Zusammenfassung der Arbeit

Dissertation zur Erlangung des akademischen Grades Dr. rer. nat.

Exakte parametrische multivariate Tests für hochdimensionale Beobachtungen mit Unterstützung der Auffindung faktorieller Strukturen

Eingereicht von:
Dipl.-Math. Mohammad Zaino

Angefertigt:
- Fakultät für Mathematik und Informatik der Universität Leipzig
- Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE) der Universität Leipzig

Betreut von:
PD Dr. S. Kropf
Prof. Dr. M. Löffler
Prof. Dr. B. Kirstein

Juni 2006

Zielstellung

Durch den Einzug neuer Untersuchungsverfahren lässt sich in medizinischen Studien zunehmend die Problematik steigender Merkmalsanzahlen bei gleichzeitig begrenzter Zahl von Patienten feststellen. Damit entsteht eine wachsende Anforderung auf die multivariate Statistik, die Diskrepanz zwischen der Variablenanzahl und den Stichprobenumfängen zu überbrücken.

In der vorliegenden Arbeit wurden verschiedene multivariate Testvarianten untersucht und weiterentwickelt, die sich für diese Zwecke eignen. Die vorgestellte Familie von exakten multivariaten parametrischen Tests geht auf Arbeiten von Läuter und Mitarbeitern aus den

1990er Jahren zurück. Dabei werden die hochdimensionalen Daten mittels datenabhängiger Gewichte in niedrigdimensionale Scores transformiert und diese Scores dann in klassischen Tests weiterverarbeitet. Wenn die Gewichte nach gewissen allgemeinen Regeln bestimmt werden, dann halten die Tests das Fehlerniveau trotz der eingeschlossenen Datenvorverarbeitung exakt ein. Einige dieser Tests sind auf diese Weise auch noch dann durchführbar, wenn die Dimension der Beobachtungen den Gesamtstichprobenumfang übersteigt. Es wird sogar angestrebt, aus der hohen Merkmalsdimension einen Gewinn an Güte zu erzielen. Die mathematische Grundlage dieser Tests ist die Theorie sphärisch verteilter Matrizen.

Innerhalb des verbliebenen Freiraumes können verschiedene Testvarianten abgeleitet werden. Das Ziel dieser Arbeit ist es, nach Testvarianten zu suchen, bei denen die Scores sich gut interpretieren lassen, um so die globale multivariate und damit mehr oder weniger abstrakte Testaussage einer besseren inhaltlichen Deutung zuzuführen. Hierzu werden vorhandene Testvarianten (der PC_q-Test und der PC_{est}-Test in zwei Untervarianten) untersucht, miteinander verglichen und erweitert. Insbesondere kommen dabei aus der Faktoranalyse bekannte Rotationsmethoden im Raum der Scores zum Einsatz. Sie werden zusammen mit Variablenselektionsmethoden so angewendet, dass sie sich in die Theoreme für Tests mit sphärisch verteilten Daten einordnen. Damit sind auch die Tests für jeden einzelnen Score im exakten Sinne durchführbar.

Neben der besseren Interpretierbarkeit der einzelnen Scores und der zugehörigen „univariaten“ Testergebnisse besteht das Ziel auch in der Erhöhung der Güte dieser einzelnen Tests, da eine bessere Dekomposition der multivariaten Effekte, in die sie bestimmenden latenten Faktoren, einer Vermischung der Effekte der einzelnen Faktoren entgegenwirken kann.

Ergebnisse und Diskussion

Die hier entwickelten Testvarianten werden an zwei realen medizinischen Datensätzen erprobt. Damit wird gleichzeitig die Anwendbarkeit der Prinzipien für verschiedene statistische Testsituationen demonstriert. Zusätzlich werden Simulationsuntersuchungen durchgeführt, in denen unter anderem die Wiedererkennung der gezielt erzeugten Faktorstrukturen kontrolliert wird.

In den Beispielen lassen sich mit den erarbeiteten Verfahren gut interpretierbare Scores ableiten. Besonders im ersten Beispiel werden die Unterschiede zwischen den untersuchten

Gruppen auf wenige Scores konzentriert. Auch in den Simulationsstudien wird der entmischende Effekt der Rotationen eindrucksvoll bestätigt.

Aus den Ergebnissen dieser Untersuchungen können Empfehlungen für die Auswahl geeigneter Testvarianten für spätere Anwendungen abgeleitet werden, wo ein Durchprobieren aller Varianten natürlich den Fehler erster Art unterlaufen würde.

Speziell lieferten der PC_q-Test und der PC_{est}-Test mit linkssymmetrischer Wurzel meist bessere Ergebnisse als der PC_{est}-Test mit einfachem Schätzer. Von den Schätzkriterien für die Faktoranzahl ist fast immer das Invers-Jolliffe-Kriterium vorzuziehen. Der Einschluss einer Rotation verbesserte die Interpretierbarkeit der Scores erheblich, der Einfluss von Variablenselektionsmethoden war eher gering.

Bei einer gegebenen praktischen Problemstellung können die Simulationsuntersuchungen der jeweiligen Aufgabe angepasst werden, indem die Gruppenanzahl, die Stichprobenumfänge und die Merkmalsanzahl übernommen und die anderen Parameter nach den zu vermutenden Gegebenheiten variiert werden. Damit erhält man Hinweise auf die, zu erzielende Güte und kann so die Auswahl steuern.

Es wird hier aber auch gezeigt, dass bei geeigneter Organisation selbst eine subjektive Bewertung der Korrelationen zwischen den Scores und den Originalvariablen ein zulässiges Hilfsmittel zur Auswahl einer Testversion aus einer Liste von mehreren Kandidaten sein kann, ohne das Testniveau zu verfälschen.

Es muss darauf hingewiesen werden, dass die „univariaten“ Tests mit einzelnen Scores lediglich zur besseren Interpretation des Ergebnisses des globalen multivariaten Tests dienen, nicht aber zu einer exakten separaten Beurteilung der vermuteten zugehörigen latenten Variablen. Selbst wenn man als globalen Test gleich den „univariaten“ Test eines anhand der Ladungsmatrix ausgewählten Scores wählt, bleibt dieser Test im exakten Sinne immer nur ein Test der globalen Nullhypothese. Allerdings ist die Power dieses Tests dann besonders auf diese latente Variable ausgerichtet und Vermischungen mit Effekten anderer latenter Variablen sind durch die Rotation weitgehend reduziert.

In allen Analysen mit realen oder simulierten Daten wurden hier Situationen betrachtet, in denen die Merkmalsanzahl fast den Gesamtstichprobenumfang erreicht. Dabei weisen klassische multivariate Tests eine unzureichende Güte auf, während die meisten der hier betrachteten Versionen recht gute Ergebnisse lieferten. Mit dem PC_q-Test wären auch Merkmalszahlen realisierbar gewesen, die über dem Stichprobenumfang liegen. Solche Situationen liegen

z.B. fast immer bei Genexpressionsanalysen mittels Microarray-Techniken vor. Da der PC_{est}-Test diese Möglichkeit nicht bietet und hier besonders der Methodenvergleich interessierte, wurden solche Strukturen in dieser Arbeit ausgeschlossen. Es ist zu vermuten, dass durch eine Rotation der Scores bei diesen Konstellationen ähnliche Effekte mit dem PC_q-Test oder mit geeignet modifizierten Versionen des PC_{est}-Tests auftreten. Die Entwicklung solcher Modifikationen des PC_{est}-Tests und der Nachweis der vermuteten Leistungen dieser Testversionen bleiben allerdings nachfolgenden Arbeiten überlassen.