Verbesserung des Chen-Qin-Test im Ein-Gruppen-Design

Diplomarbeit

vorgelegt von David Ellenberger
aus Kassel

angefertigt im
Institut für Mathematische Stochastik
der Georg-August-Universität Göttingen

2011

Inhaltsverzeichnis

1 Einleitung

1.1 Historie und Einleitung

Die vorliegende Arbeit behandelt Testverfahren für den Ein-Stichprobenfall von strukturierten verbundenen Messungen. Getestet werden soll eine über den Erwartungswert $\boldsymbol{\mu}$ formulierte Hypothese $\mathbf{H}\boldsymbol{\mu} = \mathbf{0}$. Dabei liegt das Interesse speziell auf hochdimensionalen Designs, die eine stetig wachsende Bedeutung für die Anwender erfahren. Während ursprüngliche Testverfahren wie Hotellings T^2-Test oder die ANOVA-Typ-Statistik nur im Niedrigdimensionalen geeignet sind, ließen sich für normalverteilte Zufallsvektoren starke asymptotische Resultate erzielen. So konne Werner (2004) eine modifizierte ANOVA-Typ-Statistik entwickeln, welche stabil für große Dimensionen ist.
Eine Verallgemeinerung auf hochdimensionale nichtnormale Daten konnte mit einer von Bai und Saranadasa (1996) eingeführten, speziellen multivariaten Modellierung erfolgen. Diese schränkt zwar die Abhängigkeitsstrukturen zwischen den Komponenten ein, hat sich aber mittlerweile als Minimalanforderung in der Forschung der Linearen Modelle etabliert.
Beruhend auf den Ideen des Testes für den Zweistichprobenfall von Bai und Saranadasa (1996) konnten Chen und Qin (2010) einen weiterentwickelten Test unter neuen reduzierten Annahmen herleiten. Die theoretische Grundlage dieses asymptotischen Tests beruht auf dem Zentralen Grenzwertsatz für Martingale und gilt für eine Asymptotik über den Stichprobenumfang und der Dimension gegen unendlich. Somit kommt eine Anwendung überhaupt nur für hochdimensionale Daten in Betracht. Außerdem werden zusätzliche Regularitätsannahmen an die Kovarianzmatrix gestellt.

In dieser Arbeit wurde nun ein Test entwickelt, welcher unabhängig von der Dimension angewendet werden kann und möglichst wenig Annahmen an die Verteilung der Daten stellt. Dafür wurde die klassische Theorie der ANOVA-Typ-Statistik über die Verteilung Quadratischer Formen unter Normalverteilung auf die neuen Ergebnisse von Chen und Qin angewendet. Mit Hilfe der Boxapproximation ergibt sich ein Test mit einer modifizierten Chen-Qin-Teststatistik. Eine entscheidende Veränderung stellt dabei der Ansatz dar, eine standardisierte Statistik zu bilden. Demgegenüber ist der Quotient der ANOVA-Typ-Statistik nur bezüglich des Erwartungswertes stabil, allerdings nicht bezüglich der Varianz.
Auf natürlichem Wege wird dadurch eine Momentenapproximation von Verteilungen

motiviert, welche auf dem dritten Moment beruhen. Diese Drei-Momentenapproximation wird als Pearsonapproximation bezeichnet und als Alternative zur Boxapproximation eingeführt. Die Vorteile des daraus entwickelten Tests zeigen sich vor allem bei einem strengen Testniveau $\alpha < 0.05$.

1.2 Aufbau der Arbeit

Zunächst wird ein praktischen Beispiel angegeben, welches neue Verfahren für den Ein-Stichprobenfall motivieren soll. Im folgenden Kapitel werden die Annahmen an die Verteilung der Daten formuliert und die zu testenden Hypothesen charakterisiert. Dies dient als Ausgangslage für die theoretischen Ergebnisse der darauffolgenden Kapitel 4-8. Im Kapitel 4 werden zunächst die bekannten Verfahren vorgestellt. So auch der Test von Chen und Qin, welcher verbessert werden soll. In Kapitel 5 wird dafür die Normalapproximation durch eine neue Verteilung ersetzt, die auf der Boxapproximation beruht, und der Übergang beider ineinander für ansteigende Dimensionen gezeigt. Mit den Ergebnissen lässt sich unter Entwicklung von geeigneten Schätzern ein dimensionsstabiler Test entwickeln.
Im Kapitel 7 wird die Pearsonapproximation eingeführt und die Verbesserung zur Boxapproximation diskutiert. Desweiteren wird das Problem der Schätzung der neuen Parameter im Hochdimensionalen gelöst. Die teils sehr technischen Beweise wurden in den Anhang gefügt. Mit der neuen Approximation kann somit ein weiter verbessertes Testverfahren vorgestellt werden. Einschließlich der Resultate aus Kapitel 8 können abschließend minimale Verteilungsannahmen und Möglichkeiten der Überprüfung der sogenannten Chen-Qin-Bedingung angegeben werden.

Im nächsten Teil der Arbeit werden weitere Verallgemeinerungen über das bisherige Modell hinaus untersucht. Zunächst wird eine Klasse von degenerierten multivariaten Verteilungen konstruiert, unter welcher die Güte der Schätzer und die Approximation der Teststatistik mit steigender Dimension beliebig schlecht werden. Alternative Testverfahren unter Symmetrie werden vorgestellt und diskutiert.
Exemplarisch werden Simulationen durchgeführt, um die theoretischen Resultate der Arbeit zu verifizieren und deren Güte zu überprüfen. Desweiteren wird das Beispiel mit Hilfe des im Anhang angegebenen Makros ausgewertet. Die Arbeit wird mit dem Ausblick auf weitere Fragestellungen abgeschlossen.

2 Beispiel: Cortisol-Konzentration im Blutplasma

Zunächst soll ein praktisches Beispiel für das Ein-Gruppen-Design illustriert werden. Vom Ein-Gruppen-Design spricht man, wenn die Patienten, sprich die unabhängigen Subjekte, alle aus der gleichen Grundgesamtheit stammen. Dies bedeutet, dass man keine unterschiedlichen Behandlungen an den n Subjekten vorgenommen hat. Stattdessen liegen für jedes Subjekt Mehrfachmessungen vor und man interessiert sich für Effekte innerhalb dieser abhängigen Messwiederholungen.
Im folgenden Beispiel wurden 12 Marathonläufer, welche als Leisungssportler regelmäßig trainieren, einer zweiwöchigen Trainingspause ausgesetzt und ihr Befinden vor und nach dieser Trainingspause untersucht. Die einzelnen Sportler werden dabei als stochastisch unabhängig modelliert, während sämtliche Messungen an einem Sportler als abhängig betrachtet werden.

2.1 Versuchsaufbau

Ziel der Studie ist es, ein besseres neurobiologisches Verständnis von Trainingsphasen im (Leisungs-)Sport zu erhalten. Vor allem gilt es zu untersuchen, ob und warum Leistungssportler, welche eine plötzliche Trainingspause einlegen, häufig sogenannte „Sportentzugserscheinungen“ erleiden. Diese äußern sich in Symptomen wie Schwindel, Schmerzen am Herz, Verdauungsproblemen, allgemeinem Unwohlbefinden, Schlaflosigkeit oder gar Depression. Die Trainingspause wird nun anhand von Effekten auf das Nervensystem untersucht. Dabei dient als Zielgröße der Untersuchung das Stresshormon Cortisol, welches eine große Bedeutung in Bezug auf das zentrale Nervensystem und auf den Neurotransmitter Serotonin hat.
Cortisol wird vom Körper produziert, um Aufmerksamkeit und Anspannung zu erhöhen. Die Produktion von Cortisol findet in der Nebennierenrinde statt und dessen Ausschüttung erfolgt in regelmäßigen Schüben. Diese circadianen Schübe ereignen sich 7-10 mal am Tag, wobei die maximale Cortisolkonzentration meist morgens nach dem Aufwachen gemessen wird (Cortisol Awakening Response). Wegen der starken Schwankung enthält eine einzelne Messung nur wenig Information, weshalb Messungen in regelmäßigen Abständen durchgeführt wurden. Die Cortisolkonzentration im Blutserum der Marathonläufer wurde deshalb in Abständen von 30 bis 60 Minuten gemessen. Dies geschah an 7 Zeitpunkten über einen Zeitraum von 4 Stunden, so dass

mindestens eine Ausstoßphase abgedeckt wurde.
Gesteuert wird die Produktion von Cortisol unter anderem durch körpereigene Botenstoffe wie Adrenalin, sie kann allerdings auch durch Verabreichung von Medikamenten wie beispielsweise der Substanz mCPP erhöht werden. So wurde bei sämtlichen Marathonläufern eine Messung der Cortisolkonzentration unter Verabreichung von mCPP und eine unter Placebo durchgeführt. Dies geschah jeweils vor und nach der Trainingspause, wobei die Reihenfolge der Placebo-Messung und der mCPP-Messung randomisiert wurde. Außerdem wurde darauf geachtet, dass eine hinreichend große Auswaschzeit zwischen den Messungen lag, um Interferenzen zu vermeiden.
Somit ergeben sich pro Marathonläufer 28 Messungen, welche sich aus Kombination der unterschiedlichen Faktorstufen der drei Faktoren

- Trainingspause (TP): Mit 2 Faktorstufen „vor“ und „nach“
- Stimulation (ST): Mit 2 Faktorstufen Placebo und mCPP
- 4-Std-Profil (PR): Mit 7 Faktorstufen 0 min, 30 min, 60min, 90 min, 120 min, 180 min, 240 min

ergeben. Dieses 3-faktorielles Repeated-Measures-Design wird als HD-F3 bezeichnet. Dabei steht HD für „high dimensional data“ und ersetzt die klassische Bezeichnung LD für „longitudinal data“.

2.2 Formulierung der Hypothesen

Standardmäßig werden nun sämtliche Hypothesen des HD-F3-Modells getestet. Dies sind die Haupteffekte der drei Faktoren sowie deren Wechselwirkungen.
Die primäre Zielsetzung der Studie ist es aber, zu untersuchen, welche Auswirkungen die Trainingspause auf die Leistungssportler hat. Die Messung der Cortisolkonzentration im Blut dient als ein Indikator für Stress im Hormonsystem. Stress ist dabei zunächst eine positive Reaktion des Körpers, durch welche die Leistungsfähigkeit gesteigert wird. Dies geschieht sowohl in körperlicher als auch in geistiger Hinsicht. Entscheidend für das Wohlbefinden im Körper ist nun, dass die aktivierte Energie verbraucht wird und der Hormonspiegel sich wieder auf ein Normalniveau einstellt, was beispielsweise bei intensivem Training der Fall ist. Findet also ein Ausgleich statt, spricht man von positivem Stress, auch „Eustress“ genannt. Demgegenüber liegt der negative Stress „Disstress“ dann vor, wenn bereitgestellte Energie nicht durch Aktivität abgebaut wird. Auch auf Disstress kann der Körper im Normalfall reagieren und das Hormonsystem wieder in Einklang bringen. Dauerhafte schwerwiegende Fälle von Disstress können allerdings chronischen Stress, Schlafstörungen, Depressionen und Burnout-Syndrom zur Folge haben.

In Bezug auf die Studie würde man dementsprechend erwarten, dass der Cortisolspiegel über die Trainingspause hin ansteigt (Haupteffekt TP). Dies lässt sich einerseits mit einer erhöhten Cortisol-Toleranz bei Leistungssportlern, bevor der Körper eine Gegenreaktion auf Stress zeigt, erklären. Andererseits ist es auch möglich, dass durch die abrupte Umstellung der Abbauprozess von Cortisol unter Nichtbelastung grundsätzlich gestört ist. Um Antworten auf diese Frage zu erhalten, ist die Wechselwirkung zwischen der Trainingspause und der Stimulation mit dem Cortisol-Agonisten mCPP zu betrachten ($TP \times ST$). Kann das Hormonsystem der Leistungssportler auf den durch mCPP induzierten zusätzlichen Stress reagieren, oder tritt ein gar noch extremerer Anstieg der Cortisol-Konzentration auf?
Um weiter ins Detail zu gehen, ist auch zu bewerten, ob sich Veränderungen im 4-Stunden-Profil der Cortisolkonzentration feststellen lassen ($TP \times PR$). Dies würde bedeuten, dass sich die Rhythmik der Ausschüttung von Cortisol durch die Trainingspause verändert hätte. Auch die Dreifach-Wechselwirkung ($TP \times ST \times PR$) ist hier zu untersuchen und muss mit den Rückschlüssen der anderen Effekte in Einklang gebracht werden. In der Auswertung von Brunner, Domhof, Langer 2002 [6] wurden Zeiteffekte im 4-Stunden-Profil nicht betrachtet und stattdessen über die 7 Zeitpunkte gemittelt. Für einen Test bezüglich der Dreifach-Wechselwirkung werden Verfahren benötigt, welche stabil in der Dimension sind, da man 28 Messwerte nicht als niedrigdimensional betrachten kann.
Neben den Hauptfragestellungen, welche alle über die Trainingspause formuliert werden, sind sämtliche anderen Hypothesen des voll gekreuzten Versuchsplans zu testen. Ein Anstieg der Cortisolkonzentration unter Stimulation (Haupteffekt ST) ist zu erwarten, da mCPP als Cortisol-Agonist eingesetzt wurde. Dasselbe gilt für den Zeiteffekt (Haufteffekt PR), da das Cortisoltagesprofil circadianen Schwankungen unterworfen ist und im Laufe des Tages abnimmt. Die Wechselwirkung der beiden Effekte ($ST \times PR$) ist in Einklang mit den Rückschlüssen der Studie zu bringen.

2.3 Messwerte

Die Studie inklusive der Datensätze stammt aus dem Buch „Nonparametric analysis of longitudinal data in factorial experiments“ von Brunner, Domhof und Langer, 2002 [6]. Die Profile der Cortisolkonzentration ergeben sich wie folgt:

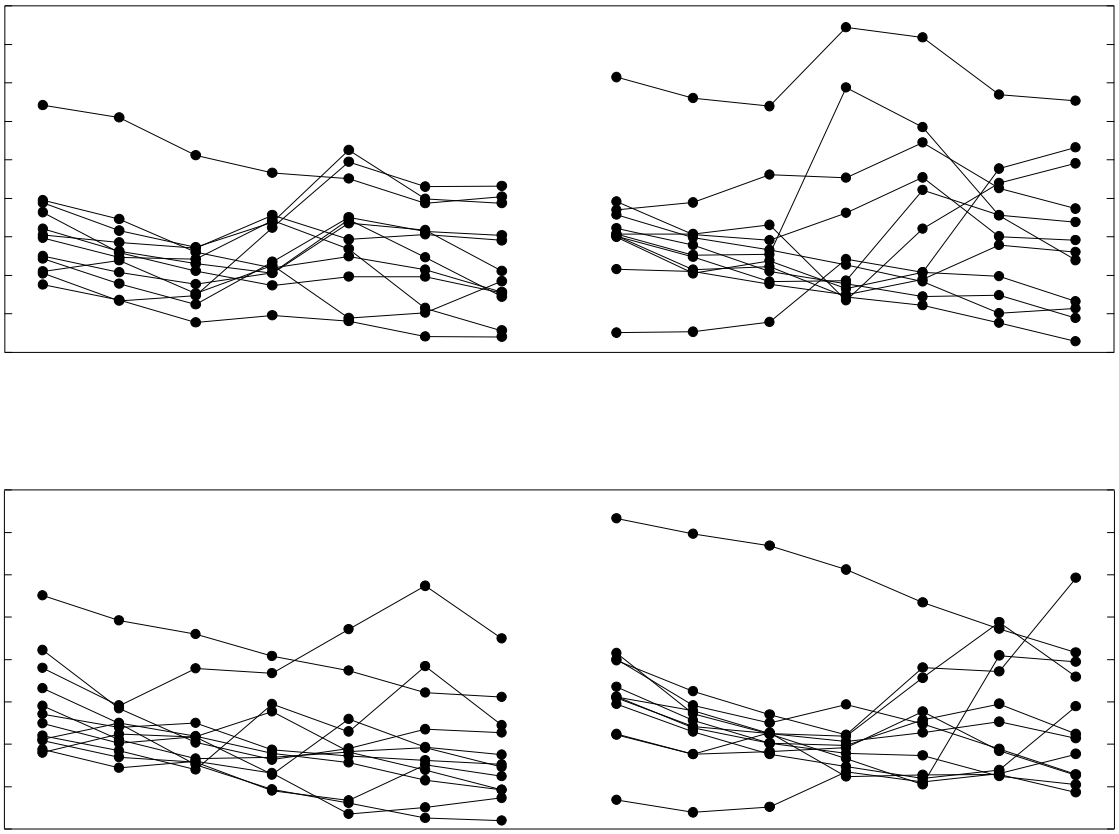


Abbildung 2.1: Profile der Cortisolkonzentration im Blutplasma in $[\mu g/dl]$ über die verschiedenen Zeitpunkte. Im oberen Bild sind die Verläufe unter Stimulation mit m-CPP der 12 Marathonläufer „vor“ und „nach“ der Trainingspause abgebildet, im unteren Bild die unter Placebo.

Anhand der Zeitprofile der Cortisolkonzentration kann man erkennen, dass zeitlich benachbarte Messwerte stark korrelieren. Es sind klare Zeitreihen erkennbar und eine autoregressive Modellierung scheint möglich. Dennoch ist es wünschenswert, hier auf eine Annahme an die Kovarianzstruktur zu verzichten und eine unstrukturierte Kovarianzmatrix zuzulassen.
Desweiteren ist auch die Annahme, dass die Ursprungsdaten einer multivariaten Normalverteilung folgen, fallenzulassen. Die Daten sind durch die 0 nach unten beschränkt und anhand der Boxplot lässt sich auch eine leichte Schiefe der Randverteilungen erkennen. Die einzige Annahme an die Marginalverteilungen soll ein beschränktes viertes Moment der Randverteilungen sein.

3 Notation und Modell

3.1 Notation

In diesem Abschnitt wird die für diese Arbeit grundlegende Notation gegeben, welche im Folgenden benutzt wird.

SKALARE werden mit kleinen, nicht fett gedruckten Buchstaben bezeichnet, während ein-dimensionale *ZUFALLSVARIABLEN* mit großen, nicht fett gedruckten Buchstaben bezeichnet werden. Für Dimensionen bezüglich (un-)beobachtbarer Zufallsvektoren wird d,p oder m benutzt und für den Stichprobenumfang stets n. Insbesondere wird für Indizes bezüglich der Dimension meist i,j,h,i',j',h' verwendet, während für Indizes bezüglich der Stichprobe vornehmlich k,l,r,k',l',r' verwendet werden.

VEKTOREN und *MATRIZEN* werden stets in Fettdruck hervorgehoben. Insbesondere wird im d-dimensionalen Raum der Nullvektor mit $\mathbf{0}_d$, der Einservektor mit $\mathbf{1}_d$ und die Einheitsmatrix mit $\mathbf{I}_d$ bezeichnet. Auch hier werden Zufallsvektoren in Großbuchstaben (meist $\mathbf{X}_k, \mathbf{Y}_k$ oder $\mathbf{Z}_k$,) und feste Vektoren in Kleinbuchstaben gesetzt.
TRANSPONIERTE Vektoren bzw. Matrizen werden mit $\mathbf{A}'$ dargestellt. Wenn nicht anders definiert, sind Vektoren stets aus $\mathbb{R}^{d \times 1}$ (zum Beispiel $\boldsymbol{\mu} = (\mu_1, \mu_2)'$).
Desweiteren sei $\mathbf{J}_d = \mathbf{1}_d'\mathbf{1}_d$ die $d \times d$-Matrix, welche nur Einsen enthält. Die Bezeichnung $\mathbf{P}$ wird sowohl als Projektor, insbesondere als $\mathbf{P}_d = \mathbf{I}_d - \frac{1}{d}\mathbf{J}_d \in \mathbb{R}^{d \times d}$, verwendet, als auch als orthogonale Basiswechselmatrix $\mathbf{P}_\mathbf{V} \in \mathbb{R}^{d \times d}$, welche eine positiv semidefinite Matrix $\mathbf{V}$ in eine Diagonalmatrix ihrer Eigenwerte überführt.

Der *MITTELWERT* von einer Stichprobe $\mathbf{Y}_k$ sei mit $\overline{\mathbf{Y}}_\cdot$ bezeichnet und die *QUADRATISCHE FORM* davon mit $Q_n = n \cdot \overline{\mathbf{Y}}_\cdot'\overline{\mathbf{Y}}_\cdot$.

KONVERGENZ IN VERTEILUNG wird mit $\mathbf{X}_n \xrightarrow{w} \mathbf{X}$,
KONVERGENZ IN WAHRSCHEINLICHKEIT mit $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$,
$\mathcal{L}_2 - KONVERGENZ$ mit $\mathbf{X}_n \xrightarrow{\mathcal{L}_2} \mathbf{X}$
und *FAST SICHERE KONVERGENZ* mit $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$ geschrieben.
ASYMPTOTISCHE ÄQUIVALENZ (in Wahrscheinlichkeit) wird mit $\mathbf{X}_n \doteqdot \mathbf{Y}_n$ be-

zeichnet und die *ASYMPTOTISCHE VERTEILUNG* einer Zufallsgröße mit $\mathbf{X}_n \rightsquigarrow F$.

Desweiteren wird die übliche *$\mathcal{O}$-NOTATION* mit den Landau-Symbolen verwendet. Es wird strikt zwischen der Asymptotik über den Stichprobenumfang $\mathcal{O}(f(n))$ und der Asymptotik über die Dimension $\mathcal{O}(g(d))$ unterschieden, um Inkonsistenzen zu vermeiden. Entsprechende Rechenregeln sind im Anhang in A.1 zu finden.

3.2 Modellierung

Gegeben ist das folgende Design: Seien

$$\mathbf{X}_k = (X_{1k},...,X_{dk})' \sim F \text{ mit } k = 1,..,n \tag{3.1}$$

unabhängige identisch verteilte Zufallsvektoren, wobei d für die verbundenen Messwiederholungen und n für den Stichprobenumfang steht. Desweiteren bezeichnet der Vektor $\boldsymbol{\mu} \in \mathbb{R}^d$ den Erwartungswert

$$E(\mathbf{X}_k) = \boldsymbol{\mu} \tag{3.2}$$

über den man im Folgenden Hypothesen formuliert und Testverfahren entwickelt, um diese zu überprüfen.
Die Messungen $X_{1k},...,X_{dk}$ können zunächst eine beliebige Abhängigkeit besitzen, allerdings gestaltet sich die Konstruktion eines Verfahrens, welches unabhängig von der Verteilung der $\mathbf{X}_k$ ist, als äußerst schwierig. Deshalb ist es nötig Annahmen an die multivariate Verteilung und somit die Abhängigkeit der $X_{1k},...,X_{dk}$ zu stellen.

3.2.1 Multivariate Normalverteilung

Definition 3.2.1 (Multivariate Normalverteilung) *Ein Vektor $\mathbf{X}_k$ heißt multivariat normalverteilt falls folgende Darstellung existiert:*

$$\mathbf{X}_k = \boldsymbol{\Upsilon} \cdot \mathbf{Z} + \boldsymbol{\mu}$$

wobei $\mathbf{Z} = (Z_1,...,Z_d)'$, mit Z_i unabhängig $\mathcal{N}(0,1)$ verteilt und $\boldsymbol{\Upsilon}$ eine d x d-Matrix ist.

Im übrigen ist die multivariate Normalverteilung, notiert als $\mathcal{N}(\boldsymbol{\mu},\mathbf{S})$, schon durch den Erwartungswert $\boldsymbol{\mu}$ und die Kovarianzmatrix $\mathbf{S} = Cov(\mathbf{X}_k) = \boldsymbol{\Upsilon} \cdot \boldsymbol{\Upsilon}'$ vollends bestimmt, wodurch die Abhängigkeiten sehr einach zu kontrollieren sind.

3.2.2 Bai-Saranadasa Modell

Eine Abschwächung der Annahmen an die Verteilung ohne die starken Einschränkungen an die Randverteilungen bietet die folgende von Bai-Saranadasa (1996) [1] eingeführte Modellierung. Diese ist sehr ähnlich der multivariaten Normalverteilung mit einer erzeugenden Matrix $\boldsymbol{\Upsilon}$, welche unter anderem die Kovarianz bestimmt, und einem erzeugenden Vektor $\mathbf{Z}$ mit unabhängigen Komponenten, aber nur mit geringen Einschränkungen an die Randverteilungen.

Definition 3.2.2 (Bai-Saranadasa Modell) *Sei*

$$\mathbf{X}_k = \boldsymbol{\Upsilon} \cdot \mathbf{Z}_k + \boldsymbol{\mu} \quad \textit{für } k = 1, \ldots, n \tag{3.3}$$

wobei $\boldsymbol{\Upsilon}$ *eine* d x m*-Matrix* ($m \geq d$) *und* $\mathbf{Z}_k = (Z_{1k}, \; ..Z_{mk})'$ *unabhängig identisch verteilt mit* $E(\mathbf{Z}_k) = \mathbf{0}_m$, $Cov(\mathbf{Z}_k) = \mathbf{I}_m$ *und sämtlichen Komponenten* $\{Z_{ik}\}_{k=1,\ldots,m}^{i=1,\ldots,d}$ *unabhängig.*

Die einzigen Einschränkungen an die Randverteilungen der Z_{ik} sind endliche vierte Momente, welche auch in für $d \to \infty$ gleichmässig beschränkt sein sollen.

$$E(Z_{ik}^4) \leq \mu_4 < \infty \tag{3.4}$$

Ein Vorteil dieser recht allgemeinen multivariaten Modellierung, die eine weite Klasse von Verteilungen jenseits der Normalverteilung zulässt, ist eine kontrollierbare Abhängigkeit der Komponenten. In der aktuellen Forschung hat sich dies als Minimalanforderung für parametrische Testverfahren für verbundene Stichproben etabliert. So wird das Bai-Saranadasa Modell neben Bai-Saranadasa (1996) [1] in anderen Arbeiten von Srivastava [19], Chen und Qin [7] und Helms [13] verwendet. Wang und Akritas (2010) [21] hingegen schränken die Abhängigkeit der Komponenten durch eine Alfa Mixing Bedingung ein, welche in dieser Arbeit aber keine Anwendung findet.

3.3 Hypothesen

Getestet werden soll im Ein-Gruppendesign nun die folgende parametrische Hypothese:

$$H_0 : \mathbf{H}\,\boldsymbol{\mu} = \mathbf{0}_d \qquad \text{vs.} \qquad H_1 : \mathbf{H}\,\boldsymbol{\mu} \neq \mathbf{0}_d \tag{3.5}$$

für eine beliebige Hypothesenmatrix $\mathbf{H}$.

In der Praxis interessiert man sich meist für spezielle Hypothesenmatrizen. So wird die Fragestellung nach einem Zeiteffekt (T) über verbundene Messungen in der HD-F1-Situation wie folgt formuliert:

$$H_0 : \ \mu_1 = ... = \mu_d \quad \Leftrightarrow \quad \mathbf{P}_d\,\boldsymbol{\mu} = \mathbf{0}_d \tag{3.6}$$

Im HD-F2 hat man zusätzlich noch einen Behandlungseffekt B mit b Faktorstufen, so dass man inklusive der Wechselwirkung zwischen dem Zeiteffekt und dem Behandlungseffekt drei Hypothesen testet:

$$H_0(T) \qquad : \overline{\mu}_{\cdot 1} = ... = \overline{\mu}_{\cdot d} \quad \Leftrightarrow \quad (\frac{1}{b}\mathbf{1}_b' \otimes \mathbf{P}_d)\boldsymbol{\mu} = \mathbf{0}_d \tag{3.7}$$

$$H_0(B) \qquad : \overline{\mu}_{1\cdot} = ... = \overline{\mu}_{b\cdot} \quad \Leftrightarrow \quad (\mathbf{P}_b \otimes \frac{1}{d}\mathbf{1}_d')\boldsymbol{\mu} = \mathbf{0}_b \tag{3.8}$$

$$H_0(T * B) \qquad : \mu_{rs} + \overline{\mu}_{\cdot\cdot} = \overline{\mu}_{r\cdot} + \overline{\mu}_{\cdot s}, \quad r = 1, \ldots, b, \ s = 1, \ldots, d \tag{3.9}$$

$$\Leftrightarrow \quad (\mathbf{P}_b \otimes \mathbf{P}_d)\boldsymbol{\mu} = \mathbf{0}_{b \cdot d} \tag{3.10}$$

Zu allgemein formulierten Hypothesen existieren häufig verschiedene Hypothesenmatrizen $\mathbf{H}_1$, $\mathbf{H}_2$, so dass jeweils $\mathbf{H}_i\boldsymbol{\mu} = \mathbf{0}_d$ ist. Diese Hypothesenmatrizen werden dann als äquivalent bezeichnet, wenn ihre Aussagen an die Beziehungen der μ_i dieselbe ist. Für praktische Verfahren ist es daher wünschenswert, wenn diese unabhängig von der expliziten Wahl der Hypothesenmatrix sind. Deshalb bietet es sich an, anstelle einer gegebenen Hypothesenmatrix $\mathbf{H}$ die dazu äquivalente eindeutige Hypothesenmatrix

$$\mathbf{T} := \mathbf{H}'(\mathbf{H}\mathbf{H}')^{+}\mathbf{H} \tag{3.11}$$

zu verwenden. Die Äquivalenz der Hypothesen ist unter anderem in Satz 1.2 auf Seite 27 im Vorlesungsskript Statistik II von Prof. Brunner ausführlich beschrieben.

3.4 Definitionen

Der Übersichtlichkeit halber wird die Hypothesenmatrix $\mathbf{H}$ und die Modellierung bei der Notation zusammengefasst. Sei nun

$$\mathbf{Y}_k = \mathbf{H} \cdot \mathbf{X}_k, \quad k = 1,..,n \tag{3.12}$$

Die Modellierung von Bai-Saranadasa aus Definition 3.2.2 der $\mathbf{X}_k$ überträgt sich nun einfach auf die $\mathbf{Y}_k$. Normalverteilte Daten werden als Spezialfall des Bai-Saranadasa Modells betrachtet. Somit lässt sich $\mathbf{Y}_k$ darstellen als

$$\mathbf{Y}_k = \mathbf{H} \cdot (\boldsymbol{\Upsilon}\, \mathbf{Z} + \boldsymbol{\mu}) \tag{3.13}$$

$$= \mathbf{H}\boldsymbol{\Upsilon}\, \mathbf{Z} + \mathbf{T}\boldsymbol{\mu} \tag{3.14}$$

$$= \boldsymbol{\Gamma}\mathbf{Z} + \boldsymbol{\theta} \tag{3.15}$$

mit $\boldsymbol{\Gamma} := \mathbf{H} \cdot \boldsymbol{\Upsilon}$ und $\boldsymbol{\theta} := \mathbf{H} \cdot \boldsymbol{\mu}$. Die Hypothese H_0 formuliert sich dann einfach als $E(\mathbf{Y}_k) = \boldsymbol{\theta} = \mathbf{0}_d$.

Die Kovarianzmatrix $\boldsymbol{\Sigma}$ und die duale Kovarianzmatrix $\mathbf{V}$ von $\mathbf{Y}_k$ ergeben sich wie folgt:

$$\boldsymbol{\Sigma} := Cov(\mathbf{Y}_k) = \boldsymbol{\Gamma} \cdot Cov(\mathbf{Z}_k) \cdot \boldsymbol{\Gamma}' \tag{3.16}$$

$$= \boldsymbol{\Gamma}\boldsymbol{\Gamma}' \tag{3.17}$$

$$\mathbf{V} := \boldsymbol{\Gamma}'\boldsymbol{\Gamma} \tag{3.18}$$

Die Kovarianzmatrix $\mathbf{\Sigma}$ und die duale Kovarianzmatrix $\mathbf{V}$ haben ähnliche Eigenschaften. So gilt beispielsweise für die Spuren sämtlicher ganzzahliger Potenzen wegen der Invarianz der Spur unter zyklischen Vertauschungen:

$$Sp(\mathbf{\Sigma}^{\delta}) = Sp((\mathbf{\Gamma} \cdot \mathbf{\Gamma}')^{\delta}) = Sp(\mathbf{\Gamma} \cdot \mathbf{\Gamma}' \cdot \mathbf{\Gamma} \cdots \mathbf{\Gamma}' \cdot \mathbf{\Gamma} \cdot \mathbf{\Gamma}') \tag{3.19}$$

$$= Sp(\mathbf{\Gamma}' \cdot \mathbf{\Gamma} \cdot \mathbf{\Gamma}' \cdot \mathbf{\Gamma} \cdots \mathbf{\Gamma}' \cdot \mathbf{\Gamma}) = Sp((\mathbf{\Gamma}' \cdot \mathbf{\Gamma})^{\delta}) \tag{3.20}$$

$$= Sp(\mathbf{V}^{\delta}) \tag{3.21}$$

Die duale Kovarianzmatrix $\mathbf{V} \in \mathbb{R}^{m \times m}$ ist unter anderem beim Berechnen von Quadrat- bzw. Bilinearformen der $\mathbf{Y}_k$ von Bedeutung. So seien die

$$A_{kl} := \mathbf{Y}_k' \cdot \mathbf{Y}_l \quad \text{mit } k,l = 1,...,n \tag{3.22}$$

$$= \mathbf{Z}_k' \cdot \mathbf{\Gamma}' \cdot \mathbf{\Gamma} \cdot \mathbf{Z}_l' \tag{3.23}$$

$$= \mathbf{Z}_k' \cdot \mathbf{V} \cdot \mathbf{Z}_l' \tag{3.24}$$

für $k \neq l$ die Bilinearformen der einzelnen Beobachtungsvektoren und analog für $k = l$ die Quadratformen. Darüber hinaus sei die Quadratform der $\sqrt{n}$ -fachen Mittelwerte

$$Q_n \quad := n \cdot \overline{\mathbf{Y}}_{\cdot}' \, \overline{\mathbf{Y}}_{\cdot} \tag{3.25}$$

4 Ergebnisse bisheriger Forschungsarbeiten

Es gibt eine weitreichende Auswahl an Lösungsansätzen für das Ein-Stichprobenpro blem. So lässt sich der T^2-Test von Hotelling (1931) für ein Ein-Gruppen-Repeated-Measures-Design anwenden. Dieser beruht auf der Schätzung der empirischen Kovarianzmatrix bzw. derer Inverse. Die Schätzung erweist sich in hochdimensionalen Designs allerdings als problematisch, da eine Berechnung nur bei sehr großen Stichprobenumfängen möglich ist. Deswegen wurde mit der ANOVA-Typ-Statistik, eingeführt in Arbeiten von Box (1954) [4] [5] ein anderer Ansatz verfogt. Bei diesem reicht die Schätzung der Spur der Kovarianzmatrix und ihrer Potenzen aus, um eine brauchbare Approximation der Quadratischen Form Q_n über ihre Momente zu erhalten. Die ursprünglich gewählten Plug-in-Schätzer $Sp(\widehat{\mathbf{\Sigma}})$ und $Sp(\widehat{\mathbf{\Sigma}}^2)$ besitzen allerdings eine Verzerrung, welche für größer werdende Dimension zunimmt. Schätzer, die stabil bezüglich der Dimension sind, konnten 2004 von Werner im Rahmen ihrer Diplomarbeit [22] entwickelt werden.
Diese ANOVA-Typ-Statistik wurde im Folgenden mehrfach weiterentwickelt. Unter anderem sind hier Arbeiten von Becker und Helms zu erwähnen.
Demgegenüber konnten Bai-Sarandasa (1996) einen neuen Test entwickeln. Auf dessen theoretischer Grundlage basiert eine große Anzahl weiterer verbesserter Testverfahren, wie das von Chen und Qin und der multivariate Test von Srivastava. Somit sind die folgenden Testverfahren aufzuführen.

4.1 ANOVA-Typ-Statistik nach Box

Seien die $\mathbf{Y}_k \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{\Sigma})$ und unter Hypothese H_0 sei $\boldsymbol{\theta} = \mathbf{H}\boldsymbol{\mu} = \mathbf{0}_d$.

Die empirische Kovarianzmatrix ist definiert als

$$\widehat{\mathbf{\Sigma}} = \frac{1}{(n-1)} \sum_{k=1}^{n} (\mathbf{Y}_k - \overline{\mathbf{Y}}_\cdot)(\mathbf{Y}_k - \overline{\mathbf{Y}}_\cdot)' \tag{4.1}$$

Beruhend auf der Box-Approximation [4] [5], lässt sich nun die ANOVA-Typ-Statistik

mittels der Plug-in-Schätzer aus $\widehat{\mathbf{\Sigma}}$ formulieren als

$$\frac{Q_n}{Sp(\widehat{\mathbf{\Sigma}})} \;\dot{\sim}\; \frac{\chi^2_{\widehat{f}}}{\widehat{f}}, \qquad \widehat{f} = \frac{Sp^2(\widehat{\mathbf{\Sigma}})}{Sp(\widehat{\mathbf{\Sigma}}^2)} \tag{4.2}$$

Diese Statistik ist in zahlreichen Statistik-Software-Paketen implementiert und diente in der Vergangenheit als Standard für Auswertungen. Sie wird daher als klassische ANOVA-Typ-Statistik bezeichnet. Schwächen dieser Statistik besonders für hohe Dimensionen wurden von Werner (2004) [22] untersucht.

4.2 Werner-Brunner-Statistik

Analog zur ATS nach Box sind die $\mathbf{Y}_k \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{\Sigma})$. Dann gilt unter $H_0: \; \boldsymbol{\theta} = \mathbf{H}\boldsymbol{\mu} = \mathbf{0}_d$:

$$\frac{Q_n}{B_0} \;\dot{\sim}\; \frac{\chi^2_f}{f} \tag{4.3}$$

$$\text{wobei} \qquad f = \frac{Sp^2\mathbf{\Sigma}}{Sp(\mathbf{\Sigma}^2)} \qquad \text{durch} \qquad \hat{f} = \frac{n}{n-1} \cdot \frac{B_1}{B_2} \tag{4.4}$$

geschätzt wird. Die Schätzer B_i seien definiert als

$$B_0 = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{Y}'_k \boldsymbol{Y}_k \qquad \text{für } Sp(\mathbf{\Sigma}) \tag{4.5}$$

$$B_1 = \frac{1}{n} \sum_{k \neq l}^{n} \boldsymbol{Y}'_k \boldsymbol{Y}_k \, \boldsymbol{Y}'_l \boldsymbol{Y}_l \qquad \text{für } Sp^2(\mathbf{\Sigma}) \tag{4.6}$$

$$B_2 = \frac{1}{n(n-1)} \sum_{k \neq l}^{n} (\boldsymbol{Y}'_k \boldsymbol{Y}_l)^2 \qquad \text{für } Sp(\mathbf{\Sigma}^2) \tag{4.7}$$

Der Korrekturfaktor $n/(n-1)$ wird mit einer Taylorapproximation zweiter Ordnung des Quotienten von B_1 und B_2 begründet. Die Konsistenz der Schätzer wird in Kapitel 6 gezeigt.

4.3 Geisser-Greenhouse-Statistik nach Becker

Mit den Schätzern B_1 und B_2 aus 4.2 ergibt sich unter den gleichen Voraussetzungen wie für den Test von Werner und Brunner die folgende Teststatistik nach Becker (2010) [2]. Sie basiert auf Arbeiten von Geisser-Greenhouse aus dem Jahr 1958 und konnte mittels neuer Schätzer so verbessert werden, dass der Test stabil bezüglich der Dimension ist.

$$\frac{Q_n}{Sp(\widehat{\mathbf{\Sigma}})} \stackrel{\cdot}{\sim} F\,(f,(n-1)f) \tag{4.8}$$

wobei F die Verteilungsfunktion der Fischerverteilung ist und der Freiheitsgrad f folgendermaßen geschätzt wird:

$$\widehat{f} = \frac{B_1}{B_2\left(1+\frac{1}{4n(n-1)}\right)} \tag{4.9}$$

4.4 Test von Bai-Saranadasa im Ein-Gruppen-Design

Als bedeutendes Ergebnis unter Nicht-Normalverteilung konnten Bai-Saranadasa (1996) [1] einen Test entwickeln, welcher als Verteilungsannahme lediglich die von ihnen angegebene Modellierung aus Definition 3.2.2 benötigt. Die für den 2-Stichprobenfall mit gleichen Kovarianzmatrizen konstruierte Teststatistik lässt sich auf den Ein-Stichprobenfall übertragen. Als Testgröße für zwei Stichproben $\{\mathbf{X}_{1k} : k = 1,...,n_1\}$, $\{\mathbf{X}_{2k} : k = 1,...,n_2\}$ mit gepoolter empirischer Kovarianzmatrix $\mathbf{S}_n$ wird

$$M_n = (\overline{\mathbf{X}}_{1\cdot} - \overline{\mathbf{X}}_{2\cdot})'(\overline{\mathbf{X}}_{1\cdot} - \overline{\mathbf{X}}_{2\cdot}) - \tfrac{n_1+n_2}{n_1\cdot n_2}\mathbf{S}_n \tag{4.10}$$

betrachtet. So lässt sich mit der Überlegung, einen Vektor $\mathbf{X}_k$ im Ein-Gruppen-Design als Differenz $\mathbf{X}_{1k} - \mathbf{X}_{2k}$ aufzufassen, eine Reduktion des Tests für eine Stichprobe umsetzen. Dabei sollen $\mathbf{X}_{1k}$ aus der ersten Stichprobe und $\mathbf{X}_{2k}$ aus der zweiten Stichprobe die gleiche Kovarianzmatrix $\frac{1}{2}\mathbf{\Sigma}$ haben. Da die Schätzung der Kovarianzmatrix $\mathbf{S}_n$ in [1] mit einem gepoolten Schätzer stattfindet, muss dieser für eine Stichprobe abgewandelt werden. Dies soll auf intuitivem Wege mit der empirischen Kovarianzmatrix erfolgen. Damit ergibt sich nun, falls die $\mathbf{X}_k$, welche nach dem Bai-Saranadasa Modell verteilt sind,

$$\lambda_{MAX} = \mathbf{o}(Sp(\mathbf{\Sigma}^2)) \tag{4.11}$$

und

$$d/n \to y > 0 \tag{4.12}$$

ein Test für $\mathbf{H}\boldsymbol{\mu} = \mathbf{0}_d$:

$$\frac{Q_n - Sp(\widehat{\boldsymbol{\Sigma}})}{\sqrt{2\frac{n}{n-1}B_n^2}} \sim \mathcal{N}(0,1) \tag{4.13}$$

Dabei sei

$$B_n^2 = \frac{(n-1)^2}{(n+1)(n-2)}\left(Sp(\widehat{\boldsymbol{\Sigma}}^2) - \frac{1}{n}Sp^2(\widehat{\boldsymbol{\Sigma}})\right) \tag{4.14}$$

ein Schätzer für $Sp(\boldsymbol{\Sigma}^2)$, welcher ebenfalls auf trivialem Wege auf den Ein-Stichprobenfall abgewandelt wurde. Dieser ist unter den genannten Voraussetzungen stabil bezüglich der Dimension.

4.5 ANOVA-Typ Statistik nach Helms

Die ATS nach Werner-Brunner unter Normalverteilung konnte Helms in seiner Diplomarbeit (2010) [13] auf Verteilungen ähnlich dem Bai-Saranadasa Modell erweitern. Unter der Annahme, dass $d^2/n \to 0$ lässt sich folgende neue ANOVA-Typ Statistik angeben

$$\frac{Q_n}{Sp(\widehat{\mathbf{V}}) \cdot \widehat{g}_1} \overset{\cdot}{\sim} \frac{\chi^2_{\widehat{f}_1}}{\widehat{f}_1} \tag{4.15}$$

mit

$$\widehat{f}_1 = \frac{((n-1)B_1 + 2B_2)^2}{B_1 \cdot B_2 \cdot n(n-1)} \tag{4.16}$$

$$\widehat{g}_1 = 1 + \frac{2B_2}{(n-1)B_1} \tag{4.17}$$

4.6 Chen-Qin-Test für eine Stichprobe

Diese Resultate von Bai-Saranadasa wurden von Chen und Qin 2010 in der Arbeit „A two-sample test for high-dimensional data with applications to gene-set testing“[7] aufgegriffen. Dabei wird neben dem Zwei-Gruppen-Fall ebenfalls ein Test der Hypothese $H_0 : \mathbf{H}\,\boldsymbol{\mu} = \mathbf{0}$ gegen $H_1 : \mathbf{H}\,\boldsymbol{\mu} \neq \mathbf{0}$ für das Ein-Gruppen-Design entwickelt. Für die Statistik

$$F_n = \frac{1}{n(n-1)} \sum_{k \neq l}^{n} \mathbf{Y}_k' \, \mathbf{Y}_l \tag{4.18}$$

lassen sich nun für große Dimension d die Ergebnisse des zentralen Grenzwertsatzes für Martingale aus Korrolar 3.1 aus Hall und Heyde (1980) [12] anwenden. Diese Asymptotik über d benötigt zusätzlich die Regularitätsannahme an die Kovarianzmatrix

$$\frac{Sp(\boldsymbol{\Sigma}^4)}{Sp^2(\boldsymbol{\Sigma}^2)} \xrightarrow{d \to \infty} 0 \tag{4.19}$$

welche im Folgenden Chen-Qin-Bedingung genannt werden soll. Unter dieser Bedingung und der Verteilungsannahme des Bai-Saranadasa Modells aus Definition 3.2.2 gilt nun nach Theorem 1 aus Chen Qin (2010) [7] folgende Asymptotische Normalität von F_n:

$$\frac{F_n - \|\boldsymbol{\mu}\|^2}{\sqrt{Var(F_n)}} \to \mathcal{N}(0,1) \tag{4.20}$$

für $n \to \infty$ und $d \to \infty$. Dabei wird keine Restriktion zwischen n und d angenommen. Für die Varianz von F_n gilt nach Abschnitt 6.1 [7]:

$$Var(F_n) = \frac{2}{n(n-1)} \cdot Sp(\boldsymbol{\Sigma}^2) \tag{4.21}$$

Für Parameter $Sp(\boldsymbol{\Sigma}^2)$ wird folgender für den Zwei-Stichprobenfall entwickelter Schätzer angegeben, welcher sich auf den Ein-Stichprobenfall überträgt

$$\widehat{Sp(\boldsymbol{\Sigma}^2)} = \frac{1}{n(n-1)} Sp\left(\sum_{k \neq l}^{n} (\mathbf{Y}_k - \overline{\mathbf{Y}}_{(kl)}) \mathbf{Y}_k' (\mathbf{Y}_l - \overline{\mathbf{Y}}_{(kl)}) \mathbf{Y}_l' \right) \tag{4.22}$$

wobei $\overline{\mathbf{Y}}_{(kl)}$ der Mittelwert ohne die k-te und l-te Stichprobe ist.

4.7 Multivariater Test von Srivastava

Ebenfalls unter der Annahme des Bai-Saranadasa Modells konnte Srivastava (2009) [19] einen multivariaten Test entwickeln, welcher invariant unter Skalierungen ist. Sei $\mathbf{D_\Sigma} = diag(\sigma_{11}^2, \sigma_{22}^2,...,\sigma_{dd}^2)$, wobei σ_{ii}^2 die Varianz der i-ten Komponente ist. Der Schätzer für $\mathbf{D_\Sigma}$ sei analog mit $\mathbf{D_S} = diag(\widehat{\sigma}_{11}^2, \widehat{\sigma}_{22}^2,...,\widehat{\sigma}_{dd}^2)$ bezeichnet und basiert auf dem gewöhnlichen empirischen Varianzschätzer. Desweiteren sei

$$\mathbf{R} = \mathbf{D}_{\mathbf{\Sigma}}^{-\frac{1}{2}} \mathbf{\Sigma} \mathbf{D}_{\mathbf{\Sigma}}^{-\frac{1}{2}}$$

die Korrelationsmatrix bezüglich $\mathbf{\Sigma}$ und

$$\widehat{\mathbf{R}} = \mathbf{D}_S^{-\frac{1}{2}} \widehat{\mathbf{\Sigma}} \mathbf{D}_S^{-\frac{1}{2}}$$

der entsprechende Pluginschätzer. Unter den Voraussetzungen des Bai-Saranadasa Modells und den Bedingungen

$$n = \mathcal{O}(d^\varepsilon), \;\; 0 < \varepsilon \leq 1 \tag{4.23}$$

und

$$\lim_{d \to \infty} \frac{Sp\left(\mathbf{R}^r\right)}{d} < \infty, \quad r = 1,..,4 \tag{4.24}$$

wird ein multivariater Test angegeben. Bei einem multivariaten Test wird im Gegensatz zu repeated-measures-Designs die Hypothese über die Einheitsmatrix formuliert als $H_0 : \boldsymbol{\mu} = \mathbf{0}_d$ gegen $H_1 : \boldsymbol{\mu} \neq \mathbf{0}_d$. Nach Theorem 3.1 ergibt sich als Teststatistik

$$\frac{n\overline{\mathbf{X}}' \mathbf{D}_S^{-1} \overline{\mathbf{X}} - \frac{(n-1)\cdot d}{n-3}}{\sqrt{2\left(Sp(\widehat{\mathbf{R}}^2) - \frac{d^2}{n-1}\right)}} \to \mathcal{N}(0,1) \tag{4.25}$$

Dabei ist $\frac{1}{d} Sp(\widehat{\mathbf{R}}^2) - \frac{d^2}{n}$ ein konsistenter, dimensionsstabiler Schätzer für $\frac{1}{d} Sp(\mathbf{R}^2)$.

5 Verbesserte Approximation der Chen-Qin-Statistik

Ziel der Arbeit ist es nun, unter Nicht-Normalverteilung geeignete robuste Verfahren für repeated measures zu erzielen. Unter der Verteilungsannahme des Bai-Saranadasa Modells aus Definition 3.2.2 als Erweiterung der Normalverteilung konnte Helms 2010 [13] ein Verfahren vorstellen, welches auf der klassischen Theorie der ANOVA-Typ-Statistik basiert. Dieses Verfahren ist für niedrige Dimensionen geeignet, während man für Designs mit großer Dimension die restriktive Annahme, dass der Stichprobenumfang n in der Größenordnung d^2 steigen muss, benötigt. Letztere Annahme ist nötig, da das Verfahren auf dem Multivariaten Zentralen Grenzwertsatz beruht, dessen Güte mit steigender Dimension schlechter wird.
Demgegenüber konnten Chen-Qin, basierend auf Ideen von Bai-Saranadasa, den Test aus 4.6 angeben. Neben der Asymptotik über den Stichprobenumfang n benötigt dieser zusätzlich die Restriktionen $d \to \infty$ und $Sp(\mathbf{\Sigma}^4)/Sp^2(\mathbf{\Sigma}^2) \to 0$, welche allerdings unabhängig von n sind. Sind diese Voraussetzungen bezüglich der Dimension nicht erfüllt, werden in niedrig-dimensionalen Designs teils liberale Ergebnisse erzielt (Siehe Abbildung 6.2).

5.1 Motivation

Für die praktische Anwendung werden Verfahren benötigt, deren Güte nicht von der tatsächlichen Dimension der Daten abhängen. Die Vorteile gegenüber einer Anwendung von verschiedenen Verfahren in Abhängigkeit der Dimension liegen auf der Hand. Zum einen entfällt die Klassifikation in niedrig und hochdimensionale Designs, welche äußerst subjektiv ist, wie in dem Besipiel mit den 28 Messungen geschildert. Zum anderen können sämtliche Hypothesen mit dem gleichen Verfahren getesten werden. Dies ist von Bedeutung um Inkonsistenzen in der Auswertung zu vermeiden.

5.2 Approximation für niedrige Dimension

Als Ausgangspunkt für Testverfahren, die in dieser Arbeit vorgestellt werden, soll die Teststatistik von Chen-Qin dienen, welche in (4.18) beschrieben wurde.
Unter Hypothese $\mathbf{H}\boldsymbol{\mu} = 0$ galt für

$$F_n = \frac{1}{n(n-1)} \sum_{k \neq l}^{n} \mathbf{Y}_k' \, \mathbf{Y}_l \tag{5.1}$$

und die zugehörige Teststatistik R_n

$$R_n := \frac{F_n}{\sqrt{\frac{2}{n(n-1)} Sp(\boldsymbol{\Sigma}^2)}} \overset{n,d \to \infty}{\rightsquigarrow} \mathcal{N}(0,1) \tag{5.2}$$

Mit der geschilderten Motivation soll nun eine bessere Approximation dieser Statistik gefunden werden, welche die Standardnormalapproximation ersetzt. Um die asymptotische Verteilung von R_n im Niedrigdimensionalen zu bestimmen, bedient man sich des Multivariaten Zentralen Grenzwertsatzes und der klassischen Theorie der ANOVA-Typ-Statistik.

5.2.1 Multivariater Zentraler Grenzwertsatz

Satz 5.2.1 (Multivariater Zentraler Grenzwertsatz)
Seien $\mathbf{Y}_i \sim F$, $i = 1,..,d$ unabhängig identisch verteilte Zufallsvektoren im $\mathbb{R}^d$ mit Erwartungswert $\boldsymbol{\mu}$, Kovarianzmatrix $\boldsymbol{\Sigma}$. Dann gilt

$$\sqrt{n}(\overline{\mathbf{Y}}_n - \boldsymbol{\mu}) \overset{n \to \infty}{\rightsquigarrow} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{5.3}$$

Beweis: Siehe Satz 2.18 in van der Vaart (1998) [20].

Der Satz erzielt ein asymptotischen Resultat einzig über den Stichprobenumfang n. Betrachtet man nun die Güte der Approximation, dann hängt diese von weiteren Parametern, wie der Dimension d und den vierten Momenten der Randverteilungen μ_{4i} ab. Letztere können weiterhin als fest ansehen werden, während die Dimension bei der Beurteilung der Güte des ZGWS berücksichtigt werden muss. Hier ist die Annahme der niedrigen Dimension also eine entscheidende Voraussetzung für die asymptotische Normalität der Mittelwertsvektoren.
In hochdimensionalen Designs kann die Dimension nicht länger als fest betrachtet

werden. Statt dessen sind Grenzwerte für $d \to \infty$ zu betrachten. Aussagen über die Güte einer Asymptotik über d und n gemeinsam wurden von Portnoy (1986) [16] erarbeitet. So gilt das Resultat des Satzes weiterhin, falls

$$\frac{d^2}{n} \to 0. \tag{5.4}$$

für $n, d \to \infty$. Gilt diese Bedingung nicht, so kann ein Beispiel angegeben werden, bei dem die Asymptotik fehlschlägt.

Ist nun aber $n \gg d^2$, so befindet man sich asymptotisch im Normalverteilungsfall und man kann auf Resultate der klassischen ANOVA-Typ-Statistik zurückgreifen.

5.2.2 Verteilung von Q_n

Satz 5.2.2 (Verteilung einer quadratischen Form unter Normalverteilung) *Seien $\mathbf{Y} = (Y_1,...,Y_d)' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ mit $\mathbf{\Sigma}$ symmetrisch, positiv semi-definit mit Eigenwerten $\lambda_i,\ i = 1,...,d$. Dann gilt für die quadratische Form*

$$Q = \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{d} \lambda_i C_i\ , \quad \textit{wobei } C_i \sim \chi_1^2 \ \textit{ unabhängig für } i = 1,...,d \tag{5.5}$$

Beweis: Siehe (S.28f) in Mathai, Provost (1992) [14].

Da die Eigenwerte in der Regel unbekannt sind und sich eine Schätzung in hochdimensionalen Designs als problematisch erweist, betrachtet man die Momente der Quadratischen Formen. Diese lassen sich stabiler schätzen.

Satz 5.2.3 (Momente einer quadratischen Form unter Normalverteilung) *Seien $Q = \sum_{i=1}^{d} \lambda_i C_i$, mit $C_i \sim \chi_1^2$ unabhängig für $i = 1,...,d$. Dann gilt für die Momente und Schiefe ν von Q:*

1. $E(Q) = \sum_{i=1}^{d} \lambda_i = Sp(\mathbf{\Sigma})$
2. $Var(Q) = 2\sum_{i=1}^{d} \lambda_i^2 = 2Sp(\mathbf{\Sigma}^2)$
3. $\nu(Q) = \frac{2\sqrt{2}\sum \lambda_i^3}{(\sum \lambda_i^2)^{3/2}}$

Beweis:

1. $E(Q) \quad = \sum_{i=1}^{d} \lambda_i \underbrace{E(C_i)}_{=1} = \sum_{i=1}^{d} \lambda_i$

2. $Var(Q) \quad = \sum_{i=1}^{d} Var(\lambda_i \cdot C_i) = \sum_{i=1}^{d} \lambda_i^2 \cdot Var(C_i)$

$= 2 \sum_{i=1}^{d} \lambda_i^2$, da die Varianz einer χ_1^2-verteilten Zufallsvariablen 2 ist.

3. $\nu(Q) \quad = \frac{E\left((Q-E(Q))^3\right)}{Var^{3/2}(Q)} = Var^{-3/2}(Q) \cdot E\left((\sum_{i=1}^{d} \lambda_i (C_i - 1))^3 \right)$

$= Var^{-3/2}(Q) \cdot \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{h=1}^{d} \lambda_i \lambda_j \lambda_h E((C_i - 1)(C_j - 1)(C_h - 1))$

Falls nun ein Index von den anderen beiden verschieden ist, ist die zugehörige Zufallsvariable von den beiden anderen unabhängig. Dann ist der Erwartungwert des Ausdrucks linear in den Erwartungswerten der unabhängigen Variablen und somit 0

$= Var^{-3/2}(Q) \cdot \sum_{i=1}^{d} \lambda_i^3 E((C_i - 1)^3)$

$= \frac{8}{2\sqrt{2}} \cdot \left(\sum_{i=1}^{d} \lambda_i^2 \right)^{-3/2} \sum_{i=1}^{d} \lambda_i^3$, da das 3. zentr. Moment der χ_1^2-Verteil. 8 ist.

□

5.2.3 Boxapproximation

Die Boxapproximation ist nun eine in der Literatur weit verbreitete Lösung, die Verteilung der Quadratischen Form aus (5.2.2) über ihr Momente zu approximieren. Sie beruht auf den Arbeiten von Box aus dem Jahr 1954 [4] [5]. Die Idee dabei ist die Summe der gewichteten χ_1^2-verteilten Zufallsvariablen durch eine um einen Faktor g gestreckte χ_f^2-Verteilung zu approximieren. Die adjustierbaren Parametern g und f sind so zu wählen, dass die ersten beiden Momente überein stimmen.

Sei $Q = \sum_{i=1}^{d} \lambda_i C_i$ mit $C_i \sim \chi_1^2$ unabhängig, $i = 1,...,d$ und $R \sim g \cdot C_0$ mit $C_0 \sim \chi_f^2$.

Dann ergeben sich Erwartungswert und Varianz wie folgt:

$$E(Q) = \sum_{i=1}^{d} \lambda_i \cdot \underbrace{E(C_1)}_{=1} = Sp(\mathbf{\Sigma}) \qquad E(R) = g \cdot E(C_0) = g \cdot f$$

$$Var(Q) = \sum_{i=1}^{d} Var(\lambda_i C_1) = 2\sum_{i=1}^{d} \lambda_i^2 = 2Sp(\mathbf{\Sigma}^2) \qquad Var(R) = g^2 \cdot 2f$$

Gleichsetzen der ersten beiden Momente ergibt

$$\left| \begin{array}{c} g \cdot f = Sp(\mathbf{\Sigma}) \\ \\ 2 \cdot g^2 \cdot f = 2Sp(\mathbf{\Sigma}^2) \end{array} \right|$$

$$\Rightarrow \qquad g = \frac{Sp(\mathbf{\Sigma}^2)}{Sp(\mathbf{\Sigma})} \qquad \text{und} \qquad f = \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)}$$

Wird nun der Quotienten $Q/E(Q)$ betrachtet, dann ist dieser invariant unter den Skalierungsfaktor g. Somit ergibt sich nun als Approximation

$$\frac{Q}{Sp(\mathbf{\Sigma})} \dot{\sim} \frac{\chi_f^2}{f}, \qquad \text{wobei } f = \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \tag{5.6}$$

Mit der Boxapproximation lässt sich nun die Verteilung von $Q_n = \overline{\mathbf{Y}}_{\cdot}'\overline{\mathbf{Y}}_{\cdot}$, einzig als Funktionen der Spur der Kovarianzmatrix und ihrer Potenzen beschreiben. Für diese lassen sich geeignete Schätzer konstruieren.

5.2.4 Der Schätzer B_0

Lemma 5.2.4 (Momente der A_{kk} im Bai-Saranadasa Modell)
Die Zufallsvektoren $\mathbf{Y}_k$, $k = 1,..,n$ seien unabhängig verteilt wie in Lemma C.2.1 (Bai-Saranadasa Modell) mit $E(\mathbf{Y}_k) = \mathbf{0}_d$ und Kovarianzmatrix $\mathbf{\Sigma}$. Die $A_{kk} := \mathbf{Y}_k'\mathbf{Y}_k$ seien die zugehörigen Quadratformen. Dann gilt für den Erwartungswertes und die Varianz für $k = 1,..,n$:

1. $E(A_{kk}) = Sp(\mathbf{\Sigma})$
2. $E(A_{kk}^2) = Sp^2(\mathbf{\Sigma}) + 2Sp(\mathbf{\Sigma}^2) + \mathcal{O}(Sp(\mathbf{\Sigma}^2))$

Beweis: Siehe Lemma C.2.1 im Anhang.

Um die Spur $Sp(\mathbf{\Sigma})$ zu schätzen, wird folgender, von Werner (2004) [22] als B_0 definierter Schätzer eingeführt:

$$B_0 = \frac{1}{n}\sum_{k=1}^{n} A_{kk} \tag{5.7}$$

Dabei ist $A_{kk} = \mathbf{Y}_k'\mathbf{Y}_k$. Unter der Hypothese H_0 schätzt dieser $Sp(\mathbf{\Sigma})$ robust bezüglich der Dimension, was im folgenden Satz gezeigt wird.

Satz 5.2.5 (Konsistenz des B_0-Schätzers) *Sei B_0 definiert wie oben mit $A_{kk} = \mathbf{Y}_k'\mathbf{Y}_k$. Sei $\boldsymbol{\mu} = \mathbf{0}_d$ der Erwartungswert unter Hypothese und $\mathbf{\Sigma}$ die Kovarianzmatrix von $\mathbf{Y}_k$. Dann ist*

1. $$E(B_0) = Sp(\mathbf{\Sigma}) \tag{5.8}$$

2. $$\frac{B_0}{Sp(\mathbf{\Sigma})} \xrightarrow{p,n\to\infty} 1, \quad \textit{gleichmäßig in } d \tag{5.9}$$

3. $$\frac{B_0}{\sqrt{Sp(\mathbf{\Sigma}^2)}} \xrightarrow{p,n\to\infty} \frac{Sp(\mathbf{\Sigma})}{\sqrt{Sp(\mathbf{\Sigma}^2)}}, \quad \textit{gleichmäßig in } d \tag{5.10}$$

Beweis:

$$E(B_0) = E\left(\frac{1}{n}\sum_{k=1}^{n} A_{kk}\right) = E(A_{11}) = E(Sp(\mathbf{Y}_1'\mathbf{Y}_1)) = E(Sp(\mathbf{Y}_1\,\mathbf{Y}_1'))$$

$$= Sp(\mathbf{\Sigma})\,, \text{ aufgrund der Invarianz der Spur unter zyklischen Vertauschungen}$$

$$Var(B_0) = Var\left(\frac{1}{n}\sum_{k=1}^{n} A_{kk}\right) = \frac{1}{n^2}\sum_{k=1}^{n} Var(A_{kk}) = \frac{1}{n}Var(A_{11})$$

$$= \frac{1}{n}\cdot(2\cdot Sp(\mathbf{\Sigma}^2) + \mathcal{O}(Sp(\mathbf{\Sigma}^2))) \quad \text{nach Lemma (5.2.4).}$$

Dann folgt für den Quotienten (5.9):

$$Var\left(\frac{B_0}{Sp(\mathbf{\Sigma})}\right) = \frac{2}{n}\frac{\mathcal{O}(Sp(\mathbf{\Sigma}^2))}{Sp^2(\mathbf{\Sigma})}$$

Desweiteren gilt $Sp(\mathbf{\Sigma}^2)/Sp^2(\mathbf{\Sigma}) \leq 1$ nach Lemma (C.1.1) im Anhang. Somit konvergiert der Quotient mit Ordnung $\mathcal{O}(n^{-1})$ in Wahrscheinlichkeit gegen 1. Der Quotient (5.10) folgt analog.

□

Die dritte Aussage (5.10) ist somit wesentlich stärker als (5.9). Außerdem erfolgt die Konsistenz des Quotienten (5.9) nicht nur für n gegen unendlich, sondern auch für d gegen unendlich bei festem n, falls das Regularitätskriterium $Sp(\mathbf{\Sigma}^2)/Sp^2(\mathbf{\Sigma}) \to 0$ für $d \to \infty$ erfüllt ist. Sind die Komponenten der $\mathbf{Y}_k$ unkorreliert mit Varianz eins, dann gilt sogar $Sp(\mathbf{\Sigma}^2) = Sp(\mathbf{I}^2) = Sp(\mathbf{I}) = d$ und $Sp^2(\mathbf{\Sigma}) = (Sp(\mathbf{I}))^2 = d^2$ und man hat eine Konvergenz des Quotienten der Ordnung $\mathbf{o}(n^{-1})\cdot\mathbf{o}(d^{-1})$ in Wahrscheinlichkeit gegen 1.

5.2.5 Verteilung von R_n

Unter Hypothese $\mathbf{H}\boldsymbol{\mu} = 0$ galt mit

$$F_n = \frac{1}{n(n-1)}\sum_{k\neq l}^{n}\mathbf{Y}_k'\,\mathbf{Y}_l \tag{5.11}$$

für die zugehörige Teststatistik R_n aus (5.2) für $n \to \infty$ und $d \to \infty$

$$R_n := \frac{F_n}{\sqrt{\frac{2}{n(n-1)}Sp(\mathbf{\Sigma}^2)}} \rightsquigarrow \mathcal{N}(0,1) \tag{5.12}$$

Dies lässt sich wie folgt umformen:

$$R_n = \frac{1}{n(n-1)} \frac{\sum_{k\neq l} \mathbf{Y}_k' \mathbf{Y}_l}{\sqrt{\frac{2}{n(n-1)} Sp(\mathbf{\Sigma}^2)}} = \frac{1}{\sqrt{n(n-1)}} \frac{\sum_{k\neq l} A_{kl}}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \tag{5.13}$$

$$= \frac{n}{\sqrt{n(n-1)}} \frac{1}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \left(\frac{1}{n} \sum_{k\neq l} A_{kl} + B_0 - B_0 \right) \tag{5.14}$$

$$= \frac{n}{\sqrt{n(n-1)}} \frac{1}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \left(\frac{1}{n} \sum_{k\neq l} A_{kl} + \frac{1}{n} \sum_{k=1}^{n} A_{kk} - B_0 \right) \tag{5.15}$$

$$= \frac{n}{\sqrt{n(n-1)}} \frac{1}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \left(\frac{1}{n} \sum_{k,l} \mathbf{Y}_k' \mathbf{Y}_l - B_0 \right) \tag{5.16}$$

$$= \frac{n}{\sqrt{n(n-1)}} \frac{1}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \left(\frac{1}{n} \cdot \sum_{k=1}^{n} \mathbf{Y}_k' \sum_{l=1}^{n} \mathbf{Y}_l - B_0 \right) \tag{5.17}$$

$$= \frac{n}{\sqrt{n(n-1)}} \frac{n \cdot \overline{\mathbf{Y}}_{\cdot}' \overline{\mathbf{Y}}_{\cdot} - B_0}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \tag{5.18}$$

$$= \frac{n}{\sqrt{n(n-1)}} \frac{Q_n - B_0}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \tag{5.19}$$

Durch einfache Umformungen lässt sich R_n also als Funktion von Q_n darstellen. Da $B_0/\sqrt{2Sp(\mathbf{\Sigma}^2)}$ nach (5.10) ein konsistenter Schätzer für $Sp(\mathbf{\Sigma})/\sqrt{2Sp(\mathbf{\Sigma}^2)}$ unabhängig von d ist, sind somit beide Ausdrücke asymptotisch äquivalent. Dementsprechend sind auch $Q_n - Sp(\mathbf{\Sigma})$ und $Q_n - B_0$ asymptotisch äquivalent. Betrachtet man die Varianzen der beiden Ausdrücke, dann ergibt sich nach Satz (5.2.3) und (4.21):

$$Var(Q_n - Sp(\mathbf{\Sigma})) = 2Sp(\mathbf{\Sigma}^2) \tag{5.20}$$

$$Var(Q_n - B_0) = \frac{n-1}{n} \cdot 2Sp(\mathbf{\Sigma}^2) \tag{5.21}$$

während der Erwartungswert bei beiden 0 ist. Um die ersten beiden Momente nicht zu verändern, wird nun $Q_n - B_0$ durch den asymptotisch äquivalenten Ausdruck $\sqrt{n-1}/\sqrt{n} \cdot (Q_n - Sp(\mathbf{\Sigma}))$ ersetzt. Dies hat zur Folge, dass der Faktor $\sqrt{n}/\sqrt{n-1}$ gekürzt wird.
Als asymtptotisch äquivalente Statistik ergibt sich:

$$R_n^* := \frac{Q_n - Sp(\mathbf{\Sigma})}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \doteqdot \frac{n}{\sqrt{n(n-1)}} \frac{Q_n - B_0}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} = R_n \tag{5.22}$$

Zunächst soll eine geeignete Approximation für R_n^* gefunden werden. Nach Boxapproximation aus Abschnitt 5.2.3 ist

$$\frac{Q_n}{Sp\mathbf{\Sigma}} \;\dot{\sim}\; \frac{\chi^2_{f_{box}}}{f_{box}} \qquad \text{mit } f_{box} = \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \tag{5.23}$$

Somit ist

$$R_n^* = \frac{Q_n}{Sp(\mathbf{\Sigma})} \cdot \frac{Sp(\mathbf{\Sigma})}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} - \frac{Sp(\mathbf{\Sigma})}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \tag{5.24}$$

$$\dot{\sim} \frac{\chi^2_{f_{box}}}{f_{box}} \cdot \frac{\sqrt{f_{box}}}{\sqrt{2}} - \frac{\sqrt{f_{box}}}{\sqrt{2}} \tag{5.25}$$

$$\dot{\sim} \frac{\chi^2_{f_{box}} - f_{box}}{\sqrt{2f_{box}}} \tag{5.26}$$

Verteilungen dieser Form sollen Grundlage für eine neue, verbesserte Approximation sein.

5.3 Approximation für beliebige Dimension

Um eine Approximation zu bekommen, die man für Designs beliebiger Dimension anwenden kann, gilt es, die niedrig- und hoch-dimensionalen Ergebnisse zu vereinigen. Nach Theorem 1 aus Chen Qin (2010) [7] ist eine Asymptotik von R_n gegen $\mathcal{N}(0,1)$ für $n \to \infty$ und $d \to \infty$ gegeben. Dabei sind keine Einschränkungen an die gemeinsame Asymptotik von n und d erforderlich. Die neue Approximation von R_n mit der standardisierten χ^2_f-Verteilung

$$\varkappa(f) := \frac{\chi^2_f - f}{\sqrt{2f}} \tag{5.27}$$

ersetzt die Normalapproximation von Chen-Qin. Um die hochdimensionalen Ergebnisse für $d \to \infty$ unter der neuen Approximation $\varkappa(f)$ anwenden zu können, bleibt die Äquivalenz zur Normalapproximation im Hochdimensionalen zu zeigen.

$$\varkappa(f_{box}) \xrightarrow{d\to\infty} \mathcal{N}(0,1) \tag{5.28}$$

Unter der Chen-Qin-Bedingung (4.19), wird nun der Übergang der beiden Approximationen ineinander gezeigt.

5.3.1 Asymptotik für $d \to \infty$

Zunächst wird die Beziehung zwischen d und f_{box} im folgenden Satz untersucht, falls $Sp(\mathbf{\Sigma}^4)/Sp^2(\mathbf{\Sigma}^2) \to 0$ für $d \to \infty$.

Satz 5.3.1 (Gemeinsame Asymptotik von f_{box} und d) *Sei $\mathbf{\Sigma}$ eine positiv-definite $d \times d$ Matrix.*

$$\frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \xrightarrow{d\to\infty} 0 \quad \Rightarrow \quad \frac{Sp(\mathbf{\Sigma}^2)}{Sp^2(\mathbf{\Sigma})} \xrightarrow{d\to\infty} 0 \tag{5.29}$$

Beweis: Siehe Satz (8.1.2) in Kapitel 8

So lässt sich zeigen, dass für $d \to \infty$ auch f_{box} gegen unendlich strebt. Zu zeigen ist nun, dass die Approximation $\varkappa(f_{box})$ unter den Voraussetzungen von Chen-Qin genau die Normalapproximation ist.

Satz 5.3.2 (Grenzverteilung von $\varkappa(f)$) *Für $f \to \infty$ folgt*

$$\varkappa^2(f) = \frac{\chi_f^2 - f}{\sqrt{2f}} \to \mathcal{N}(0,1) \tag{5.30}$$

Beweis: Sei k die größte ganze Zahl, so dass $k \leq f$

$$\frac{\chi_f^2 - f}{\sqrt{2f}} \sim \frac{\left(\sum_{i=1}^{k} \chi_1^2\right) + \chi_{f-k}^2 - f}{\sqrt{2f}}$$

$$\sim \sqrt{\frac{k}{f}} \cdot \frac{\sum_{i=1}^{k} \chi_1^2 - k}{\sqrt{2k}} + \frac{\chi_{f-k}^2 - (f-k)}{\sqrt{2f}}$$

$$\sim \underbrace{\sqrt{\frac{k}{f}}}_{\to 1} \cdot \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \varkappa(1) + \underbrace{\frac{\chi_{f-k}^2 - (f-k)}{\sqrt{2f}}}_{\to 0}$$

$$\to \mathcal{N}(0,1) \qquad \text{für } f \to \infty, \text{ nach ZGWS}$$

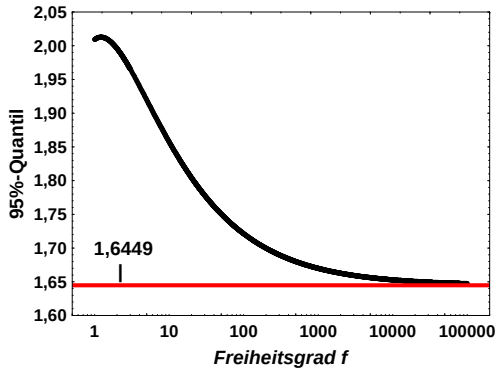
□

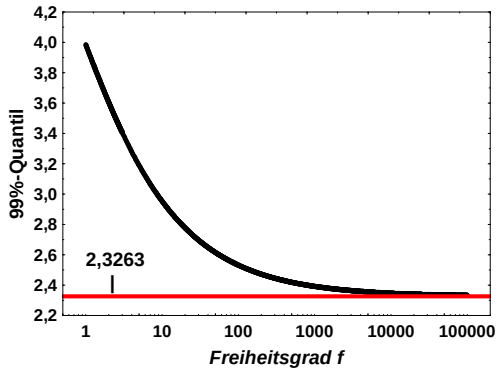
Die Resultate von Chen-Qin stellen somit die Anwendbarkeit der neuen Approximation auch für hochdimensionale Daten sicher. Mit der Grenzwertbetrachtung $d \to \infty$ ist sichergestellt, dass die neue Approximation für alle d konsistent in n ist. Für sehr große n müssen als einzige Approximationsfehler Ungenauigkeiten der Boxapproximation in Kauf genommen werden.
Des Weiteren liegen im hochdimensionalen Fall keine Einschränkungen zwischen der Asymptotik des Stichprobenumfangs n und der Dimension d vor. Im Sinne asymptotischer Statistik impliziert die Grenzwertbetrachtung für $d \to \infty$ somit, dass eine Gleichmäßigkeit in d vorliegt. Dabei lassen sich Güteaussagen dieser Asymptotik für $n \to \infty$ gleichmäßig in d treffen. So konnte im niedrigdimensionalen Fall $d \to 1$ und im hochdimensionalen Fall $d \to \infty$ die Güte der Konvergenz gezeigt werden. Für „mittelgroße“ Dimensionen d ist demgegenüber als einzige Güteaussage möglich, dass der Stichprobenumfang mit Ordnung d^2 wachsen muss. Dieser problematische Anwuchs bleibt aber in jedem Fall beschränkt, da er für weiter ansteigende Dimensionen $d \to \infty$ rückläufig wird.

In der Anwendung von asymptotischen Ergebnissen ist es gerechtfertig praxisrelevante Werte für n anzugeben, ab welcher die Approximation gut funktioniert. Diese Erfahrungswerte basieren für gewöhnlich auf Simulationsergebnissen, da eine strike Fehlerrechnung zu grob wäre. Es sei vorweggenommen, dass im Falle von Dimensionen, die weder als hochdimensional noch als niedrigdimensional klassifiziert werden können, die Güte der Approximation keine bemerkbare Beeinträchtigung erfährt. Somit lässt sich eine praxisrelevante Grenze von $n \geq 10$ angeben, so dass man für alle Dimensionen mit der $\varkappa(f_{box})$-Verteilung eine gute Approxiamtion erhält.
Durch die Relaxion der Voraussetzungen stellt die $\varkappa(f_{box})$-Approximation eine echte Verbesserung zur Normal-Approximation dar. Es sei bemerkt, dass diese bei multivariater Normalverteilung mit identischen Eigenwerten exakt ist.

5.3.2 Konservativität der neuen Approximation

Der Übergang der Verteilung $(\chi^2_f - f)/\sqrt{2f}$ zu einer $\mathcal{N}(0,1)$-Verteilung lässt sich darüber hinaus numerisch untersuchen. Dabei fällt zum einen auf, dass er sehr stabil ist und selbst über das gesamte Spektrum des Parameters f keinen großen Schwankungen unterworfen ist. So kann die Wahl des Parameters als Feineinstellung betrachtet werden. Sie ist für robuste Resultate aber nicht essentiell. Zum anderen ist festzustellen, dass für die relevanten Quantile $1 - \alpha > 0.95$ eine monotone Annäherung von oben zu den Quantilen der Normalverteilung stattfindet. Dies bedeutet, dass für sämtliche Quantile $1 - \alpha > 0.915$ das Ersetzen der Normalverteilung durch die neue Approximation zu konservativeren Ergebnissen führt.





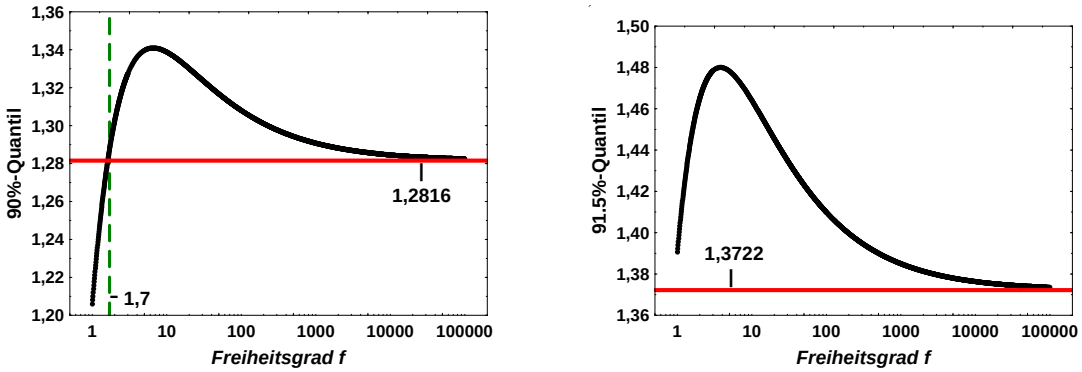


Abbildung 5.1: 95 %-Quantile (links oben) der standardisierten Chi-Quadratverteilung mit Freiheitsgrad f (schwarz, f logarithmisch skaliert) und die der Normalverteilung (rot), die analogen 99 %-Quantile (rechts oben), die 90 %-Quantile (links unten) und die 91,5 %-Quantile (rechts unten).

Die Konservativität ist eine durchaus gewünschte Eigenschaft, da das Niveau in jedem Fall strenger eingehalten wird. Ist somit die Teststatistik mit Normalapproximation nicht liberal, dann ist es die neue Approximation auch nicht.

Desweiteren lässt sich folgern, dass eine Monotonie für ansteigende f vorliegt. Auch wenn diese formal beim 95%-Quantil für kleine f noch nicht gegeben ist, kann man diese approximativ für sämtliche Quantile $\geq 95\%$ annehmen. Die Monotonie bedeutet für die weiteren Kapitel, dass eine Unterschätzung des Freiheitsgrad f zu konservativen Ergebnissen führt, während eine Überschätzung leicht liberale und somit unsichere Ergenisse bringt.
Aufgrund dieser Überlegung wird in dieser Arbeit auf eine Taylorapproximation des Schätzers für f_{box} analog zu 4.2 beruhend auf Werner (2004) [22] verzichtet. Bei der Bildung des Quotienten $\widehat{f}_{box} = B_1/B_2$ soll der Korrekturfaktor $n/(n-1) > 1$ bzw. $1+4/(n(n-1)) > 1$ keine Anwendung finden. Desweiteren wird für den Pearsonschen Freiheitsgrad gezeigt werden, dass f_{pear} stets kleiner gleich f_{box} ist.

In den Graphiken erkennt man außerdem, dass sich die Verteilungen erst für Freiheitsgrade $f >> 100$ angleichen. Dies zeigt, welch große Einschränkung die Bedingung $d \to \infty$ für die Praxis bedeutet, und erklärt die zu erwartenden Niveauüberschreitungen der Teststatistiken von Chen-Qin, Bai-Saranadasa und Srivastava.

5.4 Dimensionsstabile Statistik

Seien die $\mathbf{Y}_k$, $k = 1,..,n$ verteilt nach dem Bai-Saranadasa Modell aus Definition 3.2.2 mit Kovarianzmatrix $\mathbf{\Sigma}$. Desweiteren sei

$$R_n = \frac{\frac{1}{\sqrt{n(n-1)}} \sum_{k \neq l}^{n} \mathbf{Y}_k' \, \mathbf{Y}_l}{\sqrt{2 \cdot Sp(\mathbf{\Sigma}^2)}} \tag{5.31}$$

Es gelte die folgende Restriktion bezüglich der Dimension: Falls $d \to \infty$, dann sei die Chen-Qin-Bedingung

$$\frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \xrightarrow{d \to \infty} 0 \tag{5.32}$$

erfüllt. Dann ergibt sich unter H_0 als asymptotische Verteilung basierend auf der Boxapproximation:

$$R_n \stackrel{\cdot}{\sim} \varkappa(f_{box}) = \frac{\chi^2_{f_{box}} - f_{box}}{\sqrt{2 f_{box}}}, \quad \text{mit } f_{box} = \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \tag{5.33}$$

für $n \to \infty$ gleichmäßig in d.

6 Dimensionsstabiler Test

Um die Ergebnisse aus dem vorherigen Kapitel verwenden zu können, wird eine konsistente Schätzung von $Sp(\mathbf{\Sigma}^2)$ benötigt. Diese wird sowohl für die Varianz der Quadratischen Form als auch für den Freiheitsgrad f der standardisierten Chi-Quadrat-Verteilung $\varkappa(f)$ gebraucht. Dafür werden zunächst verschiedene Schätzer untersucht.

6.1 Schätzer für $Sp(\mathbf{\Sigma}^2)$ nach Chen und Qin

In der Arbeit von Chen-Qin wird der Parameter $Sp(\mathbf{\Sigma}^2)$ wie folgt geschätzt:

$$\widehat{Sp(\mathbf{\Sigma}^2)} = \frac{1}{n(n-1)} Sp\left(\sum_{k\neq l}^{n} (\mathbf{Y}_k - \overline{\mathbf{Y}}_{(kl)})\mathbf{Y}_k'(\mathbf{Y}_l - \overline{\mathbf{Y}}_{(kl)})\mathbf{Y}_l'\right) \tag{6.1}$$

Ursprünglich wurde der Schätzer für den Zwei-Stichprobenfall entwickelt. Die Konsistenz, gleichmäßig in d, konnte allerdings nur unter Hypothese $\mathbf{H}\boldsymbol{\mu} = \mathbf{0}$ bewiesen werden. Ist $\boldsymbol{\mu} = \mu \cdot \mathbf{1}_d$ mit $\mu \in \mathbb{R}$, dann wird die Varianz des Schätzers für $\mu \to \infty$ beliebig groß. Somit ist der Schätzer im Hochdimensionalen ohnehin nur im Ein-Gruppendesign sinnvoll anwendbar. Mit diesem Schätzer ergibt sich die Chen-Qin-Teststatistik wie folgt:

$$C_n := \frac{\frac{1}{\sqrt{n(n-1)}} \sum_{k\neq l}^{n} \mathbf{Y}_k' \mathbf{Y}_l}{\sqrt{2 \cdot \widehat{Sp(\mathbf{\Sigma}^2)}}} \stackrel{\cdot}{\sim} \mathcal{N}(0,1) \tag{6.2}$$

Es sei bemerkt, dass Becker (2010) [2] einen Schätzer für $Sp(\mathbf{\Sigma}^2)$ unter Normalverteilung entwickeln konnte, welcher unter Alternative konsistent und erwartungstreu ist. Im Simulationsteil 11.3.2 dieser Arbeit wird die Güte dieser Schätzer untersucht.

6.2 B_2-Schätzer

Demgegnüber konnte Werner (2004) [22] ebenfalls unter Normalverteilung einen unter H_0 sehr effizienten Schätzer angeben. Helms konnte in seiner Diplomarbeit (2010) [13] dessen Konsistenz für eine spezielle Klasse von nichtnormalverteilten Zufallsvektoren zeigen. Dieser Schätzer

$$B_2 := \frac{1}{n(n-1)} \sum_{k \neq l}^{n} (\boldsymbol{Y}'_k \boldsymbol{Y}_l)^2 \tag{6.3}$$

für $Sp(\boldsymbol{\Sigma}^2)$ soll auch in dieser Arbeit verwendet werden. Zunächst ist die Erwartungstreue und die Konsistenz im Bai-Saranadasa Modell zu zeigen.

Lemma 6.2.1 (Höhere Momente der A_{kl} II)

Die Zufallsvektoren $\mathbf{Y}_k$, $k = 1,..,n$ seien unabhängig verteilt wie in Lemma C.2.1 (Bai-Saranadasa Modell) mit $E(\mathbf{Y}_k) = \mathbf{0}_d$ und Kovarianzmatrix $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$. Die $A_{kl} := \mathbf{Y}'_k \mathbf{Y}_l$ seien die zugehörigen Bilinearformen für $k \neq l$. Dann gilt für deren Momente mit sämtlichen Indexkombinationen $k \neq l \neq k' \neq l'$:

3. $E(A_{kl}) = 0$
4. $E(A_{kl}^2) = Sp(\boldsymbol{\Sigma}^2)$
5. $E(A_{kl}^4) = \mathcal{O}(Sp^2(\boldsymbol{\Sigma}^2))$
6. $E(A_{kl}^2 A_{kr}^2) = \mathcal{O}(Sp^2(\boldsymbol{\Sigma}^2))$
7. $E(A_{kl}^2 A_{k'l'}^2) = Sp^2(\boldsymbol{\Sigma}^2))$

Beweis: Siehe Lemma C.2.1 im Anhang.

Satz 6.2.2 (B_2-Schätzer)
Der Schätzer $B_2 = \frac{1}{n(n-1)} \sum_{k \neq l} A_{kl}^2$ für $Sp(\boldsymbol{\Sigma}^2)$ ist erwartungstreu und konsistent und der Quotient $B_2/Sp(\boldsymbol{\Sigma}^2)$ konvergiert gleichmäßig in d gegen 1 in $\mathcal{L}_2$-Norm.

Beweis:
Für den Erwartungswert gilt:

$$E(B_2) = \frac{1}{n(n-1)} \sum_{k \neq l} E(A_{kl}^2) = Sp(\boldsymbol{\Sigma}^2)$$

Für die Varianz gilt zunächst:

$$\begin{aligned} E(B_2^2) &= E((\tfrac{1}{n(n-1)} \textstyle\sum_{k\neq l} A_{kl}^2)^2) \\ &= E(\tfrac{1}{n^2(n-1)^2} \textstyle\sum_{k\neq l} \sum_{r\neq s} A_{kl}^2 A_{rs}^2) \\ &= \tfrac{2n(n-1)}{n^2(n-1)^2} E(A_{kl}^4) + \tfrac{4n(n-1)(n-2)}{n^2(n-1)^2} E(A_{kl}^2 A_{kr}^2) + \tfrac{n(n-1)(n-2)(n-3)}{n^2(n-1)^2} E^2(A_{kl}^2) \\ &= \mathcal{O}(n^{-2}) \cdot \mathcal{O}(Sp^2(\mathbf{\Sigma}^2)) + \mathcal{O}(n^{-1})\mathcal{O}(Sp^2(\mathbf{\Sigma}^2)) + \tfrac{n(n-1)(n-2)(n-3)}{n^2(n-1)^2} Sp^2(\mathbf{\Sigma}^2) \end{aligned}$$

Dann folgt für die Varianz von $B_2/Sp(\mathbf{\Sigma}^2)$:

$$\begin{aligned} Var(B_2/Sp(\mathbf{\Sigma}^2)) &= \tfrac{1}{Sp^2(\mathbf{\Sigma}^2)} \cdot (E(B_2^2) - E^2(B_2)) \\ &= \mathcal{O}(n^{-1}) \cdot \tfrac{\mathcal{O}(Sp^2(\mathbf{\Sigma}^2)}{Sp^2(\mathbf{\Sigma}^2)} + \tfrac{n(n-1)(n-2)(n-3) - n^2(n-1)^2}{n^2(n-1)^2} \cdot \tfrac{Sp^2(\mathbf{\Sigma}^2)}{Sp^2(\mathbf{\Sigma}^2)} \\ &= \mathcal{O}(n^{-1}) + \tfrac{-4n+6}{n(n-1)} \\ &= \mathcal{O}(n^{-1}) \end{aligned}$$

□

Die Qualität der Schätzung von $Sp(\mathbf{\Sigma}^2)$ mittels dem Schätzer B_2 ist wird also nicht beliebig schlechter mit zunehmender Dimension. Ähnliche Konvergenzaussagen des Schätzers B_1 für $Sp^2(\mathbf{\Sigma})$, der nach dem Test von Werner-Brunner aus 4.2 zur Schätzung des Boxschen Freiheitsgrades f_{box} benötigt wird, werden im folgenden Abschnitt hergeleitet.

6.3 B_1-Schätzer

Für $Sp(\mathbf{\Sigma})$ wurde in (5.7) der Schätzer B_0 angegeben, dessen Konvergenzeigenschaften unabhängig von der Dimension d gezeigt wurden. Daraus lässt sich nun auf einfachem Wege ein Schätzer für $Sp^2(\mathbf{\Sigma})$ konstruieren.

Satz 6.3.1 (B_1-Schätzer) *Der Schätzer*

$$B_1 := \frac{1}{n(n-1)} \sum_{k\neq l}^{n} A_{kk} A_{ll} \tag{6.4}$$

für $Sp^2(\mathbf{\Sigma})$ ist erwartungstreu und der Quotient $B_1/Sp(\mathbf{\Sigma}^2)$ konvergiert gleichmässig in d gegen $Sp^2(\mathbf{\Sigma})/Sp(\mathbf{\Sigma}^2)$.

Beweis: Nach Satz 5.2.5 ist der Schätzer B_0 erwartungstreu und $B_0/\sqrt{Sp(\mathbf{\Sigma}^2)}$ konvergeirt gleichmässig in der Dimension gegen 1. Diese Eigenschaften übertragen sich nach Satz B.0.7 auf B_1 .

□

6.4 Teststatistik

Mit den entwickelten Schätzern lässt sich nun ein Schätzer für den Freiheitsgrad f_{box} angeben. Nach dem Satz von Slutsky (A.2.1) und dem continuous mapping theorem (A.2.2) ist

$$\widehat{f}_{box} := \frac{B_1}{B_2} = \underbrace{\frac{B_1}{Sp(\mathbf{\Sigma}^2)}}_{\to f_{box}} \cdot \underbrace{\left(\frac{B_2}{Sp(\mathbf{\Sigma}^2)}\right)^{-1}}_{\to 1} \tag{6.5}$$

$$\xrightarrow{p} \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \tag{6.6}$$

ein konsistenter Schätzer für $n \to \infty$. Da x^{-1} eine konvexe Funktion ist und darüber hinaus B_1 und B_2 nicht unabhängig sind, ist $\widehat{f}_{box}$ im allgemeinen verzerrt, also nicht erwartungstreu. Die in der Dimension gleichmäßige Konsistenz soll aber genügen.
Mit den Voraussetzungen und der Statistik R_n aus Abschnitt 5.4 ergibt sich nun mit den entwickelten Schätzern folgender Test für die Hypothese $\mathbf{H}\boldsymbol{\mu} = \mathbf{0}$:

$$Z_n := \frac{\frac{1}{\sqrt{n(n-1)}} \sum_{k \neq l}^{n} \mathbf{Y}_k' \, \mathbf{Y}_l}{\sqrt{2 \cdot B_2}} \tag{6.7}$$

$$= \frac{\sum_{k \neq l}^{n} \mathbf{Y}_k' \, \mathbf{Y}_l}{\sqrt{2 \cdot \sum_{k \neq l}^{n} (\mathbf{Y}_k' \, \mathbf{Y}_l)^2}} \stackrel{\cdot}{\sim} \varkappa(\widehat{f}_{box}) \tag{6.8}$$

Folgende Abbildungen sollen einen Eindruck über die Güte des Verfahrens vermitteln. Simuliert wurde das Niveau der vorgestellten neuen Teststatistik im Vergleich mit bisherigen Verfahren. Als Verteilung wurde zunächst die multivariate Normalverteilung gewählt, welche als Goldstandard dienen soll.

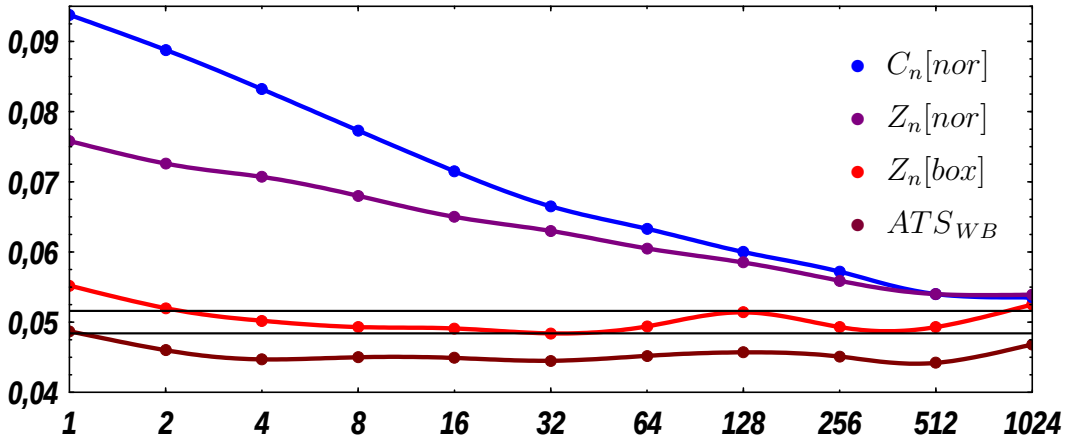


Abbildung 6.1: In der Graphik wurde das Niveau der verschiedener Testverfahren über die Dimension aufgetragen. Dabei steht $C_n[nor]$ für den Test nach Chen-Qin aus 4.6, ATS_{WB} für die ANOVA-Typ-Statistik aus 4.2, während $Z_n[box]$ bzw. $Z_n[nor]$ für den Test der Statistik Z_n mit $\varkappa(\widehat{f}_{box})$- bzw. Normal- Approximation stehen. Dabei wurde die multivariate Normalverteilung, mit Einheitsmatrix als Kovarianz- und Hypothesenmatrix simuliert. In 10000 Simulationsdurchläufen bei Stichprobenumfang von $n = 15$ wurde zum Niveau $\alpha = 0{,}05$ getestet.

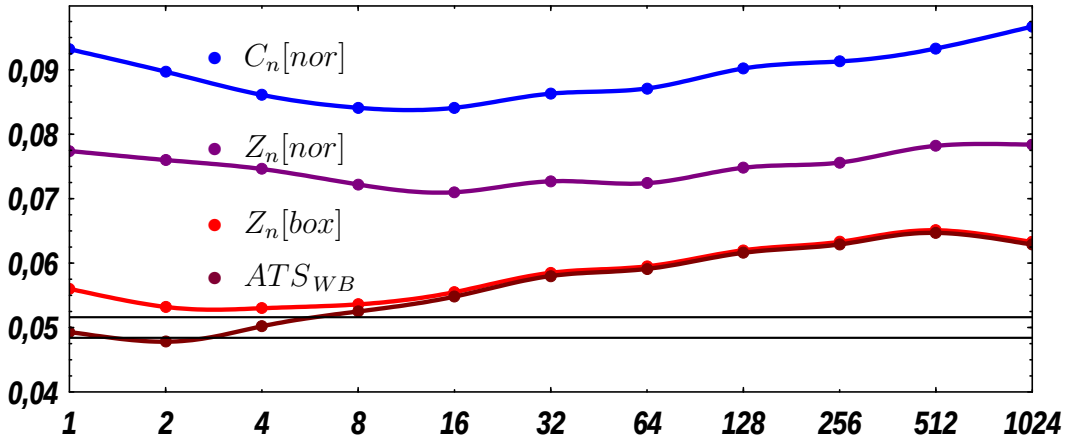


Abbildung 6.2: Simuliert wurde eine multivariate Normalverteilung mit Compound Symmetry Kovarianzstruktur $\mathbf{\Sigma} = \mathbf{I}_d + 1/2 \cdot \mathbf{J}_d$. Die multivariate Hypothese und die übrigen Parameter sind wie in Abbildung 6.1.

7 Pearsonapproximation

Die Simulationsergebnisse aus der Abbildung 6.1 zeigen, dass die Statistik Z_n mit $\varkappa(f_{box})$ bereits sehr gut approximiert wird, falls für große Dimension die Chen-Qin-Regularitätsbedingung (4.19) erfüllt ist. Leichte Niveauüberschreitungen zeigen sich einzig bei ungünstigen Eigenwertstrukturen der Kovarianzmatrizen, wie in Abbildung 6.2 zu sehen ist. So kann der Test beispielsweise bei Toeplitzstruktur mit linearem Abfall der Kovarianzen oder auch bei Compound Symmetry mit der multivariaten Hypothese $\boldsymbol{\mu} = \mathbf{0}$ etwas liberal werden. Diese Strukturen sind dadurch gekennzeichnet, dass einzelne Eigenwerte relativ zu den anderen sehr groß werden. Ist nun ein Eigenwert wesentlich größer als jeweils alle anderen, aber nicht so groß, dass er alle anderen zusammen dominiert, dann kann dies zu Approximationsungenauigkeiten bei der Boxapproximation führen.

Sind alle Eigenwerte identisch oder gleich null, was beispielsweise bei unabhängigen Messungen mit gleicher Varianz der Fall ist, ist die Boxapproximation exakt. Auch bezüglich einer leichten Inhomogenität ist die Approximation sehr robust und bietet, wie in Kapitel 5 beschrieben, eine wesentliche Verbesserung zur Approximation mit Normalverteilung. Sie findet in vielen Arbeiten Anwendung und hat sich als ein Standardverfahren zur Approximation einer Summe von gewichteten χ^2_1-verteilten Zufallsvariablen herausgebildet. Einzig bei ungünstigen Eigenwertbedingungen zeigen sich Schwächen, die schon von Werner [22] erwähnt wurden. Gerade beim Testen zu einem geringen Niveau ($\alpha < 0.05$) kann die Boxapproximation stark liberal werden. Man betrachte dazu die Quantile-Quantile-Plots in Abbildung 7.1.

Diese Nachteile der Boxapproximation motivieren dazu, sie zu verfeinern und die Information der höheren Momente auszunutzen. So soll eine Approximation gefunden werden, welche neben den ersten beiden Momenten zusätzlich das dritte mit berücksichtigt. Diese wird in dieser Arbeit als Pearsonapproximation bezeichnet, da solch eine Approximation bereits 1959 von Egon Pearson [15] veröffentlicht wurde.

7.1 Pearsonapproximation

Sei

$$Q = \sum_{i=1}^{d} \lambda_i C_i \quad \text{mit } C_i \sim \chi_1^2 \text{ unabhängig, } i = 1,...,d \tag{7.1}$$

und λ_i seien die Eigenwerte der Kovarianzmatrix $\mathbf{\Sigma}$. Die Idee der neuen Approximation ist nun, die wahre Verteilung $\sum \lambda_i \chi_1^2$ mit einer gestreckten und zentrierten Chi-Quadrat-Verteilung $g \cdot \chi_f^2 - h$ zu approximieren, so dass die ersten 3 Momente gleichgesetzt werden.
Dies ist unter anderem darin motiviert, die Teststatistik Z_n aus (6.7) als standardisierte Testgröße mit einer standardisierten Verteilung $\varkappa(f)$ zu approximieren. So stimmen die ersten beiden Momente überein, während der Freiheitsgrad f von $\varkappa(f)$ noch nicht bestimmt ist. Anstatt

$$f_{box} = \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \tag{7.2}$$

aus der Boxapproximation zu übernehmen, lässt sich das dritte Moment als weiterer Parameter hinzuziehen, um einen neuen Freiheitsgrad f für eine exaktere Approximation zu ermitteln. Gerade dieser ist essentiell, um in niedrigdimensionalen Designs eine Verbesserung zur Boxapproximation bzw. zur Normalapproximation zu erhalten, da diese bereits durch die ersten beiden Momente voll bestimmt sind.
Zur Approximation von Q mit $g \cdot \chi_f^2 - h$ ergeben sich die Momente nach Abschnitt 5.2.3 wie folgt. Betrachtet werden Erwartungswert, Varianz und Schiefe ν von Q:

$$E(Q) = Sp(\mathbf{\Sigma}), \qquad Var(Q) = 2 \cdot Sp(\mathbf{\Sigma}^2), \qquad \nu(Q) = \frac{2\sqrt{2} \cdot Sp(\mathbf{\Sigma}^3)}{(Sp(\mathbf{\Sigma}^2))^{(3/2)}}$$

$$E\left(\chi_f^2\right) = f, \qquad Var\left(\chi_f^2\right) = 2 \cdot f, \qquad \nu\left(\chi_f^2\right) = \frac{2\sqrt{2}}{\sqrt{f}}$$

Betrachte nun

$$\frac{Q - Sp(\mathbf{\Sigma})}{\sqrt{2 \cdot Sp(\mathbf{\Sigma}^2)}} \stackrel{!}{=} \frac{\chi_f^2 - f}{\sqrt{2 \cdot f}} \tag{7.3}$$

Dann sind beide Ausdrücke standardisiert, wodurch der Erwartungswert (= 0) und die Varianz (= 1) bereits übereinstimmen. Der Parameter f bleibt als Freiheitsgrad, der aus der Schiefe der Ausdrücke zu bestimmen ist.

$$\nu(Q) \stackrel{!}{=} \nu(\chi_f^2) \tag{7.4}$$

$$\Rightarrow \frac{2\sqrt{2} \cdot Sp(\mathbf{\Sigma}^3)}{Sp^{(3/2)}(\mathbf{\Sigma}^2)} = \frac{2\sqrt{2}}{\sqrt{f}} \tag{7.5}$$

$$\Rightarrow f_{pear} := \frac{(Sp\mathbf{\Sigma}^2)^3}{(Sp\mathbf{\Sigma}^3)^2} \tag{7.6}$$

Dieses Resultat lässt sich nun im niedrigdimensionalem bzw. im Normalverteilungs-Fall auf Q_n anwenden. Analog zur Anwendung der Boxapproximation aus Abschnitt 5.2.5 ist

$$R_n^* = \frac{Q_n - Sp(\mathbf{\Sigma})}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \tag{7.7}$$

$$\doteqdot \frac{\sum_{i=1}^{d} \lambda_i C_i - Sp(\mathbf{\Sigma})}{\sqrt{2Sp(\mathbf{\Sigma}^2)}} \tag{7.8}$$

$$\dot{\sim} \frac{\chi^2_{f_{pear}} - f_{pear}}{\sqrt{2f_{pear}}} \tag{7.9}$$

7.2 Vergleich mit Boxapproximation

Die Güte der Approximationen nach Box und nach Pearson lässt sich nun vegleichen. Für einen direkten Vergleich mit der Boxapproximation ist die Darstellung (7.9) in die aus (5.23) zu überführen. Lösen des Gleichungssystems nach $Q_n/Sp(\mathbf{\Sigma})$ ergibt als Analogon zur Boxapproximation

$$\frac{Q_n}{Sp(\mathbf{\Sigma})} \dot{\sim} \frac{\chi^2_{f_{box}}}{f_{box}}, \qquad f_{box} = \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \tag{7.10}$$

$$\frac{Q_n}{Sp(\mathbf{\Sigma})} \dot{\sim} \left(\frac{\chi^2_{f_{pear}} Sp(\mathbf{\Sigma}^3)}{Sp(\mathbf{\Sigma}^2)Sp(\mathbf{\Sigma})} - \frac{Sp^2(\mathbf{\Sigma}^2)}{Sp(\mathbf{\Sigma}^3)Sp(\mathbf{\Sigma})} + 1 \right), \qquad f_{pear} = \frac{Sp^3(\mathbf{\Sigma}^2)}{Sp^2(\mathbf{\Sigma}^3)} \tag{7.11}$$

Erwartungswert und Varianzen stimmen bei sämtlichen Termen überein.

Im Folgenden wurden einige Simulationen durchgeführt, um festzustellen, wie gut nun im Einzelnen die verschiedenen Approximationen die exakte Verteilung $Q = \sum_{i=1}^{d} \lambda_i C_i$ einer gewichteten Summe aus χ^2_1-verteilten Zufallsvariablen C_i annähern. Dabei wurden neben der Pearsonapproximation und Boxapproximation auch die Normalapproximation, welche ursprünglich nach Chen und Qin für Z_n bzw. C_n verwendet wird, untersucht. Letztere stellt, wie in Satz 5.3.2 gezeigt, gleichzeitig die Grenzverteilung der standardisierten χ^2_f-Approximation für $f \to \infty$ dar.

Es wurden möglichst nichttriviale Gewichte gewählt, um das Verhalten unter extremen Kovarianzbedingungen wie Compound Symmetry oder Toeplitzstruktur (mit linearem Abfall der Kovarianzen) zu untersuchen. Bei gleichen Gewichten sind die Box- und Pearsonapproximation exakt und bei ähnlich großen Gewichten somit auch dementsprechend gut, weshalb diese nicht geplottet wurden. Für extrem heterogene Gewichte kann man gut sehen, dass die neue Approximation nach Pearson die Boxapproximation stark verbessert. Bei dominierenden Eigenwerten ist bei letzterer sogar kaum noch eine wesentliche Verbesserung zur Normalverteilung erkennbar. Stellt man keine Voraussetzungen an die Kovarianzmatrix, so bietet die neue Approxmation nach Pearson eine erhebliche Verbesserung. Die Schwierigkeit dieser Approximation besteht allerdings darin, geeignete Schätzer für den Freiheitsgrad f_{pear} zu finden.

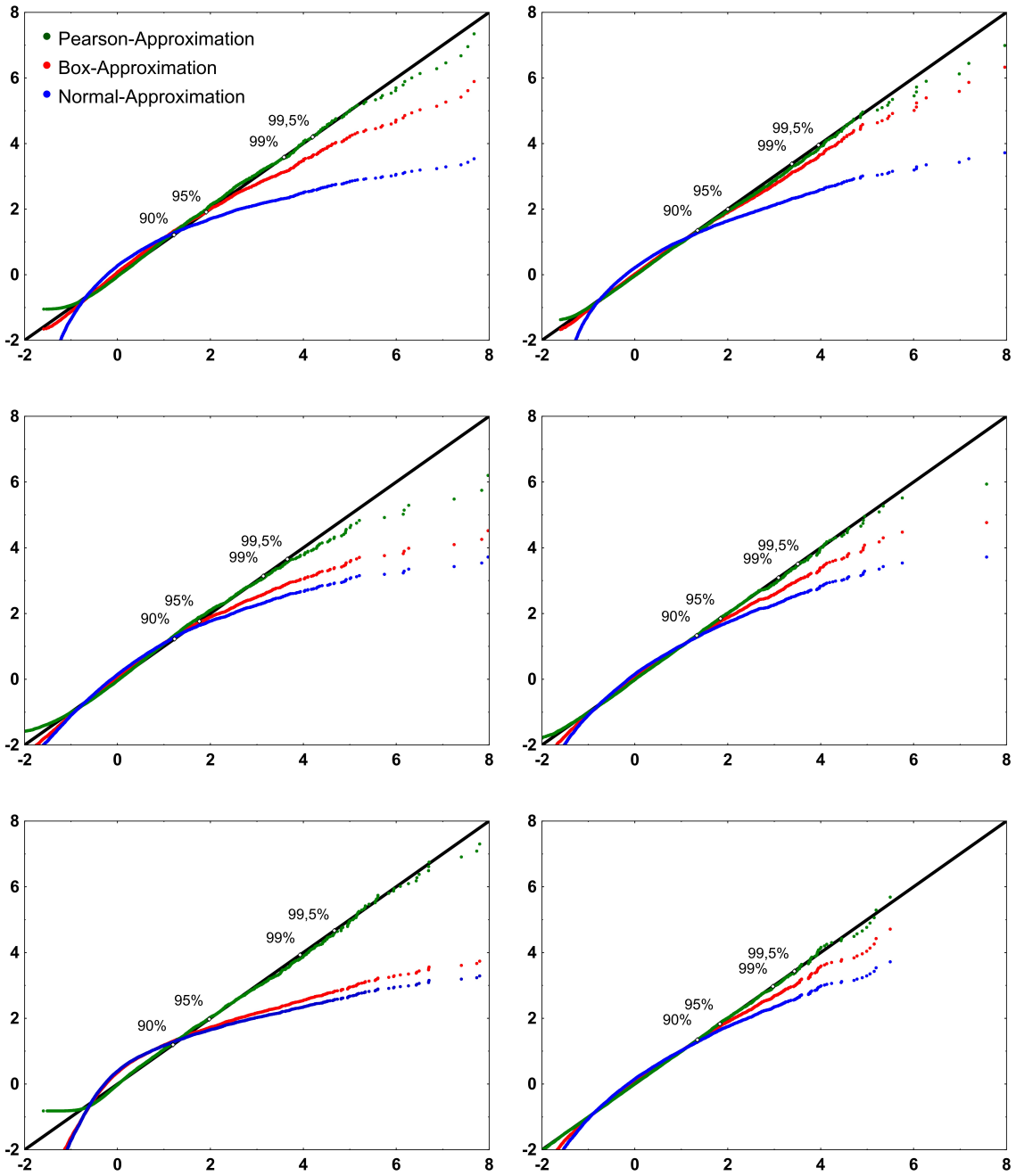


Abbildung 7.1: QQ-Plots einer Summe von gewichteten χ^2-verteilten Zufallsvariablen gegen Normalverteilung und standardisierter χ^2_f-Verteilung mit Boxschem bzw. Pearsonschen Freiheitsgrad f. Folgende Gewichte wurden verwendet:
$\lambda = (5,1,..,1) \in \mathbb{R}^{10}$ (oben links),
$\lambda = (5,5,5,1..,1) \in \mathbb{R}^{10}$ (oben rechts),
$\lambda = (10,1,..,1) \in \mathbb{R}^{100}$ (Mitte links),
$\lambda = (10 \cdot \mathbf{1}'_5, \mathbf{1}'_{95}) \in \mathbb{R}^{100}$ (Mitte rechts),
$\lambda = (100,1,..,1) \in \mathbb{R}^{1000}$ (unten links),
$\lambda = (100 \cdot \mathbf{1}'_{10}, \mathbf{1}'_{990}) \in \mathbb{R}^{1000}$ (unten rechts).

7.3 Reparametrisierung von $\varkappa(f)$

Die Verteilung der neuen Approximation

$$\varkappa(f) = \frac{\chi_f^2 - f}{\sqrt{2f}} \tag{7.12}$$

wurde über den Parameter f gestimmt. Diese stammt, als Überbleibsel der Herleitung, aus der Parametrisierung über den Freiheitsgrad f der χ_f^2-Verteilung. Da der Parameter f unter Umständen schwierig zu schätzen ist, bietet sich eine Reparametrisierung mittels eines alternativen Parameters f' an, der leichter zu schätzen ist. Dabei wird ausgenutzt, dass $\varkappa(f)$ mittels Transformationen des Argumentes $f' := u(f) : \mathbb{R} \to \mathbb{R}$ in eine stabilere Form gebracht werden kann. Stabil bedeutet in diesem Zusammenhang, dass die Verteilung $\varkappa(u^{-1}(f'))$ sich bezüglich Abweichungen von f' möglichst wenig ändert. Explizit betrachtet man dazu die Quantile von $\varkappa(u^{-1}(f'))$ und wählt u so, dass diese möglichst gleichmäßig stetig bezüglich f' sind. Auf diese Weise haben Schätzfehler von f' minimale Auswirkungen auf die Quantile von $\varkappa$.

Während sich f_{box} unabhängig von der Dimension konsistent schätzen ließ, erweist sich die Konstruktion eines Schätzers für f_{pear} mit den gleichen Eigenschaften als schwierig. Deshalb sollen Transformationen motiviert werden, so dass $u(f_{pear})$ gleichmäßig in der Dimension schätzbar wird. Als sinnvolle Transformationen $u(f)$ bieten sich an:

$$h := \frac{1}{f} \tag{7.13}$$

und

$$g := \frac{1}{\sqrt{f}} \tag{7.14}$$

Andere Transformationen wie beispielsweise $u(f) = \sqrt[3]{\frac{1}{f}}$ sind prinzipiell möglich, allerdings nur dann sinnvoll, wenn auch einfache Schätzer für $u(f)$ angegeben werden können.

Eine analytische Berechnung der α-Quantile ist nicht möglich, zumal man die Verteilungsfunktion der χ_f^2-Verteilung nicht in geschlossener Form angeben kann. Eine numerische Bestimmung soll aber genügen, um die gleichmäßige Stetigkeit bezüglich f' zu überprüfen.
Für das 90%-, 95%- und 99%-Quantil erhält man folgende Ergebnisse:

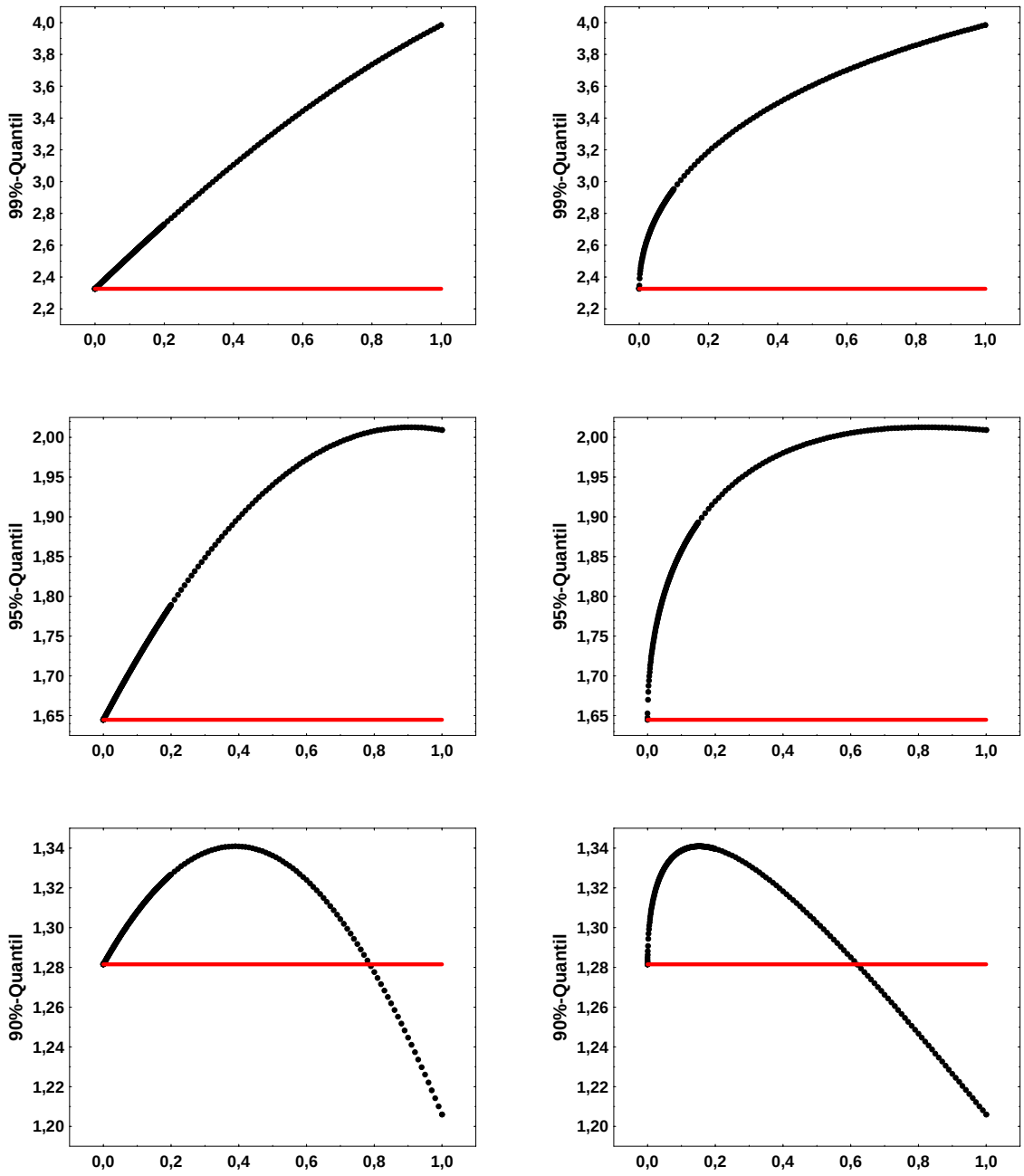


Abbildung 7.2: Auf der linken Seite sind die Quantile der $\varkappa(f)$-Verteilung mit der Transformation $f = 1/g^2$ über g aufgetragen. Auf der rechten Seite sind die Quantile mit der Transformation $f = 1/h$ über h. In der obersten Zeile wurden die 99%-Quantile abgebildet, in der mittleren die 95%-Quantile und in der unteren die 90%-Quantile. Zum Vergleich bezüglich f ist die Abbildung 5.1 zu betrachten.

Desweiteren ist

$$quantile_{\varkappa}(\alpha,u^{-1}(f')) \tag{7.15}$$

stetig bezüglich der Argumente α und $u^{-1}(f')$, da die Quantilsfunktion als Inverse der Verteilungsfunktion von $\varkappa$ eine Kombination von stetigen Funktionen darstellt. Somit lässt sich mit Blick auf die Graphen schlussfolgern, dass

$$quantile_{\varkappa}\left(\alpha,\frac{1}{g^2}\right) \tag{7.16}$$

einen gleichmäßigen Anstieg bezüglich α und g besitzt. Dabei sei $\alpha \in [0,1]$ beliebig, während für $g \in (0,1]$ die Einschränkung $g > g_0 > 0$ gefordert wird. Letztere Einschränkung ist allerdings bei gewöhnlichen Freiheitsgraden wie dem Boxschen oder dem Pearsonschen trivialerweise erfüllt, da $g_{pear} \geq g_{box} \geq 1/\sqrt{d} =: g_0$. Siehe dazu Satz 8.1.2. Somit ist die Funktion $1/g$ und damit die Quantilsfunktion gleichmäßig stetig in $g \in [g_0,1]$. Numerisch lässt sich nun die Güte dieser gleichmäßigen Stetigkeit verifizieren, indem man die Interpolation bei der Berechnung entsprechend klein wählt. Die Interpolation in Abbildung 7.2 soll als Ausschnitt genügen, um die gleichmäßige Stetigkeit zu belegen.
Demgegenüber ist für h die gleichmäßige Stetigkeit des Plots beispielsweise nicht ersichtlich. In jedem Fall scheint die Quantilsfunktion weniger glatt. Deshalb wird folgende Parametrisierung der standardisierten χ^2-Verteilung definiert:

$$\kappa(g) := \varkappa\left(\frac{1}{g^2}\right) = \frac{g}{\sqrt{2}}\left(\chi^2_{g^{-2}} - \frac{1}{g^2}\right) \tag{7.17}$$

für $g \in (0,1]$ und $\kappa(g) := \mathcal{N}(0,1)$ für $g = 0$. Der Übergang $g \to 0$ ist nach Satz 5.3.2 stetig.

Lässt sich nun

$$g_{pear} = \sqrt{\frac{1}{f_{pear}}} = \frac{Sp(\mathbf{\Sigma}^3)}{Sp^{3/2}(\mathbf{\Sigma}^2)} \tag{7.18}$$

konsistent, gleichmäßig in d, schätzen, so kann man die asymptotische Verteilung von R_n mit $\kappa(\widehat{g}_{pear})$ angeben.

7.4 Schätzung von g_{pear}

Während sich eine stabile Schätzung von f_{pear} bezüglich steigender Dimension als schwierig erweist, wird in diesem Abschnitt alternativ ein Schätzer von g_{pear} konstuiert und dessen gleichmäßige Konsistenz gezeigt.
Die Schätzung von

$$g_{pear} = \frac{Sp(\mathbf{\Sigma}^3)}{Sp^{3/2}(\mathbf{\Sigma}^2)} \tag{7.19}$$

erfolgt analog zur Boxapproximation für Zähler und Nenner getrennt. Zunächst wird ein Schätzer für $Sp(\mathbf{\Sigma}^3)$ benötigt. Sei

$$B_3 := (n(n-1)(n-2))^{-1} \sum_{k \neq l \neq r} A_{kl} A_{lr} A_{rk} \tag{7.20}$$

Dann folgt aus

$$E(B_3) = E(Sp(\mathbf{Y}_k' \mathbf{Y}_l \ \mathbf{Y}_l' \mathbf{Y}_r \ \mathbf{Y}_r' \mathbf{Y}_k)) = Sp(\mathbf{\Sigma}^3) \tag{7.21}$$

dass B_3 ein erwartungstreuer Schätzer für $Sp(\mathbf{\Sigma}^3)$ ist.

Lemma 7.4.1 (Höhere Momente der A_{kl} III)

Die Zufallsvektoren $\mathbf{Y}_k$, $k = 1,..,n$ seien unabhängig verteilt wie in Lemma C.2.1 (Bai-Saranadasa Modell) mit $E(\mathbf{Y}_k) = \mathbf{0}_d$ und Kovarianzmatrix $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}'$. Die $A_{kl} := \mathbf{Y}_k' \mathbf{Y}_l$ seien die zugehörigen Bilinearformen für $k \neq l$. Dann gilt für deren Momente mit sämtlichen Indexkombinationen $k \neq l \neq r \neq k' \neq l' \neq r'$:

8. $E(A_{kl}^2 A_{lr}^2 A_{rk}^2) = \mathcal{O}(Sp^3(\mathbf{\Sigma}^2))$
9. $E(A_{kl} A_{lr} A_{rk} A_{k'l} A_{lr} A_{rk'}) = \mathcal{O}(Sp^3(\mathbf{\Sigma}^2))$
10. $E(A_{kl} A_{lr} A_{rk} A_{kl'} A_{l'r'} A_{r'k}) = \mathcal{O}(Sp^2(\mathbf{\Sigma}^3))$
11. $E(A_{kl} A_{lr} A_{rk} A_{k'l'} A_{l'r'} A_{r'k'}) = (Sp^2(\mathbf{\Sigma}^3))$

Beweis: Siehe Lemma C.2.1 im Anhang.

Satz 7.4.2 (Varianz des B_3-Schätzer) *Unter den obigen Veraussetzungen gilt:*

$$Var(B_3 / Sp^{\frac{3}{2}}(\mathbf{\Sigma}^2)) \to 0 \quad \textit{gleichmäßig für alle } d \tag{7.22}$$

Beweis:
Für die Varianz gilt zunächst:

$$E(B_3^2) = E(((n(n-1)(n-2))^{-1} \sum_{k\neq l\neq r} A_{kl}A_{lr}A_{rk})^2)$$

$$= \tfrac{1}{n^2(n-1)^2(n-2)^2} \sum_{k\neq l\neq r} \sum_{k'\neq l'\neq r'} E(A_{kl}A_{lr}A_{rk}\ A_{k'l'}A_{l'r'}A_{r'k'})$$

Nun trennt man die Summe in Terme mit gleich vielen abhängigen Variablen bzw. gleichen Indizes:

$$= 3\cdot 2\ \cdot \tfrac{1}{n^2(n-1)^2(n-2)^2} \sum_{k\neq l\neq r} E(A_{kl}^2A_{lr}^2A_{rk}^2)$$

$$+ 3\cdot 2\cdot 3\cdot (n-3)\ \cdot \tfrac{n(n-1)(n-2)}{n^2(n-1)^2(n-2)^2} \sum_{k\neq l\neq r\neq k'} E(A_{kl}A_{lr}A_{rk}A_{k'l}A_{lr}A_{rk'})$$

$$+ 3\cdot 3\cdot (n-3)\cdot (n-4)\cdot\ \cdot \tfrac{n(n-1)(n-2)}{n^2(n-1)^2(n-2)^2} \sum_{k\neq l\neq r\neq k'\neq l'} E(A_{kl}A_{lr}A_{rk}A_{k'l'}A_{l'r}A_{rk'})$$

$$+ (n-3)\cdot (n-4)\cdot (n-5)\cdot\ \cdot \tfrac{n(n-1)(n-2)}{n^2(n-1)^2(n-2)^2} \sum_{k\neq l\neq r\neq k'\neq l'} E(A_{kl}A_{lr}A_{rk}A_{k'l'}A_{l'r'}A_{r'k'})$$

Mit Lemma 7.4.1 folgt nun:

$$= \tfrac{6}{n(n-1)(n-2)}\cdot \mathcal{O}(Sp^3\mathbf{\Sigma}^2) + \tfrac{18(n-3)}{n(n-1)(n-2)}\cdot \mathcal{O}(Sp^3\mathbf{\Sigma}^2) + \tfrac{9(n-3)(n-4)}{n(n-1)(n-2)}\cdot \mathcal{O}(Sp^2\mathbf{\Sigma}^3)$$

$$+ \tfrac{(n-3)(n-4)(n-5)}{n(n-1)(n-2)}\cdot Sp^2(\mathbf{\Sigma}^3)$$

$$= \tfrac{(n-3)(n-4)(n-5)}{n(n-1)(n-2)}\cdot Sp^2(\mathbf{\Sigma}^3) + \mathcal{O}(n^{-1})\cdot \mathcal{O}(Sp^2(\mathbf{\Sigma}^3)) + \mathcal{O}(n^{-2})\cdot \mathcal{O}(Sp^3(\mathbf{\Sigma}^2))$$

Mit $E(B_3/Sp^{\frac{3}{2}}(\mathbf{\Sigma}^2)) = Sp(\mathbf{\Sigma}^3)/Sp^{\frac{3}{2}}(\mathbf{\Sigma}^2)$ folgt für die Varianz:

$$Var(B_3/Sp^{\frac{3}{2}}(\mathbf{\Sigma}^2)) = (Sp^{\frac{3}{2}}(\mathbf{\Sigma}^2))^{-2}\cdot (E(B_3^2) - E^2(B_3))$$

$$= (Sp^3(\mathbf{\Sigma}^2))^{-1}\cdot (E(B_3^2) - Sp^2(\mathbf{\Sigma}^3))$$

$$= (Sp^3(\mathbf{\Sigma}^2))^{-1}\Bigg((\frac{(n-3)(n-4)(n-5)}{n(n-1)(n-2)} - 1)Sp^2(\mathbf{\Sigma}^3)$$

$$+ \mathcal{O}(n^{-1})\mathcal{O}(Sp^2(\mathbf{\Sigma}^3)) + \mathcal{O}(n^{-2})\mathcal{O}(Sp^3(\mathbf{\Sigma}^2)\Bigg)$$

$$= (Sp^3(\mathbf{\Sigma}^2))^{-1}\left(\frac{(n-3)(n-4)(n-5) - n(n-1)(n-2)}{n(n-1)(n-2)}Sp^2(\mathbf{\Sigma}^3)\right)$$

$$+ \mathcal{O}(n^{-1}) \cdot \mathcal{O}\Big(\underbrace{\frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)}}_{\leq 1}\Big) + \mathcal{O}(n^{-2}) \cdot 1$$

$$\leq (Sp^3(\mathbf{\Sigma}^2))^{-1}\left((\mathcal{O}(n^{-1})) \cdot Sp^2(\mathbf{\Sigma}^3)\right) + \mathcal{O}(n^{-1})$$

$$= \mathcal{O}(n^{-1})$$

□

In $\mathcal{L}_2$ - Norm bzw. in Wahrscheinlichkeit konvergiert $B_3/Sp^{3/2}(\mathbf{\Sigma}^2)$ somit gleichmäßig in d in gegen $Sp(\mathbf{\Sigma}^3)/Sp^{3/2}(\mathbf{\Sigma}^2)$.
Demensprechend konnte für den B_2-Schätzer für $Sp(\mathbf{\Sigma}^2)$ gezeigt werden, dass $B_2/Sp(\mathbf{\Sigma}^2) \xrightarrow{p} 1$ gleichmäßig in d für $n \to \infty$.

Als Schätzer für g_{pear} sei nun

$$\widehat{g}_{pear} := \frac{B_3}{B_2^{3/2}} = \underbrace{\left(\frac{B_2}{Sp(\mathbf{\Sigma}^2)}\right)^{-\frac{3}{2}}}_{\to 1} \cdot \frac{B_3}{Sp^{3/2}(\mathbf{\Sigma}^2)} \tag{7.23}$$

$$\xrightarrow{p} \frac{Sp(\mathbf{\Sigma}^3)}{Sp^{3/2}(\mathbf{\Sigma}^2)} \tag{7.24}$$

für $n \to \infty$ nach dem Slutskischen Satz und dem continuous mapping theorem. Diese lässt sich anwenden, da x^{-1} gleichmäßig stetig in 1 ist. Für $\widehat{f}_{pear}$ würde die gleiche Argumentation nicht funktionieren, da x^{-1} nicht gleichmäßig stetig bei 0 ist.

7.4.1 Schätzung von h_{pear}

Mit der gleichmäßigen Konsistenz des B_2- und des B_3-Schätzers lassen sich nun nach Satz B.0.7 auf einfachem Wege Schätzer für deren Potenzen konstruieren.

Satz 7.4.3

1. $$\widehat{Sp^3(\mathbf{\Sigma}^2)} := \frac{1}{n(n-1)\cdots(n-5)} \sum_{k\neq l\neq k'\neq l'\neq k''\neq l''}^{n} A_{kl}^2 A_{k'l'}^2 A_{k''l''}^2$$

2. $$\widehat{Sp^2(\mathbf{\Sigma}^3)} := \frac{1}{n(n-1)\cdots(n-5)} \sum_{k\neq l\neq r\neq k'\neq l'\neq r'} A_{kl}A_{lr}A_{rk}A_{k'l'}A_{l'r'}A_{r's'}$$

sind erwartungstreue Schätzer für die entsprechenden Parameter und es gilt

$$\frac{\widehat{Sp^3(\mathbf{\Sigma}^2)}}{Sp^3(\mathbf{\Sigma}^2)} \xrightarrow{p} 1 \qquad\qquad \frac{\widehat{Sp^2(\mathbf{\Sigma}^3)}}{Sp^3(\mathbf{\Sigma}^2)} \xrightarrow{p} \frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)}$$

gleichmässig in d für $n \to \infty$

Beweis: Für die Schätzer B_2, B_3 konnte gezeigt werden, dass diese erwartungstreu sind und $B_2/Sp(\mathbf{\Sigma}^2)$ gleichmässig in der Dimension gegen 1 konvergiert bzw. $B_3/Sp^{3/2}(\mathbf{\Sigma}^2)$ gegen $Sp(\mathbf{\Sigma}^3)/Sp^{3/2}(\mathbf{\Sigma}^2)$. Diese Eigenschaften übertragen sich nach Satz B.0.7 nun auf $\widehat{Sp^3(\mathbf{\Sigma}^2)}$ und $\widehat{Sp^2(\mathbf{\Sigma}^3)}$.

□

Analog zu (7.23) lässt sich ein Schätzer für h_{pear} bilden.

$$\widehat{h}_{pear} := \frac{\widehat{Sp^2(\mathbf{\Sigma}^3)}}{\widehat{Sp^3(\mathbf{\Sigma}^2)}} = \underbrace{\left(\frac{\widehat{Sp^3(\mathbf{\Sigma}^2)}}{Sp^3(\mathbf{\Sigma}^2)}\right)^{-1}}_{\to 1} \cdot \frac{\widehat{Sp^2(\mathbf{\Sigma}^3)}}{Sp^3(\mathbf{\Sigma}^2)} \tag{7.25}$$

$$\xrightarrow{p} \frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)} \tag{7.26}$$

für $n \to \infty$.

Der Schätzer besitzt im Gegensatz zu $\widehat{g}_{pear}$ die schöne Eigenschaft, dass Zähler und Nenner unverzerrt sind. Demgegenüber steht allerdings der Nachteil, dass $\widehat{Sp^3(\mathbf{\Sigma}^2)}$ und $\widehat{Sp^2(\mathbf{\Sigma}^3)}$ eine recht große Varianz haben und überhaupt erst ab einem Stichprobenumfach $n \geq 6$ berechenbar sind. Ein Vergleich der Schätzer findet in Abschnitt 11.3.3 der Simulationen statt.

7.5 Pearsonapproximation für Z_n

Seien die $\mathbf{Y}_k$, $k = 1,..,n$ verteilt nach dem Bai-Saranadasa Modell mit Kovarianzmatrix $\mathbf{\Sigma}$. Mit dem Schätzer $\widehat{g}_{pear}$ für g_{pear} lässt sich dann analog zur $\varkappa(f_{box})$-Approximation aus 6.4 die Pearsonapproximation für Z_n implementieren. Dabei soll nun $\kappa(\widehat{g}_{pear})$ die Normal-Approximation auch im Hochdimensionalen ersetzen. Die $\kappa(\widehat{g}_{pear})$- Approximation entspricht analytisch genau einer $\varkappa(f_{pear})$-Approximation, wobei f_{pear} mittels

$$\widehat{f}_{pear} := \frac{B_2^3}{B_3^2} \tag{7.27}$$

geschätzt wird. Die abgeänderte Notation über $\kappa(g)$ soll lediglich darauf hinweisen, dass f_{pear} (im Gegensatz zu g_{pear}) formal nicht stabil in der Dimension geschätzt wird. Um nun die hochdimensionalen Resultate von Chen-Qin anwenden zu können, wird der folgende Satz benötigt.

Satz 7.5.1 (Gemeinsame Asymptotik von g_{pear} und d)
Die Chen-Qin-Bedingung an die Kovarianzmatrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ formuliert sich mit

$$h_{CQ} := \frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \tag{7.28}$$

als $h_{CQ} \to 0 \quad$ für $d \to \infty$. Desweiteren sei

$$h_{pear} = \frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)} \tag{7.29}$$

und $g_{pear} = \sqrt{h_{pear}}$. Dann gilt

$$h_{CQ} \xrightarrow{d\to\infty} 0 \qquad \Leftrightarrow \qquad h_{pear}, g_{pear} \xrightarrow{d\to\infty} 0 \tag{7.30}$$

Beweis: Das Resultat folgt direkt aus Satz 8.1.2 und Korrolar 8.2.2 in Kapitel 8.

Analog zu Abschnitt 5.3.1 der Boxapproximation lässt sich nun folgern, dass die Chen-Qin-Bedingung sicherstellt, dass $\kappa(g_{pear})$ für $d \to \infty$ die Grenzverteilung

$$\lim_{d\to\infty} \kappa(g_{pear}) = \mathcal{N}(0,1) \tag{7.31}$$

annimmt. Damit lässt sich nun unabhängig von der Dimension folgende neue Teststatistik für das Ein-Gruppen-Design formulieren.

$$Z_n = \frac{\sum_{k>l}^{n} A_{kl}}{\sqrt{\sum_{k>l}^{n} A_{kl}^2}} \dot{\sim} \kappa(\widehat{g}_{pear}) \qquad \text{wobei} \qquad \widehat{g}_{pear} = \frac{B_3}{B_2^{(3/2)}} \tag{7.32}$$

7.5.1 Vergleich mit anderen Teststatistiken

Der neu formulierte Test vereint die Theorie der ANOVA-Typ-Statistiken mit den Vorteilen der Resultate von Bai-Sarandasa und Chen-Qin für hochdimensionale, nichtnormale Designs. Der Hauptvorteil ist es, dadurch robuste Resultate zu erzielen ohne Restriktionen zwischen der Aymptotik über die Dimension d und den Stichrobenumfang n hinnehmen zu müssen. So lässt sich für einen hinreichend großen Stichprobenumfang ($n \geq 10$) folgende Tabelle über die Annahmen der verschiedenen Testverfahren angeben.

		Multivariate Normalverteilung	Bai-Saranadasa Modell
Niedrigdimensional		✓ - ✓	✓ - ✓
Hochdim.	$\frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \to 0$	✓ ✓ ✓	- ✓ ✓
	$\frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \gg 0$	✓ - ✓	- - -*

Tabelle 7.1: Aufgeführt ist die Anwendbarkeit der verschiedenen Testverfahren bei unterschiedlichen Verteilungsannahmen. Das Zeichen ✓ bedeutet, dass ein jeweiliges Testverfahren zur Anwendung kommen darf. Demgegenüber induziert ein Minus-Zeichen, dass das Verfahren in der jeweiligen Parameterkombination nicht benutzt werden darf bzw. die theoretische Grundlage dafür fehlt.
Das oberste Zeichen der Dreierblöcke bezieht sich auf die ANOVA-Typ-Statistik nach Werner-Brunner, das mittlere auf den Test nach Chen-Qin, während das untere für das neue Testverfahren (7.32) steht.
* Obwohl die Anwendbarkeit nicht gezeigt wurde liefert der neue Test in Simulationen brauchbare Ergebnisse.

7.5.2 Nichtnormale Blockeffekte

Einzig nichtnormale, hochdimensionale Modelle, welche die Chen-Qin-Bedingung nicht erfüllen, stellen ein Problem bei der Auswertung dar. Ziel ist es deshalb, die Annahmen an die Verteilung weiter zu relaxieren und neue Robustheitsaussagen zu gewinnen, falls für $d \to \infty$ gerade

$$\frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \nrightarrow 0 \tag{7.33}$$

Dieser Fall tritt bei einzelnen großen Eigenwerten $\lambda_i \notin \mathcal{O}(\sqrt{d})$ auf. Meist ist dies auf große Blockeffekte der Verteilung der $\mathbf{X}_k$ zurückzuführen. Blockeffekte werden durch eine (unabhängige) Zufallsvariable beschrieben, welche auf sämtliche Komponenten der $\mathbf{X}_k$ addiert wird. Im Bai-Saranadasa Modell werden diese Blockeffekte also durch einzelne Z_i beschrieben. Diese Effekte können häufig nicht als normalverteilt betrach-

tet werden. Außerdem verhalten sie sich im Bezug auf die Auswertung sehr dominant, da ihre Ausprägung (fast) jede Komponente beeinflusst. Demgegenüber sind sie allerdings in den gewöhnlichen Modellierungen in ihrer Anzahl beschränkt. Dies motiviert dazu, ein Modell zu betrachten, welches sich als Hybrid von multivariat normalen Zufallsvariablen und einer verhältnismäßig kleinen Anzahl $e \ll m$ von nichtnormalverteilten Zufallsvariablen beschreiben lässt. Bezogen auf das Bai-Saranadasa Modell $\mathbf{X}_k = \boldsymbol{\Upsilon}\mathbf{Z}_k + \boldsymbol{\mu}$ seien diese oBdA Z_1,..,Z_e. Dann lässt sich der folgende Satz formulieren.

Satz 7.5.2 *Sei eine Abwandlung des Bai-Saranadasa Modells beschrieben durch*

$$\mathbf{X}_k = \boldsymbol{\Upsilon} \cdot \mathbf{Z}_k + \boldsymbol{\mu} \quad \textit{für } k = 1, \ldots, n \tag{7.34}$$

wobei $\boldsymbol{\Upsilon}$ *eine* d x m*-Matrix* ($m \geq d$)*. Die Vektoren* $\mathbf{Z}_k = (Z_{1k}, \ldots Z_{mk})'$ *seien unabhängig identisch verteilt mit* $E(\mathbf{Z}_k) = \mathbf{0}_m$*,* $Cov(\mathbf{Z}_k) = \mathbf{I}_m$*. Sämtliche Komponenten* Z_{ik} *sollen desweiteren unabhängig sein. Sei* $e < m$ *und es gelte die Restriktion an die gemeinsame Asymptotik von* e *und* n

$$\frac{e^2}{n} \to 0 \tag{7.35}$$

für $n, e \to \infty$*. Die Zufallsvariablen* Z_{1k},..,Z_{ek} *sollen endliche vierte Momente* $E(Z_{ik}^4) \leq \mu_4 < \infty$ *besitzen, während die restlichen Zufallsvariablen* Z_{e+1},..,Z_m *standardnormalverteilt seien. Dann gilt für* $n \to \infty$

$$\sqrt{n}\left(\overline{\mathbf{X}}_{\cdot} - \boldsymbol{\mu}\right) \;\dot{\sim}\; \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}') \tag{7.36}$$

Beweis: Zunächst wird gezeigt, dass

$$\widetilde{\mathbf{X}} := \sqrt{n}\left(\overline{\mathbf{X}}_{\cdot} - \boldsymbol{\mu}\right) \tag{7.37}$$

genau dann asymptotisch multivariat normal-verteilt ist, wenn

$$\widetilde{\mathbf{Z}} := \sqrt{n} \cdot \overline{\mathbf{Z}}_{\cdot} \tag{7.38}$$

asymptotisch multivariat normal ist. Sei $A \in \mathbb{R}^d$ eine borelmessbare Menge. Dann ist das Urbild $\boldsymbol{\Upsilon}^{-1}(A)$ bezüglich der Multiplikation mit $\boldsymbol{\Upsilon}$ ebenfalls messbar. Mit $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_m, \mathbf{I}_m)$ gilt

$$P(\widetilde{\mathbf{X}} \in A) = P(\boldsymbol{\Upsilon} \cdot \widetilde{\mathbf{Z}} \in A) = P(\widetilde{\mathbf{Z}} \in \boldsymbol{\Upsilon}^{-1}(A)) \tag{7.39}$$

$$\doteqdot P(\mathbf{Z} \in \boldsymbol{\Upsilon}^{-1}(A)) = P(\boldsymbol{\Upsilon} \cdot \mathbf{Z} \in A) \tag{7.40}$$

$$= P^{\mathcal{N}(\mathbf{0}_m, \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}')}(A) \tag{7.41}$$

Es reicht also zu zeigen, dass

$$\widetilde{\mathbf{Z}} \xrightarrow{w} \mathcal{N}(\mathbf{0}, \mathbf{I}_m) \tag{7.42}$$

konvergiert. Zunächst gilt für die wegen der Unabhängigkeit der Komponenten $\widetilde{Z}_i = \overline{Z}_{i\cdot}$ für die Verteilungsfunktion.

$$F^{\widetilde{\mathbf{Z}}}(\mathbf{a}) = \prod_{i=1}^{m} F^{\widetilde{Z}_i}(a_i) = \prod_{i=1}^{e} F^{\widetilde{Z}_i}(a_i) \cdot \prod_{i=e+1}^{m} \Phi(a_i) \tag{7.43}$$

Dabei ist $\Phi(x)$ die Verteilungsfunktion einer standard-normalverteilten Zufallsvariable. Nach dem Satz von Berry-Esseen ist

$$\left| F^{\widetilde{Z}_i}(x) - \Phi(x) \right| \leq \frac{C \cdot E|\widetilde{Z}_i^3|}{\sigma_{\widetilde{Z}_i}^3 \cdot \sqrt{n}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{7.44}$$

da die Konstante $C \leq 0{,}7655$, $\sigma_{\widetilde{Z}_i} = 1$ und $E|\widetilde{Z}_i^3| = \mathcal{O}(\mu_4)$ sind. Mit $\widetilde{\mathbf{a}} := (a_1,..,a_e)$ folgt somit

$$\prod_{i=1}^{e} F^{\widetilde{Z}_i}(a_i) = \prod_{i=1}^{e} \left(\Phi(a_i) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \right) \tag{7.45}$$

$$= \underbrace{\prod_{i=1}^{e} \Phi(a_i)}_{=\Phi_{\mathbf{0},\mathbf{I}_e}(\widetilde{\mathbf{a}})} + \sum_{j=1}^{e} \left(\mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \underbrace{\prod_{i \neq j}^{e} \Phi(a_i)}_{\leq 1} \right) \tag{7.46}$$

$$+ \sum_{j=1}^{e} \sum_{h>j}^{e} \left(\mathcal{O}\left(\frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}}\right) \underbrace{\prod_{i \neq j \neq h}^{e} \Phi(a_i)}_{\leq 1} \right) \tag{7.47}$$

$$+ ... + \mathcal{O}\left(\left(\tfrac{1}{\sqrt{n}}\right)^e \right) \tag{7.48}$$

$$= \Phi_{\mathbf{0},\mathbf{I}_e}(\widetilde{\mathbf{a}}) + \mathcal{O}\left(\frac{e}{\sqrt{n}}\right) + \mathcal{O}\left(\left(\tfrac{e}{\sqrt{n}}\right)^2\right) + \mathcal{O}\left(\left(\tfrac{e}{\sqrt{n}}\right)^3\right) + ... \tag{7.49}$$

$$= \Phi_{\mathbf{0},\mathbf{I}_e}(\widetilde{\mathbf{a}}) + \mathcal{O}\left(\sum_{i=1}^{\infty}\left(\tfrac{e}{\sqrt{n}}\right)^i\right) \tag{7.50}$$

$$= \Phi_{\mathbf{0},\mathbf{I}_e}(\widetilde{\mathbf{a}}) + \mathcal{O}\Big(\underbrace{\frac{e}{\sqrt{n}}}_{\to 0} \Big/ \Big(\underbrace{1 - \frac{e}{\sqrt{n}}}_{\to 1}\Big)\Big) \tag{7.51}$$

$$\doteqdot \Phi_{\mathbf{0},\mathbf{I}_e}(\widetilde{\mathbf{a}}) \tag{7.52}$$

$$\tag{7.53}$$

für $n \to \infty$, da e klein genug ist, so dass $e^2/n \to 0$. Somit ist

$$F^{\widetilde{\mathbf{Z}}}(\mathbf{a}) \doteqdot \Phi_{\mathbf{0},\mathbf{I}_m}(\mathbf{a}) \tag{7.54}$$

und es folgt die Behauptung.

□

Es sei bemerkt, dass dieser Satz im Fall $e = m$ einen Spezialfall der Ergebnisse von Portnoy (1986) [16] darstellt.
Mit diesem Resultat lässt sich das neue Testverfahren auf weitere Fälle anwenden, welche die Voraussetzungen für Tests nach Chen und Qin bzw. der ANOVA-Typ-Statistik nicht erfüllen.

8 Restriktionen der Freiheitsgrade

8.1 Beschränkungen bezüglich der Dimension

Im folgenden Abschnitt sollen Einschränkungen zwischen der Dimension und den Freiheitsgraden f_{box} und f_{pear} untersucht werden. Grundlage ist folgendes elementare Resultat.

Satz 8.1.1 *Sei $\mathbf{\Sigma}$ eine positiv semi-definite dxd-Matrix mit Eigenwerten λ_i, $i = 1,..d$. Dann gilt*

$$f_{box} = \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \leq Rang(\mathbf{\Sigma}) \tag{8.1}$$

Beweis: Sei $a := Rang(\mathbf{\Sigma})$ die Anzahl der Eigenwerte größer null. ObdA seien für $i = 1,..,a$ sämtliche $\lambda_i > 0$ und $\lambda_i = 0$ für $i > a$. Aus der Jensen-Ungleichung folgt mit dem Erwartungswert bezüglich gleicher Gewichte $1/a$ der λ_i, $i = 1,..,a$ und mit der konvexen Funktion $f(x) = x^2$

$$\left(\sum_{i=1}^{a} \frac{1}{a}\lambda_i\right)^2 \leq \sum_{i=1}^{a} \frac{1}{a}\lambda_i^2 \quad \Rightarrow \left(\sum_{i=1}^{d} \lambda_i\right)^2 \leq a \cdot \sum_{i=1}^{d} \lambda_i^2 \quad \Rightarrow \frac{Sp^2(\mathbf{\Sigma})}{Sp(\mathbf{\Sigma}^2)} \leq a \tag{8.2}$$

□

Diese Abschätzung lässt sich auf den Pearsonschen Freiheitsgrad und die Chen-Qin-Bedingung fortsetzen.

Satz 8.1.2 (Ordnung der Freiheitsgrade) *Seien für eine positiv-semidefinite Matrix $\mathbf{\Sigma}$*

1. $$a = Rang(\mathbf{\Sigma})$$

2. $$h_{box} = \frac{Sp(\mathbf{\Sigma}^2)}{Sp^2(\mathbf{\Sigma})}$$

3.

$$h_{pear} = \frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)}$$

4.

$$h_{CQ} = \frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)}$$

Dann gilt

$$\frac{1}{a} \leq h_{box} \leq h_{pear} \leq h_{CQ} \tag{8.3}$$

Beweis:

1.) $\frac{1}{a} \leq h_{box}$ wurde in Satz 8.1.1 gezeigt.

2.) $h_{box} \leq h_{pear}$

Zunächst ist nach Lemma C.1.1 $Sp^2(\mathbf{\Sigma}^2) \leq Sp(\mathbf{\Sigma}) \cdot Sp(\mathbf{\Sigma}^3)$

$$\Rightarrow Sp^4(\mathbf{\Sigma}^2) \leq Sp^2(\mathbf{\Sigma}) Sp^2(\mathbf{\Sigma}^3) \quad \Rightarrow \frac{Sp(\mathbf{\Sigma}^2)}{Sp^2(\mathbf{\Sigma})} \leq \frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)}$$

3.) $h_{pear} \leq h_{CQ}$

Nach Lemma C.1.1 gilt ebenfalls $Sp^2(\mathbf{\Sigma}^3) \leq Sp(\mathbf{\Sigma}^2) \cdot Sp(\mathbf{\Sigma}^4)$

$$\Rightarrow \frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)} \leq \frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)}$$

□

Insbesondere folgt für Kontrastmarizen als Hypothesenmatrix, dass

$$a \leq d - 1 \tag{8.4}$$

Auf diese Weise erhält man eine Einschränkung des zulässigen Bereiches von h. Dieser lässt sich nutzen, um die Schätzung des Freiheitsgrades h_{box} und g_{pear} zu verbessern,

indem man den Schätzer beschneidet. Außerdem sind wegen $Sp(\mathbf{\Sigma}^2) \leq Sp^2(\mathbf{\Sigma})$ und $Sp^2(\mathbf{\Sigma}^3) \leq Sp^3(\mathbf{\Sigma}^2)$ nach Satz C.1.1 die Parameter $g, h \leq 1$. Sei a der Rang der Hypothesenmatrix $\mathbf{H}$, welche $\mathbf{\Sigma}$ gemäß Abschnitt 3.3 erzeugt. Dann ist:

$$\widetilde{h}_{box} := max(1/a,\, min(1, \widehat{h}_{box})) \tag{8.5}$$

$$\widetilde{g}_{pear} := max(1/\sqrt{a}, min(1, \widehat{g}_{pear})) \tag{8.6}$$

Die dadurch induzierte Verzerrung der Schätzer ist kein Problem, da die Konsistenz hiervon nicht beeinflusst wird und außerdem eine Erhöhung von $\widehat{h}$ bzw. $\widehat{g}$ zu konservativen Ergebnissen führt, wie in Abbildung 7.2 ersichtlich.

8.2 Verifizierbarkeit der Chen-Qin-Bedingung

Ein weiterer großer Vorteil der Approximation nach Pearson ist neben der Güte der Approximation der starke Bezug zwischen dem reparametrisierten Freiheitsgrad h_{pear} und der Chen-Qin-Bedingung h_{CQ}. So lässt sich zeigen, dass

$$h_{pear} \xrightarrow{d\to\infty} 0 \qquad \Leftrightarrow \qquad h_{CQ} \xrightarrow{d\to\infty} 0 \tag{8.7}$$

Satz 8.2.1 *Sei $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ eine positiv semi-definite Matrix und $\delta < \gamma$ Konstanten. Dann gilt*

$$\frac{Sp(\mathbf{\Sigma}^{2+\delta})}{Sp^{(1+\frac{\delta}{2})}(\mathbf{\Sigma}^2)} \geq \frac{Sp(\mathbf{\Sigma}^{2+\gamma})}{Sp^{(1+\frac{\gamma}{2})}(\mathbf{\Sigma}^2)}$$

Beweis:

1. Zunächst ist $Sp(\mathbf{\Sigma}^{a\cdot b}) = \sum_{i=1}^{d} \lambda^{a\cdot b} \leq \left(\sum_{i=1}^{d} \lambda^a\right)^b = Sp^b(\mathbf{\Sigma}^a)$ für a,$b \geq 1$ und somit:

$$\frac{Sp(\mathbf{\Sigma}^{2+\delta})}{Sp^{(1+\frac{\delta}{2})}(\mathbf{\Sigma}^2)} \geq \underbrace{\left(\frac{Sp(\mathbf{\Sigma}^{2+\delta})}{Sp^{(1+\frac{\delta}{2})}(\mathbf{\Sigma}^2)}\right)}_{\leq 1}^{\underbrace{\frac{1+\frac{\gamma}{2}}{1+\frac{\delta}{2}}}_{>1}} = \frac{(Sp(\mathbf{\Sigma}^{2+\delta}))^{\frac{2+\gamma}{2+\delta}}}{Sp^{(1+\frac{\gamma}{2})}(\mathbf{\Sigma}^2)} \geq \frac{Sp(\mathbf{\Sigma}^{(2+\delta)\cdot\frac{2+\gamma}{2+\delta}})}{Sp^{(1+\frac{\gamma}{2})}(\mathbf{\Sigma}^2)}$$

$$= \frac{Sp(\mathbf{\Sigma}^{2+\gamma})}{Sp^{(1+\frac{\gamma}{2})}(\mathbf{\Sigma}^2)}$$

□

Korollar 8.2.2 *Mit den Voraussetzungen aus 8.2.1 folgt explizit:*

$$\frac{Sp(\mathbf{\Sigma}^3)}{Sp^{\frac{3}{2}}(\mathbf{\Sigma}^2)} \xrightarrow{d\to\infty} 0 \quad \Rightarrow \quad \frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \xrightarrow{d\to\infty} 0$$

Beweis:
Folgt aus Satz 8.2.1 mit $\delta = 1$ und $\gamma = 2$:

$$\frac{Sp(\mathbf{\Sigma}^4)}{Sp^2(\mathbf{\Sigma}^2)} \leq \frac{Sp(\mathbf{\Sigma}^3)}{Sp^{\frac{3}{2}}(\mathbf{\Sigma}^2)} \xrightarrow{d\to\infty} 0$$

□

Somit gilt:

$$h_{pear} \leq h_{CQ} \leq g_{pear} = \sqrt{h_{pear}} \tag{8.8}$$

Dieses Resultat ermöglicht es mittels g_{pear} die Chen-Qin-Bedingung direkt zu verifizieren. In der Praxis bietet der Schätzer $\widehat{g}_{pear}$ eine approximative Überprüfung.
Desweiteren folgt, dass für

$$h_{pear} = \frac{Sp^2(\mathbf{\Sigma}^3)}{Sp^3(\mathbf{\Sigma}^2)} \xrightarrow{d\to\infty} 0 \tag{8.9}$$

auch $\sqrt{h_{pear}} \to 0$ und somit $h_{CQ} \to 0$. Mit der Rückrichtung aus $h_{pear} \leq h_{CQ}$ folgt die Äquivalenz zur Chen-Qin-Bedingung. Diese Äquivalent überrascht nicht, da die Chen-Qin-Bedingung benötigt wurde, um die Liapounov-Bedingung des Zentralen Grenzwertsatzes für Martingale, nach Korollar 3.1 in Hall and Heyde (1980) [12] nachzuweisen. Dieser Nachweis ist dementsprechend ebenfalls mit der alternativen (Liapounov-) Bedingung $Sp(\mathbf{\Sigma}^3)/Sp^{3/2}(\mathbf{\Sigma}^2) \to 0$ möglich.

Mit einem konsistenten Schätzer für h_{pear} lässt sich somit bei hinreichend großem Stichprobenumfang prüfen, ob die Voraussetzungen des neuen Tests erfüllt sind. Auf eine Schätzung von $Sp(\mathbf{\Sigma}^4)$ kann somit verzichtet werden.

9 Grenzen des Bai-Saranadasa Modells

Das Bai-Saranadasa Modell und die damit unter Umständen verbundene Annahme an die Kovarianzmatrix bieten eine weitreichende Möglichkeit der Modellierung der $\mathbf{X}_k$. Diese Annahmen sind allerdings essentiel und nicht beliebig verallgemeinbar. Dazu wird in diesem Kapitel eine Klasse von degenerierten Verteilungen jenseits des Bai-Saranadasa Modells vorgestellt. Unter dieser Verteilung ist eine stabile Schätzung der Spuren mittels der vorgestellten Schätzer nicht mehr gegeben. Des Weiteren wird eine Verteilung konstruiert, so dass die angegebenen Testverfahren stark liberal werden.

9.1 Konstruktion einer degenerierten Verteilung

Während im Bai-Saranadasa Modell die $Z_i,\ i = 1,..,m$ die Abhängigkeiten mit $\mathbf{\Gamma}$ schon voll beschrieben werden, ist es nun Ziel, stärkere Abhängigkeiten zwischen den d Komponenten zu finden, die nicht durch die Kovarianzmatrix $\mathbf{\Sigma}$ gegeben sind. Wenn man sogar fordert, dass sämtliche Komponenten unkorreliert sein sollen, ist es möglich die einzelnen Komponenten so zu konstruieren, dass diese extrem stark abhängig sind. Die multivariate Verteilung ist in diesem Fall stark degeneriert. Eine solch degenerierte Verteilung kann mit beliebigen Randverteilungen erzeugt werden.
Formal beschreibt sich die Randverteilung der ersten Komponente X_1 als eine Zufallsvariable mit zugehörigem Wahrscheinlichkeitsraum Ω_1 und $E(X_1) = 0$. Die übrigen Komponenten $X_i,\ i = 2,..,d$ lassen sich nun aus der ersten Komponente X_1 ableiten, indem das Vorzeichen randomisiert wird, während der Betrag von X_1 übernommen wird.

Die Randomisierung der Vorzeichen erfolgt dabei nach dem folgenden Schema:
Sei $d = 2^b$ eine Potenz von 2 und sei $\mathbf{B} \in \mathbb{R}^d$ ein Zufallsvektor der an einer zufälligen Stelle 1 ist und sonst überall 0. Somit ist $\mathbf{B}$ eine zufällige Positionsziehung mit gleichen Wahrscheinlichkeiten $1/d$ für jede Position der 1. (Formal folgt $\mathbf{B}$ somit einer Multinomialverteilung für eine Ziehung $n = 1$ und den partiellen Wahrscheinlichkeiten $p_i = 1/d$ für alle $i = 1,..,d$.)
Desweiteren soll $\mathbf{B}$ unabhängig zu X_1 sein. Mit Wahrscheinlichkeitsraum Ω_2 von $\mathbf{B}$ ergibt sich als gemeinsamer Wahrscheinlichkeitsraum $\Omega := \Omega_1 \times \Omega_2$ für einen Vektor aus der Stichprobe $\mathbf{X}_k$.

Die Schwierigkeit besteht nun darin, geeignete Vorzeichen in Abhängigkeit von B zu finden, so dass sämtliche Komponenten unkorreliert sind. Dafür lassen sich Eigenschaften der sogenannten Hadamardmatrizen nutzen, welche beispielsweise in Kapitel 8.10 in Schott (2005) [18] eingeführt werden.

Definition 9.1.1 (Hadamardmatrizen) *Sei die Hadamardmatrix der Dimension* 1 *als* $\mathbf{H}_1 = 1$ *und die Hadamardmatrix der Dimension zwei als*

$$\mathbf{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{9.1}$$

gegeben. Dann lassen sich rekursiv Hadamardmarizen der Dimension $d = 2^b$ *für alle* $b \geq 0$ *definieren als*

$$\mathbf{H}_d = \mathbf{H}_{2^{b-1}} \otimes \mathbf{H}_2 = \bigotimes_{s=1}^{b} \mathbf{H}_2 \tag{9.2}$$

Damit ergibt sich beispielsweise für $d = 4$

$$\mathbf{H}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \tag{9.3}$$

Selbst wenn d keine Potenz von zwei ist, existieren häufig entsprechende Hadamardmatrizen. Im Folgenden soll der Einfachheit halber aber weiterhin $d = 2^b$ angenommen werden.
Hadamardmatrizen besitzen nun die Eigenschaft, dass sie symmetrisch sind und sämtliche Zeilen bzw. Spalten paarweise orthogonal sind. So sei die i-te Spalte von $\mathbf{H}_d$ mit $\mathbf{g}_i$ bezeichnet und der i-te Einheitsvektor mit $\mathbf{e}_i$.

$$\mathbf{g}_i' \cdot \mathbf{g}_j = \left(\bigotimes_{s=1}^{b} \mathbf{H}_2 \cdot \mathbf{e}_i \right)' \cdot \left(\bigotimes_{s=1}^{b} \mathbf{H}_2 \cdot \mathbf{e}_j \right) = \mathbf{e}_i' \left(\bigotimes_{s=1}^{b} \mathbf{H}_2 \right)' \left(\bigotimes_{s=1}^{b} \mathbf{H}_2 \right) \mathbf{e}_j$$

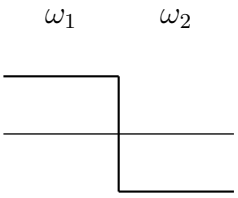
$$= \mathbf{e}_i' \left(\bigotimes_{s=1}^{b} \mathbf{H}_2' \cdot \mathbf{H}_2 \right) \mathbf{e}_j = \mathbf{e}_i' \, \mathbf{I}_d \, \mathbf{e}_j \qquad \text{da } \mathbf{H}_2 \text{ orthogonal ist}$$

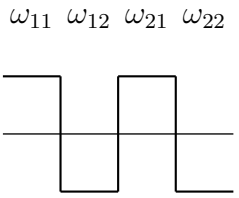
$$= \begin{cases} 1, & \text{wenn } i = j \\ 0, & \text{wenn } i \neq j \end{cases}$$

Mit dieser Überlegung lassen sich nun die $X_i : \Omega \to \mathbb{R}$ für $i = 1,..,d$ definieren:

$$X_i = X_1 \cdot (\mathbf{g}_i'\mathbf{B}) \tag{9.4}$$

Dies ist auch für X_1 wohldefiniert, da $\mathbf{g}_1 = \mathbf{1}_d$ und somit $\mathbf{g}_1'\mathbf{B} = 1$ ist.
Durch die Vorzeichenstruktur der Hadamardmatrizen sehen die $(\mathbf{g}_i'\mathbf{B})(\omega)$ nun in Abhängigkeit der Ausprägung der zufälligen Ziehung $\omega_i \in \Omega_2$ wie folgt aus:





Die linke Skizze stellt dabei die Ausprägungen von $(\mathbf{g}_2'\mathbf{B})$ für die Dimension zwei dar. Verfeinert man die Struktur mittels dem Übergang von Dimension zwei auf Dimension vier, dann wird die linke Skizze aus dem dritten Vektor von $\mathbf{H}_4$ erzeugt. Dies entspricht der Funktion $(\mathbf{g}_3'\mathbf{B})$ in Anhängigkeit der Ziehung von $\mathbf{B}$ während die rechte Skizze $(\mathbf{g}_2'\mathbf{B})$ darstellt. ω_1 steht dabei für eine Positionsziehung von 1 bis $d/2$, ω_2 für die übrigen. Induktiv steht ω_{11} dann für eine Position aus der ersten Hälfte der ersten Hälfte, analog zur Konstruktion der Hadamardmatrizen.

Wie man anhand dieser Abbildung sehr gut sehen kann, ergibt sich nun also, dass die einzelnen Komponenten der Vektoren in der Hälfte der Fälle positiv korrelieren und in der anderen Hälfte negativ. Man berechnet die Kovarianz zwischen X_i und X_j wie folgt:

$$Cov(X_i, X_j) = E(X_i \cdot X_j) \qquad \text{da } E(X_1) = 0 \tag{9.5}$$

$$= E\left(X_1^2 \cdot (\mathbf{g}_i'\mathbf{B}) \cdot (\mathbf{g}_j'\mathbf{B})\right) \tag{9.6}$$

$$= Var(X_1) \cdot E\left((\mathbf{g}_i'\mathbf{B})' \cdot (\mathbf{g}_j'\mathbf{B})\right) \tag{9.7}$$

$$= Var(X_1) \cdot E(\mathbf{B}'\mathbf{g}_i\mathbf{g}_j'\mathbf{B}) \tag{9.8}$$

$$= Var(X_1) \cdot \sum_{h=1}^{d} \mathbf{e}_h' \, \mathbf{g}_i \, \mathbf{g}_j' \, \mathbf{e}_h \cdot \underbrace{P(\mathbf{B} = \mathbf{e}_h)}_{=1/d} \tag{9.9}$$

$$= \frac{Var(X_1)}{d} \cdot Sp(\, \mathbf{g}_i \cdot \mathbf{g}_j') \tag{9.10}$$

$$= \frac{Var(X_1)}{d} \cdot Sp(\, \mathbf{g}_i' \cdot \mathbf{g}_j) \tag{9.11}$$

$$= \begin{cases} Var(X_1), & \text{wenn } i = j \\ 0, & \text{wenn } i \neq j \end{cases} \tag{9.12}$$

Somit wurde ein Zufallsvektor $\mathbf{X}$ konstruiert, welcher als Kovarianzmatrix die Einheitsmatrix besitzt und bei dem sämtliche Ränder identisch verteilt sind. Betrachte

$$\mathbf{X} = (X_1,..,X_d)' = X_1 \cdot (\mathbf{g}_1'\mathbf{B},..., \mathbf{g}_d'\mathbf{B})' \tag{9.13}$$

$$= X_1 \cdot (\mathbf{g}_1',..., \mathbf{g}_d')' \cdot \mathbf{B} \tag{9.14}$$

$$= X_1 \cdot \mathbf{H}_d \cdot \mathbf{B} \qquad \text{da } \mathbf{H}_d \text{ symmetrisch ist} \tag{9.15}$$

Der Vektor $\mathbf{X}$ ist dementsprechend nur eine zufällige Auswahl einer Spalte der Hadamardmatrix mit einer dazu unabhängigen, zufälligen Skalierung X_1.

9.2 Auswirkungen auf die Statistik

Seien die $\mathbf{X}_k$ für $k = 1,..,n$ unabhängig identisch verteilt nach der degenerierten Verteilung

$$\mathbf{X}_k = Z_k \cdot \mathbf{H}_d \cdot \mathbf{B}_k \tag{9.16}$$

wobei Z_k eine Zufallsvariable ist, $\mathbf{H}_d$ die d-dimensionale Hadamardmatrix und $\mathbf{B}_k$ eine gleichverteilte Ziehung aus der Menge der Einheitsvektoren $\{\mathbf{e}_i : i = 1,...d\}$.

Dann gilt für die Bilinearfomen A_{kl}

$$A_{kl} = \mathbf{X}_k' \cdot \mathbf{X}_l = \begin{cases} d \cdot Z_k \cdot Z_l, & \text{mit Wahrscheinlichkeit } \frac{1}{d} \\ 0, & \text{sonst} \end{cases} \tag{9.17}$$

da mit Wahrscheinlichkeit $1/d$ die Indizes k und l übereinstimmen. Ansonsten ist die Bilinearform aufgrund der Orthogonalität der Hadamardmatrizen 0. Für die Verteilung der A_{kl} sind dadurch Momente charakteristisch, die sehr stark mit der Dimension wachsen. So ergibt sich als i-tes zentrales Moment.

$$E(A_{kl}^i) = d^i \cdot E(Z_1^i) \cdot E(Z_1^i) \cdot \frac{1}{d} = d^{i-1} \cdot E^2(Z_1^i) \tag{9.18}$$

9.2.1 Vergleich mit dem Bai-Saranadasa Modell

Dazu seien im Vergleich die Momente der A_{kl} im Bai-Saranadasa Modell dargestellt, wenn die Kovarianzmatrix die Identität ist und sämtliche Ränder standardnormal verteilt sind.

	Bai-Saranadasa Modell	Degeneriertes Modell
Design	$\mathbf{X}_k = \mathbf{I}_d \cdot \mathbf{Z}_k$	$\mathbf{X}_k = Z_k \cdot \mathbf{H}_d \cdot \mathbf{B}$
Verteilung	$A_{kl} \xrightarrow[d\to\infty]{ZGWS} \mathcal{N}(0,d)$	$A_{kl} = \begin{cases} d \cdot Z_k Z_l, & \text{mit Wk } \frac{1}{d} \\ 0, & \text{sonst} \end{cases}$
$E(A_{kl})$	0	0
$E(A_{kl}^2)$	d	d
$d^{-\frac{3}{2}} E\left\|A_{kl}\right\|^3$	$\sqrt{8/\pi}$	$\sqrt{d}$
$d^{-2}E(A_{kl}^4)$	3	d

Tabelle 9.1: Vergleich der Modelle

Es zeigt sich also, dass sich die Größenordnung der dritten und vierten Momente der standardisierten Verteilung der A_{kl} grundlegend unterscheidet. Waren diese vorher beschränkt, wachsen sie nun in d.
Dadurch sind die Voraussetzungen der gleichmäßig in d beschränkten Schiefe und Kurtosis der A_{kl}, die essentiell für die Konsistenz der Schätzer B_2 und B_3 ist, nicht

mehr gegeben. Des Weiteren werden diese für die asymptotischen Resultate von Chen-Qin benötigt, siehe Beweis von Lemma 3 [7]. Aus dem Beweis von Satz 6.2.2 folgt, dass

$$Var\left(\frac{B_2}{Sp(\mathbf{\Sigma}^2)}\right) \notin \mathcal{O}(1), \tag{9.19}$$

bezüglich der Dimension und somit der B_2-Schätzer für große Dimensionen nicht konsistent ist. Simulationen der Schätzer in Abschnitt 11.3.1 bestätigen dies. Siehe dazu Abbildung 11.4.

9.3 Liberales Beispiel

Die durch eine solch degenerierte Verteilung verloren gegangenen Konvergenzeigenschaften führen schließlich zu einem Verlust der Dimensionsstabilität. So lässt sich eine spezielle Verteilung finden, so dass sämtliche bekannte Verfahren für steigende Dimensionen liberal werden.

Da die Struktur der Statistik

$$Z_n = \frac{\sum\limits_{k>l}^{n} A_{kl}}{\sqrt{\sum\limits_{k>l}^{n} A_{kl}^2}} \tag{9.20}$$

formal einer T-Statistik entspricht, bieten sich schiefe Verteilungen für die Vektoren $\mathbf{X}_k$ an, um liberale Ergebnisse zu erhalten. Obwohl die A_{kl} nicht unabhängig sind, zeigt sich in Bezug auf schiefe Verteilungen ein ähnliches Verhalten wie bei normalen T-Statistiken, welche teils starke Niveauüberschreitungen erleiden.

Als schiefe Verteilung soll eine standardisierte Bernoulliverteilung mit Parameter $p = 0.9$ zur Anwendung kommen. Somit seien die $\mathbf{X}_k$ analog zur Konstruktion in (9.16) mit Randverteilung

$$Z_k: \quad P(Z_k = 1/3) = 9/10 \quad \text{und} \quad P(Z_k = -3) = 1/10 \tag{9.21}$$

Dann ist

$$E(Z_k) = \frac{1}{3}\cdot\frac{9}{10} + (-3)\cdot\frac{1}{10} = 0 \qquad Var(Z_k) = \frac{1}{9}\cdot\frac{9}{10} + 9\cdot\frac{1}{10} = 1 \tag{9.22}$$

Dies impliziert eine ebenfalls schiefe Verteilung der A_{kl}, da $Z_k \cdot Z_l$ schief verteilt ist. Nimmt man die so gebildete Verteilung als Grundlage für Niveausimulationen, dann bekommt man beim Test der multivariaten Hypothese folgende Ergebnisse.

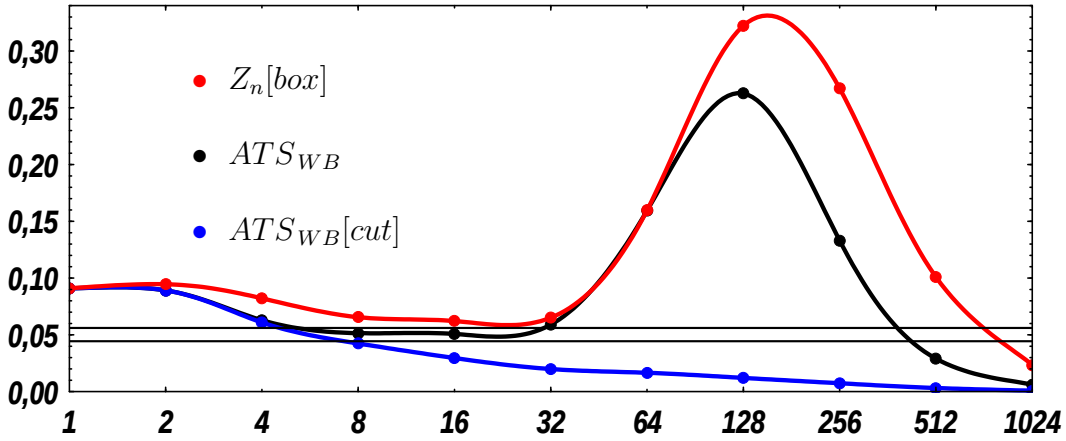


Abbildung 9.1: In der Graphik ist das Niveau des neuen Testverfahrens aus 6.4, und der ANOVA-Typ-Statistik nach Werner-Brunner über die Dimension aufgetragen. Dabei wurde die beschriebende degenerierte Bernoulliverteilung mit Stichprobenumfang $n = 40$ und multivariater Hypothese 10.000-fach simuliert. Die oberste Kurve stellt die Niveauüberschreitung der Statistik Z_n gegen das 95%-Quantil der $\varkappa(\widetilde{f}_{box})$-Verteilung dar. Die mittlere Kurve ist die ATS wie in 4.2 geschildert, während die untere die ATS ist, bei welcher der Schätzer $\widehat{f}_{box}$ auf d beschränkt wird.

Diese extremen Niveauüberschreitungen treten bei sämtlichen Statistiken auf, welche in Kapitel 4 beschrieben wurden. Einzig die ANOVA-Typ-Statistik, welche den Freiheitsgrad beschränkt, zeigt durchweg konservative Resultate. Von Stabiltität kann hier aber dennoch nicht die Rede sein, da die Beschneidung des Freiheitsgrades nur damit gerechfertigt wurde, zulässige Schätzwerte zu erhalten. Bei ANOVA-Typ-Statistiken lässt sich mit einer starken Reduzierung des Freiheitsgrades f stets ein konservatives Ergebnis erzwingen. Im übrigen tritt das selbe Ergebnis ein, wenn man die multivariate Hypothese durch die Hypothese im HD-F1 ersetzt. Dies erklärt sich mit der Invarianz der nichttrivialen Spalten der Hadamardmatrizen unter Multiplikation mit $\mathbf{P}_d$.

9.4 Praktische Auswirkungen

Das beschriebene Beispiel zeigt somit, dass Klassen von multivariaten Verteilungen existieren, auf welche die dargestellten Testverfahren nicht angewendet werden dürfen. Die größte Schwierigkeit für die Praxis besteht darin, dass dies nicht länger an den Marginalverteilungen erkenntlich sein muss. So treten ebenfalls liberale Ergebnisse auf, wenn die vierten Momente im Bai-Saranadasa Modell beliebig groß werden. Allerdings ist in diesem Fall eine Kontrollierbarkeit der Randverteilungen gegeben.
Im angegebenen Beispiel findet zwar eine äußerst schiefe Verteilung Anwendung, dennoch bleibt diese im Niedrigdimensionalen bei hinreichend großem Stichprobenumfang stabil. Es wurde gezeigt, dass bei steigender Dimension Kriterien an die Randverteilungen nicht ausreichen können.

Bezogen auf die Anwendung lassen sich die X_i für $i = 1,..,d$ beispielsweise als Zeitpunkte auffassen. So besteht bei der Modellierung mittels $\mathbf{X}_k = \boldsymbol{\Upsilon} \cdot \mathbf{Z}_k$ eine zufällige Veränderung von Zeitpunkt X_i zu X_{i+1}. Bei realen Daten muss diskutiert werden, ob unter Umständen der Zeitverlauf nicht mehr einem zufälligen Prozess unterworfen ist. Vielmehr könnte er durch gewisse Umstände vorgegeben sein. Dies würde bedeuten, dass es eine Auswahl an festen Zeitverläufen gibt, aus denen einer gezogen wird.
Im Bezug auf das Beispiel der Cortisol-Konzentration könnte solch ein Problem bei den circadianen Schwankungen des Cortisol-Spiegels auftreten. Der pulsartige Ausstoß des Hormons impliziert vorgegebene Zeitverläufe im Tagesprofil. Zu Beginn der Messung nach dem Aufstehen der Patienten ist somit unter Umständen der Verlauf der Cortisol-Konzentration über die nächsten vier Stunden schon fest vorgegeben. Probleme entstünden nun, wenn sich diese Verläufe in Abhängigkeit von bestimmten Charakteristika substantiell unterscheiden würden. Solche Charakteristika könnten beispielsweise Kombinationen aus dem Geschlecht, dem Alter oder dem Vorhandensein bestimmer Gene sein, im ungünstigsten Fall sogar unterschiedliche Rahmenbedingungen der Messung selbst.
Bei den Daten des Beispiels scheint diese Problematik dennoch nicht vorzuliegen. Das Profil des Cortisolspiegels zeigt bei sämtlichen Patienten einen ähnlichen Verlauf. So lässt sich anhand der Abbildungen verifizieren, dass zu Beginn des Tages ein erhöhter Wert vorliegt, welcher langsam abfällt und gegen Ende willkürlichen Schwankungen zu unterliegen scheint. Desweiteren würde solch eine mögliche ungünstige Abhängigkeitsstruktur auch nur je 7 der 28 Messzeitpunkte betreffen. Aufgrund dessen kann eine Auswertung mittels der Methoden aus 6.4 und 7.5 angewendet werden.

10 Alternativen

10.1 Symmetrische Verteilungen

Ein völlig anderer Ansatz um Lösungen im Ein-Stichproben-Problem zu erhalten, ist es, eine symmetrische Verteilung der $\mathbf{X}_k$ anzunehmen. Im Multivariaten bedeutet dies speziell, dass man vektoriell Symmetrie fordern muss.

$$\mathbf{X}_k \sim -\mathbf{X}_k \tag{10.1}$$

Diese Forderung ist wesentlich stärker als die komponentenweiser Symmetrie $X_{ik} \sim -X_{ki}$ für alle $i = 1,...,d$, schließlich nimmt man keine Struktur zwischen den Abhängigkeiten der Komponenten der $\mathbf{X}_k$ an.

Beispiel 10.1.1 Seien X_1, X_2 wie folgt definiert:

$$X_1 = \begin{cases} 1, & \text{mit WK } 1/3 \\ 0, & \text{mit WK } 1/3 \\ -1, & \text{mit WK } 1/3 \end{cases} \qquad X_2 = \begin{cases} 1, & \text{falls } X_1 = 1 \\ 0, & \text{falls } X_1 = -1 \\ -1, & \text{falls } X_1 = 0 \end{cases} \tag{10.2}$$

Dann ist $X_1 \sim -X_1 \sim F \sim X_2 \sim -X_2$. Aber es folgt, dass die gemeinsame Verteilung von (X_1,X_2) nicht symmetrisch um (0,0) ist, da beispielsweise

$$P((X_1,X_2) = (1,1)) = 1/3 \neq P((X_1,X_2) = (-1,-1) = 0$$

10.2 Bezug zum Zwei-Stichprobenfall

Auch wenn die Forderung nach multivariater Symmetrie sehr stark ist und nur in sehr speziellen Designs zu rechtfertigen sein mag, ist sie dennoch von großer technischer Bedeutung. Gerade bei der Erweiterung auf Verfahren für den Zwei-Stichprobenfall

sind Robustheitsaussagen unter Symmetrie von Bedeutung, falls man die Annahme $F_1 = F_2$ unterstellt. So werden Testgrößen, die eine Differenz von Stichproben $\mathbf{X}_{ik}$ aus den unterschiedlichen Gruppen $i = 1{,}2$ betrachten, wie beispielsweise folgende

$$(\overline{\mathbf{X}}_{1\cdot} - \overline{\mathbf{X}}_{2\cdot}) \tag{10.3}$$

zumindest annährend symmetrisiert.
Betrachtet man die Stichprobenumfänge n_1 und n_2 als zufällig binomialverteilt ($p = 1/2$) und unabhängig zu den $\mathbf{X}_{ik}$, so lässt sich unter $F_1 = F_2$ die Zugehörigkeit eines Samples zu einer Stichprobe auch als zufällig interpretieren. Dann kann man das dadurch induzierte Vorzeichen ebenfalls als zufällig symmetrisch verteilt betrachten. Sei durch

$$z(k) = \begin{cases} (1,\pi(k)) & \text{falls } \pi(k) \leq n1 \\ (2,\pi(k) - n1) & \text{falls } \pi(k) > n1 \end{cases}, \; k = 1,..,\, n_1 + n_2 \tag{10.4}$$

mit der Permutierung $\pi(\cdot)$ die (zufällige) Umnummerierung der Stichprobenindizes zu einer gemeinsamen Indexierung beschrieben. Dann ist

$$(\overline{\mathbf{X}}_{1\cdot} - \overline{\mathbf{X}}_{2\cdot}) = \left(n_1^{-1} \sum_{k=1}^{n_1} (\mathbf{X}_{1k} - \boldsymbol{\mu}) \; - \; n_2^{-1} \sum_{k=1}^{n_2} (\mathbf{X}_{2k} - \boldsymbol{\mu}) \right) \tag{10.5}$$

$$= n_1^{-1} \sum_{k=1}^{n_1+n_2} (\mathbf{X}_{z(k)} - \boldsymbol{\mu}) \cdot \begin{cases} 1 & \text{falls } z(k) \in (1,\cdot) \\ -(n_1/n_2) & \text{falls } z(k) \in (2,\cdot) \end{cases} \tag{10.6}$$

$$\approx n_1^{-1} \sum_{k=1}^{n_1+n_2} S_k \cdot (\mathbf{X}_{z(k)} - \boldsymbol{\mu}) \tag{10.7}$$

wobei $S_k \sim (2Ber(1/2) - 1)$ das zu $\mathbf{X}_{z(k)}$ unabhängige, symmetrisch verteilte Vorzeichen ist, da die Stichprobenzugehörigkeit unter $F_1 = F_2$ als zufällig angesehen wird. Desweiteren wurde ausgenutzt, dass für eine bionomialverteilte Zufallsvariable $\sim Bin(k|p,N)$ der Quotient $k/(N-k)$ gegen 1 konvergiert und somit n_1/n_2 zumindest nahe genug bei eins liegt, um robuste Resultate zu erzielen. Im Idealfall hat man gleiche Stichprobenumfänge $n_1 = n_2$.

10.3 Testverfahren unter Symmetrie

Die Konstruktion des liberalen Beispiels aus Abschnitt 9.3 beruhte auf ähnlichen Eigenschaften von T-Statistiken und der Statistik

$$Z_n = \frac{\sum_{k>l}^{n} A_{kl}}{\sqrt{\sum_{k>l}^{n} A_{kl}^2}} \tag{10.8}$$

Demgegenüber sind für T-Statistiken durchaus starke Robustheitsaussagen möglich, falls die Zufallsvariablen einer symmetrischen Verteilung folgen. In Simulationen der Teststatistik Z_n mit symmetrischen Verteilungen lässt sich dementsprechend ebenfalls eine erstaunliche Stabilität feststellen, welche neue Robustheitsaussagen motiviert. Zunächst soll dazu folgendes wichtiges technisches Hilfsmittel vorgestellt werden.

10.3.1 Eaton Bounds

Seien die Zufallsvariablen X_i, $i = 1,..,n$ unabhängig und symmetrisch verteilt. Desweiteren formuliert sich die Statistik S_n des Einstichproben t-Tests als

$$S_n := \sum_{k=1}^{n} \left[\sum_{l=1}^{n} X_l^2 \right]^{-1/2} X_k \tag{10.9}$$

Da die Verteilung hiervon allerdings unbekannt ist, ist es unmöglich, exakte Ergebnisse zu erzielen. Trotzdem lassen sich starke Robustheitsaussagen für die Statistik treffen. So konnte Morris L. Eaton in Arbeiten aus den Jahren 1970 [10] und 1974 [11] folgende Schranken für die Niveauüberschreitungen beim Test gegen Normalverteilung angeben.

Satz 10.3.1 (Eaton Bounds)
Seien Y_i, $i = 1,..,n$ unabhängige Zufallsvariablen mit $E(Y_i) = 0$ und $|Y_i| \leq 1$. Die Y_i müssen dabei weder symmetrisch noch identisch verteilt sein. Seien a_i, $i = 1,..,n$ Gewichte, so dass

$$\sum_{i=1}^{n} a_i^2 = 1 \tag{10.10}$$

Desweiteren seien die sogenannten Eaton Bounds über die Funktion $B_E(y) : \mathbb{R} \to \mathbb{R}$, *mit* $\varphi(x)$ *der Dichtefunktion der Normalverteilung, definiert als*

$$B_E(y) := \inf_{0 \le c < y} \int_c^\infty \left(\frac{z-c}{y-c} \right)^3 \varphi(z) dz \tag{10.11}$$

Dann gilt

$$P\left(\left| \sum_{i=1}^n a_i Y_i \right| \ge y \right) \le 2 B_E(y) \tag{10.12}$$

Beweis: Siehe Theorem 2 in Eaton (1974) [11].

Unter Symmetrie lässt sich dieses Resultat direkt auf die Statistik S_n anwenden. Sei das Vorzeichen einer Zahl x mit

$$sgn(x) = \begin{cases} 1\,, \text{ falls } x > 0 \\ 0\,, \text{ falls } x = 0 \\ -1\,, \text{ falls } x < 0 \end{cases} \tag{10.13}$$

beschrieben.
Bedingt auf die Beträge $|X_1|$,...,$|X_n|$ gilt nun die folgende Darstellung.

$$S_n := \sum_{k=1}^n \left[\sum_{l=1}^n X_l^2 \right]^{-1/2} |X_k| \cdot sgn(X_k) \tag{10.14}$$

Die Gewichte $a_i := (\sum X_l^2)^{-1/2} \cdot |X_k|$ sind dann als fest zu betrachten, während aufgrund der Symmetrie die Verteilung der Vorzeichen $sgn(X_i)$ unverändert bleibt. Außerdem ergibt sich mit der dadurch induzierten Symmetrie der Teststatistik S_n

$$P\left(S_n \ge y\right) \le B_E(y) \tag{10.15}$$

Eine numerische Berechnung der Schranken $B_E(y)$ ergibt folgende Abschätzungen für die Verteilungsfunktion von S_n:

y	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6
$B_E(y)$	.7979	.4617	.2908	.1948	.1311	.0848	.0528	.0317	.0183

Tabelle 10.1: Werte der Eaton Bounds $B_E(y)$ für einzelne y (oben) bzw. ausgewählte Quantile von Φ^{-1} (unten).

α	0.050	0.025	0.010	0.005
$y = \Phi^{-1}(1-\alpha)$	1.645	1.960	2.236	2.576
$B_E(y)$	0.1788	0.0928	0.0384	0.0195

Von Bedeutung ist vor allem eine Betrachtung der $1-\alpha$-Quantile der Normalverteilung, um einen Vergleich mit den asymptotischen Resultaten des Zentralen Grenzwertsatzes zu bekommen. So lässt sich eine strikte Abschätzung der Niveau-Überschreitungen angeben, falls für die Statistik S_n eine Normalapproximation bzw. eine T-Approximation gewählt wird. Diese Worst-Case-Abschätzung für die Ablehnwahrscheinlichkeit der Statistik S_n lässt sich gemäß der Tabelle mit knapp 18% beziffern, wenn zum Niveau 5% getestet wird. Dies mag intuitiv als sehr hoch empfunden werden, da es einer Verdrei - bis Vervierfachung des tatsächlichen Niveaus entspricht und auch bei den anderen Quantilen stets in diesem Rahmen zu liegen scheint. Dennoch bilden die Eaton Bounds eine sehr mächtige Abschätzung, da sie unabhängig vom Stichprobenumfang n sind. So gelten diese selbst für minimale Stichprobenumfänge oder für Zufallsvariablen, die mit großer Wahrscheinlichkeit 0 sind. Falls zahlreiche X_i gleich null sind entspricht dies anschaulich einer Reduktion des Stichprobenumfangs. Dadurch scheint es nicht verwunderlich, dass für steigende Stichprobenumfänge keine Verschärfung der Schranken möglich ist. Tatsächlich ist dies umgekehrt für geringe, feste Stichprobenumfänge n möglich. So konnten Dufour und Hallin (1993) [9] die Schranken $B_E(y)$ aus Satz 10.3.1 für kleine n noch weiter verbessern.

Die eigentliche Schärfe der Eaton Bounds lässt sich an Hand von folgendem Beispiel mit Stichprobenumfang $n = 3$ betrachten:
Seien X_1, X_2, X_3 unabhängig mit $P(X_i = 1) = 1/2 = P(X_i = -1)$ symmetrisch um 0.

Dann wird das Maximum der Statistik S_n genau dann angenommen, wenn sämtliche $X_i = 1$ für $i = 1,2,3$. Dies ist mit Wahrscheinlichkeit $(1/2)^3 = 0.125$ der Fall. Die Abschätzung über die Eaton Bounds hingegen ergibt:

$$\Rightarrow \qquad S_n = \sum_{i=1}^{3} (3)^{-1/2} \cdot 1 = \sqrt{3} \approx 1.732 \tag{10.16}$$

$$\Rightarrow \qquad P(S_n \geq 1.732) \leq B_E(1.732) = 0.1506 \tag{10.17}$$

Dies zeigt, wie strikt die Abschätzung nach Eaton bereits ist. Sie sind in jedem Fall dazu geeignet, asymptotische Resultate zu stützen und Aussagen zur Stabilität eines Verfahrens zu treffen. Ziel ist es nun, ähnliche Aussagen im multivariaten zu formulieren.

10.3.2 Test nach Dufour und Hallin

Dufour und Hallin stellten Abschätzungen für Testverfahren in ihrer Arbeit „Improved Eaton Bounds for Linear Combinations of Bounded Random Variables, With Statistical Applications“ [9] aus dem Jahr 1993 vor. So lassen sich die Abschätzungen nach Eaton auf einen Test der „First-Order Autocorrelation“ übertragen. Das Testverfahren prüft, ob Werte X_i, $t = 1,..n$ einer Zeitreihe eine Korrelation $p \neq 0$ besitzen und wurde von Dufour und Hallin bereits 1990 [8] vorgestellt. Für die Teststatistik

$$S_n^{AC} := \sum_{k=1}^{n} \left[\sum_{l=1}^{n} (X_l X_{l-1})^2 \right]^{-1/2} |X_k||X_{k-1}| sgn(X_k \cdot X_{k-1}) \tag{10.18}$$

konnte gezeigt werden, dass die einzelnen $sgn(X_k \cdot X_{k-1})$ der Summe als unabhängig betrachtet werden können. Dadurch lässt sich analog nachweisen, dass

$$P(S_n^{AC} \geq y) \leq B_E(y) \tag{10.19}$$

Dieser Test nimmt eine zeitliche Ordnung der n Messungen X_k an und betrachtet nur Produkte, die sich aus zeitlich benachbarten Messwerten ergeben. Trotzdem ist dieser Test auf das multivariate Einstichprobenproblem übertragbar, wenn man den Messungen eine zeitliche Ordnung in der bestehenden Reihenfolge zuweist. Ein großer Nachteil, der daraus folgt, ist, dass der Test nicht permutationsinvariant bezüglich der eingelesenen Daten ist. Ein weiterer Nachteil ist die Nichtberücksichtigung von Informationen zwischen nicht-benachbarten Messungen.

Dennoch ergibt als abgewandeltes Testverfahren für das multivariate Ein-Stichproben-Problem:

Seien $\mathbf{X}_k \in \mathbb{R}^d$, $k = 1,..,n$ multivariat symmetrisch verteilt um den Vektor $\boldsymbol{\mu} = E(\mathbf{X}_k)$. Für eine beliebige Hypothesenmatrix $\mathbf{H}$ bleibt das Modell multivariater Symmetrie erhalten, da

$$\mathbf{Y}_k = \mathbf{H}\mathbf{X}_k \sim \mathbf{H}(-\mathbf{X}_k) = -\mathbf{Y}_k \tag{10.20}$$

multivariat symmetrisch ist.

Als Teststatistik für $\mathbf{H}\boldsymbol{\mu} = \mathbf{0}_d$ sei

$$S_n^{DH} := \sum_{k=1}^{n} \left[\sum_{l=1}^{n} (\mathbf{X}_l' \mathbf{X}_{l-1})^2 \right]^{-1/2} |\mathbf{X}_k' \mathbf{X}_{k-1}| \cdot sgn(\mathbf{X}_k' \mathbf{X}_{k-1}) \tag{10.21}$$

welche asymptotisch für $n \to \infty$ als $\mathcal{N}(0,1)$ verteilt angesehen werden soll. Die asymptotische Standardnormalität ist hierbei nicht notwendigerweise gleichmäßig in d, da keine spezielle Verteilungsannahme zugrunde liegt. Einzig für das Bai-Saranadasa Modell ließe sich diese analog zur asymptotischen Normalität von Z_n zeigen. Ohne diese Annahme ist die Normalverteilung als Approximation zu betrachten. Dennoch gilt auch hier die Abschätzung:

$$P(S_n^{DH} \geq y) \leq B_E(y) \tag{10.22}$$

In Simulationen lassen sich bezüglich der Approximation $S_n^{DH} \dot{\sim} \mathcal{N}(0,1)$ gewöhnlich konservative Resultate beobachten. Problematisch ist einzig die teils sehr starke Konservativität. Diese tritt speziell dann auf, falls einzelne $A_{k(k-1)} = \mathbf{X}_k' \mathbf{X}_{k-1}$ dominieren, was beispielsweise bei degenerierten Verteilungen der Fall ist.

10.3.3 Vorzeichentest

Neben dem Problem der Permutationsinvarianz der Teststatistik S_n^{DH} erscheint es fragwürdig, eine willkürliche Anordnung der $\mathbf{X}_k$ zu unterstellen. Vielmehr sind Robustheitsaussagen der Statistik

$$F_n := \sum_{k=1}^{n} (\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_{k-1})' \cdot \mathbf{X}_k \tag{10.23}$$

aus (4.18) von Bedeutung. Die Anwendung der Eaton Bounds ist hier allerdings höchst nicht-trivial da $|(\mathbf{X}_1 + \mathbf{X}_2)' \cdot \mathbf{X}_3|$ durch die interne Summierung nicht länger unabhängig zu $sgn(\mathbf{X}_1' \mathbf{X}_2)$ ist. Als Ausweg bietet es sich an, eine Statistik lediglich über die

Vorzeichen zu formulieren. Dies ist mit dem Symmetrieargument analog zu dem Test nach Dufour und Hallin möglich. So ist $(\mathbf{X}_1 + .. + \mathbf{X}_{k-1})' \cdot \mathbf{X}_k$ bedingt auf $\mathbf{X}_1$,..,$\mathbf{X}_{k-1}$ stets symmetrisch verteilt.
Sei m die Anzahl der $(\mathbf{X}_1 + ... + \mathbf{X}_{k-1})' \cdot \mathbf{X}_k$ welche ungleich 0 sind. Dann ist

$$S_n^{SIGN} := \sum_{k=1}^{n} m^{-1/2} sgn((\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_{k-1})' \cdot \mathbf{X}_k) \tag{10.24}$$

asymptotisch normalverteilt für $m \to \infty$. Desweiteren gilt auch hier

$$P(S_n^{DH} \geq y) \leq B_E(y) \tag{10.25}$$

selbst wenn

$$n \to \infty \quad \not\Rightarrow \quad m \to \infty \tag{10.26}$$

Selbstverständlich bedarf es hierfür prinzipiell keiner asympotischen Approximation, da das Testverfahren einem Vorzeichentest entspricht. Somit ist es möglich einen exakten Test durchzuführen, allerdings soll darauf mit Hinblick auf den nächsten Abschnitt verzichtet werden.

10.4 Permutationsstabiler Vorzeichentest

Um den Vorzeichentest nun als mögliche Alternative für die parametrischen Verfahren anwenden zu können, ist primär das Problem der Instabilität unter Permutation der Daten zu lösen.

Sei $T_1 = T_1(\mathbf{X}_1,...,\mathbf{X}_n)$ eine Statistik über die Zufallsvektoren $\mathbf{X}_k$, $k = 1,..,n$. Dabei sei $T_1 \sim \mathcal{N}(\mu,\sigma^2)$ und im Allgemeinen gelte für eine beliebige Permutation π

$$T_1(\mathbf{X}_1,...,\mathbf{X}_n) \neq T_1(\mathbf{X}_{\pi(1)},...,\mathbf{X}_{\pi(n)}) \tag{10.27}$$

Im folgenden seien sämtliche $n!$ Permutationen durchnumeriert, während die 1 der Identität entspricht. Dann bezeichne

$$T_i := T_1(\mathbf{X}_{\pi(1)},...,\mathbf{X}_{\pi(n)}) \tag{10.28}$$

mit der i-ten Permutation und

$$\mathbb{X} := \{\mathbf{X}_k : \ k = 1,..,n\} \tag{10.29}$$

die Menge der Daten. Wird die Permutation nun als zufällig mit Ausprägung i betrachtet, dann gilt

$$E(T_1) = E\left(E(T_i|\mathbb{X})\right) \tag{10.30}$$

Dabei ist $E(T_i|\mathbb{X})$ das Mittel über alle Permutation zu einem gegebenen Datensatz $\mathbb{X}$. Außerdem gilt für $T_i \,|\, \mathbb{X}$ nach dem Satz über die Varianz einer bedingten Verteilung.

$$Var(T_1) = E(\underbrace{Var(T_i|\mathbb{X})}_{\geq 0}) + Var(E(T_i|\mathbb{X})) \tag{10.31}$$

$$\geq Var(E(T_i|\mathbb{X})) \tag{10.32}$$

Dabei kann $Var(T_i|\mathbb{X})$ als Indikator betrachtet werden, wie stark die Statistik unter Permutierung schwankt. Ist die Statistik gerade permutationsinvariant, dann ist die auf die Daten bedingte Varianz 0. Demgegenüber entspricht $Var(E(T_i|\mathbb{X}))$ der Varianz über das Mittel sämtlicher T_i zu einem gegebenen Datensatz. Dieser Mittelwert sei mit

$$\overline{T}_\cdot := E(T_i|\mathbb{X}) \tag{10.33}$$

bezeichnet und ist stabil unter Permutation berechenbar. Numerisch kann dies für große Stichprobenumfänge mit $n!$ Permutationen zu Problemen führen. Zur Berechnung reicht allerdings eine zufällige Auswahl, da $\overline{T}_\cdot$ nach dem starken Gesetz der großen Zahlen konvergiert.
Approximiert man nun (analog zu T_1) $\overline{T}_\cdot$ mit $\mathcal{N}(\mu,\sigma^2)$, dann wird die Testentscheidung auf Grund der verringerten Varianz $\sigma(\overline{T}_\cdot) \leq \sigma$ konservativ. Ist die Schwankung von T_1 bei Permutierung der Daten gering, so fällt die Konservativität nicht stark ins Gewicht.

Wendet man diese Überlegungen auf den Vorzeichentest an, so ergibt sich ein permutationsinvarianter Test für die Hypothese $\mathbf{H}\boldsymbol{\mu} = \mathbf{0}_d$.

Sei Π eine Teilmenge der Permutationen mit Anzahl $p = \#\{\pi : \ \pi \in \Pi\}$ und sei $\left(S_n^{SIGN}\right)_{(i)}$ die Teststatistik S_n^{SIGN} aus (10.24) angewendet auf die i-te Permutation der Ursprungsdaten. Dann sei die permutationsinvariante Version des Vorzeichentests

$$\overline{S_n^{SIGN}} := \frac{1}{p} \sum_{i \in \Pi} \left(S_n^{SIGN}\right)_{(i)} \tag{10.34}$$

Wird die Teststatistik als normalverteilt angesehen, so liefert diese Approximation in jedem Fall robuste Resultate. Die Eaton Bounds dienen auch hier als Schranke. In Simulationen wurde die Konservativität dieses Testverfahrens bestätigt. Bei degenerierten Verteilungen schöpft er dennoch das Niveau besser aus als die parametrischen Verfahren und erreicht eine bessere Power, wie in der folgenden Abbildung dargestellt ist.

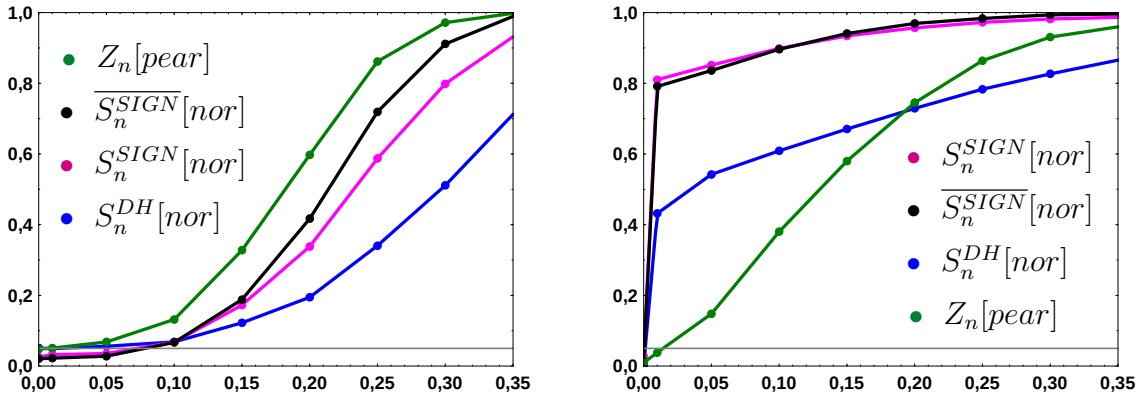


Abbildung 10.1: In der Abbildung ist die Power der verschiedenen Verfahren für ansteigenden Shifteffekt. Dabei wird die multivariate Alternative mit $\boldsymbol{\mu} = \delta \cdot \mathbf{1}_d$ simuliert, wenn zur multivariaten Hypothese getestet wird. Im linken Graph sind die Powerkurven des Tests $Z_n[pear]$ aus (7.32) und der neuen Verfahren unter multivariater Normalverteilung $\mathcal{N}(\mathbf{0}, \mathbf{I})$ zu sehen. Demgegenüber sind im rechten Graph die Powerkurven unter der degenerierten Verteilung aus (9.16) mit normalverteilten Rändern abgebildet. Desweiteren ist $d = 32$, $n = 15$ und $n_{sim} = 10000$.

11 Simulationen

In diesem Kapitel wird untersucht in welchem Rahmen die verschiedenen Teststatistiken unter der Hypothese das Niveau einhalten und welche Power sie besitzen, um die Alternative aufzudecken. Der Fehler erster Art, das Niveau α, wird für $\alpha = 10\%$, $\alpha = 5\%$, $\alpha = 1\%$ und $\alpha = 0.5\%$ untersucht und sollte nicht überschritten werden. Minimale Überschreitung dieses Niveaus, sprich Liberalität des Tests, ist natürlich zu tolerieren, zumal die Testverfahren auf asymptotischen Resultaten beruhen und teils auf Verteilungsapproximationen nach Box bzw. Pearson zurückgegriffen wird. Desweiteren werden viele verschiedene Klassen von Verteilungen abgedeckt und gerade im Bai-Saranadasa Modell ist der Übergang von Verteilungen mit hinreichend kleinen vierten Momenten bishin zu beliebig großen vierten Momenten stetig.
Genau dies ist neben der Verifikation der Theorie eine wichtige Motivation für Simulationen, um Grenzwerte der Parameter zu finden, was mit strikten Fehlerrechnungen praktisch unmöglich wäre. Dafür wird eine Vorstellung benötigt, welche Niveauüberschreitungen tolerierbar sind, damit das Verfahren unter diesen Voraussetzungen noch als robust bezeichnen werden kann. Diese Festlegung liegt beim Anwender und ist eng mit der Verteilungsannahme der statistischen Auswertung verknüpft.

11.1 Niveau-Simulationen

Die Niveau-Simulationen bestätigen nun, dass das neuen Verfahren bei der Annahme von multivariater Normalverteilung das Niveau bis auf auf eine Abweichung von kleiner $+-10\%$ zur gegebenen Fehlerwahrscheinlichkeit α einhält. Bei Nichtnormalverteilung ist eine teils leichte Konservativität bzw. Liberalität zu beobachten, während bei Verteilungen, die dem Bai-Saranadasa Modell mit endlichen vierten Momenten nicht entsprechen, ein höherer Anstieg bzw. Abfall des Fehlers ersten Art in Kauf genommen werden muss. Das Unterschreiten des Niveaus, sprich Konservativität des Testes, ist dennoch ein nicht zu vernachlässigendes Problem, das zu kontrollieren ist. Ein konservatives Verfahren kontrolliert zwar in jedem Fall die zulässige Fehlerrate, allerdings wird die Power des Testes meist im gleichen Maße schlechter, indem das Niveau nicht ausgeschöpft wird.
Das Augenmerk der Simulationen soll durchaus auch auf den höheren Quantilen $\alpha < 0{,}05$ liegen. Gerade hier sind Vorteile des neuen Testverfahrens mit Pearsonapproximation zu beobachten.

Für diese Arbeit wurde eine Auswahl an Testverfahren für das hochdimensionale Ein-Gruppen-Design und eine Auswahl möglichst verschiedenartiger Verteilungen in der (Matrix-)Programmiersprache IML in SAS Version 9.2 simuliert.
Die einzelnen Testverfahren und ihre Abkürzungen im Überblick:

- $Z_n[Pear]$ nutzt für eine Approximation von Z_n aus (7.32), die 3-Momenten-Approximation nach Pearson. Der Freiheitsgrad g_{pear} der $\kappa(g_{pear})$-Verteilung wird mit $\widetilde{g}_{pear}$ aus (8.6) geschätzt. Das Hauptaugenmerk dieser Arbeit liegt auf dieser Statistik. Sie dient als primärer Vergleich zu anderen Testverfahren.

- $Z_n[Box]$ bezieht sich auf die Approximation der Teststatistik Z_n mit der $\varkappa(f_{box})$-Verteilung. Der Freiheitsgrad f_{box} wird hierfür mit $1/\widetilde{h}_{box}$ aus (8.5) geschätzt.

- $C_n[nor]$ steht für den Test nach Chen und Qin. Dieser beruht auf einer Normalapproximation der Teststatistik C_n aus (6.2).

- ATS_{WB} stellt die ANOVA-Typ-Statistik von Werner und Brunner, mit den dimensionsstabilen Schätzern B_0, B_1 und B_2 aus 4.1 dar. Allerdings wird der Freiheitsgradschätzer $\widehat{f}_{box}$ bezüglich der Dimension beschnitten.

- ATS_{KL} stellt die klassische ANOVA-Typ-Statistik mit Boxapproximation aus 4.2 dar. Der Test verwendet für die Spuren die Plugin-Schätzer aus der empirischen Kovarianzmatrix.

- $SIGN$ steht für die permutationsinvariante Version des Vorzeichentest aus (10.34). Simuliert wurde jeweils der Mittelwert von S_n^{SIGN} für $p = 25$ Permutationen. Dieser Test soll als Vergleich des alternativen Ansatzes der multivariaten Symmetrie dienen.

Simuliert wurden ausgewählte Verteilungen, die ein möglichst breites Sprektrum abdecken. Um das Verhalten der Teststatistiken expliziter untersuchen zu können, wurde meist auf spezielle Hypothesenmatrizen wie das HD-Fi-Design Design verzichtet und stattdessen die Hypothese $\mathbf{H}_{MULT}$ mit

$$\mathbf{H}_{MULT} \cdot \boldsymbol{\mu} = \mathbf{I}_d \boldsymbol{\mu} = \mathbf{0}$$

formuliert. Durch Multiplikation mit der Einheitsmatrix werden die erzeugten Zufallsvektoren $\mathbf{X}_k$ nicht verändert. So lassen sich genauere Aussagen über das Verhalten der Approximationen treffen, während die Multiplikation mit Kontrastmatrizen die Abhängigkeiten meist abschwächt. So werden beispielsweise Blockeffekte bei Compound Symmetry, welche einzelne große Eigenwerte erzeugen, beim Testen der Hypothese $H_{HDF1} = \mathbf{P}_d$ neutralisiert.

Multivariate Standardnormalverteilung, n=15							
d	Niveau	$Z_n[pear]$	$Z_n[box]$	$C_n[nor]$	ATS_{WB}	ATS_{KL}	$SIGN$
1	0.100	0.1145	0.1145	0.1203	0.1057	0.1235	0.0772
	0.050	0.0552	0.0552	0.0938	0.0487	0.0705	0.0308
	0.010	0.0077	0.0077	0.0596	0.0062	0.0222	0.0071
	0.005	0.0029	0.0029	0.0511	0.0023	0.0142	0.0040
2	0.100	0.1073	0.1074	0.1207	0.0986	0.1117	0.0667
	0.050	0.0520	0.0520	0.0888	0.0460	0.0598	0.0280
	0.010	0.0074	0.0073	0.0504	0.0060	0.0152	0.0037
	0.005	0.0029	0.0029	0.0420	0.0022	0.0086	0.0014
4	0.100	0.1039	0.1037	0.1194	0.0965	0.0963	0.0621
	0.050	0.0501	0.0502	0.0832	0.0447	0.0464	0.0249
	0.010	0.0077	0.0077	0.0410	0.0067	0.0093	0.0025
	0.005	0.0029	0.0031	0.0320	0.0026	0.0048	0.0009
8	0.100	0.1026	0.1026	0.1180	0.0958	0.0754	0.0599
	0.050	0.049	0.0493	0.0773	0.0450	0.0312	0.0232
	0.010	0.0076	0.0078	0.0335	0.0069	0.0043	0.0023
	0.005	0.0033	0.0035	0.0245	0.0032	0.0021	0.0007
16	0.100	0.1025	0.1027	0.1156	0.0958	0.0476	0.0594
	0.050	0.0486	0.0491	0.0715	0.0449	0.0150	0.0224
	0.010	0.0078	0.0081	0.0272	0.0071	0.0010	0.0017
	0.005	0.0033	0.0035	0.0191	0.0030	0.0003	0.0005
32	0.100	0.1000	0.1002	0.1108	0.0938	0.0200	0.0582
	0.050	0.0480	0.0484	0.0665	0.0445	0.0036	0.0209
	0.010	0.0080	0.0083	0.0223	0.0074	0.0001	0.0014
	0.005	0.0031	0.0034	0.0148	0.0029	0.0000	0.0003
64	0.100	0.1007	0.1010	0.1095	0.0945	0.0043	0.0564
	0.050	0.0488	0.0494	0.0633	0.0452	0.0003	0.0199
	0.010	0.0084	0.0089	0.0198	0.0079	0.0000	0.0011
	0.005	0.0038	0.0041	0.0123	0.0035	0.0000	0.0003
256	0.100	0.0995	0.1001	0.1054	0.0936	0.0000	0.0576
	0.050	0.0484	0.0493	0.0572	0.0451	0.0000	0.0199
	0.010	0.0087	0.0094	0.0156	0.0081	0.0000	0.0010
	0.005	0.0040	0.0046	0.0088	0.0039	0.0000	0.0002
1024	0.100	0.1020	0.1029	0.1029	0.0953	0.0000	0.0556
	0.050	0.0499	0.0525	0.0526	0.0468	0.0000	0.0188
	0.010	0.0093	0.0110	0.0111	0.0086	0.0000	0.0009
	0.005	0.0043	0.0056	0.0057	0.0046	0.0000	0.0002

Tabelle 11.1: Simuliert wurden die Testverfahren unter multivariater Normalverteilung mit Einheitsmatrix als Kovarianzmatrix und Hypothesenmatrix für verschiedene Dimensionen. Der Stichprobenumfang lag bei $n = 15$ und die Anzahl der Simulationsdurchläufe bei $n_{sim} = 100.000$.

Multivariate Standardnormalverteilung, d=32							
n	Niveau	$Z_n[pear]$	$Z_n[box]$	$C_n[nor]$	ATS_{WB}	ATS_{KL}	$SIGN$
7	0.100	0.1028	0.1030	0.1210	0.0875	0.0063	0.0581
	0.050	0.0452	0.0465	0.0758	0.0374	0.0005	0.0172
	0.010	0.0031	0.0050	0.0295	0.0038	0.0000	0.0004
	0.005	0.0002	0.0012	0.0210	0.0010	0.0000	0.0000
10	0.100	0.1037	0.1039	0.1173	0.0939	0.0113	0.0592
	0.050	0.0480	0.0489	0.0713	0.0422	0.0014	0.0199
	0.010	0.0065	0.0074	0.0258	0.0062	0.0000	0.0008
	0.005	0.0022	0.0029	0.0171	0.0025	0.0000	0.0000
15	0.100	0.0988	0.0990	0.1102	0.0931	0.0199	0.0590
	0.050	0.0476	0.0482	0.0653	0.0443	0.0036	0.0232
	0.010	0.0077	0.0082	0.0221	0.0071	0.0001	0.0014
	0.005	0.0033	0.0036	0.0149	0.0032	0.0000	0.0002
20	0.100	0.1004	0.1006	0.1101	0.0962	0.0294	0.0525
	0.050	0.0496	0.0500	0.0655	0.0470	0.0070	0.0174
	0.010	0.0089	0.0094	0.0224	0.0086	0.0003	0.0008
	0.005	0.0041	0.0043	0.0148	0.0039	0.0001	0.0001
30	0.100	0.1009	0.1010	0.1094	0.0981	0.0420	0.0531
	0.050	0.0503	0.0507	0.0650	0.0485	0.0124	0.0203
	0.010	0.0098	0.0099	0.0211	0.0093	0.0008	0.0022
	0.005	0.0049	0.0051	0.0137	0.0047	0.0003	0.0010
50	0.100	0.1009	0.1010	0.1112	0.0995	0.0595	0.0531
	0.050	0.0496	0.0497	0.0656	0.0486	0.0221	0.0204
	0.010	0.0091	0.0091	0.0213	0.0088	0.0014	0.0022
	0.005	0.0045	0.0046	0.0129	0.0044	0.0002	0.0009

Tabelle 11.2: Simuliert wurden die Testverfahren unter multivariate Normalverteilung der Dimension $d = 32$ für unterschiedliche Stichprobenumfänge n. Die Kovarianzstruktur der Verteilung ist $\mathbf{S} = \mathbf{I}_d$ und getestet wurde die multivariate Hypothese $\mathbf{H}_{MULT} = \mathbf{I}_d$ in $n_{sim} = 100.000$ Simulationsdurchläufen.

Tabelle 11.3: Simuliert wurde die degenerierte Normalverteilung nach Konstruktion in (9.16). Desweiteren sei $\mathbf{\Sigma} = \mathbf{I}_d$, $\mathbf{H} = \mathbf{I}_d$, $n = 15$, $n_{sim} = 10.000$.

Tabelle 11.4: Simuliert wurde eine multivariate Normalverteilung mit Compound Symmetry Kovarianzstruktur. Desweiteren sei $\mathbf{\Sigma} = \mathbf{I}_d + 1/2 \cdot \mathbf{J}_d$, $\mathbf{H} = \mathbf{I}_d$, $n = 15$, $n_{sim} = 100.000$.

Degenerierte Normalverteilung, n=15

d	Niveau	$Z_n[pear]$	$Z_n[box]$	$C_n[nor]$	ATS_{WB}	ATS_{KL}	$SIGN$
4	0.100	0.1133	0.1137	0.1277	0.0836	0.0793	0.0616
	0.050	0.0484	0.0477	0.0852	0.0307	0.0293	0.0261
	0.010	0.0026	0.0025	0.0356	0.0012	0.0022	0.0034
	0.005	0.0003	0.0003	0.0251	0.0002	0.0007	0.0011
16	0.100	0.1132	0.1129	0.1280	0.0517	0.0120	0.0493
	0.050	0.0275	0.0266	0.0561	0.0100	0.0013	0.0168
	0.010	0.0001	0.0001	0.0057	0.0001	0	0.0009
	0.005	0	0	0.0025	0	0	0.0002
64	0.100	0.0479	0.0471	0.0537	0.0151	0.0004	0.0455
	0.050	0.0020	0.0021	0.0090	0.0008	0	0.0134
	0.010	0	0	0.0001	0	0	0.0005
	0.005	0	0	0	0	0	0.0001
256	0.100	0.0053	0.0050	0.0057	0.0020	0	0.0449
	0.050	0.0001	0.0001	0.0006	0	0	0.0135
	0.010	0	0	0	0	0	0.0004
	0.005	0	0	0	0	0	0.0001

Normalverteilung - Compound Symmetry, n=15

d	Niveau	$Z_n[pear]$	$Z_n[box]$	$C_n[nor]$	ATS_{WB}	ATS_{KL}	$SIGN$
16	0.100	0.1012	0.0987	0.1156	0.0973	0.0846	0.0637
	0.050	0.0518	0.0561	0.0852	0.0554	0.0477	0.0259
	0.010	0.0100	0.0154	0.0500	0.0153	0.0153	0.0037
	0.005	0.0043	0.0085	0.0416	0.0085	0.0100	0.0015
64	0.100	0.1043	0.1000	0.1174	0.0991	0.0808	0.0698
	0.050	0.0544	0.0614	0.0898	0.0609	0.0478	0.0305
	0.010	0.0100	0.0205	0.0563	0.0205	0.0168	0.0042
	0.005	0.0048	0.0128	0.0478	0.0129	0.0116	0.0014
256	0.100	0.1052	0.0996	0.1178	0.0989	0.0782	0.0741
	0.050	0.0546	0.0626	0.0909	0.0623	0.0468	0.0334
	0.010	0.0097	0.0220	0.0578	0.0220	0.0165	0.0058
	0.005	0.0041	0.0147	0.0498	0.0148	0.0115	0.0020
1024	0.100	0.1064	0.1003	0.1244	0.0993	0.0829	0.0752
	0.050	0.0548	0.0633	0.0967	0.0629	0.0488	0.0315
	0.010	0.0096	0.0230	0.0625	0.0230	0.0187	0.0078
	0.005	0.0040	0.0150	0.0544	0.0151	0.0130	0.0033

Normalverteilung - Autoregressiv, HDF1, n=15						
d	Niveau	$Z_n[pear]$	$Z_n[box]$	$C_n[nor]$	ATS_{WB}	ATS_{KL}
2	0.100	0.1129	0.1129	0.1191	0.1040	0.1225
	0.050	0.0558	0.0558	0.0924	0.0491	0.0709
	0.010	0.0080	0.0080	0.0598	0.0063	0.0221
	0.005	0.0029	0.0029	0.0517	0.0022	0.0143
4	0.100	0.1038	0.1034	0.1178	0.0984	0.1078
	0.050	0.0517	0.0518	0.0877	0.0485	0.0602
	0.010	0.0089	0.0091	0.0518	0.0081	0.0185
	0.005	0.0035	0.0037	0.0432	0.0032	0.0118
8	0.100	0.1022	0.1003	0.1163	0.0973	0.1009
	0.050	0.0524	0.0535	0.0869	0.0514	0.0579
	0.010	0.0101	0.0117	0.0521	0.0113	0.0189
	0.005	0.0042	0.0055	0.0434	0.0052	0.0122
16	0.100	0.1019	0.1001	0.1163	0.0975	0.0961
	0.050	0.0519	0.0538	0.0854	0.0522	0.0549
	0.010	0.0096	0.0117	0.0501	0.0114	0.0168
	0.005	0.0041	0.0060	0.0413	0.0058	0.0111
64	0.100	0.1018	0.1011	0.1172	0.0999	0.0779
	0.050	0.0505	0.0538	0.0818	0.0533	0.0382
	0.010	0.0094	0.0127	0.0421	0.0127	0.0090
	0.005	0.0041	0.0067	0.0328	0.0068	0.0053
256	0.100	0.1019	0.1029	0.1156	0.1026	0.0302
	0.050	0.0506	0.0540	0.0733	0.0540	0.0085
	0.010	0.0090	0.0118	0.0301	0.0119	0.0004
	0.005	0.0039	0.0060	0.0213	0.0060	0.0001
1024	0.100	0.1020	0.1029	0.1133	0.1029	0.0011
	0.050	0.0499	0.0525	0.0677	0.0526	0.0000
	0.010	0.0093	0.0110	0.0246	0.0111	0.0000
	0.005	0.0043	0.0056	0.0161	0.0057	0.0000

Tabelle 11.5: Simuliert wurden eine multivariater Normalverteilung mit Autoregressiver Kovarianzstruktur im $HDF1$-Design. Dabei ist $\mathbf{S} = \left((0.9)^{|i-j|}\right)_{i,j=1,..,d}$, $\mathbf{H} = \mathbf{P}_d$, $n = 15$, $n_{sim} = 100.000$.

Tabelle 11.6: Simuliert wurde unahängige lognormal-verteilte Ränder im $HDF1$-Design. Desweiteren ist $\mathbf{S} = \mathbf{I}_d$, $\mathbf{H} = \mathbf{P}_d$, $n = 15$, $n_{sim} = 100.000$.

Tabelle 11.7: Simuliert wurden Zufallsvektoren verteilt nach dem Bai-Saranadasa Modell $\mathbf{X}_k = \mathbf{\Gamma} \cdot \mathbf{Z}_k$ mit unabhängigen zentrierten exponential(1)-verteilten Erzeugern Z_i im HDF1-Design. Dabei ist $\mathbf{\Upsilon}$ die Wurzel aus einer Spektralzerlegung der zu erzeugenden Kovarianzmatrix, für welche eine Toeplitzstruktur mit linearem Abfall gewählt wurde. Somit ist $\mathbf{S} = (d - |i-j|)_{i,j=1,..,d}$, $\mathbf{H} = \mathbf{P}_d$, $n = 15$, $n_{sim} = 100.000$.

Lognormalverteilung - HDF1, n=15

d	Niveau	$Z_n[pear]$	$Z_n[box]$	$C_n[nor]$	ATS_{WB}	ATS_{KL}
16	0.100	0.0767	0.0765	0.0888	0.0632	0.0219
	0.050	0.0278	0.0274	0.0494	0.0211	0.0046
	0.010	0.0017	0.0017	0.0134	0.0012	0.0002
	0.005	0.0005	0.0005	0.0086	0.0004	0.0001
64	0.100	0.0790	0.0792	0.0881	0.0706	0.0021
	0.050	0.0312	0.0316	0.0445	0.0276	0.0002
	0.010	0.0027	0.0030	0.0101	0.0026	0.0000
	0.005	0.0008	0.0010	0.0056	0.0008	0.0000
256	0.100	0.0847	0.0851	0.0903	0.0787	0.0000
	0.050	0.0351	0.0360	0.0435	0.0328	0.0000
	0.010	0.0035	0.0041	0.0084	0.0036	0.0000
	0.005	0.0012	0.0014	0.0044	0.0013	0.0000
1024	0.100	0.0900	0.0908	0.0975	0.0852	0.0000
	0.050	0.0396	0.0405	0.0493	0.0378	0.0000
	0.010	0.0054	0.0061	0.0101	0.0056	0.0000
	0.005	0.0020	0.0024	0.0047	0.0022	0.0000

Exponentialverteilung - Toeplitzstruktur, HDF1, n=15

d	Niveau	$Z_n[pear]$	$Z_n[box]$	$C_n[nor]$	ATS_{WB}	ATS_{KL}
16	0.100	0.1043	0.1022	0.1182	0.0978	0.1024
	0.050	0.0532	0.0541	0.0897	0.0507	0.0585
	0.010	0.0096	0.0113	0.0539	0.0101	0.0191
	0.005	0.0040	0.0052	0.0454	0.0047	0.0126
64	0.100	0.1036	0.1014	0.1167	0.0976	0.1029
	0.050	0.0535	0.0546	0.0885	0.0523	0.0606
	0.010	0.0096	0.0117	0.0550	0.0110	0.0203
	0.005	0.0039	0.0056	0.0464	0.0052	0.0136
256	0.100	0.1041	0.1015	0.1176	0.0977	0.1037
	0.050	0.0529	0.0539	0.0887	0.0519	0.0599
	0.010	0.0099	0.0124	0.0538	0.0118	0.0214
	0.005	0.0043	0.0060	0.0462	0.0057	0.0144
1024	0.100	0.1029	0.1004	0.1194	0.0970	0.1061
	0.050	0.0527	0.0539	0.0902	0.0517	0.0617
	0.010	0.0094	0.0116	0.0545	0.0111	0.0215
	0.005	0.0042	0.0056	0.0466	0.0053	0.0140

11.2 Power-Simulationen

Die Untersuchung der Qualität der verschiedenen Teststatistiken, eine Hypothese H_0 unter Alternative H_1 abzulehnen, erfordert Power-Simulationen. Wird die Alternative

$$H_1 : \boldsymbol{\theta} = \mathbf{H}\boldsymbol{\mu} \neq \mathbf{0} \tag{11.1}$$

betrachtet, dann soll für ansteigende $\boldsymbol{\theta}$ die Ablehnwahrscheinlichkeit möglichst schnell gegen 1 konvergieren. Dabei sind verschiedene Klassen von Alternativen zu betrachten, welche über verschiedene Formen der Vektoren $\boldsymbol{\mu}$ charakterisiert werden. Über die einzelnen Vektoren einer Klasse mit $\mathbf{H}\boldsymbol{\mu} \neq \mathbf{0}$ soll durch

$$\delta := \|\boldsymbol{\mu}\|_\infty > 0 \tag{11.2}$$

eine eindeutige Ordnung bestimmt sein. In dieser Arbeit soll nun für folgende Klassen von Alternativen die Konvergenzgeschwindigkeit

$$P_{H_1(\delta)}(\text{Test lehnt ab}) \xrightarrow{\delta\to\infty} 1 \tag{11.3}$$

untersucht werden:

- Ein-Punkt-Alternative: OBdA soll die erste Komponente $\mu_1 = \delta$ gesetzt werden und die übrigen Komponenten $\mu_i = 0$ für $i > 1$.

- Trend-Alternative: Es soll ein gleichmäßiger Abfall der μ_i stattfinden. Die i-te Komponente erhält dabei den Shift $\mu_i = \delta \cdot (d - i + 1)/d$ für alle $i = 1,..,d$.

- Multivariate Alternative: Diese Alternative wird nur von multivariaten Tests aufgedeckt und ist im HD-Fi nicht zu betrachten. Es sei $\boldsymbol{\mu} = \delta \cdot \mathbf{1}_d$.

Untersucht wurden ausschnittweise Zufallsvektoren der Dimension $d = 32$ bei Stichprobenumfang $n = 15$. Als Verteilung der $\mathbf{Y}_k$ wurde multivariate Standardnormalverteilung für eine multivariate Hypothese und die Exponentialverteilung mit Toeplitzstrukur aus Tabelle 11.7 im HD-F1 simuliert.

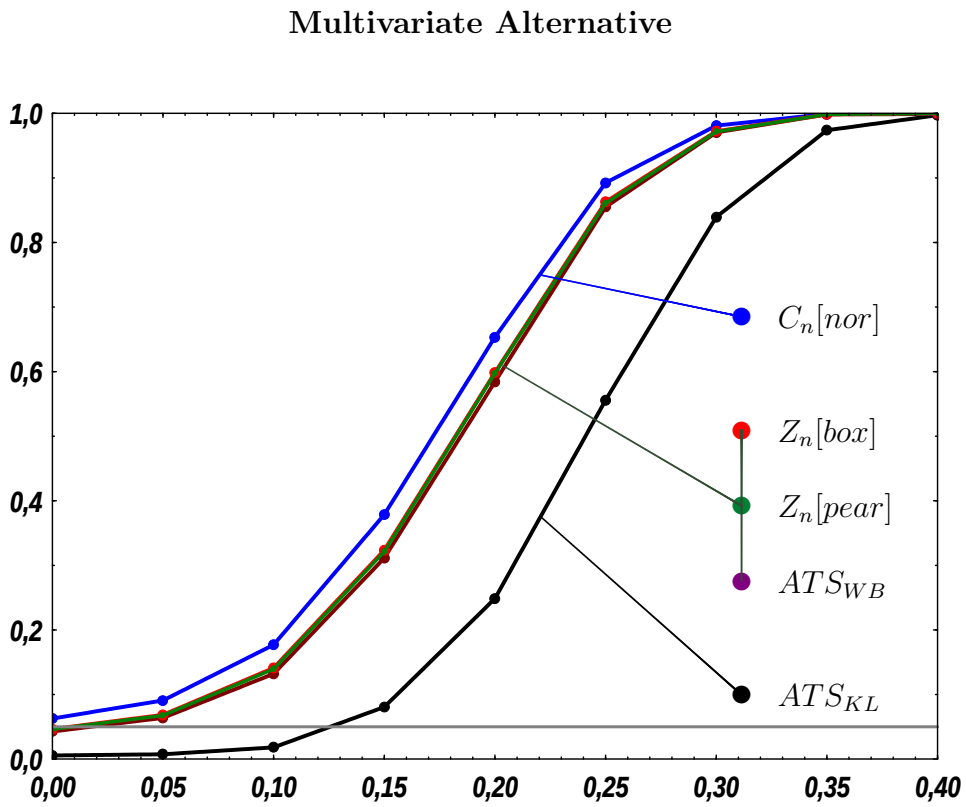


Abbildung 11.1: In der Abbildung ist die Power der verschiedenen Verfahren für eine (bezüglich δ ansteigende) Multivariate Alternative unter $\mathcal{N}(\mathbf{0}, \mathbf{I})$ zu sehen. Desweiteren ist $d = 32$, $n = 15$ und $n_{sim} = 10000$.

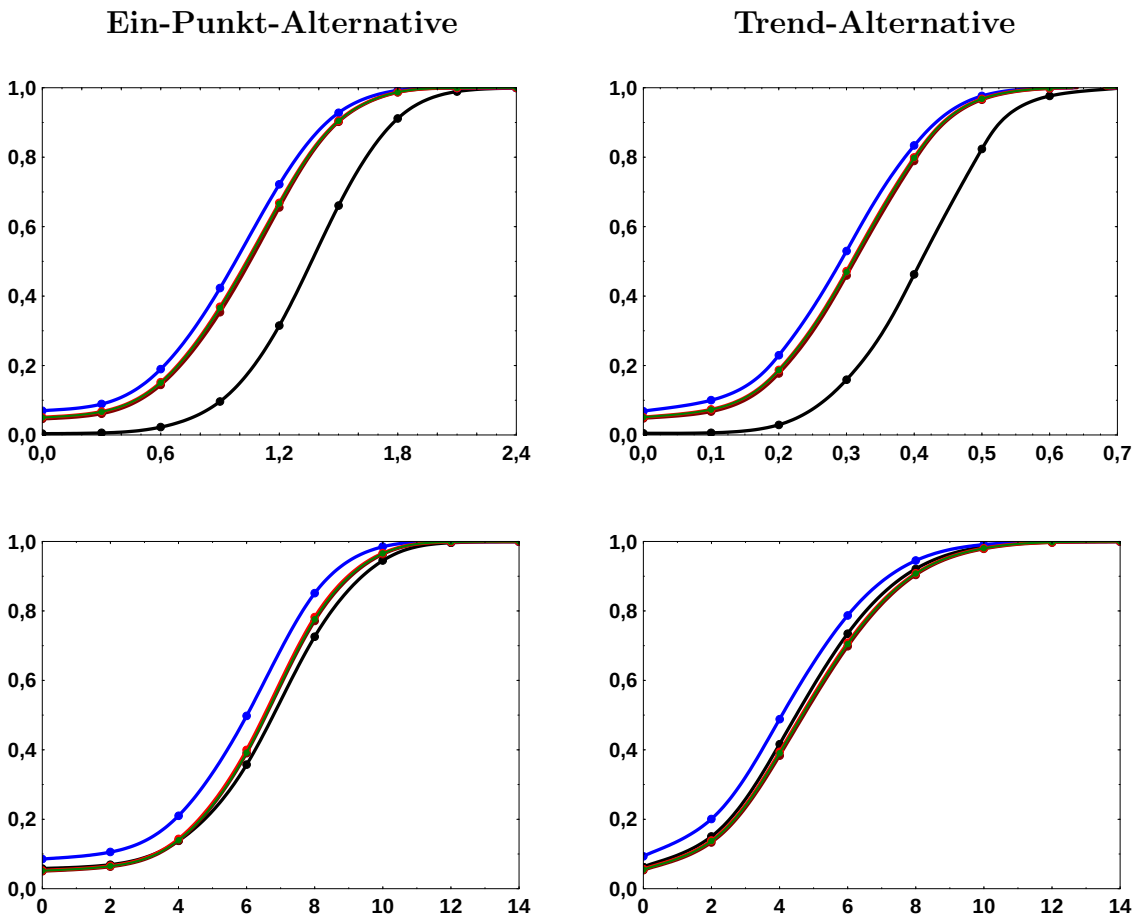


Abbildung 11.2: Aufgetragen ist die Power unter Ein-Punkt-Alternative (links) und Trend-Alternative (rechts). Die beiden oberen Graphen wurden unter $\mathcal{N}(\mathbf{0}, \mathbf{I})$-verteilten Zufallsvektoren für die multivariate Hypothese erzeugt, während die unteren, gemäß Tabelle 11.7, im Bai-Saranadasa Modell mit zentrierten exponential(1)-verteilten Z_i und Toeplitzstruktur im HDF1-Design erzeugt wurden. Die Beschriftung der Verfahren und die weiteren Parameter ergeben sich aus der Abbildung 11.1.

11.3 Gütevergleich der Schätzer

In dieser Arbeit wurden zahreiche Schätzer vorgestellt und ihre Konsistenz gezeigt. In diesem Abschnitt soll nun die tatsächliche Güte überprüft und verglichen werden.

11.3.1 Asymptotik über die Dimension

Zunächst soll die Dimensionsstabilität der hergeleiteten Schätzer im Bai-Saranadasa Modell überprüft werden. In vorherigen Arbeiten wurden Schätzer $\widehat{\theta}$ für θ mit $Var(\widehat{\theta}/\theta) \to 0$ für $n \to \infty$ als „dimensionsstabil“ bezeichnet. Diese Forderung erweist sich beispielsweise für den Schätzer B_3 für $Sp(\mathbf{V}^3)$ als zu stark. Demgegenüber konnte aber gezeigt werden, dass es ausreicht zu zeigen, dass $Var(B_3 \,/\, Sp^{3/2}(\mathbf{V}^2)) \to 0$ für $n \to \infty$ gleichmäßig in d. In Simulationen lässt sich das instabile Verhalten des B_3-Schätzers für steigende Dimension d gut erkennen. So ist im folgenden Graph die empirische Varianz der Schätzer unter multivariater Normalverteilung untersucht worden.

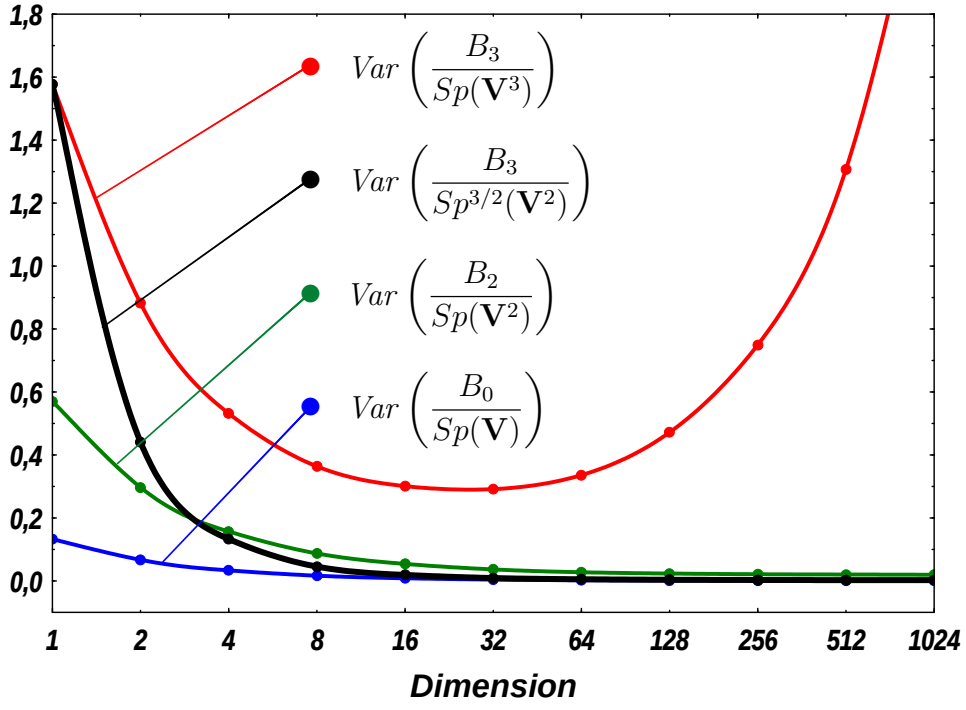


Abbildung 11.3: Empirische Varianz der Schätzer B_0, B_2 und B_3 bei Stichprobenumfang von $n = 15$, 100000 Simulationsdurchläufen und multivariater Standard-Normalverteilung mit Kovarianzmatrix und zugehöriger Hypothesenmatrix $I(d)$ für steigende Dimensionen.

Für den Schätzer B_2 konnte hingegen im Bai-Saranadasa Modell Dimensionsstabilität gezeigt werden, wie man auch in der Graphik gut erkennen kann. Das Gleiche gilt für den Schätzer B_0, für den gezeigt wurde, dass $Var(B_0 \,/\, Sp(\mathbf{V}))$ für $n \to \infty$ gleichmäßig in d gegen 1 konvergiert. Falls die Eigenwerte nicht stärker als von Ordnung $\sqrt{d}$ wachsen, folgt sogar, dass eine Konsistenz für $d \to \infty$ unabhängig von n vorliegt.
Im Allgemeinen werden im Bai-Saranadasa Modell die Schätzer für steigende Dimension stets besser, wenn die Eigenwerte beschränkt sind.

Demgegenüber lässt sich die Güte der Spur-Schätzer betrachten, falls nun anstelle des Bai-Saranadasa Modells eine degenerierte Verteilung nach Kapitel 9 simuliert wird. Dafür wurde analog die empirische Varianz der Schätzer B_0, B_2 und B_3 geplottet. Anstelle der multivariaten Normalverteilung wurde die Verteilung nach Konstruktion in (9.16) mit normalen Rändern gewählt.

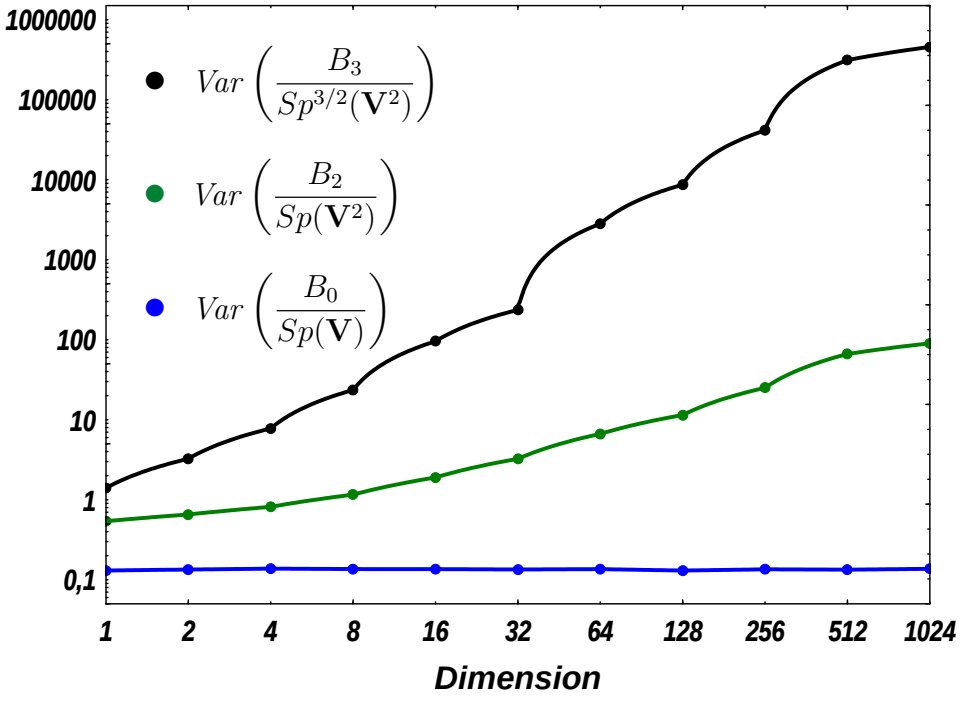


Abbildung 11.4: Empirische Varianz der Schätzer B_0, B_2 und B_3 bei Stichprobenumfang von $n = 15$, 10000 Simulationsdurchläufen und degenerierter Normalverteilung (mit Einheitsmatrix als Kovarianzmatrix und Hypothesenmatrix) für steigende Dimensionen. Die empirische Varianz wurde logarithmisch skaliert.

Es lässt sich gut erkennen, dass die Konsistenz der Schätzer B_2 und B_3 mit steigender Dimension verloren geht. Erklären lässt sich dies über die vierten Momente der standardisierten A_{kl}, welche nicht in d beschränkt sind, siehe Tabelle (9.1).
Der B_0-Schätzer bleibt hingegen stabil. Die Konvergenzgeschwindigleit nimmt allerdings nicht mehr mit steigender Dimension zu. Dies ist auch nicht zu erwarten, da die Quadratformen A_{kk} im degenerierten Modell aus (9.16) nur von der ersten Komponente abhängen, weshalb keine Veränderung der Varianz bezüglich der Dimension stattfindet.

11.3.2 Vergleich der Schätzer für $Sp(\mathbf{\Sigma}^2)$

Elementar für die Güte der Teststatistik Z_n aus 6.4 ist die Qualität der Schätzung von $Sp(\mathbf{\Sigma}^2)$. Deshalb soll nun die Varianz der folgenden verschiedenen Schätzer unter Hypothese untersucht werden:

- Der Schätzer B_2
- Der von Chen-Qin für den Zwei-Stichprobenfall entwickelte $\widehat{Sp(\mathbf{V}^2)}$. Dieser Schätzer ist nur im Ein-Stichprobenfall dimensionsstabil.
- Der von Becker (2010) [2] für den Zweistichprobenfall entwickelte Schätzer

$$B_2^{(1)} := (4n(n-1)(n-2)(n-3))^{-1} \sum_{k\neq l\neq s\neq t}^{n} \left((\mathbf{Y}_k - \mathbf{Y}_l)'(\mathbf{Y}_s - \mathbf{Y}_t)\right)^2 \quad (11.4)$$

  Dieser Schätzer ist auch unter Alternative anwendbar.

Selbst bei Stichprobenumfängen $n < 10$, ist in den folgenden Abbildungen nur eine leicht erhöhte empirische Varianz zu erkennen. Dennoch sind Teststatistiken, die nicht den B_2-Schätzer verwenden etwas liberaler, wie sich in Niveausimulationen zeigt. Dies kann neben der leicht erhöhten Varianz auch an einer Abhängigkeit zwischen den Schätzern und $Q_n - B_0$ liegen. Diese verschwindet nach dem Slutskischen Satz aber für wachsende n.

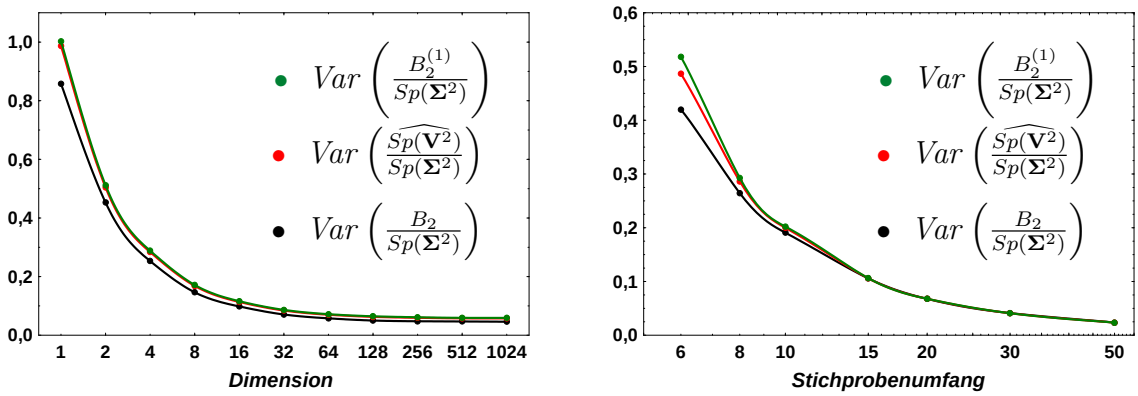


Abbildung 11.5: Im linken Graph ist die empirische Varianz der Schätzer für steigende Dimensionen zu sehen. Simuliert wurde eine multivariate Normalverteilung mit Einheitsmatrix als Kovarianzmatrix und Hypothesenmatrix bei festem Stichprobenumfang $n = 10$. Im rechten Graph hingegen wurde bei fester Dimension $d = 32$ steigende Stichprobenumfänge simuliert. Hier wurden unabhängige Exponential(1)-verteilte Zufallsvariablen im HDF1-Design betrachtet. Der Simulationsumfang lag bei 10000 Wiederholungen. Die x-Achsen sind logarithmisch skaliert.

11.3.3 Vergleich Freiheitsgradschätzer

Für die Approximation der Teststatistik ist es erforderlich, dass der Freiheitsgrad f_{box} bzw. f_{pear} möglichst exakt und effizient geschätzt wird. Um die theoretischen Resultate der Arbeit zu unterlegen, werden deshalb die Schätzer für die Freiheitsgrade simuliert. Außerdem soll ein Vergleich von unterschiedlichen Variationen der Schätzung erfolgen. Betrachtet werden nun Schätzer für $g = 1/\sqrt{f}$. Dies bietet zum einen den Vorteil im Intervall [0,1] eine einfache Vergleichbarkeit zu haben. Vor allem aber lässt sich auf diese Weise am besten den durch den Schätzfehler induzierten Fehler des Quantil für die Testentscheidung kontrollieren. In Abschnitt 7.3 wurde gezeigt, dass sich Schätzfehler von g nahezu linear auf den Fehler des Quantil der $\kappa(g)$-Verteilung übertragen, falls $\alpha \geq 0.95$.
Als Schätzer für g werden folgende untersucht:

- Der Standard-Schätzer für die Pearsonapproximation aus Abschnitt 7.4

$$\widehat{g}_{pear} := \frac{B_3}{B_2^{3/2}}$$

- Der einfach Plug-in-Schätzer der sich aus dem Boxschen Freiheitsgradschätzer $\widehat{f}_{box} = B_1/B_2$ ergibt

$$\widehat{g}_{box1} := \sqrt{\frac{B_2}{B_1}}$$

- Eine Modifikation davon, welche $\sqrt{B_1}$ durch den natürlichen B_0-Schätzer ersetzt

$$\widehat{g}_{box2} := \frac{\sqrt{B_2}}{B_0}$$

Außerdem sollen zum Vergleich dazu zusätzlich die Schätzer für $h = 1/f$ untersucht werden:

- Der in Abschnitt 7.4.1 hergeleitete Schätzer für h_{pear}

$$\widehat{h}_{pear1} := \frac{\widehat{Sp^2(\mathbf{V}^3)}}{\widehat{Sp^3(\mathbf{V}^2)}}$$

- Der Schätzer der sich durch Quadrieren von von $\widehat{g}_{pear}$ ergibt.

$$\widehat{h}_{pear2} := \frac{B_3^2}{B_2^3}$$

- Der Standard-Schätzer für die Boxapproximation aus (6.5)

$$\widehat{h}_{box} := \frac{B_2}{B_1}$$

Diese Schätzer für h stehen auch stellvertretend für Schätzer für f. Da keine Methode bekannt ist, welche einen Quotienten sinnvoll schätzt, ist man stets darauf angewiesen Zähler und Nenner getrennt zu schätzen. Als Schätzer wird schließlich der Quotient beider gebildet und gegebenenfalls noch eine Korrektur der Verzerrung eingerechnet. Somit erhält man durch Invertieren aus jedem Schätzer für h einen für f und umgekehrt. Demgegenüber unterscheiden sich Schätzer von h und g, da man für Potenzen nach Satz B.0.7 auf einfachem Wege neue Schätzer finden kann. Diese Schätzer sind unverzerrt, besitzen dafür allerdings eine größere Varianz und sind meist sehr ineffizient zu berechnen. Deshalb soll nun eine praktische Untersuchung für die Performance bei unterschiedlichen Stichprobenumfängen erfolgen. Dafür wurden Zufallsvektoren $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ der Dimension $d = 32$ simuliert und die empirische Standardabweichung der Schätzer aufgetragen.

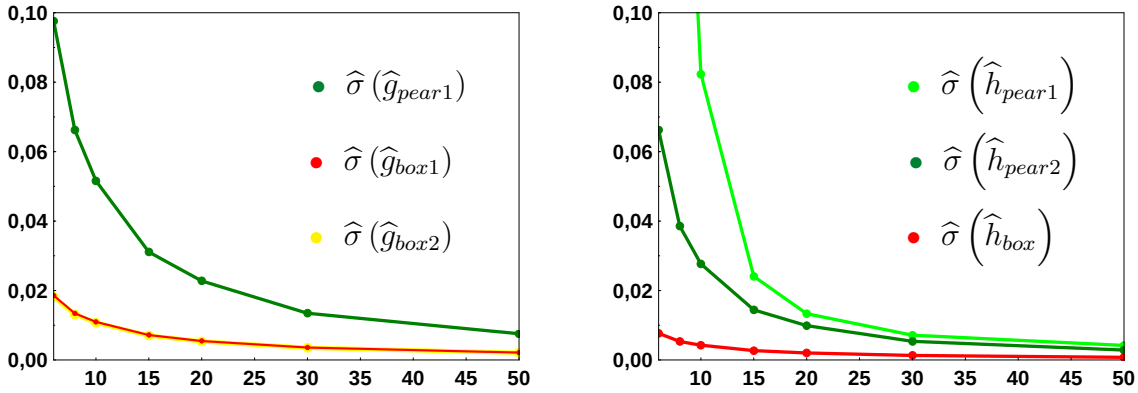


Abbildung 11.6: Im linken Graph ist die empirische Standardabweichung der Schätzer $\widehat{g}_{pear}$, $\widehat{g}_{box1}$, $\widehat{g}_{box2}$ zu sehen und im rechten die von $\widehat{h}_{pear1}$, $\widehat{h}_{pear2}$, $\widehat{h}_{box}$. Simuliert wurde mit einer multivariaten Standardnormalverteilung der Dimension 32 mit Einheitsmatrix als Kovarianz- und Hypothesenmatrix bei steigendem Stichprobenumfang.

Neben der Varianz bzw. Standardabweichung der Schätzer ist ebenfalls die angesprochene Verzerrung ein wichtiges Gütekriterium. Hier zeigt sich ein nahezu exakt gleiches Ergebnis wie bei den Abbildungen der Varianzen, weshalb auf die Graphik verzichtet wird.

Somit lässt sich schlussfolgern, dass die Schätzung von g_{box} durchaus Vorteile im Vergleich zur Schätzung von g_{pear} hat. Für Stichprobenumfänge $n > 10$ fällt dies allerdings kaum noch ins Gewicht. Außerdem haben die Simulationsergebnisse der Pearsonapproximation gezeigt, dass diese Schwankungen so gut wie keine Auswirkungen auf die Statistik haben.
Desweiteren bleibt festzustellen, dass der Vorteil der erwartungstreuen Schätzung des Nenners durch eine erhöhte Varianz zunichte gemacht werden. Insgesamt ist die erwartete Verzerrung des Quotienten dadurch sogar größer, da die erwartete Verzerrung gerade mit einer größeren Streuung des Nenners einhergeht. Der Schätzer $\widehat{h}_{pear1}$ kann somit als praxisuntauglich klassifiziert werden. Selbst der Nutzen des B_1-Schätzers erscheint fraglich, da $\widehat{g}_{box2}$ sogar minimal besser performt als $\widehat{g}_{box1}$.

12 Software

12.1 Makro HD-Fi

Um mit den Ergebnisse dieser Arbeit praktische Auswertungen durchführen zu können, wurde das SAS-Makro HD-Fi geschrieben. Der Quellcode des Makros ist in Abschnitt D im Anhang zu finden. Berechnet wird die modifizierte Chen-Qin-Teststatistik Z_n aus Abschnitt 6.4 und die p-Werte, die sich aus einer $\varkappa(f)$-Verteilung ergeben. Dabei werden diese sowohl mit dem Boxschen, als auch mit dem Pearsonschen Freiheitsgrad f berechnet. Die Freiheitsgrade werden so beschnitten, dass diese in dem zulässigen Wertebereich aus 8.1 liegen.
Zum Vergleich wird ebenfalls der p-Wert der klassischen ANOVA-Typ-Statistik angegeben, welche standardmäßig in SAS zur Auswertung bereit gestellt wird. So liefert dieser Wert bei orthogonalen, Voll-faktoriellen Designs ohne Messwiederholungen die gleichen Resultate wie die in SAS implementierte Prozedur „mixed“ mit der Option „anovaf“, die wie folgt aufzurufen ist:

```
proc mixed data=MARATHONL method=mivque0 anovaf;
class Stimulation Trainingspause Patient Zeitprofil;
model Cortisolkonz = Trainingspause|Stimulation|Zeitprofil;
repeated/ sub=Patient type=UN;
run;
```

Zum Aufruf des Makros müssen die Daten in der SAS-üblichen Form für longitudinale Daten vorliegen. Das bedeutet, dass sämtliche Messwerte in einer Spalte stehen. In den anderen Spalten sind die Ausprägungen der Faktoren gespeichert. Aufgerufen wird das Makro für die Variablen aus dem Beispiel „Cortisolkonzentration im Blutplasma“ folgendermaßen:

```
%nn_hd_fi(data=MARATHONL, var=Cortisolkonz, TIME1=Trainingspause,
TIME2=Stimulation, TIME3=Zeitprofil, SUBJECT = Patient);
```

Dabei können bis zu vier Faktoren angegeben werden. Faktorstufen von Faktoren, die nicht angegeben werden, werden als Messwiederholungen interpretiert. Somit müssen die Werte des Zeitprofils nicht von Hand gemittel werden, wenn dieser Faktor nicht in die Auswertung aufgenommen werden soll.

Diese Eingabe erzeugt den folgenden Output:

Das SAS System

16:38 Saturday, September 17, 2011

HD_Fi ––– subjects x Time_1 x ... x Time_i
Time_1, ... Time_i: fixed, subjects: random

SAS–datafile–name: MARATHONL
Response variable: Cortisolkonz

Class Level Information

class		levels	values
Time1	Trainingspause	2	0 1
Time2	Stimulation	2	0 1
Time3	Zeitprofil	7	t0 t30 t60 t90 t120 t180 t240
SUBJECT	Patient	12	1 2 3 4 5 6 7 8 9 10 11 12

Total number of observations 336
Total number of subjects 12

Chi–Quadrat–Approximation

Comparison of Z_n–Approximations with classic ANOVA

	Z_n	f_box	f_pear	p–val(ANOVA)	p–val(box)	p–val(pear)
Time1 (T1)	3.0430	1.0000	1.0000	.01397	.02128	.02128
Time2 (T2)	4.4594	1.0000	1.0000	.00024	.00687	.00687
T1*T2	−.7523	1.0000	1.0000	.88099	1.0000	1.0000
Time3 (T3)	1.4272	2.3241	1.7395	.08915	.08900	.08745
T1*T3	−.2955	3.5507	3.4510	.48356	.52532	.52398
T2*T3	3.1938	3.2494	4.1696	.01329	.01227	.01089
T1*T2*T3	−.6925	3.2458	2.3380	.65223	.72800	.73108

Abbildung 12.1: Ausgabe des SAS-Makros HD-Fi für den Datensatz der Beispiels Cortisolkonzentration im Blutserum

12.2 Auswertung des Beispiels

Für die Auswertung der Beispiels aus Kapitel 2 waren sämtliche Hypothesen des HD-F3-Designs zu testen. Wird das geschriebenen Makro auf die Daten der Cortisolkonzentration im Blutserum angewendet, so ergeben sich die in Abbildung 12.1 dargestellten p-Werte.
Zunächst sei bemerkt, dass sich die p-Werte nicht substantiell unterscheiden. Während zwischen Box- und Pearson-Approximation der Teststatistik Z_n überhaupt nur marginale Unterschiede festzustellen sind, kann man für die klassische ANOVA-Typ-Statistik bei den Haupteffekten „Trainingspause “ (T1) und „Stimulation “ (T2) leicht signifikantere p-Werte feststellen. Diese erklären sich allerdings durch eine leichte Liberalität der klassischen ANOVA-Typ-Statistik bei niedrigen Dimensionen.

Bei der Betrachtung der Mittelwerte bezüglich der Trainingspause und der Stimulation ergibt sich folgendes Bild:

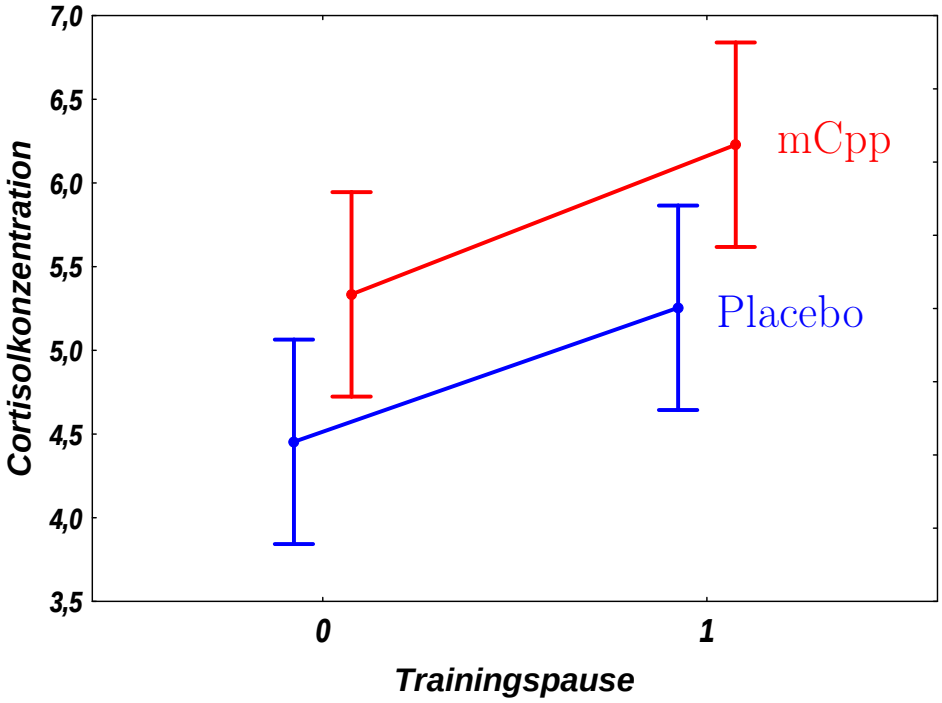


Abbildung 12.2: Geplottet wurden die geschätzten Effekte bezüglich der Wechselwirkung zwischen der Trainingspause TP und der Stimulation ST. Dabei wurde die Cortisolkonzentration über die Zeitpunkte des 4-Stunden-Profil und über die einzelnen Sportler gemittelt. Die Whiskers dienen als Indikator für die Streuung.

In Bezug auf die Studie konnte nun der vermuteten Effekt der Trainingspause nachgewiesen werden. So findet ein signifikanter Anstieg der Cortisolkonzentration über die zweiwöchige Pause statt. Der Effekt durch die Stimulation mit mCpp erweist sich als hochsignifikant. Auch im 4-Stunden-Profil des Cortisolspiegels lässt sich noch ein leichter (schwachsignifikanter) Effekt (T3) beobachten. Beide Effekte waren zu erwarten und bestätigen den Versuchsaufbau.
Demgegenüber konnte keinerlei Wechselwirkung zwischen der Trainingspause und der Stimulation festgestellt werden. Die sehr hohen p-Werte deuten darauf hin, dass hier tatsächlich kaum ein Effekt vorzuliegen scheint. An Zeitpunkten nach der Trainingspause werden durch eine Verabreichung von mCpp Extremwerte bei der Cortisolkonzentration festgestellt, die bei gesunden Menschen nicht vorkommen. Dies deutet auf eine grundsätzliche Störung des Hormonspiegels hin, welche durch die Trainingspause ausgelöst wird. Hinweise auf eine Veränderung des 4-Stunden-Profils durch die Trainingspause konnten hingegen nicht festgestellt werden. So lassen die p-Werte für $(TP \times PR)$ und für $(TP \times ST \times PR)$ keine Effekte deuten. Lediglich die Stimulation mit mCpp scheint die Zeitprofile der Cortisolkonzentration zu verändern, wie der zugehörige Effekt $(ST \times PR)$ zeigt. Dies ist mit den direkten Auswirkungen der Substanz mCpp auf die Cortisolproduktion im Körper zu erklären.

13 Zusammenfassung und Ausblick

In dieser Arbeit wurden Verfahren für den Ein-Stichprobenfall vorgestellt, welche möglichst wenig Annahmen an die zugrundeliegende Verteilung stellen. Speziell Einschränkungen bezüglich der Dimension sollten überwunden werden, so dass Testverfahren problemlos auf hochdimensionale Daten angewendet werden können.
Die Statistik Z_n wurde als Modifikation der Teststatistik von Chen und Qin entwickelt. Außerdem wurde anstelle einer Normalapproximation von Z_n eine Approximation mit einer standardisierten Chi-Quadrat-Verteilung motiviert. Mittels Boxapproximation und einer geeigneten Schätzung für den Freiheitsgrad f_{box} konnte ein sehr stabiles Verfahren angegeben werden, welches selbst bei kleinsten Stichprobenumfängen sehr gute Ergebnisse liefert. Darüber hinaus wurde mit der Pearsonapproximation eine weiter verbesserte Approximation implementiert, die gleichzeitig eine Überprüfung der Chen-Qin-Bedingung erlaubt. Die Stabilität der dafür nötigen Schätzung der Verteilung $\kappa(g)$ bzw. des Parameters g_{pear} konnte nachgewiesen werden, wodurch das Verfahren die geringsten Voraussetzungen benötigt und unter diesen nahezu optimale Ergebnisse liefert.

Für den Fall, dass im Hochdimensionalen die Chen-Qin-Bedingung nicht erfüllt ist, konnte die Lücke in der Theorie für das Bai-Saranadasa Modell weitgehend geschlossen werden. Trotzdem bleibt hier Raum für weitere Überlegungen, um diese Lücke vollständig zu schließen. Die guten Simulationsergebnisse legen den Schluss nahe, dass die Anwendbarkeit der Verfahren hier gegeben ist.
Ein Ansatz hierfür könnte es sein, das Mischmodell aus Abschnitt 7.5.2 mit normal- und nicht-normal- verteilten (Block-)Effekten zu verallgemeinern. So existieren unter Umständen schwächere Bedingungen, die gänzlich auf die Annahme von Normalverteilung verzichten. Desweiteren ist zu prüfen, inwieweit die Chen-Qin-Bedingung verallgemeinert werden kann. Ansätze hierfür bietet unter anderem die Arbeit von Bhansali, Giraitis und Kokoszka (2007) [3].

Eine weitere Aufgabe wird es sein, Verfahren zu entwerfen, welche robust bezüglich fehlender Messwerte sind. Bei repeated measures über mehrere Zeitpunkte bietet sich unter Umständen ein Interpolationsverfahren an. Interpolation ist im Bai-Saranadasa Modell generell möglich. Dennoch besteht eine große Schwierigkeit, die (zufällig) fehlenden Werte mit dem Bai-Saranadasa Modell in Einklang zu bringen.

Das neu entwickelte Testverfahren ist nicht skalierungsinvariant. Dadurch ist es im Ein-Gruppen-Design nur dann zum Test geeignet, wenn alle Messwerte in der gleichen Einheit vorliegen. Üblicherweise wird das Testverfahren für strukturierte repeated measures angewendet. Sind nun multivariate Hypothesen zu testen, wenn die Daten nicht ohne Weiteres auf die gleiche Einheit transformiert werden können, wird ein multivariater Test benötigt. Die Methodik des multivariaten Test von Srivastava bietet hier unter Umständen eine Möglichkeit, den entwickelten Test dementspechend zu erweitern.

Vor allem aber sind die Ergebnisse dieser Arbeit auf den Zwei- und Mehrstichprobenfall übertragbar. Die Ein-Gruppen-Tests von Chen-Qin und Bai-Saranadasa sind nur Nebenprodukte von Verfahren des Zwei-Stichprobenfalls. Im Kapitel 10 über Robustheitsaussagen für symmetrische Verteilungen wurde bereits der Bezug zum Zwei-Stichprobenfall diskutiert. Neben der Robustheit bezüglich Nichtnormalverteilung scheinen weitere starke Robustheitsaussagen unter der Annahme $F_1 = F_2$ möglich. In erster Linie gilt es allerdings die Normalapproximation der Teststatistik durch eine Approximation mit einer standardisierten Chi-Quadrat-Verteilung zu ersetzen, um Verfahren unabhängig von der Dimension zu erhalten.

Ebenfalls ungeklärt ist, ob sich eine bessere Adaption von schiefen Verteilungen bewerkstelligen lässt. Für einfache T-Statistiken kann man mit Hilfe von Edgeworth Expansions stark verbesserte Resultate erzielen. Außerdem sind genauere Fehlerrechnungen wünschenswert, gerade in Bezug auf den Martingalgrenzwertsatz von Hall und Heyde (1980) [12] und die Approximationen von Summen von gewichteten Chi-Quadrat-verteilten Zufallsvariablen.

A Grundlagen

A.1 $\mathcal{O}$-Notation

Die Landau Symbole $\mathcal{O}$ für asymptotisch obere Schranken und $\mathbf{o}$ für asymptotische Vernachlässigbarkeit definieren sich wie folgt:

- $f(n) = \mathcal{O}(g(n))$, dass ein c und ein n_0 existieren, so dass für alle $n > n_0 \Rightarrow |f(n)| \leq c \cdot g(n)$.
- $f(n) = \mathbf{o}(g(n))$, dass für alle c ein n_0 existiert, so dass für alle $n > n_0 \Rightarrow |f(n)| \leq c \cdot g(n)$.

Diese Notation erweist sich als äußerst hilfreich, da sich für sie folgende einfache Rechenregeln ergeben:

- $\mathcal{O}(g(n) \cdot h(n)) = \mathcal{O}(g(n)) \cdot \mathcal{O}(h(n))$
- $\mathcal{O}(g(n) + h(n)) = \mathcal{O}(g(n)) + \mathcal{O}(h(n))$

Diese Regeln lassen sich nun auch anwenden, wenn man Asymptotiken über mehrere Parameter betrachtet. So ergibt sich als gemeinsame Asymptotik über den Stichprobenumfang $\mathcal{O}(n^\alpha)$, $\alpha \in \mathbb{R}$ und über die Dimension $\mathcal{O}(g(d))$

- $\mathcal{O}(n^\alpha \cdot g(d)) = \mathcal{O}(n^\alpha) \cdot \mathcal{O}(g(d))$
- $\mathcal{O}(n^\alpha + g(d)) = \mathcal{O}(n^\alpha) + \mathcal{O}(g(d))$

Die Funktion $g(d)$ ist hierbei gewöhnlich eine Funktion in Abhängigkeit der Kovarianzmatrix $\mathbf{\Sigma}_d$. Es sei bemerkt, dass die Gültigkeit auch bei beliebigen Restriktonen zwischen n und d gegeben ist.

A.2 W-Theorie

Satz A.2.1 (Slutzky) *Seien $\mathbf{X}_n, \mathbf{X}_n, \mathbf{X}$ Zufallsvekoren. Falls $\mathbf{X}_n \xrightarrow{w} \mathbf{X}$ und $\mathbf{Y}_n \xrightarrow{w} c$, $c \in \mathbb{R}$. Dann gilt*

1. $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{w} \mathbf{X} + c$
2. $\mathbf{X}_n \mathbf{Y}_n \xrightarrow{w} c\mathbf{X}$

3. $\mathbf{Y}_n^{-1}\mathbf{X}_n \xrightarrow{w} c^{-1}\mathbf{X}$, *für* $c \neq 0$

Beweis: Siehe Satz 2.8 in van der Vaart (1998) [20]

Satz A.2.2 (continuous mapping) *Sei* $g : \mathbb{R}^d \mapsto \mathbb{R}^e$. *in jedem Punkt einer Menge* C, *so dass* $P(\mathbf{X} \in C) = 1$. *stetig. Dann gilt*

1. *Falls* $\mathbf{X}_n \xrightarrow{w} \mathbf{X}$ *dann folgt* $g(\mathbf{X}_n) \xrightarrow{w} g(\mathbf{X})$

2. *Falls* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ *dann folgt* $g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{X})$

3. *Falls* $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$ *dann folgt* $g(\mathbf{X}_n) \xrightarrow{a.s.} g(\mathbf{X})$

Beweis: Siehe Satz 2.3 in van der Vaart (1998) [20]

Satz A.2.3 (Momente der Normal-Verteilung) *Sei* $Z \sim \mathcal{N}(0,1)$. *Dann folgt für die Momente höherer Ordnung:*

1. $E(Z^3) = 0$

2. $E(Z^4) = 3$

3. $E(Z^6) = 15$

4. $E(Z^8) = 105$

Beweis: Errechnet sich durch mehrmaliges partielles Integrieren.

Satz A.2.4 (Momente der zentralen Chi-Quadrat-Verteilung) *Sei* $C \sim \chi_f^2$-*verteilt mit Freiheitsgrad* f, *dann gilt für den Erwartungswert* $\mu = E(C)$, *die zentralen Momente und die Schiefe* ν:

1. $E(C) = f$

2. $E(C - \mu)^2 = 2f$

3. $E(C - \mu)^3 = 8f$

4. $\nu(C) = \frac{2\sqrt{2}}{\sqrt{f}}$

Beweis: OBdA sei $f \in \mathbb{N}$ und $C = \sum_{i=1}^{f} Z_i^2$, mit Z_i unabhängig $\mathcal{N}(0,1)$-verteilt.

1. $E\left(\sum_{i=1}^{f} Z_i^2\right) = f \cdot Var(Z_i) = f$

2. $E\left(\sum_{i=1}^{f} Z_i^2 - f\right)^2 = E\left(\sum_{i=1}^{f}\sum_{j=1}^{f}(Z_i^2 - 1)(Z_j^2 - 1)\right) = E\left(\sum_{i=1}^{f}(Z_i^2 - 1)^2\right)$

   $f \cdot E(Z_i^4 - 2Z_i^2 + 1) = f \cdot (3 - 2 + 1) = 2f$

3. Analog zu 2. mit $E(Z_i^6 - 3Z_i^4 + 3Z_i^2 - 1) = 15 - 9 + 3 - 1 = 8$

□

Definition A.2.5 (Kroneckerprodukt) *Für zwei Matrizen*

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dm} \end{pmatrix}_{(d\times m)} \quad \textit{und} \quad \mathbf{B} = \begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{e1} & \cdots & b_{ep} \end{pmatrix}_{(e\times p)}$$

sei das Kroneckerprodukt gegeben durch

$$\mathbf{A} \otimes \mathbf{B} := \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{d1}\mathbf{B} & \cdots & a_{dm}\mathbf{B} \end{pmatrix}_{(de\times mp)} \tag{A.1}$$

Satz A.2.6 (Rechenregeln für das Kroneckerprodukt) *Es gilt:*

1. $(A \otimes B)' = (A' \otimes B')$
2. $(A \otimes B) \cdot (C \otimes D) = (A \cdot C \otimes B \cdot D)$

Beweis: Siehe Result 2.8.1 auf Seite 67 in Ravishanker und Dey (2001) [17].

B U-Statistiken

Die Methodik beim Konstruieren der Spurschätzer, wie beispielsweise B_0, B_2 und B_3, ist es, diese stets in Form einer U-Statistik darzustellen, wie sie unter anderem in Kap.12 van der Vaart (1998) [20] eingeführt werden.
Zunächst wird die Vorgehensweise noch einmal dargestellt, um dann mit der Methodik stärkere Resultate zu erzielen. Um einen Parameter $\theta \in \mathbb{R}$ mit den Datenvektoren $\mathbf{Y}_k$ zu schätzen, lässt sich eine Funktion $h : (\mathbf{Y}_1,...,\mathbf{Y}_r) \to \mathbb{R}$ mit $E(h(\mathbf{Y}_1,...,\mathbf{Y}_r)) = \theta$, welche permutationsinvariant bezüglich der r Argemente ist, nutzen. Diese Funktion wird „Kern“ genannt und dient als Grundlage, um damit einen Schätzer für den Parameter θ zu konstruieren. Dieser Schätzer wird als „U-Statistik“ bezeichnet und sei definiert als

$$U = \binom{n}{r}^{-1} \sum_{k_1<k_2<..<k_r}^{n} h\left(\mathbf{Y}_{k_1}, \mathbf{Y}_{k_2},.., \mathbf{Y}_{k_r}\right) \tag{B.1}$$

Als Beispiel wird $h(\mathbf{Y}_1,\mathbf{Y}_2) = \mathbf{Y}_1'\mathbf{Y}_2 \cdot \mathbf{Y}_1'\mathbf{Y}_2 = A_{12}^2$ betrachtet. Dann erhält man für U den B_2-Schätzer für $Sp(\mathbf{V}^2)$. Somit lassen sich allgemeine Ergebnisse über U-Statistiken auf die Schätzer anwenden. So gilt für die Varianz einer U-Statistik nach Theorem 12.3 van der Vaart (1998) [20]

$$Var(U) = \sum_{k_1<..<k_r}^{n} \sum_{l_1<..<l_r}^{n} Cov\left(h\left(\mathbf{Y}_{k_1},..,\mathbf{Y}_{k_r}\right), h\left(\mathbf{Y}_{l_1},..,\mathbf{Y}_{l_r}\right)\right) \tag{B.2}$$

$$= \binom{n}{r}^{-1} \sum_{c=0}^{r} \binom{r}{c}\binom{n-r}{r-c}\zeta_c\,, \tag{B.3}$$

wobei ζ_c die Kovarianz zwischen $h\left(\mathbf{Y}_{k_1},..,\mathbf{Y}_{k_r}\right)$ und $h\left(\mathbf{Y}_{l_1},..,\mathbf{Y}_{l_r}\right)$ ist, wenn c Indizes der beiden Ausdrücke übereinstimmen. Für $c = 0$ sind sämtliche Indizes verschieden und somit folgt aus der Unabhängigkeit $\zeta_0 = 0$. Weiter gilt somit

$$Var(U) = \sum_{c=1}^{r} \mathcal{O}(n^{-c})\zeta_c \tag{B.4}$$

Es folgt die Konsistenz der Schätzer für $n \to \infty$. Allerdings induziert dies nicht die gleich gute Konsistenz der Schätzer in Abhängigkeit von d, schließlich sind die ζ_i nicht in d beschränkt. Im Hochdimensionalen für $d \gg n$ wird in den Sätzen 5.2.5, 6.2.2 und 7.4.2 die Konsistenz der B_0-, B_2- und B_3-Schätzer deshalb explizit ausgerechnet. Alternativ kann man allerdings auch einfach zeigen, dass in diesen Fällen die ζ_c gerade die höheren Momente der A_{kl} und ihrer Linearkombinationen aus Lemma C.2.1 sind. Somit würde man aus (B.3) auch direkt die Konvergenz von beispielsweise $B_2/Sp(\mathbf{V}^2)$ und $B_3/Sp^{3/2}(\mathbf{V}^2)$ folgern können.
Ziel ist es nun, auf einfachem Wege Schätzer für ganzzahlige Potenzen α von den Spuren von $\mathbf{V}$ und ihren Vielfachen zu erhalten, wie die der Parameter $\theta = Sp^2(\mathbf{V}),\ \ Sp^3(\mathbf{V}^2)$ oder $Sp^2(\mathbf{V}^3)$. Zunächst lässt sich, sofern man einen Schätzer für θ besitzt, sofort der triviale Schätzer $\left(\widehat{\theta}\right)^{\alpha}$ für θ^α bilden. Dieser verliert zwar die Erwartungstreue des ursprünglichen Schätzers, sofern vorhanden, Konvergenzeigenschaften übertragen sich aber nach dem continuous mapping theorem für $g(x) = x^\alpha$. Auch in Simulationen erzielen die so gebildeten Schätzer gute Resultate und häufig ist diese Art der Schätzung auch die einzige bekannte Möglichkeit. Der Quotient oder die Wurzel einer Zufallsvariablen ist ein typisches Beispiel, wo dies der Fall ist und man keine Erwartungstreue mehr fordern kann. Im Falle von ganzzahligen Potenzen θ^α hingegen ist es möglich, aus den U-Statistiken für θ verbesserte Schätzer für θ^α zu konstruieren. Als Beispiel sei der Schätzer

$$B_1 := \frac{2}{n(n-1)} \sum_{k<l}^{n} A_{kk} A_{ll} \tag{B.5}$$

für $Sp^2(\mathbf{V})$ angeführt, welcher im Gegensatz zu $(B_0)^2$ erwartungstreu ist und eine kleinere Varianz hat. Dies lässt sich so verallgemeinern, dass sich aus einer U-Statistik, als konsistentem Schätzer für einen Parameter θ, eine U-Statistik für den Parameter θ^α bilden lässt. Sei

$$U_\alpha := \binom{n}{\alpha r}^{-1} \sum_{\substack{k_1<k_2<.....<k_r< \\ k_{r+1}<..<k_{2r}<...< \\ k_{(\alpha-1)r+1}<..<k_{\alpha r}}}^{n} h(\mathbf{Y}_{k_1},\mathbf{Y}_{k_2},..,\mathbf{Y}_{k_r}) \cdot h(\mathbf{Y}_{k_{r+1}},..;,\mathbf{Y}_{k_{2r}}) \cdots \cdots h(\mathbf{Y}_{k_{(\alpha-1)r+1}},..,\mathbf{Y}_{k_{\alpha r}})$$

Mit $w := \alpha \cdot r$ und $\widetilde{h}(\mathbf{Y}_1,...,\mathbf{Y}_w) := h(\mathbf{Y}_1,...,\mathbf{Y}_r) \cdots h(\mathbf{Y}_{w-r+1},...,\mathbf{Y}_w)$ folgt

$$U_\alpha = \binom{n}{w}^{-1} \sum_{k_1<k_2<..<k_w}^{n} \widetilde{h}\,(\mathbf{Y}_{k_1},..,\mathbf{Y}_{k_w}) \tag{B.6}$$

Satz B.0.7 (U-Statistiken für Potenzen) *Seien $\alpha \in \mathbb{N}$, $w = \alpha \cdot r$ und $h(\cdot)$, $\widetilde{h}(\cdot)$ symmetrische Kerne mit $E(h(\cdot)) = \theta$ und $\widetilde{h}(\cdot) = h(\cdot) \cdots h(\cdot)$ definiert wie oben.*

$$\begin{aligned} U &= \binom{n}{r}^{-1} \sum_{k_1<k_2<..<k_r}^{n} h\left(\mathbf{Y}_{k_1},..,\mathbf{Y}_{k_r}\right) \\ U_\alpha &= \binom{n}{w}^{-1} \sum_{k_1<k_2<..<k_w}^{n} \widetilde{h}\left(\mathbf{Y}_{k_1},..,\mathbf{Y}_{k_w}\right) \end{aligned}$$

Ist U ein erwartungstreuer und gleichmäßig in d konsistenter Schätzer für θ, dann ist U_α erwartungstreuer und gleichmäßig in d konsistenter Schätzer für θ^α

Beweis: Der Erwartungswert ergibt sich wie folgt

$$\begin{aligned} E(U_\alpha) &= \binom{n}{w}^{-1} \binom{n}{w} E\left(\widetilde{h}\left(\mathbf{Y}_1,..,\mathbf{Y}_w\right)\right) \\ &= E\left(h\left(\mathbf{Y}_1,..,\mathbf{Y}_r\right) \cdots h\left(\mathbf{Y}_{w-r+1},..,\mathbf{Y}_w\right)\right) \\ &= E\left(h\left(\mathbf{Y}_1,..,\mathbf{Y}_r\right)\right) \cdot E\left(h\left(\mathbf{Y}_1,..,\mathbf{Y}_r\right)\right) \cdots E\left(h\left(\mathbf{Y}_1,..,\mathbf{Y}_r\right)\right) \\ &= \theta^\alpha \end{aligned}$$

Für die Varianz von U gilt nach (B.3) bzw (B.4)

$$Var(U) = \sum_{c=1}^{r} \mathcal{O}(n^{-c})\zeta_c$$

wobei

$$\zeta_c = Cov\left(h\left(\mathbf{Y}_1,..\mathbf{Y}_c,\mathbf{Y}_{c+1},..,\mathbf{Y}_r\right) h\left(\mathbf{Y}_1,..\mathbf{Y}_c,\mathbf{Y}_{r+1},..,\mathbf{Y}_{r+c}\right)\right)$$

bezeichnet. Ist U nun ein $\mathcal{L}_2$ konsistenter Schätzer, so sind die ζ_c, $c = 1,..,r$, asymptotisch vernachlässigbar mit Ordnung

$$\zeta_c = \mathbf{o}(n^c)$$

da sonst die Varianz nicht gegen 0 konvergieren würde. Das Landau-Symbol $\mathbf{o}(n)$ bedeutet hier von echt kleinerer Ordnung als $\mathcal{O}(n)$. Dann gilt unter den gleichen Konvergenzbedingungen, welche für U gelten

$$Var(U_\alpha) = \binom{n}{w} \sum_{k_1<..<k_w}^{n} \sum_{l_1<..<l_w}^{n} Cov\left(\widetilde{h}\left(\mathbf{Y}_{k_1},..,\mathbf{Y}_{k_w}\right), \widetilde{h}\left(\mathbf{Y}_{l_1},..,\mathbf{Y}_{l_w}\right)\right)$$

Zunächst lassen sich mit der Permutationsinvarianz von $\widetilde{h}$ bezüglich der w Argumente die Indizes l_j vertauschen. So ist es möglich, diese in eine Sortierung zu bringen, so dass für alle paarweise gleichen Indizes $l_{j'} = k_j$, die $l_{j'}$ an die j-te Position plaziert werden, $j = 1,..,w$. Es kann aufgrund von $l_1 < .. < l_{j'} < .. < l_w$ und $k_1 < .. < k_j < .. < k_w$ jeweils nur maximal eine solche Paarung auftreten.
Sind die $\mathbf{Y}_{l_{j'}}$ an die j-te Position getauscht, so lässt sich das Argument von $\widetilde{h}$ als Blöcke $[1,..,r],[r+1,..,2r],...,[w-r+1,..,w]$ einteilen und sämtliche $h(\mathbf{Y}_{k_{(i-1)\cdot r+1}},..,\mathbf{Y}_{k_{i\cdot r}})$ für den i-ten Block sind unabhängig zu den $h(\cdot)$ aus den anderen Blöcken. Sei c_i die Anzahl von gleichen Indizes im i-ten Block und $c = c_1 + c_2 + .. + c_\alpha$ die Gesamtmenge an gleichen Indizes in den Mengen der k_j und l_j. Dann folgt analog zu B.3

$$Var(U_\alpha) = \binom{n}{w} \sum_{k_1<..<k_w}^{n} \sum_{l_1<..<l_w}^{n} Cov(h\left(\mathbf{Y}_{k_1},..,\mathbf{Y}_{k_r}\right) \cdots h\left(\mathbf{Y}_{k_{w-r+1}},..,\mathbf{Y}_{k_w}\right), \\ h\left(\mathbf{Y}_{l_1},..,\mathbf{Y}_{l_r}\right) \cdots h\left(\mathbf{Y}_{l_{w-r+1}},..,\mathbf{Y}_{l_w}\right))$$

$$= \binom{n}{w}^{-1} \sum_{c=0}^{w} \binom{w}{c} \binom{n-w}{w-c} \zeta_{c_1}\zeta_{c_2}\cdots\zeta_{c_\alpha}$$

$$= \sum_{c=1}^{w} \mathcal{O}(n^{-c})\zeta_{c_1}\zeta_{c_2}\cdots\zeta_{c_\alpha}$$

$$= \sum_{c=1}^{w} \mathcal{O}(n^{-c})\mathbf{o}(n^{c_1})\mathbf{o}(n^{c_2})\cdots\mathbf{o}(n^{c_\alpha})$$

$$= \sum_{c=1}^{w} \mathcal{O}(n^{-c})\mathbf{o}(n^{c})$$

$$= \mathbf{o}(1) \xrightarrow{n\to\infty} 0$$

□

Mit diesem Ergebnis lassen sich auf einfachem Wege neue konsistente Schätzer für ganzzahlige Potenzen α von $Sp(\mathbf{V})$, $Sp(\mathbf{V}^2)$ und $Sp(\mathbf{V}^3)$ konstruieren.

C Momente der A_{kl}

C.1 Resultate aus der Matrizenrechnung

Lemma C.1.1 (Eigenschaften der Spur) *Sei* $\mathbf{V}$ *eine positiv semi-definite* p×p*-Matrix. Dann gilt für die Spuren:*

1. $Sp(\mathbf{V}^2) \leq Sp^2(\mathbf{V})$
2. $Sp(\mathbf{V}^4) \leq Sp^2(\mathbf{V}^2)$
3. $Sp^2(\mathbf{V}^2) \leq Sp(\mathbf{V}) \cdot Sp(\mathbf{V}^3)$
4. $Sp^2(\mathbf{V}^3) \leq Sp(\mathbf{V}^2) \cdot Sp(\mathbf{V}^4)$
5. $Sp(\mathbf{V}^2) \cdot Sp(\mathbf{V}^4) \leq Sp^3(\mathbf{V}^2)$
6. $Sp(\mathbf{V}^6) \leq Sp^3(\mathbf{V}^2)$

Beweis: Da $\mathbf{V}$ positiv semi-definit ist, existiert eine orthogonale Zerlegung, so dass $\mathbf{V} = \mathbf{P}' \cdot diag(\lambda_1,..,\lambda_p) \cdot \mathbf{P}$ mit den zugehörigen Eigenwerten $\lambda_1,..,\lambda_p$ von $\mathbf{V}$, welche alle ≥ 0 sind. Somit ist

$$Sp(\mathbf{V}^\delta) = Sp(diag(\lambda_1,..,\lambda_p) \underbrace{\mathbf{P}\mathbf{P}'}_{=\mathbf{I}_p} diag(\lambda_1,..,\lambda_p),..., \underbrace{\mathbf{P} \cdot \mathbf{P}'}_{=\mathbf{I}_p}) = \sum_{i=1}^{p} \lambda_i^\delta$$

1. $Sp(\mathbf{V}^2) = \sum_{i=1}^{p} \lambda_i^2 \leq \sum_{i=1}^{p} \lambda_i^2 + \sum_{i \neq j} \underbrace{\lambda_i \lambda_j}_{\geq 0} = \Big(\sum_{i=1}^{p} \lambda_i \Big)^2 = Sp^2(\mathbf{V})$

2. Folgt aus 1. mit $\widetilde{\mathbf{V}} := \mathbf{V}^2$ und $\widetilde{\mathbf{V}}$ ebenfalls positiv semi-definit.

3. Beachte zunächst, dass: $(\lambda_i - \lambda_j)^2 \geq 0$ und somit $\lambda_i^2 + \lambda_j^2 \geq 2\lambda_i\lambda_j$
   Dann folgt:

   $$Sp^2(\mathbf{V}^2) = \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i^2 \lambda_j^2 = \sum_{i=1}^{p} \lambda_i^4 + 2 \sum_{j>i} \lambda_i^2 \lambda_j^2 = \sum_{i=1}^{p} \lambda_i^4 + \sum_{j>i} \lambda_i \lambda_j \cdot 2\lambda_i \lambda_j$$

   $$\leq \sum_{i=1}^{p} \lambda_i^4 + \sum_{j>i} \lambda_i \lambda_j \cdot (\lambda_i^2 + \lambda_j^2) = \sum_{i=1}^{p} \lambda_i^4 + \sum_{j>i} \lambda_i \lambda_j^3 + \sum_{j<i} \lambda_i \lambda_j^3 = \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i \lambda_j^3$$

   $$= Sp(\mathbf{V}) \cdot Sp(\mathbf{V}^3)$$

4. Analog zu 3. folgt:

   $$Sp^2(\mathbf{V}^3) = \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i^3 \lambda_j^3 = \sum_{i=1}^{p} \lambda_i^6 + 2 \sum_{j>i} \lambda_i^3 \lambda_j^3 = \sum_{i=1}^{p} \lambda_i^6 + \sum_{j>i} \lambda_i^2 \lambda_j^2 \cdot 2\lambda_i \lambda_j$$

   $$\leq \sum_{i=1}^{p} \lambda_i^6 + \sum_{j>i} \lambda_i^2 \lambda_j^2 \cdot (\lambda_i^2 + \lambda_j^2) = \sum_{i=1}^{p} \lambda_i^6 + \sum_{j>i} \lambda_i^2 \lambda_j^4 + \sum_{j<i} \lambda_i^2 \lambda_j^4 = \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i^2 \lambda_j^4$$

   $$= Sp(\mathbf{V}^2) \cdot Sp(\mathbf{V}^4)$$

5. $Sp(\mathbf{V}^2) \cdot Sp(\mathbf{V}^4) \overset{(2)}{\leq} Sp(\mathbf{V}^2) \cdot Sp^2(\mathbf{V}^2) = Sp^3(\mathbf{V}^2)$

6. Folgt aus 1. mit $\widetilde{\mathbf{V}} := \mathbf{V}^3$ und Anwenden von 4. und 2..

□

Im folgenden werden Darstellungen der Potenzen und der zugehörigen Spuren von $\mathbf{V}$ mittels der Elemente $v_{ij},\ i,j = 1,...,p$ benötigt.

Lemma C.1.2 *Sei* $\mathbf{V} = (v_{ij})_{i=1,..,m;\ j=1,..,m}$ *eine symmetrische* $m \times m$*-Matrix. Dann gilt:*

1. $\mathbf{V}^2 = \left(\sum_{h=1}^{m} v_{ih} v_{hj} \right)_{i=1,..,m;\ j=1,..,m}$

2. $\mathbf{V}^3 = \left(\sum_{h_1=1}^{m} \sum_{h_2=1}^{m} v_{ih_1} v_{h_1h_2} v_{h_2j} \right)_{i=1,..,m;\ j=1,..,m}$

3. $\mathbf{V}^4 = \left(\sum_{h_1=1}^{m} \sum_{h_2=1}^{m} \sum_{h_3=1}^{m} v_{ih_2} v_{h_2h_1} v_{h_1h_3} v_{h_3j} \right)_{i=1,..,m;\ j=1,..,m}$

4. $\mathbf{V}^6 = \left(\sum_{h_1=1}^{m} \sum_{h_2=1}^{m} \sum_{h_3=1}^{m} \sum_{h_4=1}^{m} \sum_{h_5=1}^{m} v_{ih_2} v_{h_2h_3} v_{h_3h_1} v_{h_1h_4} v_{h_4h_5} v_{h_5j} \right)_{i=1,..,m;\ j=1,..,m}$

Beweis:
Zunächst gilt für die Multiplikation von zwei Matrizen $\mathbf{A} = (a_{ij})_{i=1,..,a;\ j=1,..,m}$ und $\mathbf{B} = (b_{ij})_{i=1,..,m;\ j=1,..,b}$

$$\mathbf{A} \cdot \mathbf{B} = \left(\sum_{h=1}^{m} a_{ik} b_{kj} \right)_{i=1,..,a;\ j=1,..,b}$$

Dann folgt für symmetrische Matrizen $\mathbf{V}^r$, deren Elemente der Einfachheit halber mit $(\mathbf{V}^r)_{ij}$ bezeichnet werden, $r = 2,3,4,6$:

1. $(\mathbf{V}^2)_{ij} = \sum_{h=1}^{m} v_{ih} v_{hj}$

2. $(\mathbf{V}^3)_{ij} = \sum_{h_1=1}^{m} v_{ih_1} (\mathbf{V}^2)_{h_1j}$

$$= \sum_{h_1=1}^{m} v_{ih} v_{ih_1} \sum_{h_2=1}^{m} v_{h_1h_2} v_{h_2j}$$

$$\sum_{h_1=1}^{m} \sum_{h_2=1}^{m} v_{ih_1} v_{h_1h_2} v_{h_2j}$$

3. $(\mathbf{V}^4)_{ij} = \sum_{h_1=1}^{m} (\mathbf{V}^2)_{ih_1} (\mathbf{V}^2)_{h_1j}$

$$= \sum_{h_1=1}^{m} \Big(\sum_{h_2=1}^{m} v_{ih_2} v_{h_2h_1} \Big) \Big(\sum_{h_3=1}^{m} v_{h_1h_3} v_{h_3j} \Big)$$

$$= \sum_{h_1=1}^{m} \sum_{h_2=1}^{m} \sum_{h_3=1}^{m} v_{ih_2} v_{h_2h_1} v_{h_1h_3} v_{h_3j}$$

4. $(\mathbf{V}^6)_{ij} = \sum_{h_1=1}^{m} (\mathbf{V}^3)_{ih_1}(\mathbf{V}^3)_{h_1 j}$

$$= \sum_{h_1=1}^{m} \Big(\sum_{h_2=1}^{m} \sum_{h_3=1}^{m} v_{ih_2} v_{h_2h_3} v_{h_3h_1} \Big) \Big(\sum_{h_4=1}^{m} \sum_{h_5=1}^{m} v_{h_1h_4} v_{h_4h_5} v_{h_5j} \Big)$$

$$= \sum_{h_1=1}^{m} \sum_{h_2=1}^{m} \sum_{h_3=1}^{m} \sum_{h_4=1}^{m} \sum_{h_5=1}^{m} v_{ih_2} v_{h_2h_3} v_{h_3h_1} v_{h_1h_4} v_{h_4h_5} v_{h_5j}$$

□

Lemma C.1.3 *Sei* $\mathbf{V} = (v_{ij})_{ij}$ *eine symmetrische* $m \times m$*-Matrix. Dann gilt:*

1. $Sp(\mathbf{V}^2) = \sum_{i=1}^{m} \sum_{j=1}^{m} v_{ij}^2$
2. $Sp(\mathbf{V}^3) = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{h=1}^{m} v_{ij} v_{jh} v_{hi}$
3. $Sp(\mathbf{V}^4) = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{i'=1}^{m} \sum_{j'=1}^{m} v_{ij} v_{ji'} v_{i'j'} v_{j'i}$
4. $Sp(\mathbf{V}^6) = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{h=1}^{m} \sum_{i'=1}^{m} \sum_{j'=1}^{m} \sum_{h'=1}^{m} v_{ij} v_{jh} v_{hi'} v_{i'j'} v_{j'h'} v_{h'i}$
5. $Sp^2(\mathbf{V}^2) = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{i'=1}^{m} \sum_{j'=1}^{m} v_{ij}^2 v_{i'j'}^2$
6. $Sp^2(\mathbf{V}^3) = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{h=1}^{m} \sum_{i'=1}^{m} \sum_{j'=1}^{m} \sum_{h'=1}^{m} v_{ij} v_{jh} v_{hi} \;\; v_{i'j'} v_{j'h'} v_{h'i'}$
7. $Sp^3(\mathbf{V}^2) = \sum_{i=1}^{m} \sum_{i'=1}^{m} \sum_{j=1}^{m} \sum_{j'=1}^{m} \sum_{h=1}^{m} \sum_{h'=1}^{m} v_{ii'}^2 v_{jj'}^2 v_{hh'}^2$
8. $Sp(\mathbf{V}^2)Sp(\mathbf{V}^4) = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{i'=1}^{m} \sum_{h=1}^{m} \sum_{j'=1}^{m} \sum_{h'=1}^{m} v_{ij}^2 \;\; v_{i'h} v_{hj'} v_{j'h'} v_{h'i'}$

Beweis: Folgt aus $Sp(\mathbf{V}^r) = \sum_{i=1}^{m} (\mathbf{V}^r)_{ii}$, Symmetrie von $\mathbf{V}^r$, $r = 1,2,3,4,6$ und Anwenden von Lemma C.1.2.

□

Für die erleichterte Identifizierbarkeit der Darstellungen sei bemerkt, dass die Sequenzen der Indizes charakteristisch für einen Ausdruck sind. So stehen 2 Sequenzen der Länge 3 (ij jh hi) beispielsweise für $Sp^2(\mathbf{V}^3)$

C.2 Höhere Momente der A_{kl}

Lemma C.2.1 *Die Zufallsvektoren $\mathbf{Y}_k$, $k = 1,..,n$ seien unabhängig verteilt wie in Definition 3.2.2 (Bai-Saranadasa Modell):*

$$\mathbf{Y}_k = \mathbf{\Gamma} \cdot \mathbf{Z}_k + \mathbf{0}_d \quad \textit{für } k = 1, \ldots, n$$

wobei und $\mathbf{Z}_k = (Z_{1k}, .., Z_{mk})'$ unabhängig identisch verteilt mit $E(\mathbf{Z}_k) = \mathbf{0}_m$, $Cov(\mathbf{Z}_k) = \mathbf{I}_m$, $E(Z_{ik}^4) =: \mu_{4i} \ll \infty$ und sämtlichen Komponenten $\{Z_{ik}\}_{k=1,\ldots,m}^{i=1,\ldots,d}$ unabhängig. Die Elemente der dualen Kovarianzmatrix $\mathbf{V} = \mathbf{\Gamma}'\mathbf{\Gamma}$ seien mit v_{ij} , $i,j = 1,..,m$ bezeichnet. Die $A_{kl} := \mathbf{Y}_k'\mathbf{Y}_l$ sind die zugehörigen Bilinearformen für $k \neq l$ und A_{kk} die analogen Quadratformen. Dann gilt für deren Momente mit sämtlichen Indexkombinationen $k \neq l \neq r \neq k' \neq l' \neq r'$:

1. $E(A_{kk}) = Sp(\mathbf{V})$
2. $E(A_{kk}^2) = 2Sp(\mathbf{V}^2) + Sp^2(\mathbf{V}) + \sum_{i=1}^m (\mu_{4i} - 3) \cdot v_{ii}^2$
3. $E(A_{kl}) = 0$
4. $E(A_{kl}^2) = Sp(\mathbf{V}^2)$
5. $E(A_{kl}^4) = \mathcal{O}(Sp^2(\mathbf{V}^2))$
6. $E(A_{kl}^2 A_{kr}^2) = \mathcal{O}(Sp^2(\mathbf{V}^2))$
7. $E(A_{kl}^2 A_{k'l'}^2) = Sp^2(\mathbf{V}^2))$
8. $E(A_{kl}^2 A_{lr}^2 A_{rk}^2) = \mathcal{O}(Sp^3(\mathbf{V}^2))$
9. $E(A_{kl}A_{lr}A_{rk}A_{k'l}A_{lr}A_{rk'}) = \mathcal{O}(Sp^3(\mathbf{V}^2))$
10. $E(A_{kl}A_{lr}A_{rk}A_{kl'}A_{l'r'}A_{r'k}) = \mathcal{O}(Sp^2(\mathbf{V}^3))$
11. $E(A_{kl}A_{lr}A_{rk}A_{k'l'}A_{l'r'}A_{r'k'}) = Sp^2(\mathbf{V}^3)$

Beweis:

$$
\begin{aligned}
1.E(A_{kk}) &= E(Sp(A_{kk})) = E(Sp(\mathbf{Y}_k'\mathbf{Y}_k) = Sp(E(\mathbf{Y}_k\mathbf{Y}_k')) = Sp(\mathbf{\Sigma}) = Sp(\mathbf{V}) \\
2.E(A_{kk}^2) &= E(\textstyle\sum_{i=1}^m \sum_{j=1}^m v_{ij}\cdot Z_i \cdot Z_j)^2 \\
&= E(\textstyle\sum_{i=1}^m \sum_{j=1}^m \sum_{i'=1}^m \sum_{j'=1}^m v_{ij}\cdot v_{i'j'}\cdot Z_iZ_jZ_{i'}Z_{j'}) \\
&= \textstyle\sum_{i=1}^m v_{ii}^2 \cdot \underbrace{E(Z_i^4)}_{=\mu_{4i}} + \sum_{i\neq j} v_{ii}\cdot v_{jj}\cdot \underbrace{E(Z_i^2)\cdot E(Z_j^2)}_{=1} \\
&\quad + \textstyle\sum_{i\neq j}(v_{ij}\cdot v_{ij} + v_{ij}\cdot v_{ji})\cdot \underbrace{E(Z_i^2)\cdot E(Z_j^2)}_{=1} \\
&= \underbrace{\sum_{i=1}^m\sum_{j=1}^m v_{ii}\cdot v_{jj}}_{=Sp^2(\mathbf{V})} + 2\underbrace{\sum_{i=1}^m\sum_{j=1}^m v_{ij}^2}_{=Sp(\mathbf{V}^2)} + \underbrace{\sum_{i=1}^m(\mu_{4i}-3)v_{ii}^2}_{=\mathcal{O}(Sp(\mathbf{V}^2))} \\
3.E(A_{kl}) &= E(\mathbf{Y}_k')\cdot E(\mathbf{Y}_l) = 0 \\
4.E(A_{kl}^2) &= E(Sp(\mathbf{Y}_k'\mathbf{Y}_l\cdot\mathbf{Y}_l'\mathbf{Y}_k) = E(Sp(\mathbf{Y}_k\mathbf{Y}_k'\mathbf{Y}_l\mathbf{Y}_l') \\
&= Sp(E(\mathbf{Y}_k\mathbf{Y}_k')\cdot E(\mathbf{Y}_l\mathbf{Y}_l')) = Sp(\mathbf{\Sigma}\cdot\mathbf{\Sigma}) = Sp(\mathbf{V}^2) \\
5.E(A_{kl}^4) &= E(\mathbf{Z}_k'\mathbf{V}\mathbf{Z}_l)^4 = E(\sum_{i=1}^m\sum_{j=1}^m v_{ij}Z_{ik}Z_{jl})^4 \\
&= E(\sum_{i_1=1}^m\sum_{i_2=1}^m\sum_{i_3=1}^m\sum_{i_4=1}^m\sum_{j_1=1}^m\sum_{j_2=1}^m\sum_{j_3=1}^m\sum_{j_4=1}^m v_{i_1j_1}v_{i_2j_2}v_{i_3j_3}v_{i_4j_4}\cdot Z_{i_1k}Z_{i_2k}Z_{i_3k}Z_{i_4k}Z_{j_1l}Z_{j_2l}Z_{j_3l}Z_{j_4l}) \\
&= \sum_{i=1}^m\sum_{j=1}^m v_{ij}^4 \underbrace{\cdot E(Z_i^4)\cdot E(Z_j^4)}_{\leq\mu_4\cdot\mu_4} \\
&\quad + \sum_{i=1}^m 3\sum_{j\neq j'} v_{ij}^2v_{ij'}^2 \underbrace{\cdot E(Z_i^4)E(Z_j^2)E(Z_{j'}^2)}_{\leq\mu_4\cdot 1} + 3\sum_{i\neq i'}\sum_{j=1}^m v_{ij}^2v_{i'j}^2 \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^4)}_{\leq 1\cdot\mu_4} \\
&\quad + 3\sum_{i\neq i'}\sum_{j\neq j'} v_{ij}^2v_{i'j'}^2 \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^2)E(Z_{j'}^2)}_{=1} \\
&\quad + 6\sum_{i\neq i'}\sum_{j\neq j'} v_{ij}v_{ij'}v_{i'j}v_{i'j'} \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^2)E(Z_{j'}^2)}_{=1}
\end{aligned}
$$

$$\leq (max(3,\mu_4)^2 - 6) \left(\sum_{i=1}^{m} \sum_{j=1}^{m} v_{ij}^4 + \sum_{i=1}^{m} \sum_{j \neq j'} v_{ij}^2 v_{ij'}^2 + \sum_{i \neq i'} \sum_{j=1}^{m} v_{ij}^2 v_{i'j}^2 + \sum_{i \neq i'} \sum_{j \neq j'} v_{ij}^2 v_{i'j'}^2 \right)$$

$$+6 \left(\sum_{i=1}^{m} \sum_{j=1}^{m} v_{ij}^4 + \sum_{i=1}^{m} \sum_{j \neq j'} v_{ij}^2 v_{ij'}^2 + \sum_{i \neq i'} \sum_{j=1}^{m} v_{ij}^2 v_{i'j}^2 + \sum_{i \neq i'} \sum_{j \neq j'} v_{ij} v_{ij'} v_{i'j} v_{i'j'} \right)$$

$$= (max(3,\mu_4)^2 - 6) \cdot \left(\sum_{i=1}^{m} \sum_{j=1}^{m} v_{ij}^2 \right)^2 + 6 \cdot \left(\sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{i'=1}^{m} \sum_{j'=1}^{m} v_{ij} v_{ij'} v_{i'j} v_{i'j'} \right)$$

$$= (max(3,\mu_4)^2 - 6) \cdot Sp^2(\mathbf{V}^2) + 6 \cdot Sp(\mathbf{V}^4) = O(Sp^2(\mathbf{V}^2))$$

6. $E(A_{kl}^2 A_{kr}^2) = E((A_{kl} A_{lk} A_{kr} A_{kr})$

$$= E(\mathbf{Z}_k' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_l \cdot \mathbf{Z}_l' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_k \cdot \mathbf{Z}_k' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_r \cdot \mathbf{Z}_r' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_k)$$

$$= E(\mathbf{Z}_k' \mathbf{\Gamma}' \mathbf{\Gamma} \underbrace{\mathbf{I}_m}_{=E(\mathbf{Z}_l \cdot \mathbf{Z}_l')} \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_k \cdot \mathbf{Z}_k' \mathbf{\Gamma}' \mathbf{\Gamma} \underbrace{\mathbf{I}_m}_{=E(\mathbf{Z}_r \cdot \mathbf{Z}_r')} \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_k)$$

$$= E(\mathbf{Z}_k' \mathbf{V}' \mathbf{V} \mathbf{Z}_k \cdot \mathbf{Z}_k' \mathbf{V}' \mathbf{V} \mathbf{Z}_k) = E(\mathbf{Z}_k' \mathbf{V}' \mathbf{V} \mathbf{Z}_k)^2$$

$$= 2Sp((\mathbf{V}'\mathbf{V})^2) + Sp^2(\mathbf{V}'\mathbf{V}) + (\mu_4 - 3) \textstyle\sum_{i=1}^{d} (\mathbf{V}'\mathbf{V})_{ii}^2 \quad \text{nach (2.)}$$

$$= 2Sp((\mathbf{V}^4) + Sp^2(\mathbf{V}^2) + (\mu_4 - 3) \textstyle\sum_{i=1}^{d} (\mathbf{V}^2)_{ii}^2$$

$$= \mathcal{O}(Sp^2(\mathbf{V}^2))$$

7. $E(A_{kl}^2 A_{k'l'}^2)$

$$= E(\mathbf{Z}_k' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_l \cdot \mathbf{Z}_l' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_k \cdot \mathbf{Z}_{k'}' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_{l'} \cdot \mathbf{Z}_{l'}' \mathbf{\Gamma}' \mathbf{\Gamma} \, \mathbf{Z}_{k'})$$

$$= E(\mathbf{Z}_k' \mathbf{V}' \mathbf{V} \, \mathbf{Z}_k) \cdot E(\mathbf{Z}_{k'}' \mathbf{V}' \mathbf{V} \, \mathbf{Z}_{k'})$$

$$= Sp^2(\mathbf{V}^2) \quad \text{nach (1.)}$$

8. Der Übersichtlichkeit halber sei $(x_1,...,x_d)' := \mathbf{Y}_k$, $(y_1,...,y_d)' := \mathbf{Y}_l$ und $(z_1,...,z_d)' := \mathbf{Y}_r$. Dann gilt:

$$E(A_{kl}^2 A_{lr}^2 A_{rk}^2) = E((\sum_{i=1}^{m}\sum_{j=1}^{m} v_{ij}x_i y_j)^2 \cdot (\sum_{j=1}^{m}\sum_{h_1=1}^{m} v_{ij}y_i z_j)^2 (\sum_{h=1}^{m}\sum_{i=1}^{m} v_{ij}z_h x_i)^2)$$

$$= \sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\sum_{i_3=1}^{m}\sum_{i_4=1}^{m}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\sum_{j_3=1}^{m}\sum_{j_4=1}^{m}\sum_{h_1=1}^{m}\sum_{h_2=1}^{m}\sum_{h_3=1}^{m}\sum_{h_4=1}^{m} v_{i_1j_1}v_{i_1j_1}v_{j_3h_1}v_{j_4h_2}v_{h_3i_3}v_{h_4i_4} \cdot x_{i_1}x_{i_2}x_{i_3}x_{i_4}y_{j_1}y_{j_2}y_{j_3}y_{j_4}z_{h_1}z_{h_2}z_{h_3}z_{h_4}$$

$$\leq \quad c_6 \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}v_{jh}v_{hi'}v_{i'j'}v_{j'h'}v_{h'i}$$

$$+c_4 \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}^2 \;\; v_{i'h}v_{hj'}v_{j'h'}v_{h'i'}$$

$$+c_3 \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}v_{jh}v_{hi} \;\; v_{i'j'}v_{j'h'}v_{h'i'}$$

$$+c_2 \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}^2 \;\; v_{j'h}^2 \;\; v_{h'i'}^2$$

wobei die c_i Konstanten sind, die sich aus der kombinatorischen Vielfachheit der Terme und den vierten Momenten der Z_i ergeben ($c_i = \mathcal{O}(\mu_4^3)$).

$$= c_6 Sp(\mathbf{V}^6) + c_4 Sp(\mathbf{V}^4)Sp(\mathbf{V}^2) + c_3 Sp^2(\mathbf{V}^3) + c_2 Sp^3(\mathbf{V}^2)$$

$$= \mathcal{O}(Sp^3(\mathbf{V}^2))$$

10. $E(A_{kl}A_{lr}A_{rk}A_{kl'}A_{l'r'}A_{r'k}) = E(\mathbf{Z}_k'\mathbf{V}^3\mathbf{Z}_k)^2$

$$= 2Sp(\mathbf{V}^6) + Sp^2(\mathbf{V}^3) + (\mu_4 - 3)\sum_{i=1}^{d}(\mathbf{V}^3)_{ij}^2 \text{ nach 2.}$$

$$= \mathcal{O}(Sp^2(\mathbf{V}^3))$$

11. $E(A_{kl}A_{lr}A_{rk}A_{k'l'}A_{l'r'}A_{r'k'}) = E(A_{kl}A_{lr}A_{rk}) \cdot E(A_{k'l'}A_{l'r'}A_{r'k'})$

$$E^2(Sp(\mathbf{Y}_k\mathbf{Y}_k'\,\mathbf{Y}_l\mathbf{Y}_l'\,\mathbf{Y}_r\mathbf{Y}_r')) = Sp^2(\mathbf{V}^3)$$

9. $E(A_{kl}A_{lr}A_{rk}A_{k'l}A_{lr}A_{rk'}) = E(A_{lr}A_{lk}A_{kr}A_{lr}A_{lk'}A_{k'r})$

$$= E(\mathbf{Z}_l'\mathbf{V}\mathbf{Z}_r\mathbf{Z}_l'\mathbf{V}^2\mathbf{Z}_r)^2 = E\left((\sum_{i=1}^m\sum_{j=1}^m v_{ij}Z_{il}Z_{jr})^2 \cdot (\sum_{i=1}^m\sum_{j=1}^m (\mathbf{V}^2)_{ij}Z_{il}Z_{jr})^2\right)$$

$$= E(\sum_{i_1=1}^m\sum_{i_2=1}^m\sum_{i_3=1}^m\sum_{i_4=1}^m\sum_{j_1=1}^m\sum_{j_2=1}^m\sum_{j_3=1}^m\sum_{j_4=1}^m v_{i_1j_1}(\mathbf{V}^2)_{i_2j_2}(\mathbf{V}^2)_{i_3j_3}v_{i_4j_4} \cdot Z_{i_1l}Z_{i_2l}Z_{i_3l}Z_{i_4l} \cdot Z_{j_1r}Z_{j_2r}Z_{j_3r}Z_{j_4r})$$

analog zum Beweis von 5. gilt unter Berücksichtigung der kombinatorischen Vielfachheit:

$$= \sum_{i=1}^m\sum_{j=1}^m v_{ij}^2(\mathbf{V}^2)_{ij}^2 \underbrace{\cdot E(Z_i^4)\cdot E(Z_j^4)}_{\leq\mu_4\cdot\mu_4}$$

$$+2\sum_{i=1}^m\sum_{j\neq j'} v_{ij}(\mathbf{V}^2)_{ij}v_{ij'}(\mathbf{V}^2)_{ij'} \underbrace{\cdot E(Z_i^4)E(Z_j^2)E(Z_{j'}^2)}_{\leq\mu_4\cdot 1}$$

$$+\sum_{i=1}^m\sum_{j\neq j'} v_{ij}^2(\mathbf{V}^2)_{ij'}^2 \underbrace{\cdot E(Z_i^4)E(Z_j^2)E(Z_{j'}^2)}_{\leq\mu_4\cdot 1}$$

$$+2\sum_{i\neq i'}\sum_{j=1}^m v_{ij}(\mathbf{V}^2)_{ij}v_{i'j}(\mathbf{V}^2)_{i'j} \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^4)}_{\leq 1\cdot\mu_4}$$

$$+\sum_{i\neq i'}\sum_{j=1}^m v_{ij}^2(\mathbf{V}^2)_{i'j}^2 \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^4)}_{\leq 1\cdot\mu_4}$$

$$+2\sum_{i\neq i'}\sum_{j\neq j'} v_{ij}(\mathbf{V}^2)_{ij}v_{i'j'}(\mathbf{V}^2)_{i'j'} \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^2)E(Z_{j'}^2)}_{=1}$$

$$+\sum_{i\neq i'}\sum_{j\neq j'} v_{ij}^2(\mathbf{V}^2)_{i'j'}^2 \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^2)E(Z_{j'}^2)}_{=1}$$

$$+6\sum_{i\neq i'}\sum_{j\neq j'} v_{ij}v_{ij'}(\mathbf{V}^2)_{i'j}(\mathbf{V}^2)_{i'j'} \underbrace{\cdot E(Z_i^2)E(Z_{i'}^2)E(Z_j^2)E(Z_{j'}^2)}_{=1}$$

Mit $(\mathbf{V}^2)_{ij} = \sum_{h=1}^{m} v_{ih}v_{hj}$ folgt:

$$
\begin{aligned}
\leq \quad & \mu_4^2 \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}^2 v_{ih}v_{hj}v_{jh'}v_{h'i} \\
& +2\mu_4 \sum_{i=1}^{m}\sum_{j\neq j'} v_{ij}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}v_{jh}v_{hi}v_{ij'}v_{j'h'}v_{h'i} \\
& +\mu_4 \sum_{i=1}^{m}\sum_{j\neq j'}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}^2 v_{ih}v_{hj'}v_{j'h'}v_{h'i} \\
& +2\mu_4 \sum_{i\neq i'}\sum_{j=1}^{m}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}v_{jh}v_{hi'}v_{i'j}v_{jh'}v_{h'i} \\
& +\mu_4 \sum_{i\neq i'}\sum_{j=1}^{m}\sum_{h=1}^{m}\sum_{h'=1}^{m} v_{ij}^2 v_{i'h}v_{hj}v_{jh'}v_{h'i'} \\
& +2\sum_{i\neq i'}\sum_{j\neq j'}\sum_{h=1}^{m}\sum_{h'=1}^{m} \underbrace{v_{ij}v_{jh}v_{hi}\;\; v_{i'j'}v_{h'h'}v_{h'i'}}_{\to Sp^2(\mathbf{V}^3)} \\
& +\sum_{i\neq i'}\sum_{j\neq j'}\sum_{h=1}^{m}\sum_{h'=1}^{m} \underbrace{v_{ij}^2 v_{i'h}v_{hj'}v_{j'h'}v_{h'i'}}_{\to Sp(\mathbf{V}^2)\cdot Sp(\mathbf{V}^4)} \\
& +6\sum_{i\neq i'}\sum_{j\neq j'}\sum_{h=1}^{m}\sum_{h'=1}^{m} \underbrace{v_{ij}v_{jh}v_{hi'}v_{i'h'}v_{h'j'}v_{j'i}}_{\to Sp(\mathbf{V}^6)}
\end{aligned}
$$

Zusammenfassen zu Sechsfach-Summen mit sämtlichen Indexkombinationen ergibt

$$
\begin{aligned}
& \leq (max(3,\mu_4)^2 - 8)\cdot Sp(\mathbf{V}^2)Sp(\mathbf{V}^4) + (max(3,\mu_4)^2 - 7)\cdot Sp^2(\mathbf{V}^3) \\
& \quad +(max(3,\mu_4)^2 - 3)\cdot Sp(\mathbf{V}^6) \\
& = \mathcal{O}(Sp^3(\mathbf{V}^2))
\end{aligned}
$$

□

D SAS-Makro HD-Fi

Für die Auswertung mit SAS wurde folgendes Makro geschrieben, welches die Ergebnisse der Arbeit implementiert. Der Quellcode beruht auf den Makros HD-F1 von Carola Werner und Hans-Joachim Helms [22] [13].
Der Aufruf des Makros wird in Kapitel 12 beschrieben.

```
/*****************************************************************/
/*** Makro              NN_HD_Fi                                 ***/
/***                                                           ***/
/*** Erweitert von David Ellenberger, August 2011               ***/
/***                                                           ***/
/*****************************************************************/
/***    Fuer bis zu 4 Faktoren plus Messwiederholungen anwendbar ***/
/***    Implementiert sind die Statistik von Chen-Qin mit neuem  ***/
/***    Varianzschaetzer B_2 und Approximationen nach Box und    ***/
/***    nach Pearson und die klassische ANOVATyp-Statistik zum   ***/
/***    Vergleich.                                               ***/
/*****************************************************************/

%macro nn_hd_fi(data     =,
                var      =,
                TIME1    =,
                TIME2    =none78xz45fa,       /* Defaultwert fuer die interne Anwendung falls    */
                TIME3    =none78xz45fa,       /* nicht alle 4 Faktoren vorhanden sind (i<4)      */
                TIME4    =none78xz45fa,       /* Der Name none78xz45fa soll Verwechselungsgefahr */
                SUBJECT = );                  /* mit realen Faktoren ausschliessen und dient als */
                                              /* Dummyvariable.                                  */
data execute78xz45fa;                         /* Der Datensatz execute78xz45fa wird temporaer    */
set &data;                                    /* verwendet um die Originaldaten nicht zu ueber-  */
none78xz45fa = 1;                             /* schreiben und wird am Ende wieder geloescht     */
run;

proc sort data=execute78xz45fa;
by &subject &time1 &time2 &time3 &time4;
run;

proc iml worksize=120;
RESET LINESIZE = 80 ;
%LET faktora = &TIME1;
%LET faktorb = &TIME2;
%LET faktorc = &TIME3;
%LET faktord = &TIME4;
%LET faktorp = &SUBJECT;

/* Die Auswertung der (a*b*c*d x n)-Matrix mit klassischer ANOVA und den neuen
Approximationen fuer die Chen-Qin-Statistik nach Box und Pearson */

START box_neu(n,z,C,f_max,Rn,f_box,f_pearson,p_box,p_pearson,p_class);

    if f_max < 1 then f_max= 1;  /* Abfrage ist fuer den trivialen Fall (T * X = 0) noetig */

    T= C' * GINV(C * C') * C;
    y = T * z;
    correct=1; taylorf =1;

    m0       = y' * y;
    m        = m0 # (j(n,n,1) - i(n));
    Tn_new       = sum(m)/sqrt((n-1)*n );

    b2       = sum(m#m)/(n*(n-1)) ;
    b0 = trace(m0) / n ;
    bk       = vecdiag(m0);

    SpVhat = b0 - Tn_new/sqrt((n-1)*n);
    qn = Tn_new*sqrt(n*(n-1))/n + b0 ;
```

```
    b1      = sum((bk * bk')#(j(n,n,1) - i(n)))/(n*(n-1));

    b3 =0;
    do i11 = 1 to n;
    do i12 = i11+1 to n;
    do i13 = i12+1 to n;
    b3 = b3 + m[i11,i12]*m[i12,i13]*m[i11,i13];
    end; end; end;
    anz_counter3 = n*(n-1)*(n-2)/(3*2);
    b3 = b3 /anz_counter3;

    if b2 > 10**(-10) then do;
        Rn = Tn_new/sqrt(2*b2);
        f_box = taylorf * b1/b2;
        if f_box  < 1 then f_box=1; else if f_box  > f_max then f_box = f_max;
    end;
    else do; f_box = f_max; Rn = -sqrt(f_max/2); end;

    if b3 > 10**(-10) then do;
        sqrt_f_pear = b2**(3/2)/b3;
        if sqrt_f_pear  < 1 then sqrt_f_pear=1;
        else if sqrt_f_pear  > sqrt(f_max) then sqrt_f_pear = sqrt(f_max);
    end;
    else sqrt_f_pear = sqrt(f_max);


    pn = i(n) - j(n,n,1)/n;
    ypn     = y * pn;
    bkm     = ypn * ypn';
    scm     = bkm / (n-1);
    trs     = trace(scm);
    trs2    = trace(scm*scm);

    if trs > 10**(-10) then atsp  = qn / trs;
    else atsp =1;

    if trs2 > 10**(-10) then do;
        fp      = trs*trs/trs2;
        if fp  < 1 then fp=1; else if fp  > f_max then fp = f_max;
    end;
    else fp = f_max;

f_pearson  = sqrt_f_pear**2;

xdat = Rn * sqrt(2*f_box) + f_box;
if xdat < 0 then xdat=0;
xdat2 = Rn * sqrt(2*f_pearson) + f_pearson;
if xdat2 < 0 then xdat2=0;

p_class = 1 - probchi(atsp*fp,fp);
p_box = 1 - probchi(xdat, f_box);
p_pearson = 1 - probchi(xdat2,f_pearson);

FINISH;
reset nolog;
/*********************************************************************/

/*****Daten einlesen*************************************************/
USE execute78xz45fa;
READ ALL VAR{&var}     INTO werte;
READ ALL VAR{&subject} INTO pat_;
READ ALL VAR{&time1}   INTO t1_;
READ ALL VAR{&time2}   INTO t2_;
READ ALL VAR{&time3}   INTO t3_;
READ ALL VAR{&time4}   INTO t4_;
CLOSE execute78xz45fa;
/*********************************************************************/

lev_p = unique(pat_);                   /* Die Stufen des Faktors P ***/
lev_a = unique(t1_);
lev_b = unique(t2_);
lev_c = unique(t3_);
lev_d = unique(t4_);
a     = ncol(lev_a);                    /* Anzahl der Stufen von A  ***/
b     = ncol(lev_b);
c     = ncol(lev_c);
c     = ncol(lev_c);
d     = ncol(lev_d);
```

```
n      = ncol(lev_p);

NNreal = nrow(werte);                    /* Anzahl aller Messwerte  ***/
e = NNreal / (a*b*c*n);                  /* Anzahl der Messwiederholungen (trivialer Faktoren) ***/

ja       = j(a,a,1)/a;
jb       = j(b,b,1)/b;
jc       = j(c,c,1)/c;
jd       = j(d,d,1)/d;
je       = j(e,e,1)/e;
pa       = i(a) - ja;
pb       = i(b) - jb;
pc       = i(c) - jc;
pd       = i(d) - jd;

/*****     hypothesen-matrizen       *****/
ma       = pa @ jb @ jc @ jd @ je;
mb       = ja @ pb @ jc @ jd @ je;
mab      = pa @ pb @ jc @ jd @ je;

if ("&time3" = "none78xz45fa")=0 then do;     /* wird nur ausgewertet bei mindestens 3 Faktoren */
mc       = ja @ jb @ pc @ jd @ je;
mac      = pa @ jb @ pc @ jd @ je;
mbc      = ja @ pb @ pc @ jd @ je;
mabc     = pa @ pb @ pc @ jd @ je;
end;

if ("&time4" = "none78xz45fa")=0 then do;     /* wird nur ausgewertet bei 4 Faktoren */
md       = ja @ jb @ jc @ pd @ je;
mad      = pa @ jb @ jc @ pd @ je;
mbd      = ja @ pb @ jc @ pd @ je;
mcd      = ja @ jb @ pc @ pd @ je;
mabd     = pa @ pb @ jc @ pd @ je;
macd     = pa @ jb @ pc @ pd @ je;
mbcd     = ja @ pb @ pc @ pd @ je;
mabcd    = pa @ pb @ pc @ pd @ je;
end;


/*************************************/
****einlesen des vektors der messwerte in eine n x a*b*c matrix******;
wert_matrix = (shape (werte,n,a*b*c*e))';

/*********************************************************************
   Die BOX-Approximation: Ergebnisse in 'box'
*********************************************************************/
   RUN box_neu(n,wert_matrix,ma,a-1,Rn,f_box,f_pearson,p_box,p_pearson,p_class);                    /*Effekt A*/
    box= (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mb,b-1,Rn,f_box,f_pearson,p_box,p_pearson,p_class);                    /*Effekt B*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mab,((a-1)*(b-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class);         /*Effekt AB*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);

if ("&time3" = "none78xz45fa")=0 then do;
   RUN box_neu(n,wert_matrix,mc,c-1,Rn,f_box,f_pearson,p_box,p_pearson,p_class);                    /*Effekt C*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mac,((a-1)*(c-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class);         /*Effekt AC*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mbc,((b-1)*(c-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class);         /*Effekt BC*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mabc,((a-1)*(b-1)*(c-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class); /*Effekt ABC*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
end;

if ("&time4" = "none78xz45fa")=0 then do;
   RUN box_neu(n,wert_matrix,md,d-1,Rn,f_box,f_pearson,p_box,p_pearson,p_class);                    /*Effekt D*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mad,((a-1)*(d-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class);         /*Effekt AD*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mbd,((b-1)*(d-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class);         /*Effekt BD*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mcd,((c-1)*(d-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class);         /*Effekt CD*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mabd,((a-1)*(b-1)*(d-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class); /*Effekt ABD*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,macd,((a-1)*(c-1)*(d-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class); /*Effekt ACD*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mbcd,((b-1)*(c-1)*(d-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class); /*Effekt BCD*/
    box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
   RUN box_neu(n,wert_matrix,mabcd,((a-1)*(b-1)*(c-1)*(d-1)),Rn,f_box,f_pearson,p_box,p_pearson,p_class);/*ABCD*/
```

```
   box= box // (Rn || f_box || f_pearson || p_class || p_box ||  p_pearson);
end;


/*********************************************************************/
/***        FUNKTION                                                 ***/
/***        Ausdruck der Class Level Information (CLI)               ***/
/*********************************************************************/
start O_CLI (leva, levb, levc, levd, levp, a,b,c,d,n,NNreal);
reset center;


class  = "Time1" || "&faktora";
levels = a ;
values = leva;
if ("&time2" = "none78xz45fa")=0 then do;
class  = class //  ("Time2" || "&faktorb");
levels = levels // b;
values = values // levb;
end;
if ("&time3" = "none78xz45fa")=0 then do;
class  = class // ( "Time3" || "&faktorc");
levels = levels //  c ;
values = values // levc;
end;
if ("&time4" = "none78xz45fa")=0 then do;
class  = class // ( "Time4" || "&faktord");
levels = levels // d;
values = values // levd;
end;

class  = class //  ("SUBJECT" || "&faktorp");
levels = levels // n;
values = values // levp;

print 'LD_Fi ---  subjects x Time_1 x ... x Time_i',
      'Time_1, ... Time_i: fixed, subjects: random'     ;

print 'SAS-datafile-name:  ' "&data" ,
      'Response variable:  ' "&var" ;
print 'Class Level Information';
print class levels '  '  values;

reset noname                                                    ;
print 'Total number of observations ' NNreal ' ',
      'Total number of subjects     ' n '  '                    ;

finish                                                          ;
/* Ende von O_CLI */

/*********************************************************************/
/* nu_char FUNKTION                                                  */
/* diese funktion schreibt numerische werte in characters um         */
/* und trimmt diese durch max(...) bzw. trimmt diese, falls          */
/* sie schon characters sind.                                        */
/*********************************************************************/
start nu_char(v,v_neu);
if type (v) ='N' then v_neu = char(v,max(int (log10(max(abs(v))))+1));
 else v_neu = trim(v);
finish;
/* ende der funktion nu_char */

/*********************************************************************/
/*** ausgabe der quadratformen und der p_values                    ***/
/*********************************************************************/
start test_out(box);
print 'Chi-Quadrat-Approximation';

titcol   = {"        Z_n"  "f_box" "f_pear" "p-val(ANOVA)" "p-val(box)" "p-val(pear)" };

titrow =  "Time1 (T1)    " ;
res = box[1,];
if ("&time2" = "none78xz45fa")=0 then do;
titrow  = titrow ||  "Time2 (T2)" ||  "T1*T2    " ;
res = res // box[2:3,];
end;
if ("&time3" = "none78xz45fa")=0 then do;
titrow  = titrow ||  "Time3 (T3)" ||  "T1*T3    " || "T2*T3    " || "T1*T2*T3   " ;
res = res // box[4:7,];
end;
if ("&time4" = "none78xz45fa")=0 then do;
titrow  = titrow ||  "Time4 (T4)" ||  "T1*T4    " || "T2*T4    "  || "T3*T4    "
```

```
|| "T1*T2*T4    " || "T1*T3*T4    " || "T2*T3*T4    "   || "T1*T2*T3*T4   " ;
res = res // box[8:15,];
end;

print 'Comparison of Z_n-Approximations with classic ANOVA';
print res[r=titrow][c=titcol][format=6.5];
finish;
/*****************************************************************/

/* Aufbereitung der Variablennamen  ******************************/

run nu_char(lev_a, leva);
run nu_char(lev_b, levb);
run nu_char(lev_c, levc);
run nu_char(lev_d, levd);
run nu_char(lev_p, levp);

txta  = CAT("",leva[1]);
do i = 2 to a;
txta = CAT(txta, " ", leva[i]);
end;
txtb  = CAT("",levb[1]);
do i = 2 to b;
txtb = CAT(txtb, " ", levb[i]);
end;
txtc  = CAT("",levc[1]);
do i = 2 to c;
txtc = CAT(txtc, " ", levc[i]);
end;
txtd  = CAT("",levd[1]);
do i = 2 to d;
txtd = CAT(txtd, " ", levd[i]);
end;
txtp  = CAT("",levp[1]);
do i = 2 to n;
txtp = CAT(txtp, " ", levp[i]);
end;

/*aufruf der funktionen*********************************************/

run o_cli(txta,txtb,txtc,txtd,txtp,a,b,c,d,n,NNreal);
run test_out(box);

call delete(execute78xz45fa);

quit;
%mend nn_hd_fi;
```

Literaturverzeichnis

[1] Bai, Z. und Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, **6(2)**: 311–329. ISSN 1017-0405.

[2] Becker, B. (2010). *Test für hochdimensionale Messwiederholungen mit unbekannten Kovarianzmatrizen*. Master's thesis, Universität Göttingen.

[3] Bhansali, R. J., Giraitis, L., und Kokoszka, P. S. (2007). Convergence of quadratic forms with nonvanishing diagonal. *Statist. Probab. Lett.*, **77(7)**: 726–734. ISSN 0167-7152. doi:10.1016/j.spl.2006.11.007.

[4] Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *Ann. Math. Statistics*, **25**: 290–302. ISSN 0003-4851.

[5] Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. Math. Statistics*, **25**: 484–498. ISSN 0003-4851.

[6] Brunner, E., Domhof, S., und Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York. ISBN 0-471-44166-X.

[7] Chen, S. X. und Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, **38(2)**: 808–835. ISSN 0090-5364. doi:10.1214/09-AOS716.

[8] Dufour, J.-M. und Hallin, M. (1990). An exponential bound for the permutational distribution of the first-order autocorrelation coefficient. *Statistique et Analyse des Donnees*, **15**: 45–56.

[9] Dufour, J.-M. und Hallin, M. (1993). Improved Eaton bounds for linear combinations of bounded random variables, with statistical applications. *J. Amer. Statist. Assoc.*, **88(423)**: 1026–1033. ISSN 0162-1459.

[10] Eaton, M. L. (1970). A note on symmetric Bernoulli random variables. *Ann. Math. Statist.*, **41**: 1223–1226. ISSN 0003-4851.

[11] Eaton, M. L. (1974). A probability inequality for linear combinations of bounded random variables. *The Annals of Statistics*, **2(3)**: 609–614.

[12] Hall, P. und Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York. ISBN 0-12-319350-8. Probability and Mathematical Statistics.

[13] Helms, H.-J. (2010). *Robuste Verfahren für strukturierte hochdimensionale Repeated-Measures-Designs unter Nicht-Normalverteilung*. Master's thesis, Universität Göttingen.

[14] Mattai, A. und Provost, S. B. (1992). *Quadratic forms in random variables: Theory and applications*. Marcel Dekker, INC., New York.

[15] Pearson, E. S. (1959). Note on an approximation to the distribution of non-central χ^2. *Biometrika*, **46**: 364. ISSN 0006-3444.

[16] Portnoy, S. (1986). On the central limit theorem in $\mathbf{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields*, **73(4)**: 571–583. ISSN 0178-8051. doi:10.1007/BF00324853.

[17] Ravishanker, N. und Dey, D. K. (2001). *A First Course in Linear Model Theory*. Wiley Series in Probability and Statistics. Chapman and Hall/CRC, first edition. ISBN 1584882476.

[18] Schott, J. R. (2005). *Matrix analysis for statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition. ISBN 0-471-66983-0.

[19] Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.*, **100(3)**: 518–532. ISSN 0047-259X. doi:10.1016/j.jmva.2008.06.006.

[20] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. ISBN 0-521-49603-9 ; 0-521-78450-6.

[21] Wang, H. und Akritas, M. G. (2010). Inference from heteroscedastic functional data. *J. Nonparametr. Stat.*, **22(1-2)**: 149–168. ISSN 1048-5252. doi:10.1080/10485250903171621.

[22] Werner, C. (2004). *Dimensionsstabile Approximation für Verteilungen von quadratischen Formen im Repeated-Measures-Design*. Master's thesis, Universität Göttingen.

Danksagung

An dieser Stelle möchte ich dem Betreuer meiner Diplomarbeit Herrn Prof. Dr. Brunner herzlich danken. Vor allem gilt dies den exzellenten Arbeitsbedingungen in der Abteilung für Medizinische Statistik und der hervorragenden Betreuung. Nicht zuletzt sind hier die zahlreichen Ratschläge und Diskussionen zu nennen.
Außerdem danke ich Herrn Prof. Dr. Schather für die Übernahme des Korreferats.

Mein Dank gilt auch den übrigen Mitarbeitern der Abteilung, welche mich stets unterstützt haben, und ohne welche die Diplomarbeit in dieser Form nicht möglich gewesen wäre. Vor allem ist hier die gute Atmosphäre und die vielen freundschaftlichen Ratschläge zu erwähnen.

Schließlich bedanke ich mich bei meinen Eltern und Großeltern, die mich immer unterstützt haben und mir all dies ermöglicht haben. Mein besonderer Dank gilt Sonja, die immer für mich da war und mir bei der Rechtschreibkorrektur geholfen hat.