E. und Kropf, S. (1996). New Multivariate Tests for Data with an Inherent Structure. *Biometrical Journal*, **38**, 5-23.