Technische Universität Dortmund

Fakultät Statistik

Statistische Methoden zur Identifikation von Patientensubgruppen aus Hochdurchsatzdaten

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften

von Dipl.-Stat.

Maike Ahrens

Vorgelegt: Dortmund, den 26.08.2016

Gutachter: Prof. Dr. Jörg Rahnenführer,

Prof. Dr. Katja Ickstadt,

PD Dr. Martin Eisenacher

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Dissertation ist bisher keiner anderen Fakultät vorgelegt worden. Ich erkläre, dass ich bisher kein Promotionsverfahren erfolglos beendet habe und dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt.

Maike Ahrens

Inhaltsverzeichnis

Übersicht der wichtigsten Parameter und Abkürzungen

Allgemeine Abkürzungen

SG Subgruppe

K**,** G Die Samplegruppen krank K bzw. gesund G. Allgemeiner bezeichnet K in einem Zwei-Gruppen-Vergleich die Gruppe, die auf Subgruppen untersucht werden soll und G die als homogen angenommene Gruppe

nks Nicht-krankheitsspezifisch: Variablen mit nks Subgruppe weisen in G und K eine Samplesubgruppe auf

Univariate Methoden

OS Outlier sum

ORT Outlier robust t-statistic

PADGE Percentile analysis for differential gene expression

PACK Profile analysis using clustering and kurtosis

FS Fisher Sum

x Vektor der Beobachtungen eines Features X:
$x = (G,K) = (g_1,\ldots,g_{n_G},k_1,\ldots,k_{n_K}) = (x_1,\ldots,x_{n_G+n_K}) = (x_1,\ldots,x_N)$

med, med_K, med_G**:** $med = median(x)$ bezeichne den Median der Beobachtungen des gesamten Features. $med_K = median(K)$ den Median der Gruppe krank K und $med_G = median(G)$ den der Gruppe G

mad mediane absolute Abweichung (vom Median), engl. *median absolute deviation*

x' bezeichnet den Vektor der transformierten Beobachtungen nach robuster Standardisierung mittels *med* und *mad* bei Berechnung der OS.

$\tilde{x}$ bezeichnet den Vektor der transformierten Beobachtungen nach Zentrierung der Beobachtungswerte eines Features mit dem Median med_G der gesunden Gruppe bei Berechnung der FS. Entsprechend bezeichnen $\tilde{K} = (\tilde{k}_1,\ldots,k_{n_k})$ sowie $\tilde{G} = (\tilde{g}_1,\ldots,g_{n_G})$ die zentrierten Werte der einzelnen Gruppen.

Multivariate Methoden

T Anzahl der selektierten Variablen im ersten Schritt des neuen Workflows

OL Ordered List

J Jaccardindex

FSx Zusammenfassung der beiden Workflowvarianten FSOL und FSJ, bei denen die top-T-FS-Variablen gemäß eines Ähnlichkeitsmaßes basierend auf OL

bzw. J gruppiert werden

p_{OL} (empirischer) p-Wert zur Bewertung der Signifikanz des OL-basierten Ähnlichkeitsmaßes

t_{OL}, t_J Schwellenwerte für die jeweiligen Ähnlichkeitsmaße zur Bildung von Variablengruppen im FSx-Workflow

D Matrix der Größe $T \times T$, die die paarweisen Ähnlichkeiten (gemäß OL oder J) der top-T-Variablen enthält

$max.rk$ jeweilige Größe der Samplemengen mit den höchsten Expressionswerten, die zum Vergleich zweier Variablen herangezogen werden

r_{min} Mindestanteil von Variablen einer Variablengruppe, in denen ein Sample auf den top-$max.rk$-Rängen liegen muss, um für eine potentielle Subgruppe nominiert zu werden

$med^{FS}_{G_r}$ Das Ranking der Bedeutung der Variablengruppen und den von ihnen nominierten Samplesubgruppen in den FSx-Workflows basiert standardmäßig auf dem Median der FS-Scores der in der Gruppe G_r enthaltenen Variablen.

BC Biclustern, in dieser Arbeit meint BC immer den Plaid-Algorithmus

FSBC Anwendung des Biclusterns auf die Matrix der top-50-FS-Variablen

Simulationsstudie SimUni

n Fallzahl pro Gruppe

H_{0a}, H_{0b} die beiden möglichen Nullsituationen in SimUni: Unter H_{0a} entstammen alle Beobachtungen der Standardnormalverteilung, unter H_{0b} weisen beide Gruppen G und K eine Samplesubgruppe auf (nks).

p_{H0a} Anteil der Variablen der Nullsituation aus H_{0a}, $p_{H_{0a}} = 0.5, 1$

s Verteilungsszenario der Beobachtungen einer Subgruppe, s = I, II und III

q Subgruppenanteil der n Samples pro Gruppe

LR Likelihoodratio

z Misst den Unterschied zwischen den zugrundeliegenden Verteilungen der Subgruppe und der Standardnormalverteilung (d. h. der Verteilung der übrigen Beobachtungen). Abhängig von s ist z entweder δ, b oder σ.

ROC-Kurve *Receiver operating characteristics*-Kurve, ein Mittel zur grafischen Darstellung der Güte eines diagnostischen Verfahrens

AUC *area under the curve*, hier: Fläche unter ROC-Kurve

Simulationsstudie SimMulti

n Fallzahl pro Gruppe

n_{SG} Anzahl Samples in einer Subgruppe

p Anzahl Variablen im Datensatz

p_{SG} Anzahl Variablen, die sich auf die Subgruppe auswirken

δ Erwartungswert der Beobachtungen der Subgruppe, die aus der $N(\delta,1)$-Verteilung gezogen werden. SimMulti berücksichtigt Shifts der Größe $\delta = 2,3,4,6$.

Reale Datensätze

ParkCHIP Daten gemessen mit Autoantikörper-Microarrays von Parkinsonerkrankten und Gesundkontrollen

PD Morbus Parkinson, Abkürzung abgeleitet vom englischen *Parkinson's disease*

ALL Daten gemessen mit Affymetrix-Genexpressionschips von Patienten mit akuter lymphatischer Leukämie

NEG, BCR/ABL, E2A/PBX1 Bezeichnungen der Gruppen im ALL-Beispiel: Für Samples mit dem Label NEG liegt keine bekannte Mutation vor, die anderen beiden Gruppen sind nach ihren charakteristischen Fusionstranskripten benannt. Die E2A/PBX1-Gruppe soll von den multivariaten Verfahren detektiert werden.

LC-MS/MS *Liquid chromatography–mass spectrometry*, Flüssigchromatographie mit Massenspektrometrie-Kopplung, ein analytisches Verfahren zur Trennung und Bestimmung von Molekülen

DeNoPa Daten gemessen mit label-freier LC-MS/MS von Gesundkontrollen und therapie-naiven Parkinsonerkrankten

CSF *cerebrospinal fluid*, Gehirn-Rückenmarks-Flüssigkeit oder Liquor (cerebrospinalis)

Hb Hämoglobin, bekannt als roter Blutfarbstoff

ELISA *Enzyme Linked Immunosorbent Assay*, ein antikörperbasiertes Nachweisverfahren, im DeNoPa-Beispiel eingesetzt zur Bestimmung der Hb-Konzentration in den CSF-Proben

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Die Therapie von Krebspatienten hat sich in der letzten Jahren grundlegend verändert. Ursprünglich wurde für alle Patienten mit der gleichen Diagnose, die hauptsächlich Ursprungsorgan und Staging berücksichtigte, eine Standardtherapie gewählt, die im Mittel über alle Patienten einen guten Kosten/Nutzen-Kompromiss darstellen sollte. An die Stelle dieses „Gießkannenprinzips“ ist mittlerweile in vielen Fällen die *individualisierte* Therapie getreten [1]. Zunächst auch als *personalisierte* oder *targeted* Therapie bezeichnet, wird heutzutage der Ausdruck *precision medicine* bevorzugt. Dadurch soll der Eindruck vermieden werden, dass für jeden Patienten eine personalisierte, einzigartige Therapie entwickelt wird [2]. Unabhängig von der Terminologie ist eines der formulierten Ziele, genau die Behandlung auszuwählen, die dem individuellen Patienten bestmögliche Therapieergebnisse bei minimalen Nebenwirkungen verspricht.
Nicht immer müssen die unterschiedlichen zugrundeliegenden pathologischen Mechanismen, die für die Heterogenität einer Erkrankung verantwortlich sind, vollständig aufgeklärt sein, um diese Entscheidung treffen zu können. Eine gesunde Zelle kann auf unterschiedlichen Wegen zu einer Tumorzelle entarten, beispielsweise durch die Beteiligung verschiedener Onkogene. Hinweise auf diesen spezifischen Entstehungsweg (Pathomechanismus) bleiben etwa durch Fusionstranskripte in den Krebszellen erhalten und lassen sich in molekularen Analysen nachweisen. Dabei können sich die Unterschiede zwischen den verschiedenen Krankheitstypen auf mehreren molekularen Ebenen zeigen und so werden neben der Genexpression heutzutage auch microRNA-Expression oder Proteinabundanzen untersucht. Aus diesen Daten lassen sich entweder Rückschlüsse auf die vielversprechendste verfügbare Therapie ziehen oder Erkenntnisse über bisher unbekannte Pathomechanismen einzelner Subtypen gewinnen. Diese können dann im besten Fall zur Identifikation neuer *drug targets* und der Entwicklung neuer Therapieansätze genutzt werden.
In den vergangenen Jahren gelang es mithilfe unterschiedlicher molekularer Hochdurchsatztechnologien, verschiedene Subgruppen von Patienten innerhalb einer Krankheit zu identifizieren und zu charakterisieren. Dass sich die Patienten z. B. entsprechend ihrer Genexpressionsmuster in Subgruppen (SG) unterschiedlicher Krankheitstypen einteilen lassen, wurde bereits für verschiedene Arten von Krebs gezeigt, unter anderem für Brust-, Lungen- und Prostatakrebs [3, 4, 5] sowie für akute lymphatische Leukämie [6].

Die bisherigen Erkenntnisse in der individualisierten Medizin sind enorm, leider stehen aber nur für einen geringen Anteil von Krankheiten bereits maßgeschneiderte Therapien für den Patienten zur Verfügung. Um die Forschung auf diesem Gebiet zu fördern, wurde die individualisierte Medizin nicht nur auf Bundesebene

zu einem prioritären Aktionsfeld erklärt, sie wird auch vom Bundesministerium für Bildung und Forschung (BMBF) von 2013 bis 2016 mit bis zu 360 Mio.€ gefördert. Die EU-Fördermittel, die innerhalb des 7. Forschungsrahmenprogramms zur Verfügung gestellt wurden, belaufen sich auf rund 1.2 Mrd.€. Die Weiterentwicklung der personalisierten Medizin steht auch im Folgeprogramm Horizont 2020 (`http://www.horizont2020.de/`) weiter im Fokus. So wurde beispielsweise die CSA (*coordination and support action*) *PerMed* gegründet, um die europäischen Bestrebungen im Bereich der personalisierten Medizin zu bündeln und voranzutreiben.

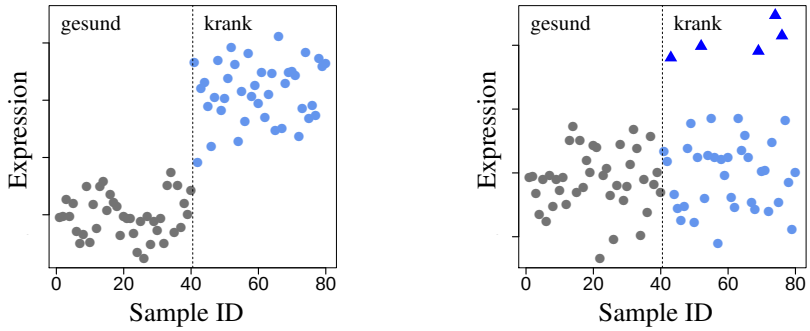
Die Basis der personalisierten Medizin ist der Einsatz von Biomarkern. Ganz allgemein bezeichnet der Begriff *Biomarker* eine objektiv messbare Größe, die zur Bewertung von normalen biologischen Prozessen, pathologischen Prozessen oder von Reaktionen auf pharmazeutische oder andere therapeutische Interventionen herangezogen werden kann (gemäß der Definition der *Biomarkers Definition Working Group*, [7]). Konkreter umfasst diese Definition beispielsweise folgende Einsatzmöglichkeiten: die Einordnung in Risikogruppen, Diagnose einer bestimmten Krankheit, Differentialdiagnose, Therapiewahl bzw. Prognose von Therapieansprechen, das Monitoring des Krankheitsverlaufs oder die Bestimmung einer Langzeitprognose.
Ähnlich vielfältig wie die Einsatzmöglichkeiten sind auch die verwendeten Messtechniken. Allein für das Beispiel Krebs reicht das mögliche Spektrum von der Patientenphysiologie über spezifische Moleküle in Körperflüssigkeiten bis hin zu Gen- oder Proteinexpressionsprofilen [8]. Doch nicht nur für Krebserkrankungen gewinnen Biomarker an Bedeutung: Während im Bereich der neurodegenerativen Erkrankungen zur Zeit intensiv an Biomarkern unter anderem zur Differentialdiagnose geforscht wird [9], wurden in der Psychiatrie bereits blutbasierte Biomarker zur Beurteilung von Selbstmordtendenzen untersucht [10].

Ein häufig verwendetes experimentelles Design zur Identifikation neuer diagnostischer Biomarker ist der Zwei-Gruppen-Vergleich *gesund* gegen *krank*. Aus der differentiellen Analyse eines entsprechenden hochdimensionalen Datensatzes einer omics-Technologie werden dabei neue Hypothesen abgeleitet und interessante Biomarkerkandidaten ausgewählt. Zur notwendigen Detektion von Lageunterschieden zwischen den beiden Gruppen kommen üblicherweise Students t-Test, Wilcoxons Rangsummentest oder Varianten wie der *moderated t-test* zum Einsatz.
Der *moderated t-test* [11] wirkt dem Effekt entgegen, dass gerade in Hochdurchsatzstudien mit kleinen Gruppengrößen Variablen mit zufällig sehr kleiner Varianz ein „zu gutes“ Ranking zugewiesen bekommen, insbesondere im Bereich niedriger Expression bzw. Intensität. Dazu wird der beobachtete Lageunterschied jeder

Variable nicht wie beim gewöhnlichen t-Test durch die zugehörige Schätzung der Standardabweichung s dividiert, sondern durch $s + s_0$, wobei die Konstante s_0 ein „kleiner“ Wert ist, der aus dem gesamten Datensatz berechnet wird. Obwohl ursprünglich für die Anwendung in Microarraystudien entwickelt, lässt sich der Ansatz auch auf andere, modernere Technologien anwenden, beispielsweise auf RNA-Seq- oder Proteomikmessungen [12, 13].

Alle genannten Lokationstests basieren auf der Annahme homogener Gruppen und sind daher am besten geeignet, um Variablen mit einem gleichmäßigen Shift zwischen den Gruppen zu detektieren. Das Expressionsmuster eines entsprechenden „optimalen“ Markerkandidaten ist in Abbildung 1(a) schematisch dargestellt. Für eine Reihe von heterogenen Krankheiten scheinen solche optimalen Marker aber schlicht nicht zu existieren. Aufgrund dieser in den letzten Jahren gereiften und akzeptierten Erkenntnis wird im Zuge der individualisierten Medizin in Hochdurchsatzdaten immer häufiger explizit nach Patientensubgruppen gesucht [14, 15, 16].
In diesem Fall ist das Ziel das Auffinden von Variablen, die als Marker für eine Subgruppe von Patienten anstatt für das gesamte Patientenkollektiv fungieren können. In diesen Variablen zeigt sich kein Expressionsunterschied zwischen den Beobachtungen der Gesunden und denen der Mehrheit der Kranken. Allein in einer Teilmenge der Kranken liegen deutlich erhöhte Werte vor (siehe Abb. 1(b)). Bei einer solchen Variable könnte es sich um eines der zuvor angesprochenen Onkogene handeln, das nur in einem kleinen Teil der Patienten aktiv ist. Ebenso könnte sich die Subgruppe in ihrer Prognose, im Krankheitsstadium oder in Bezug auf Thereapieansprechen von den übrigen Patienten unterscheiden.

Je nach Anzahl der Patienten in der Subgruppe und der Ausprägung des Unterschieds zwischen der Subgruppe und den übrigen Samples können auch die üblichen oben genannten Tests bei der Detektion solcher subgruppenanzeigenden Variablen nützlich sein. Wie bereits angesprochen, widerspricht aber das gesuchte Verteilungsmuster explizit der Annahme homogener Gruppen (genauer: identischen Verteilungen innerhalb der Gruppen). Mit zunehmender Bedeutung der individualisierten Medizin wächst der Wunsch nach speziellen Methoden zur Subgruppendetektion.
Der Begriff *Subgruppendetektion* ist in der Literatur allerdings nicht eindeutig definiert und der Bekanntheitsgrad bisher entwickelter Methoden ist gering. Viele Anwender aus den Lebenswissenschaften verstehen unter Subgruppendetektion schon die Betrachtung von Biplots oder Dendrogrammen nach einer Hauptkomponentenanalyse (PCA) bzw. nach hierarchischem Clustern. Zeigt sich dabei keine „auffällige“ Probengruppe, wird bereits der Schluss gezogen, dass die Daten keine Hinweise auf Subgruppen enthalten. Dabei wird nicht beachtet, dass die-



(a) klassischer Marker (b) Subgruppenmarker

Abbildung 1: Schematische Plots von Markerkandidaten. Bei der Darstellung von Genexpressionswerten beispielsweise repräsentiert ein Plot jeweils ein Gen, die Expression wird gegen die Probennummer aufgetragen. Als Dreieck dargestellt sind die Samples einer Subgruppe, als Kreise die übrigen Beobachtungen unabhängig von ihrer Gruppenzugehörigkeit. (a) klassischer Marker zur Trennung zweier homogener Gruppen insgesamt (homogener Shift), (b) Subgruppenmarker zur Identifikation der Patientensubgruppe (partieller Shift).

se Verfahren die Datenstruktur auf globaler Ebene darstellen und Abweichungen in kleineren Sample- und Variablengruppen vernachlässigt werden. Zusätzlich ist die Betrachtung der jeweiligen Plots subjektiv und schwer vergleichbar.
Falls sich abgegrenzte Samplesubgruppen erkennen lassen, werden die beteiligten Samples sowie das zugehörige spezifische Expressionsmuster näher untersucht. Gruppen von Proben, die mithilfe solch globaler Methoden identifiziert werden können, unterscheiden sich für gewöhnlich stark und/oder in einer größeren Variablenmenge von den übrigen Samples. Während diese Art von Samplegruppen offensichtlich relevant ist, stehen im Fokus der vorliegenden Arbeit die weniger auffälligen, kleineren Subgruppen, die nur in einer geringen Anzahl von Variablen einen Expressionsunterschied aufweisen. Da der Einfluss solcher Unterschiede für die Darstellung der Gesamtstruktur und -variation eines hochdimensionalen Datensatzes üblicherweise zu vernachlässigen ist, sind sie mit den oben beschriebenen Methoden (PCA oder hierarchisches Clustern) im Allgemeinen nicht detektierbar. Stattdessen werden in diesem Fall speziell auf den Zweck der Subgruppendetektion zugeschnittene Methoden benötigt.
Die Verwendung univariater Ansätze erlaubt dabei, das bereits erwähnte Problem der Hochdimensionalität einiger multivariater Methoden zu umgehen. Das erste Ziel ist das Ranking der Variablen im Datensatz, sodass auf den Toprängen die Variablen zu finden sind, deren Expressionsmuster am besten mit dem definier-

ten SG-anzeigenden Muster übereinstimmen. Falls die jeweilige Methode eine explizite Definition der Subgruppe beinhaltet, wird jeder einzelnen Variable eine Menge von Patienten zugeordnet, die als Subgruppenkandidaten anzusehen sind. Die Informationen über die angezeigten Subgruppen werden dabei jedoch für die einzelnen Variablen unabhängig voneinander bewertet. Die meisten bestehenden Verfahren wurden im Kontext von Genexpressionsanalysen entwickelt und vorgestellt.
Den Grundstein der Subgruppendetektion in unserem Sinne legten wohl Tomlins et al. [17] mit COPA (*cancer outlier profile analysis*). Die Idee ist stark von der Anwendung in Krebsstudien (Vergleich von Krebs gegen Kontrolle) motiviert: Das Ziel von COPA ist die Identifikation von Genen, die an Translokationen zwischen einem aktivierenden Gen und einem von möglicherweise mehreren Onkogenen beteiligt sind. Dazu werden Paare von Genen gesucht, die eine große Anzahl disjunkter „Ausreißer-“Samples mit hohen Werten in der Krebsgruppe aufweisen, aber wenig oder keine Ausreißer in der Kontrollgruppe zeigen.
Teschendorff et al. [18] schlugen das zweischrittige Verfahren PACK (*profile analysis using clustering and kurtosis*) vor. Der Clustering-Schritt dient dabei der Vorauswahl von Variablen, deren Verteilungsmuster auf das Vorliegen einer Subgruppe hinweist. Durch die anschließende Berechnung der Kurtosis lassen sich Variablen mit zwei etwa gleich großen Gruppen (z. B. höhere Werte in *krank*) von Variablen mit einer kleineren Subgruppe trennen.
Auch Tibshirani und Hastie [19] beschäftigten sich mit der Subgruppendetektion und stellten OS (*outlier sum*) als mögliche Scoringmethode vor. Unter Verwendung robuster Schätzer wird zunächst pro Variable ein Schwellenwert berechnet, der der Ausreißerdefinition dient. Die Teststatistik berechnet sich als Summe der (normierten) Beobachtungswerte in der Gruppe *krank*, die diese Schwelle übersteigen. Basierend auf der Idee von OS präsentierte Wu [20] seine Variante ORT (*outlier robust t-statistic*), bei der die ebenfalls robuste Lage- und Varianzschätzung jedoch nur auf den Beobachtungen der Kontrollgruppe basiert.
Li et al. [21] wählten einen anderen Ansatz, bei dem die klassischen statistischen Tests zum Zwei-Gruppen-Vergleich wie Students t-Test oder Wilcoxons Rangsummentest iterativ auf kleiner werdende, jeweils gleich große Anteile der höchsten Werte aus beiden Gruppen angewendet werden. Für jeden dieser Teilvergleiche werden der p-Wert und ein Maß für die Überexpression in einem Score zusammengefasst. Die Variablen können dann anhand des jeweils maximalen beobachteten Scores gerankt werden.

Bisher bietet die Literatur keinen umfassenden Vergleich der bestehenden univariaten Methoden. Von Interesse ist dabei nicht nur der Einfluss von Gesamtstichproben- und Subgruppengröße, sondern auch der Einfluss unterschiedlicher Alternativhypothesen. In den kurzen Simulationsstudien, die teilweise in den Pu-

blikationen enthalten sind, wird fast ausschließlich der Fall untersucht, dass der Großteil der Beobachtungen einer Standardnormalverteilung entstammt und für die Beobachtungen in der Patientensubgruppe eine Verschiebung des Erwartungswertes um einen einzelnen festen Wert vorliegt. Die in der vorliegenden Arbeit vorgestellte Simulationsstudie SimUni berücksichtigt hingegen verschiedene Szenarien, die unterschiedlichen Verteilungen für die Beobachtungen der Subgruppe entsprechen. Innerhalb dieser Szenarien werden dabei zusätzlich verschiedene Grade der Abweichung betrachtet. Dadurch kann überprüft werden, ob sich eine Methode als gleichmäßig überlegen zeigt oder ob für unterschiedlich starke Abweichungen unterschiedliche Ansätze am besten geeignet sind.
Die univariaten Methoden zur Subgruppendetektion werden nicht nur untereinander, sondern auch mit dem häufig in Hochdurchsatzstudien eingesetzten t-Test verglichen. Auf diese Weise können die Situationen identifiziert werden, in denen der Informationsgewinn durch die Verwendung spezifischer Subgruppentests gegenüber einer Standardauswertung besonders groß ist. Ebenfalls im Vergleich enthalten ist die neue Methode Fisher Sum (FS) [22]. Das SimUni-Design berücksichtigt zudem zwei neue Aspekte: Zum einen wird in Form des Likelihoodratios erstmals eine theoretische obere Schranke berechnet, mit der die Methoden verglichen werden können. So kann beurteilt werden, ob für eine interessierende Situation die vorhandenen Methoden als ausreichend gut betrachtet werden können, oder ob eine spezifische neue Herangehensweise etabliert werden sollte.
Desweiteren wird die Definition der Nullsituation erweitert. Bisher wurden in vergleichbaren Studien nur jeweils eine einfache Nullsituation und eine einfache Alternative berücksichtigt. Dabei stammen entweder alle Beobachtungen beider Gruppen aus der Standardnormalverteilung oder es gibt genau eine Subgruppe in der als heterogen angesehenen Gruppe mit erhöhten Werten. SimUni hingegen berücksichtigt zusätzlich eine Nullsituation mit sogenannten *nicht-krankheitsspezifischen Subgruppen*, bei denen ein ähnlicher (kleiner) Anteil erhöhter Werte in beiden Gruppen zu beobachten ist. Dadurch werden sich bereits in der Simulation Unterschiede zwischen den univariaten Methoden zeigen, die sich auch bei der Anwendung auf reale Daten bestätigen. Die wesentlichen Ergebnisse dieser Studie sind bereits in Ahrens et al. [22] veröffentlicht.

Der zweite Teil dieser Arbeit befasst sich mit der Entwicklung einer multivariaten Strategie, die die Informationen aus den SG-anzeigenden Variablen zusammenführt, die mittels eines univariaten Scores ausgewählt werden. Wird eine Patientengruppe von mehreren Variablen konsistent als SG nominiert, so stärkt das ihre Evidenz gegenüber der rein univariaten Auswertung. Unter Umständen erleichtert es auch die Charakterisierung der Gruppe und ermöglicht neue oder vertiefte Einsichten in mögliche Pathomechanismen. Der neue auf der univariaten Fisher Sum basierende FSx-Workflow wird bezüglich seiner Detektionsgüte für eine Sample-

subgruppe mit einer bereits etablierten multivariaten Methode verglichen, dem sogenannten Biclustern [23]. Ziel dieser Methode ist ebenfalls die Identifikation von Samplegruppen, die in einer Teilmenge der Variablen ein ähnliches Expressionsmuster aufweisen. Auch hier liegt der Fokus in der entsprechenden Simulationsstudie SimMulti auf kleinen Subgruppen, die erwartungsgemäß schwerer zu detektieren sind. In den Gütevergleich aufgenommen wird neben diesen beiden Verfahren auch eine Kombination der beiden, bei der das Biclustern auf eine Teilmenge der Daten angewendet wird, die mittels des univariaten FS-Scores selektiert wird.

Für den umfassenden Vergleich der jeweiligen Methoden wird neben den Simulationsstudien SimUni und SimMulti auch auf die Analyse realer Datensätze zurückgegriffen. Der Vergleich der univariaten Methoden wurde in Ahrens et al. [22] anhand eines Proteinmicroarray-Experiments gezogen, bei dem im Rahmen des ParkCHIP-Projektes Serum-Autoantikörper von Parkinsonpatienten und Gesundkontrollen verglichen wurden. Da für diesen Datensatz keine wahre Patientensubgruppe bekannt ist, wird der Datensatz in dieser Arbeit nicht im Detail behandelt. Im Gegensatz dazu werden die multivariaten Methoden anhand der beiden Datensätze ALL und DeNoPa verglichen. Zunächst wird ausführlich der Datensatz ALL [24, 25] mit Genexpressionsdaten von Patienten mit akuter lymphatischer Leukämie behandelt. Die enthaltenen Daten können aufgrund der vorliegenden Informationen über molekulare Muster der Proben unterschiedlich gruppiert werden. Insbesondere kann ein Zwei-Gruppen-Vergleich zwischen zwei Gruppen konstruiert werden, bei dem eine der Gruppen eine bekannte Subgruppe enthält. So kann an diesem Beispiel die Güte der betrachteten Methoden bezüglich der Auswahl von Patienten als potentielle Subgruppe beurteilt werden.
Die multivariaten Verfahren werden ebenfalls auf den Proteomik-Datensatz DeNoPa angewendet, der mittels label-freier Massenspektrometrie generiert wurde. Die hier untersuchte Samplesubgruppe wird basierend auf einer sogenannten ELISA-Messung zur Bestimmung des Hämoglobingehalts der Probe definiert. In diesem Beispiel ist zusätzlich von Interesse, wie gut die beobachtete Übereinstimmung zwischen den beiden Technologien ELISA und LC-MS/MS ist.

Zusammengefasst verfolgt diese Arbeit zwei Hauptziele. Zunächst soll eine Empfehlung für eine univariate Scoringmethode ausgesprochen werden, die es erlaubt, subgruppenanzeigende Variablen in einem hochdimensionalen Datensatz zu identifizieren. Dazu werden bereits publizierte Ansätze sowie eine neue Methode vorgestellt (Kapitel 3) und verglichen. Im nächsten Schritt wird basierend auf dem ausgewählten Score eine multivariate Methode entwickelt, die die Informationen aus den potentiell subgruppenrelevanten Variablen kombiniert (Kap. 4). So sollen Variablengruppen gefunden werden, die gemeinsam auf eine Gruppe von

Samples als Subgruppe hinweisen. Auch die Performanz dieser neuen multivariaten Methode wird mit der eines etablierten Verfahrens verglichen. Sowohl für den univariaten als auch für den multivariaten Abschnitt basiert die Evaluation auf Simulationsstudien (Kap. 5) und realen Datensätze (Kap. 6). Vorab werden jedoch in Kapitel 2 die formulierten Fragestellungen und Ziele konkretisiert, sowie relevante Begriffe und Annahmen erläutert.

2 Zielsetzung und Gliederung der Arbeit

Dieses Kapitel konkretisiert die in der Einleitung formulierten Ziele der vorliegenden Arbeit. Dazu wird zunächst dargestellt, an welchem Punkt eines Forschungsprojekts die entwickelten Methoden zur Untersuchung möglicher Subgruppen (SG) zum Einsatz kommen und wie die gewonnenen Ergebnisse weiter genutzt werden können. Ausgegangen wird von einem Datensatz einer quantitativen (oder semi-quantitativen) omics-Technologie, der nach eingehender Qualitätskontrolle angemessen normalisiert wurde. Die Ergebnisse der explorativen SG-Detektionsverfahren können mithilfe von Enrichmentanalysen in bekanntes biologisches Wissen eingeordnet werden, um die Formulierung neuer Forschungshypothesen zu ermöglichen. Ferner werden relevante Begriffe definiert und getroffene Annahmen erläutert. Das Kapitel schließt mit einer Gliederung der restlichen Arbeit.

In allen Hochdurchsatzstudien sind eine gewissenhafte Planung des Experiments, eine angemessene Normalisierung und die Qualitätskontrolle der Daten unerlässlich, um valide Ergebnisse zu erhalten. Eine ausführliche Darstellung aller zu berücksichtigenden Aspekte ist im Rahmen dieser Arbeit nicht möglich, aber beispielsweise zu nennen sind Matching der Gruppen bzgl. Alter und Geschlecht, die Vermeidung bzw. Adjustierung von Batcheffekten (z. B. durch unterschiedliche Produktionschargen), sowie das Erkennen und Eliminieren fehlerhafter Chips, Proben oder Läufe vor der intendierten Analyse. Gerade Probleme bezüglich der letzten beiden Punkte spiegeln sich häufig auf der globalen Ebene wider und sind dann unter Umständen mittels PCA erkennbar. Speziell auf die Behandlung von Batcheffekten gehen beispielsweise Leek et al. [26] oder Turewicz et al. [27] ein. Für die spezifischen Aufgaben rund um die Datenvorverarbeitung steht in Bioconductor (`https://www.bioconductor.org/`) eine Reihe etablierter Lösungen für unterschiedliche omics-Plattformen zur Verfügung. Hier seien nur einige stellvertretend genannt:

- das Paket `arrayQualityMetrics` [28] berechnet verschiedene Qualitätsmetriken für Microarrays und bietet die Möglichkeit der automatischen Erstellung einer Reportdatei,
- `qcmetrics` untersucht ebenfalls Qualitätsmetriken, insbesondere für Microarray- und Proteomik-Datensätze,
- die Funktion `ComBat` des `sva`-Pakets erlaubt die Korrektur bekannter Batcheffekte, die z. B. bei Messungen mit größerem zeitlichen Abstand oder aus verschiedenen Laboren auftreten können (in dieser Arbeit verwendet bei der Vorverarbeitung der DeNoPa-Daten, Abschnitt 6.3).

Im Folgenden sei stets eine bestmögliche Datenvorverarbeitung und die grundsätzliche Vergleichbarkeit der Gruppen vorausgesetzt.

Die in dieser Arbeit behandelten Methoden sind grundsätzlich auf Daten verschiedener omics-Technologien anwendbar, z. B. auf relative Proteinabundanzen in label-freien Massenspektrometrieexperimenten oder auf Expressionswerte, die auf Ebene des Transkriptoms mithilfe von Genexpressionschips gemessen wurden. Da ein Großteil der Ansätze für die Analyse von Expressionsdaten entwickelt und vorgestellt wurde, ist die Notation in diesem Bereich von den entsprechenden Begrifflichkeiten geprägt. In einigen Darstellungen in dieser Arbeit ist daher der Begriff *Expression* als Platzhalter für die jeweils von der verwendeten Technologie gemessene Größe zu verstehen. Vor allem in den Anwendungsbeispielen und Simulationen werden die allgemeinen Terme Variable oder Feature den technologiespezifischen (z. B. *probe set*) vorgezogen. Dies dient der leichteren Nachvollziehbarkeit auch in Feldern, mit denen der Leser weniger vertraut ist.
Gegeben sei also ein hochdimensionaler Datensatz, auf dessen Basis zwei Gruppen verglichen werden sollen. Dabei kann es sich ebenso um den Vergleich von Kranken und Gesunden handeln wie um den Vergleich zweier Krankheiten oder Krankheitstypen untereinander. Allgemein wird jedoch eine Gruppe als homogen angesehen (z. B. Kontrollgruppe oder Gesunde), während die andere potentiell heterogen ist und auf mögliche Samplesubgruppen untersucht wird.
Es wird grundsätzlich empfohlen, sich einen Eindruck über die globale Struktur des Datensatzes zu verschaffen, bevor spezifische Analysen durchgeführt werden. Dazu können Scatterplots der Ladungen der Samples im Datensatz bezüglich der ersten Hauptkomponenten nützlich sein. Im Verlauf der Arbeit werden diese Plots abkürzend mit „PCA-Plots“ oder „Scatter der Hauptkomponenten“ bezeichnet. Obwohl die Möglichkeit besteht, schon an dieser Stelle der Auswertung Hinweise auf Ausreißer- bzw. Subgruppensamples zu erkennen, sei nochmals betont, dass aus dem Fehlen solcher Subgruppen nicht auf die Homogenität der Gruppen im Sinne dieser Arbeit geschlossen werden kann. Die besprochenen SG-Detektionsmethoden sind nicht als Konkurrenz oder Alternative für globale Verfahren wie PCA oder hierarchisches Clustern zu sehen, sondern als ergänzende Methoden zur Beantwortung einer spezifischen Fragestellung.
Weiterhin kann mithilfe der PCA-Plots auch der globale Unterschied zwischen den beiden experimentellen Gruppen beurteilt werden. Die in dieser Arbeit beschriebenen Verfahren zur Subgruppendetektion liefern den größten Informationsgewinn über tatsächlich enthaltene Subgruppen, wenn sich die beiden Gruppen insgesamt „ähnlich“ sind. Dies meint, dass keine oder nur wenige Variablen im Datensatz die experimentellen Gruppen eindeutig trennen können. Der Grund ist schlicht, dass eine Reihe der später vorgestellten Methoden nicht zwischen homogenem und partiellem Shift unterscheidet und so beide Kandidatentypen gute Scores erhalten können. Schematische Darstellungen der Expressionsmuster von Variablen mit diesen beiden Shifttypen wurden eingangs in Abb. 1 gezeigt. Es ist

somit grundsätzlich zu empfehlen, die Natur der Verteilungsmuster der einzelnen Variablen auf den Toprängen gegebenfalls einer weiteren Prüfung zu unterziehen, falls ausschließlich Interesse an partiellen Shifts besteht. Dazu bietet sich entweder die visuelle Inspektion der Expressionsmuster an oder gerade bei größerer Variablenanzahl ein Filtern gemäß der p-Werte des t-Tests (vgl. [22]). Letzteres bietet sich besonders an, wenn bereits in den PCA-Scattern eine deutliche Abgrenzung der Gruppen erkennbar ist.
Ein weiteres Verteilungsmuster, das von einigen Methoden unerwünschterweise auf die Topränge gewählt werden kann, sind sogenannte nicht-krankheitsspezifische (nks) Subgruppen. Diese Bezeichnung geht zurück auf den typischen Vergleich *gesund vs. krank*. In dem Fall, dass stattdessen verschiedene Krankheitstypen oder -stadien verglichen werden, entspricht die „kranke“ Gruppe der, die auf mögliche Subgruppen untersucht werden soll. Die entsprechenden Expressionsmuster zeigen in beiden zu vergleichenden Gruppen eine Subgruppe von Samples, die sich beispielsweise durch erhöhte Werte von den übrigen unterscheiden. Eine solche Disregulation lässt sich gelegentlich auf Confoundervariablen wie z. B. das Geschlecht zurückführen. Weitere Möglichkeiten wären technische Varianz, die Einnahme eines Medikaments oder eine sonstige Behandlung, die nicht mit dem untersuchten Gruppenunterschied zusammenhängt.
Im Allgemeinen tragen Variablen, die eine solche nks SG anzeigen, nicht zur Identifikation und Charakterisierung unbekannter Subgruppen in der als heterogen angenommenen Gruppe bei. Daher sollten die univariaten Scoringverfahren Variablen mit krankheitsspezifischem SG-Expressionsmuster eine höhere Bedeutung zumessen als solchen mit nks Subgruppen. Bei der Anwendung auf reale Datensätze ist zu beachten, dass es in der Praxis gelegentlich zu falschen Gruppenzuordnungen kommen kann. Befindet sich ein Proband der vermeintlich gesunden Gruppe in einem sehr frühen Stadium der interessierenden Krankheit, das noch nicht diagnostizierbar ist, können sich trotzdem schon subgruppenspezifische Expressionen zeigen.

Den globalen Methoden gegenüber stehen die SG-Detektionsverfahren, bei denen auch und gerade kleinere Unterschiede in den Expressionsprofilen aufgedeckt werden sollen, die auf der übergeordneten Ebene zu vernachlässigen wären. Die Entwicklung und Anwendung solcher Verfahren liegt vielfach im Bereich der Onkologie. Im Falle univariater Methoden wird jeder Variable ein Score oder p-Wert zugewiesen, auf dessen Basis ein Ranking der Variablen im Datensatz möglich ist. Im Idealfall zeigt sich in den Expressionsmustern der Variablen auf den besten Rängen jeweils eine Gruppe von Samples, die im Vergleich zu allen übrigen Beobachtungen deutlich erhöhte Werte aufweist. Obwohl vom rein datenanalytischen Standpunkt die Regulationsrichtung einer Subgruppe unerheblich wäre, liegt der Anwenderfokus in der Biomarkersuche aus praktischen Gründen häufig zunächst

auf hochregulierten Subgruppen. Ein Grund ist, dass dies speziell auf die Klasse der Onkogene zutrifft. In einem allgemeineren Kontext erleichtern hochregulierte Marker die Analysen beispielsweise bei antikörper-basierten Färbungen. Niedrige Werte, d. h. negative Färbeergebnisse, könnten ebenso auf Probleme mit dem Antikörper zurückzuführen sein.
Univariate Ansätze zur Subgruppendetektion werden gelegentlich vorschnell als nicht angemessen kritisiert. Dies wird meist mit der Aussage begründet wird, dass der Gedanke eines univariaten Biomarkers für die heutzutage interessierenden komplexen Fragestellungen überholt sei. Dem liegt jedoch das Missverständnis zugrunde, dass das Ziel einer univariaten Auswertung grundsätzlich die Auswahl eines einzelnen Kandidaten ist. Tatsächlich ist es jedoch sinnvoll durch die univariate Vorauswahl die Datenlage für nachgeschaltete multivariate Methoden zu verbessern, indem die informationstragendsten Variablen selektiert werden.

Jeder der hier vorgestellten Ansätze zur SG-Detektion ist als Mittel zur explorativen Datenanalyse und Hypothesengenerierung zu betrachten. Der Bestimmung einer potentiellen Samplesubgruppe und/oder der auf sie hinweisenden Variablengruppen sollte in der Praxis stets eine weitergehende Analyse folgen. Falls Informationen über Kovariablen verfügbar sind (z. B. klinische Parameter, Laborwerte oder Überlebenszeiten), so könnten mögliche Assoziationen der potentiellen Samplegruppe mit diesen Kovariablen untersucht werden.
Bezüglich gefundener Variablengruppen, die eine mögliche Subgruppe anzeigen, besteht der erste Evaluationsschritt in der Einordnung in bekanntes Wissen. Ein bereits beschriebener Zusammenhang zwischen dem experimentellen Faktor und einer oder mehreren der interessierenden Variablen stärkt die Evidenz der Subgruppe. Allerdings können für die Hypothesengenerierung und die Eröffnung neuer Forschungswege auch oder gerade die Subgruppen interessant sein, deren Bedeutung bisher unklar ist.
Während bei kleineren Variablengruppen eine manuelle Literaturrecherche ausreichend sein kann, sollte für eine größere Menge von Variablen eine Enrichmentanalyse in Betracht bezogen werden. Bereits ohne die Variablengruppierung lassen sich Enrichmentansätze auf die sortierte Ergebnisliste anwenden, die das univariate Scoring der Variablen reflektiert. Möglichkeiten hierfür sind Enrichment von GO-Terms oder der Zugehörigkeit zu bestimmten biologischen Pathways. Zu den häufig genutzten frei verfügbare Tools zählen zum Beispiel

- `topGO` (*topology-based gene ontology scoring*), verfügbar für **R** über Bioconductor [29],
- Reactome (`http://www.reactome.org/`) oder
- DAVID (`https://david.ncifcrf.gov/home.jsp`).

Die Interpretation der so erhaltenen Ergebnisse sowie die Beurteilung ihrer Relevanz obliegt gewöhnlich dem klinischen oder biologischen Partner eines Projekts und wird in dieser Arbeit nicht behandelt.

Der Rest dieses Kapitels stellt nochmals die beiden Hauptfragestellungen dieser Arbeit heraus und beschreibt die zu ihrer Beantwortung verfolgten Strategien. Alle dargestellten Analysen und Grafiken wurden mithilfe der jeweils aktuellen **R**-Version erstellt. An relevanten Stellen werden die exakten Versionsnummern angegeben (Annotation mit Gennamen der realen Datensätze).
Der erste Teil der Arbeit dient der **Auswahl einer geeigneten univariaten Scoringmethode zum Ranking von subgruppenanzeigenden Variablen in einem hochdimensionalen Datensatz**. Im Fokus steht dabei die Identifikation von bisher schwer zu entdeckenden Subgruppen, die mit 10-15% nur einen kleinen Anteil der heterogenen Gruppe ausmachen. Zunächst gibt Kapitel 3 einen Überblick zum Thema Subgruppendetektion mithilfe univariater Methoden. Nach einer Literaturübersicht (Abschnitt 3.1) und der detaillierteren Beschreibung einiger bereits publizierter SG-Detektionsmethoden (3.2) wird auch der im Rahmen dieser Arbeit entwickelte Score Fisher Sum (FS) vorgestellt (3.3). Ausgewählte Methoden werden sowohl anhand einer umfassenden Simulationsstudie (SimUni, 5.1 und 5.2) als auch anhand von realen Daten verglichen (6.1). In SimUni wird die Performanz der Methoden für drei Verteilungen der Subgruppenbeobachtungen und wachsenden Unterschied z zu den übrigen Beobachtungen untersucht. Die üblicherweise zum Gütevergleich verwendeten ROC-Kurven sind durch diese zusätzliche Dimension nicht mehr praktikabel. Stattdessen ergibt sich durch Integration, d. h. durch Betrachtung der Plots AUC gegen z, eine übersichtliche Darstellung der Ergebnisse.
Aufbauend auf den Ergebnissen zu den univariaten Methoden dient der zweite Teil der Arbeit der **Entwicklung einer multivariaten Methode zur Identifikation und Charakterisierung insbesondere kleinerer Subgruppen, die sich auf wenige Variablen auswirken**. Die wesentlichen Schritte dazu sind die Vorauswahl der top FS Variablen und die anschließende Gruppierung dieser Variablen mit einem geeigneten Ähnlichkeitsmaß, das Übereinstimmungen in den angezeigten Subgruppen widerspiegelt. Zu Beginn wird eine Übersicht der Literatur zu multivariaten Ansätzen zur SG-Detektion gegeben (4.1). Der entwickelte FSx-Workflow (4.3) wird wiederum in Simulationen (SimMulti, 5.3 und 5.4) und anhand realer Daten (6.2 und 6.3) mit einer bereits etablierten Methode, dem Biclustern [23], verglichen. Vorgestellt werden zwei unterschiedliche Ähnlichkeitsmaße, die zugehörigen FSx-Varianten werden dementsprechend als FSOL (4.3.1) und FSJ (4.3.2) bezeichnet. Zum Biclustern (BC) wurde der Plaid-Algorithmus (4.2) als Referenzmethode gewählt, eine beliebte Methode zur Auswertung von Hochdurchsatzstudien vor allem im Bereich der Genexpressionsanalyse.

Zusätzlich zu den drei Methoden FSOL, FSJ und BC wird auch eine Kombination FSBC (4.4) getestet, die die Vorteile der univariaten Vorselektion mit dem etablierten Bicluster-Ansatz verbinden soll. Dazu wird der Plaid-Algorithmus auf den Teildatensatz angewendet, der basierend auf dem neuen univariaten FS-Score selektiert wird. Die vier Methoden werden jeweils in Paaren als die FSx-Workflows bzw. die Bicluster-basierten Workflows zusammengefasst. Die Kombination FSBC wird dabei explizit nicht als weitere Variante des FSx-Workflows verstanden, da sie hinsichtlich der Auswertung trotz der Vorselektion weiterhin dem Biclustern ähnlicher ist. Dies wird in den später gezeigten Anwendungsbeispielen deutlich (Kapitel 6).
Als Gütekriterium für die multivariaten SG-Detektionsmethoden dient auf realen und simulierten Daten der Jaccardindex (5.3.3), der die Größen von Schnitt und Vereinigung der wahren und vom jeweiligen Algorithmus detektierten Samplesubgruppe ins Verhältnis setzt. In SimMulti wird der Einfluss verschiedener datensatz- und methodenspezifischer Parameter auf die Detektionsgüte für verschiedene Stichprobengrößen und Subgruppengrößen untersucht (5.4.1).
Den Abschluss der Arbeit bilden in Kapitel 7 die Diskussion der erzielten Ergebnisse und die Formulierung weiterer Ziele, die im Rahmen dieser Arbeit nicht verwirklicht werden können.

3 Univariate Verfahren zur Identifikation von Patientensubgruppen

In vielen Bereichen der Lebenswissenschaften sind die Wissenschaftler und Forscher bei der Auswertung von Hochdurchsatzdaten auf die Verwendung kommerzieller Software angewiesen. Diese bietet aber in den seltensten Fällen ausreichend Flexibilität, um eine auf die jeweilige Fragestellung abgestimmte Methodenauswahl zu treffen. Die verfügbaren Methoden, üblicherweise Variationen des t-Tests, sind zur Detektion von Subgruppen nur unter bestimmten Bedingungen geeignet. Vor allem zur Detektion kleinerer Subgruppen ist die Anwendung einer speziellen SG-Detektionsmethode zu empfehlen.
Im Folgenden wird in 3.1 ein Überblick über bestehende univariate Methoden im Bereich der Subgruppendetektion gegeben. Die Gründe für den Fokus auf univariate Methoden wurden in Kapitel 2 dargelegt. Abschnitt 3.2 liefert detaillierte Darstellungen ausgewählter Methoden, die später (mehrheitlich) hinsichtlich ihrer Performanz genauer verglichen werden. Die getroffene Auswahl repräsentiert verschiedene Klassen von SG-Detektionsansätzen, um einen Eindruck von der Vielzahl der Möglichkeiten zu vermitteln. Neben an die Idee der t-Statistik angelehnten Scores werden z. B. Maßzahlen zur Beurteilung der Normalität genutzt.
Die Gruppe aus COPA, OS und ORT wird als Beispiel für die Weiterentwicklung bestehender Methoden vorgestellt. Die entsprechenden Arbeiten bauen thematisch aufeinander auf und vergleichen die Methoden explizit miteinander. Im Gegensatz dazu wurden andere Verfahren isoliert dargestellt ohne einen direkten Vergleich mit anderen spezifischen SG-Detektionsmethoden in Simulation oder Anwendung zu präsentieren. Die ausführliche Diskussion aller bisher vorgeschlagenen Methoden ist im Rahmen dieser Arbeit nicht möglich, so sei hier bei weiterem Interesse beispielsweise auf Alternativen von Lyons-Weiler et al. [30], Lian [31], Wang und Rekaya [32], Hu [33], Chen et al. [34] oder van Wieringen et al. [35] verwiesen.

3.1 Literaturübersicht

Häufig wird Students t-Test als die Standardmethode zur differentiellen Analyse hochdimensionaler Daten angesehen. Die Idee ist die Beurteilung des beobachteten Verhältnisses von Lageunterschied und Streuung der beiden zu vergleichenden Gruppen. Dabei wird innerhalb jeder Gruppe eine identische zugrundeliegende Verteilung für alle Beobachtungen angenommen. Da diese Annahme bei subgruppenanzeigenden Variablen verletzt ist, wurden für ihre Detektion alternative Methoden entwickelt, die die Eigenschaften eines (krankheitsspezifischen) SG-Expressionsmusters berücksichtigen. Eine tabellarische Übersicht der Methoden findet sich in Tabelle 1.

Methode	Jahr	Ref.	Ansatz
COPA	2005	[17]	Quantil nach robuster Standardisierung
OS	2007	[19]	Summe standardisierter Ausreißerwerte
ORT	2007	[20]	„Robustifizierter“ t-Test
PADGE	2007	[21]	Testen auf Teildatensätzen
PACK	2006	[18]	Kurtosis
MinM	2007	[36]	Minimum Fishertest
FS	2013	[22]	Differenz von Beobachtungssummen

Tabelle 1: Übersicht der im Folgenden näher besprochenen univariaten Methoden zur Subgruppendetektion.

Ein beliebter Ansatz in der SG-Detektion ist die Verwendung robuster Schätzer für Lage und Streuung um eine Verzerrung durch vorhandene Subgruppen zu vermeiden. So geschehen beispielsweise in der *cancer outlier profile analysis*, kurz COPA [17] (Abschnitt 3.2.1). Nach einer robusten Zentrierung und Skalierung der Beobachtungen einer Variable wird die Größe eines vorgegebenen Quantils (z. B. 90%-Quantil) der heterogenen Gruppe betrachtet. Anhand seiner Größe lassen sich die Variablen des Datensatzes ranken, wobei ein großer Wert für eine deutliche Ausreißergruppe spricht. Aufbauend auf COPA wurde zwei Jahre später die *outlier sum* [19] (OS, 3.2.2) vorgestellt. Statt die Variablen mittels einzelner Quantile zu bewerten, wird als Statistik die Summe der (wiederum mit robusten Methoden normierten) Beobachtungen gebildet, die nach einem gegebenen Kriterium als Ausreißer definiert werden. Schließlich wurde die *outlier robust t-statistic* [20] (ORT, 3.2.3) vorgeschlagen, die im Vergleich zur OS eine verbesserte Schätzung von Lage und Streuung und somit eine angemessene Standardisierung bieten soll. Diese Methodengruppe führte zu einer „Robustifizierung“ der t-Statistik, deren Eignung zur SG-Detektion in kleineren vergleichenden Simulationen und anhand realer Daten gezeigt wurden.
Alternativ zur Modifikation bestehender statistischer Tests wählten Li et al. [21] einen anderen Ansatz: Sie schlugen 2007 mit *percentile analysis for differential gene expression* [21] (PADGE, 3.2.4) vor, mit einem „gewöhnlichen“ Test (z. B. dem t-Test) kleiner werdende Anteile der jeweils höchsten Werte aus beiden Gruppen zu vergleichen. Für jeden Teilvergleich werden Effektgröße und p-Wert berechnet. Aus der Veränderung dieser Größen lassen sich Rückschlüsse auf das Vorliegen einer Subgruppe zu ziehen.
Weitere Ansätze ergeben sich aus der methodischen Ähnlichkeit zwischen der Suche nach Patientensubgruppen und der Suche nach Ausreißerproben. Die Beurteilung der Normalität der Daten kann beispielsweise in beiden Fragestellungen ein hilfreiches Kriterium sein. Bei der Methode *profile analysis using clustering and kurtosis* [18] (PACK, 3.2.5) erfolgt eine solche Beurteilung auf der Basis der

Kurtosis. Positive Werte weisen auf die Existenz einer kleineren Ausreißergruppe hin, wie sie in dieser Arbeit bei den gesuchten Patientensubgruppen zu finden ist. Negative Werte hingegen treten auf, wenn die beiden Gruppen etwa die gleiche Größe haben, was beispielsweise bei einer homogenen Lageverschiebung zwischen den experimentellen Gruppen der Fall sein kann.

Als weitere Klasse lassen sich die count-basierten Methoden zusammenfassen. Dabei wird im Wesentlichen die Sampleanzahl einer Gruppe bestimmt, die nach einem gewählten Kriterium als auffällig oder extrem gewertet wird. In einer der früheren Veröffentlichungen zu SG-Detektionsmethoden stellten Lyons-Weiler et al. 2004 den *permutation percentile separability test* PPST [30] vor. Die Methode dient der Erkennung von Variablen, in denen auffällig viele Werte von Samples aus einer heterogenen Gruppe in den äußeren Rändern der homogenen Vergleichsgruppe liegen, d. h. beispielsweise viele Tumorproben zeigen Werte oberhalb des 95%-Quantils der Gesundkontrollen in der jeweiligen Variable. Eine Implementierung von PPST wurde den Anwendern über die ebenfalls 2004 veröffentlichte Webanwendung caGEDA [37] zur Verfügung gestellt, die später beispielsweise in einem Review über Bioinformatik-Ressourcen für die Krebsforschung [38] vorgestellt wurde. In den oben genannten Quellen wird sie weder diskutiert noch in die Vergleiche einbezogen, es findet sich nur eine kurze Erwähnung in Tibshirani und Hastie [19] als weitere Methode mit dem Ziel der SG-Detektion.
Während beim PPST nur ein einzelnes Quantil gewählt wird, entschied sich Love [36] dafür, nacheinander alle Beobachtungen einer Gruppe als cut-off zu wählen und im Wesentlichen die Abhängigkeit der beiden binären Variablen *Beobachtung oberhalb des cut-offs* und *Gruppenzugehörigkeit* mithilfe des exakten Tests von Fisher zu bewerten. Jeder Variable wird dann der minimale p-Wert all dieser Vergleiche zugewiesen und für das Ranking herangezogen. Diese Methode mit der Bezeichnung *minimum M statistic* (MinM, 3.2.6) wurde in einer Software zur Auswertung von Proteinmicroarrays implementiert und ist außerdem über das **R**-Paket `PAA` [39] verfügbar. Ausgehend von den Vierfeldertafeln wie sie in MinM verwendet werden, werden beim Scoring mithilfe der *Fisher Sum* (FS) nicht die bloßen Anzahlen in den Zellen beurteilt, sondern auch die zugehörigen Werte der entsprechenden Samples berücksichtigt. Diese Methode wurde in [22] vorgestellt und in einer umfassenden Simulationsstudie unter anderem dem t-Test, OS und ORT verglichen.
Abschließend sei die Arbeit von Vuong et al. [40] erwähnt, in der insbesondere das Verhalten des t-Tests und OS verglichen werden. In der vorgestellten Simulation wird dazu eine neue Methode für die Generierung von subgruppenanzeigenden Expressionsmustern (*hinge function*) verwendet, mit der die differentielle Expression in den Rändern und im Zentrum der Verteilung unabhängig voneinander variiert werden können. Weiterhin wird eine quantilbasierte grafische Methode

zur Charakterisierung der Verteilungen von interessierenden Kandidaten vorgeschlagen. Die Autoren merken an, dass trotz der wachsenden Anzahl publizierter SG-Detektionsmethoden der umfassende Vergleich der falsch-positiv-Raten und der Power der Methoden unter verschiedenen Alternativen bisher vernachlässigt worden sei. Diese seien aber dringend nötig, um letztendlich auch regulatorische Behörden wie die amerikanische *Food and Drug Administration* vom Nutzen dieser spezifischen Methoden zu überzeugen.

3.2 Detaillierte Beschreibung ausgewählter univariater Methoden

Zur leichteren Vergleichbarkeit der im Folgenden beschriebenen Methoden wird eine einheitliche Notation verwendet, die von denen in den Originalmanuskripten abweichen kann. Dies bezieht sich auch auf die beiden zu vergleichenden Probengruppen. Da ein Großteil der Methoden zur Analyse von onkologischen Studien entwickelt wurde, wird häufig auf die Bezeichnungen Tumor und Kontrolle zurückgegriffen. In dieser Arbeit wird der etwas allgemeinere Vergleich Krank $K = (k_1, \ldots, k_{n_K})$ gegen Gesund $G = (g_1, \ldots, g_{n_G})$ beschrieben, generell gelten die Methoden aber für Zwei-Gruppen-Vergleiche bei Annahme jeweils einer heterogenen und einer homogenen Gruppe. Die Beobachtungen einer Variablen werden zusammengefasst im Vektor

$$x = (G,K) = (g_1, \ldots, g_{n_G}, k_1, \ldots, k_{n_K}) = (x_1, \ldots, x_{n_G+n_K}) = (x_1, \ldots, x_N).$$

3.2.1 COPA: cancer outlier profile analysis

Als eine der ersten SG-Detektionsmethoden wurde 2005 COPA im Kontext von Genexpressionsanalysen vorgeschlagen [17]. Zur Durchführung der *cancer outlier profile analysis* werden die Expressionswerte variablenweise um ihren Median zentriert und mittels *mad* (median absolute deviation bzgl. des Gesamtmedians *med*) skaliert. Das Ranking der Variablen orientiert sich in der ursprünglichen Veröffentlichung an der Größe eines gewählten Quantils der transformierten Werte der heterogenen Gruppe. So werden die Variablen beispielsweise anhand der 75-, 90- oder 95%-Quantile sortiert. Im zugehörigen **R**-Paket [41] liegt der Fokus jedoch nicht mehr auf diesem univariaten Ranking, sondern auf dem Auffinden von Variablenpaaren, die möglichst große disjunkte Mengen von Ausreißersamples in der Gruppe der Krebsproben zeigen. Die Idee dahinter ist, dass im Zusammenhang mit Krebs solche Variablenpaare (Genpaare) an bisher unbekannten Translokationen beteiligt sein könnten. Dabei wird eine Probe bzgl. einer Variable als Ausreißer angesehen, wenn der transformierte Beobachtungswert den (als „üblich“ bezeichneten) cut-off von 5 überschreitet. Aufgrund dieser Weiterentwick-

lung zu einem kombinierenden Ansatz und da beispielsweise Wu [20] bereits die Überlegenheit alternativer Methoden zeigte (z. B. ORT, 3.2.3), wird COPA in den weiteren Vergleichen der univariaten Methoden nicht berücksichtigt.

3.2.2 OS: outlier sum

Auch Tibshirani und Hastie [19] gehen für die Entwicklung ihrer *outlier sum* davon aus, dass genau eine der beiden zu vergleichenden Gruppen als homogen bzw. heterogen anzusehen ist. Zunächst werden die Beobachtungen x für jede Variable unabhängig unter Verwendung robuster Methoden standardisiert. Dazu wird (wie bei COPA, 3.2.1) nach Zentrierung mittels Median *med* mit dem *mad* (median absolute deviation bzgl. des Gesamtmedians *med*) der Variablen skaliert, sodass für die einzelnen Beobachtungen x_i gilt:

$$x_i' = (x_i - med)/mad.$$

Sei q_r das r-te Perzentil der standardisierten Werte $x' = (x_1', \ldots, x_N')$. Der Interquartilsrange *iqr* ist definiert als $q_{75} - q_{25}$ und ein x_i' wird als Ausreißer betrachtet, wenn es die Schwelle $c_{OS} = q_{75} + iqr$ überschreitet. Die Werte aller so bestimmten Ausreißer in der heterogenen Gruppe K ergeben aufaddiert die Teststatistik OS:

$$\text{OS} = \sum_{x_i' \in K, x_i' > c_{OS}} x_i', \qquad c_{OS} = q_{75} + iqr.$$

Im Unterschied zu COPA wird die Definition der Ausreißerschwelle an die Verteilung der Variablen angepasst und durch das Aufsummieren aller Ausreißerbeobachtungen der Informationsgehalt im Vergleich zum Wert eines einzelnen Quantils erhöht. Große Werte der OS-Teststatistik können durch einzelne starke Ausreißer verursacht werden, die in der Praxis meist weniger interessant sind, oder durch Ausreißergruppen.

3.2.3 ORT: outlier robust *t*-statistic

Nachdem die OS als Verbesserung von COPA vorgeschlagen wurde, motiviert Wu [20] die *outlier robust t-statistic* wiederum durch eine Verbesserung der OS: Zur Schätzung von Lage und Streuung werden bei der OS der Median *med* und das Variationsmaß *mad* (median absolute deviation wie oben) auf der Grundlage aller Beobachtungen einer Variablen berechnet. Bei ORT hingegen wird zur Zentrierung aller Beobachtungen der Median der als homogen angenommenen Gruppe verwendet. Dadurch soll auch in Fällen extrem großer Anteile von Ausreißerproben eine verzerrte Lageschätzung der homogenen Gruppe verhindert werden. Mit

ähnlicher Begründung wird die Verwendung des *mad* kritisiert, da die Abweichungen vom Gesamtmedian der Beobachtungen betrachtet werden. Stattdessen wird für ORT

$$median(\{|x_i - med_K|_{x_i \in K}, |x_i - med_G|_{x_i \in G}\})$$

als geeignetes Maß für die Variation vorgeschlagen, d. h. die Abweichungen werden vom jeweiligen Gruppenmedian $med_K = median(K)$ bzw. $med_G = median(G)$ bestimmt. Auf diese Weise soll eine Überschätzung der Variation vermieden werden, die nur auf das Vorhandensein einer Subgruppe zurückzuführen ist. Insgesamt lässt sich die Teststatistik schreiben als

$$t^* = \frac{\sum_U x_i - med_G}{median(\{|x_i - med_K|_{x_i \in K}, |x_i - med_G|_{x_i \in G}\})},$$

wobei U die Menge

$$U = \left\{x_i \in K : \quad x_i > \left(q_{75,G} + iqr_G\right)\right\}$$

der Ausreißerproben in der betrachteten Variable beschreibt. $q_{75,G}$ und iqr_G bezeichnen das 75%-Quantil bzw. den Interquartilsrange in der homogenen Gruppe G. Wu zeigte die Überlegenheit von ORT gegenüber OS bei verschiedenen Alternativen.

3.2.4 PADGE: percentile analysis for differential gene expression

Bei der *percentile analysis for differential gene expression* [21] (kurz PADGE) werden zunächst mithilfe statistischer Tests auf Lokationsunterschiede wie Students t-Test oder Wilcoxons Rangsummentest Teilmengen beider Gruppen miteinander verglichen. Dazu wird eine Menge Q von Quantilen bestimmt, beispielsweise

$$Q = \{Q_t, t = 1, \ldots, T\} = \{Q_1, Q_2, Q_3\} = \{q_{80}, q_{85}, q_{90}\},$$

wobei q_γ das γ-Quantil bezeichnet. Für die beiden zu vergleichenden Gruppen G und K definiere

$$G_t = \left\{x_i \in G : x_i > Q_{t,G}\right\},$$

K_t analog. Nach der Anwendung des gewählten Tests auf die T Paare von Teilmengen G_t und K_t werden die resultierenden p-Werte für multiple Quantile adjustiert und mit p_t bezeichnet. Zusätzlich zur Bewertung der Signifikanz wird als Maßzahl für den Lageunterschied der jeweiligen Teilmengen das Expressionratio r_t der Mengen K_t und G_t berechnet. Falls G homogen ist, und in K eine Subgruppe mit höheren Expressionswerten vorhanden ist, steigen die Expressionratios mit höheren Quantilen Q_t. Bei einem homogenen Shift zwischen den Gruppen zeigt sich idealerweise nur eine kleine Änderung. Um die untersuchten Variablen nach ihrer Relevanz zu ordnen, schlagen die Autoren einen Score vor, der sowohl den p-Wert als auch die Änderung des Lageunterschieds berücksichtigt:

$$S = \max_t \left[-\frac{r_t}{r_1} \cdot \log p_t \right],$$

Dabei ist r_1 das Expressionsratio beider Gruppen, wenn alle Beobachtungen berücksichtigt werden. Der Term r_t/r_1 beschreibt die relative Änderung des Expressionsratios vom t-ten Teilmengenvergleich zum Gesamtexpressionratio beider Gruppen.

3.2.5 PACK: profile analysis using clustering and kurtosis

Profile analysis using clustering and kurtosis [18], kurz PACK, ist ein zweistufiges Verfahren, das im ersten Schritt Variablen auswählt, bei denen es ausreichend starke Hinweise auf eine bimodale Verteilung gibt und anschließend diese Variablen entsprechend ihrer empirischen Kurtosis sortiert. Dabei hat der Anwender die Wahl zwischen einer auf- und absteigenden Sortierung. Große positive Werte treten auf, wenn eine kleinere Subgruppe sich vom Rest der Beobachtungen unterscheidet. Die Kurtosis wird hingegen negativ, wenn die Beobachtungen sich in zwei etwa gleichgroße Gruppen aufteilen, was beispielsweise bei homogenen Shifts zwischen den Gruppen auftritt. Kurtosis-Werte nahe Null treten beispielsweise bei normalverteilten Daten auf, entsprechende Variablen sollten durch den vorgeschalteten Filterschritt für die weitere Analyse nicht relevant sein.
Die Autoren schlagen insbesondere bei kleineren Fallzahlen auch die vereinfachte univariate Variante PAK vor, die auf die Vorselektion verzichtet und schlicht für alle Variablen die Kurtosis berechnet. Dementsprechend wird in dieser Arbeit die Berechnung der Kurtosis als Repräsentant für die Methode PACK verwendet. Da der Fokus in dieser Arbeit auf kleineren Subgruppen liegt, wird die absteigende Sortierung gewählt. In der Literatur werden verschiedene Schätzer für die Kurtosis verwendet, Teschendorff et al. [18] benutzen

$$Kurtosis(x) = \frac{N(N+1)\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)(N-2)(N-3)\sigma^4} - \frac{3(N-1)^2}{(N-2)(N-3)}.$$

Dabei ist $x = (x_1, \ldots, x_N)$ die Menge der insgesamt $N = n_K + n_G$ beobachteten Werte einer Variable und $\bar{x}$ und σ sind das arithmetische Mittel und die geschätzte Standardabweichung. Der gegebene Schätzer ist unverzerrt und wird häufig als Voreinstellung in gängiger Software verwendet (SAS, SPSS).

3.2.6 MinM: minimum M statistic

In der *ProtoArray Prospector* Software (Life Technologies, Carlsbad, Kalifornien, USA) wird als Teststatistik die sogenannte *minimum M statistic* (MinM) verwendet, die in Love [36] beschrieben ist. Die Verwendung dieser Methode wird mit der Sensitivität sowohl gegen homogene Unterschiede zwischen zwei Gruppen als auch gegen Subgruppen in einer der beiden Gruppen begründet. Das Vorgehen ist im Wesentlichen äquivalent zur Methodik eines *Minimum Fishers exakter Test*, die im Folgenden kurz erläutert wird: Für eine einzelne Variable (mit $n = n_K = n_G$ Samples) werden $2n$ exakte Tests nach Fisher berechnet. Dabei wird die Abhängigkeit zwischen der Gruppenzugehörigkeit jeder Beobachtung (mit Ausprägung Gesund G oder Krank K) und der Lage des beobachteten Wertes im Vergleich zu einem Schwellenwert c beurteilt. Die entsprechenden Häufigkeiten können in einer Vierfeldertafel dargestellt werden:

	Krank	Gesund	
$> c$	n_{11}	n_{12}	$n_{1\cdot}$
$\leq c$	n_{21}	n_{22}	$n_{2\cdot}$
	n_K	n_G	

.

Hier bezeichnet beispielsweise n_{11} die Anzahl der Beobachtungen in der Gruppe Krank, die über dem vorgegebenen Schwellenwert c liegen. Für den Wert für c wird nacheinander jede Beobachtung eingesetzt und der p-Wert des zugehörigen exakten Test nach Fisher bestimmt. Das anschließend bestimmte Minimum dieser p-Werte wird als p-Wert der *minimum M statistic* ausgegeben.
Die MinM-Methode wird nicht als separate Methode in die späteren Vergleiche der univariaten Methoden aufgenommen. Vielmehr wird sie hier vorgestellt, da die im Folgenden gezeigte neue Methode Fisher Sum die Idee der datenabhängigen cut-offs aufgreift.

3.3 FS: Fisher Sum

Anforderungen an die neue Methode Fisher Sum

Bei der Anwendung der bisher beschriebenen SG-Detektionsmethoden auf verschiedene reale omics-Datensätze zeigte sich, dass die Methoden generell in der Lage sind, Variablen mit dem gesuchten Expressionsmuster eines partiellen Shifts

zu erkennen. Allerdings wird je nach Methode auch solchen Variablen ein hoher Score zugewiesen, bei denen die Expressionsprofile sogenannte *nicht-krankheitsspezifische* (nks) Subgruppen zeigen: Grundsätzlich können Confoundervariablen (bekannt oder unbekannt) ebenso für erhöhte Expressionswerte in kleineren Samplegruppen verantwortlich sein, wie interessierende, biologisch relevante krankheitsspezifische Aspekte. Bei einer zufälligen Verteilung einer solchen Confoundervariablen über beide Samplegruppen wird in beiden, d. h. insbesondere auch in der Kontrollgruppe, ein SG-Muster erkennbar sein. Anhand der isolierten Betrachtung des Expressionsmusters der Gruppe Krank lässt sich nicht beurteilen, ob es sich um eine krankheitsspezifische Patientensubgruppe handelt. Erst die zusätzliche Berücksichtigung der Verteilung in der als homogen angenommenen Gruppe kann diesbezüglich Hinweise liefern. Nicht-krankheitsspezifische Subgruppen, die in beiden zu vergleichenden Gruppen auftauchen, können beispielsweise von den Methoden OS (3.2.2) oder ORT (3.2.3) fälschlicherweise als relevant bewertet werden. Um die Arbeit bei der Subgruppendetektion zu erleichtern, wird hier ein neuer Score vorgestellt, der ein **angemessenes Scoring von Variablen mit nicht-krankheitsspezifischen Subgruppen** erlaubt.
Bei der Entwicklung standen zwei weitere Punkte im Fokus, die die Subgruppengröße und die Beurteilung der Relevanz von Subgruppen betreffen. Sowohl Studien aus dem Bereich der Subgruppendetektion als auch übliche differentielle Studien zur Untersuchung homogener Shifts haben gezeigt, dass der gewöhnliche t-Test bei der Identifikation von Subgruppen hilfreich sein kann, sofern diese ausreichend groß sind. Besteht Grund zur Annahme heterogener Gruppen, ist zu beachten, dass dies im Widerspruch zur Testannahme identischer Verteilungen innerhalb der Gruppen steht und die p-Werte somit verfälscht sein können. Die neu entwickelte Methode soll gerade die **Detektion kleinerer Subgruppen bzw. solcher mit geringem Expressionsunterschied** ermöglichen.
Bei Methoden wie COPA, OS oder ORT geht der Berechnung des spezifischen Scores grundsätzlich die Zentrierung und Skalierung der Variablen mit als geeignet angesehenen Größen voraus. Dadurch kann die Abweichung der Beobachtungswerte einer möglichen Subgruppe im Kontext der Verteilung der Variablen bewertet und auch kleine absolute Änderungen aufgedeckt werden. Der möglicherweise statistischen Signifikanz eines solchen Subgruppenmusters steht die Frage der klinischen Relevanz gegenüber. Obwohl bisher keine Einigkeit über das genaue Vorgehen besteht, wird üblicherweise bei der Kandidatenauswahl aus Hochdurchsatzexperimenten nicht nur die Signifikanz sondern auch ein Effektmaß (*Fold Change*) als Filterkriterium verwendet. Auch die Erfahrung mit Anwendern aus den Lebenswissenschaften zeigt, dass häufig das Interesse an einem Kandidaten (d. h. einer Variablen) mit einem größeren absoluten Abstand der Subgruppe größer ist als bei einer Variable mit einer insgesamt sehr schmalen Verteilung. Um diese Einschätzung zu reflektieren wird bei den Standardeinstellungen

der neuen Methode explizit der Skalierungsschritt ausgelassen, sodass die **Beurteilung der Relevanz der gefundenen Subgruppen entsprechend der absoluten Abstände der Subgruppe zu den übrigen Beobachtungen** erfolgt.

Definition Fisher Sum

Die Definition der *Fisher Sum* FS [22] erfolgt am Beispiel des Vergleiches einer kranken und einer gesunden Gruppe. Bei FS handelt es sich um eine univariate Methode, ihre Berechnung erfolgt unabhängig für alle Variablen des Datensatzes. Aus Gründen der Übersichtlichkeit wird daher auf den Index des einzelnen Features verzichtet. Seien $G = \{g_1, \ldots, g_{n_G}\}$ die Werte der gesunden Gruppe G, sowie $K = \{k_1, \ldots, k_{n_K}\}$ die Beobachtungen der kranken Gruppe K für ein einzelnes Feature und x der Vektor aller Beobachtungen dieses Features:

$$x = (G,K) = (g_1, \ldots, g_{n_G}, k_1, \ldots, k_{n_K}), \qquad N = n_G + n_K.$$

Durch die Zentrierung der Werte um den Median der Werte der Gruppe G, d. h.

$$\tilde{x} = x - \mathbf{1}_{n_G+n_K} \cdot med_G = (\tilde{G}, \tilde{K}),$$

wobei $\mathbf{1}_{n_G+n_K}$ gegeben ist durch den Vektor $(1 \ldots 1)$ der Länge $n_G + n_K$, ist der Score unabhängig von der ursprüngliche Lage der Expressionswerte des Features. Die Verwendung von med_G zur Zentrierung hatte sich bereits bei ORT bewährt. Der Schwellenwert c_{FS} wird als das 90%-Quantil $q_{90,\tilde{K}}$ der Werte in $\tilde{K}$ definiert. Wie in der oben beschriebenen MinM-Methode können die beiden Merkmale *Gruppenzugehörigkeit* und *Lage zum Schwellenwert* in einer Vierfeldertafel dargestellt werden:

	Gruppe	
	krank	gesund
$> c_{FS}$	n_{11}	n_{12}
$\leq c_{FS}$	n_{21}	n_{22}

Dann berechnet sich der Score FS als (gewichtete) Summe der (zentrierten) Werte, die in die Zellen (i, j), $i, j = 1,2$, der Vierfeldertafel fallen. Mit der eingeführten Notation ist folglich

$$FS = w \sum_{\substack{\tilde{k} \in \tilde{K}, \\ \tilde{k} > c_{FS}}} \tilde{k} - \sum_{\substack{\tilde{g} \in \tilde{G}, \\ \tilde{g} > c_{FS}}} \tilde{g}. \tag{1}$$

Große Werte für FS ergeben sich, wenn ein Wert oder eine Gruppe von Werten in K einen großen absoluten Abstand zum Median der gesunden Gruppe aufweisen, während möglichst keine Beobachtung in der gesunden Gruppe den Schwellenwert c_{FS} übersteigt. Die Subtraktion des zweiten Terms stellt einen Strafterm für

hoch-regulierte Subgruppen in der gesunden Gruppe dar und bewirkt somit eine Korrektur bei nicht-krankheitsspezifischen (nks) Subgruppen. In vielen Anwendungen wird das Gewicht $w = 1$ gesetzt. Bei stark unbalancierten Gruppengrößen oder einer gewünschten stärkeren Bestrafung von nks Subgruppen kann eine Anpassung von w vorgenommen werden.

Anmerkungen und Möglichkeiten der Verallgemeinerung

Der FS-Score greift Aspekte aus den in Abschnitt 3.2 vorgestellten Methoden auf und verbindet diese mit neuen Ideen, um den eingangs definierten Anforderungen zu genügen. Die Zentrierung mittels Median der Kontrollen wurde bereits für ORT vorgeschlagen. In Anlehnung an MinM wird ein datenabhängiger cut-off zur Dichotomisierung der Daten gewählt. Die Bewertung der resultierenden Vierfeldertafel wird bei FS nicht wie bei MinM auf einen einzelnen Zellenwert gestützt, sondern auf die Summe der beitragenden Werte. Das Argument des Informationsgewinns durch Aufsummieren der extremen Werte anstelle der Betrachtung eines Einzelwertes führten schon Tibshirani und Hastie [19] beim Übergang von COPA zu OS an. Eine wesentliche Neuerung der FS ist die Korrektur für Expressionsmuster mit nks Subgruppen, die sich in der Anwendung schnell bewährt.
Die in Formel (1) angegebene Version der FS gilt für die Identifikation von Subgruppen mit erhöhten Werten. Ein Scoring der Features zur Identifikation von Subgruppen mit erniedrigten Werten ist analog möglich, indem die zentrierten Werte vor der Berechnung der FS mit (-1) multipliziert werden. Falls beide Richtungen simultan berücksichtigt werden sollen, wird jeder Variable der jeweils größere Betrag aus der Berechnung für hoch- und herunterregulierte Subgruppen zugeordnet. Dann allerdings ist auch der später vorgestellte Workflow für die Kombination der Features anzupassen. Entsprechende Details zur notwendigen Adaption sind in den jeweiligen Abschnitten (4.3.1 und 4.3.2) beschrieben.
Die standardmäßige Wahl des Schwellenwertes c_{FS} als $q_{90,\tilde{K}}$ hat sich in frühen Studien als sinnvoll erwiesen. Durch die Summierung der 10% größten Werte aus $\tilde{K}$ erreichen Variablen mit Subgruppen ab einer Größe von 10% der heterogenen Gruppe (bei gleichem Shift) bessere Scores als Variablen mit kleinerer Subgruppe. Abhängig von der minimalen als relevant erachteten Subgruppengröße kann der Parameter c_{FS} variiert werden. Bei höheren Fallzahlen wäre der Fokus auf einen kleineren Anteil der Beobachtungen denkbar (beispielsweise die höchsten 5%, d. h. $q_{95,\tilde{K}}$ für $n > 100$).
Während die Wahl von $q_{90,\tilde{K}}$ geringeren Fokus auf Subgruppen mit weniger als 10% Anteil an K legt, werden Subgruppen mit mehr als 10% eher bevorzugt: Dadurch dass nur die höchsten 10% der Werte in die Berechnung von FS eingehen, wird die Effektgröße bei Subgruppen mit mehr als 10% überschätzt, da nur die extremsten Werte berücksichtigt werden. Dies gilt auch im Fall homogener Shifts

zwischen den beiden Gruppen K und G (d. h. bei einem SG-Anteil von 1), die somit ebenfalls einen guten FS-Score zugewiesen bekommen können. Diese Eigenschaft wird nicht als nachteilig angesehen, und FS teilt sie mit vielen der oben beschriebenen Methoden. Sollte in einer Studie explizit die Untersuchung von SG-anzeigenden Variablen im Fokus stehen, die FS-gerankte Liste weist aber eine große Anzahl von Variablen mit homogenen Shifts auf, so ist ein pragmatischer Ansatz das Filtern der gerankten Liste gemäß des p-Wertes eines t-Tests. Durch Entfernen der im t-Test signifikanten Variablen wird die Liste mit Variablen des interessierenden Musters angereichert. Zur Visualisierung kann auch ein Scatterplot der $-\log_{10}$-transformierten p-Werte gegen den FS-Score erstellt werden.

An dieser Stelle wird bewusst auf die Bewertung der Signifikanz der gefundenen SG-Expressionsmuster verzichtet. Die Subgruppenanalyse sollte als explorative Verfahren und Ergänzung zu gewöhnlichen differentiellen Studien gesehen werden. In den seltensten Fällen wird eine Studie ausschließlich mit einem SG-Detektionstest ausgewertet und dann würde selbst bei einer Korrektur für multiples Testen der dafür berechneten p-Werte der Fehler erster Art insgesamt (durch die vorgehende differentielle Studie) nicht ausreichend kontrolliert werden. Daher wird hier die Darstellung als rein explorativ Verfahren bevorzugt. Grundsätzlich lässt sich jedoch durch die üblichen Simulationen unter der Nullhypothese oder durch wiederholte Permutationen der Klassenlabel eine empirische Verteilung für FS bestimmen, aus der wiederum ein p-Wert für ein einzelnes Feature abgelesen werden kann.
In der beschriebenen Zielsetzung dieser Arbeit wurde überdies bereits erklärt, dass die univariate Bewertung der Variablen nicht impliziert, dass die Inferenz über mögliche Subgruppen ausschließlich auf den einzeln berechneten Scores basiert. Stattdessen werden mithilfe der Scores die informationstragendsten Variablen ausgewählt, um die Performanz nachgeschalteter multivariater Methoden zu verbessern. Die Evidenz einer potentiellen Subgruppe wird dadurch gesteigert, dass verschiedene Variablen konsistent auf diese Samplegruppe hinweisen, auch wenn sie einzeln nicht notwendig Signifikanz zeigen.

4 Multivariate Verfahren zur Identifikation von Patientensubgruppen

Nachdem in Kapitel 3 univariate Methoden für die Identifikation von subgruppenanzeigenden Variablen vorgestellt wurden, behandelt das folgende Kapitel multivariate Ansätze zur expliziten Identifikation von Patientensubgruppen. Der Abschnitt 4.1 gibt eine Übersicht über die bisherige Literatur. Dabei wird das später als Referenzmethode für den neuen Workflow verwendete Biclustern (BC) ausführlich dargestellt (4.2). Es handelt sich um ein vor allem in Genexpressionsstudien häufig verwendetes Verfahren, dessen Ziel die Identifikation von Samplegruppen ist, deren Expression sich nur in Teilmengen von Features ähnelt.
In Abschnitt 4.3 wird der neue FSx-Workflow mit seinen beiden Varianten FSOL und FSJ vorgestellt. Das Verfahren lässt sich in drei Schritte unterteilen: die Selektion interessanter Variablen, ihre Gruppierung gemäß der angezeigten Subgruppe und die Nominierung von Samplesubgruppen aus den gebildeten Variablengruppen. Die Endungen OL bzw. *J* des FSx-Workflows bezeichnen das jeweils verwendete Ähnlichkeitsmaß im mittleren Schritt. Nach der Beschreibung dieser beiden Maße in den Abschnitten 4.3.1 und 4.3.2 folgt in 4.3.3 eine detaillierte Darstellung der einzelnen Schritte und der Workflowparameter.
Im letzten Abschnitt 4.4 dieses Kapitels wird die Möglichkeit der Kombination der univariaten FS-basierten Variablenselektion und des Biclusterns (FSBC) vorgestellt. Die Performanz dieser vier Verfahren wird in den späteren Kapiteln anhand von Simulationsstudien (Kapitel 5) und realen Daten (Kapitel 6) verglichen.

4.1 Literaturübersicht

Mit der steigenden Anzahl von Genexpressionsstudien wuchs der Bedarf an spezifischen Methoden, um wertvolle Informationen aus den Daten zu gewinnen. In der Anwendung zu Studien an heterogenen Krankheiten wie Krebs zeigte sich häufig der Nachteil einer PCA oder des hierarchischen Clusterns: Da die Ähnlichkeit von Samples über die Gesamtheit der gemessenen Variablen beurteilt wird, werden auch starke Ähnlichkeiten vernachlässigt (bzw. übersehen), wenn sie nur in kleinen Variablengruppen auftreten. Gleiches gilt umgekehrt für die Beurteilung der Ähnlichkeit von Variablen in einer kleinen Samplegruppe.
Dass durchaus Einigkeit darüber besteht, dass eine übliche PCA nicht geeignet ist, um kleine Samplesubgruppen in Hochdurchsatzexperimenten zu identifizieren, zeigt sich an der Vielzahl vorgeschlagener Variationen des gewöhnlichen Verfahrens zur Lösung dieses Problems. Ebenso wie bei den univariaten Ansätzen ist auch hier eine erschöpfende Darstellung nicht möglich. Da für die komplexeren multivariaten Verfahren eine kompakte Darstellung der Berechnungen wie bei den

univariaten Scores häufig nicht möglich ist, wird an dieser Stelle verstärkt auf die Originalliteratur verwiesen.
Neben einer Reihe projektionsbasierter Ansätze zur Identifikation unbekannter Patientensubgruppen wurden auch Varianten des *multidimensional scaling* (MDS) vorgeschlagen. Die ISIS-Methode (*Identifying splits with clear separation*) von von Heydebreck et al. [42] basiert auf dem ursprünglich von Friedman und Tukey [43] und Huber [44] vorgestellten *projection pursuit*. Dabei misst ein sogenannter *diagonal linear discriminant (DLD) score* wie deutlich sich die beiden Samplemengen einer Bipartition der Gesamtsamplemenge eines Microarrayexperiments anhand der Expressionswerte einer geeigneten Teilmenge von Variablen trennen lassen. Um die Diskriminanzgüte der Variablen zu beurteilen, wird die t-Statistik zum Vergleich der Projektionen der Beobachtungen aus den zwei Samplegruppen auf die zuvor berechnete Diskriminanzachse herangezogen. Somit liefert ISIS einen objektiv messbaren Score zur Beurteilung der Datenstruktur.
Im Gegensatz dazu soll CUMBIA ([45], *computational unsupervised method for bivisualization analysis*) die rein visuelle Identifikation kleiner Subgruppen ermöglichen. Dabei liegt besonderes Augenmerk auf der Möglichkeit, neben der Detektion auffälliger Samplegruppen auch die beteiligten Variablen zu erkennen. Im Gegensatz zu üblichen MDS-Methoden wird dazu eine gemeinsame niedrigdimensionale Darstellung von Samples und Variablen berechnet. In der zugehörigen Arbeit findet sich außerdem eine Auflistung weiterer vorgeschlagener Methoden aus dem Feld der Subgruppendetektion. Dort findet auch das Biclustern Erwähnung, das im späteren Teil der vorliegenden Arbeit unter anderem als Referenzmethode dient.

Die Idee des Biclusterns wurde bereits 1972 von Hartigan [46] publiziert, der Begriff wurde aber erst im Laufe der 1990er Jahre geprägt. Neben der mangelnden Sensitivität üblicher Clustermethoden zur Erkennung kleiner Subgruppen, die sich in einer geringen Anzahl von Variablen zeigen, gibt es ein weiteres Argument, warum speziell das hierarchische Clustern nicht geeignet ist, um die biologischen Zusammenhänge und betroffenen Prozesse abzubilden. Dass Gene häufig in mehr als einem solcher Prozesse involviert sind, kann in der Zuordnung zu jeweils genau einem Cluster bei der Partitionierung nicht berücksichtigt werden. Um mindestens in einem dieser Aspekte eine Verbesserung zu erzielen, wurde bis heute eine Vielzahl verschiedener Bicluster-Ansätze publiziert. Pontes et al. [47] listen allein 47 davon in ihrem Review *Biclustering on expression data* auf.
Da für die Analyse von Expressionsdaten bis heute gern der sogenannte Plaid-Algorithmus verwendet wird (z. B. in der Arbeit von Henriques und Madeira [48]) und auch die biologische Relevanz der Ergebnisse wiederholt gezeigt werden konnte (z. B. von Oghabian et al. [49]), wird er als Referenz für den neu entwickelten Workflow verwendet.

Es sei angemerkt, dass der Plaid-Algorithmus wie viele der alternativen Biclustermethoden ebenfalls nicht deterministisch ist und die Ergebnisse verschiedener Läufe sich durchaus stark unterscheiden können. In den letzten Jahren wurden verschiedene *Ensemblemethoden* vorgeschlagen, um die variierenden Ergebnisse verschiedener Läufe zu kondensieren und für eine weitere Analyse verwertbar zu machen. Genannt seien hier beispielsweise Ansätze von Hanczar und Nadif [50] und De Smet und Marchal [51], sowie das **R**-Paket `superbiclust` [52]. Da bisher keine umfassende Studie zum Performanzvergleich verschiedener Kombinationen von Biclusteralgorithmen und Ensemblemethoden verfügbar ist, und ihre zusätzliche Durchführung für diese Arbeit zu umfangreich ist, wird die Anwendung von Ensemblemethoden hier nicht weiter verfolgt. Allerdings wird der Jaccardindex (siehe oben) als Maß zur Beurteilung der Ähnlichkeit von Sample- oder Variablengruppen aus dem `superbiclust`-Paket für diese Arbeit übernommen.

4.2 Biclustern unter Verwendung des Plaid-Modells

Das Plaid-Modell wurde 2002 von Lazzeroni und Owen [23] vorgestellt. Gegeben sei eine Expressionsmatrix (oder die Datenmatrix einer vergleichbaren omics-Technologie) Y_{ij}, $i = 1,\ldots,p$, $j = 1,\ldots,n$, wobei der Index i das Gen (die Variable) und der Index j die Probennummer identifiziert. Im Plaid-Modell wird Y_{ij} als Summe sogenannter *Layer* aufgefasst, die sich als lineare Modelle darstellen lassen. Für ein Modell mit insgesamt L Layern wird ein einzelner Expressionswert y_{ij} von Variable i in Sample j modelliert als

$$y_{ij} = \theta_{ij0} + \sum_{l=1}^{L} \theta_{ijl}\rho_{il}\kappa_{jl} + \varepsilon_{ij},$$

wobei die Indikatoren ρ_{il} und κ_{jl} angeben, ob Variable i bzw. Probe j in Layer l, $l = 1,\ldots,L$ enthalten sind. Sowohl die Variablen als auch die Samples können in keinem, genau einem oder mehreren Layern enthalten sein, sodass es für jeden Layer $(2^n-1)(2^p-1)$ mögliche Kombinationen (beteiligter Variablen und Samples) gibt.
Aufgrund dieser hohen Anzahl von Möglichkeiten wird zur Optimierung auf einen numerischen Ansatz zurückgegriffen, bei dem in mehreren Zyklen nacheinander die Parameter θ, ρ und κ aktualisiert werden, um bei bereits bestimmten $l-1$ Layern das Modell für Layer l anzupassen. Die Lösung ist abhängig von den gewählten Startwerten für ρ und κ.
Sukzessive werden Layer zu einem Hintergrundlayer hinzufügt, solange die Summe der quadradierten Residuen „ausreichend“ reduziert wird. Dazu wird mithilfe des Maßes

$$\sigma_l^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} \rho_{il}\kappa_{jl}\theta_{ijl}^2$$

die Relevanz für jeden potentiellen Layer berechnet. Ein Layer wird nur dann akzeptiert, wenn seine Relevanz signifikant höher ist als unter H_0 (kein Effekt, Rauschen). Der *greedy* Algorithmus fügt solange Layer hinzu, bis das Relevanzmaß einen durch Permutation bestimmten Schwellenwert nicht mehr überschreitet bzw. wenn die vorgegebene maximale Anzahl L von Layern erreicht ist. Jeder vom Algorithmus ausgewählte Layer repräsentiert ein Bicluster.
Alle Bicluster-Berechnungen in dieser Arbeit beruhen auf der Verwendung des **R**-Paketes `biclust` [53]. Der implementierte Algorithmus stellt eine von Turner et al. [54] präsentierte Verbesserung der Originalversion von Lazzeroni und Owen [23] dar. Er ist weiterhin nicht deterministisch und mitunter sehr sensitiv gegenüber der Wahl der Startwerte (siehe auch Anwendungskapitel).

4.3 FSx-Workflow zur Identifikation von Patientensubgruppen

Aus einem gegebenen Datensatz zum Vergleich einer kranken und einer gesunden Gruppe sollen Informationen über möglicherweise vorhandene Patientensubgruppen gewonnen werden. Gesucht wird o. B. d. A. zunächst nach Variablen mit erhöhten Werten in der Subgruppe (SG). Zur weiteren Charakterisierung der gegebenenfalls gefundenen Subgruppe ist auch die Variablenmenge von Interesse, in denen die Samplegruppe auffällige Werte zeigt. Der zu diesem Zweck entwickelte Workflow FSx gliedert sich in drei Schritte:

Schritt 1 Vorauswahl von Variablen mit höchsten FS-Scores
Schritt 2 Gruppierung dieser Variablen gemäß der angezeigten Sample-Subgruppe
Schritt 3 Nominierung von Samples für eine potentielle SG

Als erstes wird eine vorgegebene Anzahl T von Features aus dem Datensatz selektiert, deren Expressionsmuster am stärksten auf eine Samplesubgruppe hinweisen. Die Auswahl basiert auf dem in Abschnitt 3.3 vorgestellten univariaten Score *Fisher Sum*. Anschließend werden diese Variablen unter Verwendung eines geeigneten Ähnlichkeitsmaßes gemäß der angezeigten SG gruppiert. Eine hohe Ähnlichkeit besteht, wenn die Mengen der Samples mit den höchsten Expressionswerten ähnlich sind. Diese Ähnlichkeitsdefinition schließt auch die Ähnlichkeit der Samplerangfolge über die Topränge hinaus ein. Bei Microarraydaten in Genexpressionsstudien beobachtet man z. B. bei *probe sets*, die mit demselben Gen annotiert sind, häufig starke lineare Zusammenhänge.
Die Abschnitte 4.3.1 und 4.3.2 stellen zwei Optionen zur in Schritt 2 benötigten Beurteilung der Ähnlichkeit von Variablen bezüglich der von ihnen angezeigten Subgruppe vor. Anschließend werden die einzelnen FSx-Schritte in Abschnitt 4.3.3 detailliert dargestellt. Die erste Version (FSOL) des Workflows wurde bereits in Ahrens et al. [55] vorgestellt.

4.3.1 FSOL: Variablengruppierung basierend auf Ordered List

Der *Ordered-List*-Algorithmus [56] wurde ursprünglich für den Vergleich zweier Ergebnislisten aus Genexpressionsstudien entwickelt. Die Implementierung im **R**-Paket `OrderedList` [57, 58] ermöglicht neben der eigentlichen Berechnung auch eine grafische Darstellung der Ergebnisse. Die Methode wird zunächst im ursprünglichen Kontext beschrieben, bevor der Transfer in die Anwendung zur Subgruppendetektion erklärt wird. Für die zentralen Berechnungen wird die Funktion `compareLists` verwendet.
Gegeben seien zwei geordnete Ergebnislisten aus unabhängigen Microarrayexperimenten, die im Rahmen eines Zwei-Gruppen-Vergleiches (z. B. krank gegen gesund) die gleiche Fragestellung untersuchen. Ordered List (OL) erlaubt in dieser Situation die Beurteilung der Konsistenz der Listen um signifikante Ergebnisse oder auch Trends und Hinweise aus den einzelnen Studien zu bestätigen.
Die folgende Darstellung der Methode beschränkt sich auf den Fall, dass beide Listen eine identische Menge von Features enthalten, diese allerdings in unterschiedlicher Ordnung. Für jede Liste wird diese Ordnung beispielsweise anhand eines p-Wertes bestimmt, der die Signifikanz des Lageunterschieds zwischen den beiden Gruppen bewertet. Zusätzlich kann ein Effektmaß wie der *fold change* berücksichtigt werden, dann belegen die oberen Ränge Variablen mit hochregulierten Werten in der kranken Gruppe und die untersten Ränge die Variablen mit erniedrigten Werten. Zwei Ergebnislisten werden als ähnlich betrachtet, wenn in den Toprängen beider Listen eine ähnliche Featuremenge zu finden ist.
Der erste Schritt zur Berechnung des OL-basierten Ähnlichkeitsmaßes ist die sukzessive Bestimmung der Anzahl übereinstimmender Features unter den Toprängen $r = 1, \ldots, max.rk$, d. h. die Größe O_r des Schnitts der jeweils ersten r Ränge (vgl. auch Tabelle 2, Originalanwendung). Diese Informationen werden im sogenannten *weighted overlap score*

$$wos = w_\alpha \sum_r O_r, \qquad w_\alpha = \exp(-\alpha r),$$

zusammengefasst. Dabei werden den Randbereichen der Listen höheren Gewichte zugeordnet, sodass der Wert des Scores von den Variablen mit größtem beobachteten Effekt dominiert wird. Die Anzahl *max.rk* der berücksichtigten Ränge wird im Wesentlichen durch die Wahl des Parameters α bestimmt. In der verfügbaren Implementierung kann außerdem das minimale zu berücksichtigende Gewicht `min.weight` eingestellt werden (Default 10^{-5}).
Bei Vorliegen der ursprünglichen Expressionsdaten bietet das Paket die Möglichkeit einer Optimierung von α. Liegen allerdings wie in der später interessierenden Anwendung nur zwei geordnete Resultatlisten zum Vergleich vor, ist α bzw. *max.rk* vom Anwender unter Berücksichtigung der Listenlänge und der Annahmen über die Größe einer Subgruppe zu wählen. Die Signifikanz des beobachteten

Originalanwendung

r	Liste 1	Liste 2	O_r
1	Feature A	Feature Z	0
2	Feature E	Feature A	1
3	Feature C	Feature C	2
4	Feature F	Feature H	2
...	...	...	...

$\rightarrow$

FSOL

r	Feature 1	Feature 2	O_r
1	Sample a	Sample z	0
2	Sample e	Sample a	1
3	Sample c	Sample c	2
4	Sample f	Sample h	2
...	...	...	...

Tabelle 2: Übertragung des Ordered-List-Ähnlichkeitsmaßes zur Anwendung in FSOL: Statt des Vergleichs zweier Genexpressionslisten bzgl. der relevantesten Features werden in FSOL zwei Features bzgl. der jeweils angezeigten Subgruppe verglichen. Für die obersten Ränge wird der Overlap O_r bestimmt, d. h. die Anzahl der Objekte im Schnitt der Mengen auf den Rängen 1 bis r.

Ähnlichkeitsscores kann anhand eines empirischen p-Wertes p_{OL} beurteilt werden. Dazu wird eine der Listen wiederholt permutiert, um die empirische Scoreverteilung unter der Nullhypothese zu schätzen. Der Anteil von Permutationen, der zu einem Ähnlichkeitsscore führt, der den für den vorliegenden Vergleich übersteigt, wird als p_{OL} ausgegeben. Verallgemeinerungen der Ordered-List-Methode sind nachzulesen in [56] oder [58].

Das in OL verwendete Ähnlichkeitsmaß lässt sich leicht auf die Aufgabe der Ähnlichkeitsbewertung angezeigter Subgruppen übertragen, siehe Tabelle 2. Anstatt in zwei Listen die Features auf den Toprängen zu vergleichen, werden für FSOL zwei Features bezüglich ihrer Samplemengen mit den höchsten beobachteten Werten verglichen. Für die T FS-selektierten Features wird die Ähnlichkeitsstruktur allgemein in einer Matrix $D = \left(d_{i,j}\right)_{i,j=1,\ldots,T}$, zusammengefasst, bei der der Eintrag (i, j) der p-Wert p_{OL} zum Vergleich der Features mit FS-Rängen i bzw. j ist. Für den FSOL-Workflow ist also $d_{i,j} = p_{OLi,j}$. Da kleine Werte von p_{OL} für die Ähnlichkeit der verglichenen Objekte sprechen, handelt es sich streng genommen nicht um ein Ähnlichkeitsmaß, sondern um ein Distanzmaß, das auf der Signifikanz des Ähnlichkeitsmaßes *wos* beruht. Der Einfachheit halber wird der allgemeine Term „Ähnlichkeitsmaß“ beibehalten. Wenn gewünscht, lässt sich durch Verwendung der Größe $1 - p_{OL}$ ein Ähnlichkeitsmaß ableiten.
Zu beachten ist die stochastische Natur des p-Wertes und die mögliche resultierende Variation der Ergebnisse bei zu klein gewählter Anzahl n_{perm} von Permutationen. Da bei $T = 50$ vorselektierten Variablen $\binom{T}{2} = \binom{50}{2} = 1\,225$ Vergleiche zu bestimmen sind, wird n_{perm} in dieser Arbeit aus Laufzeitgründen standardmäßig auf 1 000 gesetzt. Durch die hohe Anzahl durchgeführter Tests kommt es zur Inflation des Fehlers 1. Art und die berechneten p-Werte sollten als Scores betrachtet

werden. Der ohnehin explorative Charakter der SG-Detektion wurde in Kapitel 2 bereits erläutert.
Die Implementierung der `compareLists`-Funktion bietet beispielsweise mit dem Parameter `two.sided` die Möglichkeit weiterer Anpassungen: So können zwei Variablen auch als ähnlich definiert werden, wenn sie gegenläufige Regulationen in den Samplegruppe zeigen, d. h. einige Samples zeigen in Variable A eine Hochregulation, während sie in Variable B erniedrigte Werte aufweisen. Dies ist für die Defaultanwendung von FSx allerdings nicht relevant, weil zunächst nach Variablen mit hochregulierten Subgruppen gesucht wird.

4.3.2 FSJ: Variablengruppierung basierend auf dem Jaccardindex

Als intuitivere, ressourcen-schonende und deterministische Alternative zur Ordered-List-Methode wird nun ein Ähnlichkeitsmaß auf Basis des Jaccardindexes J vorgestellt. Dabei wird allgemein das Verhältnis der Elementanzahlen aus Schnitt und Vereinigung zweier zu vergleichenden Mengen M_1 und M_2 betrachtet:

$$J = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|}.$$

Es gilt $J = 1$ im Falle exakter Übereinstimmung der Mengen und $J = 0$, falls die Mengen disjunkt sind.
Im Schritt der Variablengruppierung bei FSx werden zwei Variablen im Sinne des Jaccardindexes als ähnlich erachtet, wenn der Index zum Vergleich der Samplemengen auf den Toprängen einen Schwellenwert t_J erreicht oder überschreitet. Wie bei der OL-basierten Methode muss dazu die Anzahl $max.rk$ der zu vergleichenden Ränge festgelegt werden. Es bezeichne $V_{1:max.rk}$ die Menge der Samples mit den $max.rk$ höchsten Expressionswerten in der Variablen V. Mit den Defaultwerten $t_J = 0.3$ und $max.rk = 10$ sind zwei Variablen A und B ähnlich im Sinne des FSJ-Ähnlichkeitsmaßes, wenn (mindestens) fünf gleiche Samples unter den jeweiligen Toprängen $A_{1:10}$ und $B_{1:10}$ zu finden sind:

$$J(A_{1:10}, B_{1:10}) = \frac{A_{1:10} \cap B_{1:10}}{A_{1:10} \cup B_{1:10}} = \frac{5}{15} = \frac{1}{3} \geq 0.3 = t_J.$$

Die Parameter $max.rk$ und t_J können abhängig von z. B. der Gruppengröße und der erwarteten bzw. interessierenden Subgruppengröße im Experiment angepasst werden. Analog zum FSOL-Ansatz werden die Ergebnisse der paarweisen Vergleiche der top-50-FS-Variablen in einer Matrix D zusammengefasst.

Wie schon beim OL-basierten Gruppieren der Variablen wird standardmäßig nach hoch-regulierten Subgruppen gesucht. Ist die gleichzeitige Berücksichtigung von Hoch- und Runterregulationen gewünscht, so muss bei der Auswahl der Variablen in Schritt 1 festgehalten werden, für welche Richtung der höhere FS-Score erzielt wurde. Bei Hinweisen auf eine herunterregulierte Subgruppe bezeichnen die mittels Jaccardindex zu vergleichenden Topränge für diese Variable dann die *max.rk* Ränge mit den niedrigsten Werten.

An dieser Stelle lassen sich Parallelen zwischen den vorgeschlagenen FSx-Ähnlichkeitsmaßen und den Ansätzen aus der *Enrichmentanalyse* von differentiellen omics-Analysen erkennen. Der wesentliche Unterschied zwischen den beiden populären Verfahren *GSEA*, d. h. der *gene set enrichment analysis*, und dem Verfahren basierend auf Fishers exaktem Test ist die Berücksichtigung der Variablenrangfolge im Gegensatz zu einer dichotomisierten Betrachtung (Signifikanz ja/nein). Ähnlich verhalten sich der *wos* bei OL und die binäre Beurteilung bei *J* zueinander.
Obwohl für die klassische Enrichmentanalyse bereits die mangelhafte Power des Fisher-Test-Ansatzes gezeigt wurde (z. B. in [59]), werden hier beide Ansätze zunächst parallel verfolgt und in Simulationen und anhand von realen Daten verglichen. Durch die deutlich geringere Komplexität bei der Beurteilung kleiner Samplemengen in der SG-Detektion im Gegensatz zur Ordnung von oft tausenden Variablen in omics-Experimenten in der klassischen Enrichmentanalyse könnte sich die einfache Methode hier beweisen.

In der Entwicklungsphase der Jaccardindex-basierten Ähnlichkeit wurde zudem ein weiterer Ansatz untersucht, der die Vorteile der simplen Berechnung des Jaccardindex mit den Ranginformationen der OL verbinden sollte: Dabei wurde der Jaccardindex sukzessive für die top-r-Ränge, $r = 3, \ldots, max.rk$, berechnet und der maximal erzielte Wert zur abschließenden Ähnlichkeitsbeurteilung mit einem Schwellenwert t_J^* verglichen. In vergleichenden Simulationsstudien zeigten sich dadurch aber keine wesentlichen Verbesserungen, sodass für die FSJ-Methode die oben gezeigte Definition des Ähnlichkeitsmaßes verwendet wird.

4.3.3 Details des FSx-Workflows

Nach der Bereitstellung der verwendeten Methoden werden in diesem Abschnitt die einzelnen Schritte des FSx-Workflows und die zugehörigen Parameter ausführlich erläutert. Abbildung 2 zeigt ein Schema des Workflows und Tabelle 3 listet die wichtigsten Parameter und ihre Standardeinstellungen auf.

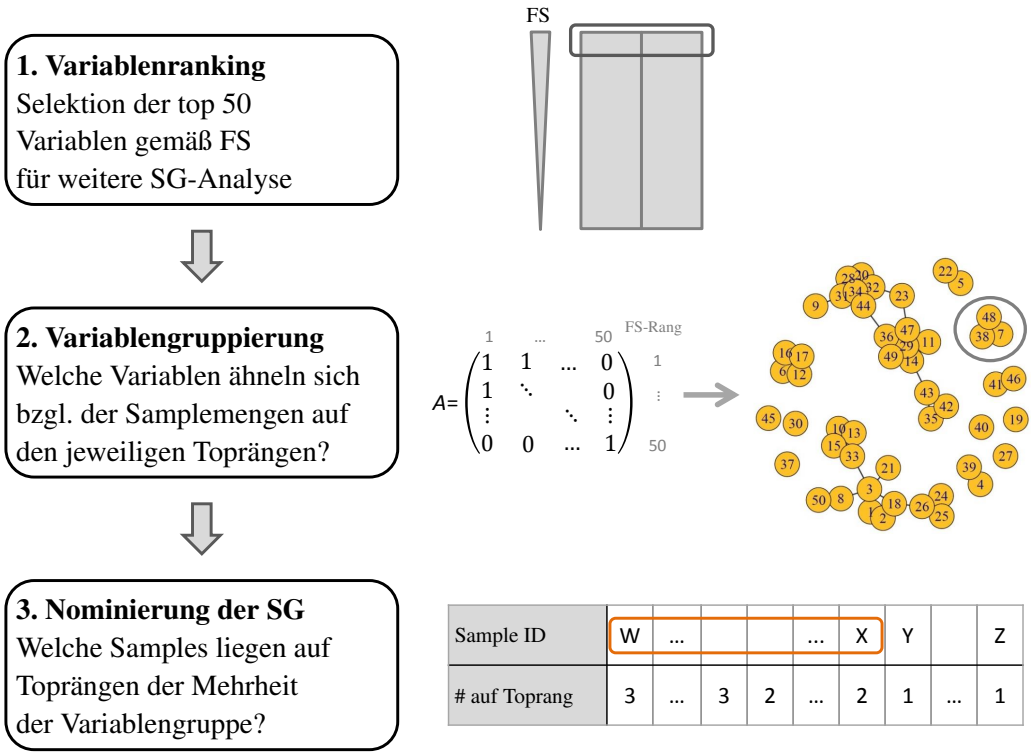


Abbildung 2: Schematische Darstellung der drei Schritte des FSx-Workflows. *Schritt 1*: Die potentiell subgruppenanzeigenden Variablen des Datensatzes werden mithilfe eines univariaten Scores identifiziert. *Schritt 2*: Mit einem geeigneten Ähnlichkeitsmaß werden zunächst die paarweisen Ähnlichkeiten der Variablen bestimmt. Diese werden mit einem Schwellenwert verglichen, um die binäre Matrix A zu erhalten, die sich als ungerichteter Graph darstellen lässt. *Schritt 3*: Für die so bestimmten Variablengruppen werden die Samples als potentielle Subgruppe nominiert, die konsistent in den Variablen einer Gruppe die höchsten Expressionswerte zeigen.

Schritt 1. Vorauswahl subgruppenanzeigender Variablen

Als erstes wird jede Variable des Datensatzes mithilfe des ausgewählten univariaten Scores hinsichtlich eines möglichen Subgruppenmusters bewertet. In umfassenden Vergleichen verschiedener Optionen (siehe 5.2 und 6.1) hat sich Fisher Sum (FS) als besonders geeignet gezeigt. Die Standardeinstellung für die Anzahl T der in den folgenden Schritten weiter betrachteten Variablen wird auf $T = 50$ festgelegt. Sie kann jedoch abhängig von Annahmen über die Subgruppenstruktur in den Daten bzgl. der Anzahl der Subgruppen und jeweils beteiligter Variablen variiert werden. Wird beispielsweise eine größere Anzahl von Subgruppen in der heterogenen Gruppe vermutet, so kann T entsprechend erhöht werden. Dies ist besonders von Vorteil, wenn eine der Subgruppen von vielen Variablen sehr deutlich angezeigt wird, d. h. wenn ein Großteil der T selektierten Variablen auf die gleiche Patientengruppe hinweist und die vorhandenen Informationen über weitere Subgruppen nicht ausgeschöpft werden. Zeigt andererseits die empirische Verteilung der FS-Scores, dass nur eine kleine Anzahl von Variablen Hinweise auf Subgruppen zu liefern scheint, so können auch kleinere Werte für T in Betracht gezogen werden.

Schritt 2. Gruppierung der selektierten Variablen gemäß der angezeigten Samplesubgruppe

Im zweiten Schritt wird für jedes der $\binom{T}{2}$ möglichen Paare der in Schritt 1 selektierten T Variablen das gewählte Ähnlichkeitmaß berechnet. Die Informationen werden in einer Matrix D der Größe $T \times T$ zusammengefasst, sodass an Position (i, j) der entsprechende Wert für die Variablen mit FS-Rängen i bzw. j steht. Die Details zur Berechnung gemäß des Ordered-List-Ähnlichkeitsmaßes bzw. mithilfe des Jaccardindex wurden oben bereits erläutert.
An dieser Stelle des Workflows ist zwischen der (manuellen) Auswertung eines einzelnen Anwendungsdatensatzes und dem automatisierten Auswertungsverfahren, wie es später für die Simulationen benötigt wird, zu unterscheiden. Letzteres erfordert die Festlegung eines fixen Schwellenwertes um die Variablengruppierung durchzuführen. Dann wird aus der Matrix D eine binäre Matrix A erzeugt, die angibt, ob die Ähnlichkeit der Variablen als hinreichend groß erachtet wird.
Für die Ähnlichkeit gemäß Ordered List bedeutet das $a_{i,j} = 1$, genau dann, wenn $p_{OL} < t_{OL}$, 0 sonst, und analog für das Jaccardindex-basierte Maß $a_{i,j} = 1$, falls $J \geq t_J$. Fasst man die resultierende Matrix A als Adjazenzmatrix auf, beschreibt sie einen ungerichteten Graphen mit insgesamt T Knoten, wobei zwischen einem Knotenpaar genau dann eine Kante existiert, wenn die repräsentierten Variablen eine ähnliche Subgruppe anzeigen. Für die Beschreibung des Graphen nützlich sind die Begriffe der (Zusammenhangs-)komponente und Clique. Eine Kompo-

FSx-Workflow: Wichtige Parameter und ihre Standardwerte		
Schritt 1	**Vorauswahl SG-anzeigender Variablen**	
	univariater Score	FS
	Featureanzahl für Gruppierung in Schritt 2	$T = 50$
Schritt 2	**Variablengruppierung**	
	Ähnlichkeit basiert auf einer/beiden Gruppen (`heterOnly`=TRUE: nur heterogene Gruppe)	beiden
	FSOL-spezifische Parameter	
	Anzahl Topränge im Vergleich	$max.rk = 10$
	Anzahl Permutationen zur p_{OL}-Schätzung	$n_{perm} = 1\,000$
	cut-off für p_{OL} (Split des Graphen)	$t_{OL} = 0.01$
	FSJ-spezifische Parameter	
	Anzahl Topränge im Vergleich	$max.rk = 10$
	cut-off für J (Split des Graphen)	$t_J = 0.30$
Schritt 3	**Bestimmung einer potentiellen Subgruppe**	
	Scoring der Variablengruppen	
	Splitten in	Komponenten
	Minimale Variablenanzahl pro Gruppe	$\min_G = 1$
	Scoring der Variablengruppen	$med^{FS}_{G_r}$
	Nominierung einer Samplesubgruppe	
	Minimaler Anteil Auftreten in Toprängen	$r_{min} > 0.5$

Tabelle 3: Übersicht der relevanten Parameter in den einzelnen FSx-Schritten und ihre Defaulteinstellungen. Siehe Abschnitt 4.3.3 für nähere Beschreibungen der Parameter und alternative Einstellungen.

nente ist dabei eine Knotenmenge, d. h. eine Menge von Variablen, in der zwischen jedem Knotenpaar ein Pfad innerhalb der gegebenen Menge existiert. Bei einer Clique ist darüber hinaus jedes Knotenpaar direkt miteinander verbunden. Man spricht von einer maximalen Clique, wenn es keinen weiteren Knoten im Graphen gibt, der durch je eine Kante mit allen anderen Knoten der Menge verbunden ist. Das **R**-Paket `igraph` [60] erlaubt die Visualisierung und die im Folgenden beschriebene Analyse des durch A dargestellten Graphen.
Für die automatisierte Auswertung werden standardmäßig die Schwellenwerte $t_{OL} = 0.01$ bzw. $t_J = 0.3$ verwendet. Bei der explorativen Analyse von Einzeldatensätzen kann es ratsam sein, die Aufteilung in Variablengruppen flexibler zu gestalten. Dazu wird die Betrachtung von Dendrogrammen oder Heatmaps empfohlen, um die Informationen in D zu visualisieren und geeignete Variablengruppen auszuwählen. Entsprechende Beispiele finden sich im Anwendungskapitel. Ebenso können die aus unterschiedlichen Schwellenwerten resultierenden Graphen verglichen werden.

Schritt 3. Identifikation und Scoring der Samplesubgruppen

Nach der Gruppierung der Variablen gemäß der angezeigten Subgruppen dient der letzte Schritt des FSx-Workflows der Identifizierung, also der expliziten Benennung, eben dieser Samplegruppen. Dazu werden (standardmäßig) die Zusammenhangskomponenten herangezogen, die sich durch die Dichotomisierung der Ähnlichkeit in Schritt 2 bei Anwendung des jeweiligen cut-offs ergeben. Für jede der $v = |G|$ beteiligten Variablen einer Gruppe G werden die Samples notiert, die die $max.rk$ höchsten Beobachtungswerte zeigen. Die Häufigkeit des Auftretens jedes dieser Samples unter den Toprängen wird in einer Tabelle zusammengefasst (siehe Tab. 4).
Zur Nominierung einer Subgruppe wird ein Mindestanteil r_{min} von Variablen jeder Variablengruppe festgelegt, in denen ein Sample auf den Toprängen auftauchen muss, um für die jeweilige Subgruppe ausgewählt zu werden. Der FSx-Standardwert liegt bei $r_{min} = 0.5$, sodass ein Sample als nominiert gilt, wenn es in mindestens der Hälfte der Variablen auf einem der $max.rk$ Topränge liegt. Die konservativste Möglichkeit der Nominierung besteht in der Auswahl der Samples, die in jeder der v Variablen einen der Topränge belegen. Der liberalste Ansatz nominiert die Gesamtheit der m Samples, die für mindestens eine der Variablen einen Toprang belegen.
Durch die fehlende Transitivität der Ähnlichkeitsrelation kann bei großen Komponenten die Übereinstimmung zwischen einzelnen Variablenpaaren relativ gering ausfallen, was dazu führt, dass nur sehr wenige Samples als potentielle Subgruppe nominiert werden. In diesem Fall kann statt der Variablengruppierung gemäß der Zusammenhangskomponenten auch eine feinere Gruppierung gewählt wer-

Sample ID	S_1	S_2	...	S_i	S_{i+1}	...	S_m
abs. Hfkt. in Toprängen	h_1	h_2	...	h_i	...		h_m
rel. Hfkt. in Toprängen	r_1	r_2	...	r_{min}	...		r_m
Nominiert	ja	ja	...	ja	nein	...	nein

Tabelle 4: Häufigkeitstabelle zur Nominierung einer Samplesubgruppe: Für jedes Sample, das in einer der Variablen unter den Toprängen der Variablen in einer zuvor identifizierten Variablengruppe G auftaucht, werden die absoluten (abs.) und relativen (rel.) Häufigkeiten (Hfkt.) ihres Auftretens in den Toprängen aller Variablen der Gruppe tabelliert. Der Parameter r_{min} bestimmt den zur Nominierung erforderlichen Mindestanteil.

den. Dazu werden die bestehenden Komponenten weiter in maximale Cliquen unterteilt. Die resultierenden Variablengruppen sind im Vergleich zu den Zusammenhangskomponenten kleiner und in Bezug auf die angezeigte Variablengruppe homogener. Ebenso können kleinere Gruppen auch durch einen strengeren Schwellenwert in Schritt 2 erreicht werden.

Für die gewählte Gruppierung in Zusammenhangskomponenten oder maximale Cliquen ist ein Scoring der Variablengruppen möglich, das die Evidenz der jeweils angezeigten Subgruppe widerspiegelt. Ein sinnvolles Maß dafür ist der Median der FS-Scores der beteiligten Variablen. Denkbar sind auch das arithmetische Mittel oder das Maximum der Scores.
Zusätzlich zur Sortierung der Variablengruppen ist an dieser Stelle auch ein Filtern nach der Gruppengröße möglich. Der Vorteil der Nominierung einer Samplegruppe aus einer Gruppe von Variablen besteht in der verbesserten Genauigkeit der Nominierung, die durch die Konsensbildung der Variablen erreicht werden soll. Da bei ein-elementigen Variablengruppen keine Konsensentscheidung möglich ist, ist eine Behelfslösung durch die Nominierung aller top $max.rk$ Samples gegeben. Im oben genannten Kriterium zur Nominierung geschieht das durch die Festlegung der Mindestanzahl des Auftretens eines Samples in den Toprängen als aufgerundetes Produkt der Variablenanzahl und r_{min}. Es wird in diesen Fällen jedoch die genauere Betrachtung der Variable empfohlen, bei der die empirische Verteilung mit den üblichen Methoden auf Ausreißer untersucht wird (z. B. ähnlich wie bei der Outlier-Sum-Methode).
Obwohl durch das Gruppieren der SG-anzeigenden Variablen die Evidenz der Subgruppe erhöht werden soll, sollten die in einzelnen Variablen enthaltenen Informationen nicht von vornherein ignoriert werden. Eine fehlende Bestätigung durch weitere Variablen kann ebenso auf einen zu strengen Schwellenwert in

Schritt 2 zurückzuführen sein oder auf eine zu geringe Anzahl T der in Schritt 1 ausgewählten Variablen. Letzteres gilt besonders, falls eine größere Anzahl von Subgruppen vorliegt, die jeweils von einer größeren Anzahl von Variablen angezeigt wird, sodass auch die Variablen auf den Rängen jenseits von T noch subgruppenrelevante Informationen liefern.

4.4 FSBC: Biclustern nach FS-Selektion

Als vierte Methode in den Vergleich multivariater Ansätze zur Subgruppendektion wird mit FSBC die Kombination aus Fisher Sum und Biclustern aufgenommen. Dabei wird der gewählte Bicluster-Algorithmus, hier also der Plaid-Algorithmus, auf den reduzierten Datensatz aus den top-T-FS-Variablen angewendet. Die Anzahl der möglichen lokalen Optima wird im Gegensatz zur Standardanwendung so drastisch verringert und der Informationsgehalt über die wahre Subgruppe gesteigert, falls diese von den ausgewählten Variablen erfasst wird. Die Variante FSBC soll so zur Senkung der bereits angesprochenen hohen Variabilität der Biclusterergebnisse führen und ggf. sogar den Bedarf nach Ensemblemethoden senken.

5 Simulationsstudien

Das folgende Kapitel gliedert sich in zwei Abschnitte, die jeweils eine größere Simulationsstudie vorstellen. Der erste Teil (SimUni) befasst sich mit dem Vergleich verschiedener univariater Methoden, die das Ranking der Variablen eines hochdimensionalen Datensatzes gemäß der enthaltenen Informationen über eine potentielle, bisher unbekannte Patientensubgruppe ermöglichen sollen. Die wesentlichen Ergebnisse wurden in Ahrens et al. [22] bereits publiziert und werden hier nochmals dargestellt. In SimUni werden im Gegensatz zu zuvor in diesem Feld publizierten Simulationen verschiedene Verteilungsszenarien mit jeweils unterschiedlich deutlich auftretenden Subgruppen betrachtet. Außerdem werden die Methoden auch auf Variablen mit nicht-krankheitsspezifischen Subgruppen angewendet. Eine zusätzliche Erweiterung stellt die Berücksichtigung eines Likelihoodratios (LR) dar, sodass die Methoden nicht nur untereinander, sondern auch mit einer theoretisch optimalen Methode verglichen werden können.
Die zweite vorgestellte Studie (SimMulti) vergleicht vier multivariate Workflows zur Identifikation von Patientensubgruppen. Zwei davon stellen Varianten des hier neu entwickelten FSx-Workflows dar, eine etablierte Methode dient als Referenz und bei der vierten Methode handelt es sich um eine Kombination aus neuem und etablierten Ansatz. Für verschiedene Kombinationen aus Stichprobengröße und Subgruppengröße wird der Lokationsshift, d. h. der mittlere Abstand der Subgruppe zum Rest der Samples, variiert.

5.1 Simulationsstudie zum Vergleich univariater Subgruppendetektionsmethoden (SimUni)

Die hier beschriebene Simulationsstudie SimUni dient der Beantwortung der Frage, **welche univariate Methode am besten geeignet ist, um subgruppenanzeigende Variablen aus einem hochdimensionalen Datensatz zu identifizieren** und so Hinweise auf Patientensubgruppen aufzudecken. Dabei wird insbesondere untersucht, welchen Einfluss die den Beobachtungen der Subgruppe zugrundeliegende Verteilung auf das Ranking der Methoden hat.
Dazu werden nach der Festlegung der benötigten Notation die untersuchten Verteilungen der Subgruppen definiert und die Struktur der simulierten Datensätze beschrieben. Nach der Auflistung der sieben zu vergleichenden univariaten Methoden wird das Likelihoodratio definiert, das als (theoretisch optimale) Referenzmethode für die konkurrierenden Methoden dient. Abschließend erfolgt die Beschreibung des Gütekriteriums, auf dessen Basis die insgesamt acht Methoden beurteilt werden.

5.1.1 Notation und Generierung der Daten

Gemäß der eingangs definierten Fragestellung sollen Variablen identifiziert werden, bei denen eine Subgruppe in der heterogenen Gruppe K (z. B. krank) einer eigenen Verteilung folgt, während sich die Verteilungen der übrigen Beobachtungen in K und der gesamten homogenen Gruppe G (z. B. gesund) nicht unterscheiden. Dieses Verteilungsmuster beschreibt eine krankheitsspezifische (ks) Subgruppe und wird als interessierende Alternative mit H_1 bezeichnet. Seien f_1 und f_0 die Dichten der Werte einer Gruppe mit bzw. ohne vorhandene Subgruppe. Bei Vorliegen von H_1 beschreibt folglich f_1 die Gruppe K und f_0 die Gruppe G mit $f_0 \neq f_1$.
Im Gegensatz dazu stehen die Variablen aus der Nullsituation H_0, in der sich die Dichten der beiden Gruppen nicht unterscheiden, beide folgen entweder f_0 oder f_1. Liegt keinerlei Subgruppenstruktur vor, gilt f_0 für G und K. Diese Situation sei mit H_{0a} bezeichnet. In vergleichbaren Studien wurde H_{0a} bisher als einzige Nullsituation betrachtet.
Bei der Auswertung echter omics-Datensätze mit SG-Detektionsmethoden findet sich häufig ein weiteres Verteilungsmuster, das als Nullsituation interpretiert werden kann: Sowohl in K als auch in G liegt eine Subgruppe vor, deren Verteilung sich offenbar von der der übrigen Beobachtungen unterscheidet (d. h. f_1 gilt in G und K). Dieses Muster einer nicht-krankheitsspezifischen (nks) Subgruppe wird mit H_{0b} bezeichnet.
Anhand der zusätzlichen Berücksichtigung von H_{0b} in der Simulationsstudie lässt sich die Robustheit der verschiedenen Methoden gegenüber solchen Verteilungsmustern beurteilen. Der Anteil der Variablen der Nullsituation mit dem üblichen Muster sei $p_{H_{0a}} \in [0,1]$. Dieser Anteil ist für reale Datensätze unbekannt und kann stark variieren. In der durchgeführten Simulationsstudie wurde einmal zum Vergleich mit anderen Studien $p_{H_{0a}} = 1$ gewählt (einfache Nullsituation), sowie zur Veranschaulichung des Einflusses der nks Subgruppen ein relativ geringer Anteil von $p_{H_{0a}} = 0.5$ für die kombinierte Nullsituation.

Zur Simulation von Genexpressionswerten oder Daten aus vergleichbaren Technologien werden die Beobachtungen üblicherweise aus der Standardnormalverteilung Φ, $N(0,1)$, gezogen, die das Rauschen in echten Daten reflektieren sollen. f_0 bezeichne die Dichte von Φ und beschreibe die Beobachtungen einer homogenen Gruppe ohne Subgruppen. Die alternative Dichte f_1 einer Gruppe mit einer enthaltenen Subgruppe ist abhängig von deren Verteilung d_s. Hält man an der Annahme von Φ für die übrigen Beobachtungen fest, ergibt sich allgemein folgende Mischverteilung

$$D_s = (1-q) \cdot N(0,1) + q \cdot d_s \tag{2}$$

für die Gesamtheit der Beobachtungen einer Gruppe mit einer SG vom Anteil q.

Im Folgenden werden drei verschiedene Szenarien $s =$ I, II, III, betrachtet. Diese entsprechen (I) einer einfachen Lokationsänderung, (II) einer gleichzeitigen Erhöhung von Varianz und Erwartungswert sowie (III) einer Varianzerhöhung in der SG. Tabelle 5 zeigt die in den verschiedenen Szenarien verwendeten Verteilungen d_s zur Generierung der Beobachtungen aus einer Subgruppe sowie die betrachteten Parameterbereiche Z.
In den üblichen Simulationen, die bisher im Bereich der univariaten SG-Detektion zur Gütebeurteilung vorgestellt wurden, wurde nur Szenario I berücksichtigt. In diesem folgen die Beobachtungen der Subgruppe einer verschobenen Normalverteilung $N(\delta,1), \delta > 0$, es handelt sich um eine reine Shift-Alternative. Erst in Ahrens et al. [22] wurde eine größere Anzahl von Alternativen untersucht.
Aus biologischer Sicht scheint es genauso plausibel, dass mit einem Shift in der Subgruppe auch eine Erhöhung der Varianz einhergeht (Shift-Scale-Alternative). In Szenario II werden durch die Addition rechteckverteilter Werte aus $U[0,b]$ gleichzeitig der Erwartungswert und die Variation der Subgruppe erhöht. Obwohl diese Mischverteilung nicht als tatsächlich zugrunde liegende, wahre Verteilung anzunehmen ist, ist sie für den Zweck dieser Studie gut geeignet.
Durch die Wahl der unteren Grenze als 0 ergibt sich kein „Rand“ zwischen der Subgruppe und den übrigen Beobachtungen, was die Detektion der Subgruppe erschwert. Außerdem zeigt die Erfahrung mit Datensätzen unterschiedlicher omics-Technologien, dass die generierten Verteilungsmuster mit denen in echten Daten durchaus vergleichbar sind. Die Szenarien I und II stellen zwei Fälle dar, in denen die Subgruppe einmal relativ klar (für großes δ) und einmal gar nicht vom Rest der Beobachtungen abgegrenzt ist.
Unabhängig von der konkreten Verteilung der Subgruppe wird diese in der Regel eine erhöhte Varianz in der betroffenen Gruppe bewirken. Szenario III untersucht eben diese Gemeinsamkeit der weiteren denkbaren Möglichkeiten für eine Subgruppenverteilung unter Verwendung von $N(0,\sigma^2), \sigma > 1$. Im Gegensatz zu den Szenarien I und II entsteht durch die Subgruppe hier keine Änderung des Erwartungswertes.

s	z	Z	d_s	Erhöhung von
I	δ	$[0,7]$	$N(\delta,1)$	Erwartungswert
II	b	$[0,10]$	$N(0,1)+U[0,b]$	Erwartungswert und Varianz
III	σ	$[1,6]$	$N\left(0,\sigma^2\right)$	Varianz

Tabelle 5: Übersicht über die in SimUni verwendeten Verteilungen d_s für die Generierung von Beobachtungen einer Subgruppe in den verschiedenen Szenarien s. Der Unterschied der Subgruppe wird allgemein mit z bezeichnet.

Als Maß für den Unterschied zwischen der Verteilung der Subgruppe und der Verteilung der übrigen Beobachtungen der Gruppe dient allgemein der Parameter z. Den verschiedenen Szenarien s entsprechend, beschreibt z entweder δ, b oder σ.

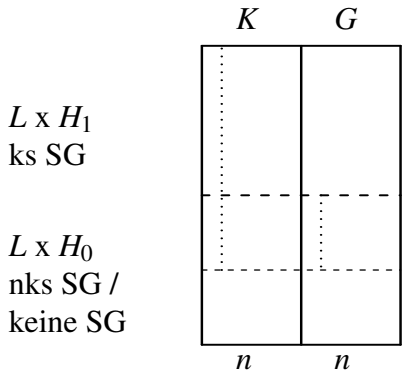
In bisher veröffentlichten Arbeiten wurden bei variierenden Stichproben- und Subgruppengrößen meist nur wenige ausgewählte Shifts δ betrachtet (d. h. insbesondere nur Szenario I). Im Gegensatz dazu werden in der Simulationsstudie in Ahrens et al. [22] alle Kombinationen der folgenden Größen über einen breiten Parameterbereich (vgl. Tabelle 5) berücksichtigt:

- die einfache und die kombinierte Nullsituation, $p_{H_{0a}} = 1, 0.5$
- SG-Verteilungen $s =$ I, II und III
- Gruppengrößen $n = 20, 30, 50, 70, 100$
- SG-Anteile $q = 0.1, 0.2, 0.3, 0.5, 0.75, 1$

Der Fall $q = 1$ wurde der Vollständigkeit halber aufgenommen und entspricht einer homogenen Gruppe, in der alle Beobachtungen der gleichen Verteilung gemäß Szenario s folgen. In der vorliegenden Arbeit werden die Ergebnisse nur für eine Auswahl dieser möglichen Parameterkonstellationen gezeigt, und zwar für $n = 30, 70, 100$ und $q = 0.1, 0.2, 0.3$. Für die Plots der übrigen Konstellationen sei auf das online verfügbare ergänzende Material zu Ahrens et al. [22] verwiesen.
Als Datenbasis für SimUni werden pro Parameterkombination (s, n, q, z) nach dem in Abbildung 3 gezeigten Schema insgesamt $2L = 2\,000$ Variablen generiert, die jeweils zur Hälfte aus Nullsituation und Alternative stammen.

5.1.2 Univariate Methoden im Vergleich

In der SimUni-Studie wird die neue Methode Fisher Sum (FS) mit der *outlier robust t-statistic* (ORT), *outlier sum* (OS), *percentile analysis for differential gene expression* (PADGE), Kurtosis im Sinne von PAK, Students t-Test und Bartletts Varianztest verglichen. Details zu FS, ORT, OS, PADGE und PAK wurden in Abschnitt 3.2 vorgestellt. Der t-Test wird als üblicherweise in differentiellen Studien verwendete Referenzmethode in die Simulationen aufgenommen. Bartletts Test auf Varianzhomogenität dient als Referenzmethode in Szenario III. Von Interesse ist seine Sensitivität gegenüber erhöhter Varianz, die nur in einer Subgruppe vorliegt.
Für FS, ORT, OS, PADGE und den t-Test wird jeweils die einseitige Variante zur Suche nach hochregulierten Subgruppen in der heterogenen Gruppe verwendet. Bei der Berechnung von PADGE werden insgesamt vier Quantile berücksichtigt (90%-, 85%-, 80%- und 50%-Quantil). Im entsprechenden Artikel [21] wurde zwar die Grundidee vermittelt, nicht aber alle Details der Scoreberechnung



(a) Simulationsschema SimUni

Situation	G	K	# Variablen	Verteilungsmuster
H_{0a}	f_0	f_0	500	keine Subgruppe vorhanden
H_{0b}	f_1	f_1	500	Subgruppe in G und in K (nks)
H_1	f_0	f_1	1000	Subgruppe nur in K (ks)

(b) Anzahl der Variablen aus Nullsituation und Alternative

Abbildung 3: Schema der SimUni-Studie: (a) Struktur der Datenmatrix in der kombinierten Nullsituation mit $p_{H_{0a}} = 0.5$ (b) Variablenanzahlen aus Nullsituation und Alternative pro Setting

dargestellt. Daher sind Abweichungen zwischen der Originalversion und der in der vorliegenden Arbeit (und somit auch in Ahrens et al. [22]) verwendeten Implementierung nicht auszuschließen. Die Kurtosis wird wie in Teschendorff et al. [18] beschrieben für die Menge aller Beobachtungen einer Variable gemeinsam berechnet. Die Berechnung der Kurtosis erfolgt mithilfe der Funktion `kurtosis` aus dem **R**-Paket `e1071` [61]. Dabei ist die Einstellung `type=2` zu wählen.

5.1.3 Likelihoodratio

Für ein Paar aus einfacher Null- und Alternativhypothese erlaubt das Likelihoodratio (LR) bzw. der daraus abzuleitende optimale Likelihoodratio-Test (Neyman-Pearson-Lemma) die Beurteilung der Evidenz der Alternative anhand einer gegebenen Stichprobe. In SimUni ist von Interesse, ob die Beobachtungen einer Variable aus einer Verteilung mit einer ks Patientensubgruppe (H_1) stammen oder ob kein Unterschied zwischen den Gruppen besteht (H_{0a} oder H_{0b}).
Da in der Simulation die zu den beiden Hypothesen gehörigen Dichten vollständig bekannt sind, kann das entsprechende LR für jede Variable des simulierten Datensatzes bestimmt werden. Das LR wird aufgrund seiner Optimalitätseigenschaften daher als weiterer univariater Score in den SimUni-Vergleich aufgenommen. Da

an dieser Stelle ein Ranking der Variablen gemäß der Evidenz für die Alternative ausreichend ist, kann auf die übliche (monotone) Logtransformation und das Bestimmen eines p-Wertes verzichtet werden.
Durch die Berücksichtigung des LR können die konkurrierenden SG-Detektionsmethoden nicht nur untereinander verglichen werden, sondern ihre Performanz auch in Relation zu einer optimalen Referenz gesetzt werden. So können ggf. Parameterbereiche identifiziert werden, in denen eine Verbesserung der momentan verfügbaren Methoden am sinnvollsten wäre. Da die Berechnung des LR die Kenntnis aller Verteilungsparameter voraussetzt, ist diese Methode auf echte Datensätze jedoch nicht direkt übertragbar.

Wie oben definiert, bezeichnet f_0 die Dichte einer Variablen in einer homogenen Gruppe, in der keinerlei Subgruppen vorliegen, d. h. f_0 ist die Dichte der Standardnormalverteilung. Im Gegensatz dazu kennzeichnet f_1 die gemischte Dichte aus $N(0,1)$ und einer abweichenden Verteilung d_s (vgl. Formel (2)), die je nach Szenario variiert. Für die Berechnung des LR wird der Quotient von L_1 und L_0 gebildet, wobei L_1 die Likelihood basierend auf dem gesuchten Verteilungsmuster H_1 ist, und L_0 die Likelihood unter der Nullsituation H_0. Mit obigen Definitionen lässt sich das LR schreiben als

$$\begin{aligned} LR = L_1/L_0 &= \frac{\prod_G f_0 \cdot \prod_K f_1}{p_{H_{0a}} \prod_G f_0 \prod_K f_0 + (1-p_{H_{0a}}) \prod_G f_1 \prod_K f_1} \\ &= \frac{1}{p_{H_{0a}} \prod_K \frac{f_0}{f_1} + (1-p_{H_{0a}}) \prod_G \frac{f_1}{f_0}}. \end{aligned}$$

In dieser Kurzschreibweise steht $\prod_G f_0$ beispielsweise für die Produktbildung der Werte von f_0, die an den in der Gruppe G beobachteten Werten ausgewertet wird, ausführlicher also $\prod_{x_i \in G} f_0(x_i)$. Der Zähler repräsentiert die interessierende Alternative, in der die Dichte f_0 der Standardnormalverteilung für die Beobachtungen in G und die subgruppenanzeigende gemischte Dichte f_1 in der Gruppe K gilt. Die Linearkombination im Nenner spiegelt die Nullsituation wider. Die beiden Summanden stellen die Fälle dar, in denen die Dichten der beiden Gruppen übereinstimmen: für beide gilt f_0 (H_{0a}) bzw. für beide Gruppen gilt f_1 (H_{0b}). Ist der Anteil $p_{H_{0a}}$ des ersten Fall gleich 1, vereinfacht sich die Darstellung zu

$$LR = \prod_K \frac{f_1}{f_0} = \prod_K \left[(1-q) + q \frac{f_z}{f_0} \right],$$

wobei f_z die Dichte der SG-Verteilung mit Parameterwert z ist. In Szenario I wäre f_2 demnach die Dichte der $N(2,1)$-Verteilung. Hier beruht die Beurteilung der Evidenz also nur noch auf den Beobachtungen der Gruppe K. Der Parameter q gibt den wahren Anteil der Subgruppe in K an. Große Werte des Likelihoodratios

sprechen für das Vorliegen einer krankheitsspezifischen Subgruppe. Pseudocode für die wesentlichen Schritte der Berechnung des LR in den drei Verteilungsszenarien in **R** ist im Anhang zu finden.

5.1.4 Qualitätskriterium

Eine Reihe von Simulationsstudien im Bereich der univariaten SG-Detektion nutzt zur Beurteilung der Güte der unterschiedlichen Verfahren *receiver operating characteristics*-Kurven, kurz ROC-Kurven. Dazu definiert man die Sensitivität (Richtig-Positiv-Rate) als die Wahrscheinlichkeit, eine Variable mit vorhandener krankheitsspezifischer (ks) Subgruppe in K zu identifizieren (H_1) und die Spezifität (1 – Falsch-Positiv-Rate) als die Wahrscheinlichkeit, dass eine Variable aus der Nullsituation (H_{0a} oder H_{0b}) korrekt klassifiziert wird. Für die ROC-Kurve wird dann die Richtig-Positiv-Rate gegen die Falsch-Positiv-Rate der jeweiligen Methoden abgetragen, um die resultierenden Kurven zu vergleichen.
Die Betrachtung solcher ROC-Kurven ist allerdings nur für eine geringe Zahl untersuchter Parameterkonstellationen praktikabel. Aufgrund des breiten Spektrums an Alternativen, das in dieser Arbeit berücksichtigt wird, ist dieses Vorgehen hier nicht empfehlenswert. Stattdessen wird die in der ROC-Kurve enthaltene Information auf die Fläche unter der Kurve (AUC, *area under the curve*) komprimiert und auf die nähere Betrachtung der einzelnen ROC-Kurven verzichtet.
Die Verteilungen in den drei Simulationsszenarien wurden so gewählt, dass sie durch einen einzelnen Parameter $z, z \in \{\delta, b, \sigma\}$, parametrisierbar sind. Dies ermöglicht einen übersichtlichen Vergleich der Methoden für eine feste Parameterkombination (s,n,q) anhand der Plots der AUC-Werte gegen z. Abbildung 26 im Anhang zeigt exemplarisch vergleichende Plots der ROC-Kurven. Zur Berechnung der ROC-Kurven und AUC-Werte wurde das **R**-Paket `pROC` [62] verwendet.

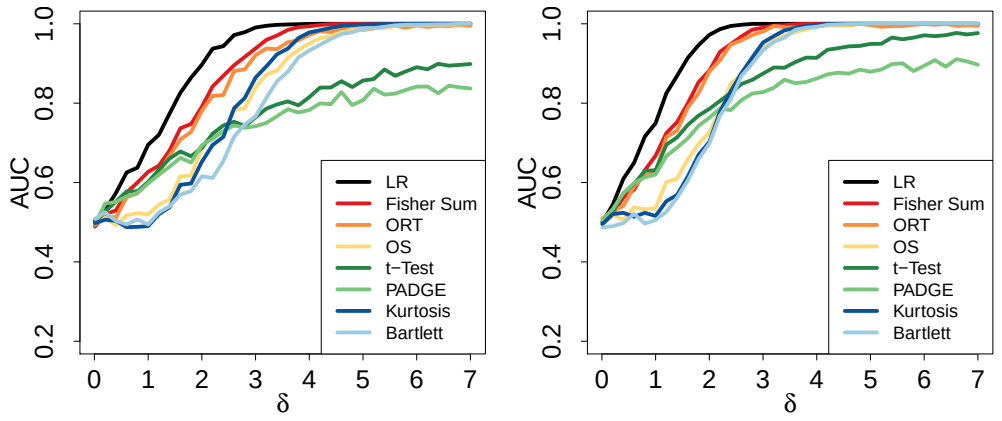
5.2 Ergebnisse der SimUni-Studie

Anhand ihrer AUC-Verläufe werden die SG-Detektionsmethoden nun im Hinblick auf ihre Fähigkeit verglichen, die Gesamtheit der Variablen eines Datensatzes anhand des jeweils zugeordneten Scores bzw. p-Werts korrekt in die Klassen „H_1 : ks Subgruppe vorhanden“ bzw. „H_0 : kein Unterschied zwischen den beiden Gruppen vorhanden“ trennen zu können. Die folgenden Erläuterungen behandeln hauptsächlich die Kombinationen von $n = 30, 70$ und $q = 0.1, 0.2, 0.3$ für die drei Verteilungsszenarien. Begonnen wird mit der ausführlicheren Darstellung der Ergebnisse des klassischen Designs ($p_{H_{0a}} = 1$) in den Szenarien I und II. Es folgt der Vergleich mit der kombinierten Nullsituation ($p_{H_{0a}} = 0.5$, $s =$ I, II) bevor das allgemeine Szenario III mit erhöhter Varianz zusammengefasst wird.

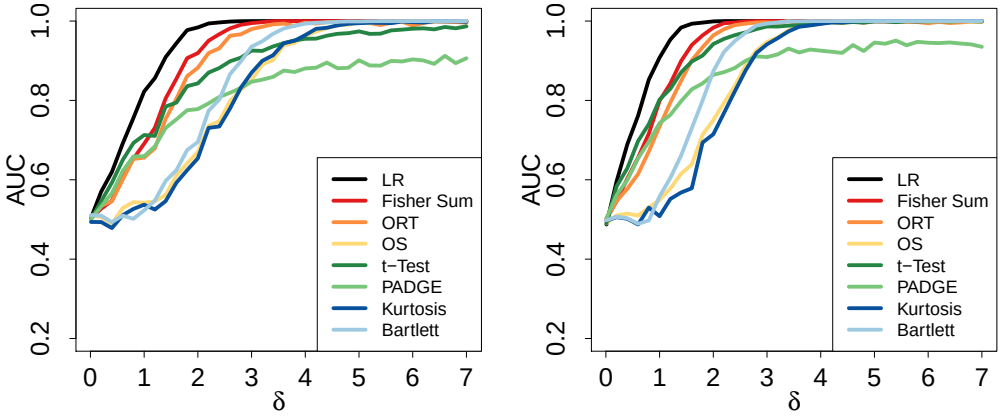
Szenarien I und II, einfache Nullsituation $p_{H_{0a}} = 1$

Zunächst werden die SimUni-Ergebnisse in den Szenarien I und II für das konventionelle Design besprochen, d. h. mit der einfachen Nullsituation, in der entweder gar keine Subgruppe vorliegt (H_{0a}) oder eine krankheitsspezifische Subgruppe mit variierender Abweichung von $N(0,1)$ auftritt. Die zugehörigen AUC-Kurven für Szenario I sind in Abbildung 4 dargestellt, die für Szenario II sind im Anhang zu finden (Abb. 27). Die Ergebnisse bezüglich der Methoden mit ähnlichen Verläufen und dem Verhalten dieser Gruppen sowohl zueinander als auch zum LR unterscheiden sich im Wesentlichen nicht zwischen den Szenarien I und II. Auch die Unterschiede zwischen den Fallzahlen $n = 30$ und 70 sind für feste Kombinationen (s,q) gering. Die Performanz ist für größere Fallzahlen generell höher, die AUC-Kurven sind demnach etwas steiler bzw. nach links verschoben.

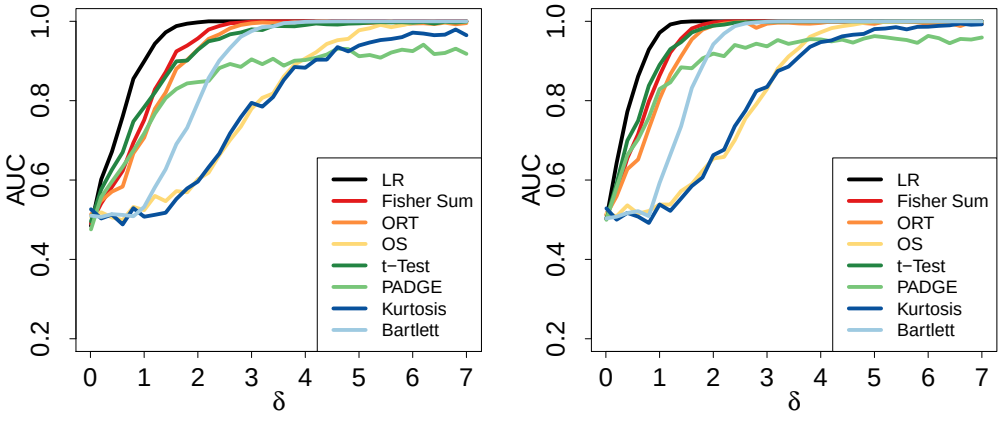
Fisher Sum und ORT zeigen in allen (n,q)-Kombinationen einen sehr ähnlichen Verlauf. Sie erreichen unter den konkurrierenden Methoden für kleine Subgruppenanteile ($q = 0.1$) die besten AUC-Werte im Bereich kleinerer bis moderater Abweichungen. In Szenario I trennt sich das Paar FS/ORT für $q = 0.1$ deutlicher von den übrigen Methoden als in Szenario II.
Mit wachsendem q schließt der t-Test zu FS/ORT auf und ist für einen Subgruppenanteil q über 30% die beste Wahl (Plots hier nicht gezeigt). Scheinbar erreicht das Paar aus t-Test und PADGE nur auf dem Rand der Hypothese, d. h. für z nahe Null, immer die besten AUC-Werte. Der Bereich vergrößert sich mit q. Für größeres z werden PADGE und t-Test dann von Fisher Sum und ORT überholt. Der übliche t-Test ist der abgewandelten Variante in PADGE in unseren Simulationen stets überlegen, wobei der Unterschied für kleine Werte von z gering ausfällt. PADGE übertrifft im Gegensatz zum t-Test jedoch das Paar Fisher Sum/ORT auch für größere q und z nicht.
Als letztes ist die Gruppierung aus Kurtosis, Bartletts Test und OS zu erkennen. Im Gegensatz zu den übrigen Verfahren steigt die AUC nicht direkt für $z > 0$, sondern die Kurven formen zunächst ein Plateau beim AUC-Ausgangswert von 0.5 bis etwa $\delta = 1$ bzw. $b = 1$. Dadurch erreicht diese Gruppe für kleine bis moderate Werte von z im Vergleich die niedrigsten AUC-Werte. Dies lässt sich z. B. für die OS dadurch erklären, dass nur Werte zur Teststatistik beitragen, die oberhalb eines Schwellenwertes liegen ($q_{75} + iqr$). Solange der Shift klein ist, d. h. der Erwartungswert der Subgruppe nur minimal über dem der übrigen Beobachtungen, ergibt sich kein Unterschied in der Verteilung der Scores zwischen H_0 und H_1.
Mit wachsendem q erzielt Bartletts Test höhere AUC-Werte als Kurtosis und OS, die sich dann kaum unterscheiden. In Szenario I ist der Varianztest dabei schon für $q = 0.3$ deutlich überlegen, in Szenario II, d. h. bei schwächerer Trennung der Subgruppe, ist diese Trennung erst für $q = 0.5$ erkennbar. In Szenario I fällt



(a) $n = 30$, $q = 0.1$ (b) $n = 70$, $q = 0.1$



(c) $n = 30$, $q = 0.2$ (d) $n = 70$, $q = 0.2$



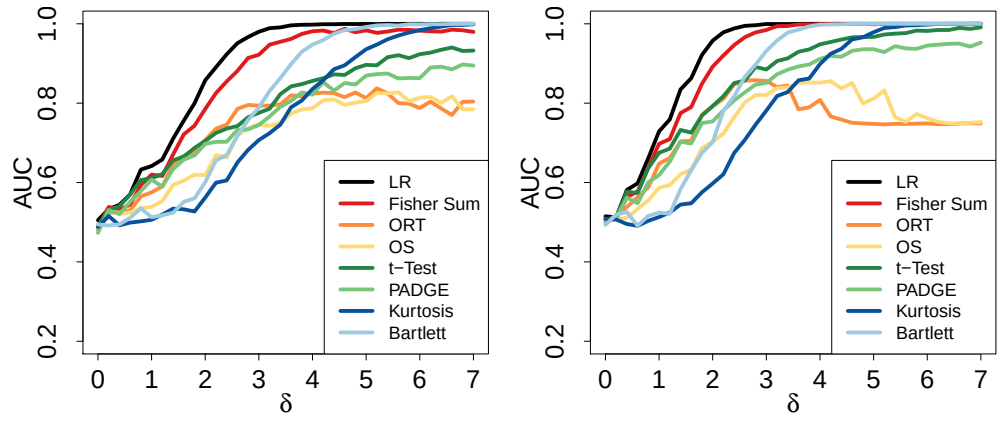
(e) $n = 30$, $q = 0.3$ (f) $n = 70$, $q = 0.3$

Abbildung 4: Ergebnisse für ausgewählte Parameterkonstellationen in SimUni, Szenario I, $p_{H_{0a}} = 1$.

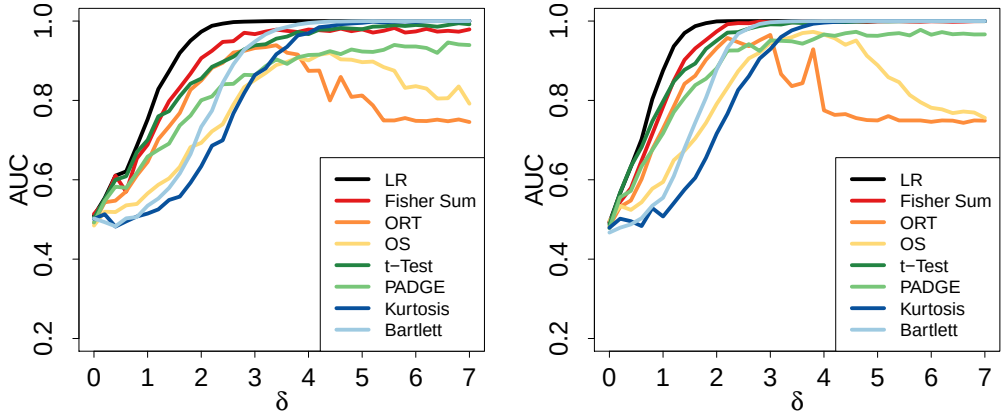
zudem auf, dass Kurtosis und OS im Gegensatz zu den anderen Methoden mit wachsendem q flachere Kurvenverläufe zeigen, d. h. es wird bei gleichem z für $q = 0.3$ eine geringere AUC erreicht als bei $q = 0.2$.
Die bisherigen Ergebnisse lassen sich also wie folgt zusammenfassen: In den **Szenarien I und II mit der einfachen Nullsituation zeigen Fisher Sum und ORT im Vergleich sowohl zu den übrigen Methoden als auch zum Likelihoodratio eine sehr gute Performanz. Dies gilt insbesondere im Bereich kleinerer bis moderater Subgruppen**, die weder von der Gruppe der t-Tests (t-Test und PADGE) noch von der Kurtosis/OS/Bartlett-Gruppe gut erfasst werden.

Szenarien I und II, zusammengesetzte Nullsituation, $p_{H_{0a}} = 0.5$

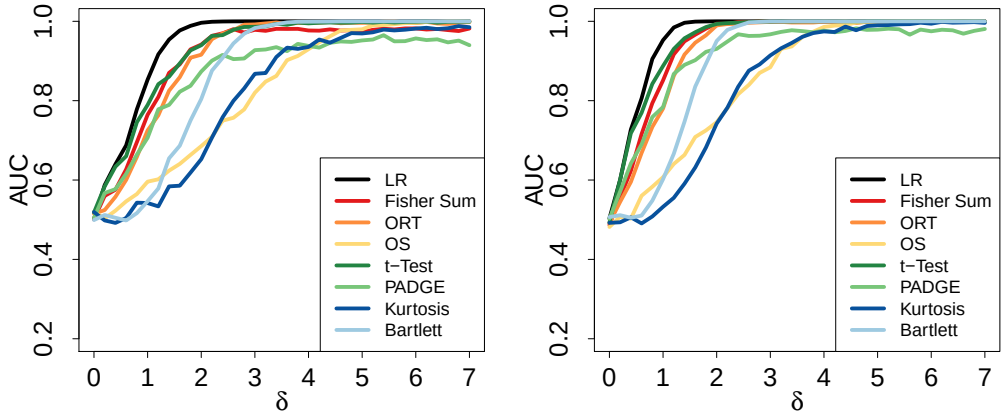
Der wesentliche Unterschied zwischen den beiden vielversprechendsten Methoden Fisher Sum und ORT wird bei Berücksichtigung nicht-krankheitsspezifischer Subgruppen in der kombinierten Nullsituation ersichtlich (s. Abb. 5): Während Fisher Sum auch für einen großen Anteil von Variablen mit nks Subgruppen eine vergleichbar gute Performanz aufweist, verliert ORT deutlich und ähnelt in seinem Verlauf stark der OS.
Besonders zu beachten ist dabei, dass im Fall der gemischten Nullsituation bei kleinen bis moderaten Subgruppenanteilen ($q = 0.1, 0.2$) die AUC-Werte von ORT und OS mit wachsendem Unterschied z der Subgruppe nicht gegen 1 konvergieren. In diesen Fällen werden sowohl den Variabeln mit krankheitsspezifischer als auch mit nks Subgruppe hohe Scores zugewiesen und die als Nullsituation definierten Muster erhalten zu hohe Scores. Erst für größeres $q = 0.3$ konvergieren die AUC-Werte der beiden Methoden wieder wie gewünscht gegen 1. In diesem Fall zeigen 30% der Werte pro Gruppe und somit auch 30% aller Beobachtungen für die nks Variablen (durchschnittlich) erhöhte Werte, was die Erhöhung des 75%-Quantils und somit einen größeren Schwellenwert für die OS-Berechnung nach sich zieht.
Der t-Test und seine Abwandlung PADGE verhalten sich im Vergleich zu FS wie bei der einfachen Nullsituation: Abgesehen von der direkten Nähe zur Nullhypothese ist FS für kleinen Subgruppenanteil $q = 0.1$ bei moderaten Unterschieden z beiden Methoden deutlich überlegen. Diese Überlegenheit verringert sich mit wachsendem Anteil. Ab etwa 30-50% Subgruppenanteil werden vom t-Test höhere AUC-Werte als von FS erreicht. PADGE erzielt zwar mit wachsendem q ebenfalls höhere AUCs, reicht aber für $q = 0.3$ nicht an FS und den t-Test heran.
Wie in der einfachen Nullsituation zeigen sich keine wesentlichen Unterschiede zwischen den AUC-Verläufen beim Vergleich der betrachteten Fallzahlen $n = 30$ bzw. 70. Ebenso erreichen die Methoden in Szenario I bessere Performanz (verglichen mit dem jeweiligen LR) als in Szenario II.



(a) $n = 30$, $q = 0.1$ (b) $n = 70$, $q = 0.1$



(c) $n = 30$, $q = 0.2$ (d) $n = 70$, $q = 0.2$



(e) $n = 30$, $q = 0.3$ (f) $n = 70$, $q = 0.3$

Abbildung 5: Ergebnisse für ausgewählte Parameterkonstellationen in SimUni, Szenario I, $p_{H_{0a}} = 0.5$.

Szenario III

Szenario III untersucht die erhöhte Varianz in einer Subgruppe ohne eine zusätzliche Verschiebung des Erwartungswertes. In diesem Fall weist im Gegensatz zu den Szenarien I und II somit im Schnitt nur (maximal) die Hälfte der Werte aus der Subgruppe erhöhte Werte auf, die Änderung der Expression ist bidirektional. Als mögliche Referenz für dieses Szenario wurde daher Bartletts Test auf Varianzhomogenität in die Vergleiche aufgenommen.
In Szenario III zeigt sich für $p_{H_{0a}} = 1$ eine fast identische Gruppierung der untersuchten Methoden wie in den Szenarien I und II: Die Gruppe aus FS, ORT und OS zeigt für kleines $q = 0.1$ in der einfachen Nullsituation sehr ähnliche Kurven. Mit wachsendem q verbessern sich FS und ORT gleichermaßen; der Abstand zur OS vergrößert sich. Insgesamt ist diese Gruppe aber für kleine und moderate Subgruppen Bartletts Test und der Kurtosis deutlich unterlegen. Für $q = 0.1$ besteht noch eine Überlegenheit der Kurtosis; ab $q = 0.3$ erreicht Bartletts Test die höchsten AUC-Werte. Da keine Änderung des Erwartungswerts vorliegt, liegen die beobachteten AUC-Werte des t-Tests für alle Standardabweichungen σ bei etwa 0.5, für PADGE nur leicht darüber.
In Szenario III der einfachen Nullsituation ist der Abstand des jeweils besten verfügbaren Tests (Kurtosis bzw. Bartletts Test) zum Likelihoodratio relativ gering und zumindest für größeres q steigen die Kurven von FS und ORT für kleinere σ schnell. Beim Wechsel zur kombinierten Nullsituation fällt die Performanz der Kurtosis stark ab, etwa auf das Niveau von OS und ORT. Wie in den Szenarien I und II verhalten letztere sich sehr ähnlich und erreichen im betrachteten Wertebereich für σ maximal etwa 0.8. Bartletts Test als generell beste Methode trennt Nullsituation und Alternative erst für $q = 0.3$ ausreichend zuverlässig. Als zweitbeste Methode zeigt sich die Fisher Sum. Im Setting $(s,n,q) = (\text{III}, 30, 0.1)$ mit $p_{H_{0a}} = 0.5$ liegt der größte in SimUni beobachtete Unterschied zwischen dem LR und jeweils der besten Methode vor.

5.3 Simulationsstudie zum Vergleich multivariater Subgruppendetektionsmethoden (SimMulti)

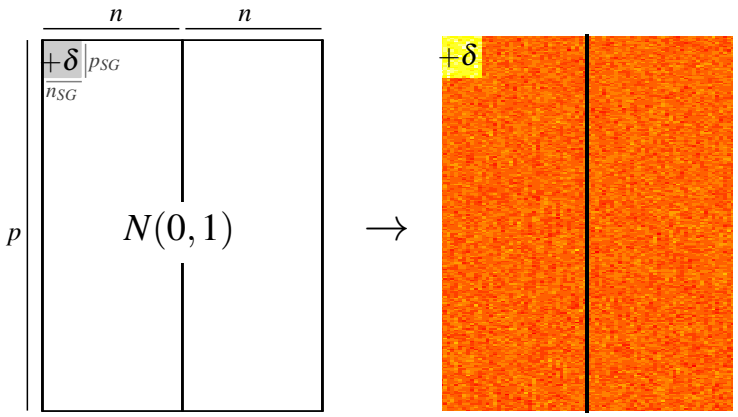
Das Hauptziel der SimMulti-Studie ist der **Vergleich der Detektionsgüte neuer und etablierter multivariater SG-Detektionsmethoden mit Fokus auf kleinere Patientensubgruppen, die sich nur in einer geringen Anzahl von Variablen zeigen**. Zu Beginn wird das allgemeine Schema der in der Simulation generierten Datensätze beschrieben (Abschnitt 5.3.1). Ein großer Teil der SimMulti-Studie ist den Sensitivitätsanalysen gewidmet, in denen der Einfluss wesentlicher methoden- und datensatzspezifischer Parameter auf die Detektionsgüte untersucht wird. Die

variierten Parameter werden ebenfalls in 5.3.1 erläutert. Auf die Auflistung der vier in SimMulti verglichenen Methoden (5.3.2) folgt die Definition des verwendeten Gütekriteriums (5.3.3). Die Darstellung der Ergebnisse zum Performanzvergleich der neuen FSx-Methode mit den Bicluster-basierten Ansätzen werden im anschließenden Abschnitt 5.4 vorgestellt. Erste Analysen mit diesem Design wurden in Ahrens et al. [55] zum Vergleich von FSOL und dem Biclustern anhand weniger Parameterkombinationen gezeigt.

5.3.1 Generierung der Daten

Allgemeines Datenschema in SimMulti

Basierend auf den Ergebnissen der Studie SimUni wird in der nun vorgestellten SimMulti-Studie nur eines der drei Verteilungsszenarien betrachtet. Alle Beobachtungen entstammen einer Normalverteilung und die erhöhte Expression der Subgruppe in einigen Variablen wird durch einen Lokationsshift δ reflektiert (d. h. $N(\delta,1)$ statt $N(0,1)$). Eine Patientensubgruppe der Größe n_{SG} bewirkt in p_{SG} Variablen einen Shift δ. Abbildung 6 zeigt die Struktur der Datensätze, die in den einzelnen SimMulti-Läufen generiert werden.
Pro Parameterkonstellation werden $L = 500$ Datensätze erzeugt. Für jede Kombination aus Fallzahl $n = 40, 70$ pro Gruppe und der Subgruppengröße $n_{SG} = 5, 10$ werden die Lokationsshifts $\delta = 2, 3, 4, 6$ in $p_{SG} = 5$ Variablen untersucht. Die Parameterkombination $(n, n_{SG}) = (40, 5)$ leitet sich aus dem später verwendeten Datenbeispiel ALL ab. Zusätzlich wird auch eine höhere Fallzahl und Subgruppengröße betrachtet.



(a) Simulationsschema (b) Heatmap

Abbildung 6: Schema zur Simulation der Datenmatrix eines Simulationslaufes in SimMulti und die Heatmap eines resultierenden Datensatzes.

Sensitivitätsanalysen

Tabelle 6 zeigt einen Überblick der Simulationen zur Sensitivitätsanalyse der vier SG-Detektionsmethoden gegenüber datensatz- und verfahrensspezifischen, sowie allgemeinen Parametern. Untersucht werden die Anzahl p der insgesamt im Datensatz enthaltenen Variablen und die Anzahl p_{SG} der Variablen, in denen die Subgruppe auftritt. Bezüglich der FSx-spezifischen Parameter wird der Einfluss der Anzahl T der im ersten Schritt selektierten Variablen, sowie die Anzahl *max.rk* der beim anschließenden paarweisen Vergleich der Variablen einbezogenen Topränge betrachtet.
Abschließend wird der mit `heterOnly` bezeichnete Parameter variiert. Er gibt an, ob die Ähnlichkeit der Variablen nur basierend auf der als heterogen angenommenen Gruppe beurteilt werden soll oder auf der Gesamtheit aller verfügbaren Samples aus beiden Gruppen. Dies könnte insbesondere für das Biclustern bezüglich der Variabiliät der Ergebnisse vorteilhaft sein, da der Suchraum für den Optimierungsprozess begrenzt wird. Es gilt aber zu bedenken, dass damit die Möglichkeit verloren geht, Ähnlichkeiten zwischen den gefundenen Subgruppen mit beispielsweise falsch gelabelten Samples aus der als homogen angenommenen Gruppe aufzudecken.
Der Einfluss der Parameter auf die Detektionsgüte wird dabei aus Gründen der Übersichtlichkeit hauptsächlich nach dem OFAT-Prinzip (*one-factor-at-a-time*) untersucht: Soweit nicht anders beschrieben, wurden für die jeweils festen Parameter die Standardeinstellungen gewählt, die in den detaillierten Methodenbeschreibungen gelistet sind. Da vermutlich auch die Interaktion von Fallzahl und Subgruppengröße einen Einfluss auf die Detektionsgüte hat, wird jede der Parametereinstellungen aus der Sensitivitätsanalyse für vier Kombinationen aus Fallzahl und Subgruppengröße betrachtet. Alle Ergebnisse werden in Form von Abbildungen vollständig im Anhang gezeigt und die wesentlichen Aussagen werden im Hauptteil der Arbeit zusammengefasst.

5.3.2 Multivariate Methoden im Vergleich

Im Rahmen der SimMulti-Studie werden vier multivariate Ansätze zur Subgruppendetektion verglichen. Dabei handelt es sich um die beiden Varianten FSOL und FSJ des neuen FSx-Workflows, der in Abschnitt 4.3 ausführlich beschrieben wurde. Als etablierte Referenzmethode dient der Plaid-Algorithmus als Vertreter für das Biclustern (4.2). Aufgrund der im zugehörigen Abschnitt angesprochenen Problematik der mitunter hohen Variabilität der Ergebnisse wird auch die in dieser Arbeit vorgeschlagene Kombination FSBC (4.4) als mögliche Alternative untersucht.

Symbol	Beschreibung	Werte
datensatzspezifische Parameter		
p	Variablenanzahl im Datensatz	1 000, 10 000, 50 000
p_{SG}	Anzahl der SG-anzeigenden Variablen	5, 20, 50
FSx-spezifische Parameter		
T	Anzahl FS-selektierter Variablen	25, 50, 100
max.rk	Anzahl zu vergleichender Topränge	5, 10, 15
allgemeine Parameter		
`heterOnly`	Samples zur Ähnlichkeitsbeurteilung	eine/beide Gruppen

Tabelle 6: Übersicht der durchgeführten Simulationen zur Analyse der Sensitivität der untersuchten SG-Detektionsmethoden gegenüber datensatz- und verfahrensspezifischen Parametern.

Der gewählte Bicluster-Algorithmus ermöglicht die Zuordnung von Variablen und Samples zu mehr als einem Bicluster. Die FSx-Ansätze sind diesbezüglich bei den Variablengruppen weniger flexibel. Die durch die verschiedene Variablenmengen nominierten Samplesubgruppen sind jedoch nicht notwendig disjunkt, d. h. Samples können wie beim Biclusteransatz in keiner, einer oder mehreren Subgruppen auftreten.
Für die Bicluster-Berechnungen wird auf die Funktion `BCPlaid` aus dem **R**-Paket `biclust` [53] zurückgegriffen. Dabei werden die Standardeinstellungen der verfügbaren Parameter übernommen. Soweit nicht anders beschrieben, werden bei den FSx-Ansätzen für die jeweils nicht variierten Workflowparameter die Standardeinstellungen gewählt, die in den detaillierten Methodenbeschreibungen gelistet sind (siehe z. B. Tabelle 3).

5.3.3 Gütekriterium

Da in den Simulationsstudien die tatsächlich vorhandene Samplesubgruppe bekannt ist, lässt sich die Güte der SG-Detektionsverfahren durch einen direkten Vergleich der wahren Subgruppe G_W mit der jeweils nominierten Subgruppe G_N beurteilen. In SimMulti erfolgt der Vergleich mithilfe des Jaccardindexes

$$J(G_W, G_N) = \frac{|G_W \cap G_N|}{|G_W \cup G_N|},$$

der schon als Ähnlichkeitsmaß in FSJ vorgestellt wurde. Der Jaccardindex berücksichtigt sowohl die Sensitivität als auch die Spezifität des SG-Detektionsverfahrens: Ausgehend von der optimalen Detektion (G_W=G_N und $J = 1$) nimmt J

sowohl bei nicht entdeckten Samples (falsch-negativ) als auch bei der Nominierung zusätzlicher Samples (falsch-positiv) ab. Sein Wertebereich liegt zwischen 0 und 1.
Der Jaccardindex wird im Bereich des Biclusterns bereits häufig eingesetzt, beispielsweise als Standardähnlichkeitsmaß zur Beurteilung der Ähnlichkeit verschiedener Bicluster-Outputs im **R**-Paket `superbiclust` [52] oder in Studien zum Vergleich verschiedener Bicluster-Algorithmen (z. B. Eren et al. [63]). Er ist daher eine naheliegende Wahl für das Gütekriterium der SimMulti-Studie.

Die Gütebeurteilung der untersuchten SG-Detektionsverfahren basiert in jedem Simulationslauf auf der jeweils „besten“ Subgruppe. Bei der praktischen Anwendung der Verfahren auf reale Datensätze wird eine solche Limitation auf eine einzelne Subgruppe nicht empfohlen. Unter Berücksichtigung des Simulationsdesigns mit genau einer vorhandenen wahren Patientensubgruppe ist sie hier durchaus sinnvoll. Bei den Bicluster-Varianten betrifft die Wertung jeweils das erste reportete Bicluster. Falls kein Bicluster gefunden wurde, wird die leere Menge als nominierte Subgruppe behandelt, was zu einem Jaccardindex von 0 führt. Das Scoring der Variablengruppen im FSx-Workflow wurde in Abschnitt 4.3.3 genauer beschrieben. Im Gegensatz zum Biclustern gibt es hier immer eine „beste“ Variablengruppe, die jedoch abhängig von den Nomierungskriterien nicht zwingend zur Nominierung einer Samplesubgruppe führt.
Zur übersichtlichen Darstellung der Ergebnisse werden die für eine feste Parameterkonstellation berechneten $L = 500$ Jaccardindizes in Form von Boxplots dargestellt. Dabei wird der Jaccardindex auf der y-Achse dargestellt und gegen den Shift $\delta = 2,3,4,6$ abgetragen.

5.4 Ergebnisse der SimMulti-Studie

Die Darstellung der Ergebnisse der SimMulti-Studie umfasst zwei Abschnitte. Zunächst werden die Sensitivitätsanalysen vorgestellt, die unter anderem im Rahmen der Festlegung der Defaultwerte der Workflowparameter durchgeführt wurden. Dabei werden die zugehörigen Plots hier jeweils nur für die Kombination (n, n_{SG})=(70, 10) und ggf. eine Auswahl der vier Methoden gezeigt. Die übrigen Plots finden sich dieser Arbeit angehängt. Nach dieser Untersuchung einzelner Parameter wird die Gesamtperformanz der vier Methoden FSOL, FSJ, Biclustern (BC) und FSBC für jeweils feste Parameterwerte ausführlich für die vier betrachteten Kombinationen (n, n_{SG}) verglichen.

5.4.1 Sensitivitätsanalysen

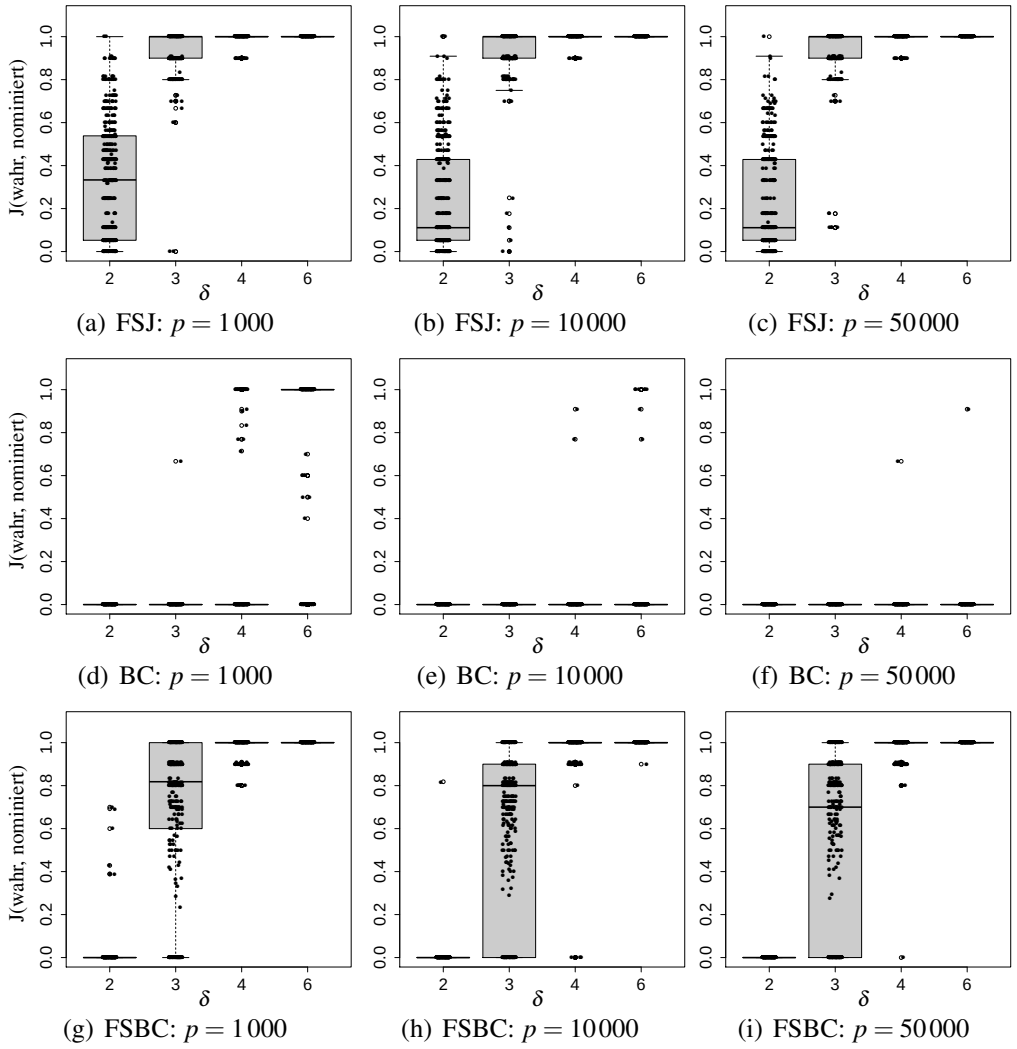
Der folgende Abschnitt fasst die Ergebnisse der Sensitivitätsanalysen der untersuchten Methoden gegenüber den in Tabelle 6 genannten allgemeinen, datensatz- und methodenspezifischen Parametern zusammen.

Einfluss der Datensatzgröße: Variablenanzahl p

Die Variation der Variablenanzahl im Datensatz über $p = 1\,000, 10\,000, 50\,000$ reflektiert unterschiedliche omics-Technologien, die Datensätze von typischer Größe erzeugen. Während bei einem Hochdurchsatz-Proteomikexperiment in der Regel etwa 1 000 Variablen analysiert werden, liegt diese Anzahl bei einem Microarraydatensatz aus beipielsweise mRNA-Chip-Experimenten mit 10 000-50 000 eine Größenordnung höher.
Abbildung 7 zeigt beispielhaft den Vergleich zwischen FSJ, BC und FSBC für die drei Datensatzgrößen. Die vollständigen Ergebnisse finden sich im Anhang in den Abbildungen 35 und 36 ($n = 40$), sowie in Abbildungen 37 und 38 ($n = 70$).
Die Erhöhung der Variablenanzahl p im Datensatz zeigt nur geringe Auswirkungen auf die Performanz der drei Methoden mit FS-Selektion FSOL, FSJ und FSBC. Im Gegensatz dazu hat die Variation von p einen deutlichen Einfluss auf das Biclustern: Während bei einer Datensatzgröße von $p = 1\,000$ Variablen für einen großen Lokationsshift von $\delta = 6$ in 75% der Simulationsläufe ein optimaler Jaccardindex von 1 erreicht wird, liegen für $p = 50\,000$ fast alle entsprechenden Werte bei 0 (für $(n, n_{SG}) = (70, 10)$). Das gleiche Muster zeigt sich in den drei übrigen (n, n_{SG})-Settings.
Aufgrund dieser Ergebnisse des Biclusterns werden die nachfolgenden Analysen für $p = 1\,000$ dargestellt, um Performanzänderungen besser beobachten zu können. Das Problem fehlender Sensitivität des Biclusterns bei der Detektion tatsächlich vorliegender Subgruppen lässt sich durch den Schritt der univariaten Vorselektion erheblich reduzieren: Durch die Anwendung der Methode auf die Teilmatrix der top-50-FS-Variablen (FSBC) wird eine ähnliche Güte und Stabilität gegenüber p wie bei den FSx-Verfahren erreicht.

Einfluss der Anzahl p_{SG} der subgruppenanzeigenden Variablen

Im Falle kleiner Patientensubgruppen, die sich nur in wenigen Variablen auswirken, ist die Performanz der etablierten Bicluster-Methode nicht zufriedenstellend. Selbst bei großer Abweichung der Subgruppe von den übrigen Daten werden (außer für $p = 1\,000$) in fast keinem Lauf überhaupt Bicluster detektiert, d. h. der Jaccardindex ist gleich 0 und es werden keine Hinweise auf die tatsächlich vorhandene Subgruppe aufgedeckt.
Betrifft die Subgruppe aber eine größere Variablenanzahl wie $p_{SG} = 20$ oder 50, so verbessert sich die Performanz des Biclusterns deutlich für alle (n, n_{SG})-Kom-



(a) FSJ: $p = 1\,000$ (b) FSJ: $p = 10\,000$ (c) FSJ: $p = 50\,000$

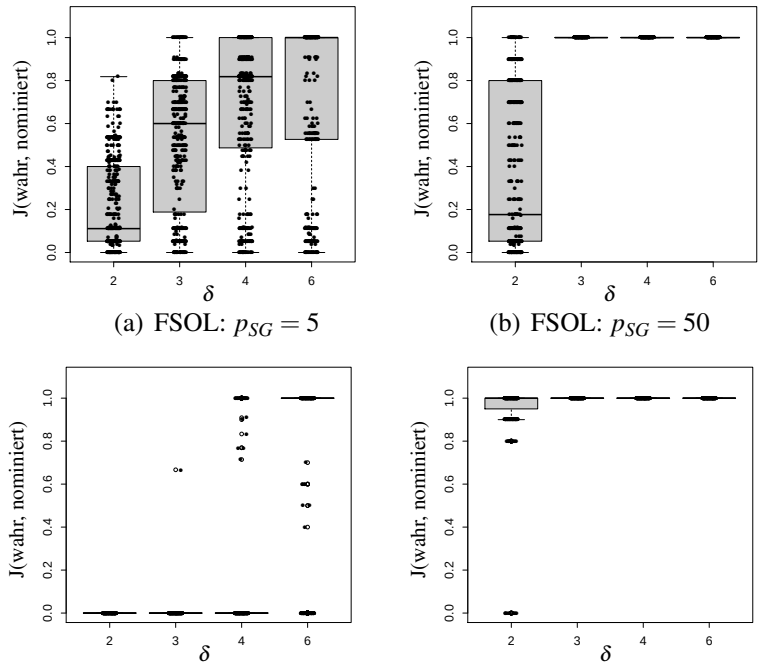
(d) BC: $p = 1\,000$ (e) BC: $p = 10\,000$ (f) BC: $p = 50\,000$

(g) FSBC: $p = 1\,000$ (h) FSBC: $p = 10\,000$ (i) FSBC: $p = 50\,000$

Abbildung 7: Einfluss der Featureanzahl p im Datensatz auf den erzielten Jaccardindex zum Vergleich der nominierten und wahren Subgruppe bei den Ansätzen FSJ, Biclustern und FSBC. Das Beispiel zeigt die Situation $n = 70$, $n_{SG} = 10$ für Lokationsshifts $\delta = 2,3,4,6$ in jeweils $L = 500$ simulierten Datensätzen mit $p = 1\,000, 10\,000, 50\,000$ Variablen. Die Abbildungen 35-38 im Anhang zeigen entsprechende Plots für die vier (n, n_{SG})-Kombinationen und die Methoden FSOL, FSJ, BC und FSBC. Aus Platzgründen wurde teilweise auf die Beschriftung der Ordinate verzichtet. Für alle Plots wird der Jaccardindex zum Vergleich der nominierten und wahren Subgruppe abgetragen.

binationen und BC überholt die neuen FSx-Workflows für $p_{SG} = 50$ (s. Abb. 8). In der SimMulti-Studie wird bewusst eine geringe Anzahl $p_{SG} = 5$ subgruppenanzeigender Variablen gewählt, um das Verhalten der Methoden in diesem interessierenden Bereich zu beurteilen. Die Abbildungen 39-42 im Anhang zeigen die entsprechenden SimMulti-Ergebnisse für die vier betrachteten Kombinationen aus Fallzahl $n = 40, 70$ und Subgruppengröße $n_{SG} = 5, 10$ und mit jeweils $p_{SG} = 5, 20, 50$ subgruppenanzeigenden Variablen.

Eine Besonderheit zeigt sich bei der Erhöhung von p_{SG} bei der FSBC-Methode: Im Gegensatz zu den anderen Methoden wächst die Performanz nicht generell mit der Anzahl der SG-anzeigenden Variablen. Bei der Erhöhung von p_{SG} von 5 auf 20 verbessert sich die Performanz für alle (n, n_{SG})-Settings. Im Vergleich dazu wird eine weitere Verbesserung bei $p_{SG} = 50$ wenn überhaupt für kleine bis mittlere Shifts erzielt. Für größeres δ bricht die Performanz zusammen und die Jaccardindizes fallen auf 0.



(a) FSOL: $p_{SG} = 5$ (b) FSOL: $p_{SG} = 50$

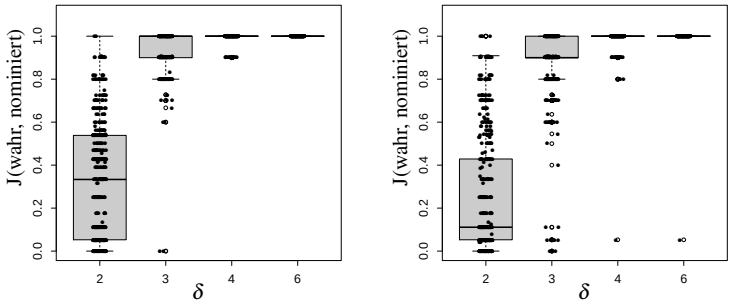
(c) BC: $p_{SG} = 5$ (d) BC: $p_{SG} = 50$

Abbildung 8: Einfluss der Anzahl p_{SG} der subgruppenanzeigenden Variablen auf die Methoden FSOL und Biclustern (BC) im Setting $n = 70$, $n_{SG} = 10$ der SimMulti-Studie. Gezeigt wird der erzielte Jaccardindex für Subgruppen mit Lokationsshifts $\delta = 2, 3, 4, 6$ aus jeweils $L = 500$ simulierten Datensätzen. Die Abbildungen 39-42 im Anhang zeigen entsprechende Plots für $p_{SG} = 5, 20, 50$ in den vier (n, n_{SG})-Kombinationen für die Methoden FSOL, FSJ, BC und FSBC.

Dieses Verhalten hängt mit der Anzahl T der vorselektierten Variablen zusammen. Für $p_{SG} = 50 = T$ weisen erwartungsgemäß bei ausreichend großem Shift alle Variablen der selektierten Teilmatrix in den Spalten, die die Subgruppe repräsentieren, hohe Werte auf. Somit hat die Subgruppe in der Teilmatrix nur Auswirkungen auf den Spalteneffekt, während sich kein Unterschied in den Zeilen ergibt. In der Simulation zeigt sich die sehr gute Performanz der FSBC-Methode daher wieder, sobald das zu schätzende Plaid-Modell entsprechend angepasst wird.
Dazu wird der Parameter `fit.model` der `biclust`-Funktion abweichend vom Default `y=m+a+b` auf `y=m+a` gesetzt. Für die praktische Anwendung wurde bereits die Betrachtung der empirischen Verteilung aller berechneten FS-Scores empfohlen, um eine angemessene Wahl für den Parameter T sicherzustellen. Zeigt sich eine hohe Anzahl von Variablen mit großen Scores, so sollte T ggf. erhöht werden. Allerdings ist zu bedenken, dass anders als im hier betrachteten Simulationsdesign in der Anwendung die Möglichkeit multipler und zumindest teilweise disjunkter Subgruppen besteht. Die beschriebene Änderung im anzupassenden Bicluster-Modell wäre in diesem Falle nicht notwendig, da dann wieder Zeilen- und Spalteneffekte vorliegen würden. Abbildung 43 des Anhangs zeigt die Effekte der angesprochenen Modifikationen bei FSBC für $p_{SG} = 50$ beispielhaft für zwei (n, n_{SG})-Settings.

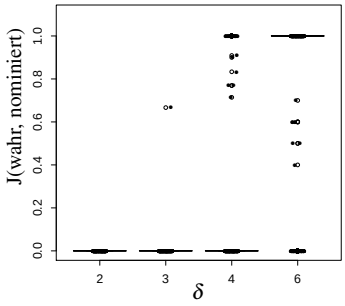
Einfluss der Sampleauswahl zur Ähnlichkeitsbeurteilung

Für jede der SG-Detektionsmethoden ist die generelle Entscheidung zu treffen, ob die Ähnlichkeit der Variablen nur basierend auf den Beobachtungen der als heterogen angenommenen Gruppe beurteilt werden soll oder über beide Gruppen hinweg. Abbildung 9 zeigt die Ergebnisse für FSJ und Biclustern für (n, n_{SG})=$(70, 10)$, während der Vergleich aller vier Methoden und Settings im Anhang in den Abbildungen 44-47 zu sehen ist.
Die Einschränkung der Samplemenge auf die heterogene Gruppe führt bei den Bicluster-basierten Methoden im Allgemeinen zu einer verbesserten Performanz. Abgesehen vom $(40, 10)$-Setting lassen sich durch Einschränkung des Datensatzes im Schnitt höhere Jaccardindizes erzielen. Diese Beobachtung deckt sich mit den Ergebnissen aus den vorherigen Abschnitten, die die Sensitivität des Biclusterns gegenüber Variationen von Datensatz- und Subgruppengröße (bzgl. der Variablenanzahl p_{SG}) zeigten. Bei FSOL führt die Einschränkung bei größerer Subgruppe ($n_{SG} = 10 [= max.rk]$) zu einer höheren Detektionsgüte, allerdings weniger ausgeprägt als beim Biclustern.
Eine Auffälligkeit zeigt sich bei FSJ in den beiden Settings mit $n = 40$ (in schwächerer Form auch bei FSOL für $(40, 5)$). Die sonst überlegene Methode zeigt in diesen Fällen bei ausschließlicher Berücksichtigung der heterogenen Gruppe weit unterlegene Performanz. Dies ist begründet in der Kombination der Default-

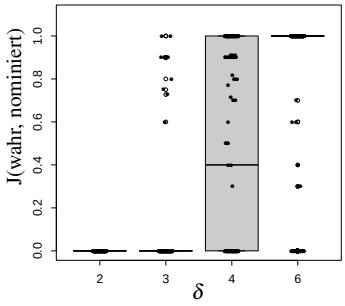


(a) FSJ: beide Gruppen

(b) FSJ: nur heterogene Gruppe



(c) BC: beide Gruppen



(d) BC: nur heterogene Gruppe

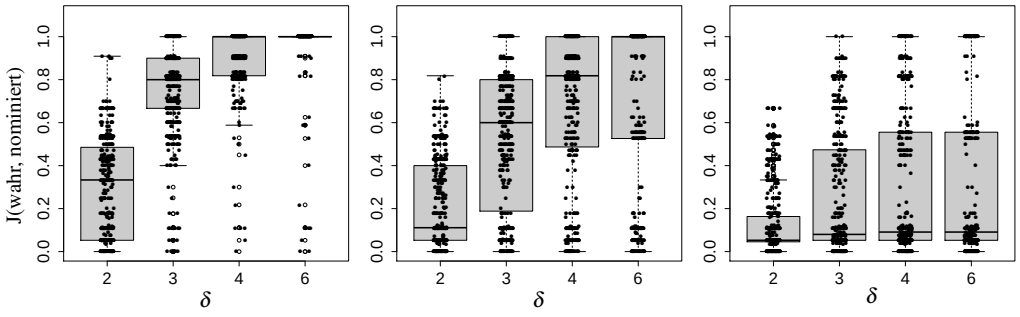
Abbildung 9: Einfluss der Sampleauswahl (nur heterogene Gruppe oder beide) auf die SG-Detektionsmethoden in der SimMulti-Studie. Gezeigt wird der erzielte Jaccardindex der Ansätze FSJ und Biclustern für Subgruppen mit Lokationsshifts $\delta = 2,3,4,6$ aus jeweils $L = 500$ Datensätzen in der Situation $(n, n_{SG}) = (70, 10)$. Die Abbildungen 44-47 im Anhang zeigen entsprechende Plots für die jeweils vier (n, n_{SG})-Kombinationen und Methoden.

einstellungen, die für diese (n, n_{SG})-Paare nicht geeignet sind. Zunächst werden beim Vergleich der Samples auf den top $max.rk = 10$ Rängen bereits durch Zufall größere Ähnlichkeiten zu erwarten sein, da immerhin jeweils ein Viertel der Samplemenge berücksichtigt wird. Dann führt die Verwendung des Default-cutoffs von $t_J = 0.3$ bei der Dichotomisierung dazu, dass ein Großteil der selektieren Variablen in einer einzigen Zusammenhangskomponente zusammengefasst wird. Die anschließende Auswahl der Komponente für die Nominierung einer Samplesubgruppe geschieht anhand des medianen FS-Scores der enthaltenen Variablen. Da in der Simulation nur $p_{SG} = 5$ Variablen auf die Subgruppe hinweisen und demnach einen höheren FS-Score erreichen sollten, kann diese Information bei der Medianbildung in großen Gruppen verloren gehen. Wird trotzdem die Komponente zur Nominierung herangezogen, die die informationstragenden Variablen enthält, so liegen bei kleinem p_{SG} und großer Variablengruppe die tatsächlichen Subgruppensamples nicht in einer ausreichend hohen Anzahl der Variablen auf den Toprängen, um letztendlich nominiert zu werden. Dies wäre allerdings bei ei-

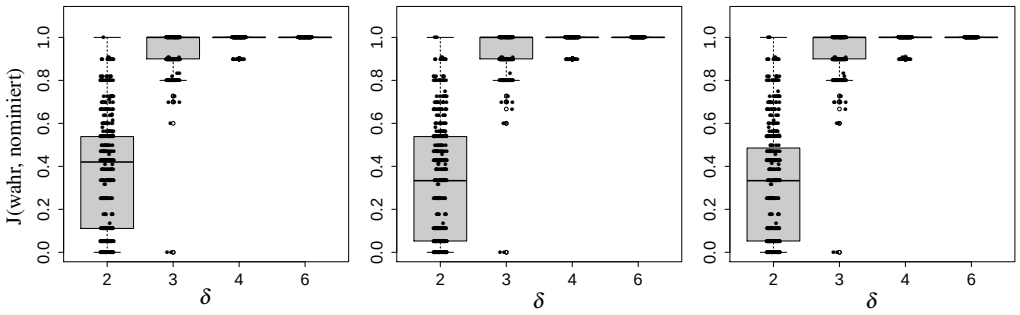
ner Anpassung des Parameters r_{min} möglich. In der praktischen Anwendung könnte der Schwellenwert t_J leicht angepasst werden, wenn ein „ungünstiges“ Variablengrouping zu beobachten ist. In der Simulationsstudie führt bei Einschränkung auf die heterogene Gruppe beispielsweise eine Änderung von $t_J = 0.3$ auf 0.35 für $n = 40$ dazu, die Performanz des Biclusterns zu übertreffen (vgl. Abb. 48).
In der Anwendung auf reale Daten wird grundsätzlich zur Berücksichtigung beider Gruppen geraten. So kann die gemeinsame Nominierung eines Samples aus der als homogen angenommenen Gruppe mit einer Menge von Samples aus der heterogenen Gruppe Hinweise auf eine möglicherweise fehlerhafte Gruppenzuordnung liefern. Falls nicht anders angegeben, wird daher in dieser Arbeit die Defaulteinstellung *beide Gruppen* verwendet.

Einfluss des FSx-Parameters *T*

Weiterhin wird der Einfluss der Anzahl T der Variablen untersucht, die basierend auf den berechneten FS-Scores für die weiteren Analysen in den FSx-Workflows und FSBC ausgewählt werden. Der Anhang zeigt in den Abbildungen 49-52 die Ergebnisse für die vier (n, n_{SG})-Settings für die Werte $T = 25, 50, 100$. Ein Auszug ist in Abbildung 10 dargestellt.
Generell zeigt sich die FSJ-Methode sehr robust gegenüber Variation des Parameters T. Etwas mehr Einfluss hat der Parameter auf FSBC, allerdings unterscheiden sich die Effekte zwischen den (n, n_{SG})-Settings. Während für die $(70, 5)$-Kombination die Performanz mit T abnimmt, scheint für die anderen drei Settings $T = 50$ die beste Wahl zu sein. Die Variation von T hat im $(40, 10)$-Setting kaum Auswirkungen, für $(40, 5)$ und $(70, 10)$ zeigen sich die Unterschiede vor allem für den Shift $\delta = 3$.
FSOL reagiert am deutlichsten auf die Änderung von T. Für $n = 40$ fällt der Performanzverlust bei der Erhöhung von T geringer aus als für die größere Fallzahl. In allen vier Settings zeigt sich beim Wechsel von $T = 50$ auf $T = 100$ eine auffällige Verringerung der Jaccardindizes. Den Extremfall stellt dabei das Setting $(70, 5)$ dar, in dem die Performanz zusammenbricht.
Im Folgenden wird anhand eines Beispiellaufs erläutert, an welcher Stelle des Workflows Probleme auftauchen. Ausgewählt wurde ein exemplarischer Lauf mit Shift $\delta = 6$, in dem alle $p_{SG} = 5$ Variablen deutlich die gesuchte Subgruppe aus $n_{SG} = 5$ Variablen anzeigen. Alle Variablen erhalten einen hohen FS-Score, sind in der selektierten Teilmatrix der top-100-FS-Variablen enthalten und bei der Betrachtung der zugehörigen Heatmap ist die wahre Subgruppe aus fünf Variablen und fünf Samples klar zu erkennen (s. Anhang, Abb. 53(a)). Schritt 1 des Workflows funktioniert demnach wie gewünscht. Die `igraph`-Darstellung des Gruppierungsschritts zeigt, dass die fünf interessierenden Variablen gemeinsam in einer Zusammenhangskomponente aus insgesamt 45 Variablen liegen (Abb. 53(b)).



(a) FSOL: $T = 25$ (b) FSOL: $T = 50$ (c) FSOL: $T = 100$



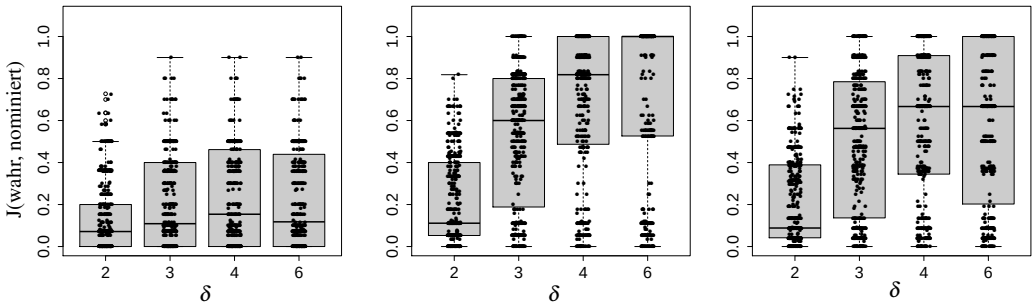
(d) FSJ: $T = 25$ (e) FSJ: $T = 50$ (f) FSJ: $T = 100$

Abbildung 10: Einfluss des Parameters T zur Auswahl der Anzahl FS-selektierter Variablen auf die Detektionsgüte der multivariaten Methoden gemessen am erreichten Jaccardindex. Für verschiedene Lokationsshifts $\delta = 2,3,4,6$ und $T = 25,50,100$ werden jeweils 500 Datensätze simuliert. Hier werden die Ergebnisse beispielhaft für zwei Methoden und das Setting $(n, n_{SG}) = (70,10)$ gezeigt. Die Gesamtergebnisse finden sich im Anhang (Abb. 49-52).

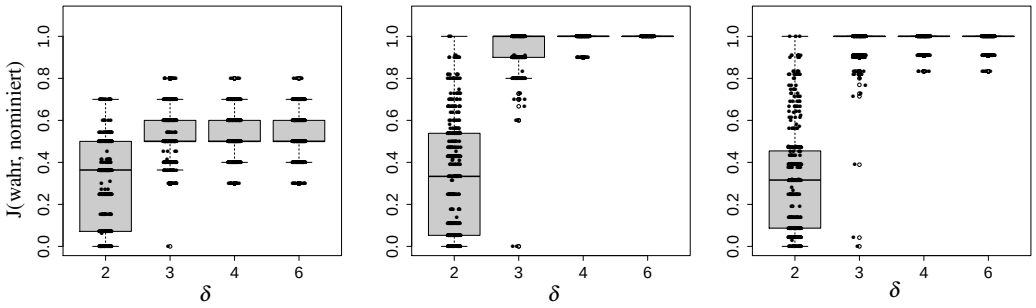
Damit können sie nicht zu einer Erhöhung des medianen FS-Scores der Gruppe beitragen, die Variablengruppe erreicht nur Rang 16. Eine strengere cut-off-Wahl von $t_{OL} = 0.005$ unterteilt die Komponente in kleinere Variablengruppen. Für die neue Aufteilung belegt eine der interessierenden Variablen allein Rang 1, die übrigen vier Variablen finden sich in der Gruppe auf Rang 3.
Ähnlich wie bei der Sampleauswahl zur Variablengruppierung wird die Defaulteinstellung für T nicht ausschließlich auf Grundlage der Simulationsergebnisse festgelegt. SimMulti zufolge wäre bereits die Auswahl von $T = 25$ Variablen für die nachfolgende Analyse ausreichend. In der Anwendung auf realen Daten könnten allerdings z. B. im Falle mehrerer Subgruppen deutlich mehr Variablen subgruppenrelevante Informationen enthalten als hier mit $p_{SG} = 5$. Da sich für $T = 50$ noch keine Reduktion der Detektionsgüte erkennen lässt, wird in den übrigen Simulationen dieser Wert als Default gewählt. Für praktische Anwendungen wird grundsätzlich die Betrachtung der empirischen Verteilung der FS-Scores empfohlen, sodass T gegebenenfalls angepasst werden kann. Auch die Visualisierung der Zwischenergebnisse der einzelnen Schritte sollte stets berücksichtigt werden.

Einfluss des FSx-Parameters *max.rk*

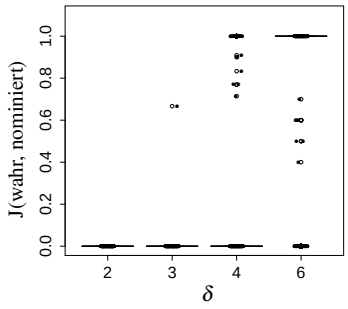
Zuletzt wird der Parameter *max.rk* der FSx-Methoden untersucht. Bei den betrachteten Settings wird fast immer die jeweils beste Performanz erreicht, wenn der gewählte Parameterwert mit der wahren Subgruppengröße n_{SG} übereinstimmt. In diesen Fällen nähern sich die Jaccardindizes mit wachsendem δ an $J = 1$. Eine Ausnahme zeigt sich im Setting $(n, n_{SG}) = (70, 5)$. FSJ verliert im mittleren Shiftbereich an Performanz, wenn der Parameter *max.rk* von $n_{SG} = 5$ auf 10 erhöht wird und fällt bei weiterer Erhöhung auf 15 auf Werte, die im Schnitt unter 0.4 liegen. FSOL hingegen zeigt hier bei Übereinstimmung der Subgruppengröße und eingestelltem Parameterwert vergleichsweise schlechte Jaccardindizes, erreicht bei einer Erhöhung auf $max.rk = 10, 15$ aber eine deutliche Verbesserung.
Es sei angemerkt, dass auch für Kombinationen, bei denen der Parameterwert *max.rk* nicht der wahren Größe n_{SG} entspricht, die Detektionsgüte der FSx-Methoden besonders für kleine bis moderate Shifts δ trotzdem höher ist als die des Biclusterns. Die detaillierten Ergebnisse sind den Abbildungen 54-57 zu entnehmen, ein Beispielsetting ist in Abbildung 11 gezeigt.
Der Defaultwert wird als $max.rk = 10$ gewählt, da der durchschnittliche Verlust bei der Suche nach kleinen Subgruppen so gering gehalten wird. Bei Studien mit sehr großer Fallzahl, in denen auch anteilig „kleine“ Subgruppen größer als $n_{SG} = 10$ sein können, sollte unter Umständen *max.rk* entsprechend erhöht werden. Analog wird ein kleinerer Wert für kleine Gruppen empfohlen, insbesondere falls für die Ähnlichkeitsbetrachtungen nur eine Gruppe herangezogen wird (s.o.).



(a) FSOL: $max.rk = 5$ (b) FSOL: $max.rk = 10$ (c) FSOL: $max.rk = 15$



(d) FSJ: $max.rk = 5$ (e) FSJ: $max.rk = 10$ (f) FSJ: $max.rk = 15$



(g) BC

Abbildung 11: Einfluss des Parameters *max.rk* bei der Anwendung von FSOL und FSJ im Setting $(n, n_{,SG}) = (70, 10)$. Gezeigt wird der erzielte Jaccardindex der Ansätze bei unterschiedlichen *max.rk*-Werten und zum Vergleich für das Biclustern. Details zur Darstellung sind in der Beschreibung zu Abbildung 7 zu finden.

5.4.2 Vergleich der vier Methoden FSOL, FSJ, BC und FSBC bei Verwendung der Standardparameter

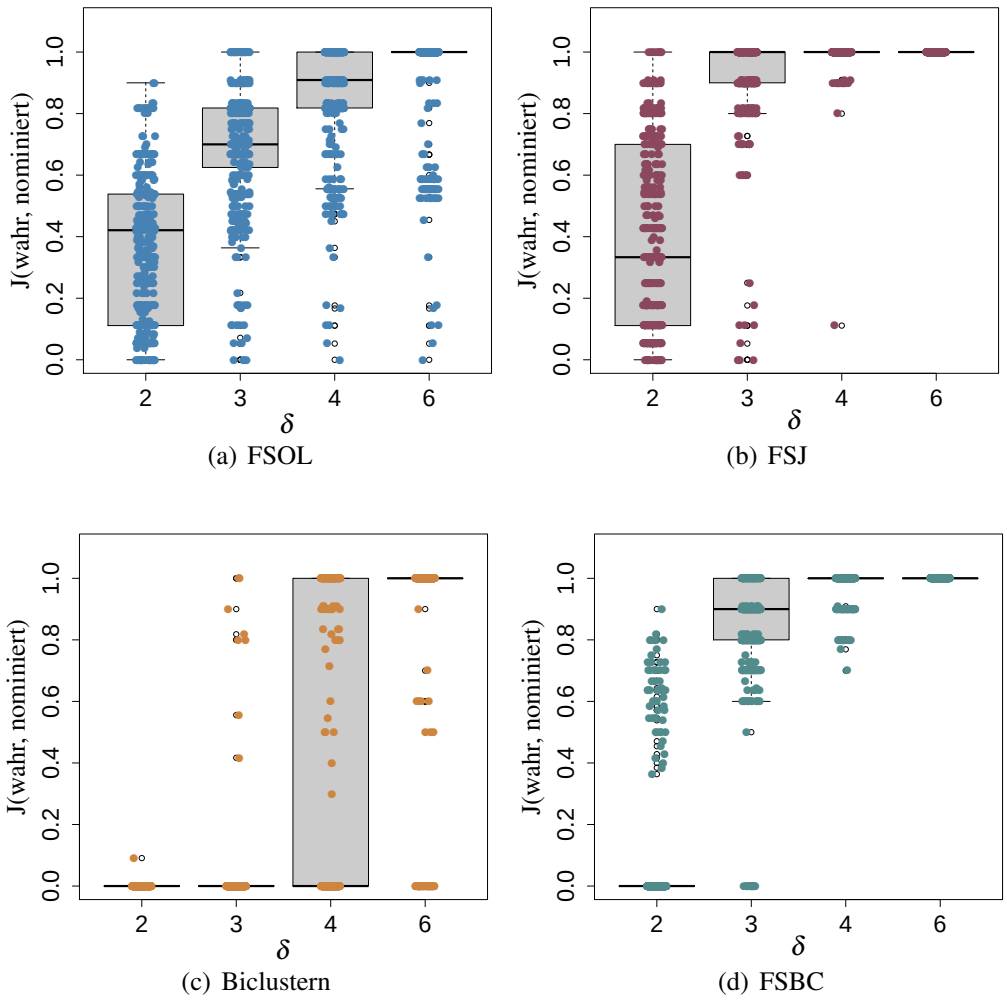
Basierend auf den vorhergehenden Sensitivitätsanalysen werden die SimMulti-Ergebnisse nun mit Fokus auf folgende Parameterwerte besprochen:

p	p_{SG}	heterOnly	T	$max.rk$
1000	5	FALSE	50	10

.

Die zugehörigen Boxplots der von den vier Methoden erreichten Jaccardindizes sind in den Abbildungen 12 und 13 noch einmal im direkten Vergleich für die (n, n_{SG})-Settings $(40, 10)$ und $(70, 5)$ zu sehen. Dabei handelt es sich gemessen am Subgruppenanteil um das günstigste bzw. um das ungünstigste Setting. Die Methoden FSOL, BC und FSBC zeigen bei $(40, 10)$ ihre jeweils beste Performanz. FSJ erreicht die besten Ergebnisse in beiden Settings mit $n_{SG} = 10$. Die entsprechenden Darstellungen für die beiden übrigen Settings finden sich im Anhang (Abb. 58 und 59).

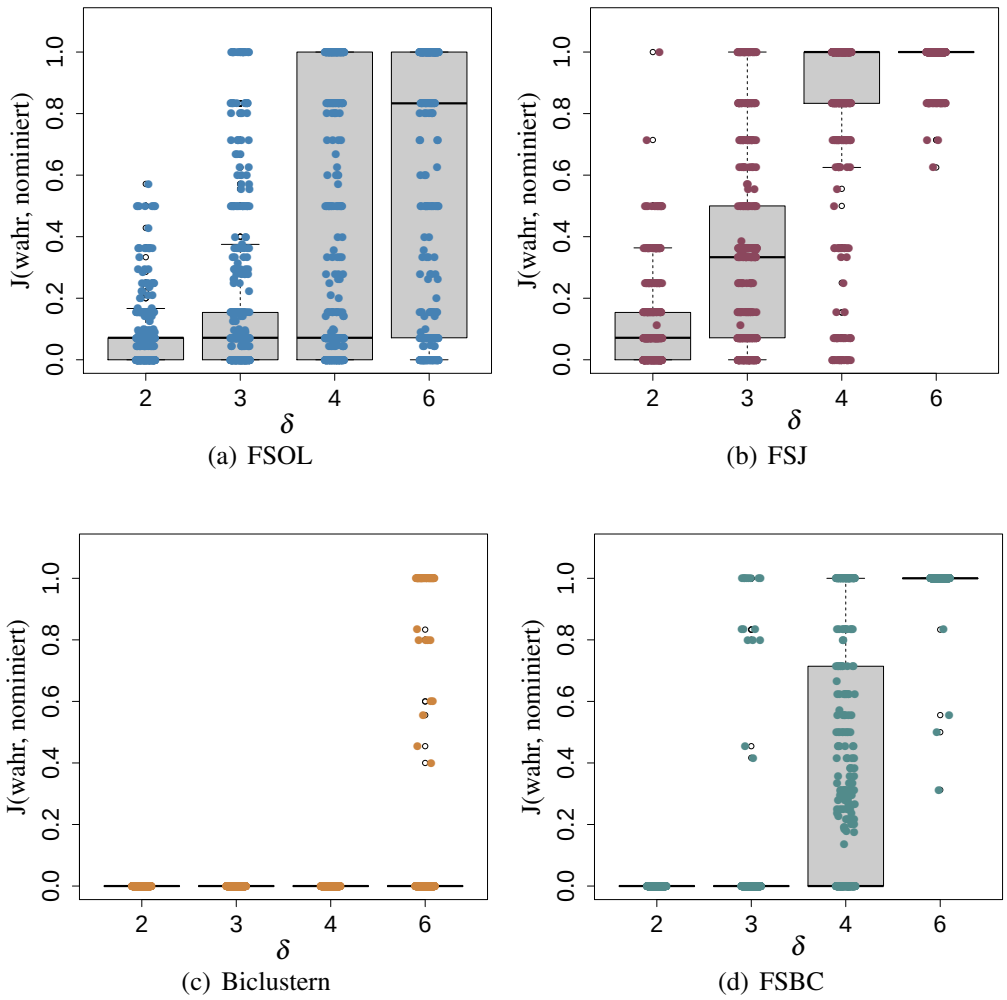
Für alle betrachteten (n, n_{SG})-Kombinationen zeigt sich das Biclustern als ungeeignet für die Detektion von Subgruppen mit kleinen bzw. moderaten Shifts ($\delta = 2, 3$). Es wird zumeist kein Bicluster gefunden, somit ist $J = 0$. Falls ein Bicluster detektiert wird, ist mindestens ein Sample aus der wahren Subgruppe enthalten. Die besten Ergebnisse werden für große Subgruppenanteile erreicht: Im $(40, 10)$-Setting kann für $\delta = 4$ schon in einem beträchtlichen Anteil der Läufe die Subgruppe recht zuverlässig bestimmt werden (rund 41% der Läufe mit $J > 0.8$). Für $\delta = 6$ liegt der Jaccardindex für 75% der Läufe bei $J = 1$. Letzteres gilt auch für das Setting $(n = 70, n_{SG} = 10)$ mit Shift $\delta = 6$.
Durch die Anwendung des Biclustern auf die Teilmatrix der top-50-FS-Variablen (FSBC) wird eine große Performanzsteigerung im Vergleich zur ursprünglichen Variante erreicht, insbesondere im Bereich moderater Shifts. Außer für $(n, n_{SG}) = (70, 5)$ liegen bei FSBC bereits bei $\delta = 3$ die zentralen 50 Prozent der Jaccardindizes deutlich im positiven Bereich. Wie beim klassischen Biclusteransatz zeigen sich die höchsten Werte bei $(n, n_{SG}) = (40, 10)$. Ausschließlich in dieser Konstellation werden deutlichere Hinweise auf die enthaltene Subgruppe auch schon in etwa 11% der Läufe mit kleinem Shift $\delta = 2$ gefunden.
Außer im $(40, 5)$-Setting erzielt FSJ höhere Jaccardindizes als FSBC, vor allem bei kleinen Shifts $\delta = 2$. Für $(n, n_{SG}) = (40, 5)$ ist die Shiftgröße ausschlaggebend: für $\delta = 2, 3$ ist FSJ überlegen, für größere Shifts hingegen FSBC.



(a) FSOL (b) FSJ

(c) Biclustern (d) FSBC

Abbildung 12: Vergleich der vier Methoden FSOL, FSJ, BC und FSBC für $(n, n_{SG}) = (40, 10)$ für eine fest gewählte Parameterkombination. Details sind dem Text zu entnehmen.



(a) FSOL (b) FSJ

(c) Biclustern (d) FSBC

Abbildung 13: Vergleich der vier Methoden FSOL, FSJ, BC und FSBC für $(n, n_{SG}) = (70, 5)$ für eine fest gewählte Parameterkombination. Details sind dem Text zu entnehmen.

6 Anwendung auf reale Datensätze

Für einen umfassenden Vergleich der verschiedenen Methoden zur Identifikation von Patientensubgruppen ist zusätzlich zu den bereits beschriebenen Simulationsstudien auch ein Vergleich anhand von realen Daten erforderlich. Dieses Kapitel behandelt beispielhaft drei Datensätze aus unterschiedlichen omics-Technologien. Das erste Beispiel zeigt die Analyse eines Teildatensatzes aus dem ParkCHIP-Projekt, das unter anderem eine Studie mit Antikörpermicroarrays beinhaltet. Die Daten dienten in Ahrens et al. [22] dem Vergleich univariater Methoden zur Identifikation subgruppenanzeigender Variablen und der Entwicklung des Fisher-Sum-Scores. Da für dieses Beispiel keine wahre Patientensubgruppe bekannt ist, werden nur die wesentlichen Ergebnisse der genannten Arbeit dargestellt. Im Gegensatz dazu werden ausführliche Analysen u. a. mittels des neuen FSOL-Workflows an zwei anderen Datensätzen durchgeführt, die durch die Kenntnis einer definierten Subgruppe für die Beurteilung der multivariaten SG-Detektionsverfahren insgesamt besser geeignet sind.
Der im Folgenden mit ALL bezeichnete Datensatz enthält Genexpressionsdaten von Patienten, die an akuter lymphatischer Leukämie (ALL) erkrankt sind. Die zusätzlich verfügbaren Kovariablen (Metadaten) über die molekulare Biologie der Tumore ermöglichen dabei die Konstruktion eines Zwei-Gruppen-Vergleiches, bei dem eine der Gruppen homogen ist, während die andere eine kleine Gruppe eines abweichenden Tumortyps enthält. Der hier verwendete Datensatz wurde bereits in Ahrens et al. [55] als Anwendungsbeispiel für den FSOL-Workflow gezeigt. Die im Folgenden gezeigten Analysen gehen für das Biclustern deutlich über die bisherige Darstellung hinaus. Weiterhin werden erstmals die beiden Methoden FSJ und FSBC eingesetzt.
Anschließend werden die Ergebnisse der multivariaten SG-Detektionsverfahren für den DeNoPa-Datensatz vorgestellt und verglichen. Für diesen Proteomikdatensatz wurde eine Samplesubgruppe auf Basis von ELISA-Messungen definiert, d. h. mit einer von der Massenspektrometrie unabhängigen Methode. Für die ALL- und DeNoPa-Daten gilt es jeweils zu überprüfen, ob die verschiedenen Verfahren Hinweise auf die definierten Subgruppen liefern.
Auf den ALL- und den DeNoPa-Datensatz werden jeweils die vier Methoden FSOL, FSJ, FSBC und BC angewendet. Die Darstellung der Ergebnisse der FSx-Workflows FSOL und FSJ unterscheidet sich dabei von der für die Bicluster-basierten Verfahren FSBC und BC. Während der FSJ-Workflow deterministischer Natur ist und die zufällige Komponente in FSOL nur geringen Einfluss auf das Ergebnis zeigt, ist beim Biclustern eine mitunter erhebliche Variation der Ergebnisse in Abhängigkeit der Startwerte im Optimierungsprozess zu beobachten. Für beide Datensätze wird zunächst jeweils das Ergebnis eines Laufes für die FSx-Ansätze ausführlicher diskutiert. Für FSBC und BC werden hingegen die Ergebnisse aus

jeweils 1 000 Läufen mit unterschiedlichen Startwerten deskriptiv analysiert. Aufgrund der Ähnlichkeit der Ergebnisse hinsichtlich der Subgruppendetektion zwischen FSBC und den FSx-Ansätzen werden nach den FSx-Ergebnissen als erstes die FSBC-Analysen gezeigt und abschließend die der Referenzmethode. Die für die Bicluster-basierten Methoden angegebenen Laufzeiten wurden gemessen auf einem System mit Intel® Core™i7-2600 CPU @ 3.40GHz, 8GB RAM.

6.1 ParkCHIP

ParkCHIP bezeichnet ein bereits abgeschlossenes Verbundprojekt, das vom Ministerium für Innovation, Wissenschaft und Forschung des Landes NRW gefördert wurde (ParkCHIP, FZ 280381102). Ziel des Projekts, das am Medizinischen Proteom-Center der Ruhr-Universität Bochum koordiniert wurde, war die Identifikation diagnostischer Biomarker für die Parkinsonerkrankung mithilfe von Proteinmicroarrays. Auf dem verwendeten ProtoArray® v5.0 der Firma Life Technologies (Carlsbad, CA, USA) waren rund 9 500 Proteinspots vorhanden. Das Ziel war die spezifische Detektion von Autoantikörpern aus dem Blutserum. Von den insgesamt drei Versuchsgruppen werden im Folgenden nur die Gesundkontrollen und die Parkinsonerkrankten (kurz PD für *Parkinson's disease*) berücksichtigt. Die Gruppen bestehen aus jeweils 72 unabhängigen Serumproben, die bezüglich Alter und Geschlecht gematcht wurden. Dieser Teildatensatz ist online über die Homepage des Medizinischen Proteom-Centers verfügbar[1]. Da PD als heterogene Krankheit bekannt ist (vgl. z. B. [64, 65]), waren die Ergebnisse bzgl. subgruppenrelevanter Marker auf der neuen Ebene von Autoantikörpern von besonderem Interesse.
Die ParkCHIP-Daten liegen quantilnormalisiert vor und werden sowohl für die Analysen als auch für die graphische Darstellung $\log_2$-transformiert. Abbildung 14 zeigt den Scatterplot der ersten beiden Hauptkomponenten. Die Samples teilen sich scheinbar in Richtung der ersten Hauptkomponente in zwei Gruppen auf. Die Ermittlung des Ursprungs einer solchen Trennung ist häufig nur unter Verwendung von Metadaten möglich (Daten hier nicht gezeigt). Die Vermutung, dass es sich um einen durch die beiden verwendeten Produktionschargen der Arrays verursachten Effekt handeln könnte, bestätigte sich nicht. Stattdessen besteht offenbar ein Zusammenhang mit den Tagen, an denen die Arrays gescannt wurden. Eine mögliche Erklärung sind somit Reaktionen des empfindlichen Messsystems auf Schwankungen in den äußeren Bedingungen wie z. B. der Luftfeuchtigkeit.
Der interessierende Zwei-Gruppen-Vergleich zwischen PD-Erkrankten und Gesundkontrollen wird an dieser Stelle ohne zusätzliche Adjustierung analysiert, da zum einen kein Zusammenhang mit den experimentellen Gruppen zu bestehen

[1] `http://www.medizinisches-proteom-center.de/Ahrens_et_al/index.html`

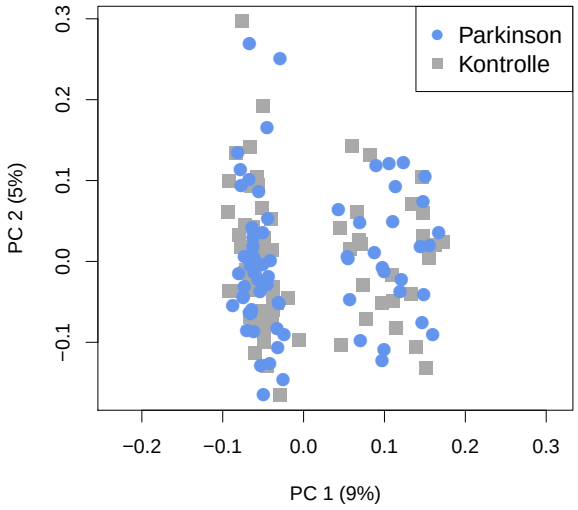


Abbildung 14: Scatterplot der ersten beiden Hauptkomponenten für den ParkCHIP-Datensatz. In Klammern angegeben ist der Anteil der durch die jeweilige Hauptkomponente erklärten Varianz. In blauen Kreisen ist die Gruppe der Parkinsonpatienten dargestellt, die Gruppe der Gesundkontrollen in grauen Quadraten.

scheint und zum anderen solch detaillierte Metadaten in der Praxis nur selten verfügbar sind. Die Anwendung der univariaten Scoringmethoden zum Auffinden subgruppenanzeigender Variablen lässt trotz des beobachteten Effekts interessante Ergebnisse erwarten.

6.1.1 Ergebnisse ParkCHIP

In Ahrens et al. [22] wurden die univariaten Methoden, die bereits im Rahmen der SimUni-Studie besprochen wurden, zusätzlich auf den ParkCHIP-Datensatz angewendet. Dabei handelt es sich um FS, ORT, OS, den t-Test, PADGE, Kurtosis im Sinne von PAK, sowie um Bartletts Test auf Varianzhomogenität. Während in der durchgeführten Simulationsstudie nur drei definierte Verteilungsmuster auftreten (keine Subgruppe, krankheitsspezifische (ks) oder nicht-ks Subgruppe), findet sich in echten Daten typischerweise eine größere Anzahl solcher Muster. Ziel der im Folgenden zusammengefassten Analyse war die Charakterisierung der von den Methoden bevorzugt gefundenen Verteilungsmuster. Dazu wurden alle Variablen des Datensatzes entsprechend der jeweils untersuchten Methode gerankt und die ($\log_2$)-Intensitätsprofile der besten 15 Kandidaten beurteilt.
Abschließend wurde speziell für die neue, favorisierte FS-Methode die Annotation der bekannten biologischen Funktionen der Variablen auf den Toprängen be-

trachtet, um eine Assoziation der Variablen mit der Parkinsonerkrankung zu untersuchen. Der Vergleich der p-Werte und Ränge gemäß des t-Tests mit den FS-Ergebnissen einiger Kandidaten verdeutlicht den Unterschied zwischen einer gewöhnlichen differentiellen Analyse und der spezifischen Subgruppenanalyse.

Charakterisierung der bevorzugten Verteilungsmuster

Die Intensitätsplots der jeweils besten 15 Kandidaten der verglichenen Methoden finden sich im Anhang in den Abbildungen 28-34. Die größte Anzahl von Variablen mit dem interessierenden Verteilungsmuster einer krankheitsspezifischen (ks) Subgruppe findet sich unter den Topkandidaten von FS und PADGE. Die von FS ausgewählten Variablen zeigen im Vergleich zu denen von PADGE durchgehend breitere Verteilungen, die einen größeren Intensitätsbereich überspannen. Die Begründung liegt in der FS-Defaulteinstellung, nach der die einzelnen Variablen nicht skaliert werden, um Variablen mit absolut größeren Abweichungen einen höheren Score zuzuweisen. Während PADGE in den Simulationsstudien im Vergleich zu den übrigen Methoden nicht kompetitiv schien, entsprechen die empirischen Verteilungen der Variablen auf den Toprängen zu einem großen Teil dem gewünschten Muster. Im Gegensatz dazu zeigen die top OS bzw. top ORT Variablen des ParkCHIP-Datensatzes überwiegend nicht-ks Subgruppen an. Durch die der Scoreberechnung vorgeschalteten Skalierung werden im Gegensatz zu FS auch Variablen mit sehr schmalen Verteilungen und nur relativ betrachtet größeren Abweichungen der Subgruppe gute Scores zugewiesen.
Keine der 15 Variablen mit kleinstem p-Wert im t-Test weist einen deutlich erkennbaren homogenen Lokationsunterschied zwischen den beiden verglichenen Gruppen auf. Dies legt nahe, dass Variablen mit diesem Verteilungstyp nicht in den Daten enthalten sind, da der t-Test gemäß der SimUni-Studie solche Variablen besser detektiert als Variablen mit nur einer Teilmenge hochregulierter Samples in einer der Gruppen, wie sie hier auftreten. Sowohl die Kurtosis als auch der Bartletttest wählen vorrangig Variablen mit sehr schmalen Grundverteilungen und einzelnen deutlichen „Ausreißer“-Werten aus. Die Variablen mit einer Gruppe erkennbar hoch-regulierter Samples in der Parkinsongruppe tauchen ebenfalls in den Topkandidaten von OS und/oder FS auf.

Bekannte biologische Funktionen der top FS Variablen

Tabelle 7 zeigt die Fisher-Sum-Scores und p-Werte des t-Tests sowie die jeweiligen Ränge der Variablen mit höchsten FS-Scores. Für die Literaturrecherche wurden die einzelnen Variablen zunächst mit den verfügbaren Informationen über die zugehörigen Gene annotiert. Einige der Gene auf den Toprängen wurden bereits ausführlich im Zusammenhang mit neurodegenerativen Erkrankungen gene-

rell oder mit Parkinson im Speziellen untersucht. Zunächst fällt auf, dass die ersten beiden Positionen von zwei Variablen belegt werden, die beide mit dem Gen Paralemmin-2 annotiert sind. Im t-Test-Ranking liegen zwischen den beiden Variablen dagegen etwa 100 Plätze. Ein Zusammenhang dieses Gens mit Parkinson ist nicht unwahrscheinlich: Paralemmin (PALM) interagiert nach Basile et al. [66] mit Dopaminrezeptoren, die unter anderem bei Parkinson von Bedeutung sind. Ähnliches könnte auch für die hier gefundene Form PALM2 zutreffen.
Auf Position 3 des FS-Rankings liegt MTHFR (Methylentetrahydrofolat-Reduktase). Im Jahre 2005 konnten de Lau et al. [67] ein stark erhöhtes PD-Risiko in Rauchern mit einem bestimmten MTHFR-Genotyp zeigen. Auf den oberen FS-Rängen finden sich außerdem mit z. B. CALB2 oder CAMK4 einige Gene, die auf Prozesse der Calciumregulation hinweisen. Deren Zusammenhänge mit PD wurden vor mehreren Jahren (z. B. [68, 69]) aber auch aktueller von Schapira [70] oder Calì et al. [71] diskutiert.
Insgesamt scheint ein Zusammenhang der Variablen auf den FS-Toprängen mit potentiellen Subtypen der Parkinsonerkrankung durchaus möglich. Das Auffinden dieser bekanntermaßen assoziierten Variablen lässt sich vorsichtig als „proof of principle“ für die Relevanz der FS-Ergebnisse interpretieren. In diesem Falle sollte gerade die bislang nicht mit PD assoziierten Kandidaten in weiteren Analysen betrachtet werden, da diese zu neuen Erkenntnissen über mögliche PD-Subtypen führen könnten.

FS-Rang	FS-Score	t-Test-Rang	p-Wert t-Test	Genname
1	25.67	141	0.027	PALM2
2	25.43	42	0.010	PALM2
3	24.94	217	0.037	MTHFR
...				
13	18.69	636	0.083	CALB2
17	17.50	336	0.053	CAMK4

Tabelle 7: Vergleich der Rankings gemäß Fisher Sum und t-Test auf den ParkCHIP-Daten für einige FS-Topkandidaten. Angegeben ist auch der Name des annotierten Gens (Abkürzung) zur Beurteilung einer möglichen Assoziation mit der Parkinsonerkrankung.

6.2 ALL

Der Datensatz ALL ist über das gleichnamige **R**-Paket `ALL` öffentlich verfügbar [25] und enthält Genexpressionsdaten von 128 Affymetrixchips (Affymetrix, Inc., Santa Clara, California, U.S.). Genauer handelt es sich um RMA-normalisierte Werte von 12 625 Features, sogenannter *(Affymetrix) probe sets*, gemessen mithilfe des Chip-Typs `hgu95av2`. Informationen zur Probenaufbereitung sowie Analyseergebnisse von Teildatensätzen sind nachzulesen in Chiaretti et al. (2004 [24], 2005 [72]). Untersucht wurden Patienten mit akuter lymphatischer Leukämie (ALL). Von klinischer Relevanz ist dabei die Unterscheidung in B- bzw. T-Zell-Typ, d. h. die Unterscheidung gemäß der Zelllinie, in der sich die maligne Transformation manifestiert. Die Gruppe der B-ALL-Fälle lässt sich entsprechend chromosomaler Veränderungen weiter unterteilen, im Folgenden relevant sind die beiden Translokationen E2A/PBX1 und BCR/ABL. Letztere beschreibt eine Veränderung auf dem sogenannten Philadelphia-Chromosom.
Das `ALL`-Paket enthält neben den Genexpressionsdaten auch Angaben zu Alter, Geschlecht, Zelltyp und Translokation sowie weitere Informationen zum Krankheitsverlauf. Die verfügbaren Kovariablen ermöglichen die Konstruktion spezifischer Vergleiche, sodass beispielsweise die Güte der verschiedenen Subgruppendetektionsmethoden beurteilt werden kann. Für die hier dargestellte Auswertung wird der Datensatz zunächst auf den B-Zell-Typen eingeschränkt. Entsprechend der Variable `mol.biol`, d. h. der Molekularbiologie des Tumors, ist für diesen Typen eine weitere Unterteilung der Proben möglich (siehe Tabelle 8). Der klinische Verlauf der Krankheit kann sich zwischen diesen einzelnen Subtypen bedeutend unterscheiden. Während der ALL-Typ mit dem E2A/PBX1-Fusionsgen relativ gut behandelbar ist, galt die Prognose für Patienten mit dem BCR/ABL-Fusionstranskript lange Zeit als eher ungünstig. Erst in den letzten Jahren konnte die Prognose durch den Einsatz von *Imatinib* und *Dasatinib*, sogenannten Tyrosinkinaseinhibitoren, verbessert werden [73].
Für die Beurteilung der SG-Detektionsverfahren werden zwei ausreichend große Gruppen benötigt; eine davon homogen, eine mit einer bekannten Subgruppe von Samples. Die gewählte heterogene Gruppe besteht mit 37 Samples im Wesentlichen aus der BCR/ABL-Gruppe und wird ergänzt durch den Subtyp E2A/PBX1 ($n_{SG} = 5$). Als homogene Gruppe wird die Gruppe NEG aus 42 Samples gewählt, sodass die beiden zu vergleichenden Gruppen vom gleichen Umfang sind. Alle gezeigten Auswertungen beziehen sich auf diesen Teildatensatz, d. h. auf den Vergleich

42 NEG vs. (37 BCR/ABL und 5 E2A/PBX1).

Die Gruppe NEG ist dadurch gekennzeichnet, dass keine bekannten Mutationen vorliegen. Sie könnte theoretisch unbekannte Mutationen enthalten, auf die folg-

	Ausprägungen der Kovariable `mol.biol`					
	ALL1/AF4	**BCR/ABL**	**E2A/PBX1**	**NEG**	NUP-98	p15/p16
B-Zell	10	**37**	**5**	**42**	0	1
T-Zell	0	0	0	32	1	0

Tabelle 8: Verteilung der Kovariable `mol.biol` über die Molekularbiologie der Samples im ALL-Datensatz. Hervorgehoben sind die Gruppen, die in der hier vorgestellten Analyse verwendet werden.

lich bislang in der klinischen Routine nicht getestet wird. Letztendlich könnte diese Gruppe sogar heterogener sein als die durch ein Fusionstranskript charakterisierte BCR/ABL-Gruppe. Im Laufe der Auswertung könnte sich zeigen, ob es sich bei der Homogenität der Gruppe NEG um eine gerechtfertigte Annahme handelt. Selbst falls dies nicht der Fall sein sollte, sollte die Anwendung der Subgruppendetektionsverfahren trotzdem nützlich sein, um Hinweise auf die interessierende Subgruppe E2A/PBX1 zu finden.
Die gesuchten Effekte könnten für die FSx-Workflows allenfalls dadurch maskiert werden, dass eine der gegebenenfalls vorhandenen Subgruppen in NEG bei ausreichender Sampleanzahl erhöhte Expression in denselben Variablen zeigt wie die E2A/PBX1-Samples. Dann könnte die Subgruppe im Selektionsschritt als nichtkrankheitsspezifisch gewertet werden und so in der weiteren Subgruppendetektion nicht berücksichtigt werden. Die FSx-Workflows suchen in einer vorgegebenen experimentellen Gruppe (hier BCR/ABL und E2A/PBX1) nach hochregulierten Subgruppen. Dadurch wären NEG-Subgruppen für die SG-Detektion nur hinderlich, sollten die betroffenen Variablen auch mit der E2A/PBX1-Subgruppe assoziiert sein. Trotz der unter Umständen nicht gerechtfertigten Annahme wird für die Beschreibung der Auswertung an der bislang verwendeten Terminologie *homogen* für die Referenzgruppe festgehalten. Folglich wird die Gruppe, die die gesuchte Subgruppe enthält, als *heterogen* bezeichnet.
Die Annotation der *probe sets* hinsichtlich der assoziierten Gene wurde mithilfe des über Bioconductor verfügbaren **R**-Pakets `hgu95av2.db` durchgeführt ([74], Paketversion 3.2.3, Entrez Gene Version vom 27.09.2015, Bioconductor Version 3.3, **R**-Version 3.3.1).

Der Scatterplot der ersten beiden Hauptkomponenten für den ALL-Datensatz ist in Abbildung 15 zu sehen. Auf dieser globalen Ebene zeigt sich keine Trennung der Samples gemäß der wahren Tumorsubtypen.

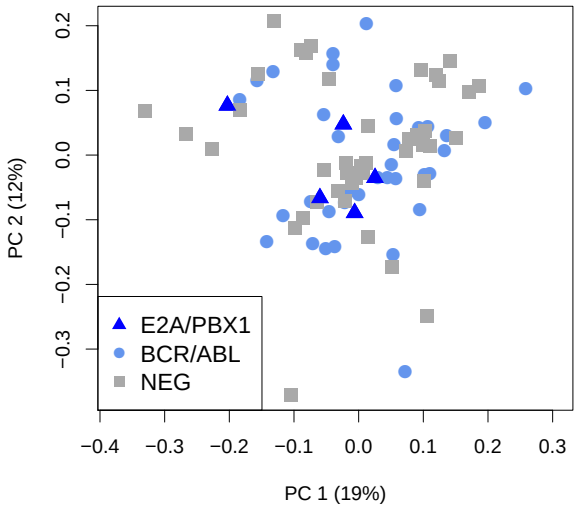


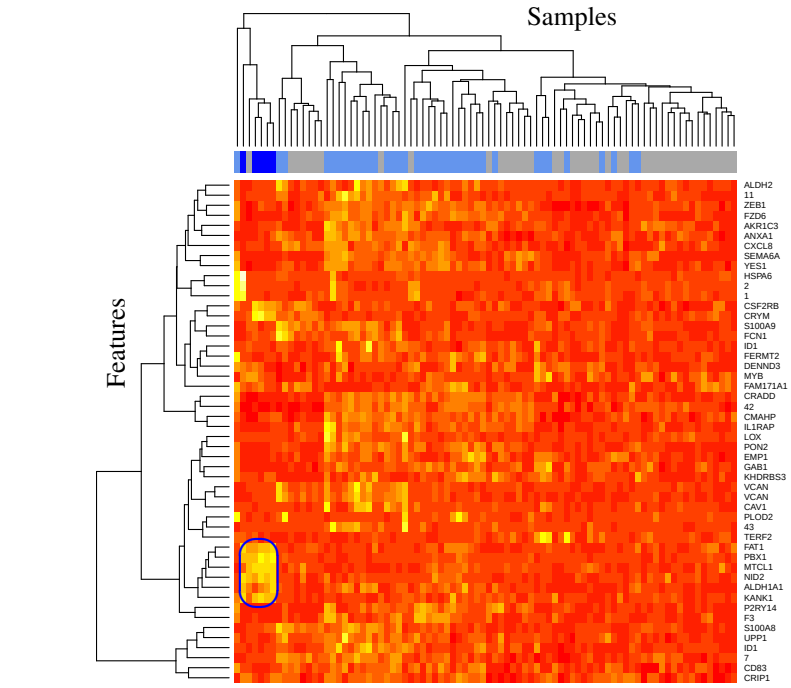
Abbildung 15: Scatterplot der ersten beiden Hauptkomponenten (PC) für den ALL-Datensatz. In Klammern angegeben ist der Anteil der durch die jeweilige Hauptkomponente erklärten Varianz. Die Samples sind entsprechend ihres Tumorsubtyps markiert.

6.2.1 Ergebnisse der FSx-Verfahren

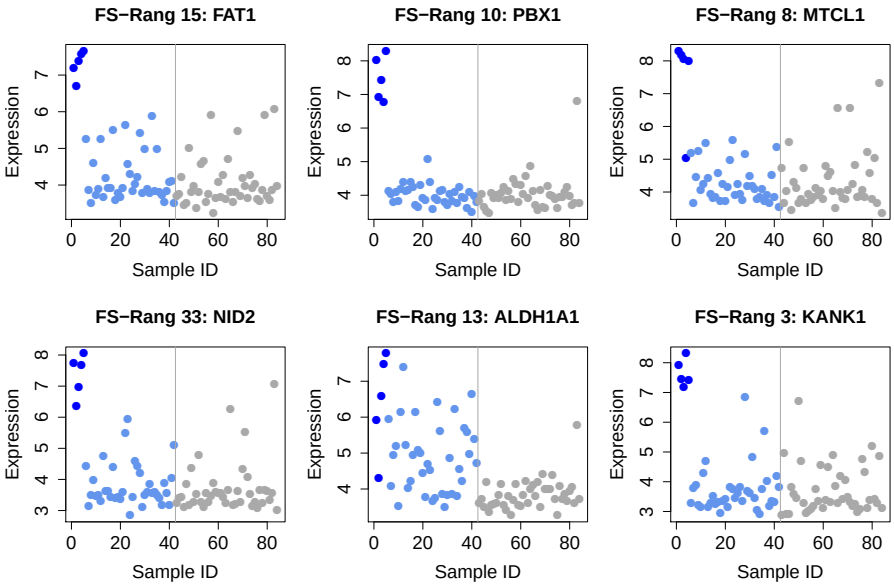
Im Folgenden wird die Anwendung der beiden Varianten des FSx-Workflows auf den ALL-Datensatz präsentiert. Der erste Schritt zur Vorauswahl interessanter Features ist für beide Ansätze identisch, die Unterschiede ergeben sich erst ab dem zweiten Schritt durch die unterschiedlichen Ähnlichkeitsmaße. Aufgrund der ähnlichen Ergebnisse der beiden Varianten wird die ausführliche Darstellung möglicher Analysen in Schritt 2 hier nur für FSOL dargestellt. Für FSJ wird lediglich die zur Auswahl der interessierenden Variablengruppe genutzte `igraph`-Darstellung beschrieben. Für den Vergleich der Ergebnisse des letzten Workflowschritts werden die besten Variablengruppen gelistet und jeweils die von der PBX1-enthaltenden Variablengruppe nominierte Subgruppe mit der wahren Subgruppe verglichen.

FSx Schritt 1: Vorauswahl der Features gemäß FS

Zunächst werden die $T = 50$ Variablen ausgewählt, für die sich die höchsten FS-Werte ergeben. Eine Heatmap der Expressionswerte dieses Teildatensatzes findet sich in Abbildung 16(a). Es zeigt sich, dass bereits der erste Schritt des FSx-Workflows wesentlich zur Identifikation von Patientensubgruppen beitragen kann. Nach der Vorauswahl der Variablen tritt nach Anwendung des üblichen hierar-



(a) Heatmap



(b) Expressionsplots der Variablen, die in (a) eine Subgruppe anzeigen

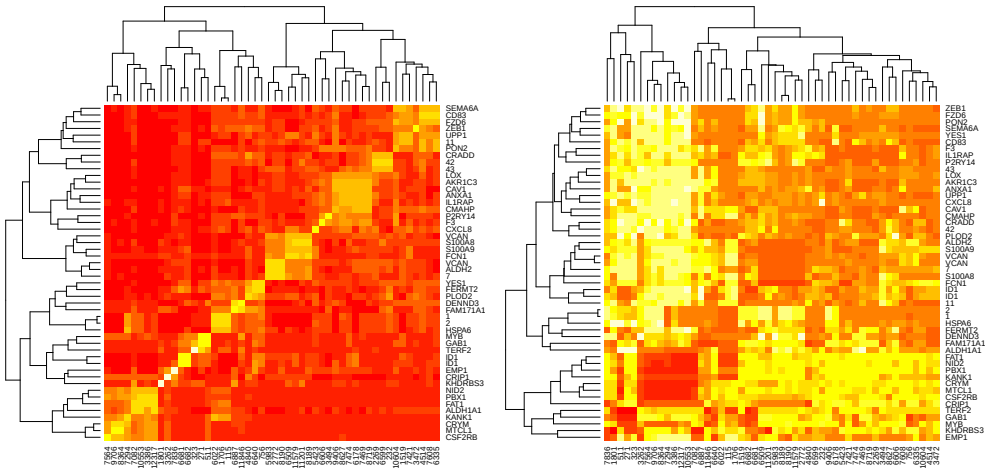
Abbildung 16: (a) Heatmap der Expressionswerte der top-50-FS-Features des ALL-Datensatzes (hierarchisches Clustern, Euklidische Distanz und *complete linkage*). Spalten repräsentieren Samples und sind entsprechend des Tumortyps farbcodiert: grau - homogene Gruppe NEG, Blautöne - heterogene Gruppe aus BCR/ABL (hellblau) und der gesuchten Subgruppe E2A/PBX1 (dunkelblau). Zeilen repräsentieren Features; angegeben sind die abgekürzten Gennamen (falls annotiert, sonst FS-Ränge). Das blaue Oval markiert das auffälligste Cluster hoher Werte. (b) Expressionsplots der im markierten Cluster enthaltenen Variablen. Farbcodierung wie in (a).

chischen Clusterns (d. h. mit einem gewöhnlichen Distanzmaß, hier Euklidische Distanz) eine Subgruppe von sechs Variablen und sechs Samples deutlich hervor. Die Expressionsmuster dieser Variablen sind in Abbildung 16(b) zu sehen. Der Abgleich mit dem wahren Tumortyp zeigt, dass die Samplegruppe alle fünf zu identifizierenden Fälle mit der E2A/PBX1-Mutation enthält. Eine der Variablen misst PBX1 selbst, hier ist die Abgrenzung der wahren Subgruppe zu den beiden anderen Mutationstypen am deutlichsten.
Beim eben beschriebenen Clustern wird die Ähnlichkeit der Expressionswerte der 50 Features über die Menge aller Samples beurteilt. Abhängig vom Datensatz kann eine tatsächlich enthaltene Subgruppe deshalb in dieser Darstellung weniger deutlich erkennbar sein als in diesem Beispiel oder sogar verborgen bleiben. In jedem Fall kann die Anwendung der weiteren Schritte des FSx-Workflows zusätzliche Informationen über die Subgruppenstruktur in den Daten liefern, wie die folgenden Abschnitte zeigen.

FSx Schritt 2: Variablengruppierung

Die Ähnlichkeit der in Schritt 1 ausgewählten Variablen bzgl. der von ihnen angezeigten Subgruppe wird beim **FSOL**-Ansatz mithilfe des p-Wertes p_{OL} beurteilt. Einen guten Eindruck der Ähnlichkeitsstruktur erhält man durch Anwendung von hierarchischem Clustern, wenn die Matrix $-log_{10}D$ als Distanzmatrix verwendet wird, wobei $D = (p_{OL})_{(i,j)}, i, j \in \{1, \ldots, T\}$. Aus technischen Gründen werden Einträge aus D mit dem Wert 0 vor der Transformation durch das Minimum der beobachteten positiven Werte ersetzt. Die Häufigkeit des Auftretens von Nullwerten hängt mit dem Parameter n_{perm} zusammen, der die Anzahl der Permutationen und somit die Genauigkeit bei der Simulation von p_{OL} bestimmt. Die Heatmap in Abbildung 17(a) zeigt (wie schon Abb. 16(a)) eine Variablengruppe, die neben PBX1 auch NID2, ALDH1A1 und FAT1 enthält. Darüber hinaus sind durch die Verwendung von p_{OL} als geeignetes Ähnlichkeitsmaß aber auch zusätzliche Variablengruppen erkennbar. Die von diesen Variablen angezeigten Samplesubgruppen könnten Hinweise auf weitere molekulare Unterschiede zwischen den ALL-Patienten liefern.

Eine weitere Möglichkeit bei der Auswertung ist die fokussierte Betrachtung von einzelnen Variablen, die unter den Toprängen in Schritt 1 identifiziert wurden. Dabei könnte ein Feature aufgrund seiner allgemeinen Funktion oder einer bekannten Assoziation mit einer relevanten Samplesubgruppe in einer anderen Krebsart als interessant erachtet werden. Beispielsweise handelt es sich beim Gen PBX1 (FS-Rang 10) um ein (Proto-)Onkogen, das Zellwachstum und -überleben unter anderem in Eierstockkrebs [75] und bestimmten Typen von Brustkrebs [76] beeinflusst. Somit ist es sicher von Interesse, die Features zu betrachten, die eine



(a) $-\log_{10} p_{OL}$ (b) p_{OL}

Abbildung 17: Heatmaps zur Darstellung der Matrix $(p_{OL})_{(i,j)}$ (Clusterparameter wie oben): (a) nach $-\log_{10}$-Transformation, (b) zum Vergleich ohne Transformation.

ähnliche Samplesubgruppe anzeigen wie PBX1. Wird dazu direkt die auf p_{OL} basierende Distanzmatrix D herangezogen, liegt im Gegensatz zur Betrachtung des resultierenden Dendrogramms kein Informationsverlust vor.
Abbildung 18 zeigt die Scatterplots der sechs Features, die gemäß des Ordered-List-p-Wertes p_{OL} die größte Ähnlichkeit zu PBX1 aufweisen. Dazu zählen unter anderem FAT1 und KANK1, bekannte bzw. vermutete Suppressorgene. Für NID2 wurde bereits ein Zusammenhang zwischen gemeinsamer hoher Expression mit PBX1 und dem *t(1;19)*-Typ in kindlicher ALL berichtet [77]. Dass dieser *t(1;19)*-Typ das unsere gesuchte Subgruppe definierende Fusionstranskript E2A/PBX1 produziert, untermauert den aus den FSOL-Ergebnissen abgeleiteten Hinweis auf einen wahren aufgedeckten Zusammenhang.
Die resultierenden Muster deuten insofern auf interessante Paare von SG-anzeigenden Variablen hin, als dass für den Großteil der Samples kein offensichtlicher (z. B. linearer) Zusammenhang zwischen den betrachteten Variablen besteht. Nur eine kleine Gruppe von Samples zeigt in beiden Variablen auffällig hohe Werte und lässt sich deutlich vom Rest abgrenzen. Je nach zugrundeliegender Technologie zeigen hohe Ähnlichkeiten mitunter auch starke lineare Zusammenhänge an, wenn beispielweise verschiedene *probe sets* mit dem gleichen Gen annotiert sind und im Wesentlichen dieselbe Information messen. Die zugehörigen Scatterplots liegen in diesen Fällen um die Winkelhalbierende.

Um die Nominierung einer potentiellen Subgruppe im nächsten Schritt des Workflows zu ermöglichen, ist zum Abschluss von Schritt 2 noch die Aufteilung der Variablen in einzelne Variablengruppen notwendig. Bei Anwendung harter Kriterien (vgl. Abschnitt 4.3.3) ergibt sich mit den Defaultparametern die in Abb. 19(a) dargestellte Gruppierung.

Für **FSJ**, d. h. bei Anwendung des Jaccardindex-basierten Ähnlichkeitsmaßes, ergibt sich am Ende von Schritt 2 die in Abbildung 19(b) gezeigte Gruppierung der zuvor ausgewählten Variablen. Für die weitere Betrachtung wird wiederum die Zusammenhangskomponente ausgewählt, die mit PBX1 das SG-definierende Gen enthält. Diese enthält bei Verwendung des Standard-cut-offs für die Ähnlichkeit ($J \geq 0.3$) mit FAT1, KANK1, NID2, MTCL1, CRYM und CSF2RB eine Teilmenge der bei FSOL betrachteten Komponente.

FSx Schritt 3: Nominierung

Das Scoring der jeweils in Schritt 2 bestimmten Komponenten G_r geschieht gemäß dem Median $med_{G_r}^{FS}$ der FS-Scores der enthaltenen Variablen. Die resultierenden besten Gruppen beider FSx-Methoden sind in Tabelle 9 dargestellt.
In diesem Beispiel ist der beste Rang für **FSOL** von einer einzelnen Variable besetzt. Rang zwei belegt eine Gruppe aus zwei Variablen, die beide mit demselben Gen ID1 annotiert sind. Dieses wurde in anderen Leukämietypen bereits mit Krankheitsbeginn und -progression in Verbindung gebracht [78]. Die Betrachtung der Expressionsplots der Variablen in den Gruppen auf Rang 1 und 2 zeigt, dass sich die hohen FS-Scores durch erhöhte Werte in Samples des BCR/ABL-Typs ergeben, nicht in denen der E2A/PBX1-Samples (Plots hier nicht gezeigt).
Die Komponente auf Rang 3 enthält u. a. die bereits erwähnten Variablen zur Messung der Genexpression von FAT1, NID2, PBX1 und KANK1, die die hier interessierende (da bekannte) Samplegruppe des Typs E2A/PBX1 anzeigen. Die erhöhte Expression von FAT1 und NID2 im E2A/PBX1-Typ der kindlichen ALL wurde in [79] bereits beschrieben. Die Vereinigung der Samples mit den jeweils zehn höchsten Expressionswerten der elf Variablen umfasst insgesamt 48 Samples, davon sind nur fünf mindestens $\lceil 11 \cdot r_{min} \rceil = \lceil 11 \cdot 0.5 \rceil = 6$ mal vertreten und somit für die Subgruppe nominiert. Bei diesen fünf nominierten Samples handelt es sich genau um die zu identifizierende Subgruppe E2A/PBX1, sodass ein Jaccardindex von 1 erreicht wird.

Im Ranking der Variablengruppen bei **FSJ** liegt die PBX1 beinhaltende Gruppe aus den sieben Variablen CSF2RB, CRYM, MTCL1, NID2, FAT1, PBX1 und KANK1 auf Rang 4. Nominiert werden insgesamt sieben Samples, darunter die

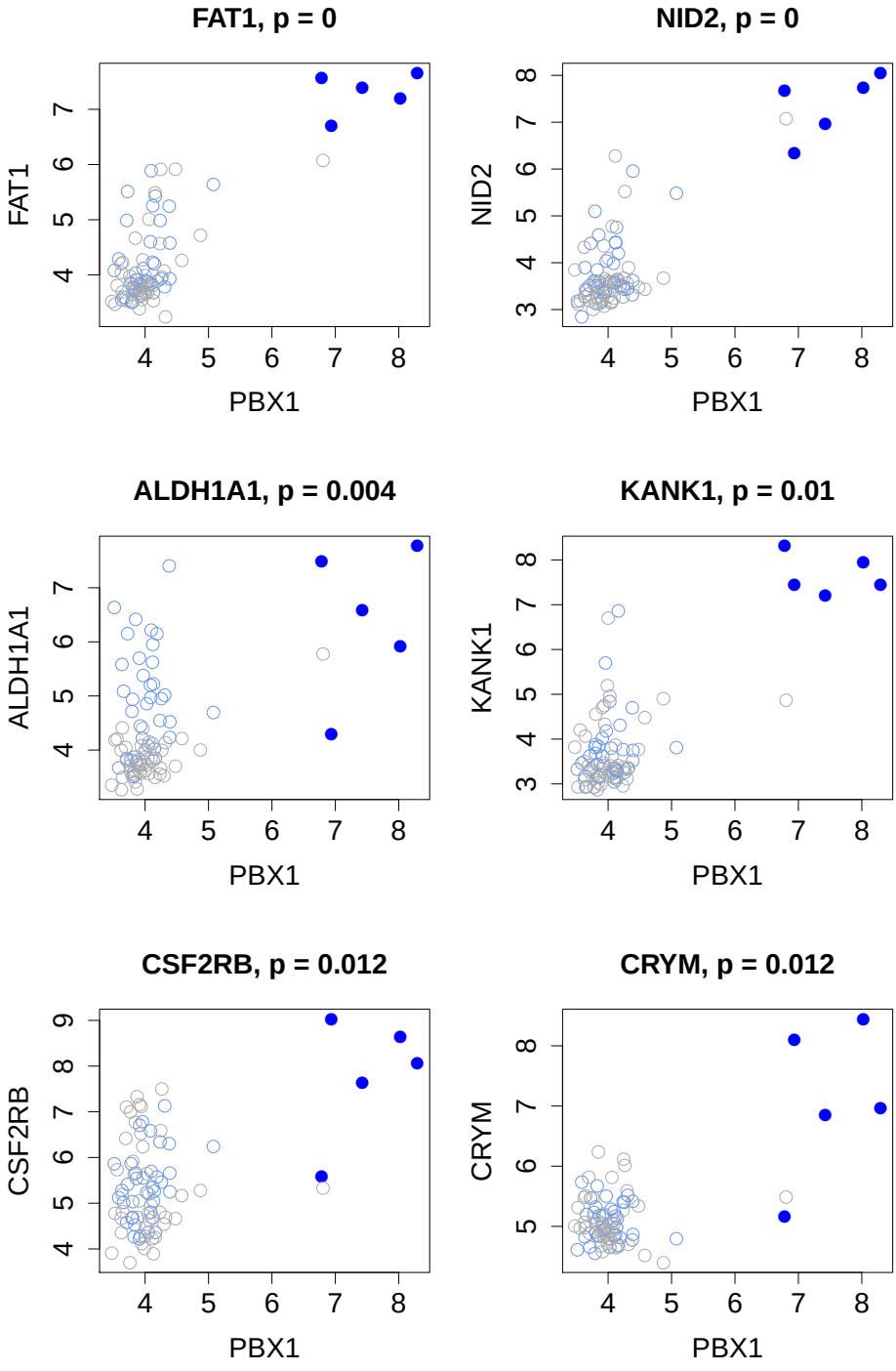
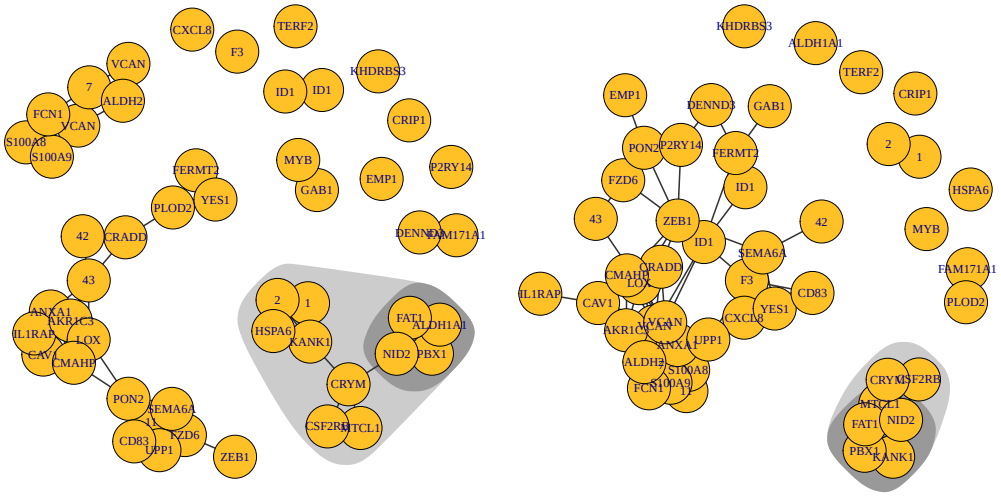


Abbildung 18: Scatterplots der Expressionswerte der Variablen mit größter (Ordered-List)-Ähnlichkeit zu PBX1, geplottet gegen PBX1 selbst. Wie oben sind Samples entsprechend des Tumortyps farblich codiert: grau - homogene Gruppe NEG, hellblau - BCR/ABL, dunkelblau - die gesuchte Subgruppe E2A/PBX1.



(a) FSOL (b) FSJ

Abbildung 19: Gruppierung der top-50-FS-Variablen im Vergleich für die beiden FSx-Varianten unter Verwendung der jeweiligen Standardparameterwerte: (a) für FSOL basierend auf p_{OL}, (b) für FSJ basierend auf dem Jaccardindex. Knoten repräsentieren Variablen; falls keine Genannotation verfügbar ist, ist der FS-Rang der Variable eingetragen. Ein Variablenpaar ist verbunden, wenn die Ähnlichkeit der angezeigten Subgruppe ausreichend hoch ist, genauer wenn der zugehörige $p_{OL} < 0.01$ bzw. wenn $J \geq 0.3$. Dunkel hinterlegt ist jeweils eine maximale Clique, in hellerem grau ist die übergeordnete Zusammenhangskomponente markiert. Weitere Erläuterungen sind im Text zu finden.

fünf gesuchten E2A/PBX1-Samples, was zu einem Jaccardindex von 0.71 führt. Vier der E2A/PBX1-Samples tauchen dabei in allen sieben Variablen unter den Toprängen auf, die übrigen 3 Samples hingegen nur in jeweils vier. Eine strengere Wahl des Mindestanteils r_{min} würde in diesem Fall also die Nominierung von vier der fünf gesuchten Samples und einen Jaccardindex von 0.8 bedeuten.

Ähnlich zu den Scatterplots in Abbildung 18, bei denen der Zusammenhang zwischen einer ausgewählten Variable und jeweils denen mit der höchsten Ähnlichkeit untersucht wurde, können die paarweisen Scatterplots der Variablen in einzelnen Variablengruppen betrachtet werden. Beispielhaft zeigt Abbildung 20 einen solchen Plot für die drei Variablen FAT1, PBX1 und KANK1. Diese Darstellung gibt Aufschluss darüber, ob die Variablen sich bzgl. der Ordnung der Expressionswerte nur in den Toprängen ähneln oder ob die Variablen über alle Samples hinweg beispielsweise einen linearen Zusammenhang aufweisen. Im untersuchten Tripel

(a) Top-Variablengruppen FSOL, cut-off $t_{OL} = 0.01$

Rang r	$\lvert G_r \rvert$	FS-Ränge	$med^{FS}_{G_r}$	Gen(e)
1	1	9	12.05	F3
2	2	5, 22	11.77	ID1, ID1
3	11	15, 13, 33	11.36	FAT1, ALDH1A1, NID2,
		10, 1, 2, 18, 3		PBX1, 1, 2, HSPA6, KANK1,
		50, 21, 8		CRYM, CSF2RB, MTCL1

(b) Top-Variablengruppen FSJ, cut-off $t_J = 0.3$

Rang r	$\lvert G_r \rvert$	FS-Ränge	$med^{FS}_{G_r}$	Gen(e)
1	2	1, 2	18.19	1, 2
2	2	4, 25	11.89	FAM171A1, PLOD2
3	1	13	11.36	ALDH1A1
4	7	21, 50, 8,	11.19	CSF2RB, CRYM, MTCL1,
		33, 15, 10, 3		NID2, FAT1, PBX1, KANK1

Tabelle 9: Vergleich der Variablengruppen mit höchstem medianen FS-Score $med^{FS}_{G_r}$ bei den FSx-Workflows. Gezeigt werden jeweils die obersten Ränge bis zur Variablengruppe G_r, die offenbar mit der gesuchte Subgruppe (Samples des Typs E2A/PBX1) assoziiert sind. r bezeichnet den Rang der Variablengruppe im $med^{FS}_{G_r}$-Ranking, $\lvert G_r \rvert$ die Anzahl enthaltener Variablen.

scheint sich die Ähnlichkeit der Variablen darauf zu begrenzen, dass die Samples mit jeweils höchsten Werten übereinstimmen (jeweils rechte obere Ecke). Die übrigen Samples bilden eine zufällig wirkende Punktwolke in der jeweils unteren linken Ecke. Im allgemeinen sind Variablenpaare mit diesem Zusammenhangsmuster von größerem Interesse für die Charakterisierung der identifizierten Subgruppen hinsichtlich ihrer Pathologie und beteiligter Pathways.
Im Anhang ist exemplarisch ein Scatterplot dargestellt, der im Gegensatz dazu eher einen linearen Zusammenhang über alle Samples hinweg vermuten lässt (Abb. 60). Es handelt sich um die beiden Variablen mit FS-Rängen 1 und 2, die im FSJ-Workflow die Komponente auf Rang 1 bilden.

6.2.2 Ergebnisse der Bicluster-basierten Verfahren

Wie eingangs erläutert, wurden für die Analyse der beiden Bicluster-basierten Verfahren jeweils 1 000 Läufe mit unterschiedlichen Startwerten deskriptiv analysiert. Der Vergleich der Laufzeiten von FSBC und BC findet sich in Tabelle 10(a). Es wird überprüft, ob sich auch ohne die Anwendung automatisierter Ensembleme-

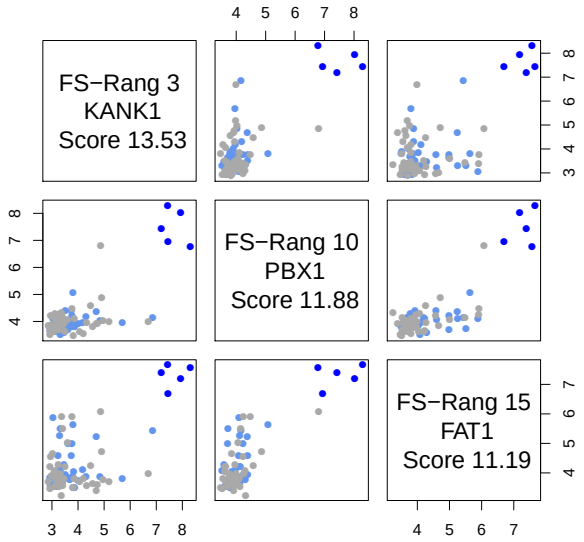


Abbildung 20: Paarweise Scatterplots der Expressionswerte für drei Variablen aus der wahren interessierenden Variablengruppe im ALL-Beispiel. In grau die Samples der homogenen Gruppe NEG, in hellblau die Samples der BCR/ABL-Gruppe und in dunkelblau die gesuchte Subgruppe vom Typ E2A/PBX1.

thoden Hinweise auf die vorliegende Subgruppe aufdecken lassen. Die wichtigsten Ergebnisse der im Folgenden dargestellten Anwendung der Bicluster-basierten Verfahren auf den ALL-Datensatz sind in Tabelle 10 zusammengefasst.

FSBC: Biclustern auf top-50-FS-Variablen

In 1 000 Läufen wurde mittels FSBC jeweils mindestens ein Cluster gefunden, dabei pro Lauf zwischen einem und sechs Clustern, wobei die Häufigkeit mit der Clustergröße abnimmt. Gut 400 Läufe lieferten ein einzelnes Cluster als Ergebnis. Insgesamt wurden in den 1 000 Läufen 2 007 Cluster detektiert (Tab. 10(b)). Diese enthalten zwischen zwei und 34 Samples, die mediane Größe beträgt 15, am häufigsten wurde eine Größe von 21 Samples beobachtet (306 Läufe). Die Cluster enthalten zwischen zwei und 19 Variablen, wobei der Großteil mit knapp 95% aus zwei bis fünf Variablen besteht.
Um einen Eindruck über die möglicherweise subgruppenrelevanten Variablen im Datensatz zu gewinnen, werden die jeweiligen Variablenkombinationen der Cluster mit typischer Größe (zwei bis fünf Variablen) tabelliert und absteigend nach ihrer Häufigkeit geordnet (Tab. 10(e)). Die beiden häufigsten Variablenkombinationen (insg. 915) in den so analysierten 1 896 Clustern enthalten verschiedene Kombinationen der Gene KANK1, PBX1, FAT1, NID2, MTCL1, CRYM und TERF2.

(a) Laufzeit der Bicluster-basierten Methoden

	Laufzeit	Variablenanzahl
FSBC	$<$ 1min	$T = 50$
BC	6h 23min	12 625

(b) Häufigkeit der Anzahl gefundener Bicluster in einem Lauf

	0	1	2	3	4	5	6	7	8	9
FSBC	0	409	312	173	81	19	6	0	0	0
BC	6	32	57	77	482	232	78	31	3	2

(c) maximale Anzahlen korrekt nominierter Samples aus wahrer E2A/PBX1-Subgruppe

	0	1	2	3	4	5
FSBC	189	39	9	0	55	708
BC	6	222	772	0	0	0

(d) Jaccardindex des Clusters mit maximaler Anzahl korrekt nominierter E2A/PBX1-Samples

	Min.	q_{25}	q_{50}	MW	q_{75}	Max.
FSBC	0.00	0.28	0.50	0.45	0.71	1.00
BC	0.00	0.07	0.18	0.15	0.18	0.18

(e) Häufigste Variablenkombinationen in Clustern typischer Größe für FSBC

	Variablen im Cluster	
Hfkt	FS-Ränge	Gen(e)
552	3, 10, 15, 33	KANK1, PBX1, FAT1, NID2
363	8, 10, 15, 30, 50	MTCL1, PBX1, FAT1, TERF2, CRYM
164	7, 12, 16	7, S100A8, S100A9

Tabelle 10: Charakterisierung der Ergebnisse der Bicluster-basierten Verfahren in 1 000 Wiederholungen mit unterschiedlichen Startwerten auf dem ALL-Datensatz. (a) zeigt die gemessene Laufzeit für die insgesamt 1 000 Biclusterläufe. In (b) und (c) werden Anzahlen gefundener Bicluster bzw. maximale Anzahlen nominierter Samples aus der E2A/PBX1-Subgruppe tabelliert. Für die Beschreibung der Jaccardindizes in (d) werden die Extrema, die 25%, 50% und 75%-Quantile, sowie das arithmetische Mittel (MW) angegeben. (e) zeigt für die FSBC-Methode die häufigsten Variablenkombinationen für Cluster typischer Größe (zwei bis fünf Variablen).

Mit dieser manuellen (eindimensionalen) Zusammenfassung der verschiedenen Clusterläufe lässt sich am ALL-Beispiel somit mittels FSBC eine ähnliche Variablengruppe als SG-relevant identifizieren wie mit den FSx-Ansätzen. Lediglich TERF2 trat bei den FSx-Ergebnissen nicht in Erscheinung. Im Gegensatz zu den Variablen, die ebenfalls von den FSx-Ansätzen ausgewählt wurden, lässt sich bei Betrachtung der Expressionswerte von TERF2 keine Abgrenzung der interessierenden Subgruppe erkennen (siehe Abb. 21).
Zur Beurteilung der Nützlichkeit der FSBC-Methode bei der Generierung von konkreteren Hinweisen auf die wahre E2A/PBX1-Subgruppe wurde pro Cluster die Anzahl der enthaltenen SG-Samples berechnet (Tab. 10(c)). Wurden in einem Lauf mehrere Cluster identifiziert, wurde das mit der maximalen Anzahl berücksichtigt. In rund 700 der 1 000 Läufe waren alle fünf gesuchten Samples in dem jeweils betrachteten Cluster enthalten. In knapp 200 Läufen wurde keines der interessierenden Samples für ein Cluster ausgewählt. Für die Cluster mit maximaler Anzahl gefundener SG-Samples pro Lauf wurden auch die Jaccardindizes zur Übereinstimmung mit der wahren Subgruppe (fünf E2A/PBX1-Samples) bestimmt und die resultierende empirische Verteilung in Kennzahlen zusammengefasst (Tab. 10(d)). Das dritte Quartil liegt beispielsweise bei 0.71.
Nachdem die Kombination der Variablen KANK1, PBX1, FAT1 und NID2 als typisches Cluster identifiziert wurde, ist der fehlende Schritt zur Nominierung einer Subgruppe die Betrachtung der jeweils im Bicluster enthaltenen Samples. In den häufigsten Fällen enthalten die 552 Cluster sechs oder zehn Samples. Zunächst sei angemerkt, dass alle Cluster die gesuchten fünf Samples enthalten. Im wesentlichen handelt es sich pro Sampleanzahl immer um die gleiche Samplemenge. In den 203 Fällen mit sechs Samples handelt es sich beispielsweise immer um die fünf gesuchten und ein zusätzliches Sample (immer dasselbe). Bei letzterem handelt es sich um eines der beiden Samples, das auch von FSJ identifiziert wurde. Auch die Menge aus zehn Samples (291 der 552 betrachteten Cluster) ist immer identisch: es ist eine Obermenge der Sechsergruppe, d. h. das reportete Bicluster enthält noch vier zusätzliche Samples. Die zugehörigen Jaccardindizes für die so hergeleiteten Samplesubgruppen liegen zwischen 0.5 und 0.83.

BC: Biclustern auf gesamter Expressionsmatrix

Anders als bei FSBC wurde nicht in jedem der 1 000 BC-Läufe ein Bicluster gefunden (Tab. 10(b)). Insgesamt wurden zwischen null und neun Cluster detektiert, wobei in über der Hälfte der Fälle (714 Läufe) vier oder fünf Cluster ausgegeben werden. Über alle Läufe hinweg ist die Clusteranzahl mit 4 192 in etwa doppelt so hoch wie bei FSBC. Dieser Faktor scheint klein im Vergleich zum Verhältnis der verwendeten Datensätze: Mit 50 Variablen liegt der Anteil der Datenmatrix von FSBC verglichen mit dem BC-Datensatz (12 625 Variablen) unter einem Prozent.

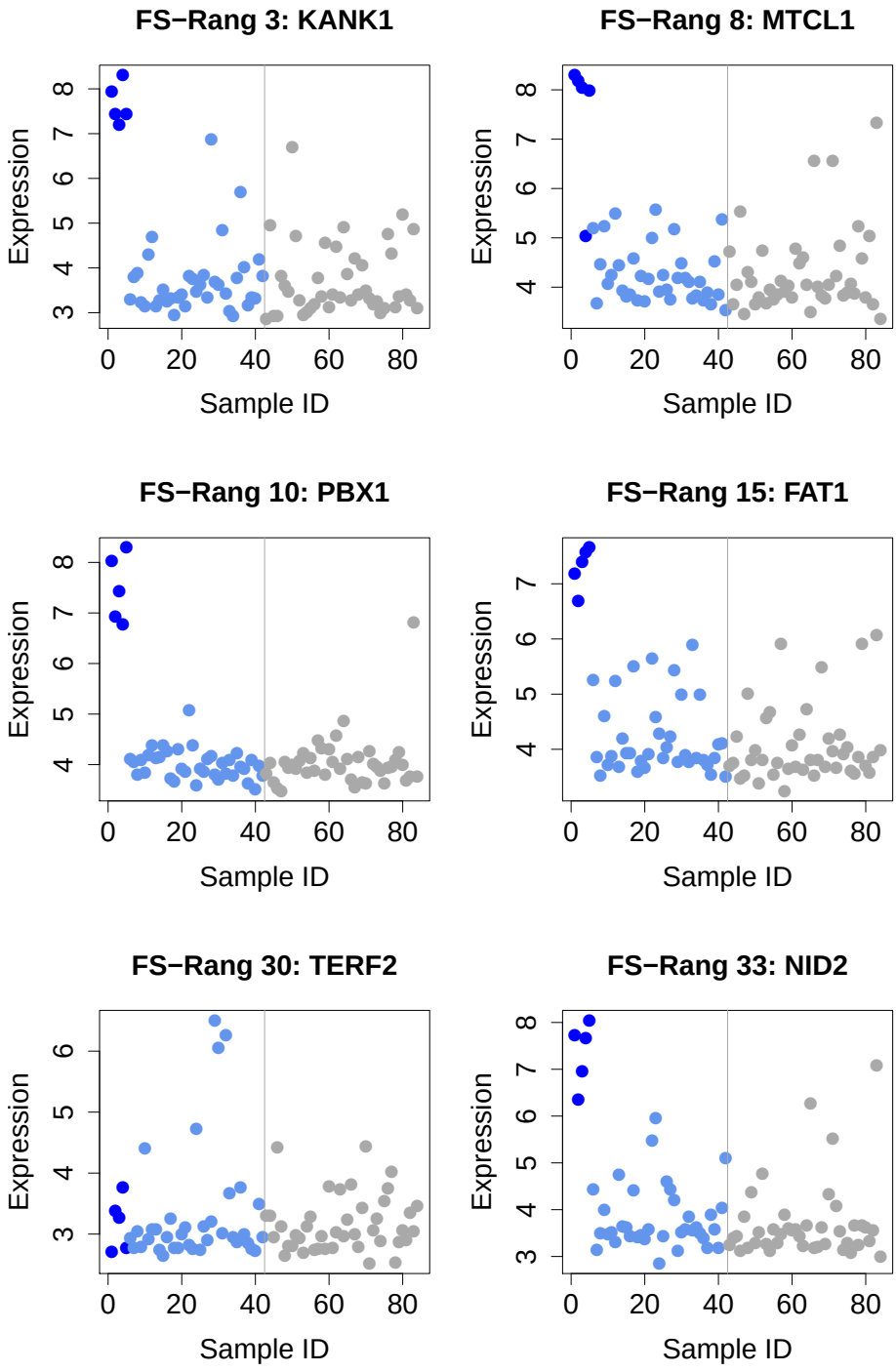


Abbildung 21: Auswahl potentiell interessanter Variablen bei manueller Auswertung von 1 000 FSBC-Läufen auf den ALL-Daten. Bis auf TERF2 zeigen alle Variablen die gesuchte Hoch-Regulation in der wahren Subgruppe. Die übrigen fünf Variablen wurden auch mit den FSx-Ansätzen identifiziert.

Die gefundenen Cluster enthalten zwischen acht und 36 Samples mit deutlichen Häufungspunkten bei einer Größe von zehn und elf Samples (zusammen 2 371 der 4 192 Cluster). Der größte Unterschied zu den FSBC-Ergebnissen ergibt sich für die empirische Verteilung der Clustergrößen hinsichtlich enthaltener Variablen, da diese bei FSBC durch $T = 50$ beschränkt ist. Insgesamt sind zwischen zwei und 339 Variablen beteiligt, wobei sich deutliche Häufungspunkte für kleinere Cluster mit bis zu 20, mittlere mit meist 64 und große mit knapp 340 Variablen abzeichnen.
Aufgrund des Verzichts auf automatisierte Ensemblemethoden und aus Gründen der Übersichtlichkeit wird im Folgenden die Menge der Variablen in den kleineren Clustern analysiert. Anders als bei FSBC handelt es sich bei den häufigsten Clusterzusammensetzungen weniger offensichtlich um unterschiedliche Kombinationen einer interessanten Variablengruppe. Die fünf häufigsten Kombinationen in dieser Clustergröße bestehen beispielsweise aus disjunkten Variablenmengen. Die siebthäufigste Kombination enthält allerdings zusätzlich zu den 17 Variablen der dritthäufigsten Kombination nur eine zusätzliche Variable. Ähnlichkeiten sind auch für die weiteren Variablenkombinationen erkennbar. In diesem Beispiel wäre es womöglich ebenfalls durch Deskription möglich, relativ konsistente Variablengruppen über die Bicluster-Läufe hinweg zu identifizieren. Bedingt durch die hohe Anzahl verfügbarer Variablen ist der Aufwand ohne automatisierte Ensemblemethoden jedoch ungleich höher als für FSBC.
Der wesentliche Unterschied zwischen dem Biclustern und den drei bisher angewendeten Methoden FSOL, FSJ und FSBC zeigt sich im Hinblick auf die Identifikation der interessierenden E2A/PBX1-Subgruppe. In keinem der 1 000 Läufe wurden mehr als zwei der fünf wahren SG-Samples in ein Bicluster nominiert. Bereits ohne eingängige Beurteilung der vom Biclustern identifizierten Samplegruppen ist somit ersichtlich, dass das Biclustern auf diesem Datensatz nicht in der Lage ist, nützliche Hinweise auf die Existenz der gesuchten wahren Subgruppe zu liefern.

6.3 DeNoPa

DeNoPa ist eine longitudinale Studie zur Erforschung der unterschiedlichen Verlaufsformen der Parkinsonerkrankung (kurz PD für *Parkinson's disease*) ab einem sehr frühen Stadium. Verschiedene Testverfahren und Untersuchungsmethoden werden hinsichtlich ihrer Eignung zur Frühdiagnose beurteilt. Die Rekrutierung von Parkinsonfällen und Gesundkontrollen lief von 2008 bis 2012 in der Paracelsus-Elena Klinik (Zentrum für Parkinson-Syndrome und Bewegungsstörungen) in Kassel. Neben allgemeinen Kovariablen wie Alter und Geschlecht wurden von allen Probanden auch parkinsonspezifische klinische Variablen erfasst. Dazu gehören zum Beispiel das Schlafverhalten oder Ergebnisse eines olfaktori-

schen Tests. Weitere Informationen sind auf der offiziellen Homepage `http://www.denopa.de/` nachzulesen. Erste Analysen der Studie zu nicht-motorischen Störungen und der sogenannten REM-Schlaf-Verhaltensstörung wurden in Mollenhauer et al. [80] vorgestellt.

Der hier als DeNoPa bezeichnete Datensatz wurde mittels labelfreier Massenspektrometrie aus CSF-Proben (*cerebrospinal fluid* für Zerebrospinalflüssigkeit oder auch Gehirn-Rückenmarks-Flüssigkeit) am Medizinischen Proteom-Center der Ruhr-Universität Bochum generiert. Details zum Quantifizierungsworkflow der sogenannten LC-MS/MS-Messungen sind beispielsweise nachzulesen in Levin und Bahn [81]. Eine Subgruppenanalyse dieses Datensatzes ist von besonderem klinischen Interesse, da die Parkinsonerkrankung sehr heterogen ist und unterschiedliche Verlaufsformen zeigt. Um eine für die Subgruppendetektion ausreichende Fallzahl zu erreichen, wurden zwei separat gemessene Datensätze vereinigt, die jeweils etwa gleich viele Patienten und Kontrollen enthielten. Für die folgenden Analysen stehen somit Daten von 40 Gesundkontrollen und 42 Parkinsonpatienten zur Verfügung.
Die Menge der 904 Variablen des vereinigten Sets ist der Schnitt der in beiden Experimenten identifizierten und quantifizierten Proteine bzw. Proteingruppen. Der Vergleich wurde auf Basis der Proteinaccessions (bei Gruppen: Accession eines Repräsentanten) durchgeführt. Um die Effekte möglicher technisch bedingter Verzerrungen (z. B. in Form von Shifts zwischen den beobachteten Abundanzlevels der Teildatensätze) zu minimieren, wurde vor den eigentlichen Analysen eine Batchadjustierung mithilfe der `ComBat`-Funktion des **R**-Pakets `svn` durchgeführt. Auf der globalen Ebene der PCA sind nach der Adjustierung keine Untergruppen von Samples erkennbar (Abb. 22).
Für die untersuchten Proben liegen neben den massenspektrometrischen Daten auch die per ELISA (*Enzyme Linked Immunosorbent Assay*) bestimmten Hämoglobinwerte vor. Die ELISA-Messungen wurden im Anschluss an die Probenentnahme im Rahmen der Laborroutine unabhängig von den Proteomanalysen durchgeführt. Für vier der 82 Proben wurden erhöhte Hämoglobinwerte festgestellt. Für dieses Anwendungsbeispiel werden die entsprechenden Samples als interessierende „Hb-Subgruppe“ aufgefasst.
Hämoglobin (Hb) ist als Blutfarbstoff bekannt und sollte grundsätzlich im CSF nicht vorhanden sein, weder in den Kontrollen noch bei den PD-Patienten. Es ist daher anzunehmen, dass sich das gefundene Hb nicht ursprünglich im CSF befand, sondern dass die CSF-Probe während der Entnahme mittels Lumbalpunktion kontaminiert wurde. In diesem Beispiel sind drei der vier Hb-Samples aus der erkrankten PD-Gruppe und eines aus den Kontrollen.
Bei der im Folgenden dargestellten Anwendung der SG-Detektionsmethoden auf die massenspektrometrischen Daten ist vorwiegend von Interesse, ob Hinwei-

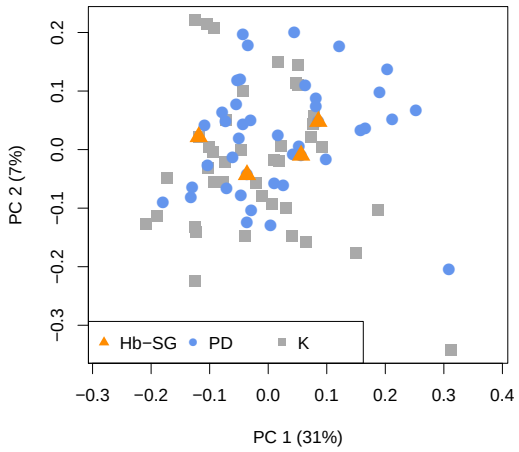


Abbildung 22: Scatterplot der zweiten gegen die erste Hauptkomponente des DeNoPa-Datensatzes. Die im ELISA unauffälligen Samples der Gruppen Parkinson (PD) und Kontrolle (K) sind in blau bzw. grau dargestellt. Die interessierende Subgruppe mit erhöhten Hb-Werte in der ELISA-Messung ist in orange gekennzeichnet.

se auf die Hb-Subgruppe generiert werden können. Wie schon im ParkCHIP-Beispiel erwähnt, ist Parkinson als heterogene Krankheit bekannt. Es besteht daher die Möglichkeit, dass weitere wahre Subgruppen im Datensatz vorhanden sind. Werden also andere Samplesubgruppen von den getesteten Verfahren identifiziert und als relevanter eingestuft, so spricht das nicht gegen die Methode insgesamt. Ob sich ggf. vorliegende Unterschiede zwischen Patienten und Kontrollen oder zwischen den Patienten untereinander auf der Ebene des Proteoms des CSF nachweisen lassen, ist bislang unklar.
Die Annotation der Proteinaccessions hinsichtlich der assoziierten Gene wurde mithilfe des über Bioconductor verfügbaren **R**-Pakets `UniProt.ws` durchgeführt ([82], 29.07.2016, Paketversion 2.12.0, Bioconductor Version 3.3, **R**-Vers. 3.3.1).

6.3.1 Ergebnisse der FSx-Verfahren

FSx Schritt 1: Vorauswahl der Features gemäß FS

Im ersten Workflowschritt werden die FS-Scores für alle Variablen des Datensatzes berechnet, um die $T = 50$ Variablen mit den höchsten Scores für die weitere Analyse in einer Teildatenmatrix zusammenzufassen. Die Visualisierung dieser Teilmatrix in einer Heatmap (Abb. 23) lässt verglichen mit der entsprechenden Darstellung des ALL-Datensatzes (Abb. 16(a)) keine so deutlich abgegrenzte Patientensubgruppe erkennen.

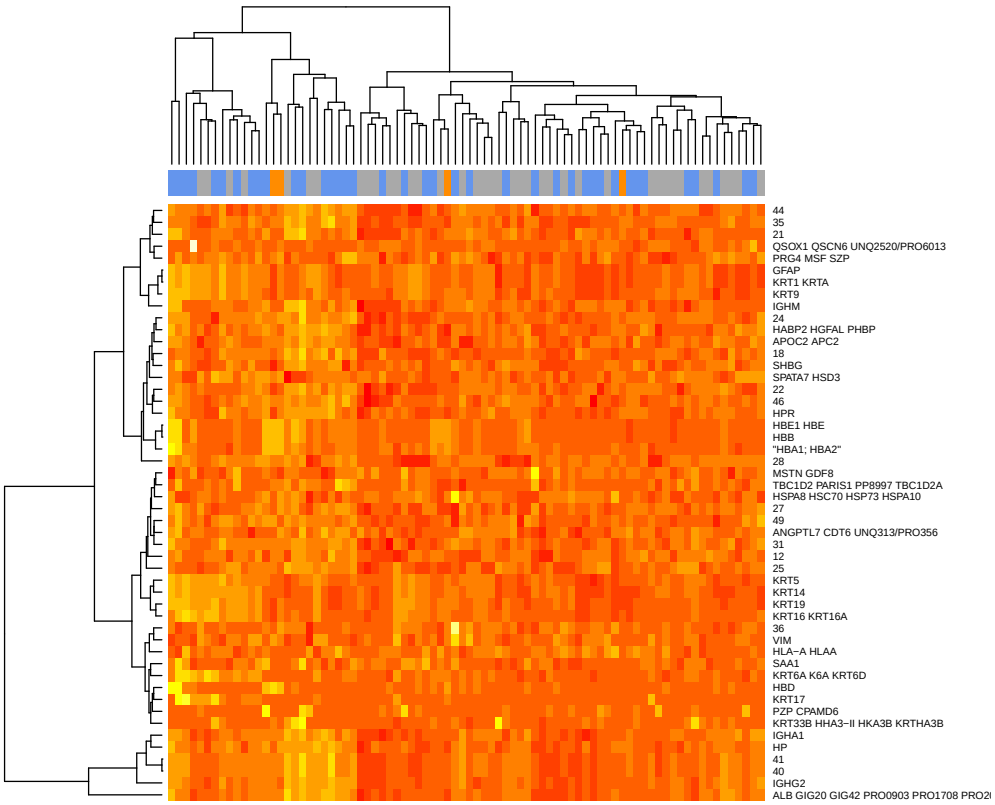


Abbildung 23: Heatmap der Expressionswerte der top-50-FS-Features des DeNoPa-Datensatzes (hierarchisches Clustern, Euklidische Distanz und *complete linkage*). Spalten repräsentieren Samples und Zeilen Features (Proteingruppen). Die Farbcodierung der Samples zwischen Dendrogramm und Heatmap entspricht der in Abb. 22: in orange die Hb-Subgruppe, Kontrollen in grau, PD-Patienten in blau.

Bei einem Split des Sampledendrogramms (Spaltendendrogramm) in zwei Cluster scheinen die Samples mit höheren Werten vermehrt im kleineren (linken) Cluster zu liegen. Die 26 Samples in diesem Cluster stammen etwa zur Hälfte aus den beiden Sets, sodass der Effekt scheinbar nicht auf ein Batchartefakt zurückzuführen ist. Mit 19 Patienten und sieben Kontrollen ist die Trennung offenbar eher mit der Gruppenzugehörigkeit assoziiert.

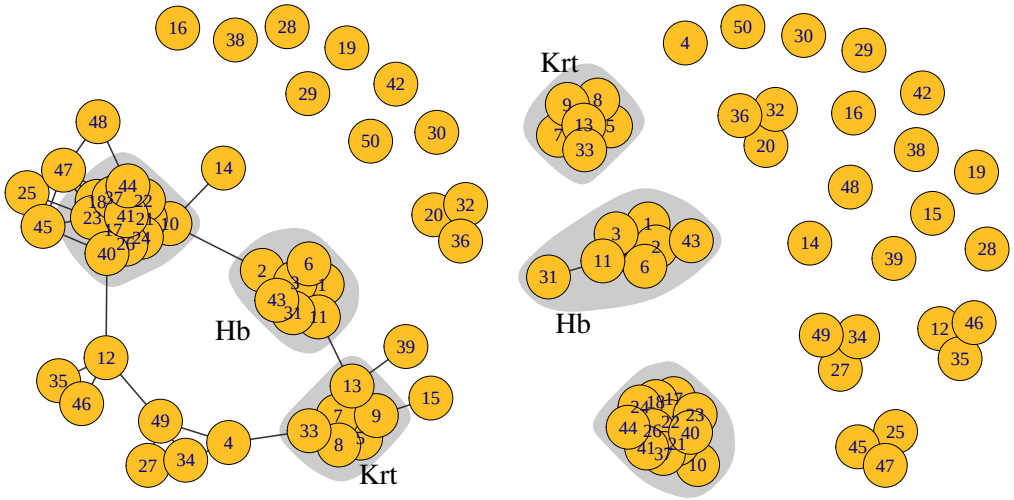
FSx Schritt 2: Variablengruppierung

Nach der Selektion potentiell subgruppenanzeigender Features werden die beiden Ähnlichkeitsmatrizen für FSOL und FSJ zunächst unter Verwendung der Stan-

dardeinstellungen berechnet. Für **FSOL** zeichnen sich in der `igraph`-Darstellung drei Variablengruppen ab, die allerdings noch in einer Komponente verbunden sind (Abb. 24(a)). Um die von den jeweiligen Gruppen angezeigten Samplesubgruppen genauer analysieren zu können, wird der Schwellenwert für die Dichotomisierung der Ähnlichkeit strenger gewählt und von 0.01 auf 0.005 gesenkt (Abb. 24(b)). Tabelle 11 zeigt Details zu den drei Variablengruppen mit höchsten medianen FS-Scores. Anhand der Genannotationen der Variablen lässt sich bereits eine Hämoglobin- und eine Keratingruppe identifizieren.
Auch für **FSJ** ergibt sich mit dem Standardwert für den cut-off t_J insgesamt keine gute Aufteilung der Variablen. Es zeichnet sich ähnlich zu FSOL eine Keratingruppe ab. Bei einer Erhöhung des cut-offs auf $t_J = 0.5$ bleiben im Wesentlichen drei Variablengruppen bestehen (`igraph`-Darstellungen im Anhang, Abb. 61): zum einen die Keratingruppe, eine Hämoglobingruppe, sowie eine weitere Komponente, deren Charakterisierung aufgrund fehlender Annotationen der beteiligten Variablen hier nicht zu bestimmen ist. In der Anwendung sollten auch die Samplegruppen näher betrachtet werden, die von der Keratingruppe und der dritten Variablenmenge nominiert werden, um möglicherweise neue Erkenntnisse über PD-Subgruppen zu gewinnen. Die Details zu den drei Variablengruppen mit höchstem medianen FS-Score sind im Vergleich mit FSOL in Tabelle 11 zu sehen.

FSx Schritt 3: Nominierung

Von Interesse ist nun der Vergleich der mittels FSx nominierten Hämoglobinsubgruppe (basierend auf den LCMS-Daten) mit der Samplegruppe, für die mittels ELISA erhöhte Hämoglobinkonzentrationen detektiert wurden. In diesem Schritt wird für beide FSx-Versionen die Defaulteinstellung zur Samplenominierung verwendet ($r_{min} = 0.5$).
Von **FSOL** werden von den sieben Variablen der Hb-Gruppe auf Rang 2 sieben Samples nominiert. Darunter befinden sich drei der vier gesuchten Samples, was einem Jaccardindex von 0.375 entspricht. Wie auch auf dem ALL-Datensatz kommen die FSx-Varianten für das DeNoPa-Beispiel zu ähnlichen Ergebnissen (vgl. dazu Tabelle 12): Bei Anwendung von **FSJ** werden von den vier Variablen der gefundenen Hämoglobin(variablen)gruppe zusätzlich zu den sieben von FSOL nominierten Samples noch drei weitere Samples nominiert (falsch-positiv, Jaccardindex 0.27). Diese treten im Overlap der Variablen jedoch seltener auf. Die Anwendung des konservativen Nominierungsansatzes würde zu identischen Samplegruppen führen, d. h. nur die sieben von FSOL nominierten Samples finden sich in den top-$max.rk$-Rängen aller vier Variablen der Gruppe.



(a) FSOL, cut-off 0.01 (b) FSOL, cut-off 0.005

Abbildung 24: Einfluss des cut-offs t_{OL} auf die Variablengruppierung in Schritt 2 des FSOL-Workflows auf den DeNoPa-Daten. Um die Informationen aus den sich in (a) mit dem cut-off 0.01 abzeichnenden kleineren Variablengruppen nutzen zu können, wird der cut-off auf 0.005 gesenkt (b). Die Knoten sind mit den FS-Rängen gelabelt. Details folgen in Tabelle 11(a).

6.3.2 Ergebnisse der Bicluster-basierten Verfahren

Zur Beurteilung der Bicluster-basierten Verfahren BC und FSBC wurden je 1 000 Bicluster-Läufe mit unterschiedlichen Startwerten durchgeführt und interessierende Charakteristiken zusammengefasst. Tabelle 13(a) gibt diese zusammen mit die Laufzeiten der beiden Methoden an. Im Folgenden wird wie für die FSx-Ansätze untersucht, ob die Verfahren grundsätzlich in der Lage sind, Hinweise auf die bekannte Hämoglobin-Subgruppe zu liefern.

FSBC: Biclustern auf top-50-FS-Variablen

Ähnlich wie im ALL-Beispiel ergibt sich auf dem DeNoPa-Datensatz für FSBC kein eindeutiger Häufungspunkt in Bezug auf die Anzahl der gefundenen Bicluster in den verschiedenen Läufen. Die Anwendung der FSBC-Methode führte in den meisten Fällen (937 Läufe) zur Detektion von einem bis drei Biclustern (in jeweils rund 31, 37 bzw. 26 Prozent der 1 000 Läufe, Tab. 13(b)). In 32 Fällen wurde kein Bicluster gefunden, selten traten vier, fünf oder sechs Cluster auf. Bei FSBC war in jedem Lauf mit mindestens einem Bicluster auch mindestens eines der Hb-SG-Samples unter den nominierten Samples (Tab. 13(c)). In etwa der

(a) Top-Variablengruppen FSOL, cut-off $t_{OL} = 0.005$

Rang r	$\lvert G_r \rvert$	FS-Ränge	$med^{FS}_{G_r}$	Gen(e)
1	1	4	13.08	KRT19
2	7	1, 2, 3, 6, 11, 31, 43	9.21	HBB, HBE1/HBE, HBA1/HBA2, HBD, KRT9, 31, HP
3	6	5, 7, 8, 9, 13, 33	8.32	KRT14, KRT16/KRT16A, KRT1/ KRTA, GFAP, KRT5, KRT17

(b) Top-Variablengruppen FSJ, cut-off $t_J = 0.5$

Rang r	$\lvert G_r \rvert$	FS-Ränge	$med^{FS}_{G_r}$	Gen(e)
1	4	1, 2, 3, 6	16.66	HBB, HBE1/HBE, HBA1/HBA2, HBD
2	9	4, 5, 7	8.32	KRT19, KRT14, KRT16/KRT16A,
		8, 9, 11		KRT1/KRTA, GFAP, KRT9
		13, 15, 33	7.85	KRT5, KRT6A/K6A/KRT6D
3	1	10		SAA1

Tabelle 11: Vergleich der Variablengruppen mit höchstem medianen FS-Score bei den FSx-Workflows für den DeNoPa-Datensatz. Bezeichnung der Spalten jeweils wie in Tabelle 9.

Hälfte dieser Fälle (477 Läufe) waren es drei der vier gesuchten Samples. Beim Vergleich der nominierten und wahren Subgruppe mittels Jaccardindex beträgt das Maximum über alle Läufe für FSBC 0.3.
Die Anzahl der verfügbaren Variablen ist bei FSBC durch die Anzahl T der FS-selektierten Variablen im Gegensatz zur ursprünglichen Bicluster-Methode stark beschränkt. Die gefundenen FSBC-Cluster weisen eine Größe zwischen zwei und neun Variablen auf, wobei drei und sieben die häufigsten Werte sind. Die Gruppen sind somit von ähnlicher Größe wie bei den FSx-Ansätzen. Aufgrund der Variabilität der Clusterergebnisse liegt keine eindeutige Lösung vor. Wie im ALL-Beispiel wird daher an dieser Stelle ein exemplarisches Cluster hinsichtlich der beteiligten Variablen betrachtet. Ohne die Verwendung von automatisierten Ensemblemethoden wird zunächst für jede auftretende Variablenkombination die Anzahl der Cluster bestimmt, an der exakt diese Variablenmenge beteiligt ist. Diese Betrachtung erfolgt für die Gesamtmenge der identifizierten Cluster aus allen Läufen. Bei Verwendung von FSBC ergibt sich so als häufigstes Cluster (600 von insgesamt 1 948 Clustern) die Kombination von drei mit Hämoglobin annotierten Variablen. Auf den folgenden Plätzen finden sich verschiedene Kombinationen der mit Keratin annotierten Variablen.
Bei stark ausgeprägten Häufungspunkten der Variablenkombinationen lassen sich mit der FSBC-Methode also unter Umständen selbst ohne Ensemblemethoden

(a) FSOL, cut-off 0.005

ID	P12	P17	P19	*P38*	P42	*K40*	*P36*
Hfkt	7	5	5	5	5	4	4

(b) FSJ, cut-off 0.5

ID	*K40*	P12	P17	P19	*P36*	*P38*	P42	K20	P26	P1
Hfkt	4	4	4	4	4	4	4	3	3	2

Tabelle 12: Ergebnisse des Nominierungsschritts der FSx-Workflows für die DeNoPa-Daten. Die Samples sind entsprechend ihrer Häufigkeit im Overlap der jeweiligen Hb-Variablengruppe aufgeführt. Kursiv gesetzt wurden die korrekt nominierten Samples mit erhöhten Hb-Werten in den ELISA-Messungen.

nützliche Hinweise auf Variablengruppen gewinnen. Die Variablengruppen ermöglichen in diesem Beispiel mithilfe von Annotationen die Formulierung von Hypothesen über enthaltene Subgruppen. Neben den gesuchten Hinweisen auf eine Hämoglobinsubgruppe zeichnet sich wie auch bei den FSx-Methoden eine Keratingruppe ab.
Zum Abschluss werden noch die Samplekombinationen, die im häufigsten Cluster (Hb-Variablengruppe aus drei Variablen) am häufigsten auftreten, tabelliert um diese mit der wahren Subgruppe und mit den nominierten Samples der FSx-Methoden zu vergleichen. Die häufigste Kombination aus 48 Samples weist keine große Ähnlichkeit mit den FSx-Nominierungen auf. Bei den nächsten fünf Gruppen, die elf bis 25 Samples enthalten, besteht allerdings die Schnittmenge aus zehn Samples, die bis auf eines der nominierten Samples identisch mit der von FSJ nominierten Hb-Subgruppe ist. Betrachtet man diese Samplegruppe als nominiert, führt das zu einem Jaccardindex von 0.27 bzgl. der gesuchten Subgruppe.

BC: Biclustern auf gesamter Expressionsmatrix

Zunächst zeigt Tabelle 13(b) die Verteilung der Anzahl gefundener Bicluster über die 1 000 Läufe. Im Gegensatz zu FSBC wird bei BC immer mindestens ein Cluster gefunden und es gibt ein deutliches Maximum bei der Anzahl gefundener Bicluster. In den meisten Fällen (888 Läufe) liefert die BC-Methode zwei Bicluster, seltener drei (108 Läufe) und nur in Ausnahmen vier oder fünf (je zwei Läufe). Die Betrachtung der Läufe mit genau zwei Clustern zeigt, dass diese nicht nur in ihrer Größe übereinstimmen, sondern im Wesentlichen bzgl. der enthaltenen Variablen und Samples über die Läufe hinweg identisch sind: Das erste Cluster besteht in 874 der 888 Läufe aus denselben 99 Variablen und 14 Samples. Das zweite Cluster setzt sich dabei fast immer (872 der 874 Läufe) aus 27 Variablen

(a) Laufzeit der Bicluster-basierten Methoden

	Laufzeit	Variablenanzahl
FSBC	< 1min	$T = 50$
BC	9min 8s	904

(b) Häufigkeit der Anzahl gefundener Bicluster in einem Lauf

	0	1	2	3	4	5	6	gesamt
FSBC	32	311	368	258	29	1	1	1 948
BC	0	0	888	108	2	2	0	2 118

(c) maximale Anzahlen korrekt nominierter Samples aus wahrer Hb-Subgruppe

	0	1	2	3	4
FSBC	0	293	197	477	1
BC	872	66	30	32	0

(d) maximaler Jaccardindex bzgl. Hb-Subgruppe (pro Lauf mit gefundenen Clustern)

	Min.	q_{25}	q_{50}	MW	q_{75}	Max.
FSBC	0.04	0.05	0.09	0.13	0.20	0.30
BC	0.00	0.00	0.00	0.01	0.00	0.19

Tabelle 13: Charakterisierung der Ergebnisse der Bicluster-basierten Verfahren in 1 000 Wiederholungen mit unterschiedlichen Startwerten auf dem DeNoPa-Datensatz. (a) zeigt die Reduktion der Laufzeit von FSBC bei 1 000 Wiederholungen im Vergleich zum gewöhnlichen Biclustern. In (b) und (c) werden Anzahlen gefundener Bicluster bzw. maximale Anzahlen nominierter Samples aus der Hb-Subgruppe tabelliert. Für die Beschreibung des maximalen Jaccardindizes in (d) werden die Extrema, die 25%-, 50%- und 75%-Quantile, sowie das arithmetische Mittel (MW) angegeben.

zusammen, die elf Samples nominieren, bei denen es sich um eine echte Teilmenge des ersten Clusters handelt. Keines der nominierten Samples ist in der interessierenden Hb-Subgruppe.
Losgelöst vom am häufigsten beobachteten Ergebnis der verschiedenen Läufe listet Tabelle 13(c) die maximale Anzahl korrekt nominierter Samples in den Clustern eines Laufes. Nur in 128 der 1 000 Läufe ist überhaupt eines der gesuchten Samples enthalten, davon in etwa der Hälfte nur genau eines. Entsprechend gering fallen auch die maximalen Jaccardindizes pro Lauf aus (Tabelle 13(c)). In diesem Beispiel zeigt das Biclustern zwar eine sehr geringe Variation über die Läufe, allerdings ist es ohne die vorherige Variablenselektion wie beim FSBC-Ansatz nicht möglich, die interessierende Subgruppe aufzuspüren.

7 Zusammenfassung und Diskussion der erzielten Ergebnisse

Hintergrund der Subgruppendetektion

Aufgrund der wachsenden Bedeutung der personalisierten Medizin ist das klinische und pharmazeutische Feld fortwährend an der Untersuchung neuer Hypothesen über bisher unbekannte Patientensubgruppen in heterogenen Krankheiten interessiert. Im Idealfall können im Zuge der weiteren Überprüfung dieser Hypothesen für eine Subgruppe (SG) spezifische Pathomechanismen aufgedeckt und zugehörige *drug targets* identifiziert werden. Für die Aufklärung pathologischer Mechanismen entfernt sich die Forschung zunehmend von bisherigen, starren Krankheitsgrenzen und lenkt ihren Fokus auf krankheitsübergreifende Studien. Diese Studien könnten dazu führen, dass Krebsarten beispielsweise nicht mehr primär nach ihrem Ursprungsorgan klassifiziert werden, sondern nach ihrer molekularen Pathologie. Ähnliche Ansätze gibt es auch im Feld der neurodegenerativen Krankheiten, bei denen beispielsweise Gemeinsamkeiten von Alzheimer, Huntington und Parkinson von Interesse sind.
Ein wichtiger Schritt, der die Entwicklung einiger *targeted* Therapien überhaupt erst ermöglicht hat, ist die Etablierung und der verstärkte Einsatz von Hochdurchsatz-Technologien (*omics*) wie beispielsweise Genexpressionsarrays. Die Analyse von Tausenden von Biomolekülen stellte dabei neue Herausforderungen an statistische Inferenz- und explorative Methoden. Vor diesem Hintergrund kam es beispielsweise zur Entwicklung des sogenannten q-values nach Storey und Tibshirani [83] zur Kontrolle der *false discovery rate* oder Methoden des maschinellen Lernens mit besonderem Fokus auf der Variablenselektion [84].
Allerdings konnte bisher kein Verfahren zur Detektion von Samplesubgruppen etabliert werden, das seinen Weg in die klinische Anwendung gefunden hätte. Stattdessen besteht häufig der Irrglaube, dass allgemeine Methoden wie das hierarchische Clustern oder eine Hauptkomponentenanalyse der Daten geeignet sind, um auch kleinere Samplesubgruppen mit Abweichungen in einer geringen Anzahl von Features zu detektieren. Im Allgemeinen ist dies aber nicht der Fall, weil diese Methoden gerade darauf abzielen, die globale Struktur der Daten zu beschreiben und kleinere Abweichungen zu vernachlässigen.
Um dieses Problem zu umgehen, wurde seit Anfang der 2000er Jahre eine Reihe von univariaten Methoden veröffentlicht, die es erlauben sollen, aus hochdimensionalen Datensätzen die Variablen zu identifizieren, die auf eine Patientensubgruppe hinweisen (z. B. [17]-[21]). Allerdings wurden die Artikel vorwiegend in statistischen Journalen veröffentlicht, sodass weder die Notwendigkeit für spezifische Methoden zur Subgruppendetektion noch die Diskussion um Vor- und Nachteile unterschiedlicher Ansätze im Feld der klinischen Anwender bekannt wurde.

Univariate Methoden zur Identifikation von subgruppenanzeigenden Variablen

In dieser Arbeit wurden zunächst verschiedene bestehende Ansätze zur univariaten Subgruppendetektion in einer umfassenden Simulationsstudie (SimUni) verglichen. Dazu zählen unter anderem die *profile analysis using clustering and kurtosis* [18], die *outlier sum* [19] und die *outlier robust t-statistic* [20] oder die *percentile analysis for differential gene expression* [21]. Außerdem wurde die **neue univariate Methode Fisher Sum (FS)** definiert und in den Vergleich aufgenommen. Die wesentlichen Ergebnisse wurden zuvor in Ahrens et al. [22] präsentiert. Für die FS-Berechnung wird zunächst nach der Zentrierung aller Beobachtungen um den Median der als homogen angenommenen Gruppe (z. B. gesund) in einem Zwei-Gruppen-Vergleich ein Quantil der potentiell heterogenen Gruppe (z. B. krank) als cut-off ausgewählt. Als Score herangezogen wird die Differenz der pro Gruppe aufsummierten Werte oberhalb dieses Quantils (Standardwert ist das 90%-Quantil). Somit ergeben sich große Werte für FS, falls die höchsten Werte der Gruppe krank im Schnitt „deutlich“ über den höchsten Werten der gesunden Gruppe liegen. Dies gilt sowohl im Falle von Subgruppen, wenn sich also nur eine kleine Menge der Samples der kranken Gruppe von allen übrigen Beobachtungen unterscheidet, als auch bei homogenen Shifts zwischen den beiden Gruppen. Im Gegensatz zu einigen vergleichbaren Methoden (z. B. *outlier sum* oder *outlier robust t-statistic*) wird eine Skalierung der Variablen bei FS standardmäßig nicht vorgenommen. So liegt der Fokus auf größeren absoluten Abweichungen der Subgruppe zu den übrigen Beobachtungen anstatt auf dem relativen Abstand der Subgruppe bezüglich der Verteilung der restlichen Werte.

In **SimUni** wurden erstmals unterschiedliche Verteilungen für die Beobachtungen der Subgruppe untersucht und für jedes dieser Szenarien ein breiter Bereich von Parameterwerten berücksichtigt. Zum Vergleich der Methoden wurden zunächst die ROC-Kurven (*receiver operating characteristics*-Kurven) aus dem methodenspezifischen Score und wahren SG-Status (SG vorhanden/nicht vorhanden) der einzelnen Variablen berechnet. Da für jede betrachtete Kombination aus der Verteilung der SG, der Fallzahl und dem SG-Anteil die Performanz für wachsende Unterschiede zwischen der Subgruppe und den übrigen Beobachtungen von Interesse war, wurde für die bessere Vergleichbarkeit als Gütekriterium die AUC (*area under the curve*), die Fläche unter der ROC-Kurve, verwendet.
In der Simulation zeigte sich, dass der Wechsel einer reinen Shift-Alternative zu einer Shift-Scale-Alternative für die Subgruppe zwar zu quantitativen Unterschieden in der Trennungsgüte zwischen Variablen mit bzw. ohne vorliegende Subgruppe führt, dass sich am Ranking der Methoden dadurch allerdings keine wesentlichen Änderungen ergeben. Durch die Abdeckung eines größeren Spektrums von

Parameterwerten konnten unterschiedliche Bereiche abgegrenzt werden, in denen jeweils verschiedene Methoden die beste Performanz zeigten.
Im Gegensatz zu den üblichen Vergleichen der konkurrierenden Methoden berücksichtigt SimUni als Referenzmethode zusätzlich das Likelihood-Ratio (LR). So können die unterschiedlichen Verfahren nicht nur untereinander, sondern auch mit einer theoretisch optimalen Methode verglichen werden. Dies ermöglicht die Identifikation von Parameterbereichen mit schwacher Performanz der Methoden, um entweder bestehende Verfahren anzupassen oder neue Methoden für spezifische Bereiche, zum Beispiel kleine Abweichungen der Subgruppe, zu entwickeln.
Eine weitere Besonderheit des SimUni-Designs im Vergleich zu bisher veröffentlichten Simulationsstudien im Bereich der Subgruppendetektion ist die Berücksichtigung eines zusätzlichen Expressionsprofils, das als *nicht-krankheitsspezifische (nks) Subgruppe* bezeichnet wird.
In den Simulationen zeigte sich, dass die neue Methode FisherSum (FS) für die Detektion kleiner Subgruppen bis etwa 30% zu empfehlen ist, besonders aufgrund ihrer höheren Robustheit gegenüber nks Subgruppen. Bei größeren Subgruppenanteilen erzielte meist der gewöhnliche t-Test die besten Ergebnisse. Bei festem Subgruppenanteil zeigte sich bei größeren Fallzahlen insgesamt eine bessere Performanz der Methoden, jedoch keine Änderungen in der Rangfolge. In der Anwendung auf einen realen Datensatz (ParkCHIP) bestätigten sich die Simulationsergebnisse hinsichtlich der fehlenden Robustheit einiger Methoden gegenüber den nks Subgruppen.

Multivariate Methoden zur Subgruppendetektion

Aufgrund der vielversprechenden Ergebnisse der Fisher Sum auf simulierten und realen Daten wurde diese univariate Scoringmethode als Basis für die **Entwicklung des multivariaten FSx-Workflows** zur Subgruppendetektion ausgewählt. Zunächst werden die Variablen mit höchsten FS-Scores hinsichtlich übereinstimmend angezeigter Subgruppen verglichen, um möglicherweise die Evidenz für die Subgruppe zu verstärken. Vorgeschlagen werden für diesen Vergleichsschritt zwei verschiedene Ähnlichkeitsmaße, die zugehörigen Workflowvarianten werden als FSOL [55] bzw. FSJ bezeichnet. OL steht dabei für Ordered List, eine ursprünglich für den Vergleich geordneter Genlisten entwickelte Methode [56]. Alternativ werden bei FSJ zwei Variablen anhand des Verhältnisses von Schnitt und Vereinigung der Samples mit den jeweils höchsten Expressionswerten verglichen. Nach der Bestimmung der paarweisen Ähnlichkeiten für die top-FS-Variablen wird durch die Wahl eines Schwellenwertes eine Aufteilung der Variablen in einzelne Gruppen erreicht. Für jede dieser Gruppen kann dann eine Samplemenge als potentielle Subgruppe nominiert werden.
Die so entwickelte Methode FSx wurde wiederum mithilfe einer Simulationsstu-

die und durch die Anwendung auf reale Datensätze charakterisiert. Als Referenzmethode für den FSx-Workflow wurde das Biclustern (BC) ausgewählt. Dabei handelt es sich um einen Ansatz, der genau für den Zweck entwickelt wurde, aus hochdimensionalen Daten Mengen von Samples zu detektieren, die sich nur in einem Teil der Variablen ähneln. Für die Anwendung auf Genexpressionsdaten oder andere omics-Technologien hat sich dabei der Plaid-Algorithmus bewährt. Dieser ist jedoch nicht deterministisch und die Ergebnisse verschiedener Läufe unterscheiden sich mitunter stark. In der Literatur wurde daher die wiederholte Berechnung des BC-Algorithmus mit verschiedenen Startwerten empfohlen. Die anschließende Verwendung sogenannter Ensemblemethoden soll dann die Ergebnisse der unterschiedlichen Läufen kondensieren und ein repräsentatives Konsens-(Bi-)Clusterergebnis liefern.
Die Optimierung des Referenzworkflows hinsichtlich verschiedener Kombinationen von Clusteralgorithmen und nachgeschalteten Ensemblemethoden konnte im Rahmen dieser Arbeit nicht geleistet werden. Stattdessen wurde jeweils für die Simulation und die Anwendung auf reale Daten ein vereinfachter Ansatz gewählt. Die hier gezeigten Ergebnisse des Biclusters bieten daher sicherlich Raum für Verbesserungen. Trotzdem lassen sich wesentliche Eigenschaften des Biclusterns beispielsweise hinsichtlich der zu detektierenden Gruppengröße erkennen.
Als vierte multivariate SG-Detektionsmethode wurde mit FSBC die Kombination aus FSx und Biclustern in den Vergleich aufgenommen. Dabei wird der Biclusteralgorithmus auf die Teilmatrix der top-50-FS-Variablen angewendet, die auch im FSx-Workflow betrachtet wird. Durch die erhebliche Reduktion der Dimension der Datenmatrix und damit des Suchraums im Optimierungsprozess sollte die Variation der Biclusterergebnisse wesentlich gesenkt werden.

In der Simulationsstudie **SimMulti** zum Vergleich der multivariaten SG-Detektionsmethoden wurde pro Lauf eine Datenmatrix generiert, die den Vergleich zweier Gruppen basierend auf einer Hochdurchsatztechnologie repräsentiert. Die p Zeilen wurden als Features aufgefasst und die jeweils n Spalten als Samples einer Gruppe. Die Einträge wurden zum größten Teil aus der Standardnormalverteilung gezogen, für eine kleine Teilmatrix jedoch aus einer Normalverteilung mit positivem Erwartungswert δ ($\delta = 2,3,4,6$). Die Zeilen der Datenmatrix, in die diese Teilmatrix fiel, stellten somit die subgruppenanzeigenden Variablen dar: Bis auf eine kleine Gruppe von Beobachtungen mit erhöhten Werten besteht kein Unterschied zwischen den beiden verglichenen Gruppen. Auf jede so generierte Matrix der Größe $p \times (2n)$ wurden die Methoden FSOL, FSJ, BC und FSBC angewendet.
Jede Methode kann sowohl subgruppenanzeigende Variablengruppen als auch potentielle Samplesubgruppen nominieren. Für den Gütevergleich der Methoden

wurde die Menge der jeweils nominierten Samples mithilfe des Jaccardindizes mit der Menge der Samples verglichen, deren Beobachtungen tatsächlich aus der Verteilung mit höherem Erwartungswert gezogen wurden. Für jeden Shift μ wurden auf diese Weise 500 Datensätze generiert und die Verteilungen der von den Methoden erzielten Jaccardindizes verglichen.
In den Simulationen wurden verschiedene Parameter variiert und ihr Einfluss auf die vier Methoden untersucht. Zum einen handelte es sich um datensatzspezifische Parameter, die vom Anwender im Allgemeinen nicht beeinflusst werden können: die Featureanzahl p des Datensatzes, die Anzahl n der Samples in den zu vergleichenden Gruppen, die Größe n_{SG} der Samplesubgruppe, sowie die Anzahl p_{SG} der Variablen, in denen der Erwartungswert erhöht wurde. Außerdem wurden für die neuen Workflows FSOL und FSJ verschiedene workflowspezifische, vom Anwender wählbare Parameter auf die gleiche Weise untersucht. Dazu zählten die Anzahl T der im ersten Schritt ausgewählten Variablen mit interessantestem Verteilungsmuster, sowie zwei Parameter zur Berechnung der Variablenähnlichkeit hinsichtlich der angezeigten Subgruppe.

In den durchgeführten Simulationen zeigte das etablierte Biclustern enttäuschende Performanz, wenn sich Samplesubgruppen nur in einer geringen Anzahl von Variablen zeigen. Dabei scheint die Ursache nicht in der fehlenden Optimierung der Parameter oder nicht verwendeter Ensemblemethoden zu liegen, sondern grundsätzlich in der mangelnden Sensitivität gegenüber Subgruppen in kleinen Variablenmengen. Dies wurde durch die Variation des Parameters $p_{SG} = 5, 20, 50$ sehr deutlich. Für die maximale betrachtete Anzahl wurde für das Biclustern die beste Detektionsgüte über die wachsenden Shifts in der gesamten SimMulti-Studie beobachtet. In realen Daten wird allerdings eher selten eine so große Anzahl von Variablen mit subgruppenanzeigenden Expressionsmustern gefunden. Daher lag der Fokus der SimMulti-Studie auf den vom Biclustern nur unzureichend zu detektierenden kleinen Variablengruppen mit $p_{SG} = 5$.
Die drei neuen Methoden FSOL, FSJ und FSBC mit univariater Vorselektion der Variablen hingegen zeigten sich bei moderaten Shifts ($\delta \geq 3$) durchaus in der Lage, die gesuchten Subgruppen auch für kleines p_{SG} recht zuverlässig zu identifizieren. Bei passenden Parametereinstellungen zeigte FSJ gute Performanz und Robustheit gegenüber den meisten untersuchten Parametern. Allen drei Methoden ist jedoch gemein, dass es bei ausnehmend ungünstigen Parameterkombinationen zu deutlichen Einbußen in der Performanz kommt. In der praktischen Anwendung lassen sich diese Fälle jedoch im Allgemeinen durch Betrachtung der Zwischenergebnisse der Workflows erkennen und die Workflowparameter entsprechend anpassen.

Anwendungsbeispiele

In dieser Arbeit wurden insgesamt drei reale Datensätze verwendet. Zunächst wurden verschiedene univariate Methoden auf den **ParkCHIP**-Datensatz angewendet um die Auswahl einer univariaten Methode zu unterstützen, die als Ausgangspunkt für die Entwicklung des Subgruppendetektionsworkflows FSx dient. Für die Beurteilung der Performanz dieses neuen Workflows sowie des Referenzworkflows wurde auf Datensätze zurückgegriffen, für die bereits im Vorhinein eine interessierende Samplesubgruppe definiert war. Die von diesen Subgruppen betroffenen Variablenmengen waren im Vorhinein nicht vollständig bekannt. Dadurch kann die Detektionsgüte nur hinsichtlich der Samplemengen erfolgen, nicht jedoch für die Variablen. Die verwendeten Datensätze ALL und DeNoPa umfassten jeweils insgesamt etwa 80 Samples, unterschieden sich entsprechend der zugrundeliegenden omics-Technologie jedoch in der Variablenanzahl. Die ALL-Daten stammten von einem Genexpressionschip und enthielten rund 12 000 Variablen, während für die DeNoPa-Samples mittels labelfreier Massenspektrometrie die relative Abundanz von etwa 900 Proteinen analysiert wurde.

Aufgrund der verfügbaren Annotation mit molekularbiologischen Informationen der enthaltenen Samples konnte aus dem **ALL**-Datensatz ein Zwei-Gruppen-Vergleich konstruiert werden, der sich besonders für die Gütebeurteilung der Subgruppendetektionsmethoden eignete. Benötigt wurden zwei ausreichend große zu vergleichende Gruppen, von denen eine genau eine bekannte Subgruppe enthält, d. h. eine Menge von Samples, die sich in einer Menge von Features in ihrem Expressionslevel von allen übrigen Samples unterscheidet. Ausgewählt wurden die drei Gruppen NEG, E2A/PBX1 und BCR/ABL des Faktors `mol.biol`, der die molekularen Rearrangements der Samples codiert. NEG bezeichnet dabei die Gruppe ohne bekannte Mutation, E2A/PBX1 und BCR/ABL die Gruppen, bei denen die jeweiligen Fusionstranskripte gefunden wurden. Alle vier untersuchten Subgruppendetektionsmethoden (FSOL, FSJ, BC und FSBC) wurden auf den Vergleich

42 NEG vs. (37 BCR/ABL und 5 E2A/PBX1)

angewendet, um Subgruppen in der zusammengesetzten Gruppe zu identifizieren. Die nominierten Subgruppen wurden jeweils auf Hinweise auf die E2A/PBX1-Subgruppe untersucht. Da im Gegensatz zu den Simulationsstudien die Existenz weiterer Subgruppen nicht auszuschließen war, war es nicht notwendig, dass im Falle mehrerer potentieller Subgruppen die gesuchte Gruppe den besten Rang belegt. Stattdessen war von Interesse, ob grundsätzlich Hinweise auf die Subgruppe gefunden werden konnten.
Die beiden FSx-Varianten führten zu ähnlichen Ergebnissen bzgl. der Gruppie-

rung der Variablen und der anschließenden Nominierung. Mit den Defaulteinstellungen ergab sich in diesem Beispiel die Variablengruppe für FSOL aus elf Variablen als Obermenge der sieben Variablen, die in FSJ gruppiert wurden. FSOL nominierte exakt die fünf gesuchten Samples als potentielle Subgruppe, bei FSJ wurden zwei zusätzliche Samples nominiert. Somit erreichten die FSx-Methoden bzgl. der gesuchten Subgruppe Jaccardindizes von 1.0 bzw. 0.71.
Bei der wiederholten Anwendung der FSBC-Methode zeichnete sich deutlich die häufig reportete Variablengruppe aus KANK1, PBX1, FAT1, NID2, MTCL1, CRYM und TERF2 ab. Bei der Betrachtung der Häufigkeiten von zunächst Clustergrößen (bzgl. Variablen) und dann Variablen- sowie zugehörigen Samplekombinationen zeigte sich die gesuchte Subgruppe in diesem Fall auch ohne automatisierte Ensemblemethoden recht deutlich. In diesem Beispiel stimmten somit die von den multivariaten Methoden nominierten potentiellen Subgruppen zumeist gut mit der gesuchten, zuvor definierten Subgruppe überein.
Im Gegensatz zu den drei anderen Methoden schien das Biclustern nicht geeignet, um die interessierende Subgruppe aufzuspüren. Von den fünf gesuchten Samples wurden in allen detektierten Biclustern der 1 000 durchgeführten Läufe maximal zwei gemeinsam nominiert, die Jaccardindizes der entsprechenden Cluster liegen unter 0.2.
Eine Möglichkeit für zukünftige Analysen des ALL-Datensatzes stellt die Untersuchung der Gruppe NEG auf eventuell enthaltene Subgruppen dar. Dass für die Samples keine bekannten Mutationen vorliegen, schließt die Tatsache nicht aus, dass Subgruppen von bislang unbekannten und somit nicht getesteten Mutationen enthalten sind. In der hier durchgeführten Analyse konnte die gesuchte Subgruppe trotz möglicherweise in NEG vorhandenen Subgruppen von FSOL, FSJ und FSBC detektiert werden. Im Gegensatz zu diesen drei Ansätzen ist das gewöhnliche Biclustern nicht auf Variablen beschränkt, die Hinweise auf Subgruppen in einer vorgebenen Gruppe zeigen. Hier könnten Heterogenitätseffekte in der als homogen betrachteten Gruppe NEG zur schlechten Performanz beigetragen haben.

Die vier Methoden zur Subgruppendetektion wurden außerdem anhand des **DeNoPa**-Datensatzes untersucht. Hierbei handelt es sich um Messungen der Proteinabundanz mithilfe massenspektrometrischer Methoden in Gesundkontrollen und Parkinson-Patienten. Die interessierende Samplesubgruppe wurde dabei über die Konzentration von Hämoglobin (Hb) in den vermessenen CSF-Proben definiert, die zuvor mit einer unabhängigen Technologie (ELISA) bestimmt wurden. Von der resultierenden Hb-Subgruppe lagen drei Samples in der Gruppe der Patienten und eines in den Kontrollen. Hier war daher von besonderem Interesse, ob die standardmäßige Korrektur bei nicht-krankheitsspezifischen (nks) Subgruppen der FS-Methode bereits so streng gewählt ist, dass die SG-anzeigenden Variablen

(hauptsächlich sogenannte Untereinheiten des Hämoglobins) einen geringen Score zugewiesen bekommen.
Der Vergleich der Methoden führte zu einem ähnlichen Ergebnis wie auf dem ALL-Datensatz. Das Biclustern war nicht geeignet, um die Hb-Subgruppe zu detektieren. Meist befand sich keines der Samples in einem Bicluster, nur in 128 der 1 000 Läufe war überhaupt eines der gesuchten Samples in einem Cluster enthalten. Im Gegensatz dazu enthielten die nominierten Subgruppen der Methoden nach FS-Selektion (FSOL, FSJ, FSBC) die gleichen drei Samples der gesuchten Subgruppe. Durch die Nominierung zusätzlicher Samples lagen die erreichten Jaccardindizes mit 0.27 bzw. 0.375 jedoch deutlich unterhalb der Werte im ALL-Beispiel. Die Betrachtung der annotierten Variablen lieferte für FSOL, FSJ, FSBC nicht nur deutliche Hinweise auf die gesuchte Hämoglobingruppe, sondern auch auf eine Keratin-Subgruppe. Keratin ist ein in der obersten Hautschicht des Menschen enthaltener Stoff, der häufig als Kontaminante während der Aufbereitung im Labor in die Proben gelangt. In diesem Fall könnte es sich, wie auch beim Hämoglobin vermutet, um eine Verunreinigung durch Hautzellen des jeweiligen Patienten handeln, die während der Lumbalpunktion eintritt. Allerdings wurden schon häufiger verschiedene Keratinformen als diagnostische Biomarker vorgeschlagen, unter anderem KRT9 für Multiple Sklerose und Neuromyelitis optica [85] oder für die ebenfalls neurodegenerative Krankheit Alzheimer [86]. Für letztere wurde das KRT9 wie in der DeNoPa-Studie in CSF-Proben nachgewiesen. Es könnte sich demnach bei der zusätzlich von den Methoden gefundenen Variablengruppe (die ebenfalls KRT 9 enthält) um relevante Ergebnisse für die Charakterisierung einer bisher unbekannten Parkinsonsubgruppe handeln.
Beim ALL-Beispiel zeigte vor allem der Expressionsplot der subgruppendefinierenden Variable PBX1 das gewünschte Subgruppenmuster, in dem sich alle Samples der definierten Subgruppe deutlich vom Rest der Samples unterschieden. Betrachtet man hingegen die Abundanzplots der vier mit Hämoglobin annotierten Variablen im DeNoPa-Datensatz (s. Anhang, Abb. 62), so zeigt sich eine geringere Übereinstimmung zwischen der als Goldstandard behandelten ELISA-Messung und den labelfreien Daten. Das einzige von den Methoden nicht nominierte Subgruppensample zeigt in keiner der relevanten Variablen eine erhöhte Abundanz, dafür haben einige weitere Samples aus der PD-Gruppe konsistent auffällig hohe Werte. Nach der visuellen Inspektion der Plots kommt man somit zu dem Schluss, dass die Nominierungen entsprechend der Datenbasis durchaus nachvollziehbar sind. Die geringen Jaccardindizes sind demnach auch auf eine geringe Konkordanz der ELISA- und MS-Messungen zurückzuführen. An dieser Stelle kann nicht geklärt werden, ob die ELISA-Messung nicht sensitiv genug ist, um die im MS-Experiment erkennbare Subgruppe zu detektieren oder ob es sich dabei wirklich um falsch positive Kandidaten handelt.

Mögliche Adaptionen des FSx-Workflows und der Simulationen

Sowohl für die in dieser Arbeit vorgestellten Workflows als auch für die Designs der Simulationsstudien sind aufgrund der komplexen Zielsetzung verschiedenste Adaptionen denkbar. Darunter fallen kleinere Änderungen wie eine Anpassung der Standardeinstellungen der Workflowparameter oder größere wie der Wechsel zu anderen Referenzmethoden, Gütekriterien oder einer aufwendigeren Struktur der Datenmatrix in der Simulation. Im Folgenden werden kurz einige Anregungen für weitergehende Analysen gegeben.

Bezüglich der neuen **Workflows FSOL und FSJ** wurden mögliche Anpassungen der Parameterwerte in den entsprechenden methodischen und anwendungsbezogenen Abschnitten dieser Arbeit beschrieben. Bei Kombination der FS-Selektion mit der Ordered-List-Methode kann statt des simulierten p-Wertes p_{OL} auch auf die (deterministische) Teststatistik *wos* (*weighted overlap score*) zurückgegriffen werden. So könnten die möglicherweise auftretenden Änderungen bei der Ähnlichkeitsbestimmung umgangen und die benötigte Rechenzeit verkürzt werden. Vorstudien zeigten jedoch, dass diese Variationen im Ergebnis bei ausreichend hoher Anzahl von Permutationen klein sind. Zudem lag die Berechnungsdauer in beiden Datenbeispielen unter einer Minute (je rund 80 Samples und 1 000 bzw. 12 000 Variablen). Ein wesentlicher Vorteil bei der Verwendung des p-Wertes statt des *wos* besteht in der deutlicheren Trennung der Variablengruppen im jeweils resultierenden Dendrogramm, was sowohl die manuelle Auswertung einzelner Datensätze als auch die automatisierte Auswertung in den Simulationen erleichtert.

Auch wenn sich in den Vergleichen der univariaten Methoden Fisher Sum sowohl in der Simulation als auch in der Anwendung auf reale Daten bewährt hat, ließe sich zur Vorauswahl der Variablen auch auf eine andere **univariate Scoringmethode** zurückgreifen. Unterschiede bei der Variablenauswahl würden sich dabei z. B. für die *outlier robust t-statistic* oder *outlier sum* durch die höhere Sensitivität der gegenüber nks Subgruppen oder der Beurteilung des relativen statt absoluten Abstands der Subgruppen (durch Skalierung) ergeben. Ein Vorteil der *outlier sum* würde ggf. bei der Nominierung der Subgruppen aus einzelnen Variablen entstehen, da sie im Gegensatz zur Fisher Sum direkt zur Nominierung einer Subgruppe flexibler Größe geeignet ist.

Wie beispielsweise im ALL-Beispiel deutlich wurde, kann schon die Reduktion des Gesamtdatensatzes mithilfe eines univariaten Scores zur Gewinnung von Hinweisen auf zuvor unentdeckte Subgruppen beitragen. Jede der zu diesem Zweck gewählten Methode kann anschließend mit einem der hier gezeigten oder einem alternativen geeigneten Ähnlichkeitsmaß zum paarweisen Vergleich oder wie bei FSBC mit einer eigenständigen multivariaten Methode kombiniert werden.

Die für das Ähnlichkeitsmaß in FSx verwendete Ordered-List-Methode ist nur ei-

ne von mehreren möglichen zum Vergleich geordneter Genlisten. Bis heute werden Arbeiten zu diesem Thema verfasst, da nun auch vermehrt technologieübergreifende Vergleiche z. B. aus RNA-Seq und Microarrayexperimenten von Interesse sind [87]. Alternativen zum Ordered-List-Algorithmus stellen beispielsweise CORaL, *comparison of ranked lists for analysis of gene expression data* von Antosh et al. [88] oder *rank-rank hypergeometric overlap* (RRHO) von Plaisier et al. [89] dar. Ein wesentliches Merkmal von CORaL ist die Berücksichtigung von unterschiedlich großen Mengen differentieller Gene in den beiden zu vergleichenden Listen. So werden die top m Ränge der ersten Liste mit den top n Rängen der zweiten Liste verglichen und in einem Optimierungsprozess wird die „beste“ Kombination der Listenlängen bestimmt.
Im Rahmen der vorliegenden Arbeit sollen Variablenpaare hinsichtlich der Samplemenge mit den jeweils höchsten Expressionswerten verglichen werden. Das Ziel ist die Bestimmung von Variablengruppen, die auf die gleichen Subgruppen, insbesondere also auf Subgruppen gleicher Größe hinweisen. Daher scheint der speziellere Vergleich in Ordered List mit der Betrachtung von Sets jeweils gleicher Größe für diese Fragestellung besser geeignet. Generell liegt der Fokus bei RRHO eher auf der Visualisierung der gefundenen Ähnlichkeiten und wird als Ergänzung zu anderen vergleichenden Methoden vorgeschlagen. Wie bereits für CORaL beschrieben, ist auch der bei RRHO vorgesehene Vergleich unterschiedlich großer Mengen von Toprängen für die zu beantwortende Fragestellung nicht notwendig. Bei großen Stichproben und einer komplexen Subgruppenstruktur (größere erwartete Anzahl von Subgruppen, ggf. Teilmengen voneinander) könnten die allgemeineren Ähnlichkeitsvergleiche von CORaL und RRHO unter Umständen nützlich sein. RRHO ist über Bioconductor als gleichnamiges **R**-Paket verfügbar [90].

Das **Design der SimMulti-Studie** lässt sich in weiterführenden Arbeiten nahezu beliebig komplex gestalten, hier seien nur einige Möglichkeiten aufgeführt:

- stochastischer Shift der Subgruppe in verschiedenen Variablen
- Abhängigkeitsstruktur der Variablen (sodass diese z. B. Gruppen von korrelierten Gene darstellen)
- mehrere Patientensubgruppen (disjunkt oder überlappend, mit gleichen oder unterschiedlichen Shifts). Diese Erweiterung würde Änderungen an den Bewertungskriterien in der Simulation erfordern.

Zusätzlich zu den in dieser Arbeit betrachteten Shiftwerten $\delta = 2,3,4,6$ wäre auch eine Analyse kleinerer Shifts denkbar, um zu bestimmen, an welchem Punkt die verglichenen Methoden nicht mehr in der Lage sind, die wahren Samplesubgruppen und die auf sie hinweisenden Variablengruppen zuverlässig zu identifizieren. Wie in den Sensitivitätsanalysen in Abschnitt 5.4.1 bereits deutlich am Beispiel des Biclusterns gezeigt wurde, ist dieser „Bruchpunkt“ jedoch abhängig von der

Kombination verschiedener Parameter, neben δ spielen vor allem die Anzahlen n_{SG} und p_{SG} der betroffenen Sample- bzw. Variablenanzahl eine Rolle.
Nicht nur in den Simulationen sondern auch bei der Bewertung der Detektionsgüte in überwachten Anwendungsbeispielen mit bekannten Subgruppen können andere Optionen für das **Gütekriterium** in Betracht gezogen werden. Gerade für die in dieser Arbeit untersuchten Subgruppen mit einer geringen Anzahl von Samples ist der Effekt einzelner falsch-positiv oder falsch-negativ nominierter Samples relativ groß. Im DeNoPa-Beispiel erreichte FSOL mit drei korrekt und vier falsch-positiv nominierten Samples einen Jaccardindex von 0.375. Mit den üblicherweise in der klinischen Diagnostik verwendeten Maßen Sensitivität und Spezifität stellt sich das Ergebnis wesentlich positiver dar: Bei drei von vier als Subgruppensamples erkannten Proben ergibt sich eine Sensitivität von 0.75. Da von den 78 Samples ohne erhöhte Hb-Werte 74 korrekterweise nicht in der nominierten Subgruppe liegen, berechnet sich die Spezifität der SG-Detektion sogar zu 0.95.

Empfehlungen zur Wahl einer Methode in der Anwendung

Bisher wurde in der Literatur eine Vielzahl von Methoden zur Subgruppendetektion vorgeschlagen. Mithilfe von Simulationsstudien kann festgestellt werden, welche der Methoden welche spezifischen Subgruppenmuster am besten identifizieren können. Eine praktische Entscheidungshilfe ist dadurch allerdings nur bedingt gegeben, da für reale Datensätze die gegebenenfalls zugrundeliegende Subgruppenstruktur nicht bekannt ist. Eine mögliche Empfehlung lautet dann, einen Ansatz zu wählen, der für eine Vielzahl von Mustern ausreichend gute Ergebnisse liefert. Innerhalb der Klasse der Biclusterverfahren haben das Eren et al. [63] beispielsweise für den Plaid-Algorithmus gezeigt, weshalb er in dieser Arbeit als Repräsentant für die Biclusterverfahren gewählt wurde.
Die guten Ergebnisse des Plaid-Algorithmus konnten in der vorliegenden Arbeit für ein ähnliches Setting wie in Eren et al. [63] mit einer großen Anzahl subgruppenanzeigender Variablen ($p_{SG} = 50$) bestätigt werden. Das Biclustern schien in der Simulationsstudie gut geeignet, um Subgruppen auch mit kleiner Sampleanzahl zu erkennen, falls diese Gruppe in einer ausreichend großen Zahl von Variablen auffällige Werte zeigt. Der **Plaid-Algorithmus zeigte jedoch auf simulierten und realen Daten mangelnde Performanz bei der Detektion kleiner Subgruppen in wenigen Variablen, vor allem bei kleinem Lokationsshift der Subgruppe**. Für diesen speziellen Fall könnten vergleichende Studien für weitere Biclusteralgorithmen durchgeführt werden, um innerhalb dieser Methodenklasse einen besser geeigneten Ansatz zu finden.
Allerdings konnten in dieser Arbeit bereits drei alternative Workflows präsentiert werden, die sowohl in Simulationen als auch auf realen Daten gute Performanz zeigten. Ausgehend vom etablierten Biclusteransatz konnte durch die vorgeschal-

tete univariate FS-Selektion mit **FSBC** eine wesentliche Verbesserung erzielt werden. Eine hohe Detektionsgüte wird dabei in den Simulationen allerdings weiterhin erst für größere Shifts der Subgruppe erreicht. Weiterhin bleibt die Frage nach der besten Wahl einer Ensemblemethode offen. Diese werden im Allgemeinen benötigt, um die ggf. stark variierenden Ergebnisse unterschiedlicher Biclusterläufe zu kombinieren. Es konnte aber anhand der realen Daten gezeigt werden, dass unter Umständen auch durch einfache deskriptive Analysen einer größeren Anzahl von Clusterläufen relativ konsistente Hinweise auf vorliegende Subgruppen extrahiert werden können.
Neben den Bicluster-basierten Methoden BC und FSBC konnten in dieser Arbeit mit FSOL und FSJ zwei weitere multivariate Methoden zur Subgruppendetektion vorgestellt werden. Basierend auf den in dieser Arbeit erzielten Ergebnissen können **FSOL und FSJ für explorative Analysen zur Subgruppendetektion in hochdimensionalen Datensätzen** empfohlen werden. Dies gilt **insbesondere, wenn Subgruppen identifiziert werden sollen, die sich nur auf eine kleine Anzahl von Variablen auswirken und kleinere Shifts bewirken**. In den Simulationen zeigte der Vergleich der beiden FSx-Varianten eine Überlegenheit von FSJ. Berücksichtigt man außerdem die simple Implementierung, kürzere Rechenzeit und deterministische Natur der Methode, so scheint FSJ gegenüber FSOL vorzuziehen zu sein. Aufgrund des zunächst schlicht gehaltenen SimMulti-Designs und der geringen Anzahl bisheriger Anwendungsbeispiele sollten wenn möglich aber bis auf Weiteres beide Methoden in Betracht gezogen werden. In seltenen Fällen zeigte sich durch ungünstige Wahl der Workflowparameter eine stark eingeschränkte Performanz. Methoden zur Erkennung solcher Fälle und notwendiger Adaptionen wurden in der Arbeit besprochen. Die wichtigsten Punkte dieser Betrachtung sind in der Übersicht in Abbildung 25 zusammengefasst.
In explorativen Studien könnte auch die Eigenschaft der FSx-Ansätze als Vorteil gegenüber dem Biclustern gewertet werden, dass durch das FS-Ranking im ersten Schritt zumindest die potentiell subgruppenrelevanten Variablen des Datensatzes ersichtlich werden. Beim Biclustern hingegen wäre für Läufe ohne gefundene Bicluster kein Ansatzpunkt für weitere Analysen gegeben.

Wesentliche Ergebnisse des Vergleichs der vier Methoden BC, FSBC, FSOL und FSJ

- Für Subgruppen mit kleiner Variablenanzahl $p_{SG} = 5$ erzielen alle drei Methoden mit FS-basierter Vorselektion ähnliche, gute Performanz.
- FSBC zeigt dabei Schwächen im Bereich kleiner Lokationsshifts.
- Das Biclustern findet die Subgruppen im Gegensatz zu den anderen drei Methoden nur für sehr große Shifts.
- FSJ ist FSOL in den Simulationen überlegen, FSOL erzielte in der Anwendung höhere Jaccardindizes.
- Das Biclustern mit dem Plaid-Algorithmus erzielt unter allen vier Methoden die besten Ergebnisse für größere Mengen subgruppenanzeigender Variablen, dann sogar für kleine Shifts.
- Außer bei den kleineren betrachteten Shifts ist in diesem Fall aber die Performanz der FSJ-Methode ebenso gut.
- FSBC und BC erfordern zusätzliche Ensemblemethoden oder manuelle Extraktion der angezeigten Cluster, was die Variabilität der Ergebnisse unter Umständen zusätzlich erhöht.

Abbildung 25: Zusammenstellung der wichtigsten Aspekte aus Simulationen und Anwendungsbeispielen zum Vergleich der Subgruppendetektionsmethoden.

Literaturverzeichnis

[1] Das Prinzip Gießkanne hat bei der Behandlung von Krebspatienten ausgedient. http://www.aerztezeitung.de/medizin/krankheiten/krebs/article/590242/prinzip-giesskanne-behandlung-krebspatienten-ausgedient.html. Interview mit Prof. Wolff Schmiegel, veröffentlicht am: 24.02.2010, besucht am 19.08.2016.

[2] Committee on a Framework for Development a New Taxonomy of Disease; National Research Council. Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease. http://www.ncbi.nlm.nih.gov/books/NBK91503, 2011.

[3] D.J. Slamon, G.M. Clark, S.G. Wong, W.J. Levin, A. Ullrich und W.L. McGuire. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177–182, 1987. doi: 10.1126/science.3798106. URL http://www.sciencemag.org/content/235/4785/177.abstract.

[4] M.S. Tockman, J.L. Mulshine, S. Piantadosi, Y.S. Erozan, P.K. Gupta, J.C. Ruckdeschel, P.R. Taylor, T. Zhukov, W.H. Zhou, Y.L. Qiao und S.X. Yao. Prospective detection of preclinical lung cancer: results from two studies of heterogeneous nuclear ribonucleoprotein A2/B1 overexpression. *Clinical Cancer Research*, 3(12):2237–2246, 1997. URL http://clincancerres.aacrjournals.org/content/3/12/2237.abstract.

[5] R.B. Shah, R. Mehra, A.M. Chinnaiyan, R. Shen, D. Ghosh, M. Zhou, G.R. MacVicar, S. Varambally, J. Harwood, T.A. Bismar, R. Kim, M.A. Rubin und K.J. Pienta. Androgen-independent prostate cancer is a heterogeneous group of diseases: Lessons from a rapid autopsy program. *Cancer Research*, 64(24):9209–9216, 2004. doi: 10.1158/0008-5472.CAN-04-2442. URL http://cancerres.aacrjournals.org/content/64/24/9209.abstract.

[6] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield und E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286 (5439):531–537, 1999. doi: 10.1126/science.286.5439.531. URL http://www.sciencemag.org/content/286/5439/531.abstract.

[7] K. Strimbu und J.A. Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463–466, 2010.

[8] W.S. Dalton und S.H. Friend. Cancer Biomarkers – An invitation to the table. *Science*, 312(5777):1165–1168, 2006. doi: 10.1126/science.1125948. URL http://www.sciencemag.org/content/312/5777/1165.abstract.

[9] D.J. Irwin, J.Q. Trojanowski und M. Grossman. Cerebrospinal fluid biomarkers for differentiation of frontotemporal lobar degeneration from Alzheimer's disease. *Frontiers in Aging Neuroscience*, 5:6, 2013. ISSN 1663-4365. doi: 10.3389/fnagi.2013.00006. URL http://journal.frontiersin.org/article/10.3389/fnagi.2013.00006.

[10] H. Le-Niculescu, D.F. Levey, M. Ayalew, L. Palmer, L.M. Gavrin, N. Jain, E. Winiger, S. Bhosrekar, G. Shankar, M. Radel, E. Bellanger, H. Duckworth, K. Olesek, J. Vergo, R. Schweitzer, M. Yard, A. Ballew, A. Shekhar, G.E. Sandusky, N.J. Schork, S.M. Kurian, D.R. Salomon und A.B. Niculescu. Discovery and validation of blood biomarkers for suicidality. *Molecular Psychiatry*, 18:1249–1264, 2013. doi: 10.1038/mp.2013.95. URL http://dx.doi.org/10.1038/mp.2013.95.

[11] G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:1–25, 2004. doi: 10.2202/1544-6115.1027.

[12] C.W. Law, Y. Chen, W. Shi und G.K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15 (2):1–17, 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-2-r29. URL http://dx.doi.org/10.1186/gb-2014-15-2-r29.

[13] K. Kammers, R.N. Cole, C. Tiengwe und I. Ruczinski. Detecting significant changes in protein abundance. *EuPA Open Proteomics*, 7:11 – 19, 2015. ISSN 2212-9685. doi: http://dx.doi.org/10.1016/j.euprot.2015.02.002. URL http://www.sciencedirect.com/science/article/pii/S2212968515000069.

[14] M.H.W. Starmans und P.C. Boutros. Biomarkers and subtypes of cancer. *Aging*, 7(5):280–281, 11 2015.

[15] M.Y. Wu, D.Q. Dai, X.F. Zhang und Y. Zhu. Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm. *PLoS ONE*, 8(6):1–15, 6 2013. doi: 10.1371/journal.pone.0066256. URL http://dx.doi.org/10.1371%2Fjournal.pone.0066256.

[16] J.S. Ikonomidis, C.R Ivey, J.B. Wheeler, A.W. Akerman, A. Rice, R.K. Patel, R.E. Stroud, A.A. Shah, C.G. Hughes, G. Ferrari, R. Mukherjee und Jones J.A. Plasma biomarkers for distinguishing etiologic subtypes of thoracic aortic aneurysm disease. *The Journal of thoracic and cardiovascular surgery*, 145(5):1326–33, 2013. doi: 10.1016/j.jtcvs.2012.12.027.

[17] S.A. Tomlins, D.R. Rhodes, S. Perner, S.M. Dhanasekaran, R. Mehra, X.W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J.E. Montie, R.B. Shah, K.J. Pienta, M.A. Rubin und A.M. Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, 2005. doi: 10.1126/science.1117679. URL `http://www.sciencemag.org/content/310/5748/644.abstract`.

[18] A.E. Teschendorff, A. Naderi, N.L. Barbosa-Morais und C. Caldas. PACK: Profile analysis using clustering and kurtosis to find molecular classifiers in cancer. *Bioinformatics*, 22(18):2269–2275, 2006. doi: 10.1093/bioinformatics/btl174. URL `http://bioinformatics.oxfordjournals.org/content/22/18/2269.abstract`.

[19] R. Tibshirani und T. Hastie. Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1):2–8, 2007. doi: 10.1093/biostatistics/kxl005. URL `http://biostatistics.oxfordjournals.org/content/8/1/2.abstract`.

[20] B. Wu. Cancer outlier differential gene expression detection. *Biostatistics*, 8(3):566–575, 2007. doi: 10.1093/biostatistics/kxl029. URL `http://biostatistics.oxfordjournals.org/content/8/3/566.abstract`.

[21] L. Li, A. Chaudhuri, J. Chant und Z. Tang. PADGE: analysis of heterogeneous patterns of differential gene expression. *Physiological Genomics*, 32 (1):154–159, 2007. doi: 10.1152/physiolgenomics.00259.2006. URL `http://physiolgenomics.physiology.org/content/32/1/154.abstract`.

[22] M. Ahrens, M. Turewicz, S. Casjens, C. May, B. Pesch, C. Stephan, D. Woitalla, R. Gold, T. Brüning, H.E. Meyer, J. Rahnenführer und M. Eisenacher. Detection of patient subgroups with differential expression in omics data: A comprehensive comparison of univariate measures. *PLoS ONE*, 8 (11):e79380, 11 2013. doi: 10.1371/journal.pone.0079380. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0079380`.

[23] L. Lazzeroni und A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2002.

[24] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz und R. Foa. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778, 2004. doi: 10.1182/blood-2003-09-3243. URL `http://bloodjournal.hematologylibrary.org/content/103/7/2771.abstract`.

[25] X. Li. *ALL: A data package*, 2009. R package version 1.4.16.

[26] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly und R.A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739, 2010.

[27] M. Turewicz, C. May, M. Ahrens, D. Woitalla, R. Gold, S. Casjens, B. Pesch, T. Brüning, H.E. Meyer, E. Nordhoff, M. Böckmann, C. Stephan und M. Eisenacher. Improving the default data analysis workflow for large autoimmune biomarker discovery studies with ProtoArrays. *Proteomics*, 13:2083–2087, 2013.

[28] A. Kauffmann, R. Gentleman und W. Huber. arrayQualityMetrics – a Bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, 2009. doi: 10.1093/bioinformatics/btn647. URL `http://bioinformatics.oxfordjournals.org/content/25/3/415.abstract`.

[29] A. Alexa, J. Rahnenführer und T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 2006. doi: 10.1093/bioinformatics/btl140. URL `http://bioinformatics.oxfordjournals.org/content/22/13/1600.abstract`.

[30] J. Lyons-Weiler, S. Patel, M.J. Becich und T.E. Godfrey. Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics*, 5(1):1–9, 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-110. URL `http://dx.doi.org/10.1186/1471-2105-5-110`.

[31] H. Lian. MOST: detecting cancer differential gene expression. *Biostatistics*, 9(3):411–418, 2008. doi: 10.1093/biostatistics/kxm042. URL `http://biostatistics.oxfordjournals.org/content/9/3/411.abstract`.

[32] Y. Wang und R. Rekaya. LSOSS: Detection of cancer outlier differential gene expression. *Biomarker Insights*, 5:69–78, 2010.

[33] J. Hu. Cancer outlier detection based on likelihood ratio test. *Bioinformatics*, 24(19):2193–2199, 2008.

[34] L.A. Chen, D.T. Chen und W. Chan. The distribution-based p-value for the outlier sum in differential gene expression analysis. *Biometrika*, 97 (1):246–253, 2010. doi: 10.1093/biomet/asp075. URL `http://biomet.oxfordjournals.org/content/97/1/246.abstract`.

[35] W.N. van Wieringen, M.A. van de Wiel und A.W. van der Vaart. A test for partial differential expression. *Journal of the American Statistical Association*, 103(483):1039–1049, 2008. doi: 10.1198/016214507000001319. URL `http://dx.doi.org/10.1198/016214507000001319`.

[36] B. Love. *Functional Protein Microarrays in Drug Discovery*. CRC Press, 2007. Chapter 21. The Analysis of Protein Arrays.

[37] S. Patel und J. Lyons-Weiler. caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer. *Applied Bioinformatics*, 3(1):49–62, 2004.

[38] D. Kihara, Y.D. Yang und T. Hawkins. Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. *Cancer Informatics*, 2:25–35, 2006.

[39] M. Turewicz, M. Ahrens, C. May, K. Marcus und M. Eisenacher. PAA: An R/Bioconductor package for biomarker discovery with protein microarrays. *Bioinformatics*, 2016. doi: 10.1093/bioinformatics/btw037. URL `http://bioinformatics.oxfordjournals.org/content/early/2016/01/22/bioinformatics.btw037.abstract`.

[40] H. Vuong, K. Shedden, Y. Liu und D.M. Lubman. Outlier-based differential expression analysis in proteomics studies. *Journal of Proteomics and Bioinformatics*, 4(6):116–122, 2011.

[41] J.W. MacDonald und D. Ghosh. COPA – cancer outlier profile analysis. *Bioinformatics*, 22(23):2950–2951, 2006. doi: 10.1093/bioinformatics/btl433. URL `http://bioinformatics.oxfordjournals.org/content/22/23/2950.abstract`.

[42] A. von Heydebreck, W. Huber, A. Poustka und M. Vingron. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, 17(suppl 1):107–114, 2001. doi: 10.1093/bioinformatics/17.suppl_1.S107. URL `http://bioinformatics.oxfordjournals.org/content/17/suppl_1/S107.abstract`.

[43] J.H. Friedman und J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23(9):881–890, 1974. ISSN 0018-9340. doi: http://doi.ieeecomputersociety.org/10.1109/T-C.1974.224051.

[44] P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 06 1985. doi: 10.1214/aos/1176349519. URL http://dx.doi.org/10.1214/aos/1176349519.

[45] C. Soneson und M. Fontes. A method for visual identification of small sample subgroups and potential biomarkers. *The Annals of Applied Statistics*, 5 (3):2131–2149, 2011. doi: 10.1214/11-AOAS460.

[46] J.A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972. ISSN 01621459. URL http://www.jstor.org/stable/2284710.

[47] B. Pontes, R. Giráldez und J.S. Aguilar-Ruiz. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180, 2015. ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2015.06.028. URL http://www.sciencedirect.com/science/article/pii/S1532046415001380.

[48] R. Henriques und S.C. Madeira. Biclustering with flexible plaid models to unravel interactions between biological processes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4):738–752, July 2015. ISSN 1545-5963. doi: 10.1109/TCBB.2014.2388206.

[49] A. Oghabian, S. Kilpinen, S. Hautaniemi und E. Czeizler. Biclustering methods: Biological relevance and application in gene expression analysis. *PLoS ONE*, 9(3):1–10, 03 2014. doi: 10.1371/journal.pone.0090801. URL http://dx.doi.org/10.1371%2Fjournal.pone.0090801.

[50] B. Hanczar und M. Nadif. Ensemble methods for biclustering tasks. *Pattern Recognition*, 45(11):3938–3949, 2012. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2012.04.010. URL http://www.sciencedirect.com/science/article/pii/S0031320312001677.

[51] R. De Smet und K. Marchal. An ensemble biclustering approach for querying gene expression compendia with experimental lists. *Bioinformatics*, 27(14):1948–1956, 2011. doi: 10.1093/bioinformatics/btr307. URL http://bioinformatics.oxfordjournals.org/content/27/14/1948.abstract.

[52] T. Khamiakova. *superbiclust: Generating Robust Biclusters from a Bicluster Set (Ensemble Biclustering)*, 2014. URL http://CRAN.R-project.org/package=superbiclust. R package version 1.1.

[53] S. Kaiser, R. Santamaria, T. Khamiakova, M. Sill, R. Theron, L. Quintales, F. Leisch und E. De Troyer. *biclust: BiCluster Algorithms*, 2015. URL http://CRAN.R-project.org/package=biclust. R package version 1.2.0.

[54] H. Turner, T. Bailey und W. Krzanowski. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data Analysis*, 48(2):235–254, 2005. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2004.02.003. URL http://www.sciencedirect.com/science/article/pii/S0167947304000295.

[55] M. Ahrens, M. Turewicz, K. Marcus, H.E. Meyer, C. May, M. Eisenacher und J. Rahnenführer. FSOL - a workflow for the detection of patient subgroups and affected molecular features in high-throughput omics data. *PeerJ PrePrints*, 3:e1604, 2015. doi: 10.7287/peerj.preprints.1305v1. URL https://peerj.com/preprints/1305v1/.

[56] X. Yang, S. Bentink, S. Scheid und R. Spang. Similarities of ordered gene lists. *Journal of Bioinformatics and Computational Biology*, 4(3):693–708, 2006. doi: 10.1142/S0219720006002120.

[57] X. Yang, S. Scheid und C. Lottaz. *OrderedList: Similarities of Ordered Gene Lists*, 2008. URL http://compdiag.molgen.mpg.de/software/index.shtml. R package version 1.38.0.

[58] C. Lottaz, X. Yang, S. Scheid und R. Spang. OrderedList – a Bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics*, 22(18):2315–2316, 2006. doi: 10.1093/bioinformatics/btl385. URL http://bioinformatics.oxfordjournals.org/content/22/18/2315.abstract.

[59] L. Abatangelo, R. Maglietta, A. Distaso, A. D'Addabbo, T.M. Creanza, S. Mukherjee und N. Ancona. Comparative study of gene set enrichment methods. *BMC Bioinformatics*, 10(1):1–12, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-275. URL http://dx.doi.org/10.1186/1471-2105-10-275.

[60] G. Csardi und T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.org.

[61] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel und F. Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-2.

[62] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez und M. Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77, 2011.

[63] K. Eren, M. Deveci, O. Küçüktunç und Ü.V. Çatalyürek. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 14(3):279–292, 2013. doi: 10.1093/bib/bbs032. URL `http://bib.oxfordjournals.org/content/14/3/279.abstract`.

[64] S.J.G. Lewis, T. Foltynie, A.D. Blackwell, T.W. Robbins, A.M. Owen und R.A. Barker. Heterogeneity of Parkinson's Disease in the early clinical stages using a data driven approach. *Journal Of Neurology Neurosurgery and Psychiatry*, 76(3):343–348, 2005.

[65] R. Erro, C. Vitale, M. Amboni, M. Picillo, M. Moccia, K. Longo, G. Santangelo, A. De Rosa, R. Allocca, F. Giordano, G. Orefice, G. De Michele, L. Santoro, M.T. Pellecchia und P. Barone. The heterogeneity of early Parkinson's Disease: A cluster analysis on newly diagnosed untreated patients. *PLoS ONE*, 8(8):e70244, 08 2013. doi: 10.1371/journal.pone.0070244. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0070244`.

[66] M. Basile, R. Lin, N. Kabbani, K. Karpa, M. Kilimann, I. Simpson und M. Kester. Paralemmin interacts with D3 dopamine receptors: Implications for membrane localization and cAMP signaling. *Archives of Biochemistry and Biophysics*, 446(1):60–68, 2006. ISSN 0003-9861. doi: 10.1016/j.abb.2005.10.027. URL `http://www.sciencedirect.com/science/article/pii/S0003986105004431`.

[67] L.M.L. de Lau, P.J. Koudstaal, J.B.J. van Meurs, A.G. Uitterlinden, A. Hofman und M.M.B. Breteler. Methylenetetrahydrofolate reductase C677T genotype and PD. *Annals of Neurology*, 57:927–930, 2005.

[68] A. Mouatt-Prigent, Y. Agid und E.C. Hirsch. Does the calcium binding protein calretinin protect dopaminergic neurons against degeneration in Parkinson's disease? *Brain Research*, 668(1–2):62–70, 1994. ISSN 0006-8993. doi: 10.1016/0006-8993(94)90511-8. URL `http://www.sciencedirect.com/science/article/pii/0006899394905118`.

[69] R.J. Phillips, G.C. Walter, S.L. Wilder, E.A. Baronowsky und T.L. Powley. Alpha-synuclein-immunopositive myenteric neurons and vagal preganglionic terminals: Autonomic pathway implicated in Parkinson's disease? *Neuroscience*, 153(3):733 – 750, 2008. ISSN 0306-4522. doi: http://dx.doi.org/10.1016/j.neuroscience.2008.02.074. URL http://www.sciencedirect.com/science/article/pii/S0306452208003783.

[70] A.H.V. Schapira. Calcium dysregulation in Parkinson's disease. *Brain*, 136 (7):2015–2016, 2013. ISSN 0006-8950. doi: 10.1093/brain/awt180. URL http://brain.oxfordjournals.org/content/136/7/2015.

[71] T. Calì, D. Ottolini und M. Brini. Calcium signaling in Parkinson's disease. *Cell and Tissue Research*, 357(2):439–454, 2014. ISSN 1432-0878. doi: 10.1007/s00441-014-1866-0. URL http://dx.doi.org/10.1007/s00441-014-1866-0.

[72] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, K.S. Wang, F. Mandelli, R. Foà und J. Ritz. Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *American Association for Cancer Research*, 11(20):7209–7219, 2005. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-04-2165. URL http://clincancerres.aacrjournals.org/content/11/20/7209.

[73] R. Foà, A. Vitale, M. Vignetti, G. Meloni, A. Guarini, M.S. De Propris, L. Elia, F. Paoloni, P. Fazi, G. Cimino, F. Nobile, F. Ferrara, C. Castagnola, S. Sica, P. Leoni, E. Zuffa, C. Fozza, M. Luppi, A. Candoni, I. Iacobucci, S. Soverini, F. Mandelli, G. Martinelli und M. Baccarani. Dasatinib as first-line treatment for adult patients with philadelphia chromosome–positive acute lymphoblastic leukemia. *Blood*, 118(25):6521–6528, 2011. ISSN 0006-4971. doi: 10.1182/blood-2011-05-351403. URL http://www.bloodjournal.org/content/118/25/6521.

[74] M. Carlson. *hgu95av2.db: Affymetrix Human Genome U95 Set annotation data (chip hgu95av2)*, 2016. R package version 3.2.3.

[75] J.T. Park, I.M. Shih und T.L. Wang. Identification of Pbx1, a potential oncogene, as a Notch3 target gene in ovarian cancer. *Cancer Research*, 68(21): 8852–8860, 2008. doi: 10.1158/0008-5472.CAN-08-0517. URL http://cancerres.aacrjournals.org/content/68/21/8852.abstract.

[76] L. Magnani, E.B. Ballantyne, X. Zhang und M. Lupien. PBX1 genomic pioneer function drives ERα signaling underlying progression in breast

cancer. *PLoS Genetics*, 7(11):e1002368, 11 2011. doi: 10.1371/journal.pgen.1002368. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1002368.

[77] U.R. Kees, J. Ford, M. Watson, A. Murch, M. Ringńer, R.L. Walker und P. Meltzer. Gene expression profiles in a panel of childhood leukemia cell lines mirror critical features of the disease. *Molecular Cancer Therapeutics*, 2(7):671–677, 2003. URL http://mct.aacrjournals.org/content/2/7/671.abstract.

[78] L. Wang, N. Man, X.-J. Sun, Y. Tan, M. García-Cao, F. Liu, M. Hatlen, H. Xu, G. Huang, M. Mattlin, A. Mehta, E. Rampersaud, R. Benezra und S.D. Nimer. Regulation of AKT signaling by Id1 controls t(8;21) leukemia initiation and progression. *Blood*, 126(5):640–650, 2015. ISSN 0006-4971. doi: 10.1182/blood-2015-03-635532. URL http://www.bloodjournal.org/content/126/5/640.

[79] Z. Li, W. Zhang, M. Wu, S. Zhu, C. Gao, L. Sun, R. Zhang, N. Qiao, H. Xue, Y. Hu, S. Bao, H. Zheng und J.J. Han. Gene expressionâ€“based classification and regulatory networks of pediatric acute lymphoblastic leukemia. *Blood*, 114(20):4486–4493, 2009. URL https://doi.org/10.1182/blood-2009-04-218123.

[80] B. Mollenhauer, E. Trautmann, F. Sixel-Döring, T. Wicke, J. Ebentheuer, M. Schaumburg, E. Lang, N.K. Focke, K.R. Kumar, K. Lohmann, C. Klein, M.G. Schlossmacher, R. Kohnen, T. Friede und C. Trenkwalder. Nonmotor and diagnostic findings in subjects with de novo Parkinson disease of the DeNoPa cohort. *Neurology*, 81(14):1226–1234, 2013. doi: 10.1212/WNL.0b013e3182a6cbd5.

[81] Y. Levin und S. Bahn. *Quantification of Proteins by Label-Free LC-MS/MS*, pages 217–231. Humana Press, Totowa, NJ, 2010. ISBN 978-1-60761-780-8. doi: 10.1007/978-1-60761-780-8_13. URL http://dx.doi.org/10.1007/978-1-60761-780-8_13.

[82] M. Carlson. *UniProt.ws: R Interface to UniProt Web Services*, 2016. R package version 2.12.0.

[83] J.D. Storey und R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003. doi: 10.1073/pnas.1530509100. URL http://www.pnas.org/content/100/16/9440.abstract.

[84] R. Tibshirani, T. Hastie, B. Narasimhan und G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002. doi: 10.1073/pnas.082099299. URL `http://www.pnas.org/content/99/10/6567.abstract`.

[85] S. Jiang, J. Wu, Y. Yang, J. Liu, Y. Ding und M. Ding. Proteomic analysis of the cerebrospinal fluid in multiple sclerosis and neuromyelitis optica patients. *Molecular Medicine Reports*, 6:1081–1086, 2012. doi: 10.3892/mmr.2012.1025.

[86] J.L. Richens, H.L. Spencer, M. Butler, F. Cantlay, K.A. Vere, N. Bajaj, K. Morgan und P. O'Shea. Rationalising the role of keratin 9 as a biomarker for Alzheimer's disease. *Scientific Reports*, 6, 2016. doi: 10.1038/srep22962.

[87] S. Fabrizio, R. Chiara und F. Federico. Similarity measures based on the overlap of ranked genes are effective for comparison and classification of microarray data. *Journal of Computational Biology*, 23(7):603–614, 2016.

[88] M. Antosh, D. Fox, L.N. Cooper und N. Neretti. CORaL: comparison of ranked lists for analysis of gene expression data. *Journal of Computational Biology*, 20(6):433–443, 2013. doi: 10.1089/cmb.2013.001.

[89] S.B. Plaisier, R. Taschereau, J.A. Wong und T.G. Graeber. Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Research*, 38(17):e169, 2010. doi: 10.1093/nar/gkq636.

[90] J.D. Rosenblatt und J.L. Stein. *RRHO: Test overlap using the Rank-Rank Hypergeometric test*, 2014. R package version 1.12.0.

Anhang

Das im Folgenden präsentierte ergänzende Material zum Hauptteil der Arbeit behandelt Aspekte der univariaten und multivariaten Subgruppendetektionsmethoden.

Univariate Methoden

Der folgende Abschnitt enthält das ergänzende Material zu den univariaten Subgruppendetektionsmethoden. Es beschäftigt sich sowohl mit der der Simulationsstudie SimUni als auch mit dem ParkCHIP-Datensatz.

Pseudocode zur LR-Berechnung in SimUni

Im Folgenden werden die wesentlichen Schritte zur Berechnung des LR als Referenzmethode in der SimUni-Studie als Pseudocode dargestellt. Mit den im Hauptteil verwendeten Notationen ist das LR zu berechnen als

$$\begin{aligned} LR &= L_1/L_0 \\ &= \frac{\prod_G f_0 \cdot \prod_K f_1}{p_{H_{0a}} \prod_G f_0 \prod_K f_0 + (1 - p_{H_{0a}}) \prod_G f_1 \prod_K f_1} \end{aligned}$$

In den Szenarien I und III ist diese Berechnung unter Verwendung der Funktion `dnorm()` in **R** für einen konkreten Datenvektor `c(G, K)` aus den Beobachtungen der gesunden und kranken Gruppe in wenigen Schritten möglich:

```
f0 <-          dnorm( c(G, K))
f1 <- (   q  * dnorm( c(G, K), mean = delta ) +
       (1-q) * dnorm( c(G, K)) )
```

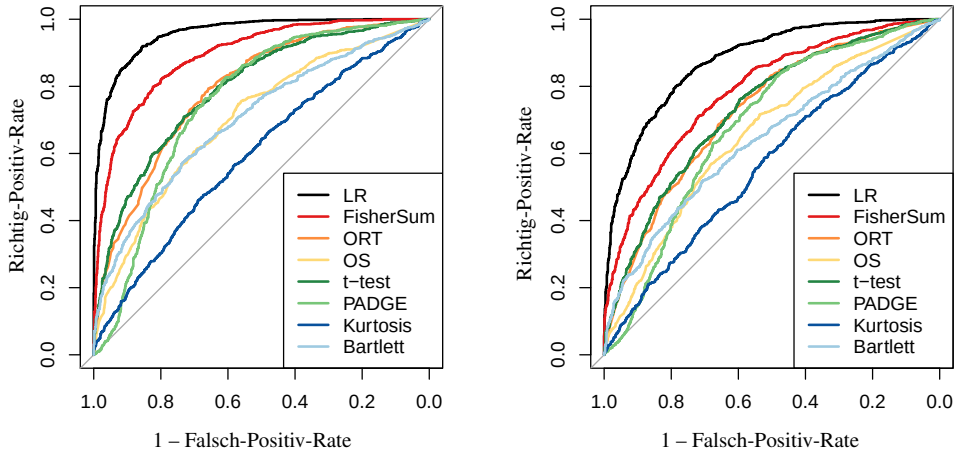
für Szenario I, und analog für Szenario III durch Einstellung des `dnorm`-Parameters `sd` als σ.
Im Gegensatz dazu wird für die in Szenario II benötigte Dichte keine entsprechende Funktion bereit gestellt. Stattdessen wird diese Dichte mittels `density()` aus einer ausreichend großen Stichprobe der Größe `iter.sim` aus der simulierten Mischverteilung (`mixture`) geschätzt:

```
# Szenario II
mixture  <- rnorm( iter.sim ) +
            c( runif( iter.sim*q, min=0, max=b),
               rep( 0, iter.sim - iter.sim*q)  )
dmix.est <- density( mixture )
```

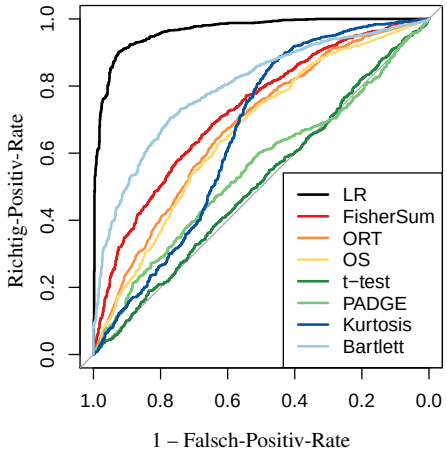
Die Werte dieser Dichtefunktion `dmix.est` werden an den entsprechenden Stellen abgelesen. In jedem Szenario folgt abschließend

```
# Likelihoodratio
LR <- ( prod( f0[ind.G] )  * prod( f1[ind.K] ))/
      ( p.H0a *  prod( f0 ) + (1-p.H0a) * prod( f1 ))
```

ROC-Kurven in SimUni



(a) Szenario I (b) Szenario II

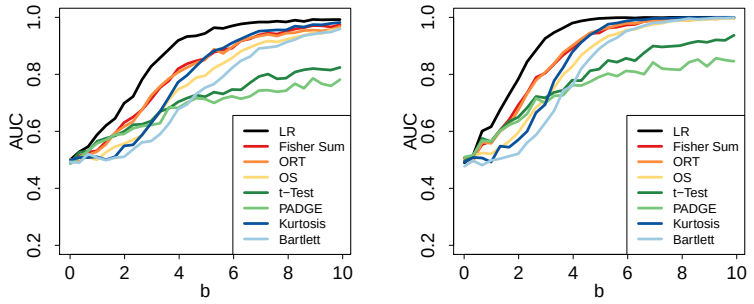


(c) Szenario III

Abbildung 26: Exemplarische Darstellung der ROC-Kurven, die für jede Kombination (s,n,q,z) aus der Verteilung s der SG, der Fallzahl n pro Gruppe, dem SG-Anteil q für den verteilungsspezifischen Unterschied z berechnet werden. Aus den ROC-Kurven wird die AUC, das Gütekriterium in der SimUni-Studie, bestimmt.

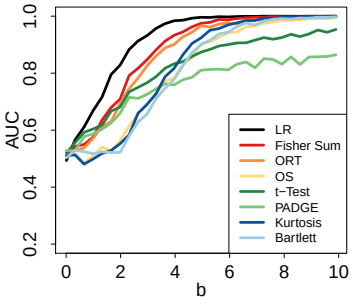
AUC-Verläufe in Abhängigkeit von z (SimUni)

In diesem Abschnitt werden ergänzende Plots zu den SimUni-Ergebnissen aus Abschnitt 5.2 gezeigt.

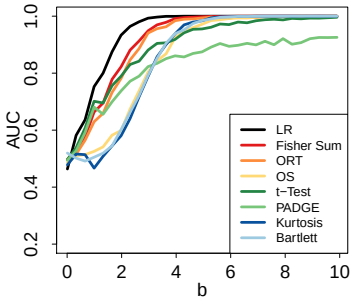


(a) $n = 30$, $q = 0.1$

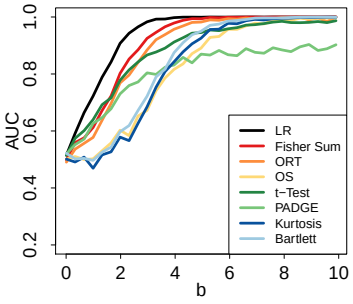
(b) $n = 70$, $q = 0.1$



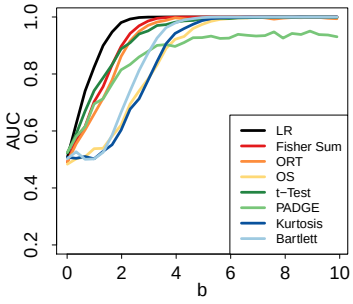
(c) $n = 30$, $q = 0.2$



(d) $n = 70$, $q = 0.2$



(e) $n = 30$, $q = 0.3$



(f) $n = 70$, $q = 0.3$

Abbildung 27: Ergebnisse für ausgewählte Parameterkonstellationen in SimUni, Szenario II, $p_{H_{0a}} = 1$.

Vergleich der Topkandidaten univariater Methoden (ParkCHIP)

Die folgenden Seiten zeigen die Expressionsprofile der jeweils 15 besten Kandidaten von *t*-Test, FS, ORT, OS, PADGE, Kurtosis und Bartletts Test auf dem ParkCHIP-Datensatz. Eine Beschreibung der bevorzugten Verteilungsmuster findet sich im Abschnitt 5.2 *Ergebnisse der SimUni-Studie*. Zur besseren Vergleichbarkeit sind alle Plots mit der gleichen Skala abgetragen. Die Überschriften der Einzelplots geben jeweils den annotierten Gennamen an.

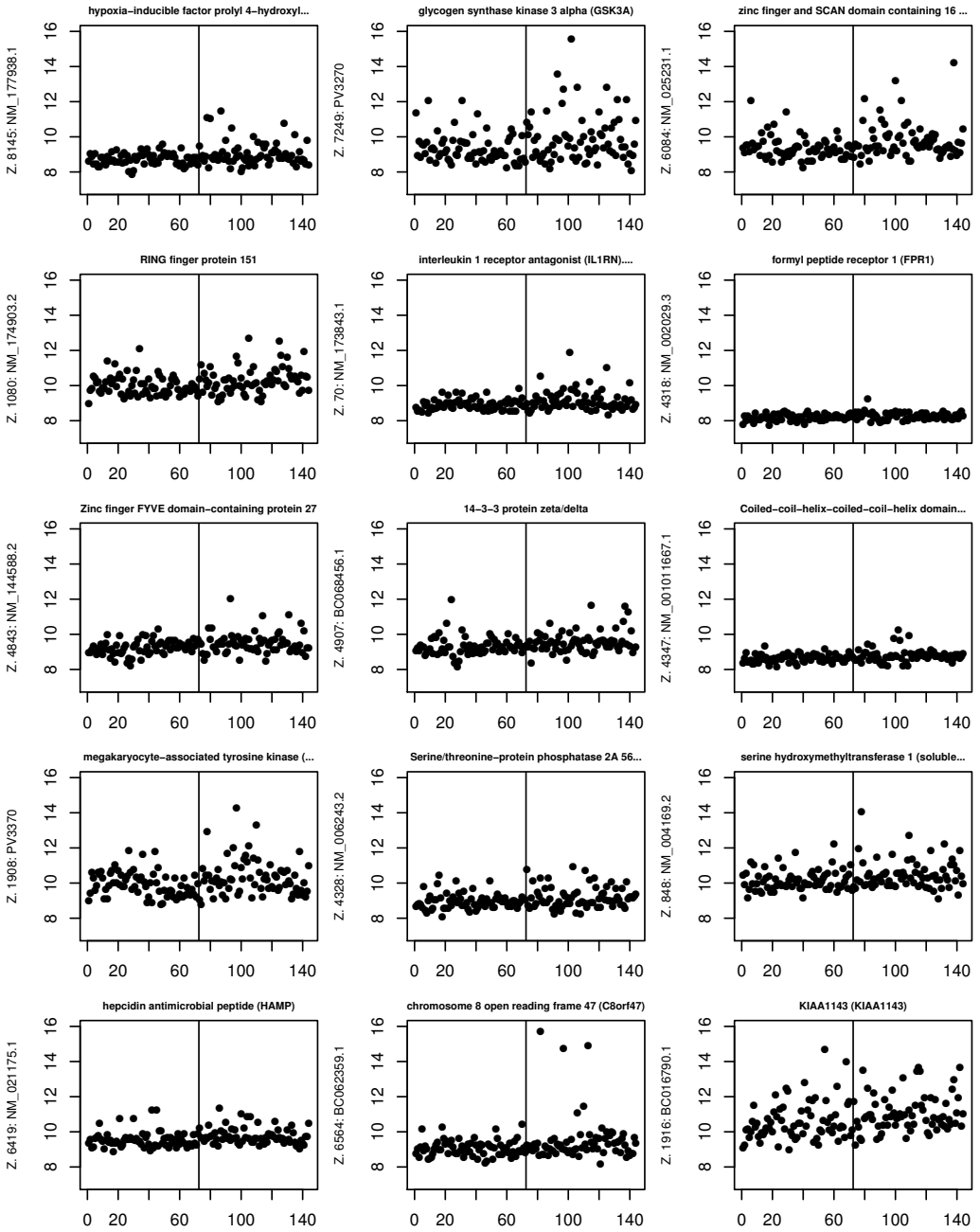


Abbildung 28: Intensitätsplot der t-Test-Topkandidaten auf dem ParkCHIP-Datensatz: Auf der Abszisse sind die Sample IDs aufgetragen, auf der Ordinate die $\log_2$-Intensitäten.

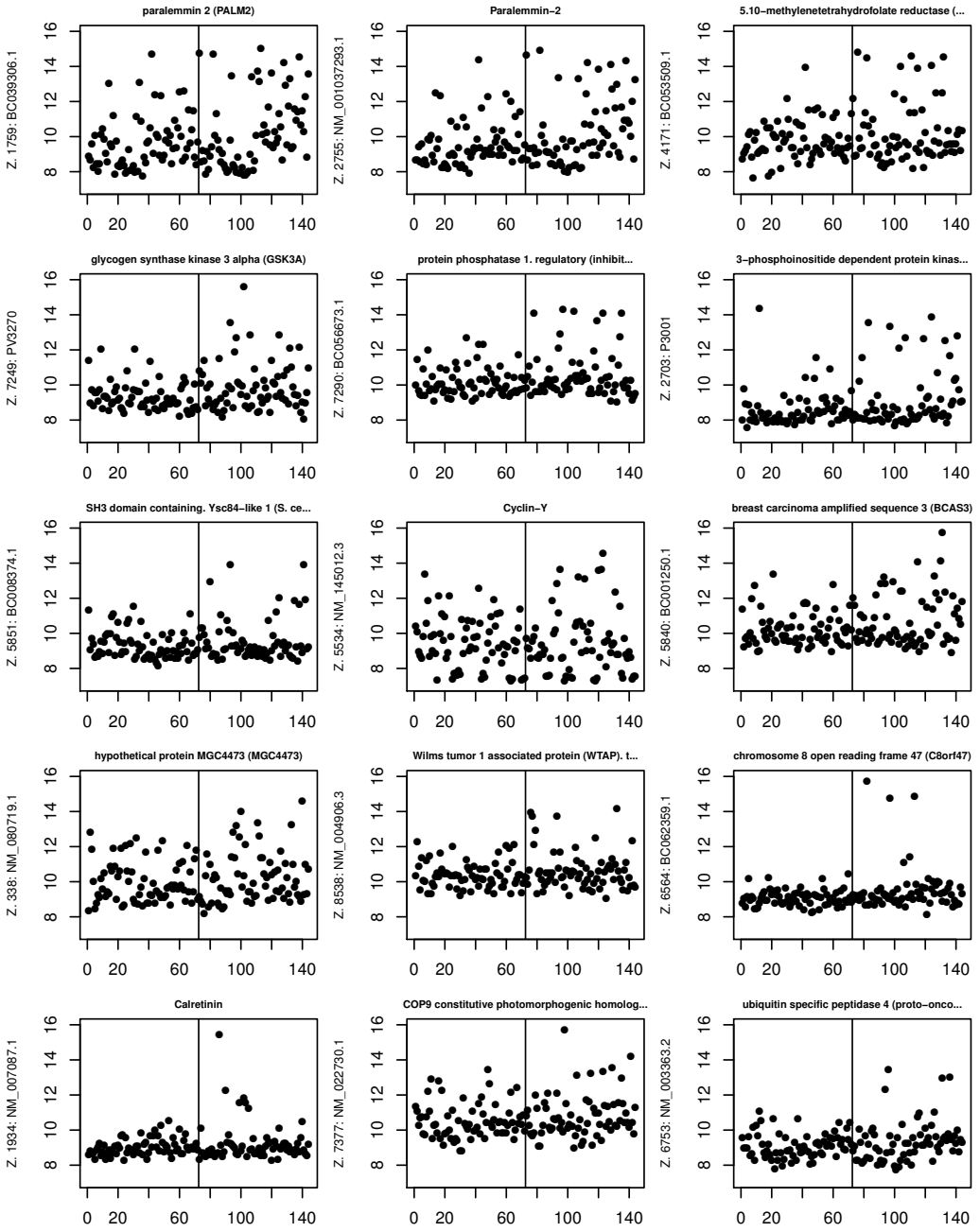


Abbildung 29: Intensitätsplot der Fisher-Sum-Topkandidaten auf dem ParkCHIP-Datensatz: Auf der Abszisse sind die Sample IDs aufgetragen, auf der Ordinate die $\log_2$-Intensitäten.

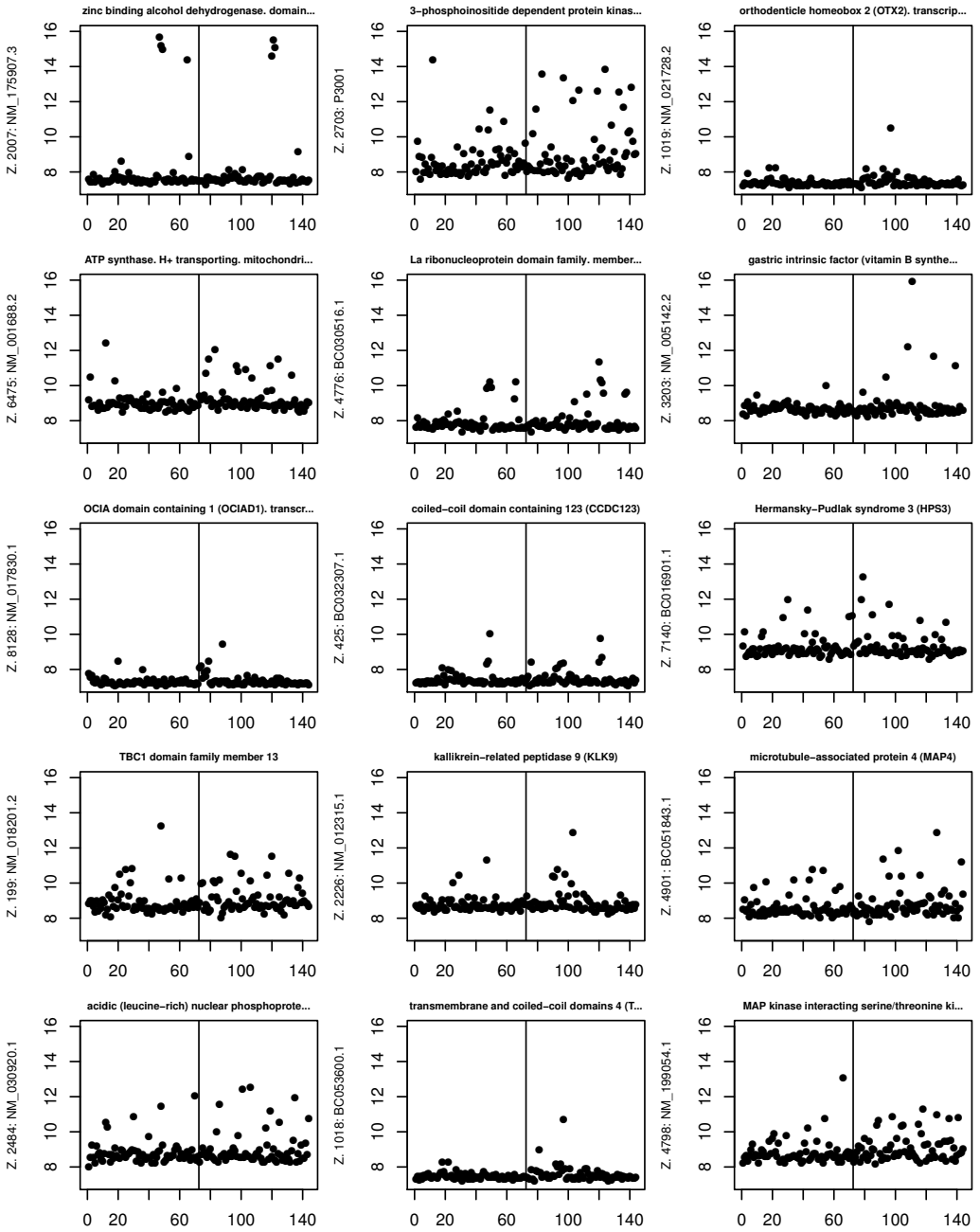


Abbildung 30: Intensitätsplot der ORT-Topkandidaten auf dem ParkCHIP-Datensatz: Auf der Abszisse sind die Sample IDs aufgetragen, auf der Ordinate die $\log_2$-Intensitäten.

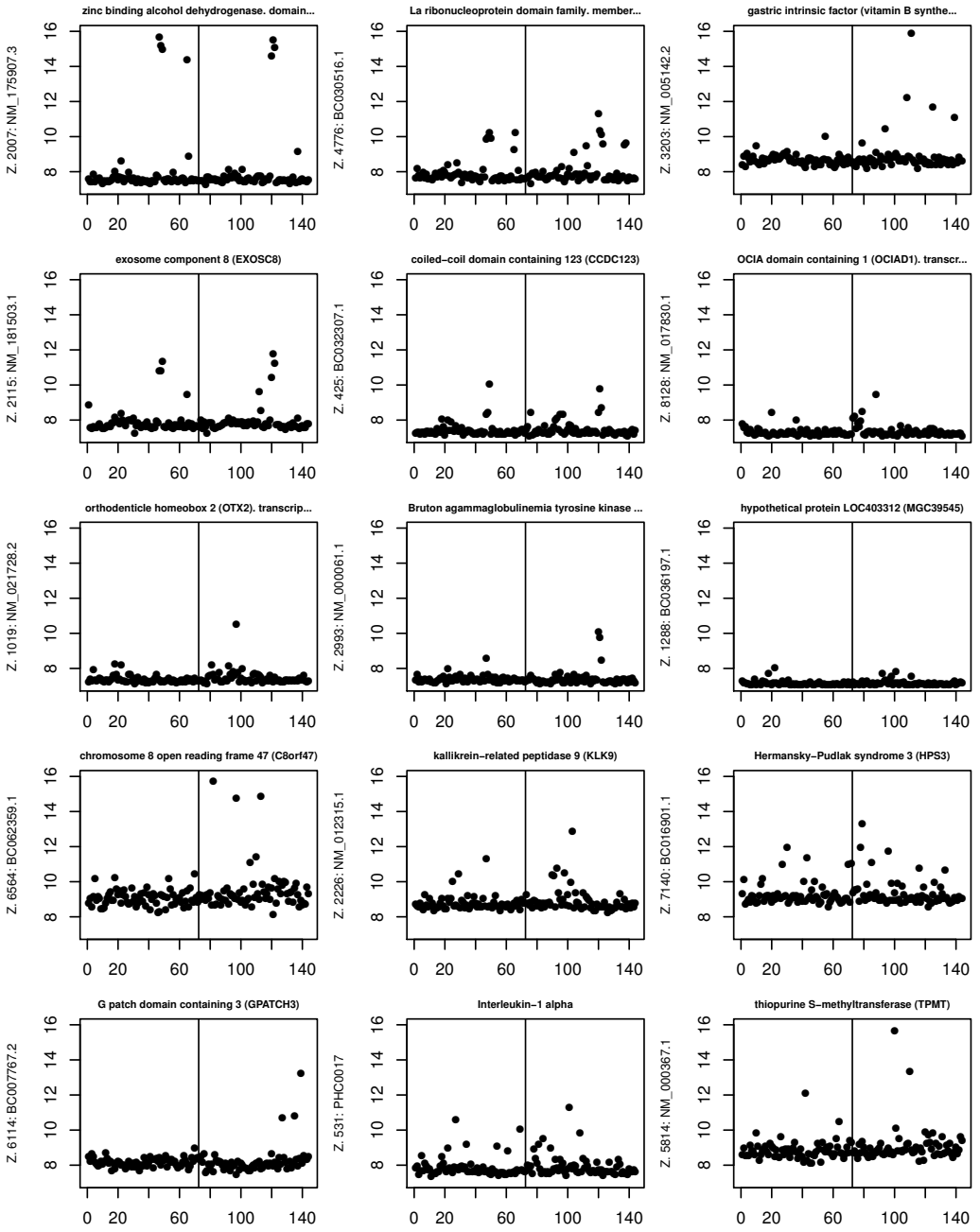


Abbildung 31: Intensitätsplot der OS-Topkandidaten auf dem ParkCHIP-Datensatz: Auf der Abszisse sind die Sample IDs aufgetragen, auf der Ordinate die $\log_2$-Intensitäten.

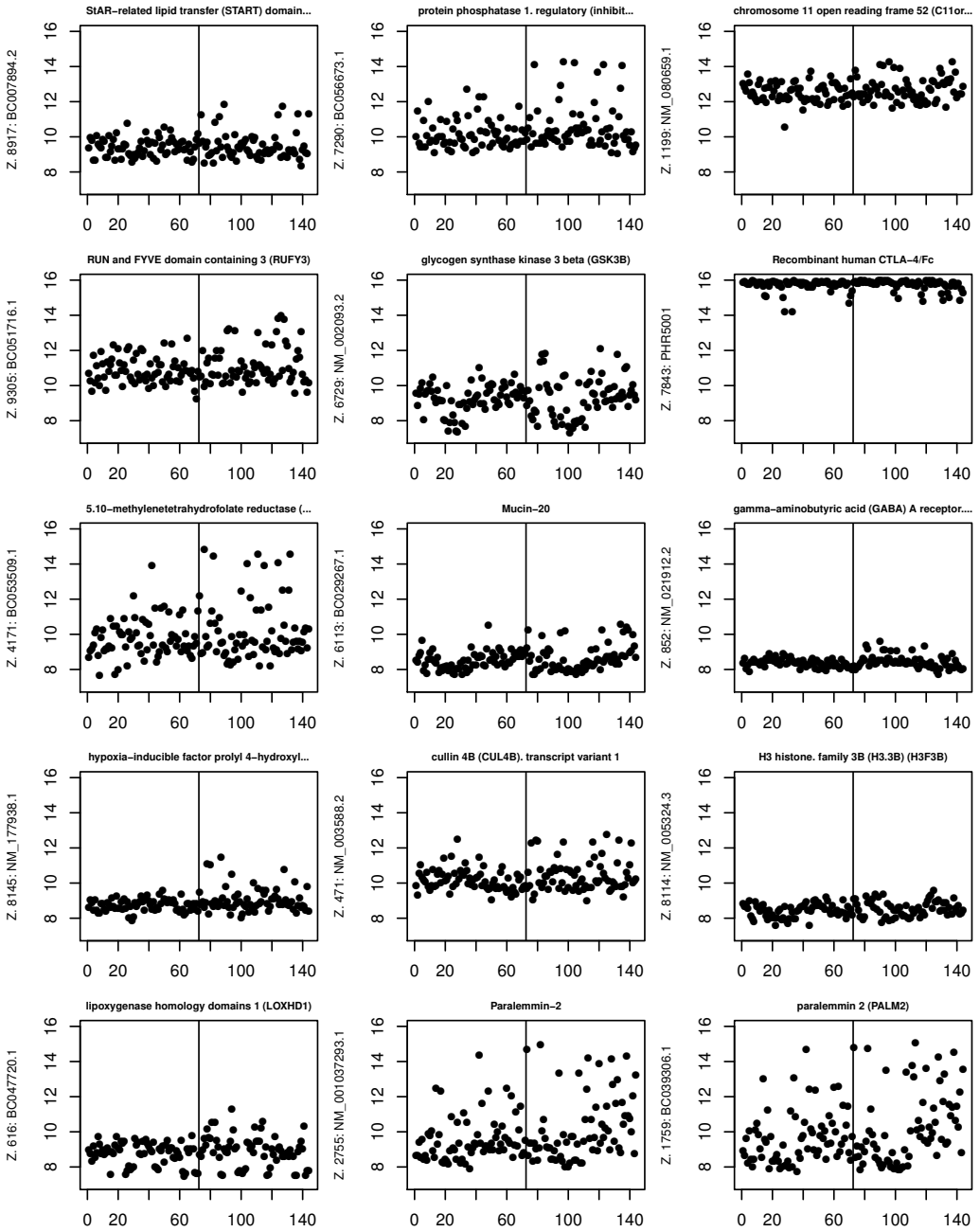


Abbildung 32: Intensitätsplot der PADGE-Topkandidaten auf dem ParkCHIP-Datensatz: Auf der Abszisse sind die Sample IDs aufgetragen, auf der Ordinate die $\log_2$-Intensitäten.

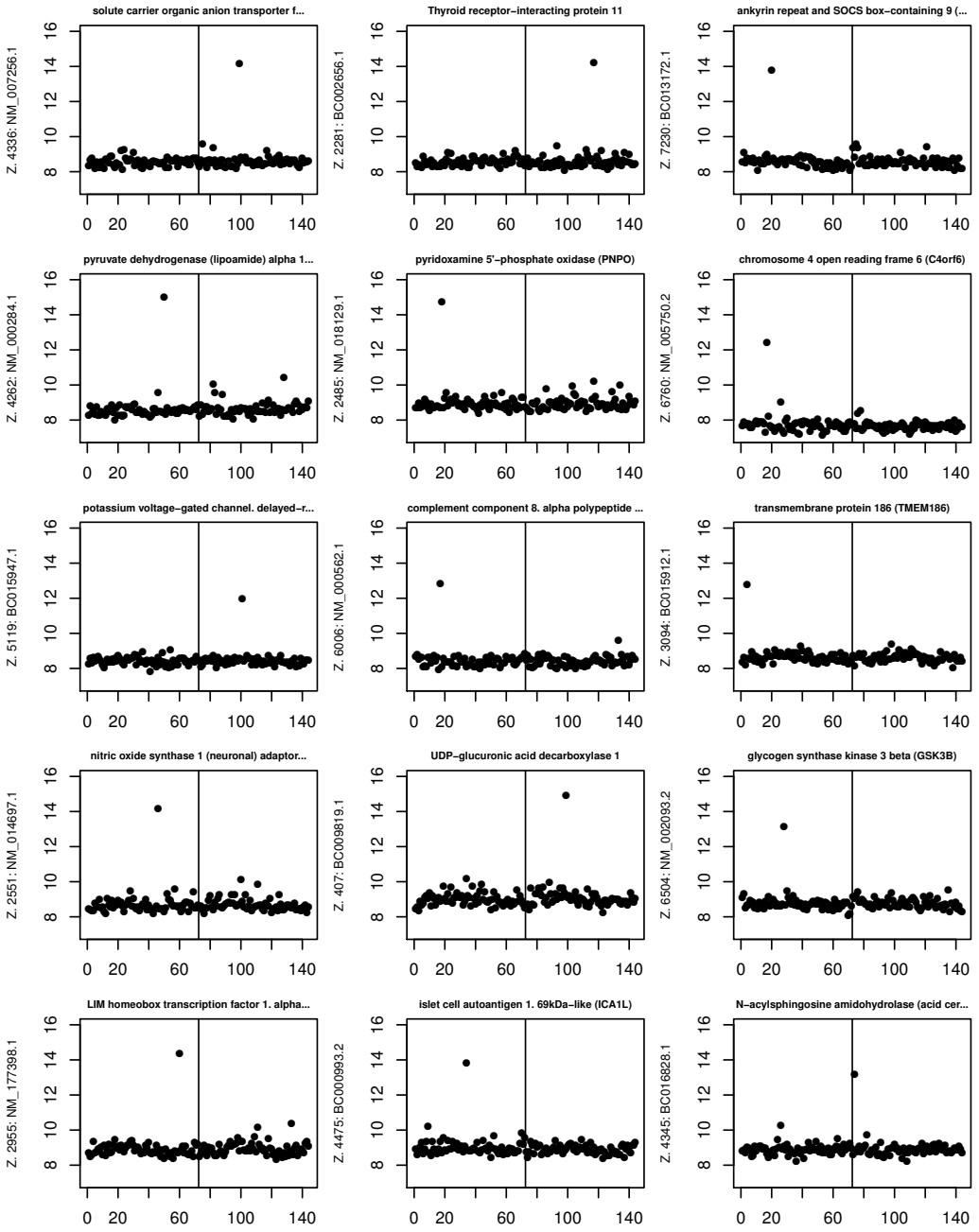


Abbildung 33: Intensitätsplot der Kurtosis-Topkandidaten auf dem ParkCHIP-Datensatz: Auf der Abszisse sind die Sample IDs aufgetragen, auf der Ordinate die $\log_2$-Intensitäten.

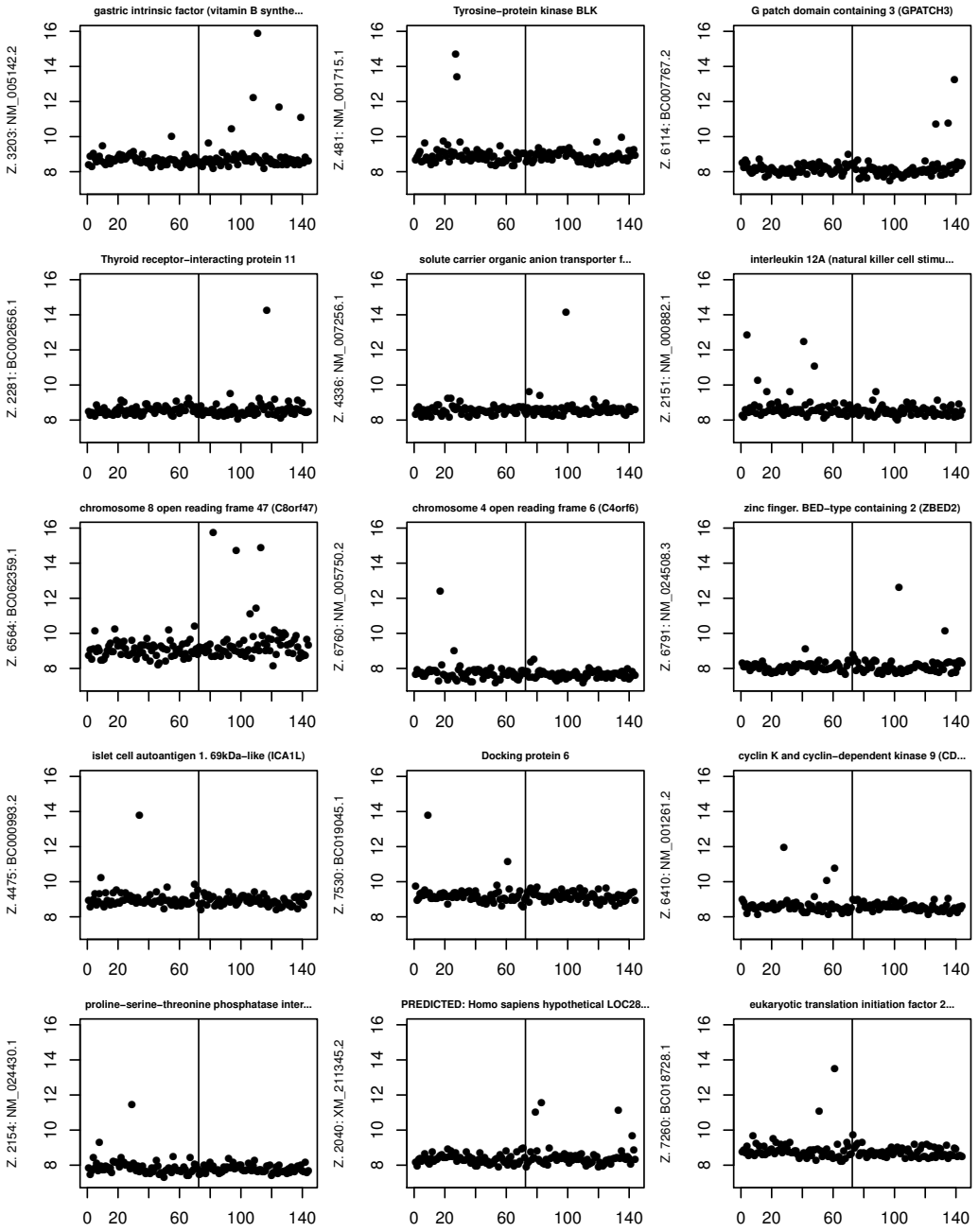


Abbildung 34: Intensitätsplot der Topkandidaten von Bartletts Test auf dem ParkCHIP-Datensatz: Auf der Abszisse sind die Sample IDs aufgetragen, auf der Ordinate die $\log_2$-Intensitäten.

Multivariate Methoden

Sensitivitätsanalysen in SimMulti: Einfluss verschiedener Parameter auf die Performanz der SG-Detektionsmethoden

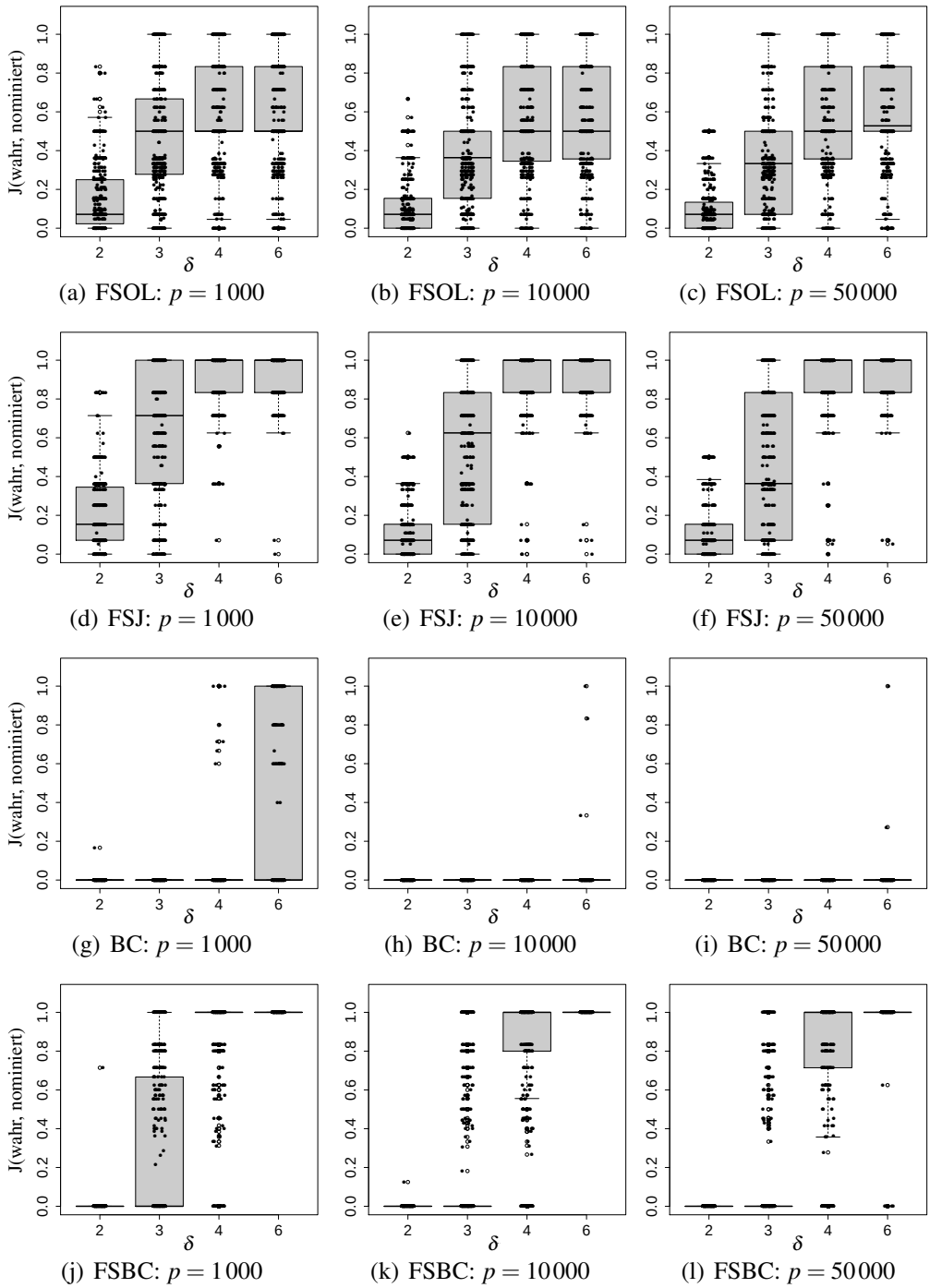
In diesem Kapitel werden zusätzliche Ergebnisse der SimMulti-Studie präsentiert. Die Variation einzelner Parameter dient der Sensitivitätsanalyse der verschiedenen Methoden gegenüber datensatzspezifischen Parametern sowie wählbaren methodenspezifischen Parametereinstellungen. Dabei wird nur der jeweils interessierende Parameter variiert, während für die übrigen die im Hauptteil der Arbeit definierten Standardeinstellungen gelten. Die verschiedenen Parameterwerte werden jeweils für die vier möglichen Kombinationen aus Fallzahl $n = 40, 70$ pro Gruppe und einer Subgruppengröße $n_{SG} = 5, 10$ betrachtet. Die folgende Aufstellung gibt einen Überblick über die durchgeführten Vergleiche:

Parameter	Beschreibung	Abbildung(en)
p	Gesamtanzahl Features	Abb. 35 – Abb. 38
p_{SG}	Anzahl SG-anzeigender Features	Abb. 39 – Abb. 42
	Ergänzung zu FSBC, $p_{SG} = 50$	Abb. 43
`heterOnly`	Sampleauswahl	Abb. 44 – Abb. 47
T	Anzahl FS-selektierter Features	Abb. 49 – Abb. 52
	Ergänzung zu FSOL, $T = 100$	Abb. 53
max.rk	Anzahl verglichener Topränge	Abb. 54 – Abb. 57

Im Anschluss an die Sensitivitätsanalysen finden sich in Abbildungen 58 und 59 (ergänzend zu den Abb. 12 und 13) die Plots zum direkten Vergleich der vier Methoden für die feste Parameterkonstellation

p	p_{SG}	`heterOnly`	T	*max.rk*
1000	5	`FALSE`	50	10

in den (n, n_{SG})-Settings $(40, 5)$ bzw. $(70, 10)$.



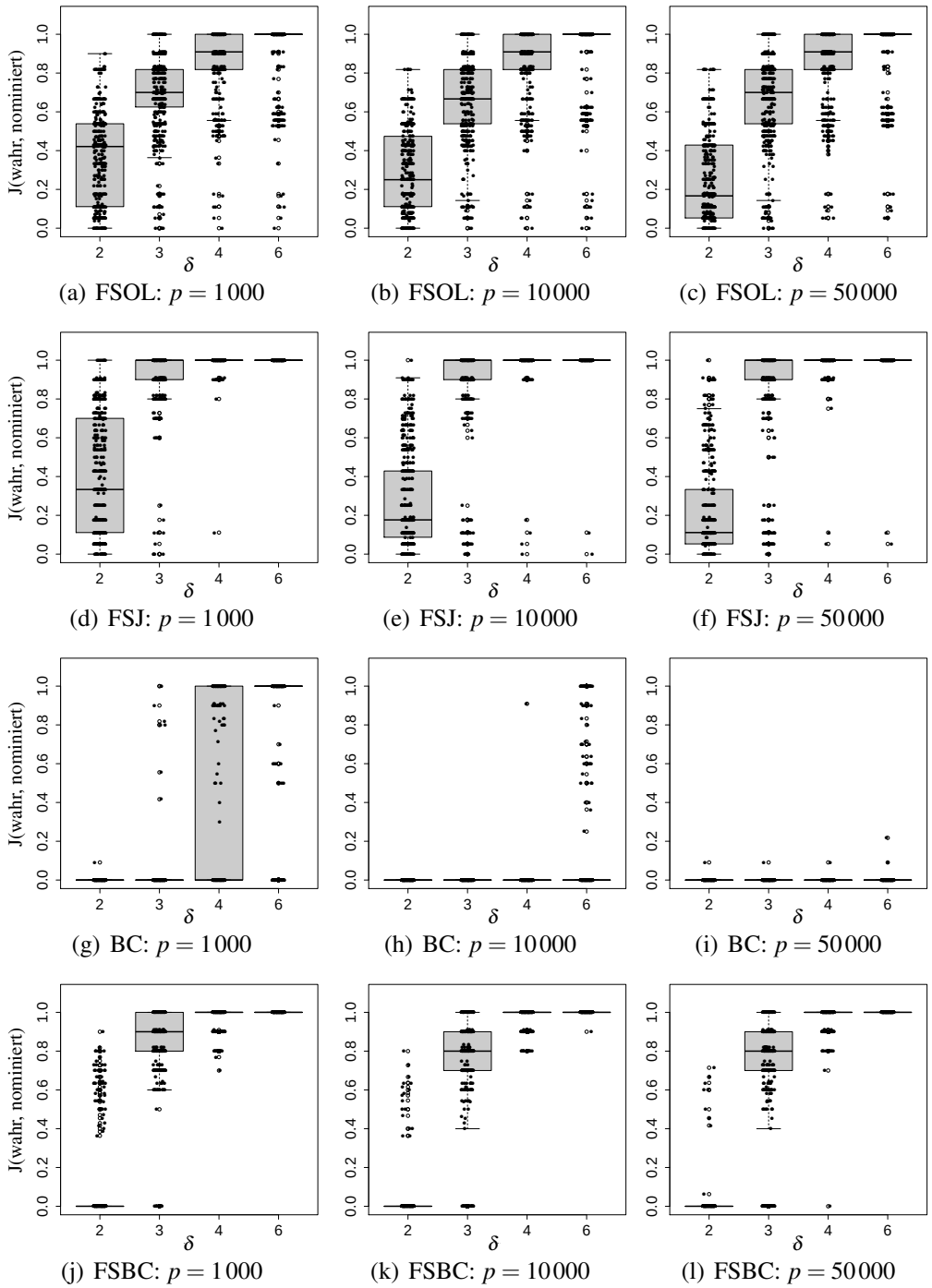
(a) FSOL: $p = 1000$ (b) FSOL: $p = 10000$ (c) FSOL: $p = 50000$

(d) FSJ: $p = 1000$ (e) FSJ: $p = 10000$ (f) FSJ: $p = 50000$

(g) BC: $p = 1000$ (h) BC: $p = 10000$ (i) BC: $p = 50000$

(j) FSBC: $p = 1000$ (k) FSBC: $p = 10000$ (l) FSBC: $p = 50000$

Abbildung 35: Einfluss der Variablenanzahl p im Originaldatensatz auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ, Biclustern (BC) und die Kombination FSBC.
Setting: $(n, n_{SG}) = (40, 5)$



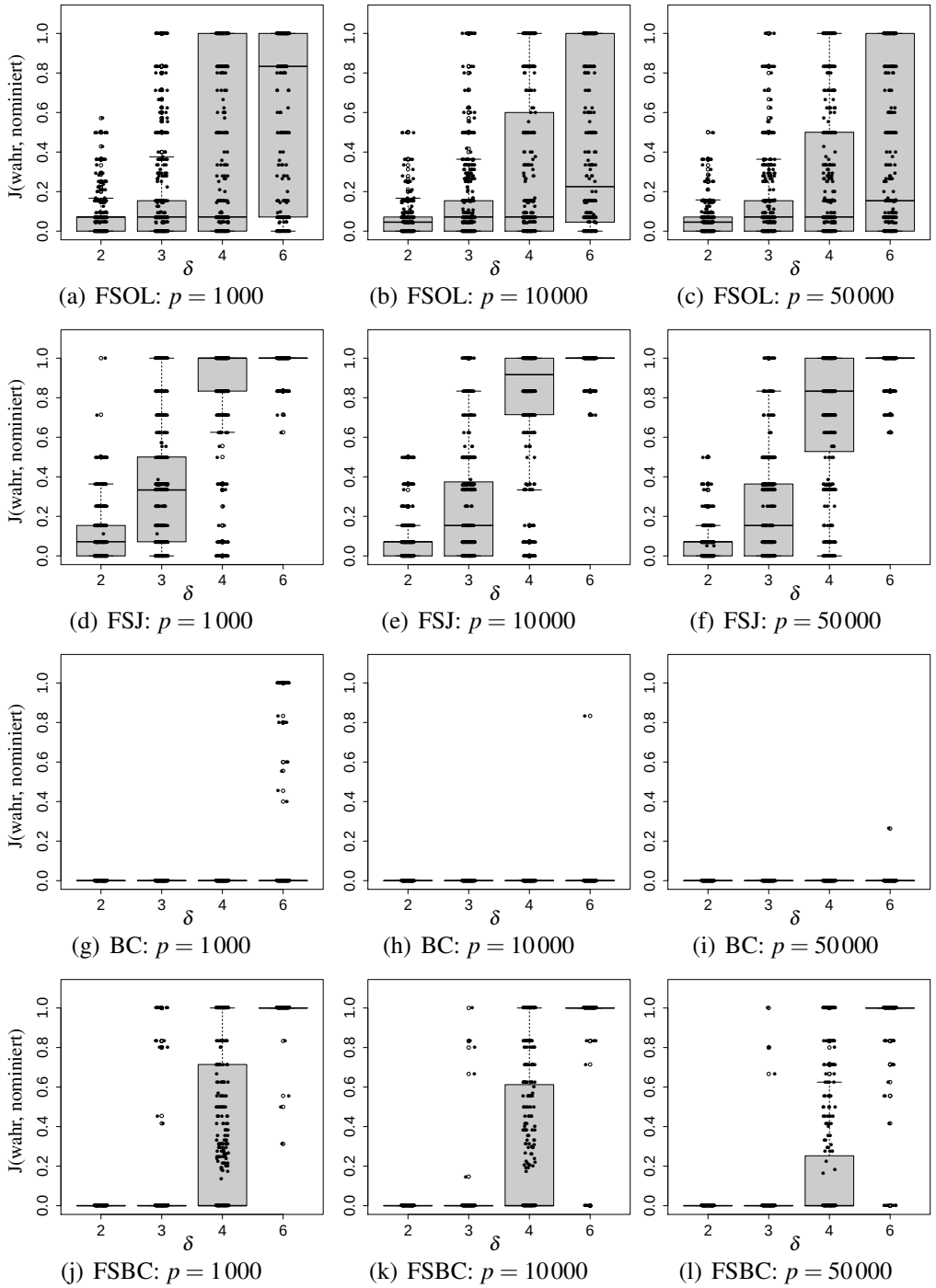
(a) FSOL: $p = 1000$ (b) FSOL: $p = 10000$ (c) FSOL: $p = 50000$

(d) FSJ: $p = 1000$ (e) FSJ: $p = 10000$ (f) FSJ: $p = 50000$

(g) BC: $p = 1000$ (h) BC: $p = 10000$ (i) BC: $p = 50000$

(j) FSBC: $p = 1000$ (k) FSBC: $p = 10000$ (l) FSBC: $p = 50000$

Abbildung 36: Einfluss der Variablenanzahl p im Originaldatensatz auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ, Biclustern (BC) und die Kombination FSBC.
Setting: $(n, n_{SG}) = (40, 10)$



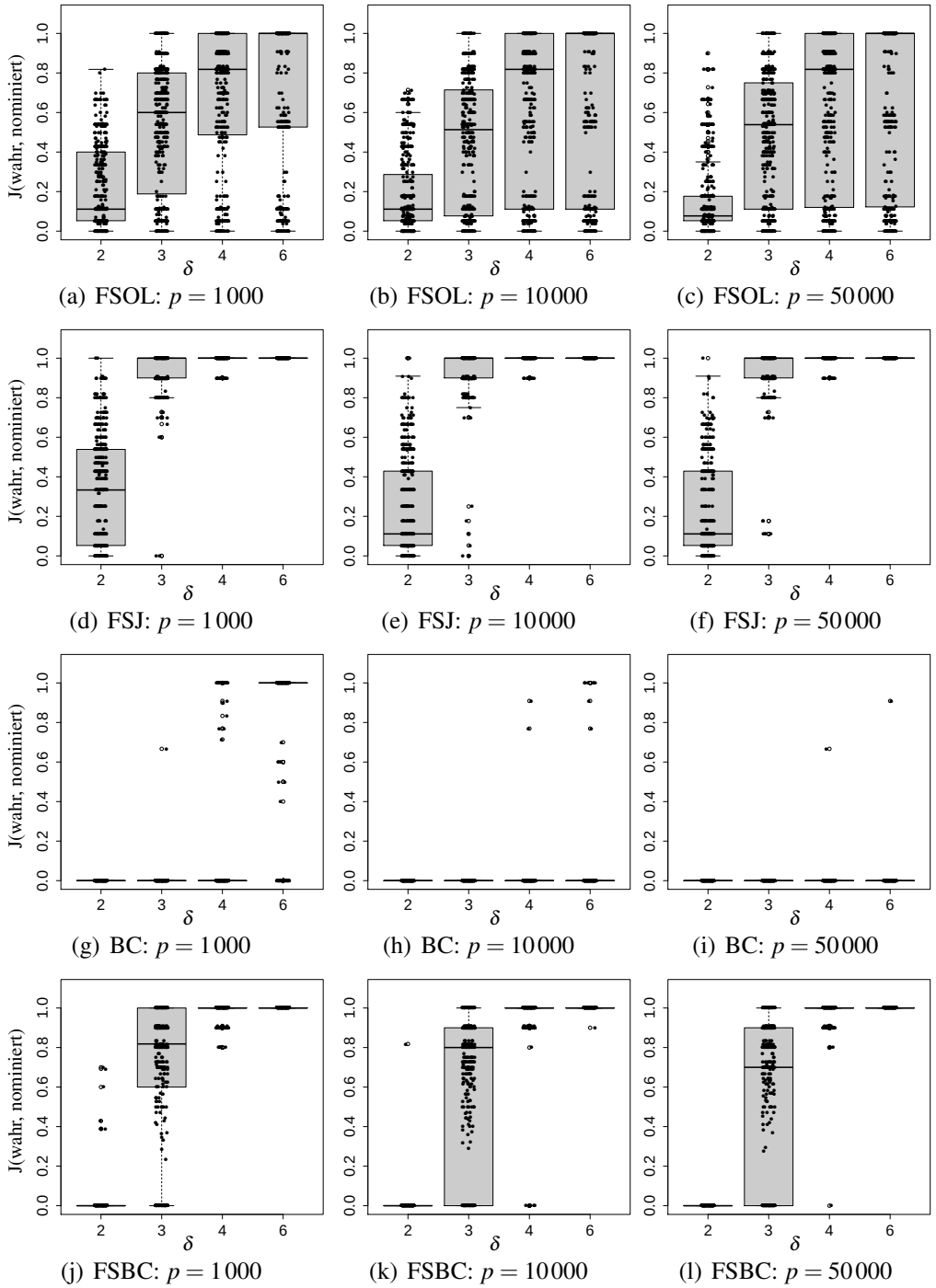
(a) FSOL: $p = 1000$ (b) FSOL: $p = 10000$ (c) FSOL: $p = 50000$

(d) FSJ: $p = 1000$ (e) FSJ: $p = 10000$ (f) FSJ: $p = 50000$

(g) BC: $p = 1000$ (h) BC: $p = 10000$ (i) BC: $p = 50000$

(j) FSBC: $p = 1000$ (k) FSBC: $p = 10000$ (l) FSBC: $p = 50000$

Abbildung 37: Einfluss der Variablenanzahl p im Originaldatensatz auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ, Biclustern (BC) und die Kombination FSBC.
Setting: $(n, n_{SG}) = (70, 5)$



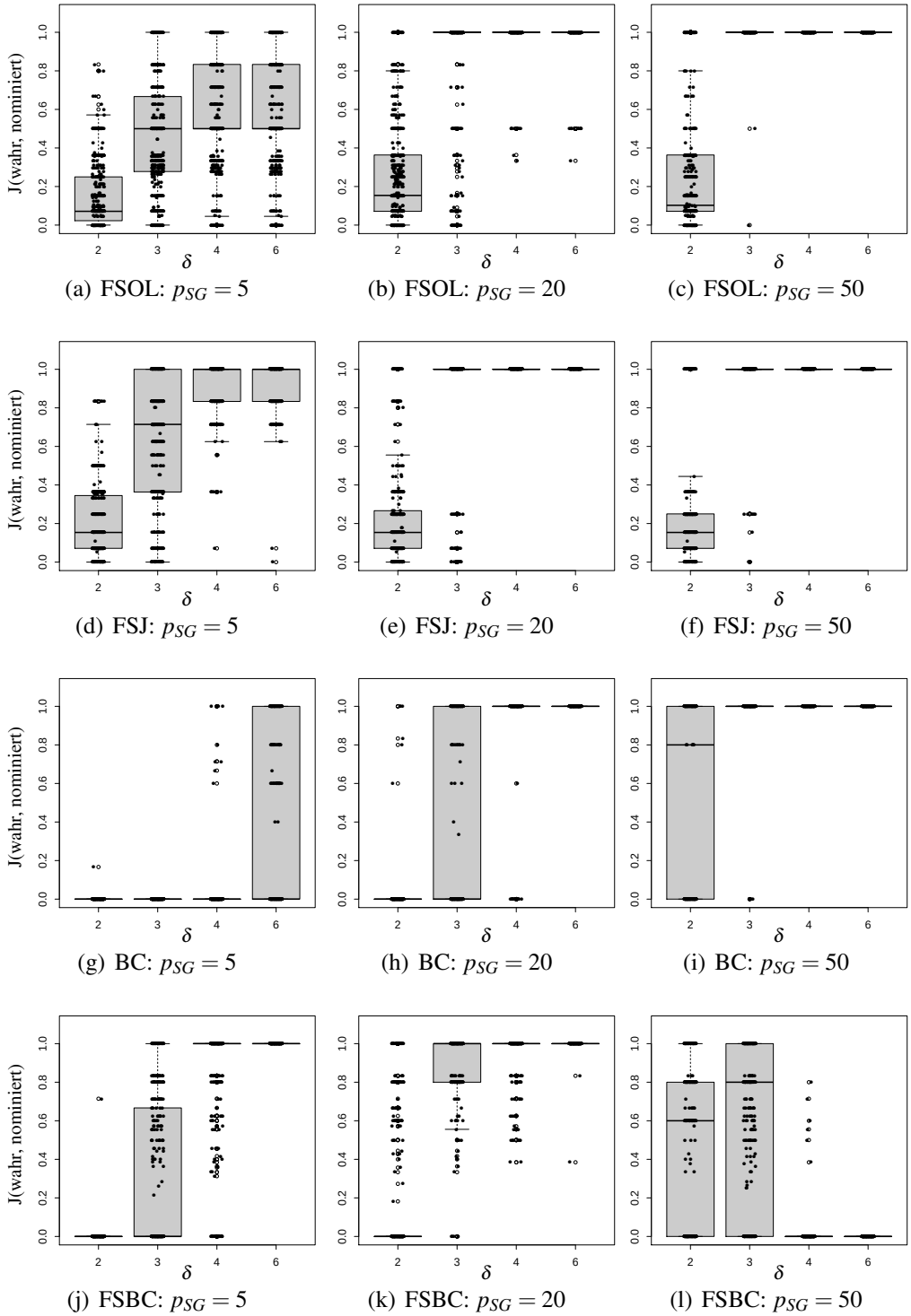
(a) FSOL: $p = 1000$ (b) FSOL: $p = 10000$ (c) FSOL: $p = 50000$

(d) FSJ: $p = 1000$ (e) FSJ: $p = 10000$ (f) FSJ: $p = 50000$

(g) BC: $p = 1000$ (h) BC: $p = 10000$ (i) BC: $p = 50000$

(j) FSBC: $p = 1000$ (k) FSBC: $p = 10000$ (l) FSBC: $p = 50000$

Abbildung 38: Einfluss der Variablenanzahl p im Originaldatensatz auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ, Biclustern (BC) und die Kombination FSBC.
Setting: $(n, n_{SG}) = (70, 10)$



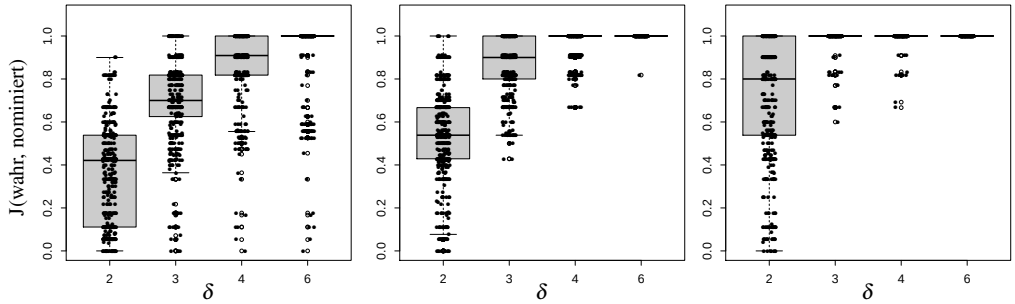
(a) FSOL: $p_{SG} = 5$ (b) FSOL: $p_{SG} = 20$ (c) FSOL: $p_{SG} = 50$

(d) FSJ: $p_{SG} = 5$ (e) FSJ: $p_{SG} = 20$ (f) FSJ: $p_{SG} = 50$

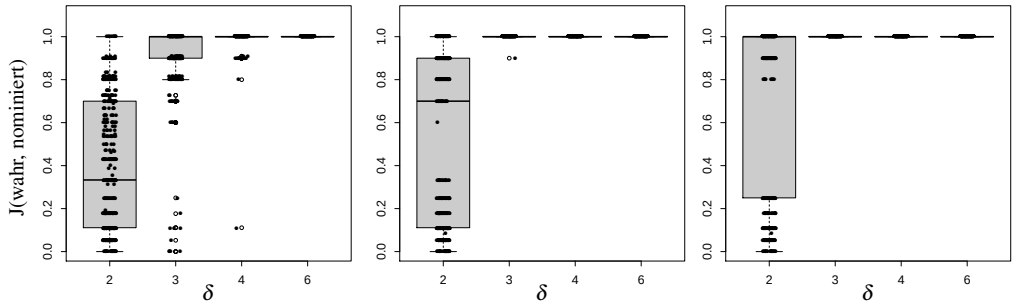
(g) BC: $p_{SG} = 5$ (h) BC: $p_{SG} = 20$ (i) BC: $p_{SG} = 50$

(j) FSBC: $p_{SG} = 5$ (k) FSBC: $p_{SG} = 20$ (l) FSBC: $p_{SG} = 50$

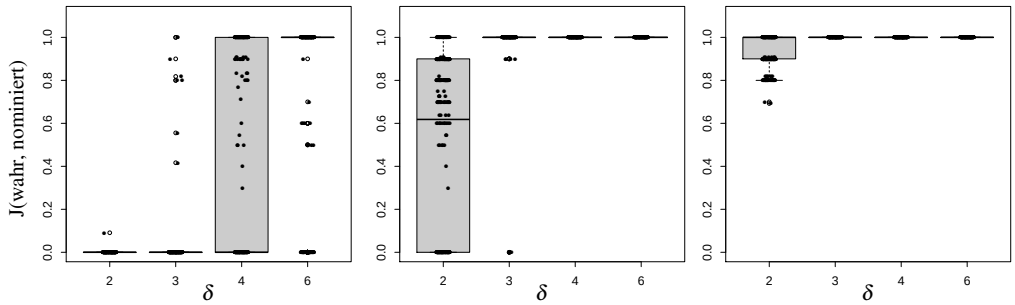
Abbildung 39: Einfluss der Anzahl p_{SG} der subgruppenanzeigenden Variablenanzahl auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ und Biclustern (BC) und FSBC.
Setting: $(n, n_{SG}) = (40, 5)$



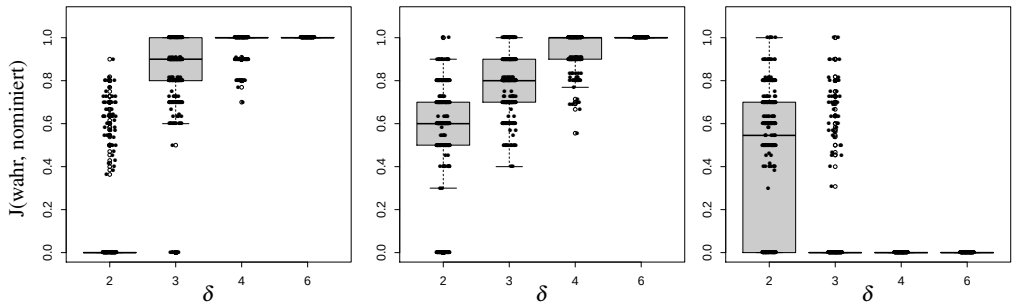
(a) FSOL: $p_{SG} = 5$ (b) FSOL: $p_{SG} = 20$ (c) FSOL: $p_{SG} = 50$



(d) FSJ: $p_{SG} = 5$ (e) FSJ: $p_{SG} = 20$ (f) FSJ: $p_{SG} = 50$

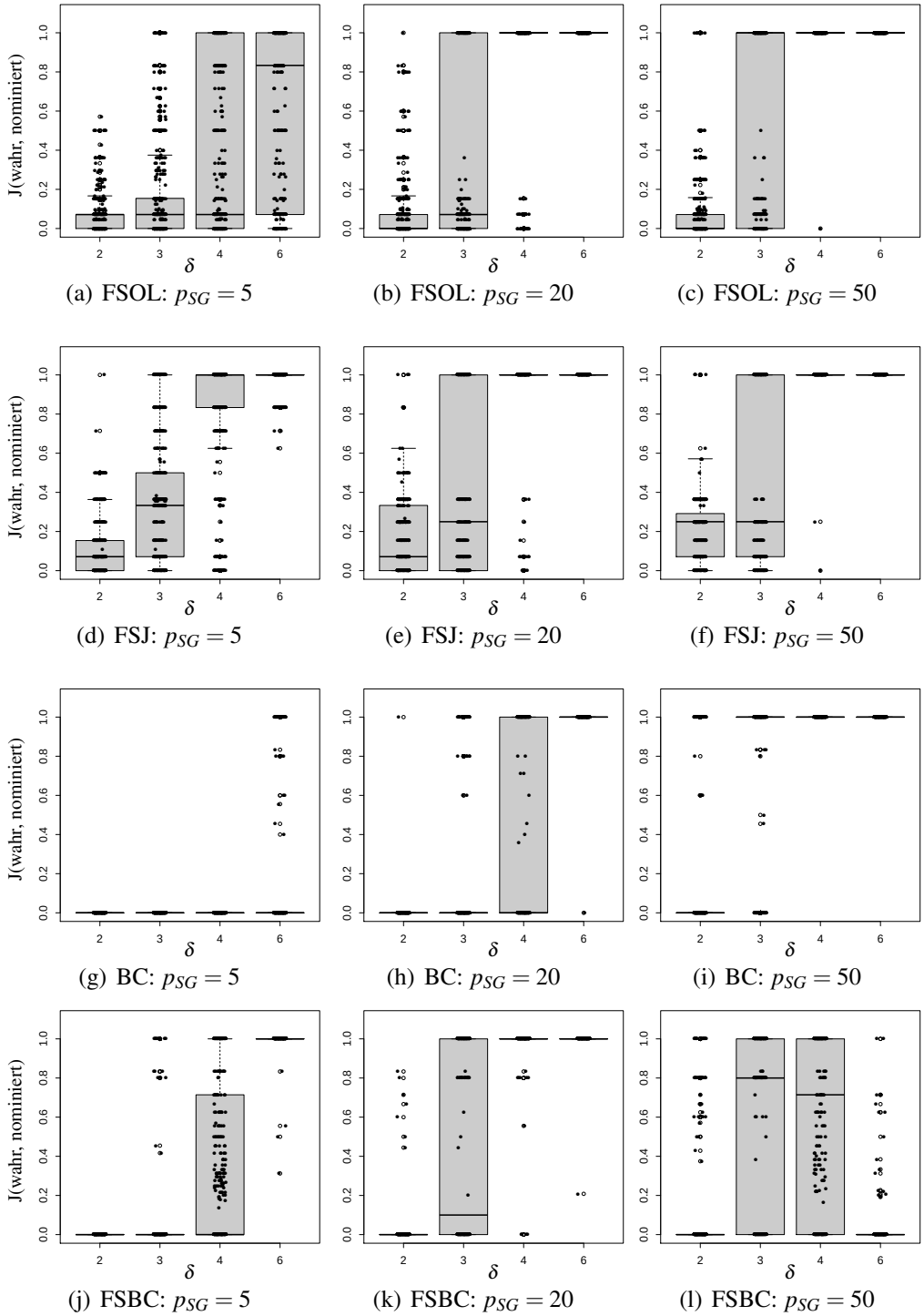


(g) BC: $p_{SG} = 5$ (h) BC: $p_{SG} = 20$ (i) BC: $p_{SG} = 50$



(j) FSBC: $p_{SG} = 5$ (k) FSBC: $p_{SG} = 20$ (l) FSBC: $p_{SG} = 50$

Abbildung 40: Einfluss der Anzahl p_{SG} der subgruppenanzeigenden Variablenanzahl auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ und Biclustern (BC) und FSBC.
Setting: $(n, n_{SG}) = (40, 10)$



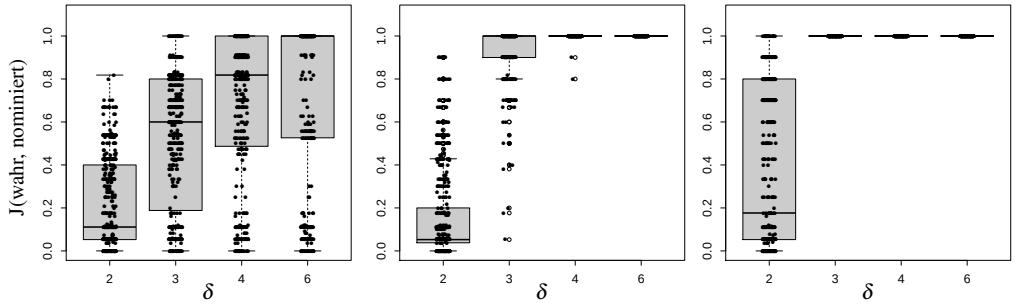
(a) FSOL: $p_{SG} = 5$ (b) FSOL: $p_{SG} = 20$ (c) FSOL: $p_{SG} = 50$

(d) FSJ: $p_{SG} = 5$ (e) FSJ: $p_{SG} = 20$ (f) FSJ: $p_{SG} = 50$

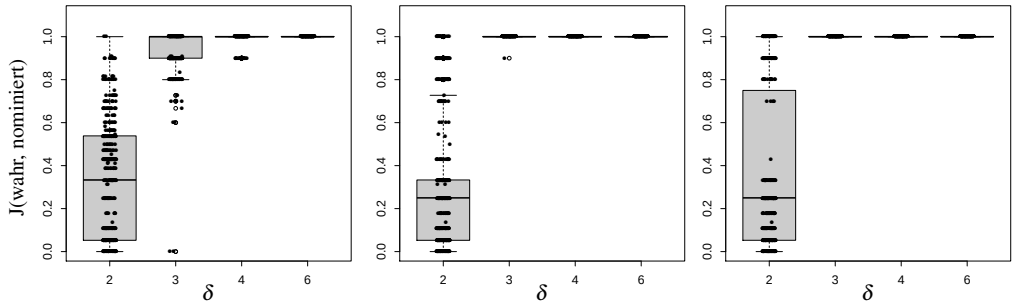
(g) BC: $p_{SG} = 5$ (h) BC: $p_{SG} = 20$ (i) BC: $p_{SG} = 50$

(j) FSBC: $p_{SG} = 5$ (k) FSBC: $p_{SG} = 20$ (l) FSBC: $p_{SG} = 50$

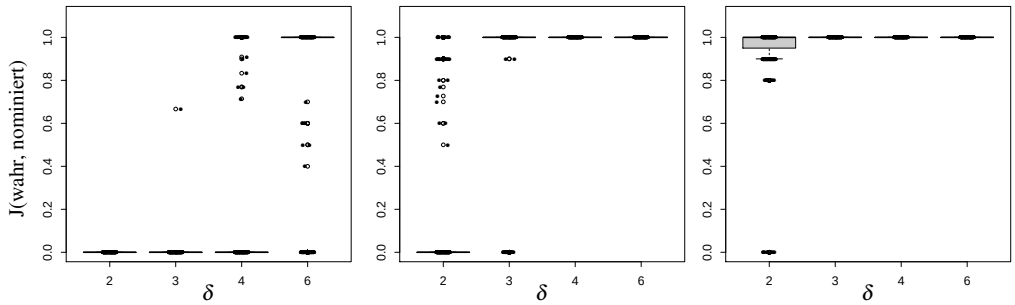
Abbildung 41: Einfluss der Anzahl p_{SG} der subgruppenanzeigenden Variablenanzahl auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ und Biclustern (BC) und FSBC.
Setting: $(n, n_{SG}) = (70, 5)$



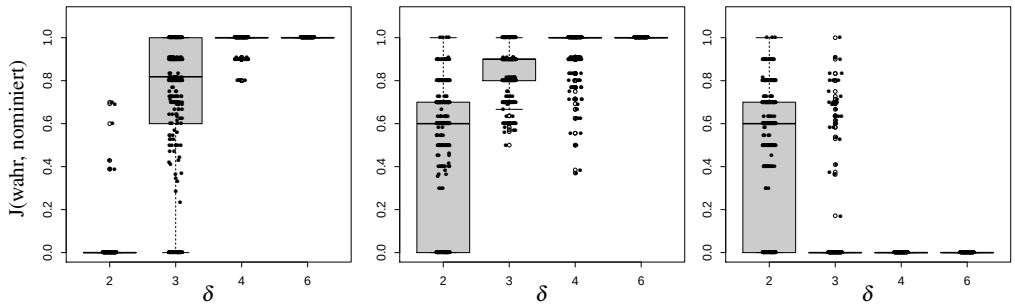
(a) FSOL: $p_{SG} = 5$ (b) FSOL: $p_{SG} = 20$ (c) FSOL: $p_{SG} = 50$



(d) FSJ: $p_{SG} = 5$ (e) FSJ: $p_{SG} = 20$ (f) FSJ: $p_{SG} = 50$

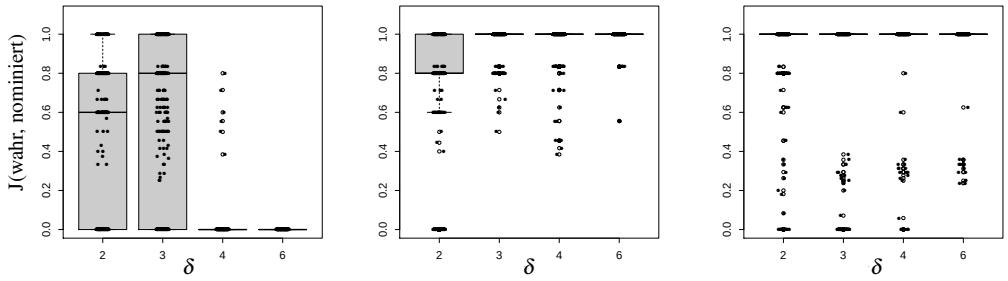


(g) BC: $p_{SG} = 5$ (h) BC: $p_{SG} = 20$ (i) BC: $p_{SG} = 50$



(j) FSBC: $p_{SG} = 5$ (k) FSBC: $p_{SG} = 20$ (l) FSBC: $p_{SG} = 50$

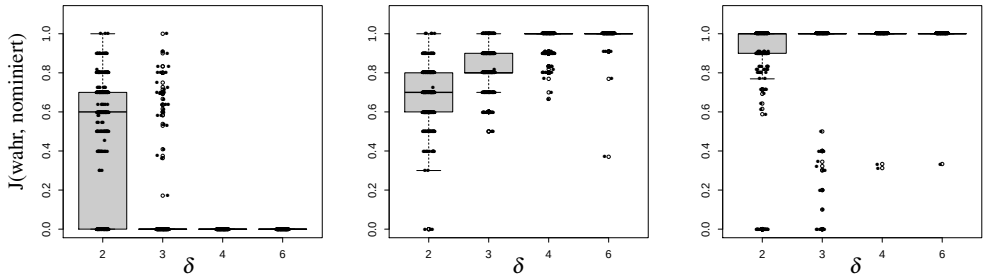
Abbildung 42: Einfluss der Anzahl p_{SG} der subgruppenanzeigenden Variablenanzahl auf die untersuchten Subgruppendetektionsmethoden FSOL, FSJ und Biclustern (BC) und FSBC.
Setting: $(n, n_{SG}) = (70, 10)$



(a) (n, n_{SG}, T)=(40,5,50), $y = m + a + b$

(b) (n, n_{SG}, T)=(40,5,100), $y = m + a + b$

(c) (n, n_{SG}, T)=(40,5,50), $y = m + a$

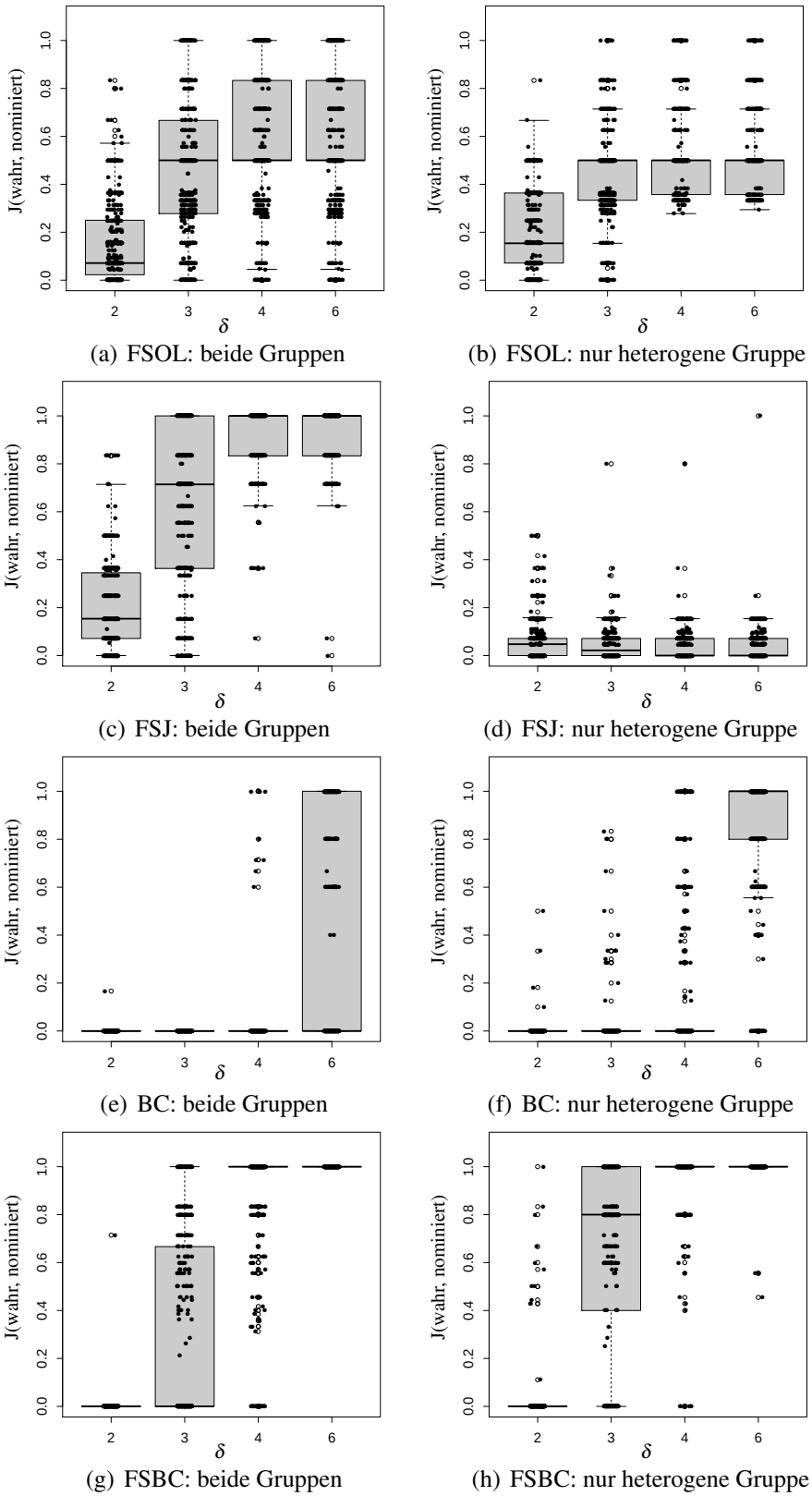


(d) (n, n_{SG}, T)=(70,10,50), $y = m + a + b$

(e) (n, n_{SG}, T)=(70,10,100), $y = m + a + b$

(f) (n, n_{SG}, T)=(70,10,50), $y = m + a$

Abbildung 43: Ergänzende Analyse zu Abb. 39-42: Dem Einbruch der Performanz von FSBC bei $p_{SG} = T = 50$ (linke Spalte) kann entgegengewirkt werden durch Erhöhung des FS-Parameters T (mittlere Spalte) und/oder Änderung des zu schätzenden Modells beim Biclustern (rechte Spalte). Hier beispielhaft gezeigt für zwei (n, n_{SG})-Settings.



(a) FSOL: beide Gruppen

(b) FSOL: nur heterogene Gruppe

(c) FSJ: beide Gruppen

(d) FSJ: nur heterogene Gruppe

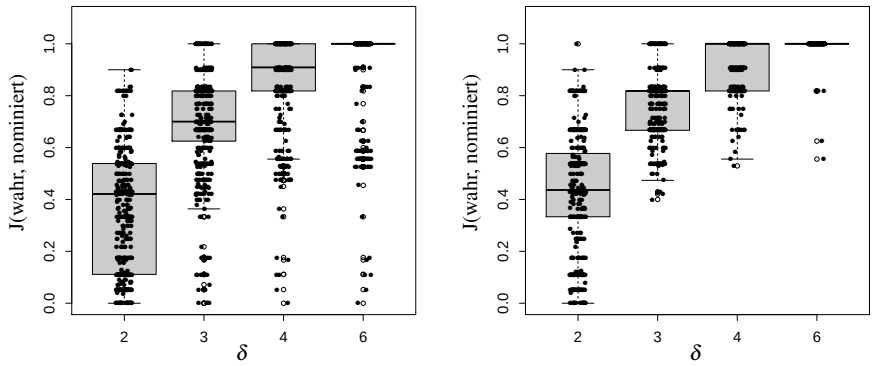
(e) BC: beide Gruppen

(f) BC: nur heterogene Gruppe

(g) FSBC: beide Gruppen

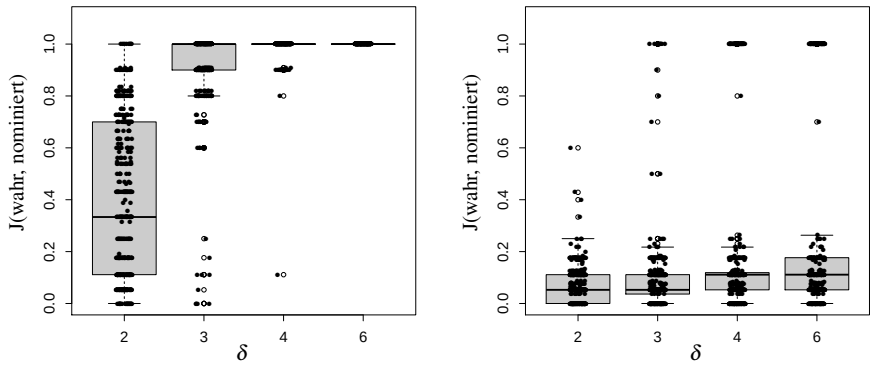
(h) FSBC: nur heterogene Gruppe

Abbildung 44: Einfluss der Sampleauswahl (Parameter `heterOnly`, nur eine oder beide Gruppen) auf FSOL, FSJ, Biclustern und FSBC für $(n, n_{SG}) = (40, 5)$.



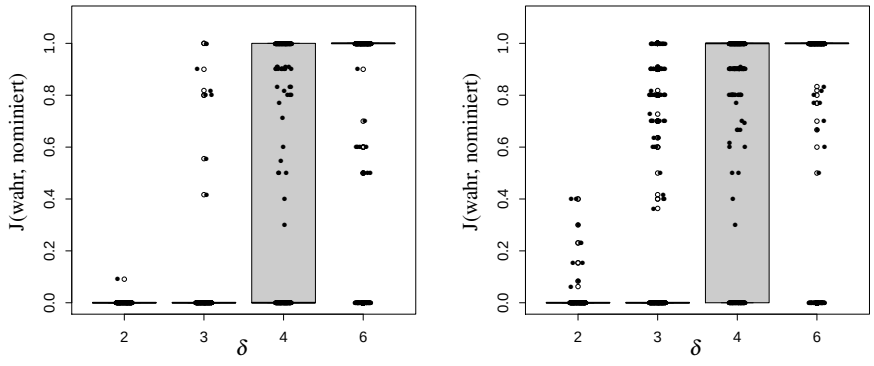
(a) FSOL: beide Gruppen

(b) FSOL: nur heterogene Gruppe



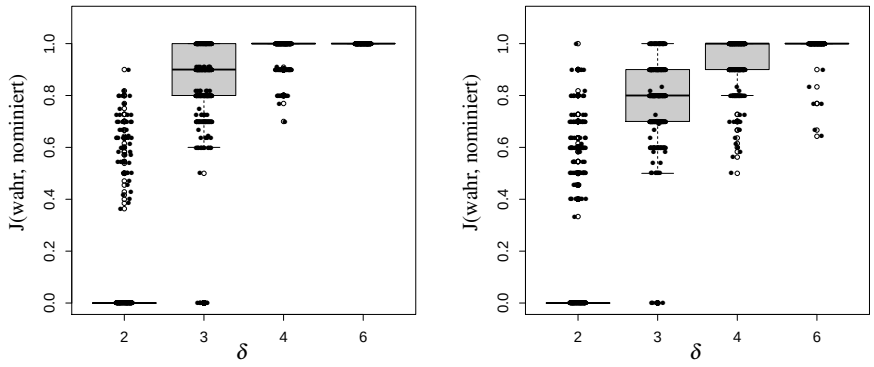
(c) FSJ: beide Gruppen

(d) FSJ: nur heterogene Gruppe



(e) BC: beide Gruppen

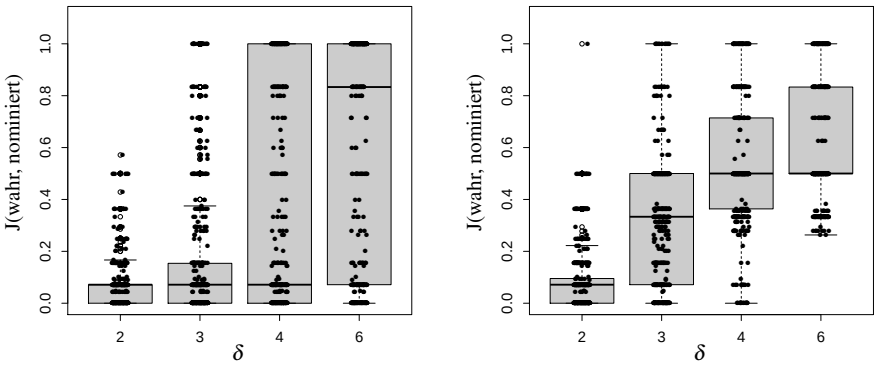
(f) BC: nur heterogene Gruppe



(g) FSBC: beide Gruppen

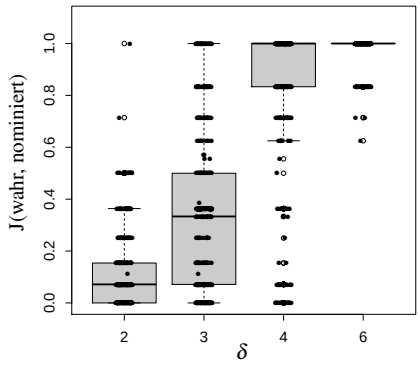
(h) FSBC: nur heterogene Gruppe

Abbildung 45: Einfluss der Sampleauswahl (Parameter `heterOnly`, nur eine oder beide Gruppen) auf FSOL, FSJ, Biclustern und FSBC für $(n, n_{SG}) = (40, 10)$.

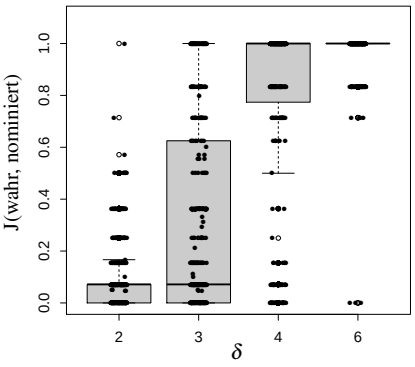


(a) FSOL: beide Gruppen

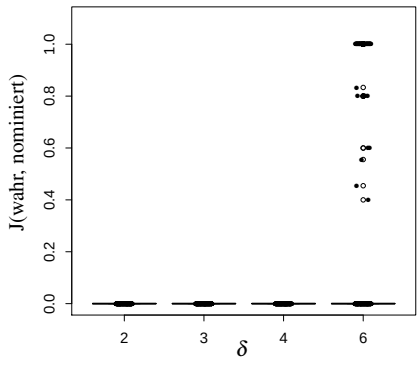
(b) FSOL: nur heterogene Gruppe



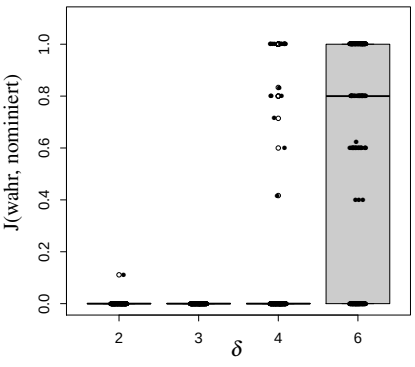
(c) FSJ: beide Gruppen



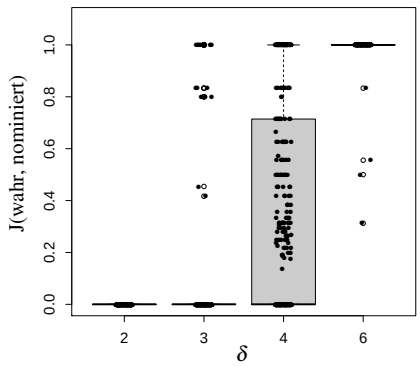
(d) FSJ: nur heterogene Gruppe



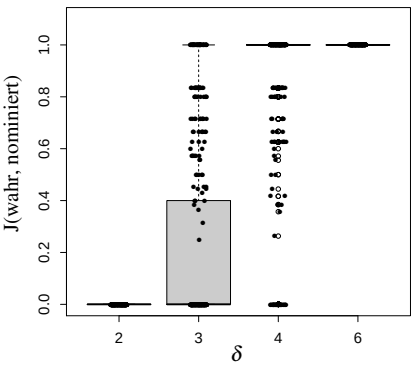
(e) BC: beide Gruppen



(f) BC: nur heterogene Gruppe

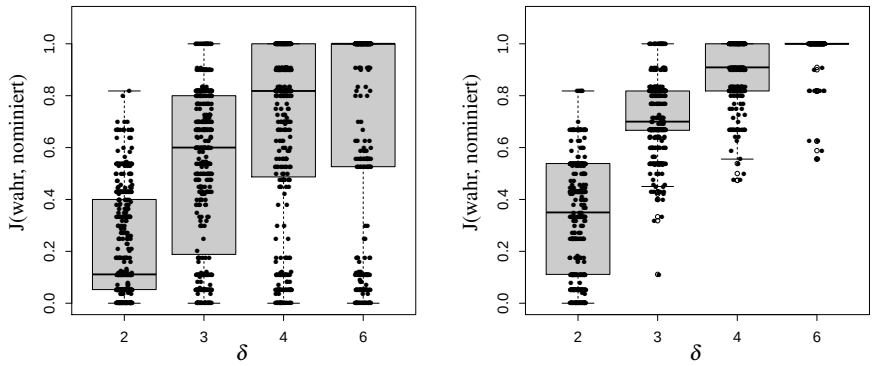


(g) FSBC: beide Gruppen



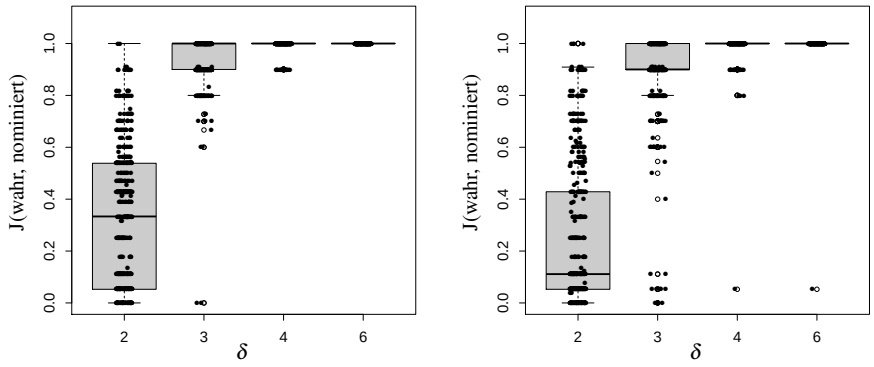
(h) FSBC: nur heterogene Gruppe

Abbildung 46: Einfluss der Sampleauswahl (Parameter `heterOnly`, nur eine oder beide Gruppen) auf die untersuchten SG-Detektionsmethoden FSOL, FSJ, Biclustern und die Kombination FSBC.
Setting: $(n, n_{SG}) = (70, 5)$



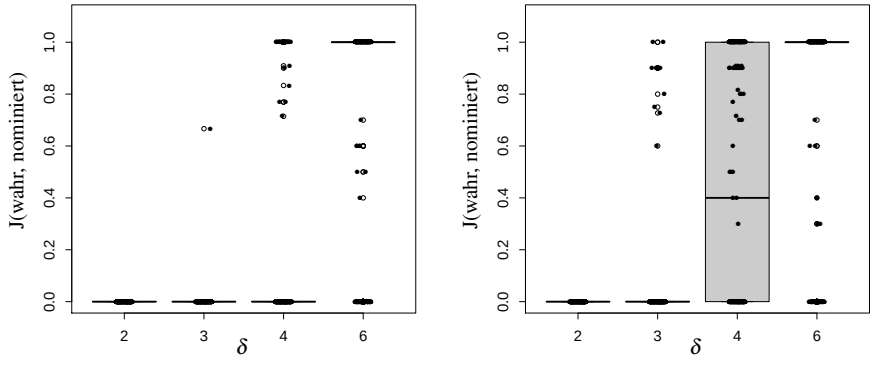
(a) FSOL: beide Gruppen

(b) FSOL: nur heterogene Gruppe



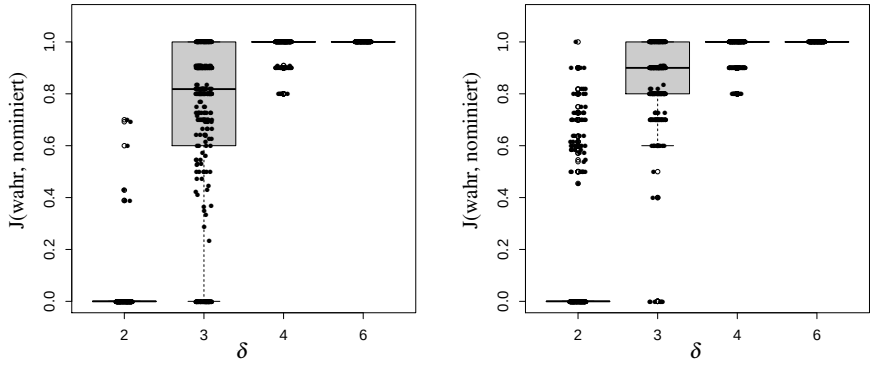
(c) FSJ: beide Gruppen

(d) FSJ: nur heterogene Gruppe



(e) BC: beide Gruppen

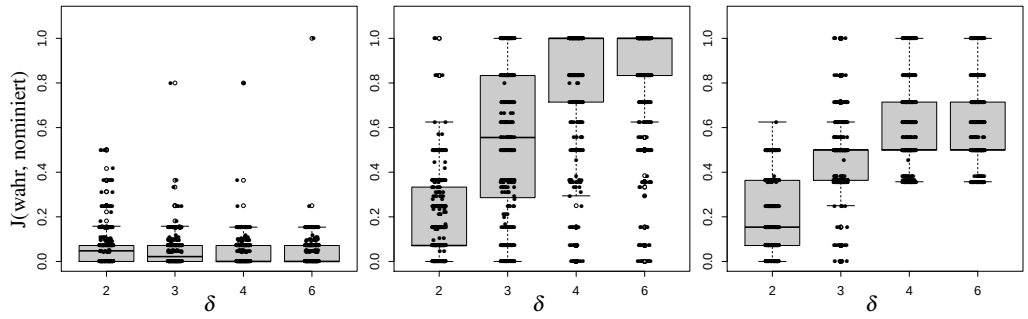
(f) BC: nur heterogene Gruppe



(g) FSBC: beide Gruppen

(h) FSBC: nur heterogene Gruppe

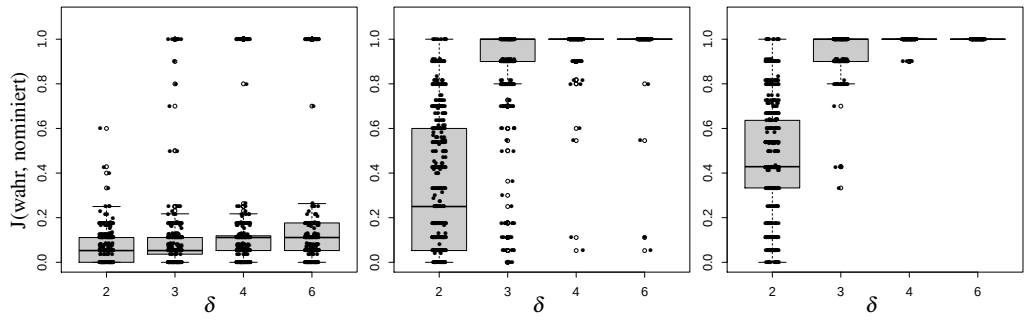
Abbildung 47: Einfluss der Sampleauswahl (Parameter `heterOnly`, nur eine oder beide Gruppen) auf die untersuchten SG-Detektionsmethoden FSOL, FSJ, Biclustern und die Kombination FSBC.
Setting: $(n, n_{SG}) = (70, 10)$



(a) $n_{SG} = 5, t_J = 0.30$, nur heterogene Gruppe

(b) $n_{SG} = 5, t_J = 0.35$, nur heterogene Gruppe

(c) $n_{SG} = 5, t_J = 0.35$, beide Gruppen

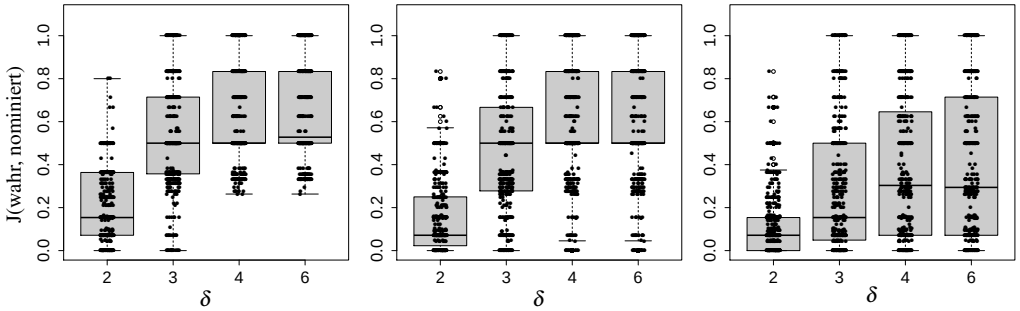


(d) $n_{SG} = 10, t_J = 0.3$, nur heterogene Gruppe

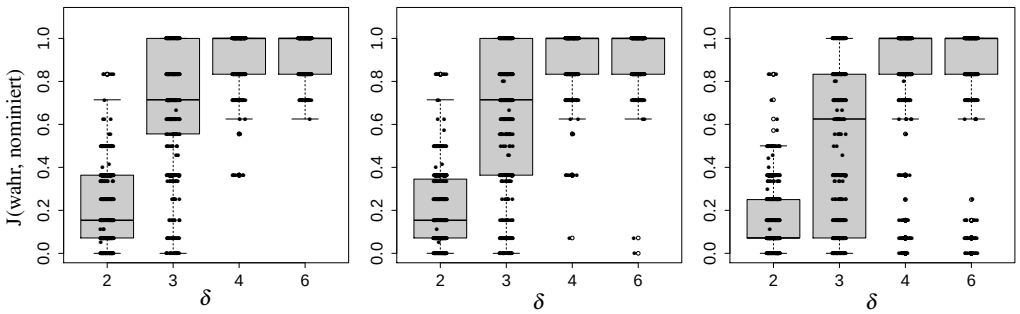
(e) $n_{SG} = 10, t_J = 0.35$, nur heterogene Gruppe

(f) $n_{SG} = 10, t_J = 0.35$, beide Gruppen

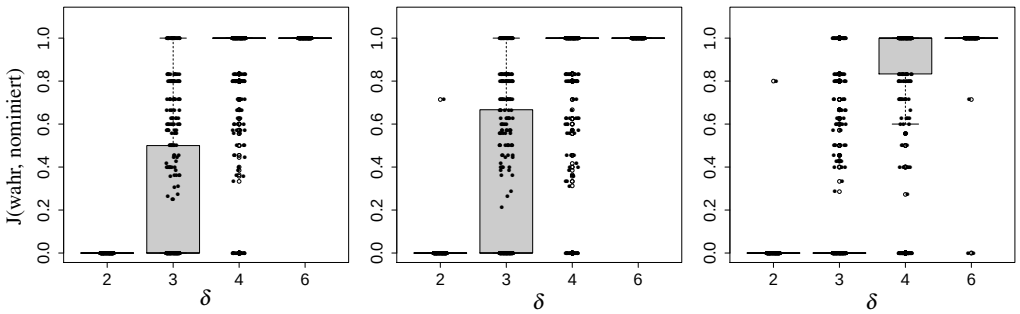
Abbildung 48: Ergänzende Analyse zu Abb. 44 und 45 zur Notwendigkeit der Anpassung des FSJ-cut-offs t_J bei ausschließlicher Ähnlichkeitsbeurteilung der Variablen anhand der heterogenen Gruppe bei $n = 40$.



(a) FSOL: $T = 25$ (b) FSOL: $T = 50$ (c) FSOL: $T = 100$

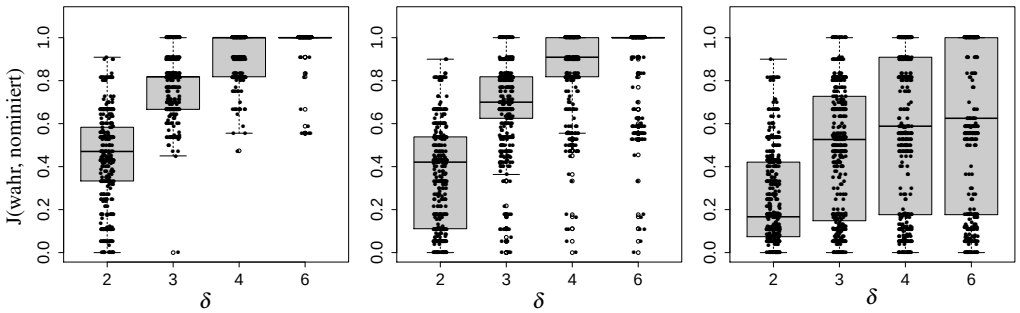


(d) FSJ: $T = 25$ (e) FSJ: $T = 50$ (f) FSJ: $T = 100$

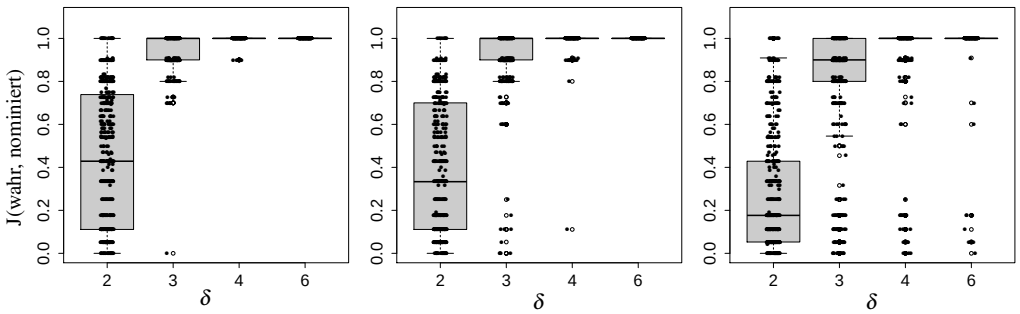


(g) FSBC: $T = 25$ (h) FSBC: $T = 50$ (i) FSBC: $T = 100$

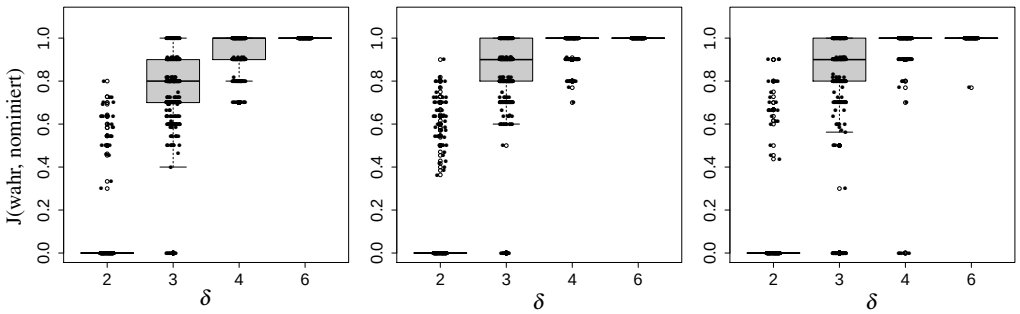
Abbildung 49: Einfluss des Parameters T zur Auswahl der FS-selektierten Variablenanzahl bei der Verwendung von FSOL, FSBC und FSJ.
Setting: $(n, n_{SG}) = (40, 5)$



(a) FSOL: $T = 25$ (b) FSOL: $T = 50$ (c) FSOL: $T = 100$

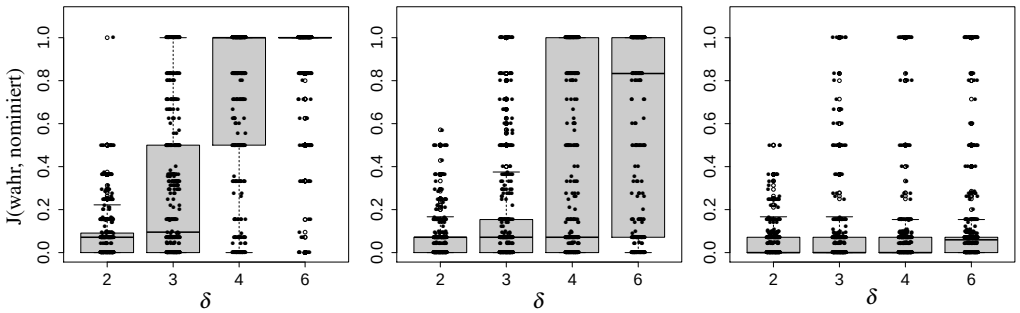


(d) FSJ: $T = 25$ (e) FSJ: $T = 50$ (f) FSJ: $T = 100$

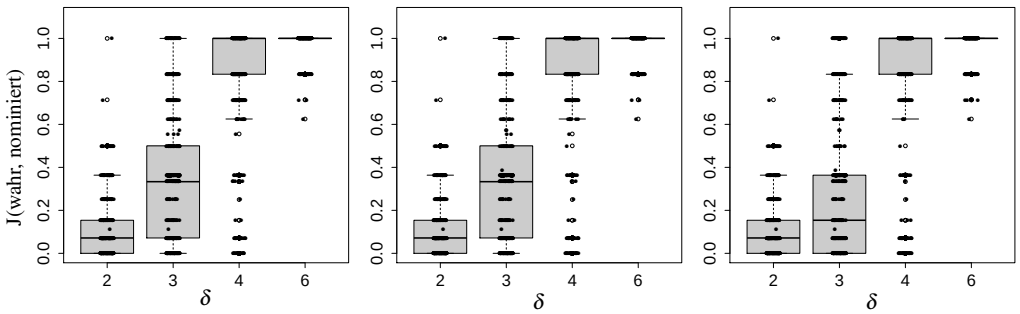


(g) FSBC: $T = 25$ (h) FSBC: $T = 50$ (i) FSBC: $T = 100$

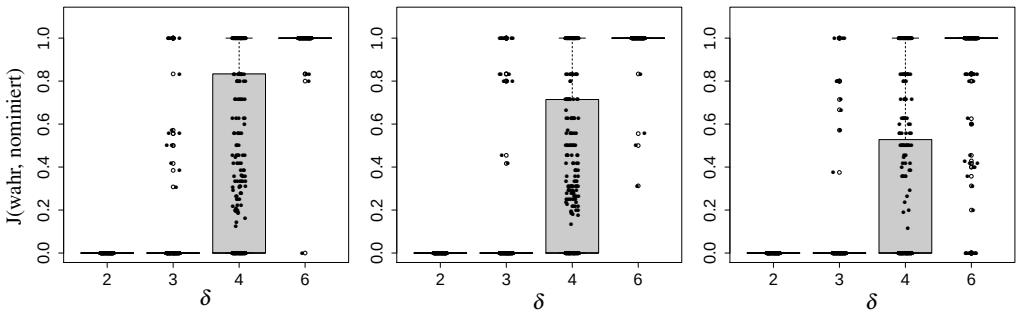
Abbildung 50: Einfluss des Parameters T zur Auswahl der FS-selektierten Variablenanzahl bei der Verwendung von FSOL, FSBC und FSJ.
Setting: $(n, n_{SG}) = (40, 10)$



(a) FSOL: $T = 25$ (b) FSOL: $T = 50$ (c) FSOL: $T = 100$

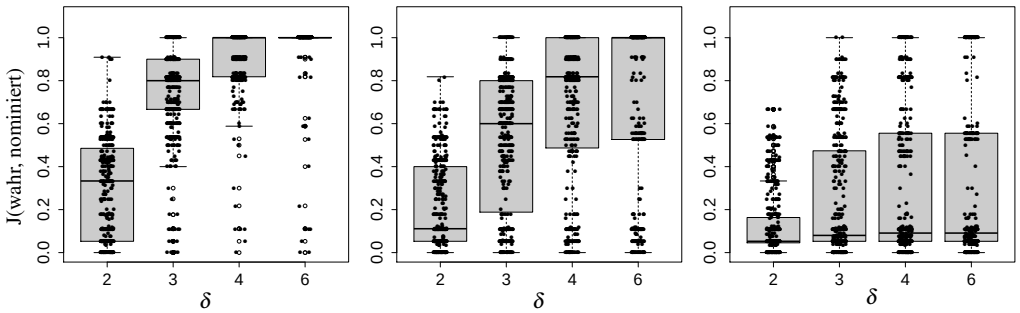


(d) FSJ: $T = 25$ (e) FSJ: $T = 50$ (f) FSJ: $T = 100$

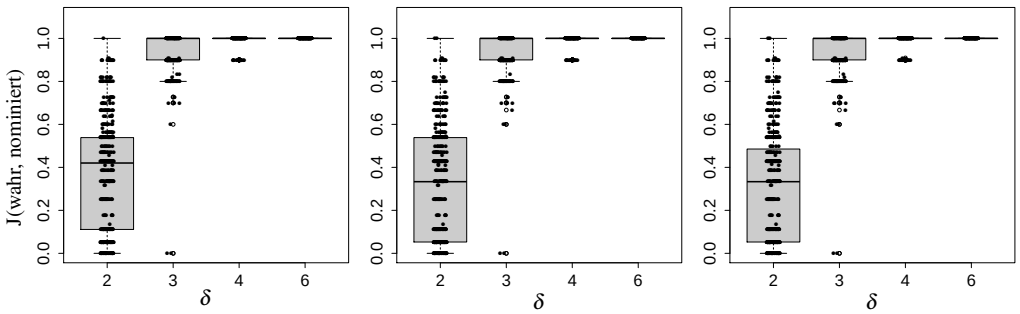


(g) FSBC: $T = 25$ (h) FSBC: $T = 50$ (i) FSBC: $T = 100$

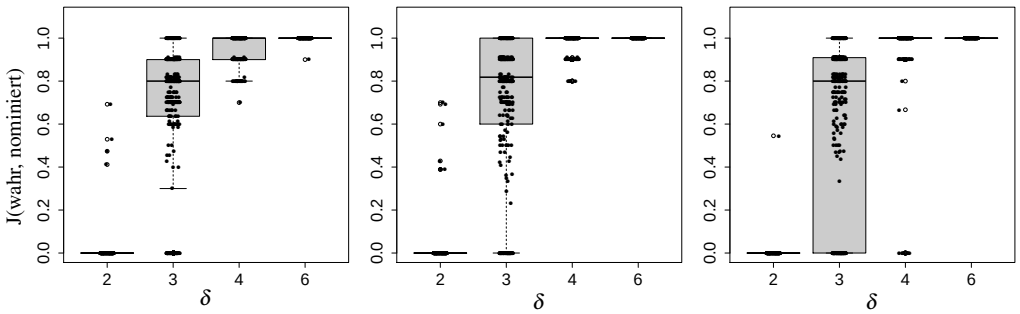
Abbildung 51: Einfluss des Parameters T zur Auswahl der FS-selektierten Variablenanzahl bei der Verwendung von FSOL, FSBC und FSJ.
Setting: $(n, n_{SG}) = (70, 5)$



(a) FSOL: $T = 25$ (b) FSOL: $T = 50$ (c) FSOL: $T = 100$

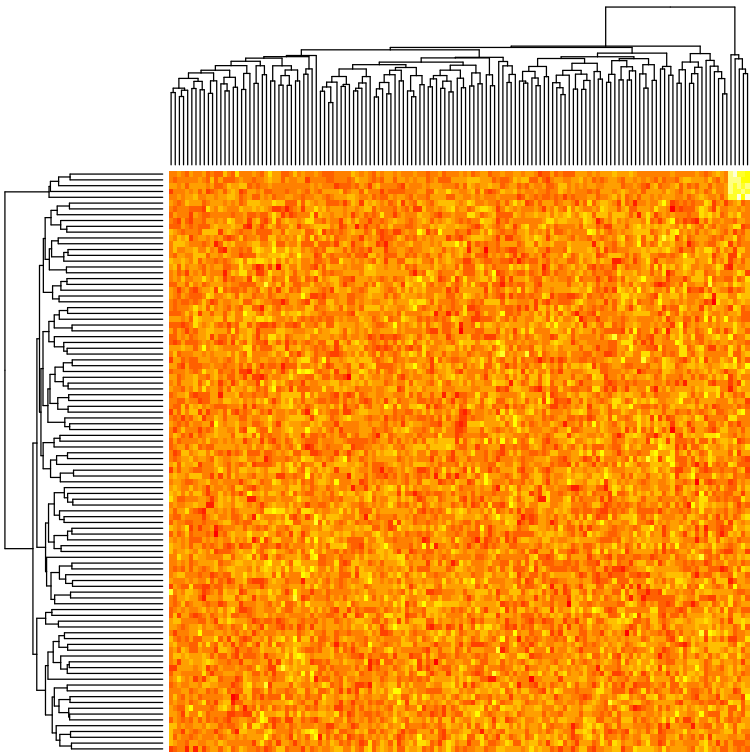


(d) FSJ: $T = 25$ (e) FSJ: $T = 50$ (f) FSJ: $T = 100$

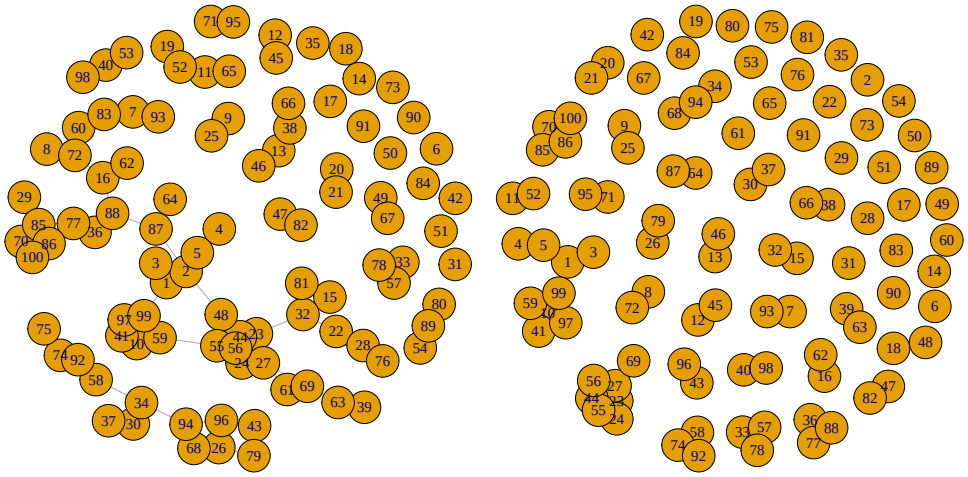


(g) FSBC: $T = 25$ (h) FSBC: $T = 50$ (i) FSBC: $T = 100$

Abbildung 52: Einfluss des Parameters T zur Auswahl der FS-selektierten Variablenanzahl bei der Verwendung von FSOL, FSBC und FSJ.
Setting: $(n, n_{SG}) = (70, 10)$

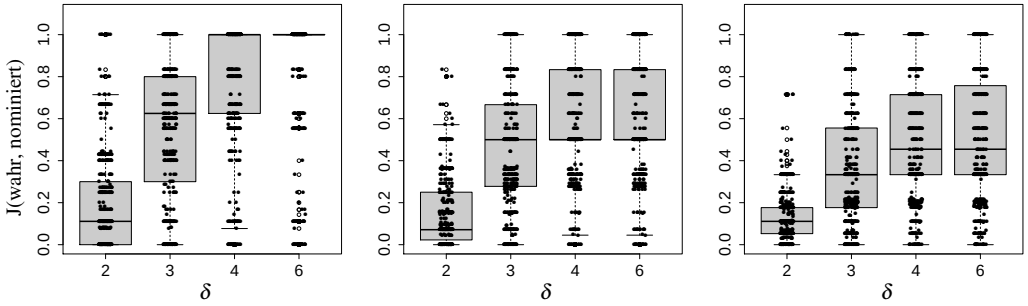


(a) Heatmap top-T-FS-Variablen

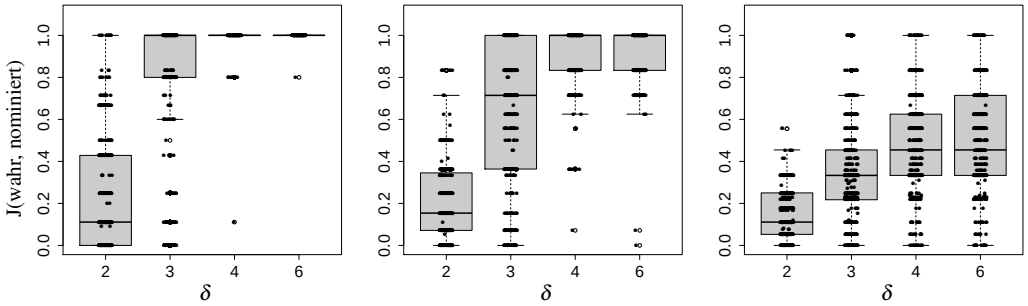


(b) igraph-Darstellungen bei Variation von t_{OL}.
Links mit Standardeinstellung ($t_{OL} = 0.01$), rechts mit $t_{OL} = 0.005$.

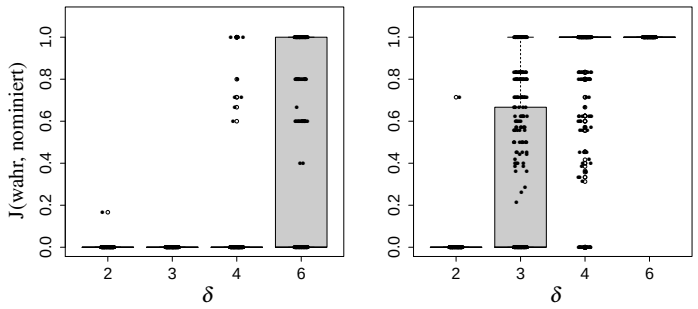
Abbildung 53: Ergänzende Analyse zum Performanzeinbruch von FSOL beim Übergang von $T = 50$ zu $T = 100$, $(n, n_{SG}) = (70, 5)$. (a) Die Heatmap der top-T-FS-Variablen zeigt eine deutliche Abgrenzung der gesuchten Subgruppe. (b) `igraph`-Darstellungen für verschiedene cut-offs $t_{OL} = 0.01, 0.005$. Von Interesse sind die Variablen mit Labels 1 bis 5.



(a) FSOL: $max.rk = 5$ (b) FSOL: $max.rk = 10$ (c) FSOL: $max.rk = 15$

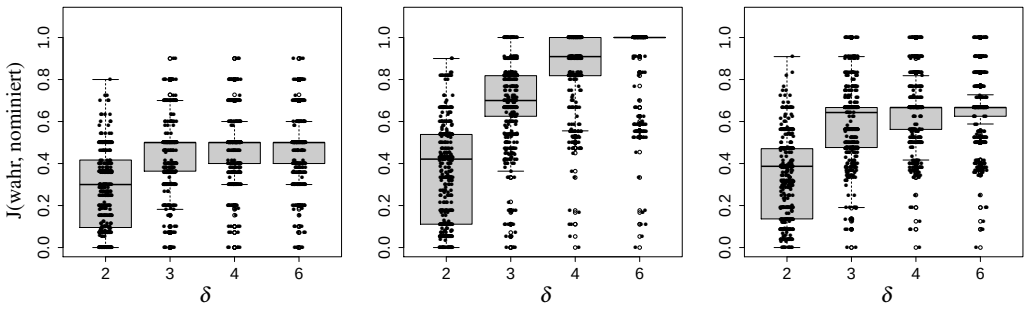


(d) FSJ: $max.rk = 5$ (e) FSJ: $max.rk = 10$ (f) FSJ: $max.rk = 15$

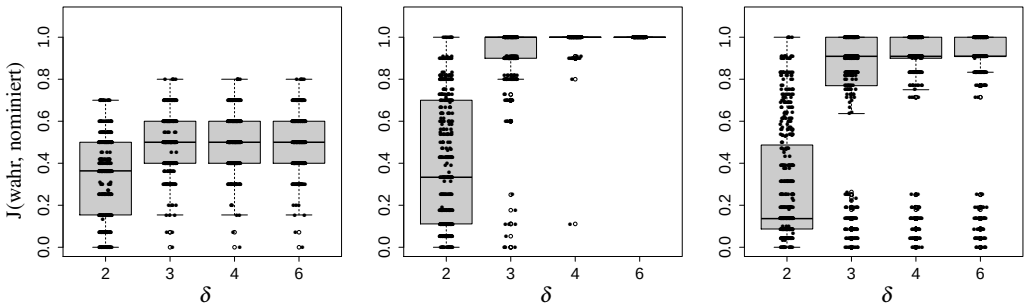


(g) BC (h) FSBC

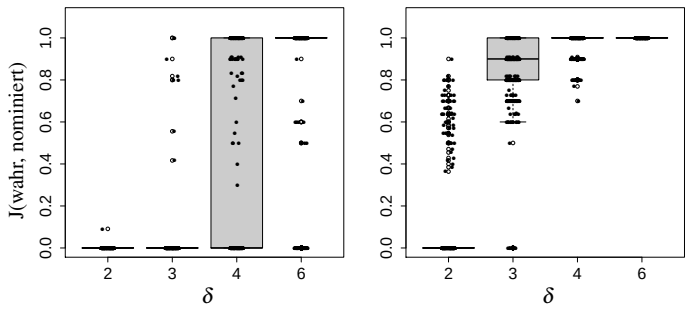
Abbildung 54: Einfluss des Parameters *max.rk* bei der Verwendung von FSOL und FSJ. Zum Vergleich die Ergebnisse der beiden Methoden BC und FSBC, die invariant gegenüber *max.rk* sind.
Setting: $(n, n_{SG}) = (40, 5)$



(a) FSOL: $max.rk = 5$ (b) FSOL: $max.rk = 10$ (c) FSOL: $max.rk = 15$

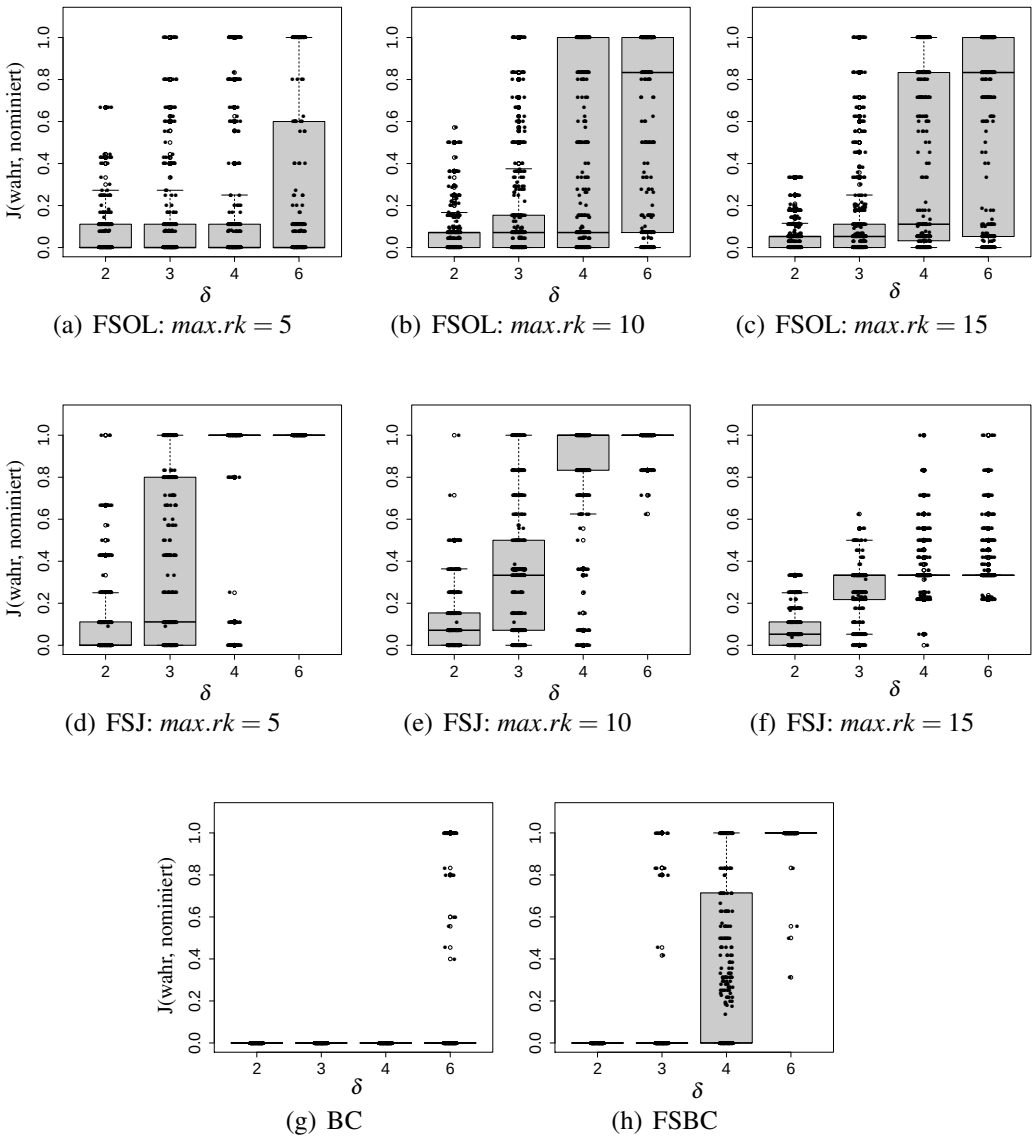


(d) FSJ: $max.rk = 5$ (e) FSJ: $max.rk = 10$ (f) FSJ: $max.rk = 15$



(g) BC (h) FSBC

Abbildung 55: Einfluss des Parameters *max.rk* bei der Verwendung von FSOL und FSJ. Zum Vergleich die Ergebnisse der beiden anderen Methoden BC und FSBC, die invariant gegenüber *max.rk* sind.
Setting: $(n, n_{SG}) = (40, 10)$

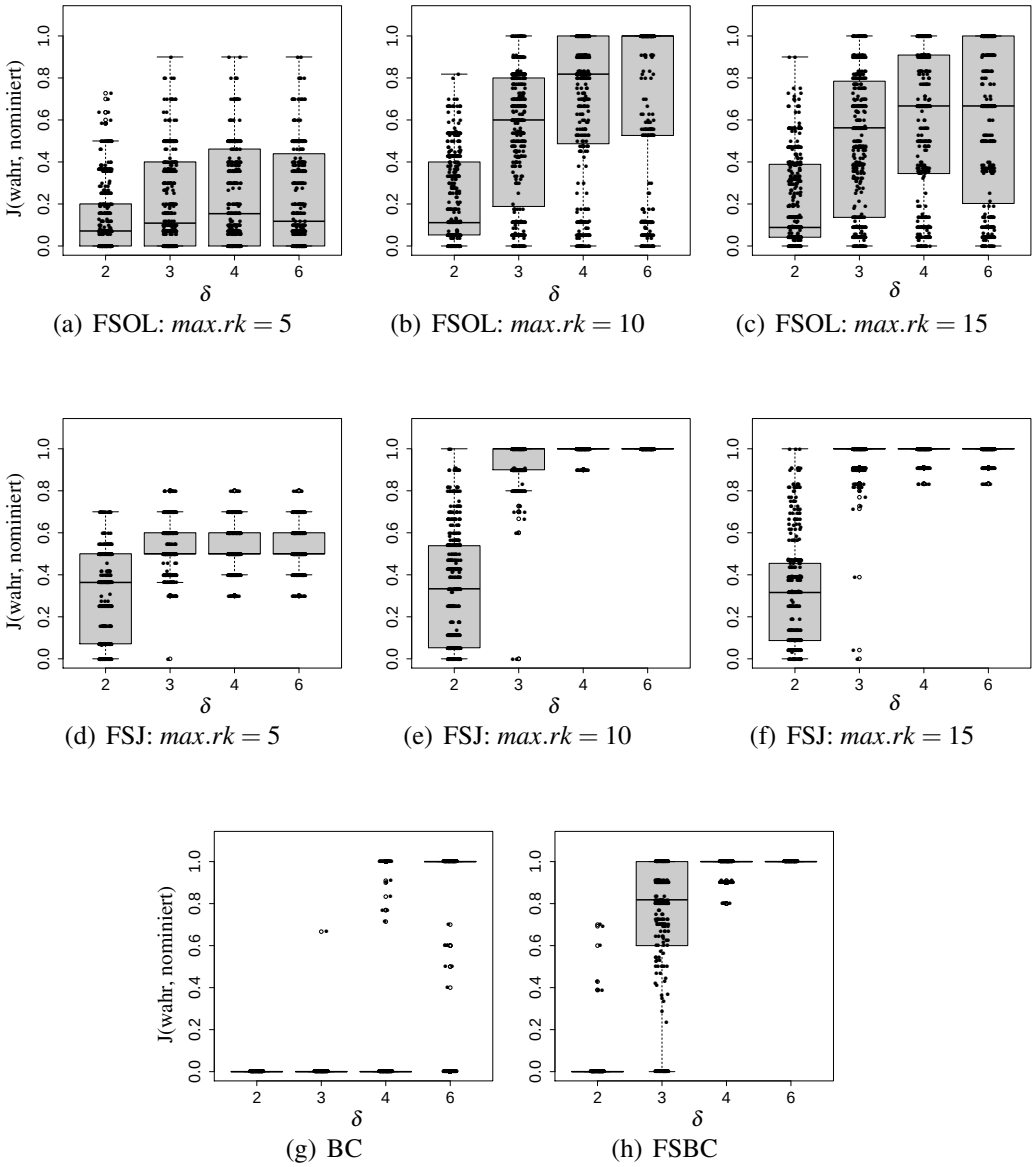


(a) FSOL: $max.rk = 5$ (b) FSOL: $max.rk = 10$ (c) FSOL: $max.rk = 15$

(d) FSJ: $max.rk = 5$ (e) FSJ: $max.rk = 10$ (f) FSJ: $max.rk = 15$

(g) BC (h) FSBC

Abbildung 56: Einfluss des Parameters *max.rk* bei der Verwendung von FSOL und FSJ. Zum Vergleich die Ergebnisse der beiden anderen Methoden BC und FSBC, die invariant gegenüber *max.rk* sind.
Setting: $(n, n_{SG}) = (70, 5)$

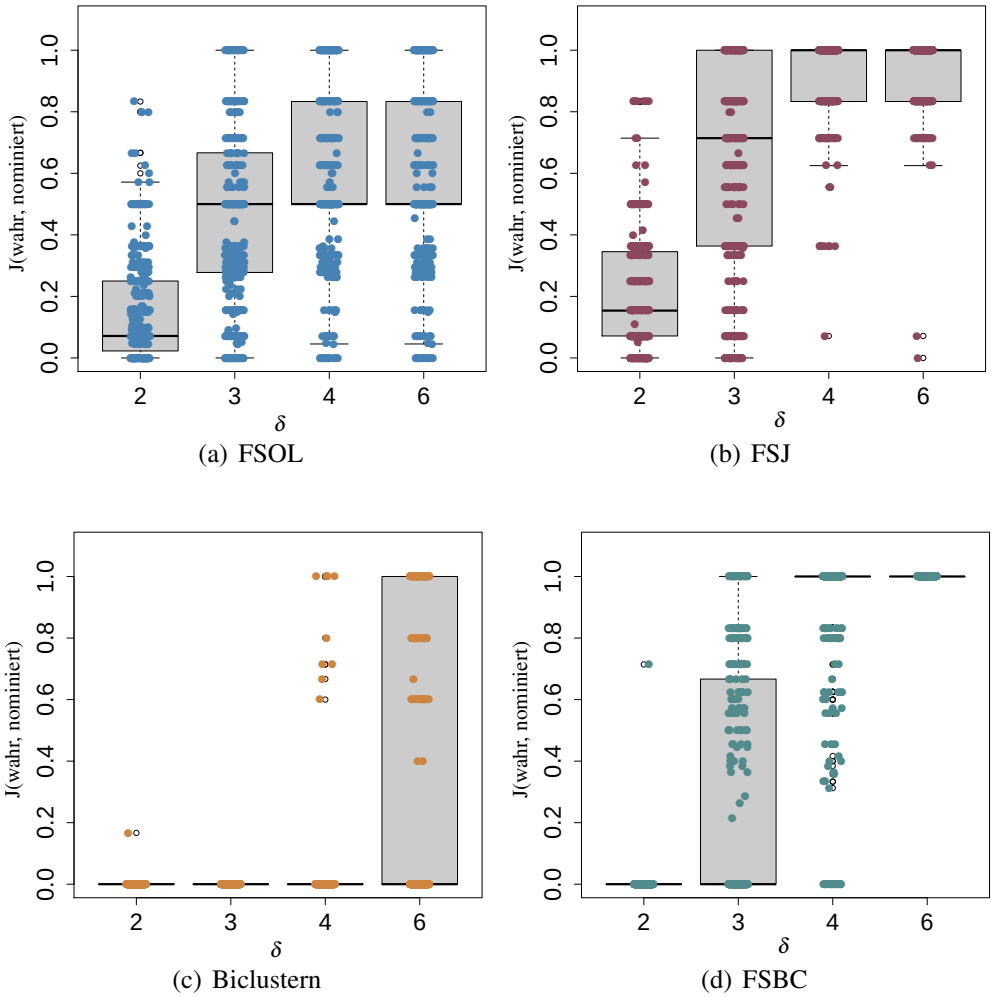


(a) FSOL: $max.rk = 5$ (b) FSOL: $max.rk = 10$ (c) FSOL: $max.rk = 15$

(d) FSJ: $max.rk = 5$ (e) FSJ: $max.rk = 10$ (f) FSJ: $max.rk = 15$

(g) BC (h) FSBC

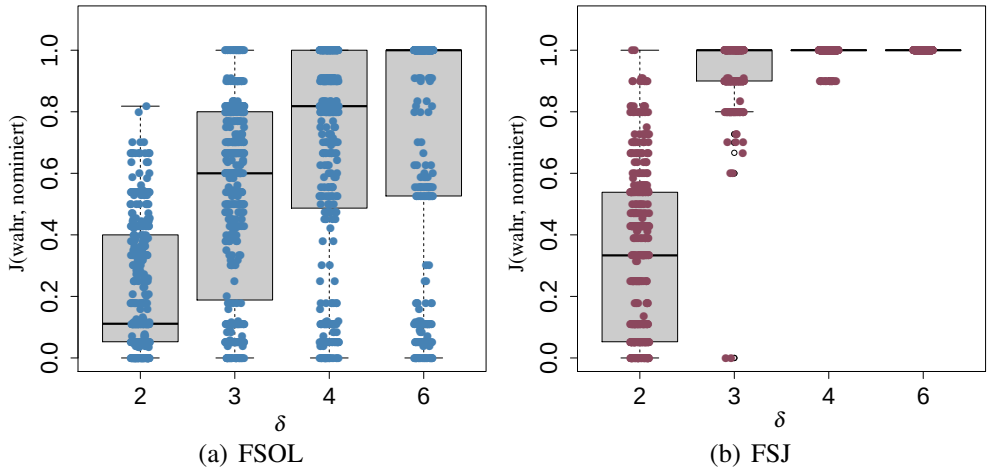
Abbildung 57: Einfluss des Parameters *max.rk* bei der Verwendung von FSOL und FSJ. Zum Vergleich die Ergebnisse der beiden anderen Methoden BC und FSBC, die invariant gegenüber *max.rk* sind.
Setting: $(n, n_{SG}) = (70, 10)$



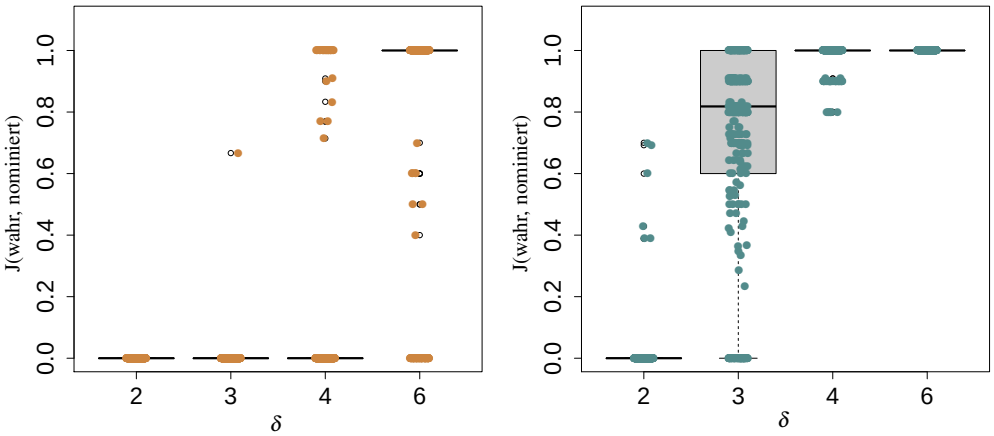
(a) FSOL (b) FSJ

(c) Biclustern (d) FSBC

Abbildung 58: Vergleich der vier Methoden FSOL, FSJ, BC und FSBC für eine fest gewählte Parameterkombination.
Setting: $(n, n_{SG}) = (40, 5)$



(a) FSOL (b) FSJ

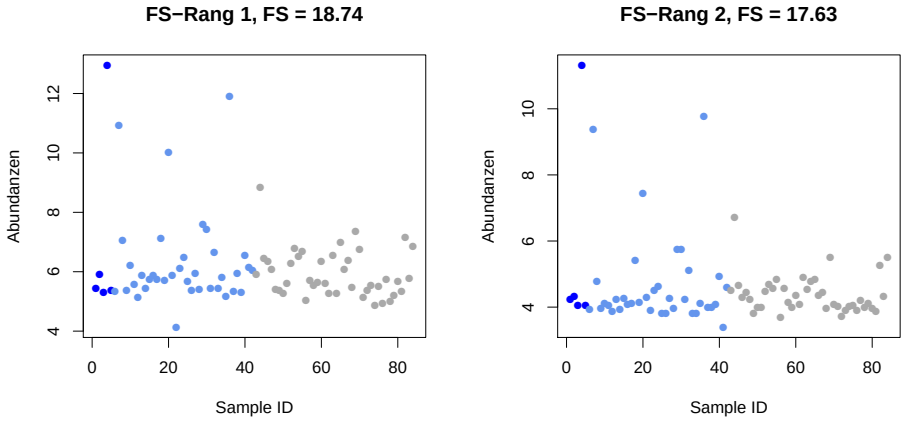


(c) Biclustern (d) FSBC

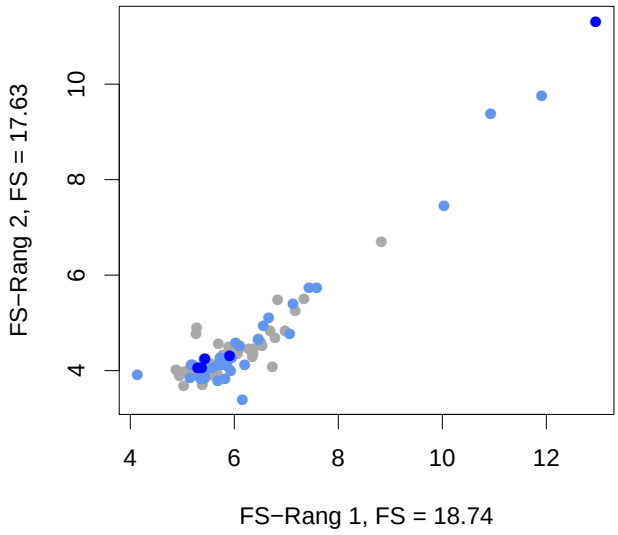
Abbildung 59: Vergleich der vier Methoden FSOL, FSJ, BC und FSBC für eine fest gewählte Parameterkombination.
Setting: $(n, n_{SG}) = (70, 10)$

Anwendung multivariater Workflows auf reale Daten

In diesem Abschnitt werden ergänzende Plots zu den Auswertungen der Datenbeispiele ALL und DeNoPa aufgeführt.
Für ALL beinhaltet das einen Scatterplot der Variablen einer FSJ- Komponente (Abb. 60). Für das DeNoPa-Beispiel werden zunächst die `igraph`-Darstellungen des FSJ-Workflows für zwei verschiedene cut-offs t_J gezeigt (Abb. 61). Abschließend folgt die Betrachtung der Expressionsplots der interessierenden Variablengruppe und der Lage der nominierten Subgruppensamples in Abbildung 62.

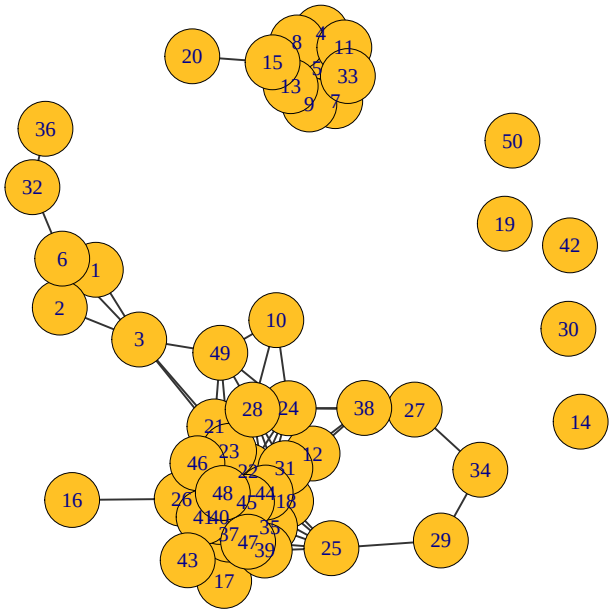


(a) Expressionsplots der Variablen mit FS-Rängen 1 und 2

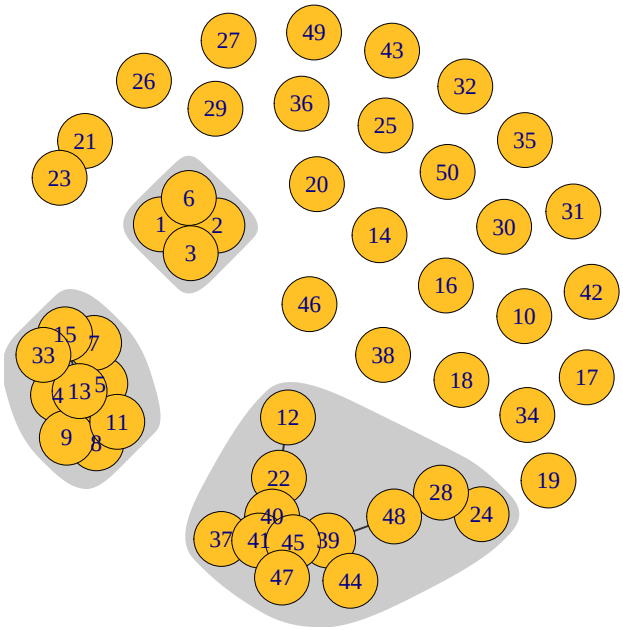


(b) Scatterplot des Variablenpaares

Abbildung 60: Expressionsplots und Scatterplot der Expression der FSOL-Komponente auf Rang 2 für den ALL-Datensatz. Wie in den vorigen Darstellungen codiert grau für die Gruppe NEG, hellblau für BCR/ABL und dunkelblau für die gesuchte Subgruppe vom Typ E2A/PBX1.

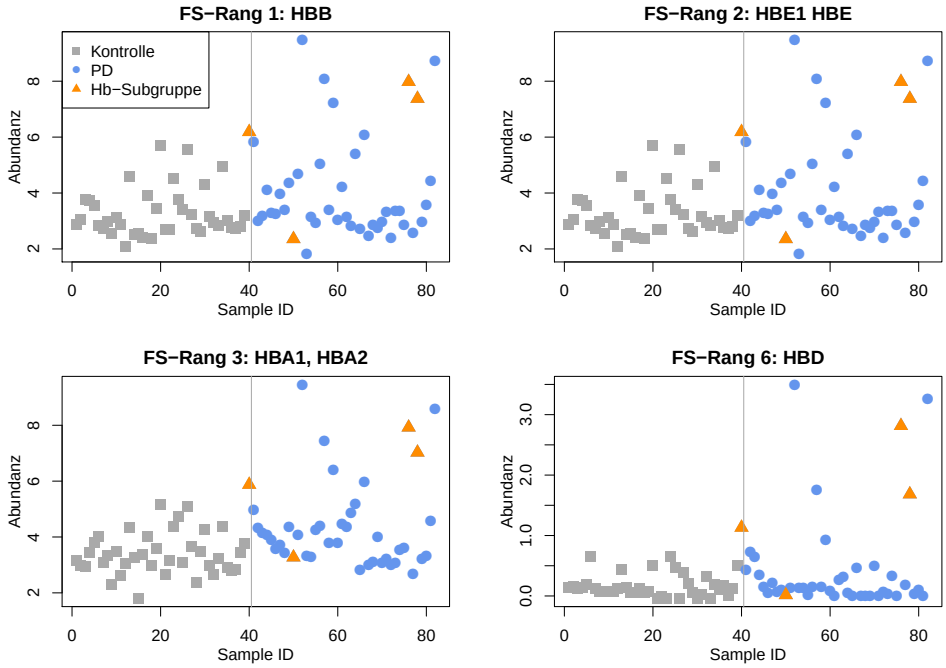


(a) FSJ, cut-off $t_J = 0.3$

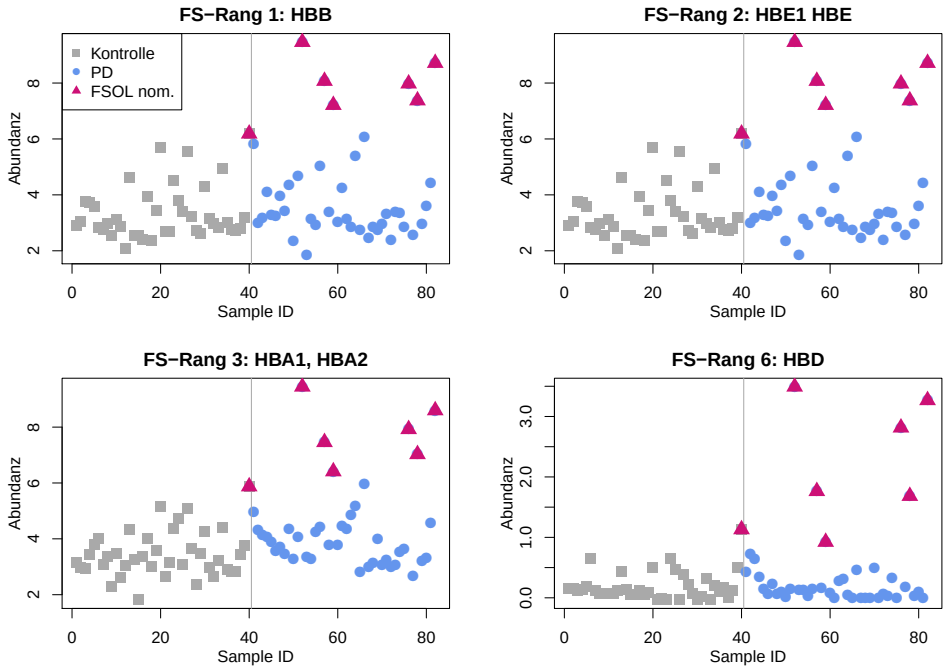


(b) FSJ, cut-off $t_J = 0.5$

Abbildung 61: Gezeigt werden die `igraph`-Darstellungen der Variablengruppierung bei FSJ für den DeNoPa-Datensatz. (a) Bei Verwendung des Standard-cut-offs $t_J = 0.3$ bleibt eine große Variablenmenge in einer großen Komponente verbunden. (b) Bei Erhöhung des cut-offs auf $t_J = 0.5$ können die angezeigten Subgruppen für die einzelnen Variablengruppen bestimmt werden.



(a) Markierung der ELISA-Hb-Subgruppe



(b) Markierung der von FSOL nominierten Subgruppe

Abbildung 62: Gezeigt werden die Abundanzplots der vier mit Hämoglobin annotierten Variablen in den top-50-FS-Variablen des DeNoPa-Datensatzes. (a) Eines der vier Samples aus der ELISA-Hb-Subgruppe zeigt auf Ebene der label-freien Daten keine erhöhten Werte. (b) Markierung der von FSOL nominierten Samples.