Schnelle Diskriminanzanalyse mit vielen Variablen

Bernd Heinen
SAS Institute GmbH
In der Neckarhelle 162
Heidelberg
Bernd.heinen@jmp.com

Zusammenfassung

Diskriminanzanalyse erfordert die Schätzung der Kovarianzmatrix der Variablen, für die die Diskriminanzfunktion berechnet werden soll. So unterschiedliche Bereiche wie Genomanalyse und Halbleiterfertigung haben gemeinsam, dass in einem einzigen Datensatz sehr viele Variablen analysiert werden sollen, häufig 5.000 und mehr. Häufig beträgt in dieser Situation auch die Zahl der Beobachtungen nur einen Bruchteil der Zahl der Variablen. Durch eine integrierte Hauptkomponentenanalyse lässt sich die Dimensionalität der Aufgabe deutlich reduzieren. So werden auch umfangreiche Aufgabenstellungen problemlos auf dem PC lösbar.

Schlüsselwörter: Diskriminanzanalyse, Hauptkomponenten, Singulärwertzerlegung, JMP Pro

1 Einleitung

Diskriminanzanalyse dient dazu, einzelne Beobachtungen mit Hilfe der Ausprägungen stetiger Variablen einer von mehreren möglichen Gruppen zuzuordnen. Diese Zuordnung wird an einem Testdatensatz entwickelt und kann später dazu dienen, neue Objekte zu klassifizieren. Anwendungsbereiche finden sich in Soziologie, Ökonometrie und den Naturwissenschaften. Gerade in der Genetik kommt eine spezielle Herausforderung hinzu: oftmals sind die vorhandenen Datensätze ehr breit als lang, d.h. sie haben mehr Variablen als Beobachtungen. In dieser Situation versagen die meisten Analyseverfahren, aber Hauptkomponentenanalyse und Diskriminanzanalyse sind auch unter diesen Bedingungen anwendbar.

2 Herleitung

Für einen kurzen Abriss der verwendeten Verfahren nehmen wir an, dass k Gruppen unterschieden werden sollen, wofür p Merkmalsvariablen zur Verfügung stehen. Jede Gruppe wird durch ihren Mittelwertsvektor ($\bar{x}_{i1}, \bar{x}_{i2, \dots,} \bar{x}_{ip}$) repräsentiert. Die Abstände zwischen einer Beobachtung und dem Mittelpunkt jeder Gruppe wird mit dem Mahalanobis Abstand $d_i = (x - \bar{x}_i)^T \Sigma^{-1} (x - \bar{x}_i)$ berechnet, wobei Σ die Kovarianzmatrix bezeichnet. Eine einzelne Beobachtung wird dann der Gruppe zugeordnet, zu deren Mittelpunkt ihr Abstand am geringsten ist. Um zur Zuordnung eines Punktes zu einer

der Gruppen nicht jedes Mal die Mahalanobis Distanz ausrechnen zu müssen, kann man einen Satz von k-1 kanonischen linearen Diskriminanzfunktionen berechnen:

$D_j = \beta j_0 + \beta_{j1} X_1 + \beta_{j2} X_2 + ... + \beta_{jp} X_p$

Die linearen Diskriminanzfunktionen stellen die einfachste Struktur zur Trennung von Gruppen dar (zweidimensional: Geraden, dreidimensional: Ebenen, höherdimensional: Hyperebenen als trennende Mengen).

Ähnlich der kleinsten Quadrate Schätzung für lineare Modelle werden die Koeffizienten bestimmt durch eine Optimierungsaufgabe:

$$D = \frac{SQ_{zwischen}}{SQ_{innerhalb}} ! = Max$$

Die Schätzung der Diskriminanzfunktion erfordert die Invertierung einer p x p Matrix. Die im Folgenden vorgestellte Methode für eine schnelle Berechnung der Diskriminanzfunktion bei umfangreichen Merkmalsvektoren bedient sich der Hauptkomponentenanalyse. Dabei werden orthogonale Linearkombinationen der Variablen gesucht, die ein neues Koordinatensystem aufspannen, in dem die Datenpunkte ebenfalls dargestellt werden können. Die Hauptkomponenten sind unkorreliert und ermöglichen meistens auch ein Reduktion der Dimensionen, die benötigt werden, um die Beobachtungen darzustellen.

3 Beispieldaten und lineare Diskriminanzanalyse

Die Daten, die dem begleitenden Beispiel zu Grunde liegen, stammen von einem Datensatz, der Brustkrebsmarker und Microarraydaten von 10.787 spots enthält. Er weist eine für solche Studien durchaus charakteristische Form auf da er mehr Spalten als Zeilen (230 Patienten) umfasst. Die Ergebnisse der einzelnen Marker sind zu einem Profil zusammengefasst (Combined Status, s. Abb. 1), dessen unterschiedliche Ausprägungen die Gruppen für die Diskriminanzanalyse beschreiben.

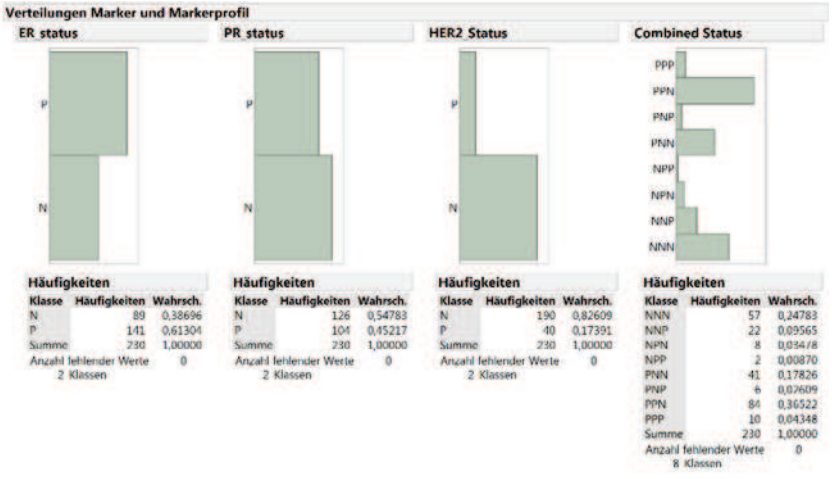


Abb. 1: Verteilung der Zielgröße „Combined Status“

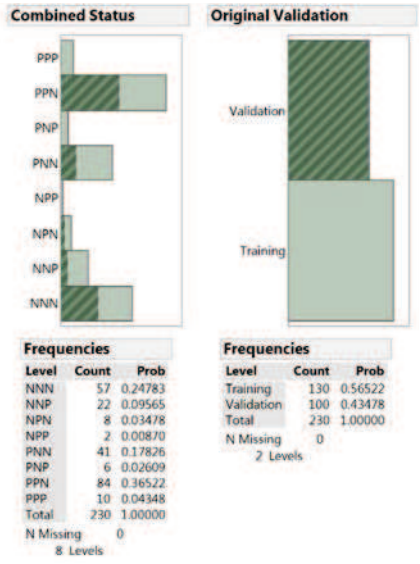
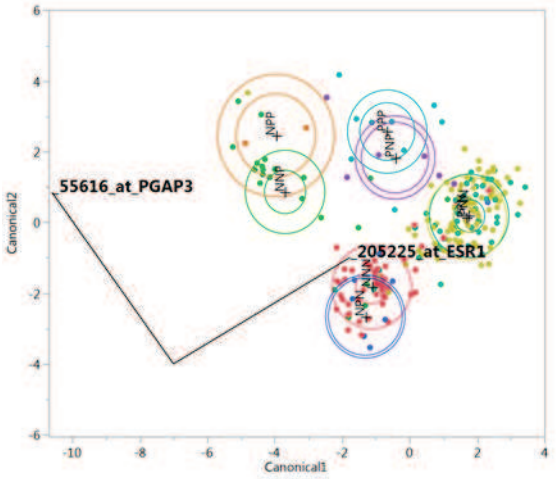


Abb. 2: Validierung

Die Güte der Diskriminanzfunktion wird an 100 Validierungsdatensätzen überprüft, die nicht zur Anpassung der Funktion herangezogen werden. Leider deckt die zufällige Zuordnung der Beobachtungen zu dem Validierungsdatensatz nicht alle Gruppierungen ab. Für diese Betrachtung ist das allerdings unerheblich.
Führt man eine Diskrimanzanalyse mit zwei Parametern durch, so erhält man zwei Diskriminanzfunktionen

Canonical 1 = -0,4 + 0,67 * :_205225_at_ESR1 – 0,8* :_55616_at_PGAP3
und
Canonical 2 = -10,66 + 0,39 * :_205225_at_ESR1 + 1,06 * :_55616_at_PGAP3

wobei die mit Doppelpunkt beginnenden Terme die Expressionen der Gene (die Variablen) darstellen.
In der nebenstehenden Grafik (Abb. 3) sind die Beobachtungen des Datensatzes als Punkte in einem Koordinatensystem dargestellt, das durch diese beiden Funktionen aufgespannt wird, weshalb die Achsen mit „Canonical1“ und „Canonical2“ bezeichnet sind. Die Geraden stellen die beiden Variablen dar und zeigen, wie stark sie mit den Diskriminanzfunktionen korreliert sind, wobei der Winkel der Geraden zu der jeweiligen Achse den Grad der Korrelation angibt, die Richtung der Geraden die Art der Korrelation. Beide Variablen sind mit beiden Funktionen positiv korreliert, _205225_at_ESR1 stärker mit Canonical 1 als mit Canonical 2, bei _55616_at_PGAP3 ist es umgekehrt. Der Ursprung dieser beiden Geraden ist der Punkt (0, 0) im kanonischen Koordinatensystem,



Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	130	61	46.9231	0.39218	282.718
Validation	100	51	51.0000	-0.2732	

Abb. 3: Diskriminanzfunktion

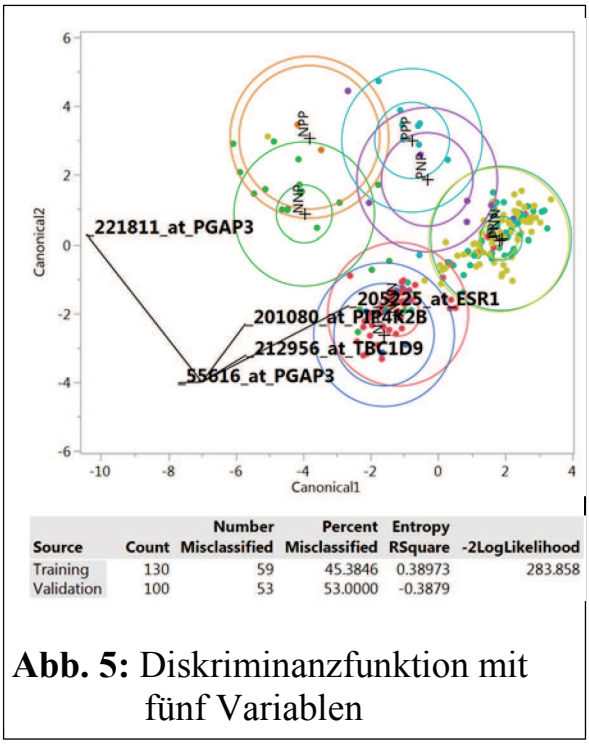
was dem Gesamtmittelwertsvektor der beiden Variablen entspricht. Der besseren Lesbarkeit halber ist die Darstellung aus dem Ursprung verschoben.

Combined Status		Predicted							
		NNP	NPP	PNP	PPP	NNN	NPN	PNN	PPN
Actual	NNP	7	5	1	0	1	2	0	0
	NPP	0	2	0	0	0	0	0	0
	PNP	0	1	4	0	0	0	1	0
	PPP	0	0	3	6	0	0	0	0
	NNN	0	0	0	0	12	12	1	2
	NPN	0	0	0	0	1	4	0	0
	PNN	0	0	0	0	1	0	15	12
	PPN	0	0	0	0	0	0	18	19

Abb. 4: Klassifikationshäufigkeiten

Aus der Grafik sieht man gleich, dass es zwei Gruppen von Statuskombinationen gibt, deren Mitglieder man untereinander schlecht unterscheiden kann, aber gut gegenüber den anderen Kombinationen. Vergleicht man jeweils den tatsächlichen Status mit dem prognostizierten, bestätigt sich dieser Befund (Abb. 4). In den grünen Bereichen (die jeweils andere Gruppe) befinden sich nur wenige Fehlklassifikationen, in den orangenen Bereichen (innerhalb der eigenen Gruppe) befinden sich wesentlich mehr. Die größten Zellhäufigkeiten erwartet man in den weißen Zellen der Hauptdiagonalen. Das trifft auch zu, aber die geradezu identische Lage des Paares PNN und PPN spiegelt sich auch in den häufigen wechselseitigen Fehlklassifikationen wieder.

Fasst man die Fehklassifikationen zusammen, so ergibt sich eine Fehlklassifikationsrate von 47%. Alle Aussagen zur Fehlklassifikation betreffen den Trainingsdatensatz. Man kann die gleichen Betrachtungen auch für den Validierungssatz anstellen mit einem ähnlichen Ergebnis. Die Fehlklassifikationsrate liegt dort aber bei 51%.

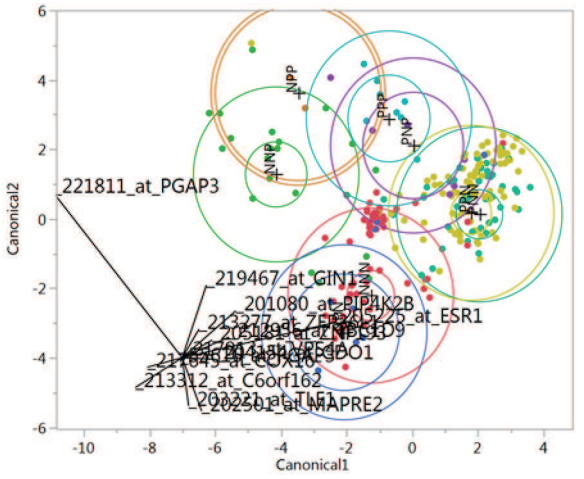


Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	130	59	45.3846	0.38973	283.858
Validation	100	53	53.0000	-0.3879	

Abb. 5: Diskriminanzfunktion mit fünf Variablen

Erhöht man die Zahl der Variablen – hier auf fünf (Abb. 5) - können entsprechend mehr kanonische Funktionen gebildet werden. Auf die ersten beiden, die den größten Teil der Gesamtvariabilität erklären, wird in der Grafik Bezug genommen. Man erkennt schon in der zweidimensionalen Projektion, dass die Gruppen besser getrennt werden. Dennoch verbessert sich die Fehlklassifikationsrate nur auf 45%, im Validierungsdatensatz verschlechtert sie sich sogar auf 53%.

Die Zahl der Variablen kann man beliebig erhöhen, die Zahl der kanonischen Diskriminanzfunktionen kann höchstens k-1 betragen, wenn k Gruppen zu unterscheiden sind



Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	130	30	23.0769	0.64591	164.698
Validation	100	63	63.0000	-0.9262	

Abb. 6: Diskriminanzfunktion mit 14 Variablen

Das nebenstehende Beispiel zeigt die Ergebnisse bei einer Anpassung von 14 Variablen. Von den maximal sieben Funktionen, die gebildet werden können, tragen nur die ersten fünf signifikant zur Erklärung der Gesamtstreuung bei, sie decken insgesamt fast 98% der Streuung ab. Die Fehlklassifikationsrate sinkt dabei auf 23% im Trainingsdatensatz, steigt aber auf 63% im Validierungssatz.
An diesen drei willkürlich gewählten Beispielen erkennt man gleich, dass eine Erweiterung der Diskriminanzfunktion um weitere Variablen durchaus sinnvoll sein kann, um ihre Trennschärfe zu erhöhen.

Hierbei muss aber auch der Validierungsdatensatz zur Kontrolle herangezogen werden. Das Umzusetzen ist im Prinzip einfach, schließlich rechnen Computer geduldig und schnell. Die Schätzung der Parameter für p Variablen erfordert aber die Invertierung einer p x p Matrix, was sich bei großen Variablenzahlen durchaus auch in der Rechenzeit bemerkbar macht. Man sieht den exponentiellen Anstieg, hier getestet für 2.000 – 5.200 Variablen (Abb. 7).

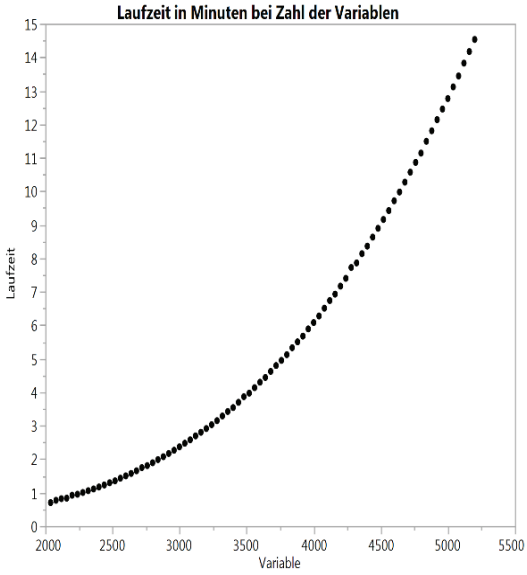


Abb. 7: Laufzeit in Abhängigkeit von der Zahl der Variablen

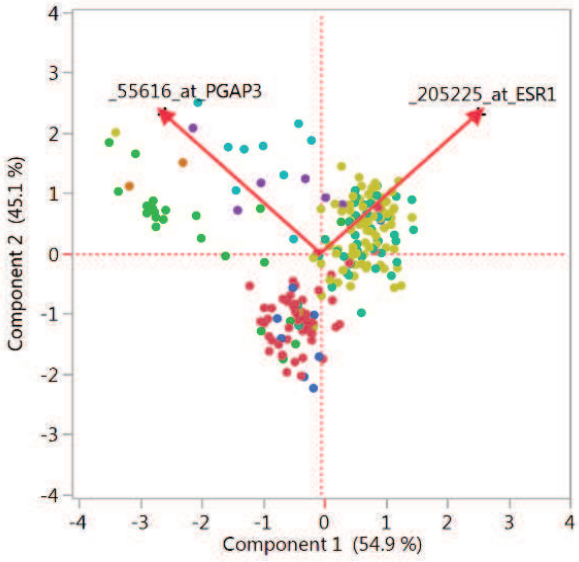


Abb. 8: Hauptkomponenten

4 Alternative Berechnung

Hier setzt die Überlegung an, die Berechnung zu beschleunigen, indem man durch Singulärwertzerlegung der gemeinsamen Kovarianzmatrix eine Matrix für die Hauptkomponententransformation erhält. Singulärwertzerlegung bedeutet ja, dass man eine reelle Matrix A in der Form A = U Σ V‘ darstellen kann, wobei Σ eine Matrix mit Diagonalelementen $\sigma_{11} \geq \sigma_{22} \geq \cdots \geq \sigma_{rr} > 0$ und 0 sonst ist. U und V sind orthonormale Matrizen, r ist der Rang von A. Formt man die gemeinsame Kovarianzmatrix der p Variablen entsprechend um, so enthält die Matrix V‘ gerade die Hauptkomponenten, die Diagonalelemente der Matrix Σ sind die Varianzen der einzelnen Variablen. Abb. 8 zeigt die Ladung der Variablen in Bezug auf die Hauptkomponenten. Beide Vektoren stehen annähernd senkrecht zueinander, d.h. ihre Korrelation ist gering. Bei p Variablen können maximal p Hauptkomponenten gebildet werden, von denen so viele beibehalten werden, dass sie mindestens einen Anteil von 0,9999 der Quadratsumme der Singulärwerte umfassen. Der Originaldatensatz wird damit transformiert, d.h. die Originaldaten werden als Koordinaten in dem System der Hauptkomponenten dargestellt.

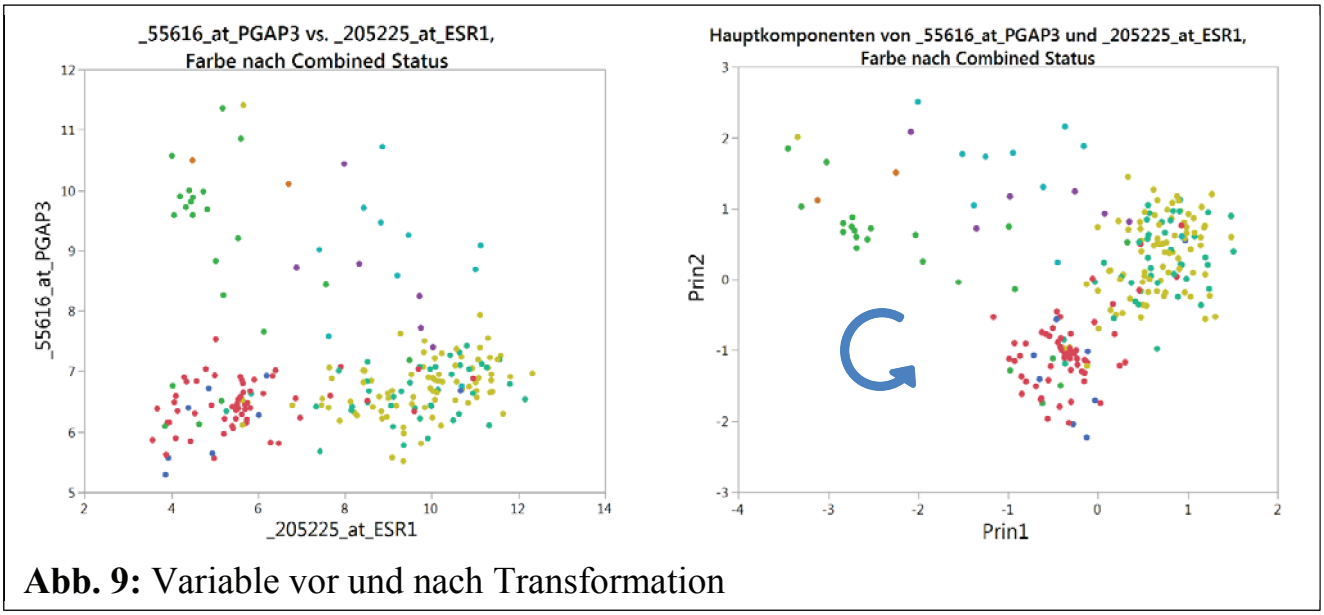


Abb. 9: Variable vor und nach Transformation

Abb. 9 zeigt die Originalmesswerte (links) und die transformierten Werte (rechts). Wegen der geringen Korrelation der Variablen untereinander erscheinen die Punktewolken einfach gegeneinander gedreht. Die Hauptkomponenten sind orthogonal und damit ist deren gepoolte Kovarianzmatrix eine Diagonalmatrix, wie aus der Gegenüberstellung in Abb. 10 zu sehen ist. Die Matrix ist leicht zu invertieren, was die gesamte Berechnung beschleunigt, so dass mehr Variablen in akzeptablen Zeiträumen in eine Diskriminanzanalyse eingeschlossen werden können.

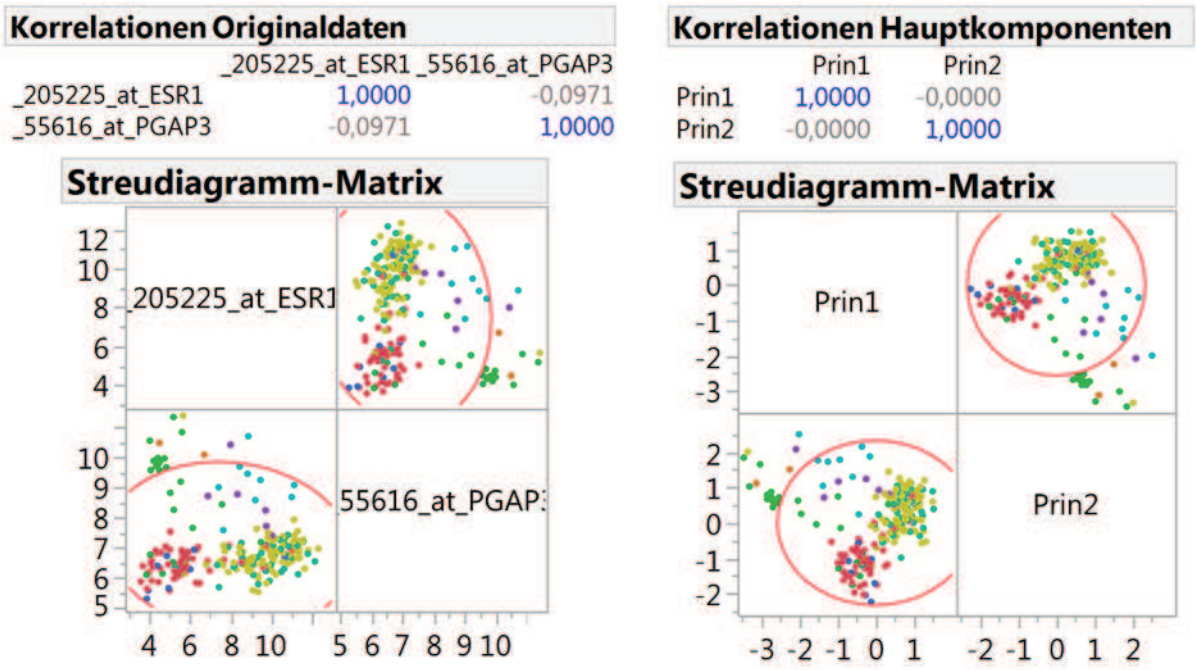
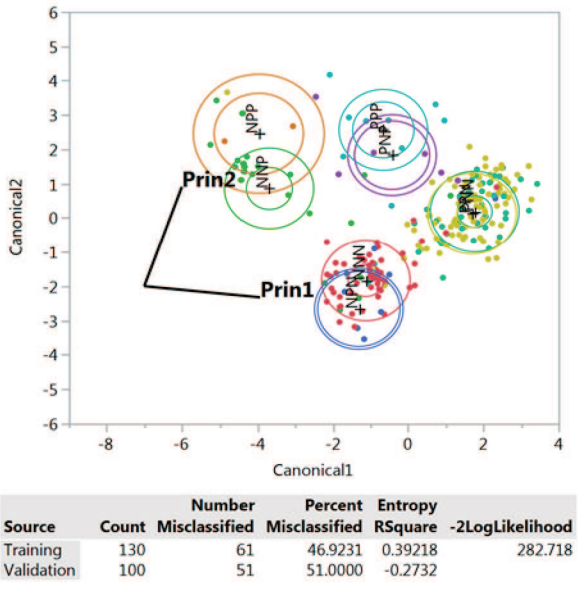


Abb. 10: Korrelationen



Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	130	61	46.9231	0.39218	282.718
Validation	100	51	51.0000	-0.2732	

Abb. 11: Diskriminanzanalyse mit Hauptkomponenten

Hinsichtlich der Trennung der Zielgruppen führt die Diskriminanzanalyse der Hauptkomponenten zu Ergebnissen, die identisch sind mit denen der Originalwerte, wie die Grafiken und Daten aus Abb. 3 im Vergleich mit Abb. 11 zeigen.

5 Anwendung

Die Vorteile des Verfahrens liegen auf der Hand. In der Diskriminanzanalyse, wie in allen Modellierungsverfahren, ermöglicht die Hinzunahme mehrerer (oder sehr vieler) Variablen eine höhere Präzision der Vorhersage. Ein schnellerer Algorithmus eröffnet zunächst einfach die Möglichkeit, mehr Variablen in die Analyse einzuschließen. Die Zeitersparnis des beschriebenen Verfahrens ist enorm, bei 5.200 Variablen benötigt die herkömmliche Analyse 15 Minuten, der erweiterte Algorithmus 3 Sekunden. Außerdem

wächst die Rechenzeit linear mit der Zahl der Variablen, selbst 10.000 Variablen werden in etwas mehr als 5 Sekunden berechnet.

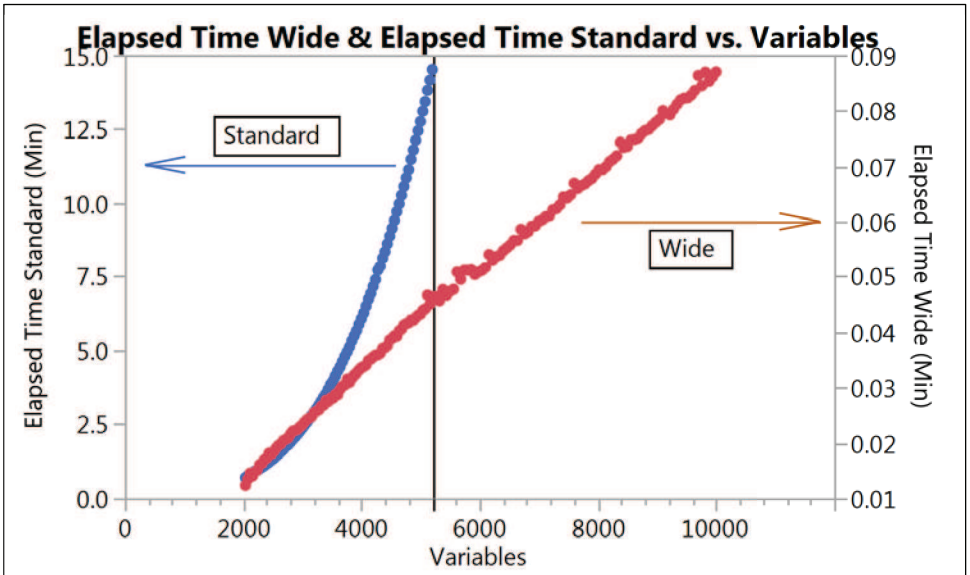


Abb. 12: Laufzeitvergleich der Diskriminanzverfahren

Wie in dem einleitenden Beispiel schon aufgefallen ist, ist eine unabhängige Kontrolle z.B. über einen unabhängigen Validierungsdatensatz empfehlenswert. Schnelle Algorithmen erlauben das Durchspielen vieler Alternativen, und erhöhen die Wahrscheinlichkeit, ein treffsicheres Modell zu finden.

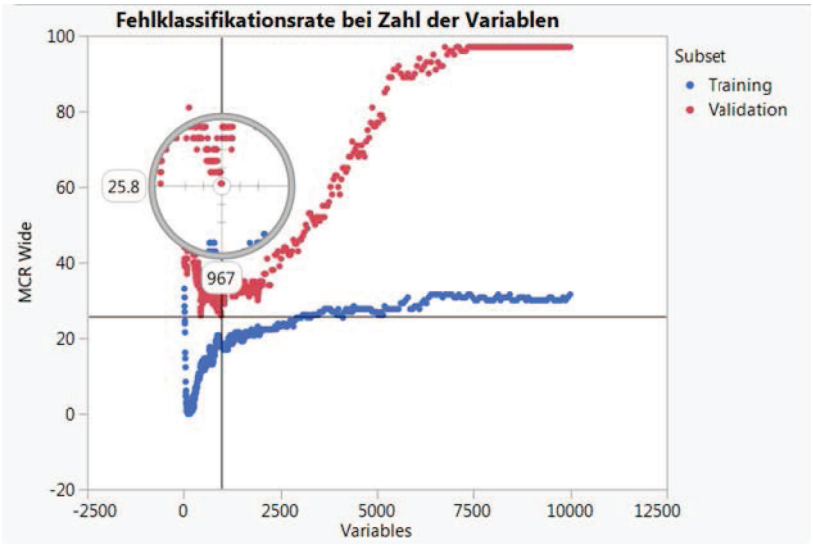


Abb. 13: Fehlklassifikationsraten

So ist in diesem Fall (Abb. 13) das beste Modell dasjenige, das 967 Variablen einschließt. In diesem Beispiel habe ich zunächst einen Random Forest berechnet und die Variablen dann in der Reihenfolge ihrer Beiträge zu diesem Erklärungsmodell in die Diskriminanzanalyse aufgenommen. Durch die kurzen Rechenzeiten sind aber noch andere Strategien denkbar, z.B. Permutationen der Variablen durchzuspielen.

Literatur

[1] Benjamin Doerr, Vorlesung Mathematik für Informatiker II, Sommersemester 2010, Max Planck Institut für Informatik

[2] Lindsay I Smith, A tutorial on Principal Components Analysis, February 26, 2002, Université de Montréal

[3] Jun Liu, Songcan Chen, Xiaoyang Tan, A study on three linear discriminant analysis based methods in small sample size problem, Pattern Recognition 41 (2008) 102–116, Elsevier

[4] SAS Institute Inc. 2015, JMP 12 Multivariate Methods, 2015, Cary, NC