Test für hochdimensionale Messwiederholungen mit unbekannten Kovarianzmatrizen

Diplomarbeit

vorgelegt von

Benjamin Markus Becker

aus Stuttgart

angefertigt am

Institut für Mathematische Stochastik

Georg-August-Universität Göttingen

2010

Inhaltsverzeichnis

Aber mit zauberisch fesselndem Blicke
winken die Frauen den Flüchtling zurücke,
warnend zurück in der Gegenwart *Spur*.
– Friedrich Schiller

1 Einleitung

Das Ziel dieser Arbeit ist die Ausarbeitung eines hochdimensionalen Zweistichprobentests für repeated measures (Messwiederholungen), der unbekannte, unstrukturierte und ungleiche Kovarianzmatrizen sowie gleichzeitig auch ungleiche Stichprobenumfänge zulässt.

Ein Test wird hochdimensional genannt, wenn der Stichprobenumfang nicht größer als die Anzahl der abhängigen Beobachtungen je Versuchseinheit ist, d. h. wenn die Beobachtungsvektoren in einem Raum einer Dimensionalität liegen, die mindestens so groß wie die Anzahl der Vektoren ist. Dies ist vor allem bei Microarrayanalysen oder bei Fragestellungen der klinischen Forschung der Fall, bei denen sehr viele Messungen an nur wenigen Versuchseinheiten durchgeführt werden. Meistens sind diese Messungen alle gleich skaliert, beispielsweise indem bei Microarrayanalysen stets Fluoreszenzintensitäten beobachtet werden oder bei bestimmten Medikamentenstudien regelmäßig wiederholte Messungen eines bestimmten Blutparameters der Versuchstiere analysiert werden. In diesem Fall werden die Daten als repeated measures bezeichnet. Typische Hypothesen betreffen die Gleichheit des Einflusses verschiedener Messzeitpunkte, verschiedene Stichprobenzugehörigkeiten oder die Wechselwirkung zwischen Stichprobenzugehörigkeit und Messzeitpunkt.

Abweichend vom üblichen Sprachgebrauch müssen in dieser Arbeit Messwiederholungen von multivariaten Daten unterschieden werden, bei denen an jeder Versuchseinheit völlig verschiedene Größen beobachtet werden können wie Körpergewicht, Blutdruck oder Blutzuckerspiegel. In diesem Falle haben Hypothesen über die Gleichheit der Beobachtungen natürlich keinen Sinn. Außerdem müssen Teststatistiken für diese Daten andere Invarianzeigenschaften als für repeated measures haben. Statistiken für Messwiederho-

lungen sollen invariant unter skalaren linearen Transformationen sein, damit der willkürliche Übergang zu einem anderen Größensystem –etwa vom metrischen zum angloamerikanischen– kein anderes Ergebnis liefert. Bei multivariaten Daten ist die allgemeinere Anforderung nach Invarianz auch unter komponentenweise verschiedenen skalaren Transformationen zu stellen, weil die verschiedenen Größen auch verschiedenen Einheitentransformationen unterworfen werden können. In beiden Fällen ist selbstverständlich auch Invarianz unter Permutation der Komponenten notwendig, d. h. unter der Anwendung der symmetrischen Gruppe auf die Reihenfolge der Komponenten, da der Sachverhalt typischerweise nicht von der Nummerierung der Beobachtungen abhängt. Für bestimmte Arten von Messwiederholungen, etwa wenn man mit Magnetresonanztomographie gewonnene Diffusionstensoren des Hirngewebes auswerten will, benötigt man Invarianz unter der Gruppe der orthogonalen Transformationen, weil dort die willkürliche Drehung der Patienten in eine beliebige Richtung einer orthogonalen Transformation des Beobachtungsvektors entspricht. Da die symmetrische Gruppe eine Untergruppe der orthogonalen ist, erfordern in diesem Fall repeated measures die allgemeinere Eigenschaft.

Im Hochdimensionalen ist die Unterscheidung zwischen multivariaten Daten und Messwiederholungen zwingend erforderlich, weil es keine nicht trivialen Teststatistiken geben kann, die sowohl unter komponentenweise verschiedenen skalaren als auch unter orthogonalen Transformationen invariant sind (siehe Lehmann [20], S. 318). Im Niedrigdimensionalen ist beides miteinander vereinbar, und es gibt seit langem Statistiken für beide Fälle, die dort die Unterscheidung überflüssig machen. So erbt Hotellings T^2-Statistik seine Invarianzeigenschaften von der Mahalanobis-Distanz, da in der T^2-Statistik nur das Inverse der unbekannten Kovarianzmatrix gegen das der empirischen Kovarianzmatrix ausgetauscht wird. Dies ist im Hochdimensionalen nicht länger möglich, weil dort die empirische Kovarianzmatrix singulär ist.

Stattdessen kann mit der sogenannten Wald-Typ-Statistik durch die Wahl einer verallgemeinerten Inversen der empirischen Kovarianzmatrix ein Test konstruiert werden, der nur noch Invarianz unter orthogonalen Transformationen zulässt. Die Verteilung dieser Statistik ist leider nur asymptotisch

bekannt. Dies scheint keine gute Annäherung für die Praxis zu sein, denn Simulationen zufolge wird bei wachsender Dimension der Test erst extrem liberal und später extrem konservativ.

Die ANOVA-Typ-Statistik, die auf Arbeiten von Box ([5], [6]) für ein- und zweifaktorielle Blockmodelle basiert, verzichtet auf die Invertierung der empirischen Kovarianzmatrix und setzt direkt an der Verteilung der quadratischen Form des Erwartungswertvektors an, die durch eine gestreckte χ^2-Verteilung gut approximiert werden kann. Geisser und Greenhouse haben diese Statistik für multivariate Mehrstichprobenmodelle angepasst und auf ihre Anwendbarkeit für hochdimensionale Tests hingewiesen. Ungelöst ist in diesen Arbeiten die befriedigende Schätzung des sogenannten Box'schen ϵ, des Freiheitsgrades der χ^2-Verteilung, der von der Spur des Quadrates der unbekannten Kovarianzmatrix und dem Quadrat ihrer Spur abhängt. Die Zuverlässigkeit des verbreiteten Ansatzes, eine bestimmte Kovarianzstruktur zu unterstellen und anschließend die Varianzkomponenten mit den bekannten Verfahren (z. B. REML, MINQUE, ANOVA) zu schätzen und mit diesem Ergebnis weiterzurechnen, hängt zu sehr von der unüberprüfbaren Annahme über die Kovarianzstruktur ab. Deswegen ist unbedingt ein Ansatz zu bevorzugen, der ohne Annahmen über die Kovarianzmatrix auskommt. Eine Option wäre es, die Spuren durch das naive Einsetzen der empirischen Kovarianzmatrix konsistent zu schätzen. Konsistenz ist aber im Hochdimensionalen kein ausreichendes Kriterium, weil insbesondere der Spurschätzer für das Quadrat der Kovarianzmatrix in diesen Situationen sehr verzerrt sein kann. Dies würde zu einer konservativen Statistik führen.

Für das hochdimensionale Zweistichprobenproblem bei identischen, unbekannten Kovarianzmatrizen schlagen Bai und Saranadasa [4] deswegen einen Freiheitsgradschätzer auf Basis der empirischen Kovarianzmatrix vor, der stochastisch gegen den echten Wert konvergiert, wenn der Quotient aus Stichprobenumfang und Dimensionalität bei wachsendem Stichprobenumfang gegen einen festen Wert konvergiert. Diese Asymptotik ist mittlerweile in der Literatur über hochdimensionales Testen weit verbreitet, obwohl man mit ihr in der Praxis stets zwei Approximationsfehler zu befürchten hat, nämlich wegen des Stichprobenumfanges und der Dimensionalität.

Eine strengere Asymptotik liegt der Arbeit von Werner [31] über das Einstichprobenproblem zugrunde. Hier wird zusätzlich zur Konsistenz und Erwartungstreue gefordert, dass das Verhältnis zwischen dem mittleren quadratischen Fehler und dem wahren Wert für jede Dimensionalität gleichmäßig beschränkt ist, so dass eine zu hohe Dimensionalität die Approximation nicht verderben kann. Diese Eigenschaft wird Dimensionsstabilität genannt. Mit Spurschätzern, die ausnutzen, dass unter der Nullhypothese die Hypothesenmatrix die Beobachtungsvektoren zentriert und dass quadratische und Bilinearformen dieser Vektoren gerade den gewünschten Erwartungswert haben, kann die Dimensionsstabilität erfüllt werden.

Nach genau diesem Schema sind in der Arbeit von Ahmad [2] die entsprechenden Schätzer für einen Zweistichprobentest mit identischen Kovarianzmatrizen oder identischen Stichprobenumfängen konstruiert. Der für die Praxis relevante Fall, dass Stichprobenumfänge und Kovarianzmatrizen zugleich verschieden sein können, kann damit aber nicht befriedigend gelöst werden, weil er eine Zentrierung unter der Alternative erfordert. In der vorliegenden Arbeit wird dies Problem umgangen, indem die identische Verteilung der Zufallsvektoren einer Stichprobe zur Zentrierung ausgenutzt wird. Das Ergebnis ist ebenfalls dimensionsstabil und eröffnet eine Erweiterungsmöglichkeiten für sehr allgemeine Mehrstichprobentests.

Die Gliederung dieser Arbeit sieht die Erklärung der benutzten Notation im nächsten Abschnitt vor. Danach werden im dritten Abschnitt die Modelle nebst Hypothesen formuliert und die Invarianzeigenschaften dargestellt, anhand derer Messwiederholungen von multivariaten Daten unterschieden werden müssen. Im vierten Abschnitt werden ausgehend von geeigneten Abstandsdefinitionen Teststatistiken vorgestellt und die approximative Verteilung der Statistik hergeleitet, die in dieser Arbeit konstruiert werden soll. Der fünfte Abschnitt ist bisherigen Arbeiten über verwandte Fragestellungen gewidmet. Die Spurschätzer für die neue Statistik werden im sechsten Abschnitt ausgearbeitet. Der siebte Abschnitt enthält eine Simulationsstudie, in der die neue Statistik untersucht wird und auch mit bestehenden, weniger allgemeinen Statistiken verglichen wird. Der darauffolgende Abschnitt führt in den Gebrauch und die Interna der Makros ein, die zur komfortablen Aus-

wertung von Daten mit der neuen Statistik programmiert worden sind. Im neunten Abschnitt wird damit beispielhaft eine solche Auswertung durchgeführt. Der letzte Abschnitt fasst die wichtigsten Inhalte der Arbeit zusammen und bietet einen Ausblick auf mögliche Weiterentwicklungen.

2 Notation

Matrizen werden mit fetten Großbuchstaben bezeichnet. Vektoren werden klein und fett geschrieben, Zeilen- oder Spaltenvektoren einer Matrix bleiben aber in Großschrift und erhalten einen Spalten- bzw. Zeilenindex. Skalare werden mit normalen Kleinbuchstaben abgekürzt. Quadratische- und Bilinearformen sowie Teststatistiken sind Ausnahmen. Für sie sind normale Großbuchstaben vorgesehen.

Die Standardbasisvektoren werden mit $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d \in \mathbb{R}^d$ und der Einservektor mit $\mathbf{1}_k = \sum_{j=1}^{k} \boldsymbol{e}_j \in \mathbb{R}^k$ bezeichnet. Wichtige Matrizen sind $\boldsymbol{J}_k = \mathbf{1}_k \mathbf{1}_k'$ und $\boldsymbol{P}_k = \boldsymbol{I}_k - \frac{1}{k}\boldsymbol{J}_k$. Man kann nachrechnen, dass die zentrierende Matrix $\boldsymbol{P}_k$ idempotent und symmetrisch ist. Ihr Rang ist $k - 1$.

Gruppen und ihre Repräsentationen als Mengen von Matrizen mit der Matrixmultiplikation oder als lineare Abbildungen mit Hintereinanderausführung werden miteinander identifiziert. Es wird also nicht unterschieden zwischen der orthogonalen Gruppe, der Menge der orthogonalen Matrizen und der Menge der orthogonalen linearen Abbildungen, denn wegen Isomorphie kann einheitlich O_k geschrieben werden. Entsprechend wird die symmetrische Gruppe $\mathbb{S}_k$ mit der Menge der Permutationsmatrizen identifiziert (siehe Fischer [15]).

Weitere wichtige Matrixgruppen sind $\mathbb{D}_d$, die Menge der Diagonalmatrizen, und GL_d die Menge der regulären Matrizen bzw. der invertierbaren linearen Abbildungen.

Wichtige Funktionen einer Matrix $\boldsymbol{A}$ sind ihre Spur $\mathrm{Sp}\boldsymbol{A}$ und ihr Rang $\mathrm{rg}\boldsymbol{A}$.

Stichproben werden mit i, i' usw. durchnummeriert, Versuchseinheiten mit k, l, s, t und die Stufen von Sub-Plot-Faktoren mit j. n_1 und n_2 sind die Stichprobenumfänge von Stichprobe 1 und 2. Der Gesamtstichprobenumfang ist n. Dies wird auch beim Einstichprobenlayout geschrieben.

3 Das Testproblem

3.1 Modelle und Hypothesen

In dem zugrunde liegenden Modell gibt es für jede der $n_1 + n_2$ Versuchseinheiten der Stichproben 1 und 2 jeweils d abhängige Messungen. Die d-Dimensionen Beobachtungsvektoren sind unabhängig und in jeder Stichprobe jeweils identisch normalverteilt mit unbekannten Erwartungswerten $\boldsymbol{\mu}_i$ und unbekannten Kovarianzmatrizen, die positiv semidefinit sind, d. h. $\boldsymbol{\Sigma}_i \geq 0$:

$$\boldsymbol{Y}_{j(i)} \sim \mathcal{N}_d\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right),\ i = 1, 2\, j = 1, \ldots, n_i \tag{1}$$

Als multivariates Modell betrachtet, bei dem jede der d Messungen eines Individuums unterschiedlich skaliert sein kann, würde nur die Hypothese $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ interessieren. Bei Messwiederholungen würde diese Hypothese bedeuten, dass der Verlauf der Messwerte in beiden Gruppen gleich ist. Mit solchen Daten sind aber auch Hypothesen über eine Strukturierung der d-dimensionalen Erwartungswertvektoren möglich, etwa wenn die Messwiederholungen einem oder mehreren Zeiteffekten unterworfen sind. Eine entsprechende Modellierung bieten *Split-Plot-Designs*. Das sind partiell hierarchische Versuchspläne, bei denen der zufällige Faktor Versuchseinheit unter einen festen Faktor A (Whole-Plot-Faktor) verschachtelt ist und beide mit mindestens einem festen Sub-Plot-Faktor gekreuzt sind. Der Sub-Plot-Faktor ist meistens die Zeit, die Bezeichnung verweist allerdings auf eine frühe Anwendung der Split-Plot-Designs in der Landwirtschaft, bei denen die Stufen der Sub-Plot-Faktoren ähnlichen, aber unterschiedlich behandelten Unterparzellen verschiedener Ackerflächen entsprachen. Der häufigste Fall mit nur einem Sub-Plot-Faktor wird *SP-a.b* abgekürzt, bei zweien schreibt man *SP-a.bc*, wobei a für die Anzahl der Stufen des Whole-Plot-Faktors steht und b und c für die Anzahl der Stufen der Sub-Plot-Faktoren B und C stehen. Im Zweistichprobenlayout ist folglich $a = 2$, und wegen der gekreuzten Faktoren B und C ist stets $bc = d$.

Das nichtsinguläre Modell ist folglich:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{\eta} \text{ mit } \boldsymbol{X} = (\mathbf{1}_{n_1} \oplus \mathbf{1}_{n_2}) \otimes \boldsymbol{I}_d,\ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \tag{2}$$

$\boldsymbol{\mu}_1$ und $\boldsymbol{\mu}_2$ sind entsprechend den Sub-Plot-Faktoren strukturiert, genauso $\boldsymbol{y} \in \mathbb{R}^N$, wobei $N = nd$. Bei SP-a.bc ist mit allen denkbaren Wechselwirkungen also

$$\mu_{ijl} = \mu + \alpha_i + \beta_l + \gamma_{l'} + (\alpha\beta)_{il} + (\alpha\gamma)_{il'} + (\beta\gamma)_{ll'} + (\alpha\beta\gamma)_{ill'}$$

mit $l = 1, \ldots, b$, $l' = 1, \ldots, c$ und bei SP-a.b einfach $\mu_{ij} = \mu + \alpha_i + \beta_l + (\alpha\beta)_{il}$, $l = 1, \ldots, b$. Die Hypothesen über die Parameter μ, α_i, β_j usw. werden in solchen linearen Modellen konventionell mit einer Hypothesenmatrix $\boldsymbol{H}$ als $H_0 : \boldsymbol{H}\boldsymbol{\mu} = \mathbf{0}$ dargestellt. Diese Matrix ist für äquivalente Hypothesen über die Parameter aber nicht eindeutig. Deswegen ist abzusichern, dass sinnvolle Statistiken nicht von der in diesem Rahmen willkürlichen Wahl der Hypothesenmatrix abhängen.

Lemma 3.1. *Sei $\boldsymbol{H}$ eine Hypothesenmatrix. Dann ist $\boldsymbol{T} = \boldsymbol{H}' (\boldsymbol{H}\boldsymbol{H}')^- \boldsymbol{H}$ eine symmetrische und idempotente Matrix, für die gilt:*

$$\boldsymbol{T}\boldsymbol{\mu} = \mathbf{0} \Leftrightarrow \boldsymbol{H}\boldsymbol{\mu} = \mathbf{0}.$$

Die Matrix $\boldsymbol{T}$ ist maximale Invariante unter der Wahl der Hypothesenmatrix, es gilt also:

$$(\boldsymbol{H}_1\boldsymbol{\mu} = \mathbf{0} \Leftrightarrow \boldsymbol{H}_2\boldsymbol{\mu} = \mathbf{0}) \Leftrightarrow \boldsymbol{T}(\boldsymbol{H}_1) = \boldsymbol{T}(\boldsymbol{H}_2)$$

Beweis. Siehe [7], Abschnitt 4.2. □

Daraus folgt, dass eine Statistik, die eine symmetrische und idempotente Hypothesenmatrix voraussetzt, stets invariant unter der genauen Wahl dieser Matrix ist ([20], S.216). Insbesondere ist es gleichgültig, wie die verallgemeinerte Inverse berechnet wird. Weil das Verfahren, das in dieser Arbeit entwickelt wird, symmetrische und idempotente Hypothesenmatrizen benötigt, wird die Matrixschreibweise der wichtigsten Hypothesen zu den Split-Plot-

Layout gleich in dieser Form eingeführt. Bei SP-2.b kann die Hypothese über den Haupteffekt der Behandlungsgruppe (A) wie folgt getestet werden:

$$\begin{aligned} & H_0\left(A\right): \mu_{1j} = \mu_{2j}\, j = 1, \ldots, d \\ \Leftrightarrow\, & H_0\left(A\right): \frac{1}{d}\left(\boldsymbol{P}_2 \otimes \boldsymbol{J}_d\right)\boldsymbol{\mu} = \mathbf{0} \\ \Leftrightarrow\, & H_0\left(A\right): \frac{1}{d}\boldsymbol{J}_d\boldsymbol{\mu}_1 - \frac{1}{d}\boldsymbol{J}_d\boldsymbol{\mu}_2 = \mathbf{0} \end{aligned}$$

Die letzte Zeile bietet eine Vereinfachung für das Zweistichprobenproblem, denn für $a > 2$ müsste die Hypothesenmatrix $\frac{1}{d}\left(\boldsymbol{P}_a \otimes \boldsymbol{J}_d\right)$ gewählt werden. So aber kann jede Hypothese als Differenz oder Summe zweier transformierter Erwartungswertvektoren aufgefasst werden, wie man bei der Hypothese über den Haupteffekt des Sub-Plot-Faktors Zeit (B) sieht:

$$\begin{aligned} & H_0\left(B\right): \mu_{ij} = \mu_{ij'},\, i = 1, 2;\, j, j' = 1, \ldots, d \\ \Leftrightarrow\, & H_0\left(B\right): \mu_{ij} - \overline{\mu_{i\cdot}} = 0 \\ \Leftrightarrow\, & H_0\left(B\right): \boldsymbol{P}_d\boldsymbol{\mu}_i = 0 \\ \Leftrightarrow\, & H_0\left(B\right): \boldsymbol{P}_d\boldsymbol{\mu}_1 + \boldsymbol{P}_d\boldsymbol{\mu}_2 = \mathbf{0} \end{aligned}$$

Völlig analog ist es auch bei der Wechselwirkung zwischen Zeit und Behandlungsgruppe (AB):

$$\begin{aligned} & H_0\left(AB\right): \mu_{ij} - \mu_{i'j'} = 0,\, i, i' = 1, 2;\, j.j' = 1, \ldots, d \\ \Leftrightarrow\, & H_0\left(AB\right): \mu_{ij} - \overline{\mu_{i\cdot}} - \overline{\mu_{\cdot j}} + \overline{\mu_{\cdot\cdot}} = 0 \\ \Leftrightarrow\, & H_0\left(AB\right): \mu_{1j} - \overline{\mu_{1\cdot}} - \mu_{2j} + \overline{\mu_{2\cdot}} = 0 \\ \Leftrightarrow\, & H_0\left(AB\right): \boldsymbol{P}_d\boldsymbol{\mu}_1 - \boldsymbol{P}_d\boldsymbol{\mu}_2 = \mathbf{0}. \end{aligned}$$

Die Formulierung der Hypothese für den Haupteffekt A ist im SP-2.bc-Layout identisch. Wegen der zweifaktoriellen Strukturierung von

$$\boldsymbol{\mu}_i = \left(\mu_{i11}, \ldots, \mu_{i1c}, \mu_{i21}, \ldots, \mu_{ibc}\right)'$$

stellen sich die Hypothesenmatrizen als Kroneckerprodukte dar:

$$\begin{aligned} H_0(B) :& \mu_{ijl} - \mu_{ij'l} = 0 \\ \Leftrightarrow& \mu_{ijl} - \overline{\mu_{i\cdot l}} = 0 \\ \Leftrightarrow& \frac{1}{c}\left(\boldsymbol{P}_b \otimes \boldsymbol{J}_c\right)\boldsymbol{\mu}_1 + \frac{1}{c}\left(\boldsymbol{P}_b \otimes \boldsymbol{J}_c\right)\boldsymbol{\mu}_2 = \mathbf{0} \end{aligned}$$

Genauso berechnet man die anderen Hypothesenmatrizen:

$$\begin{aligned} H_0(C) :& \frac{1}{b}\left(\boldsymbol{J}_b \otimes \boldsymbol{P}_c\right)\left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right) = \mathbf{0} \\ H_0(AB) :& \frac{1}{c}\left(\boldsymbol{P}_b \otimes \boldsymbol{J}_c\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) = \mathbf{0} \\ H_0(AC) :& \frac{1}{b}\left(\boldsymbol{J}_b \otimes \boldsymbol{P}_c\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) = \mathbf{0} \\ H_0(BC) :& \left(\boldsymbol{P}_c \otimes \boldsymbol{P}_b\right)\left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right) = \mathbf{0} \\ H_0(ABC) :& \left(\boldsymbol{P}_c \otimes \boldsymbol{P}_b\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) = \mathbf{0} \end{aligned}$$

Man kann leicht überprüfen, dass die genannten Hypothesenmatrizen alle symmetrisch und idempotent sind. Ganz analog können auch Hypothesen gebildet werden, wenn drei oder noch mehr Sub-Plot-Faktoren vorliegen. Diese Fälle sind aber so selten, dass sie nicht weiter berücksichtigt werden. Gleiches gilt für die Formulierung spezieller Hypothesen über Kontraste der Sub-Plot-Faktorstufen. Liegt erst eine entsprechende Kontrastmatrix vor, kann sie mit Lemma 3.1 in die benötigte Gestalt gebracht werden, so dass sich für die Theorie nichts weiter ändert.

Der Einfachheit halber soll fortan jede Hypothese durch $H_0 : \boldsymbol{T}\boldsymbol{\mu}_1 - \boldsymbol{T}\boldsymbol{\mu}_2 = \mathbf{0}$ dargestellt werden. Für die Hypothesen über den Gruppeneffekt oder über eine Wechselwirkung mit der Stichprobenzugehörigkeit passt dies offensichtlich. Die Hypothesen nur über Sub-Plot-Faktoren fügen sich in diesen Rahmen, wenn in der zweiten Stichprobe einfach alle Vorzeichen umgedreht werden. An der Kovarianzmatrix von $\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)$ ändert das nichts. Im Rest dieser Arbeit kann also ohne Einschränkung $\boldsymbol{T}\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)$ geschrieben werden, wo für Hypothesen über Sub-Plot-Faktorstufen $\boldsymbol{T}\left(\overline{\boldsymbol{Y}}_{1\cdot} + \overline{\boldsymbol{Y}}_{2\cdot}\right)$ stehen müsste.

3.2 Schätzung von vielen Varianzkomponenten

Der Term $\boldsymbol{\eta}$ wird für das SP-2.b-Design in der Literatur fast immer sinngemäß (siehe etwa [23] oder [25]) als

$$\eta_{ijl} = z_{ij} + \epsilon_{ijl}$$

mit unabhängig und identisch verteilten Zufallsvariablen $z_{ij} \sim \mathcal{N}(0, \sigma_V^2)$, $\epsilon_{ijl} \sim \mathcal{N}(0, \sigma^2)$ modelliert. Das Ergebnis ist eine Kovarianzmatrix des Beobachtungsvektors mit einer sogenannten Compound-Symmetry-Struktur:

$$\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{I}_n \oplus \left(\sigma^2 \boldsymbol{I}_d + \sigma_V^2 \boldsymbol{J}\right)$$

Dieses Modell ist auch bei hoher Dimensionalität für Varianzkomponentenverfahren zugänglich. Allerdings können die Annahmen von gleichen Korrelationen der Messungen untereinander bei echten Daten sehr unrealistisch sein. Aldworth und Hoffman ([3]) mahnen sogar an, die Möglichkeit negativer Kovarianzen in bestimmten Fällen einzubeziehen. Stimmen all diese Annahmen nicht, kann das Niveau verfehlt werden. Bei hochdimensionalen Daten wird dieses Problem noch schlimmer, da ein Varianzkomponentenverfahren mit der falschen Struktur in sehr vielen falsch geschätzten Parametern mündet. Als Beispiel sind in der Tabelle 1 Niveausimulationen der ANOVA-Methode für eine eine Compound-Symmetry-Struktur angeführt, bei denen in Wahrheit eine autoregressive Kovarianzstruktur

$$\boldsymbol{\Sigma}_i = \left(\rho^{|l-j|}\right)_{l,j=1,\ldots,d}$$

vorlag.

d	2	5	10	50	100
$\rho = 0,6$	0.0555	0,0675	0,0905	0,108	0,114
$\rho = 0,9$	0,05	0,098	0,136	0,229	0,2405

Tabelle 1: Simuliertes 5%-Niveau von der ANOVA-Methode bei angenommener Compound-Symmetry-Struktur, obwohl autoregressive Struktur vorliegt. $n_1 = n_2 = 10$

Man sieht in der Tabelle, dass der Test liberaler wird, je mehr falsch geschätzte Parameter in der Kovarianzmatrix vorkommen. Unterscheiden sich die Parameter besonders stark ($\rho = 0,9$), dann wird dies noch schlimmer.

3.3 Multivariates Modell

Multivariate lineare Modelle werden konventionell durch

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{3}$$

beschrieben, wobei $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ die beobachteten Daten sind, $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ die Designmatrix und $\boldsymbol{B} \in \mathbb{R}^{k\times d}$ die Parametermatrix ist, sowie $\boldsymbol{E} = \left(\boldsymbol{E}_j'\right)_{j=1,\ldots,N}$ mit $\boldsymbol{E}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ die Fehlerterme enthält. Diese Schreibweise kann das Split-Plot-Design aus (2) aufnehmen, wenn $\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 \end{pmatrix}'$ und $\boldsymbol{X} = \boldsymbol{1}_{n_1} \oplus \boldsymbol{1}_{n_2}$ sind. Zwar testet man mit multivariaten Verfahren hauptsächlich Hypothesen der Form $H_0 : \boldsymbol{H}\boldsymbol{B} = \boldsymbol{0}$, z. B. $H_0 : \boldsymbol{P}_2\boldsymbol{B} = \boldsymbol{0}$, aber wie in Abschnitt 3.1 gesehen, können die Hypothesen im Split-Plot-Modell durch die Multiplikation mit einer Hypothesenmatrix auf diese Form gebracht werden. Bei SP-2.b sind beispielsweise die Schreibweisen $H_0\,(AB) : \boldsymbol{P}_2\boldsymbol{B}\boldsymbol{P}_d = \boldsymbol{0}$ und $H_0\,(B) : \boldsymbol{I}_2\boldsymbol{B}\boldsymbol{P}_d = \boldsymbol{0}$ möglich. Im wesentlichen müssen beim multivariaten Modell nur auf beiden Seiten die Matrizen transponiert werden und der vec-Operator angewendet werden, der die Spalten einer Matrix untereinander schreibt, so dass ein langer Vektor entsteht, um eine univariate Darstellung zu erhalten.

Da in der multivariaten Statistik anders als bei den Verfahren mit Varianzkomponentenschätzung für gemischte Modelle kaum Annahmen über die Struktur der Kovarianzmatrizen gemacht werden und gerade gezeigt wurde, dass das Split-Plot-Modell in den Rahmen eines multivariaten Modells (sogenannte GMANOVA, siehe [22, 24]) gesetzt werden kann, liegt es nahe, multivariate Verfahren anwenden zu wollen. Im folgenden Unterabschnitt soll erklärt werden, warum die multivariate Notation bei hochdimensionalen Messwiederholungen trügerisch ist.

3.4 Konkurrierende Invarianzeigenschaften hochdimensionaler Tests

Ansatzpunkt ist die Beachtung des Äquivarianzprinzips, das bei [11] zu finden ist. Es fordert, dass eine Transformation der Daten, die den zugrundeliegenden Sachverhalt nicht verfälscht, auch das Ergebnis eines Tests nicht verändern darf. Im folgenden werden Transformationen vorgestellt, die bei Messwiederholungen oder multivariaten Daten den Sachverhalt nicht verändern. Man wird sehen können, dass diese Transformationen allesamt Gruppenstrukturen haben.

Definition 3.2. Sei G eine Gruppe, die auf $\mathbb{R}^{n\times d}$ operiert. Dann heißt der Test ϕ *invariant unter der Gruppe* G, genau dann wenn

$$\phi\left(f\left(\boldsymbol{Y}\right)\right) = \phi\left(\boldsymbol{Y}\right) \text{ für alle } f \in G \text{ und für alle } \boldsymbol{Y} \in \mathbb{R}^{n\times d}.$$

Da die Spalten von $\boldsymbol{Y}$ die Messungen darstellen, besorgt die Operation der symmetrischen Gruppe

$$\mathbb{S}_d = \{\pi | \pi : \{1,\ldots,d\} \to \{1,\ldots,d\},\ \pi \text{ ist bijektiv}\}$$

auf den Spalten von $\boldsymbol{Y}$ die Umnummerierung der Messungen. Dies darf das Ergebnis des Tests freilich nicht verändern, ein Test muss also die Invarianz unter dieser Gruppe einhalten. Aus der Theorie zur Gauß-Elimination ist bekannt (siehe [14], S.12), dass eine Matrixrepräsentation der symmetrischen Gruppe die Menge der Permutationsmatrizen

$$\mathbb{P}_d = \left\{ \boldsymbol{P}_\pi = \begin{pmatrix} \boldsymbol{e}_{\pi(1)} & \ldots & \boldsymbol{e}_{\pi(d)} \end{pmatrix} | \pi \in \mathbb{S}_d \right\}$$

mit der Matrixmultiplikation ist. Die Operation auf den Spalten von $\boldsymbol{Y}$ wird so durch $(\boldsymbol{P}_\pi, \boldsymbol{Y}) \mapsto \boldsymbol{Y}\boldsymbol{P}_\pi$ beschrieben. Die Umformung von (3) zu $\boldsymbol{Y}\boldsymbol{P}_\pi = (\boldsymbol{X}\boldsymbol{B} + \boldsymbol{E})\boldsymbol{P}_\pi$ erhält also alle Eigenschaften der Modellierung. Dieselbe Überlegung ist auf die Nummerierung der n Versuchseinheiten anzuwenden. Analog zur Permutation der Spalten durch Rechtsmultiplikation mit Matrizen aus $\mathbb{P}_d$ erreicht man die Umnummerierung der Zeilen durch Links-

multiplikation mit Elementen aus $\mathbb{P}_n$.

Die Gruppe der Permutationsmatrizen $\mathbb{P}_n$ ist eine Untergruppe der orthogonalen Gruppe O_n, es gilt also stets $\boldsymbol{P}_\pi \boldsymbol{P}'_\pi = \boldsymbol{I}$. Tatsächlich ist es sinnvoll, sofort Invarianz unter O_n zu fordern, da die grundlegenden Schätzer für die Parameter, über die die linearen Hypothesen gebildet werden, orthogonal invariant sind. Die folgende Definition der kanonischen Form basiert auf dieser Invarianzeigenschaft.

Definition 3.3. Seien $\boldsymbol{X} \in \mathbb{R}^{n\times k}$, $\boldsymbol{B} \in \mathbb{R}^{k\times d}$ und $\mathrm{Cov}\,(\boldsymbol{E}_{i\cdot}) = \boldsymbol{\Sigma}$. Die Zufallsmatrix $\boldsymbol{Y}^* \in \mathbb{R}^{n\times d}$ heißt *multivariate kanonische Form* zum multivariaten linearen Modell $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ mit der Hypothese $H_0 : \boldsymbol{H}\boldsymbol{B} = \boldsymbol{0}$, $\mathrm{rg}\boldsymbol{H} = r \leq k$, falls es eine Matrix $\boldsymbol{Q} \in O_n$ gibt, so dass $\boldsymbol{Y}^* = \boldsymbol{Q}\boldsymbol{Y}$ und $\boldsymbol{Y}^* = \begin{pmatrix} \boldsymbol{Y}^1 \\ \boldsymbol{Y}^2 \\ \boldsymbol{Y}^3 \end{pmatrix}$ gelten, wobei $E\left(\boldsymbol{Y}^1\right) \overset{H_0}{=} \boldsymbol{0} \in \mathbb{R}^{r\times d}$, $E\left(\boldsymbol{Y}^2\right) \in \mathbb{R}^{(k-r)\times d}$ und $E\left(\boldsymbol{Y}^3\right) = \boldsymbol{0} \in \mathbb{R}^{(n-k)\times d}$ sind.

Anschaulich gesagt werden auf diese Weise die Komponenten, die das Modell spezifizieren, die „Erfüllung" der Hypothese angeben und die nur die Varianz in den Daten beinhalten (entspricht dem Residuenraum, $\boldsymbol{Y}^1$ und $\boldsymbol{Y}^2$ bilden den Schätzraum), voneinander getrennt. Die kanonische Form des multivariaten Zweistichprobenproblems oder der Hypothese $H_0\,(A)$ aus dem Split-Plot-Modell (2) wird durch eine orthogonale Matrix $\boldsymbol{Q}$ erreicht, deren erste beiden Zeilen lauten:

$$\begin{aligned} \boldsymbol{Q}_{1\cdot} &= \begin{pmatrix} \sqrt{\frac{n_2}{n_1 n}}\boldsymbol{1}'_{n_1} & -\sqrt{\frac{n_1}{n_2 n}}\boldsymbol{1}'_{n_2} \end{pmatrix} \\ \boldsymbol{Q}_{2\cdot} &= \frac{1}{\sqrt{n}}\boldsymbol{1}'_n \end{aligned} \tag{4}$$

Die übrigen Zeilen können durch ein Orthonormalisierungsverfahren gewonnen werden. Mit der Anpassung von Split-Plot-Hypothesen an das multivariate Modell ist $\boldsymbol{Y}^* = \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{P}_d$ die kanonische Form zur Hypothese $H_0\,(AB)$ im Modell SP-2.b. Ändert man die erste Zeile von $\boldsymbol{Q}$ auf

$$\boldsymbol{Q}_{1\cdot} = \begin{pmatrix} \sqrt{\frac{n_2}{n_1(n_1+n_2)}}\boldsymbol{1}_{n_1} & \sqrt{\frac{n_1}{n_2(n_1+n_2)}}\boldsymbol{1}_{n_2} \end{pmatrix},$$

dann gehört $\boldsymbol{Y}^* = \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{P}_d$ zur Hypothese $H_0(B)$.

Die Definition der kanonischen Form findet sich bei [20], S. 294, und [22], S. 433. Sie kann als eine Verallgemeinerung der univariaten kanonischen Form ($d = 1$) für feste Effekte verstanden werden, wie sie in [27], S. 37ff., und [20], S. 265, angegeben werden. In [25] wird diese univariate kanonische Form auch auf lineare Modelle mit zufälligen Effekten, wie das Split-Plot-Modell in der Schreibweise (2), verallgemeinert. [30] enthebt diese Definition der Einschränkung, nur reguläre Kovarianzmatrizen zulassen zu können, so dass folgende Definition auch beim Split-Plot-Design (2) verwendet werden kann, wenn man $N = nd$ auffasst:

Definition 3.4. Sei $\boldsymbol{X} \in \mathbb{R}^{N\times k}$, $\boldsymbol{b} \in \mathbb{R}^k$ und $\mathrm{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Sigma} \geq 0$. Der Zufallsvektor $\boldsymbol{y}^* \in \mathbb{R}^N$ heißt *kanonische Form* zum linearen Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\eta}$ mit der Hypothese $\boldsymbol{H}\boldsymbol{b} = \boldsymbol{0}$, $\mathrm{rg}\boldsymbol{H} = r$, falls es eine Matrix $\boldsymbol{Q} \in O_N$ gibt, so dass $\boldsymbol{y}^* = \boldsymbol{Q}\boldsymbol{y}^*$ und $E(\boldsymbol{y}^*) = \begin{pmatrix} \boldsymbol{y}^1 \\ \boldsymbol{y}^2 \\ \boldsymbol{y}^3 \end{pmatrix}$ gelten, wobei $E(\boldsymbol{y}^1) \overset{H_0}{=} \boldsymbol{0}$, $\boldsymbol{y}^2 \in \mathbb{R}^{(k-r)}$ und $E(\boldsymbol{y}^3) = \boldsymbol{0} \in \mathbb{R}^{(N-k)}$ sind.

Tests für das Split-Plot-Design auf Grundlage dieser Definition setzen also stets Invarianz unter Linksmultiplikation von $\boldsymbol{y}$ mit O_N voraus. Übersetzt man dies durch $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y}')$ in die multivariate Schreibweise, folgt somit Invarianz unter Rechtsmultiplikation mit O_d. Bei multivariaten Tests kann aber nur Invarianz unter Rechtsmultiplikation mit $\mathbb{P}_d$ motiviert werden.

Ein entscheidender Unterschied zwischen beiden kanonischen Formen ist also, dass für das Split-Plot-Modell in der Schreibweise (3) für die Definition 3.3 keine weiteren Invarianzeigenschaften über die Spalten von $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ vorausgesetzt werden. Hingegen beruht die kanonische Form laut Definition 3.4 auf Invarianz unter Linksmultiplikation von O_N.

Neben der Nummerierung ist die Skalierung der Messungen durch ein bestimmtes Einheitensystem ein unerwünschter Einflussfaktor, den eine entsprechende Invarianzeigenschaft ausschalten muss. Der Übergang von einer linearen Skalierung zu einer anderen wird durch eine affin-lineare Transformation $y \mapsto gy + h$ mit $g \neq 0$ erreicht. Wenn die Gruppe der nichtsingulären

Diagonalmatrizen durch

$$\mathbb{D}_d = \left\{ \boldsymbol{G} = \begin{pmatrix} g_1 & & 0 \\ & \ddots & \\ 0 & & g_d \end{pmatrix}, \, g_i \neq 0 \right\} \subset GL_d$$

gegeben ist und $\mathbb{R}^d$ mit der Vektoraddition betrachtet wird, dann ist die Menge der komponentenweise verschiedenen affin-linearen Transformationen

$$\mathbb{A}_d = \left\{ (\boldsymbol{G}, \boldsymbol{h}) : \boldsymbol{Y} \mapsto (\boldsymbol{Y} + \mathbf{1}_n \boldsymbol{h}') \, \boldsymbol{G}, \, \boldsymbol{G} \in \mathbb{D}_d, \, \boldsymbol{h} \in \mathbb{R}^d \right\}$$

eine Gruppe mit der Verknüpfung

$$(\boldsymbol{G}_1, \boldsymbol{h}_1) \circ (\boldsymbol{G}_2, \boldsymbol{h}_2) : \boldsymbol{Y} \mapsto (\boldsymbol{Y} + \mathbf{1}_n \boldsymbol{h}_1' + \mathbf{1}_n \boldsymbol{h}_2') \, \boldsymbol{G}_1 \boldsymbol{G}_2,$$

denn Assoziativität, das neutrale Element $(\boldsymbol{I}_d, \mathbf{0}_d)$ und das zu $(\boldsymbol{G}, \boldsymbol{h})$ inverse Element $\left(\boldsymbol{G}^{-1}, -\boldsymbol{h}\right)$ werden von den Gruppen $\mathbb{D}_d$ und $\mathbb{R}^d$ geerbt. Die Operation ist offensichtlich.

Man kann aber auch selbstinverse Abbildungen von mehreren Variablen bilden. Beispielsweise wäre es der Anwendungssicherheit dienlich, wenn Messwiederholungen auch als Baselinewert mit Differenzen zu diesem Wert dargestellt werden können, ohne dass dies ein abweichendes Ergebnis ergäbe. Lineare Transformationen dieser Art haben die Eigenschaft, zu sich selbst invers zu sein.

Um einen für alle Arten von hochdimensionalen Daten universell einsetzbaren Test zu haben, wäre es wünschenswert, Invarianz unter einer Gruppe von linearen Abbildungen zu haben, die alle genannten zugleich umfasst. Dies läuft auf die Gruppe der invertierbaren linearen Abbildungen hinaus, die durch reguläre Matrizen repräsentiert wird. Das folgende Theorem von Lehmann ([20], S. 318) besagt, dass diese Anforderung kein hochdimensionaler Test erfüllen kann:

Theorem 3.5. *Sei $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ eine Zufallsmatrix, deren Zeilen stochastisch unabhängig sind, mit existierenden positiv definiten Kovarianzmatrizen stetig*

verteilt sind und Erwartungswert $E(\boldsymbol{Y}) = \begin{pmatrix} \boldsymbol{\mu}_1 \mathbf{1}'_{n_1} & \boldsymbol{\mu}_2 \mathbf{1}'_{n_2} \end{pmatrix}$ *haben. Dann gibt es für* $n_1 + n_2 = n \leq d \in \mathbb{N}$ *keinen nichttrivialen, hochdimensionalen Test zur Hypothese* $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$*, der zugleich unter* GL_d *und* O_n *invariant ist.*

Zum Beweis wird ein Lemma benötigt:

Lemma 3.6. *Sei* $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ *mit* $n \leq d$ *eine Zufallsmatrix, deren Zeilen stochastisch unabhängig sind. Die Zeilen seien stetig verteilt, so dass ihre Kovarianzmatrizen existieren und positiv definit sind. Dann sind die Zeilen von* $\boldsymbol{Y}$ *fast sicher linear unabhängig.*

Beweis. Da wegen der stetigen Verteilung der Komponenten fast sicher keiner der Zeilenvektoren genau im Ursprung liegt, würde lineare Abhängigkeit bedeuten, dass ein Zeilenvektor $\boldsymbol{Y}'_{j\cdot}$ in dem Unterraum $\mathbb{U}$ liegt, der von den anderen $n-1$ Zeilenvektoren aufgespannt wird:

$$\boldsymbol{Y}'_{j\cdot} \in \operatorname{span}\left(\ldots, \boldsymbol{Y}'_{(j-1)\cdot}, \boldsymbol{Y}'_{(j+1)\cdot}, \ldots\right) = \mathbb{U} \subsetneq \mathbb{R}^d$$

Das bedeutet aber, dass $\boldsymbol{Y}'_{i\cdot}$ auf das Komplement $\mathbb{U}^{\mathsf{C}}$ von $\mathbb{U}$ projiziert Null ist. Da aus den $\boldsymbol{Y}'_{j'\cdot}$, $j' \neq j$ eine Basis für $\mathbb{U}$ ausgewählt werden kann und diese zu einer Basis für $\mathbb{U} \oplus \mathbb{U}^{\mathsf{C}}$ ergänzt werden kann, muss sich also mit der Basistransformation $\boldsymbol{B}$ von der Standardbasis auf diese Basis der Vektor $\boldsymbol{Y}'_{j\cdot}$ als

$$\boldsymbol{B}\boldsymbol{Y}'_{i\cdot} = \begin{pmatrix} * & \mathbf{0}_{1 \times \dim\left(\mathbb{U}^{\mathsf{C}}\right)} \end{pmatrix}'$$

schreiben lassen können. Die Kovarianzmatrix von $\boldsymbol{B}\boldsymbol{Y}'_{i\cdot}$ kann mit einer Matrix $\boldsymbol{Q}_d \in O_d$ diagonalisiert werden. Weil die Kovarianzmatrix von $\boldsymbol{B}\boldsymbol{Y}'_i$ auch nichtsingulär ist, sind ihre Eigenwerte positiv. Wegen der stetigen Verteilung von $\boldsymbol{B}\boldsymbol{Y}'_{i\cdot}$ und der Unabhängigkeit von den anderen $\boldsymbol{Y}'_{k\cdot}$ gilt nun:

$$\begin{aligned} &P\left(\boldsymbol{Q}_d \boldsymbol{B}\boldsymbol{Y}'_{i\cdot} = \boldsymbol{Q}_d \begin{pmatrix} * & \mathbf{0}_{1 \times \dim\left(\mathbb{U}^{\mathsf{C}}\right)} \end{pmatrix}' \middle| \boldsymbol{Y}'_{k\cdot}, k \neq i\right) \\ =&P\left(\boldsymbol{Q}_d \boldsymbol{B}\boldsymbol{Y}'_{i\cdot} = \boldsymbol{Q}_d \begin{pmatrix} * & \mathbf{0}_{1 \times \dim\left(\mathbb{U}^{\mathsf{C}}\right)} \end{pmatrix}'\right) = 0 \end{aligned}$$

□

Beweis. Zu Theorem 3.5. Es kann angenommen werden, dass die Daten in kanonischer Form sind. Durch Addition eines Vektors $\boldsymbol{h} \in \mathbb{R}^d$ kann $\boldsymbol{Y}^2$ bei Invarianz unter $\mathbb{A}_d$ eliminiert werden. Weil die Zeilen von $\boldsymbol{Y}^1$ und $\boldsymbol{Y}^3$ fast sicher linear unabhängig sind, gibt es ein $\boldsymbol{A} \in GL_d$, so dass

$$\begin{pmatrix} \boldsymbol{Y}^1 \\ \boldsymbol{0} \\ \boldsymbol{Y}^3 \end{pmatrix} \boldsymbol{A} = \begin{pmatrix} 1 & 0 & \dots & & & \\ 0 & 0 & \dots & & & \\ 0 & 1 & 0 & \dots & & \\ \vdots & & \ddots & & & \\ 0 & \dots & 0 & 1 & 0 & \dots \end{pmatrix}.$$

Die Testentscheidung ist also fast sicher immer dieselbe. □

Da GL_d von O_d und $\mathbb{D}_d$ erzeugt wird –dies folgt aus der Existenz einer Singulärwertzerlegung– liegt es nahe, für multivariate Daten auf Invarianz unter O_d zu verzichten und sich mit $\mathbb{A}_d$ zu begnügen. Für Messwiederholungen ist die Addition eines konstanten Vektors ohnehin unverträglich mit den Hypothesen über die Sub-Plot-Faktoren. Invarianz unter $\mathbb{D}_d$ ist ebenfalls unnötig, da identische Größen nicht durch unterschiedliche Skalen festgelegt werden müssen. Auch Invarianz unter selbstinversen linearen Transformationen ist nicht notwendig. Letztlich muss ein Test für Messwiederholungen nur unter O_d und $\mathbb{R}^*$ invariant sein.

4 Ansätze für Teststatistiken

4.1 Nachteile bekannter Statistiken

Eine sinnvoller Test für die Gleichheit zweier Stichproben verwirft die Nullhypothese, wenn der Abstand zwischen den Mittelwertvektoren beider Stichproben zu groß wird. Wie am Ende vom Unterabschnitt 3.1 erklärt, kann dieser Gedanke auch für andere Hypothesen angewendet werden. Eine etablierte Definition für den Abstand zweier Vektoren ist durch den Mahalanobis-Abstand gegeben. Haben zwei Zufallsvektoren $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^d$ dieselbe Kovarianzmatrix $\boldsymbol{S} = E\left(\left(\boldsymbol{Z}_i - E\boldsymbol{Z}_i\right)\left(\boldsymbol{Z}_i - E\boldsymbol{Z}_i\right)'\right) > 0$ für $i = 1, 2$, dann ist die Mahalanobis-Distanz durch

$$\delta_{\boldsymbol{S}}\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2\right) = \left(\boldsymbol{Z}_1 - \boldsymbol{Z}_2\right)' \boldsymbol{S}^{-1}\left(\boldsymbol{Z}_1 - \boldsymbol{Z}_2\right)$$

definiert. Man sieht sofort, dass für jede Matrix $\boldsymbol{A} \in GL_d$ gilt

$$\delta_{\boldsymbol{ASA}'}\left(\boldsymbol{AZ}_1 - \boldsymbol{AZ}_2\right) = \delta_{\boldsymbol{S}}\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2\right).$$

In der Realität ist die Kovarianzmatrix natürlich unbekannt. Wird sie durch die empirische Kovarianzmatrix ersetzt, dann ergibt sich die Statistik von Hotelling [18]. Benutzt man eine Hypothesenmatrix $\boldsymbol{H}$ mit vollem Rang, die beispielsweise aus den Hypothesenmatrizen in 3.1 durch das Entfernen der richtigen Anzahl von Zeilen erzeugt werden kann, und wird mit $\hat{\boldsymbol{S}}_{n-2} = \frac{1}{n-2}\sum_{i=1}^{2}\left(n_i - 1\right)\sum_{t=1}^{n_i}\left(\boldsymbol{Y}_{i(t)} - \overline{\boldsymbol{Y}}_{i(t)}\right)'\left(\boldsymbol{Y}_{i(t)} - \overline{\boldsymbol{Y}}_{i(t)}\right)$ die gepoolte empirische Kovarianzmatrix bezeichnet, so basiert die Hotelling-Statistik (siehe [22], S.216) auf dem Term

$$T^2 = \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)\boldsymbol{H}'\left(\boldsymbol{H}\hat{\boldsymbol{S}}_{n-2}\boldsymbol{H}'\right)^{-1}\boldsymbol{H}\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)'.$$

Dies ist wie die Mahalanobis-Distanz invariant unter GL_d und damit für Messwiederholungen und multivariate Daten zugleich geeignet. Diese Eigenschaft erhält sich auch in dem modifizierten Test für den Fall nichtidentischer Kovarianzmatrizen und ungleicher Stichprobenumfänge, der bei Krishnamoorthy und Yu [19] ausgearbeitet ist.

Wie Theorem 3.5 nahelegt, kann die Hotelling-Statistik aber nur für niedigdimensionale Daten definiert sein. Ein offensichtlicherer Grund ist, dass fast sicher $\mathrm{rg}\hat{\boldsymbol{S}}_{n-2} = \min(n, d)$ gilt (siehe Satz A.1) und somit $\boldsymbol{H}\hat{\boldsymbol{S}}_{n-2}\boldsymbol{H}$ nicht invertierbar ist. Als Ausweg kann man die Inverse der empirischen Kovarianzmatrix durch eine andere geeignete Matrix ersetzen. Srivastava und Du [28] schlagen eine studentisierung der Komponenten vor, indem die Matrix $\boldsymbol{R}$ der Wurzeln der Diagonaleinträge von $\boldsymbol{S}_n$ invertiert eingesetzt wird:

$$T_{SD} = \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right) \boldsymbol{R}^{-1} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)'$$

Dieser Term verändert sich nicht, wenn $\boldsymbol{Y}$ mit $\boldsymbol{D} \in \mathbb{D}_d$ transformiert wird, wohl aber, wenn eine Transformation mit einem Element aus O_d durchgeführt wird. Das macht diesen Ansatz geeignet für hochdimensionale multivariate Daten, nicht für Messwiederholungen. Deswegen soll darauf nicht weiter eingegangen werden.

Eine andere Möglichkeit, das Inverse der empirischen Kovarianzmatrix zu ersetzen, ist die Moore-Penrose-Inverse $(\cdot)^+$. Dies geschieht in der sogenannten Wald-Typ-Statistik:

$$W_n = \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right) \boldsymbol{H}' \left(\boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}'\right)^+ \boldsymbol{H} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)'$$

Sie ist invariant unter O_d, nicht aber unter $\mathbb{D}_d$, ist also für Messwiederholungen und nicht für multivariate Daten brauchbar.

Diese Statistik hat allerdings gravierende Nachteile. Zum einen gibt es Alternativenvektoren, für die die Macht gleich bleibt, wie lang der Vektor auch sein mag. Man findet sie durch

$$\left(\boldsymbol{I}_d - (\boldsymbol{HSH}')^+ \boldsymbol{HSH}'\right) \boldsymbol{Z},\ \boldsymbol{Z} \in \mathbb{R}^d,\ \boldsymbol{HZ}, \boldsymbol{Z} \neq \boldsymbol{0}, .$$

Anschaulich erklärt sich das durch die säulenförmige Gestalt mit ellipsoidalen Grundriss, die der Annahmebereich hat.

Zum anderen ist die Verteilung dieser Teststatistik im hochdimensionalen

Fall unbekannt. Man weiß nur, dass

$$\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right) \boldsymbol{H}' \left(\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}'\right)^{+} \boldsymbol{H} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)' \sim \chi^2_{\mathrm{rg}(\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}')}$$

und $\hat{\boldsymbol{\Sigma}}$ ein konsistenter Schätzer für $\boldsymbol{\Sigma}$ ist, aber selbst wenn $\boldsymbol{\Sigma}$ nichtsingulär ist, ist die Konvergenz der Verteilung von W_n gegen diese Chiquadratverteilung so schlecht, dass der Test nur bei niedriger Dimensionalität sein Niveau gut einhält, bei wachsender Dimensionalität aber erst extrem liberal und anschließend extrem konservativ wird. In folgender Abbildung wird das veranschaulicht. Auf der Abszisse ist die Dimensionalität d aufgetragen, auf der Ordinate das simulierte Quantil. Als Kovarianzmatrix wurde $\boldsymbol{I}_d$ gewählt, die Stichprobenumfänge waren $n_1 = n_2 = 20$ und die Anzahl der Simulationsdurchläufe waren jeweils 10000.

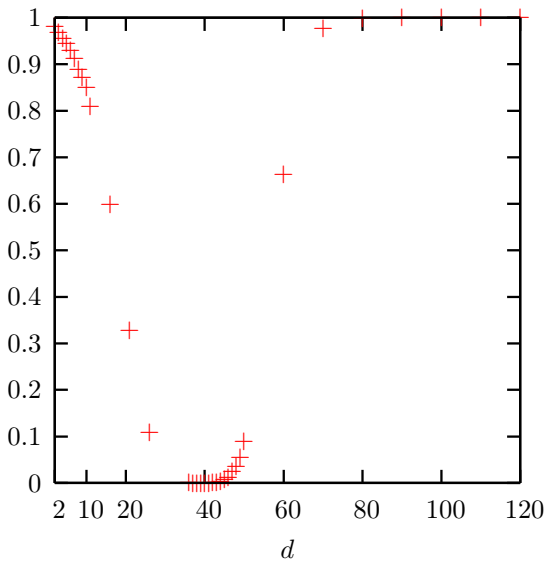

Abbildung 1: Simuliertes 95%-Quantil der Wald-Typ-Statistik

4.2 ANOVA-Typ-Statistik

4.2.1 Zwei Stichproben

Ein dritter Ansatz lässt die invertierte empirische Kovarianzmatrix ganz weg. Das Testkriterium ist somit die quadrierte euklidische Norm der Mittelwertvektoren, also eine quadratische Form, deren exakte Verteilung laut Repräsentationstheorem A.3 gegeben ist durch die Summe gestreckter und unabhängiger Chiquadratverteilungen.

$$\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right) \boldsymbol{T} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)' = \sum_{j=1}^{d} \lambda_j Z_j.$$

λ_i bezeichnet hierbei die Eigenwerte von $\boldsymbol{\Sigma} = n_1^{-1}\boldsymbol{\Sigma}_1 + n_2^{-1}\boldsymbol{\Sigma}_2$ und $Z_j \overset{uiv}{\sim} \chi_1^2$. Um hieraus den kritischen Wert zu berechnen, möchte man aber nicht gestreckte Chiquadratverteilungen falten, damit man die Verteilung erhält, weil das viel zu kompliziert und aufwendig wäre. Stattdessen setzt man Erwartungswert und Varianz des Ausdruckes auf der rechten Seite mit einer um den Faktor g gestreckten Chiquadratverteilung mit f Freiheitsgraden gleich, um so eine approximative Verteilung zu finden.

$$\begin{aligned} E\left(\sum_{j=1}^{d} \lambda_j \chi_1^2\right) &= \sum_{j=1}^{d} \lambda_j = \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma} = gf \\ \mathrm{Var}\left(\sum_{j=1}^{d} \lambda_j \chi_1^2\right) &= 2\sum_{j=1}^{d} \lambda_j^2 = 2\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}\right)^2 = 2\,g^2 f \end{aligned} \tag{5}$$

Man erhält also $f = \frac{\mathrm{Sp}^2 \boldsymbol{T}\boldsymbol{\Sigma}}{\mathrm{Sp}(\boldsymbol{T}\boldsymbol{\Sigma})^2}$ und $g = \frac{\mathrm{Sp}(\boldsymbol{T}\boldsymbol{\Sigma})^2}{\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}}$. Wegen der Idempotenz und Symmetrie von $\boldsymbol{T}$ kann in die Spuren jeweils $\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}$ statt $\boldsymbol{T}\boldsymbol{\Sigma}$ eingesetzt werden, siehe unten Lemma 6.1. Mit dem Zeichen „ $\dot{\sim}$ ” soll abgekürzt werden, dass die linke Seite des Zeichens approximativ die Verteilung auf der rechten Seite hat. Es kann also geschrieben werden:

$$\begin{aligned} &\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right) \boldsymbol{T} \left(\overline{\boldsymbol{Y}_{1\cdot}} - \overline{\boldsymbol{Y}_{2\cdot}}\right)' \doteqdot g\chi_f^2 \\ \Leftrightarrow &\frac{\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right) \boldsymbol{T} \left(\overline{\boldsymbol{Y}_{1\cdot}} - \overline{\boldsymbol{Y}_{2\cdot}}\right)'}{\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}} \doteqdot f^{-1}\chi_f^2 \end{aligned} \tag{6}$$

Als nächstes soll der unbekannte Parameter $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}$ auf der linken Seite durch einen beobachtbaren Term ersetzt werden, nämlich durch die Spur der empirischen Kovarianzmatrix $\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1}\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_1 + \frac{1}{n_2}\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_2$.

Lemma 4.1. *Die Spur der empirischen Kovarianzmatrix einer multivariat normalverteilten Stichprobe ist stochastisch unabhängig von ihrem Mittelwertsvektor.*

Beweis. Wenn man die Beobachtungen der Stichprobe i in einem langen Vektor $\boldsymbol{y}_i$ untereinander schreibt, und $\boldsymbol{T}_{n_i d} = (\boldsymbol{I}_{n_i} \otimes \boldsymbol{T})$ abkürzt, dann ist

$$\boldsymbol{T}_{n_i d}\boldsymbol{y}_i \sim \mathcal{N}\left(\mathbf{1}_{n_i} \otimes \boldsymbol{T}\boldsymbol{\mu}, \boldsymbol{I}_{n_i} \otimes \boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\right). \tag{7}$$

Der Mittelwertsvektor mehrmals untereinandergeschrieben ist

$$\boldsymbol{T}_{n_i d}\overline{\boldsymbol{y}} = n_i^{-1}\left(\boldsymbol{J}_{n_i} \otimes \boldsymbol{I}_d\right)\boldsymbol{T}_{n_i d}\boldsymbol{y}.$$

Die Spur der empirischen Kovarianzmatrix kann man schreiben als eine quadratische Form:

$$\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_i = (n_i - 1)^{-1}\,\boldsymbol{y}_i'\boldsymbol{T}'_{n_i d}\left(\boldsymbol{P}_{n_i} \otimes \boldsymbol{I}_d\right)\boldsymbol{T}_{n_i d}\boldsymbol{y}_i \tag{8}$$

Die Matrizen $(\boldsymbol{J}_{n_i} \otimes \boldsymbol{I}_d)$ und $(\boldsymbol{P}_{n_i} \otimes \boldsymbol{I}_d)$ sind symmetrisch und positiv semidefinit. Weil

$$\begin{aligned} &\left(\boldsymbol{J}_{n_i} \otimes \boldsymbol{I}_d\right)\left(\boldsymbol{I}_{n_i} \otimes \boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\right)\left(\boldsymbol{P}_{n_i} \otimes \boldsymbol{I}_d\right) \\ =&\left(\boldsymbol{J}_{n_i} \otimes \boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\right)\left(\boldsymbol{P}_{n_i} \otimes \boldsymbol{I}_d\right) = \mathbf{0} \end{aligned}$$

ist, ist die quadratische Form $\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_i$ laut dem Satz von Craig und Sakamoto (siehe Satz A.2) unabhängig von $\mathbf{1}_{n_i} \otimes \boldsymbol{T}\overline{\boldsymbol{Y}}_{i\cdot}$. Insbesondere ist sie unabhängig

von einem beliebigen Teil dieses Vektors, namentlich $\overline{\boldsymbol{Y}}_{i\cdot}$. □

Die Spuren der empirischen Kovarianzmatrizen sind also jeweils unabhängig von den Mittelwertvektoren derselben Stichprobe und von den Mittelwertvektoren der anderen Stichprobe sowieso. Wenn jetzt mit derselben Idee wie beim Zähler eine approximative Chiquadratverteilung von der Spur der empirischen Kovarianzmatrix gefunden werden kann, muss der gesamte Bruch approximativ einer F-Verteilung folgen.

Das Repräsentationstheorem A.3 liefert für $n_1^{-1}\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_1 + n_2^{-1}\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_2$:

$$\sum_{i=1}^{2} n_i^{-1}\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_i = \sum_{i=1}^{2} \frac{1}{n_i\left(n_i-1\right)} \boldsymbol{y}_i'\boldsymbol{T}'_{n_i d}\left(\boldsymbol{P}_{n_i}\otimes\boldsymbol{I}_d\right)\boldsymbol{T}_{n_i d}\boldsymbol{y}_i \quad \sim \sum_{i=1}^{2}\sum_{j=1}^{n_i d}\lambda_{j(i)}Z_j \tag{9}$$

Die $\lambda_{j(i)}$ sind die Eigenwerte von $\boldsymbol{W}_i = \frac{1}{n_i(n_i-1)}\left(\boldsymbol{P}_{n_i}\otimes\boldsymbol{I}_d\right)\left(\boldsymbol{I}_n\otimes\boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\right)$ und $Z_j \overset{uiv.}{\sim} \chi_1^2$. Man setzt wiederum die ersten beiden Momente gleich und errechnet wegen der Unabhängigkeit der Chiquadratverteilungen:

$$\begin{aligned}
E\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i d}\lambda_{j(i)}\chi_1^2\right) &= \sum_{i=1}^{2}\tfrac{1}{n_i(n_i-1)}\mathrm{Sp}\left(\left(\boldsymbol{P}_{n_i}\otimes\boldsymbol{I}_d\right)\left(\boldsymbol{I}_n\otimes\boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\right)\right) \\
&= \sum_{i=1}^{2} n_i^{-1}\left(n_i-1\right)^{-1}\mathrm{Sp}\left(\boldsymbol{P}_{n_i}\otimes\boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\right) \\
&= \sum_{i=1}^{2} n_i^{-1}\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_i \qquad = g_0 f_0 \\
\mathrm{Var}\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i d}\lambda_{j(i)}\chi_1^2\right) &= 2\sum_{i=1}^{2}\tfrac{1}{n_i^2(n_i-1)^2}\mathrm{Sp}\left(\left(\boldsymbol{P}_{n_i}\otimes\boldsymbol{I}_d\right)\left(\boldsymbol{I}_n\otimes\boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\right)\right)^2 \\
&= 2\sum_{i=1}^{2} n_i^{-2}\left(n_i-1\right)^{-1}\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 \qquad = 2g_0^2 f_0
\end{aligned}$$

Es ergeben sich folgende Freiheitsgrade und folgende Verteilung:

$$f_0 = \frac{\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}}{\sum_{i=1}^{2} n_i^{-2}\left(n_i-1\right)^{-1}\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}_i}$$

$$g_0 = \frac{\sum_{i=1}^{2} n_i^{-2} \left(n_i - 1\right)^{-1} \mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2}{\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}}$$

$$\frac{\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}}{\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}} \mathrel{\dot{\sim}} f_o^{-1}\chi^2_{f_0}.$$

Zusammen mit (6) kürzt sich $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}$ weg und man erhält approximativ eine F-Verteilung:

$$Q_F = \frac{\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right) \boldsymbol{T} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)'}{\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}} \mathrel{\dot{\sim}} F\left(f, f_0\right) \tag{10}$$

Weil für den Fall $n_1 = n_2$ und $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma\boldsymbol{I}_d$ sich auf diese Weise die bekannte ANOVA-Statistik ergibt, wird eine Statistiken mit solchen Approximationen in der Literatur ([8]) auch ANOVA-Typ-Statistik genannt. In Abschnitt 7 wird untersucht, ob die gefundene Approximation tauglich ist.

4.2.2 Eine Stichprobe

Den Rechenweg kann man auch auf den Fall einer einzigen Stichprobe bestehend aus

$$\boldsymbol{Y}_k \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \, k = 1, \ldots, n \tag{11}$$

übertragen. Die Statistik mit der F-Approximation lautet dann einfach

$$\frac{\overline{\boldsymbol{Y}}_{\cdot}'\boldsymbol{T}\overline{\boldsymbol{Y}}_{\cdot}}{n^{-1}\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}} \mathrel{\dot{\sim}} F\left(f, \left(n-1\right) f\right) \tag{12}$$

mit $f = \frac{\mathrm{Sp}^2\boldsymbol{\Sigma}}{\mathrm{Sp}\boldsymbol{\Sigma}^2}$. Der Einstichprobenfall wird die Simulation in dieser Arbeit vereinfachen, sonst aber nicht weiter thematisiert werden. Für den Freiheitsgrad kann der Schätzer aus [31] eingesetzt werden, der im nächsten Abschnitt vorgestellt wird.

5 Betrachtung bisheriger Arbeiten

5.1 Approximation und Versuche zur Freiheitsgradschätzung

Mit der Idee, die Verteilungsfunktion einer Summe von χ^2-verteilten Zufallsvariablen so durch eine Chiquadratverteilung zu approximieren, dass die Varianz bei der Approximation dieselbe ist, wurde schon von Satterthwaite [26] der Student-t-Test modifiziert, um ungleiche Stichprobenumfänge und ungleiche Varianzen zuzulassen (Behrens-Fischer-Problem). Bei Box [5, 6] wurden Erwartungswert und Varianz einer quadratischen Form mit denen einer gestreckten chiquadratverteilen Zufallsvariable gleichgesetzt, um die approximative Verteilung zu erhalten. Ähnlich der obigen Rechnung wurde dies auf zwei unabhängige quadratische Formen angewendet, und so entstanden approximative Teststatistiken für ein- und mehrfaktorielle Blockmodelle. Die Ergebnisse ähneln denen von (12).

Geisser und Greenhouse [16, 17] übertragen diese Idee auf Split-Plot-Designs mit gleichen Kovarianzmatrizen oder gleichen Stichprobenumfängen. Sie erwähnen in [16] beiläufig auch die Eignung der quadrierten euklidischen Norm für hochdimensionale Daten, da sie nicht erfordert, die empirische Kovarianzmatrix zu invertieren. Für die Freiheitsgrade der F-Verteilung schlagen die Autoren vor, eine Abschätzung nach unten einzusetzen, nämlich $f \geq 1$. Diese Abschätzung gilt auch für den Fall ungleicher Kovarianzmatrizen. Weil die Quantile der F-Verteilung $F_{f,(n-1)f}$ in f für $f \geq 1$ monoton fallend sind, ergibt sich daraus ein konservativer Test. Den naiven Ansatz, die empirische Kovarianzmatrix für die Freiheitsgrade zu verwenden, verwerfen die Autoren. Schon für bekannte Kovarianzmatrizen könne der Rechenaufwand –geschrieben im Jahr 1958– umständlich sein. Der Effekt geschätzter Kovarianzmatrizen auf die Approximatitionsqualität war den Autoren nicht absehbar.

In gleichen Jahr befasste sich Dempster [12, 13] speziell mit hochdimensionalen Tests. Sein Ausgangspunkt ist dieselbe F-Approximation, insbesondere ebenfalls mit dem quadrierten euklidischen Abstand der Mittelwertvektoren

beider Stichproben. Weil er nur zwei Stichproben mit identischen Kovarianzmatrizen betrachtet, hat die Verteilung der Statistik die Freiheitsgrade f und $(n-2)\,f$. Für Freiheitsgrad f gewinnt er Schätzer auf Grundlage der kanonischen Form aus Definition 3.3, indem er heuristisch zwei Gleichungen aus den Längen und Winkeln der Zeilenvektoren von $\boldsymbol{Y}^2$ und $\boldsymbol{Y}^3$ zueinander konstruiert. Beide Gleichungen können nach einem Wert aufgelöst werden, der f schätzen soll. Erwartungstreue und Konsistenz sind bei diesen Schätzern nicht eindeutig, weil sie von der willkürlichen Basisergänzung bei der kanonischen Form und der Wahl einer der beiden gewonnenen Gleichungen abhängen.

Wesentlich später (1996) greifen Bai und Saranadasa [4] die Statistik und das Modell von Dempster auf. Sie benutzen den Term

$$W_2 = \tfrac{(n_1+n_2-2)^2}{(n_1+n_2)(n_1+n_2+1)} \left(\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{S}_{n_1+n_2-2}\right)^2 - \tfrac{1}{n_1+n_2-2}\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{S}_{n_1+n_2-2}\right) \right),$$

der mit der Asymptotik $\frac{d}{n} \to c > 0$, $\frac{n}{n_i} < N_0$ und $n \to \infty$ gegen $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}^2$ konvergiert. Damit wird die quadratische Form $\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)' \boldsymbol{T} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)$ standardisiert, um eine asymptotisch normalverteilte Teststatistik zu erhalten.

Aus W_2 und derselben Asymptotik machen Srivastava und Fujikoshi [29] eine F-Statistik für mehrere Stichproben mit identischen Kovarianzmatrizen, die bei zwei Stichproben die approximative Verteilung

$$\frac{\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)' \boldsymbol{T} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)}{\mathrm{Sp}\boldsymbol{T}\boldsymbol{S}_n} \;\dot{\sim}\; F\left(f_S, (n-2)\,f_S\right), \tag{13}$$

annimmt und den Freiheitsgrad $f_S = \frac{\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}}{\mathrm{Sp}(\boldsymbol{T}\boldsymbol{\Sigma})^2}$ durch $\frac{\mathrm{Sp}^2\boldsymbol{S}_n}{W_2}$ ersetzt, der in ihrer Asymptotik konsistent ist.

Nachteilig an dieser Asymptotik ist, dass die Dimensionalität als genauso flexibel betrachtet wird wie der Stichprobenumfang, denn in der Realität ist es viel einfacher, einen Versuch an zusätzlichen Versuchseinheiten zu wiederholen, als ihn mit einer anderen Anzahl an Messwiederholungen neu zu gestalten. Man wird also nicht die Anzahl der Messungen verändern, damit sich die Statistik besser an das asymptotische Verhalten annähert.

Bei keinem der genannten Autoren wurde der Fall unterschiedlicher Kovarianzmatrizen bei gleichzeitig verschiedenen Stichprobenumfängen gelöst, obwohl die Ähnlichkeit der Box-Approximation für nichtsphärische Kovarianzmatrizen mit der Satterthwaite-Approximation für unbalancierte heteroskedastische Stichproben eine Kombination anbietet.

Problematisch ist weiterhin der Umgang mit unterschiedlichen Invarianzeigenschaften. Geisser und Greenhouse stellen ihr Verfahren zwar von Anfang an in den Split-Plot-Kontext, empfehlen in [17] aber trotzdem diesen Ansatz, falls multivariate Tests, insbesondere für hochdimensionale Daten, nicht mehr anwendbar seien. Dass verschiedene Größen so nicht mit eindeutigen Ergebnissen getestet werden können, bleibt unberücksichtigt. In [29] wird zwar auf die Unmöglichkeit hingewiesen, einen unter GL_d invarianten hochdimensionalen Test zu finden, der Test basierend auf der quadrierten euklidischen Norm wird aber trotzdem wie ein multivariater Test behandelt und sogar mit solchen verglichen, obwohl er eher eine Alternative für Varianzkomponenten-ANOVA als für multivariate Tests ist.

5.2 Dimensionsstabile Lösung

Gezielt für Daten mit Messwiederholungen und mit einer allgemeineren Asymptotik wurden die Freiheitsgrade in den Arbeiten über einen Einstichprobentest von Werner [31] sowie Ahmad, Werner und Brunner [1] konstruiert. Die im nächsten Kapitel konstruierten Schätzer ähneln den Schätzern aus diesen beiden Arbeiten bis auf das für den Zweistichprobentest nötige. Deswegen werden diese Vorarbeiten hier genauer dargestellt.

Modell (18) wird sinngemäß zum Einstichprobenmodell, wenn $\boldsymbol{\mu}$, n und $\boldsymbol{\Sigma}$ Erwartungswertvektor, Stichprobenumfang und Kovarianzmatrix darstellen:

$$\boldsymbol{Y}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ j = 1, \ldots, n \tag{14}$$

Da der Faktor A nicht existiert, gibt es nur noch Hypothesen über die Strukturierung der Messwiederholungen, etwa

$$H_0(B):\ \boldsymbol{P}_d\boldsymbol{\mu} = \mathbf{0}.$$

Bei mehreren Zeitfaktoren sehen die Matrixformulierungen der Hypothesen analog aus, also ohne die Differenzen aber mit denselben Matrizen für $\boldsymbol{T}$ wie auf S. 11. Mit derselben Technik, mit der in (6) die Verteilung von $\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)' \boldsymbol{T} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)$ approximiert wurde, geschah es schon in [31] für den Einstichprobenfall:

$$\frac{\overline{\boldsymbol{Y}}_{\cdot}' \boldsymbol{T} \overline{\boldsymbol{Y}}_{\cdot}}{n^{-1}\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}} \doteqdot f_1^{-1}\chi^2_{f_1},\ f_1 = \frac{\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}\right)}{\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}\right)^2}$$

Der unbekannte Parameter $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}$ im Nenner wurde allerdings nicht durch die Multiplikation einer Box-Approximation von $\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}$ eliminiert, sondern er wurde durch einen anderen Schätzer $B_0 = n^{-1}\sum_{k=1}^{n} \boldsymbol{Y}_k \boldsymbol{T} \boldsymbol{Y}_k$ ersetzt, der konsistent und unter der Nullhypothese erwartungstreu ist. Anschließend wurden Erwartungswert und Varianz des gesamten Bruches $\frac{\overline{\boldsymbol{Y}}_{\cdot}' \boldsymbol{T} \overline{\boldsymbol{Y}}_{\cdot}}{n^{-1}B_0}$ mit folgender Taylor-Approximation berechnet, die für Zufallsvariablen X und Y gilt :

$$E\left(\frac{X}{Y}\right) \doteq \frac{EX}{EY}\left(1 + \frac{\mathrm{Var}\left(Y\right)}{E^2Y} - \frac{\mathrm{Cov}\left(X,Y\right)}{EXEY}\right) \tag{15}$$

$$\mathrm{Var}\left(\frac{X}{Y}\right) \doteq \frac{E^2X}{E^2Y}\left(\frac{\mathrm{Var}\left(X\right)}{E^2X} + \frac{\mathrm{Var}\left(Y\right)}{E^2Y} - 2\frac{\mathrm{Cov}\left(X,Y\right)}{EXEY}\right) \tag{16}$$

Danach werden Erwartungswert und Varianz einer F-Verteilung mit diesen Annäherungen gleichgesetzt, um eine F-Approximation zu gewinnen. Der erste Freiheitsgrad f_I muss dabei $\frac{f_I-2}{f_I} = 1$ erfüllen, was nur asymptotisch für $f_I \to \infty$ möglich ist. Da die Dichte einer $F_{f_I,f_{II}}$-Verteilung für $f_I \to \infty$ gegen die Dichte einer um den Faktor f_{II}^{-1} gestreckten $\chi^2_{f_{II}}$-verteilten Zufallsvariable konvergiert, ist also

$$\frac{\overline{\boldsymbol{Y}}_{\cdot}' \boldsymbol{T} \overline{\boldsymbol{Y}}_{\cdot}}{n^{-1}B_0} \doteqdot f_{II}^{-1}\chi^2_{f_{II}}$$

Man kann auch –wie in [1] geschehen– die Approximation $E\left(\frac{X}{Y}\right) \doteq \frac{E(X)}{E(Y)}$ benutzen, ohne dass sich am Ergebnis etwas ändert.

Das entscheidende Problem ist aber die adäquate Schätzung des Freiheitsgrades. Das Einsetzen der empirischen Kovarianzmatrix anstelle von $\boldsymbol{\Sigma}$

wurde verworfen. Für den hochdimensionalen Fall nützt die Konsistenz dieses Schätzers nichts, denn der asymptotische Fall hinreichend großer Stichprobenumfänge ist nie hochdimensional. In Simulationen der zitierten Arbeit wurde gezeigt, dass $\mathrm{Sp}\left(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\right)^2$ für $n < d$ den Parameter $\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}\right)^2$ stark überschätzt. Bei einer Compound-Symmetry-Struktur konnte sogar bewiesen werden, dass der Quotient $\frac{\mathrm{Sp}\left(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\right)^2}{\mathrm{Sp}(\boldsymbol{T}\boldsymbol{\Sigma})^2}$ bei festem Stichprobenumfang mit d schrumpft, die relative Verzerrung also immer größer wird. Die Sorge von Geisser und Greenhouse, wie sich ein Test durch das Einsetzen von Schätzern für die Freiheitsgrade verhalten würde, beantwortet sich in diesem Layout also genau wie bei ihrem Vorschlag für den Split-Plot-Fall, dem Einsetzen einer Untergrenze.

Sinnvoller als Konsistenz ist also stochastische Konvergenz, die für alle $d \in \mathbb{N}$ gleichmäßig gilt. Wegen der Tschebyschjow-Ungleichung wird die Konvergenz schon erreicht, wenn der Schätzer asymptotisch erwartungstreu ist und der quadrierte Variationskoeffizient verschwindet.

Lemma 5.1. *$\hat{\theta}_{n,d}$ konvergiert für alle $d \in \mathbb{N}$ gleichmäßig stochastisch gegen θ_d, wenn $\left|E\left(\frac{\hat{\theta}_{n,d}}{\theta_d}\right) - 1\right| \overset{n\to\infty}{\longrightarrow} 0$ und $Var\left(\frac{\hat{\theta}_{n,d}}{\theta_d}\right) \overset{n\to\infty}{\longrightarrow} 0$.*

Beweis. Der Beweis ist in [31] zu finden. □

Auf diese Weise kann das entscheidende Kriterium formuliert werden:

Definition 5.2. Der skalare, erwartungstreue Schätzer $\hat{\theta}_{n,d}$ heißt *dimensionsstabil*, wenn eine Folge z_n existiert, so dass

$$\theta_d^{-2}\mathrm{Var}\left(\hat{\theta}_{n,d}\right) \leq z_n$$

für alle d gilt.

Intuitiv gesagt sichert die Dimensionsstabilität, dass Erwartungstreue und Konsistenz des Schätzers mit wachsender Dimensionalität nicht unbegrenzt schlecht werden können. Konsistenz ergibt sich bei Dimensionsstabilität, wenn z_n eine Nullfolge ist.

Der Vorteil dieser Definition gegenüber der Asymptotik aus [4, 29] ist, dass man in der praktischen Anwendung eines Tests mit Schätzern, die diese Eigenschaft erfüllen, nicht darüber nachdenken muss, ob die Anzahl der

Messwiederholungen zu groß oder zu klein sein könnte, um das nominelle Niveau befriedigend gut zu erreichen.

Dimensionsstabile Spurschätzer wurden in [1, 31] aus quadratischen Formen, ihrem Produkt und aus Bilinearformen der Beobachtungsvektoren konstruiert, die unter der Nullhypothese $H_0 : \boldsymbol{T}\boldsymbol{\mu} = \mathbf{0}$ gerade die zu schätzenden Parameter $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}$, $\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}$ und $\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}\right)^2$ als Erwartungswert haben (siehe unten Lemma 6.3 und notiere $k \neq s$):

$$\begin{aligned} E\left(\boldsymbol{Y}'_k\boldsymbol{T}\boldsymbol{Y}_k\right) &= \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma} \\ E\left(\boldsymbol{Y}'_k\boldsymbol{T}\boldsymbol{Y}_k\boldsymbol{Y}'_s\boldsymbol{T}\boldsymbol{Y}_s\right) &= \mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma} \\ E\left(\boldsymbol{Y}'_k\boldsymbol{T}\boldsymbol{Y}_s\right)^2 &= \mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}\right)^2 \end{aligned}$$

Die Terme

$$\begin{aligned} B_0^{(W)} &= n^{-1}\sum_{k=1}^{n}\boldsymbol{Y}'_k\boldsymbol{T}\boldsymbol{Y}_k \\ B_1^{(W)} &= n^{-1}\left(n-1\right)^{-1}\sum_{k=1}^{n}\boldsymbol{Y}'_k\boldsymbol{T}\boldsymbol{Y}_k\boldsymbol{Y}'_s\boldsymbol{T}\boldsymbol{Y}_s \\ \text{bzw. } B_2^{(W)} &= n^{-1}\left(n-1\right)^{-1}\sum_{k=1}^{n}\left(\boldsymbol{Y}'_k\boldsymbol{T}\boldsymbol{Y}_s\right)^2 \end{aligned}$$

erweisen sich als konsistente Schätzer für die jeweiligen Spuren und sind darüber hinaus dimensionsstabil. Mit der Taylorapproximation (15) wird der Freiheitsgradschätzer $f_W = \frac{nB_1^{(W)}}{(n-1)B_2^{(W)}}$ approximativ erwartungstreu. Es wurde in Simulationen gezeigt, dass die Statistik

Definition 5.3.

$$C_W = \frac{\overline{\boldsymbol{Y}}'_{\cdot}\boldsymbol{T}\overline{\boldsymbol{Y}}_{\cdot}}{n^{-1}B_0^W} \mathrel{\dot{\sim}} f_W^{-1}\chi^2_{f_W} \tag{17}$$

mit diesem Freiheitsgradschätzer das Niveau sehr gut einhält.

Auf den Zweistichprobenfall mit entweder gleichen Stichprobenumfängen oder gleichen Kovarianzmatrizen wurde diese Idee in der Arbeit von Ahmad [2] übertragen, indem die Zentrierung $\boldsymbol{T}E\left(\boldsymbol{Y}_{1k} - \boldsymbol{Y}_{2l}\right) \overset{H_0}{=} \mathbf{0}$ ausgenutzt wird.

Entsprechend der Differenz der Beobachtungen können Schätzer für die Terme $\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1 + \boldsymbol{T}\boldsymbol{\Sigma}_2\right)$, $\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1 + \boldsymbol{T}\boldsymbol{\Sigma}_2\right)$ und $\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1 + \boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2$ mit derselben Idee und denselben Eigenschaften gefunden werden wie schon bei $B_0^{(W)}$, $B_1^{(W)}$ und $B_2^{(W)}$. Beim Gebrauch der Abkürzungen

$$\begin{aligned} A_{k,l} &= \left(\boldsymbol{Y}_{1k} - \boldsymbol{Y}_{2l}\right)' \boldsymbol{T} \left(\boldsymbol{Y}_{1k} - \boldsymbol{Y}_{2l}\right) \\ \text{und } A_{k,l;r,s} &= \left(\boldsymbol{Y}_{1k} - \boldsymbol{Y}_{2l}\right)' \boldsymbol{T} \left(\boldsymbol{Y}_{1r} - \boldsymbol{Y}_{2s}\right) \end{aligned}$$

lautet die Statistik

$$C_A = \frac{2\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)' \boldsymbol{T} \left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)}{n B_0^{(A)}} \mathrel{\dot{\sim}} \tilde{f}^{-1} \chi^2_{f_A} \tag{18}$$

mit $\tilde{B}_0 = \sum_{k=1}^{n_1} \sum_{l=1}^{n_2} A_{k,l}$ und dem Freiheitsgradschätzer $\tilde{f} = \frac{\tilde{B}_1}{\tilde{B}_2}$ bestehend aus $\tilde{B}_1 = n_1^{-1} n_2^{-1} \left(n_1 - 1\right)^{-1} \left(n_2 - 1\right)^{-1} \sum_{k \neq r}^{n_1} \sum_{l \neq s}^{n_2} A_{k,l} A_{r,s}$ und $\tilde{B}_2 = n_1^{-1} n_2^{-1} \left(n_1 - 1\right)^{-1} \left(n_2 - 1\right)^{-1} \sum_{k \neq r}^{n_1} \sum_{l \neq s}^{n_2} A_{k,l;r,s}^2$. Der große Nachteil ist, dass nur gleiche Stichprobenumfänge oder gleiche Kovarianzmatrizen zugelassen werden können.

Im folgenden Abschnitt sollen ähnliche Spurschätzer ausgearbeitet werden, die diese Beschränkung nicht haben.

6 Spur- und Freiheitsgradschätzer

6.1 Grundsätzliche Lemmata

Lemma 6.1. *(Spur einer Matrix) Sei* $\boldsymbol{C}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$, $\boldsymbol{Q} \in O_d$

1. $\mathrm{Sp}\left(\boldsymbol{BC}\right) = \mathrm{Sp}\left(\boldsymbol{CB}\right)$
2. *Ist* $\boldsymbol{B} = \boldsymbol{B}'$, *dann gilt* $\mathrm{Sp}\boldsymbol{QBQ}' = \mathrm{Sp}\boldsymbol{B}$
3. *(Spurungleichung) Ist* $\boldsymbol{B}$ *außerdem positiv semidefinit, dann gilt*

$$\mathrm{Sp}\left(\boldsymbol{B}^2\right) \leq \mathrm{Sp}^2\left(\boldsymbol{B}\right).$$

Beweis.

1. $\mathrm{Sp}\left(\boldsymbol{BC}\right) = \sum_{l,l'=1}^{d} \boldsymbol{B}_{l'l}\boldsymbol{C}_{ll'} = \mathrm{Sp}\left(\boldsymbol{CB}\right)$
2. $\mathrm{Sp}\left(\boldsymbol{QBQ}'\right) = \mathrm{Sp}\left(\left(\boldsymbol{QB}\right)\boldsymbol{Q}'\right) = \mathrm{Sp}\left(\boldsymbol{Q}'\boldsymbol{QB}\right) = \mathrm{Sp}\left(\boldsymbol{B}\right)$
3. Wiedergegeben wird der Beweis aus [32], S. 166. Mit einer orthogonalen Matrix $\boldsymbol{Q}$, die $\boldsymbol{B}$ diagonalisiert, gilt

$$\mathrm{Sp}\left(\boldsymbol{B}\right)^2 = \mathrm{Sp}\left(\boldsymbol{QBQ}'\right)^2 = \sum_{j=1}^{d} \lambda_j^2 \leq \left(\sum_{j=1}^{d} \lambda_j\right)^2 = \mathrm{Sp}^2\left(\boldsymbol{B}\right)$$

□

Durch dieses Lemma wird deutlich, dass die ANOVA-Typ-Statistik mit durch Skalarprodukte geschätzten Freiheitsgraden invariant unter orthogonalen Transformationen sein kann. Außerdem lässt sich die Verzerrung von $\mathrm{Sp}\left(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\right)^2$ anschaulich erklären. Da es $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in O_d$ gibt, die $\boldsymbol{T\Sigma T}$ und $\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\boldsymbol{T}$ diagonalisieren, braucht man jeweils nur die Eigenwerte λ_j bzw. $\hat{\lambda}_j$ zu betrachten. Es ist stets

$$E\left(\sum_{j=1}^{n}\hat{\lambda}_j\right) = E\mathrm{Sp}\left(\boldsymbol{Q}_2\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\boldsymbol{T}\boldsymbol{Q}_2'\right) = E\mathrm{Sp}\left(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\boldsymbol{T}\right) = E\mathrm{Sp}\left(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\right)$$
$$= \mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}\right) = \mathrm{Sp}\left(\boldsymbol{Q}_1\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}\boldsymbol{Q}_1'\right) = \sum_{j=1}^{d}\lambda_j \tag{19}$$

und $\mathrm{Sp}\left(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\right)^2 = \sum_{j=1}^{d}\hat{\lambda}_j^2$. Laut Satz A.1 ist im Hochdimensionalen aber fast sicher $\mathrm{rg}\boldsymbol{T}\hat{\boldsymbol{\Sigma}} = \min\left(n, \mathrm{rg}\boldsymbol{T}\boldsymbol{\Sigma}\right)$. Manche Eigenwerte von $\boldsymbol{T}\hat{\boldsymbol{\Sigma}}$ verschwinden also. Weil (19) immer gilt, müssen also Eigenwerte überschätzt werden. Deswegen wird auch die Quadratsumme $\mathrm{Sp}\left(\boldsymbol{T}\hat{\boldsymbol{\Sigma}}\right)^2$ zu groß.

Lemma 6.2. *(Cauchy-Schwarz'sche Ungleichung für Zufallsvariablen) Seien X, Y zwei Zufallsvariablen, deren zweite Momente existieren. Dann gilt:*

$$Cov(X,Y) \leq \sqrt{Var(X)\ Var(Y)}$$

Beweis. $(X - EX)$ und $(Y - EY)$ in die bekannte Cauchy-Schwarz-Ungleichung eingesetzt erbringen den Beweis. □

Die Hauptlast der Beweise im folgenden Abschnitt liegt auf der Spurungleichung und der Chauchy-Schwarz'schen Ungleichung. Begonnen werden die Beweise aber mit Resultaten über die Momente von quadratischen und Bilinearformen.

Lemma 6.3. *(Momente quadratischer Formen) Seien $\boldsymbol{z}_1 \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{V}_1\right)$ und unabhängig davon $\boldsymbol{z}_2 \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{V}_2\right)$ und $Q_1 = \boldsymbol{z}_1'\boldsymbol{T}\boldsymbol{z}_1$ und $Q_2 = \boldsymbol{z}_2'\boldsymbol{T}\boldsymbol{z}_2$ quadratische Formen, d. h. $\boldsymbol{T}$ symmetrisch. Dann gelten:*

1. $E\left(Q_1\right) = \mathrm{Sp}\boldsymbol{T}\boldsymbol{V}_1$ *(Satz von Lancaster)*
2. $E\left(Q_1^2\right) = 2\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{V}_1\right)^2 + \mathrm{Sp}^2\boldsymbol{T}\boldsymbol{V}_1$
3. $E\left(Q_1^4\right) = 48\,\mathrm{Sp}\boldsymbol{T}\boldsymbol{V}_1^4 + 32\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{V}_1\right)^3\mathrm{Sp}\boldsymbol{T}\boldsymbol{V}_1 + 12\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{V}_1\right)^2\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{V}_1 + 12\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{V}_1\right)^2 + \mathrm{Sp}^4\boldsymbol{T}\boldsymbol{V}_1$

4. $Var(Q_1^2) = 48\,\mathrm{Sp}(\boldsymbol{TV}_1)^4 + 32\,\mathrm{Sp}(\boldsymbol{TV}_1)^3 \cdot \mathrm{Sp}\boldsymbol{TV}_1 + 8\,\mathrm{Sp}(\boldsymbol{TV}_1)^2 \cdot \mathrm{Sp}^2\boldsymbol{TV}_1 + 8\,\mathrm{Sp}^2(\boldsymbol{TV}_1)^2$

5. $Var(Q_1) = 2\,\mathrm{Sp}(\boldsymbol{TV}_1)^2$

6. $Var(Q_1Q_2) = 4\,\mathrm{Sp}(\boldsymbol{TV}_1)^2 \cdot \mathrm{Sp}(\boldsymbol{TV}_2)^2 + 2\,\mathrm{Sp}^2\boldsymbol{TV}_1 \cdot \mathrm{Sp}(\boldsymbol{TV}_2)^2 + 2\,\mathrm{Sp}(\boldsymbol{TV}_1)^2 \cdot \mathrm{Sp}^2\boldsymbol{TV}_2$

Beweis. Stillschweigend wurde schon in (5) durch das Repräsentationstheorem der Erwartungswert und die Varianz einer quadratischen Form in normalverteilten Zufallsvariablen ausgerechnet. Mit der induktiven Formel

$$E\left(Q_1^0\right) = \eta(0) \text{ mit } \eta(m) = 2^m m!\mathrm{Sp}(\boldsymbol{TV}_1)^{m+1}$$
$$E(Q_1^r) = \sum_{\iota=1}^{r-1} \binom{r-1}{\iota} E(Q_1^\iota)\,\eta(r-1-\iota)$$

aus Mathai und Provost [21], S. 55, können $E(Q_1^2)$ und $\mathrm{Var}(Q_1)$ eleganter berechnet werden. Diese Resultate setzt man in $\mathrm{Var}(Q_1Q_2) = E(Q_1Q_2)^2 - E(Q_1^2)\,E(Q_2^2)$ ein, und der Beweis ist komplett. □

Lemma 6.4. *(Momente von Bilinearformen)*

Seien $\boldsymbol{z}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_1)$ und $\boldsymbol{z}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_2)$ unabhängige Zufallsvariablen und $\boldsymbol{T}$ eine symmetrische, idempotente Matrix. Dann gelten für die Bilinearform $Q = \boldsymbol{z}_1'\boldsymbol{T}\boldsymbol{z}_2$:

1. $EQ = 0$

2. $EQ^2 = \mathrm{Sp}\boldsymbol{TV}_1\boldsymbol{TV}_2$

3. $EQ^4 = 6\,\mathrm{Sp}(\boldsymbol{TV}_1\boldsymbol{TV}_2)^2 + 3\,\mathrm{Sp}^2(\boldsymbol{TV}_1\boldsymbol{TV}_2)$

4. $Var(Q^2) = 6\,\mathrm{Sp}(\boldsymbol{TV}_1\boldsymbol{TV}_2)^2 + 2\mathrm{Sp}^2(\boldsymbol{TV}_1\boldsymbol{TV}_2)$

Beweis. Mit $\boldsymbol{T}_T = \frac{1}{2}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{T} \\ \boldsymbol{T} & \boldsymbol{0} \end{pmatrix}$ und $\boldsymbol{z}' = \begin{pmatrix} \boldsymbol{z}_1' & \boldsymbol{z}_2' \end{pmatrix} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_1 \oplus \boldsymbol{V}_2)$ ist $Q = \boldsymbol{z}'\boldsymbol{T}_T\boldsymbol{z}$ eine quadratische Form. Den Rest besorgen die korrespondierenden Nummern aus Lemma 6.3. □

6.2 Spurschätzer

Für das Durchzählen der Versuchtseinheiten werden die Indizes k, l, s und t benutzt, gegebenenfalls mit „Verzierungen", also k', k'' usw.. Die Indizes k, l, s und t mit derselben Verzierung seien fortan als verschieden zu betrachten, bei verschiedenen Verzierungen sind gleiche Werte möglich. Es kann also $l' = l$ sein. Definiere folgende quadratische bzw. Bilinearformen:

$$\begin{aligned} A^{(i)}_{klst} &= \left(\boldsymbol{Y}_{k(i)} - \boldsymbol{Y}_{l(i)}\right)' \boldsymbol{T} \left(\boldsymbol{Y}_{s(j)} - \boldsymbol{Y}_{t(i)}\right) \\ A^{(i)}_{kl} &= \left(\boldsymbol{Y}_{k(i)} - \boldsymbol{Y}_{l(i)}\right)' \boldsymbol{T} \left(\boldsymbol{Y}_{k(i)} - \boldsymbol{Y}_{l(i)}\right) \\ A_{klk'l'} &= \left(\boldsymbol{Y}_{k(1)} - \boldsymbol{Y}_{l(1)}\right)' \boldsymbol{T} \left(\boldsymbol{Y}_{k'(2)} - \boldsymbol{Y}_{l'(2)}\right) \end{aligned}$$

Daraus werden erwartungstreue Schätzer für die Terme $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_2$, $\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)$, $\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2$ und $\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)$ als Bausteine für die Freiheitsgrade zusammengesetzt. Anders als in den angeführten Vorarbeiten ([1, 2, 31]) wird die Zentrierung also nicht durch die Hypothese sondern durch die identische Verteilung in derselben Stichprobe erreicht. Der Nachteil ist, dass Mindeststichprobenumfänge von $n_i \geq 4$ vorausgesetzt werden.

Theorem 6.5. *(Erwartungstreue) Es gilt:*

1. $E\left(A^{(i)}_{kl} A^{(i)}_{st}\right) = 4\,\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}_i$
2. $E\left(A^{(1)}_{kl} A^{(2)}_{k'l'}\right) = 4\,\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1 \cdot \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_2$
3. $E\left(A^{(i)}_{klst}\right)^2 = 4\,\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_i^2$
4. $E\left(A^2_{klj'k'}\right) = 4\,\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2$

Beweis. Wegen $\left(\boldsymbol{Y}_{k(i)} - \boldsymbol{Y}_{l(i)}\right) \sim \mathcal{N}\left(\boldsymbol{0}, 2\,\boldsymbol{\Sigma}_i\right)$ und der Unabhängigkeit der beiden quadratischen Formen erhält man aus Lemma 6.3.1

$$\begin{aligned} E\left(A^{(i)}_{kl} A^{(i)}_{st}\right) &= E\left(A^{(i)}_{kl}\right) E\left(A^{(i)}_{st}\right) = 4\,\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}_i \\ \text{und } E\left(A^{(1)}_{kl} A^{(2)}_{k'l'}\right) &= E\left(A^{(1)}_{kl}\right) E\left(A^{(2)}_{k'l'}\right) = 4\,\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_2. \end{aligned}$$

Die letzten beiden Behauptungen beweist man entsprechend mit Lemma 6.4.2. □

Unabhängig von der Dimensionalität können $\mathrm{Sp}^2\boldsymbol{\Sigma}_i$, $\mathrm{Sp}\boldsymbol{\Sigma}_i^2$, $\mathrm{Sp}\boldsymbol{\Sigma}_1 \cdot \mathrm{Sp}\boldsymbol{\Sigma}_2$ bzw. $\mathrm{Sp}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2$ also durch folgende Schätzer erwartungstreu geschätzt werden:

$$B_1^{(i)} = \left(4n_i\left(n_i-1\right)\left(n_i-2\right)\left(n_i-3\right)\right)^{-1} \sum_{k,l,s,t=1}^{n_i} A_{kl}^{(i)} A_{st}^{(i)}$$

$$B_2^{(i)} = \left(4n_i\left(n_i-1\right)\left(n_i-2\right)\left(n_i-3\right)\right)^{-1} \sum_{k,l,s,t=1}^{n_i} \left(A_{klst}^{(i)}\right)^2$$

$$C_1 = \left(4n_1\left(n_1-1\right)n_2\left(n_2-1\right)\right)^{-1} \sum_{k,l=1}^{n_1}\sum_{k',l'=1}^{n_2} A_{kl}^{(i)} A_{k'l'}^{(i)}$$

$$C_2 = \left(4n_1\left(n_1-1\right)n_2\left(n_2-1\right)\right)^{-1} \sum_{k,l=1}^{n_1}\sum_{k'l'=1}^{n_2} A_{klk'l'}^{2}$$

Man kann ausrechnen, dass $\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_1 \cdot \mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_2$, $\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_1\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_2$ und $\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_i$ identisch zu C_1, C_2 bzw. $n_i^{-1}\left(n_i-1\right)^{-1}\sum_{k=1}^{n_i} A_{kl}^{(i)}$ sind. Dazu sei im Anhang A.2.1 auf (20) und in A.2.2 auf (21) verwiesen.

Um Konsistenz und Dimensionsstabilität zu beweisen, wird die Varianz benötigt.

Lemma 6.6. *1.* $Var\left(A_{kl}^{(i)} A_{st}^{(i)}\right) = 64\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 + 64\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 \cdot \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)$

2. $Var\left(A_{kl}^{(1)} A_{j'k'}^{(2)}\right) = 64\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^2 \mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2 + 32\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^2 \cdot \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right) + 32\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right) \cdot \mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2$

3. $Var\left(A_{klst}^{(i)2}\right) = 96\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^4 + 32\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2$

4. $Var\left(A_{klk'l'}^2\right) = 96\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2 + 32\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)$

Die ersten beiden Nummern beweist man mit Lemma 6.3.6, die anderen beiden mit Lemma 6.4.4.

Theorem 6.7. *Die erwartungstreuen Schätzer*

1. $B_1^{(i)}$ *für* $\mathrm{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}_i$

2. $B_2^{(i)}$ *für* $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_i^2$

3. C_1 *für* $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1 \cdot \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_2$

4. C_2 *für* $\mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2$.

sind in der Asymptotik $n_1, n_2 \to \infty$ *konsistent und dimensionsstabil.*

Beweis. Zu zeigen ist für jeden der Schätzer, dass seine Varianz bei jeder Dimensionalität und Kovarianzmatrix nicht größer als ein Vielfaches vom Quadrat seines Erwartungswertes ist. Die entscheidende Ungleichung ist, dass $\mathrm{Cov}(X,Y) \leq \mathrm{Var}(X)$ falls $\mathrm{Var}(X) = \mathrm{Var}(Y)$. Es werden die Abkürzungen

$$\nu_i = \frac{n_i(n_i-1)(n_i-2)(n_i-3) - (n_i-4)(n_i-5)(n_i-6)(n_i-7)}{16n_i(n_i-1)(n_i-2)(n_i-3)} = o(1)$$

und

$$\nu = \frac{n_1(n_1-1)n_2(n_2-1) - (n_1-2)(n_1-3)(n_2-2)(n_2-3)}{16n_1n_2(n_1-1)(n_2-1)} = o(1)$$

vereinbart.

In der folgenden Ungleichungskette verschwinden die Kovarianzen, wenn alle Indices verschieden sind:

$$\begin{aligned}
\mathrm{Var}B_1^{(i)} &= \tfrac{1}{(4n_i(n_i-1)(n_i-2)(n_i-3))^2} \sum_{k,l,s,t=1}^{n_i} \sum_{k',l',s',t'=1}^{n_i} \mathrm{Cov}\left(A_{kl}^{(i)}A_{st}^{(i)}, A_{k'l'}^{(i)}A_{s't'}^{(i)}\right) \\
&\leq \nu_i \mathrm{Var}\left(A_{kl}^{(i)}A_{rs}^{(i)}\right) \\
&= \nu_i\left(64\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 + 64\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 \cdot \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)\right) \\
&\leq 128\,\nu_i \mathrm{Sp}^4\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)
\end{aligned}$$

In $(n_1-2)(n_1-3)(n_2-2)(n_2-3)$ Fällen verschwinden alle Kovarianzen, weil die Indices alle verschieden sind. In den anderen Fällen kommt die

Cauchy-Schwarz-Umgleichung zum Einsatz. Dies erklärt die erste Ungleichung. Die letzte Ungleichung nutzt aus, dass $\mathrm{Sp}\boldsymbol{A}^2 \leq \mathrm{Sp}^2\boldsymbol{A}$ für jede symmetrische positiv semidefinite Matrix $\boldsymbol{A}$ gilt. Völlig analog gehen die Beweise für die anderen Schätzer.

$$
\begin{aligned}
\mathrm{Var}B_2^{(i)} &= \left(4\,n_i\,(n_i-1)\,(n_i-2)\,(n_i-3)\right)^{-2} \sum_{k,l,s,t}^{n_i} \sum_{k',l',s',t'}^{n_i} \mathrm{Cov}\left(A_{klst}^{(i)2}, A_{k'l's't'}^{(i)2}\right) \\
&\leq \nu_i \mathrm{Var}\left(A_{klst}^{(i)2}\right) \\
&= \nu_i \left(96\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^4 + 32\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2\right) \\
&\leq 128\,\nu_i \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 \\
\mathrm{Var}C_1 &= \tfrac{1}{(4n_1(n_1-1)n_2(n_2-1))^2} \sum_{k,l}^{n_1} \sum_{k',l'}^{n_2} \sum_{k'',l''}^{n_1} \sum_{k''',l'''}^{n_2} \mathrm{Cov}\left(A_{kl}^{(1)} A_{k'l'}^{(2)}, A_{k''l''}^{(1)} A_{k'''l'''}^{(2)}\right) \\
&\leq \nu \mathrm{Var}\left(A_{kl}^{(1)} A_{k'l'}^{(2)}\right) \\
&= \nu \left(64\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^2 \cdot \mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2 + 32\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^2 \cdot \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right)\right. \\
&\quad \left. +32\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right) \cdot \mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2\right) \\
&\leq 128\,\nu \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right) \cdot \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right) \\
\mathrm{Var}C_2 &= \tfrac{1}{(4n_1(n_1-1)n_2(n_2-1))^2} \sum_{k,l}^{n_1} \sum_{k',l'}^{n_2} \sum_{k'',l''}^{n_1} \sum_{k''',l'''}^{n_2} \mathrm{Cov}\left(A_{klk'l'}^2, A_{k''l''k'''l'''}^2\right) \\
&\leq \nu \mathrm{Var}\left(A_{klk'l'}^2\right) \\
&= \nu \left(96\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2 + 32\,\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)\right) \\
&\leq 128\,\nu \mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)
\end{aligned}
$$

Dimensionsstabilität folgt, weil der Quotient aus Varianz und dem Quadrat des jeweiligen Erwartungswertes für alle d beschränkt ist, nämlich durch $128\nu_i$ bzw. 128ν. Konsistenz folgt, da $\nu_i \to 0$ für $n_i \to \infty$ und

$$
\nu = \frac{1}{16} - \frac{1}{16} \underbrace{\frac{(n_1-2)\,(n_1-3)}{n_1\,(n_1-1)}}_{\substack{n_1\to\infty \\ \longrightarrow 1}} \cdot \underbrace{\frac{(n_2-2)\,(n_2-3)}{n_2\,(n_2-1)}}_{\substack{n_2\to\infty \\ \longrightarrow 1}} \overset{n_1,n_2\to\infty}{\longrightarrow} 0.
$$

□

6.3 Zusammensetzen der Summe

Als nächstes ist zu prüfen, ob die offensichtlich erwartungstreuen Schätzer $B_1 = n_1^{-2}B_1^{(1)} + 2n_1^{-1}n_2^{-1}C_1 + n_2^{-1}B_1^{(2)}$ und $B_2 = n_1^{-2}B_2^{(1)} + 2n_1^{-1}n_2^{-1}C_2 + n_2^{-1}B_2^{(2)}$ auch dimensionsstabil und konsistent für $\mathrm{Sp}^2(\boldsymbol{T\Sigma})$ bzw. $\mathrm{Sp}(\boldsymbol{T\Sigma})^2$ sind. Zur Abkürzung sei fortan $i \neq i'$.

Lemma 6.8. *(Abschätzung der Kovarianzen von Termen aus $B_1^{(i)}$ und C_1 sowie $B_2^{(i)}$ und C_2)*

1. $$\left.\begin{array}{l} Cov\left(A_{kl}^{(i)}A_{rs}^{(i)}, A_{kl}^{(i)}A_{r''s''}^{(i')}\right) \\ Cov\left(A_{kl}^{(i)}A_{rs}^{(i)}, A_{kr}^{(i)}A_{r''s''}^{(i')}\right) \\ Cov\left(A_{kl}^{(i)}A_{rs}^{(i)}, A_{kl'}^{(i)}A_{r''s''}^{(i')}\right) \end{array}\right\} \leq 128\,\mathrm{Sp}^3(\boldsymbol{T\Sigma}_i)\cdot \mathrm{Sp}\boldsymbol{T\Sigma}_{i'},\ l \neq l'$$

2. $$Cov\left(A_{klst}^{(1)2}, A_{kls't'}^2\right) \leq 128\,\mathrm{Sp}(\boldsymbol{T\Sigma}_1)^2\cdot \mathrm{Sp}\boldsymbol{T\Sigma}_1\boldsymbol{T\Sigma}_2$$

3. $$Cov\left(A_{klst}^{(1)2}, A_{kl's''t''}^2\right) \leq 128\,\mathrm{Sp}(\boldsymbol{T\Sigma}_1)^2\cdot \mathrm{Sp}\boldsymbol{T\Sigma}_1\boldsymbol{T\Sigma}_2$$

4. $$Cov\left(A_{klst}^{(2)2}, A_{k'l'st}^2\right) \leq 128\,\mathrm{Sp}(\boldsymbol{T\Sigma}_2)^2\cdot \mathrm{Sp}\boldsymbol{T\Sigma}_1\boldsymbol{T\Sigma}_2$$

5. $$Cov\left(A_{klst}^{(2)2}, A_{k'l's''t}^2\right) \leq 128\,\mathrm{Sp}(\boldsymbol{T\Sigma}_2)^2\cdot \mathrm{Sp}\boldsymbol{T\Sigma}_1\boldsymbol{T\Sigma}_2$$

Beweis.

1. Mit der Cauchy-Schwarz'schen Ungleichung, den Varianzen aus Lemma 6.6 und der Spurungleichung erhält man

$$\begin{aligned} \left.\begin{array}{l} \mathrm{Cov}\left(A_{kl}^{(i)}A_{rs}^{(i)}, A_{kl}^{(i)}A_{r''s''}^{(i')}\right) \\ \mathrm{Cov}\left(A_{kl}^{(i)}A_{rs}^{(i)}, A_{kr}^{(i)}A_{r''s''}^{(i')}\right) \\ \mathrm{Cov}\left(A_{kl}^{(i)}A_{rs}^{(i)}, A_{kl'}^{(i)}A_{r''s''}^{(i')}\right) \end{array}\right\} &\leq \left(\mathrm{Var}\left(A_{kl}^{(i)}A_{rs}^{(i)}\right)\mathrm{Var}\left(A_{kl}^{(i)}A_{r''s''}^{(i')}\right)\right)^{\frac{1}{2}} \\ &\leq \left(128\,\mathrm{Sp}^4\boldsymbol{T\Sigma}_i \cdot 128\,\mathrm{Sp}^2\boldsymbol{T\Sigma}_i\mathrm{Sp}^2\boldsymbol{T\Sigma}_{i'}\right)^{\frac{1}{2}} \\ &\leq 128\,\mathrm{Sp}^3\boldsymbol{T\Sigma}_i\mathrm{Sp}\boldsymbol{T\Sigma}_{i'} \end{aligned}$$

2. Ebenfalls mit der Cauchy-Schwarz'schen Ungleichung schätzt man ab:

$$\begin{aligned}
\operatorname{Cov}\left(A^{(1)2}_{klst}, A^{2}_{kls't'}\right) &\leq \left(\operatorname{Var}\left(A^{(1)}_{klst}\right)\operatorname{Var}\left(A_{kls't'}\right)\right)^{\frac{1}{2}} \\
&= \left(96\operatorname{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^4 + 32\operatorname{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^2\right)^{\frac{1}{2}} \\
&\quad \cdot \left(96\operatorname{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2 + 32\operatorname{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^{\frac{1}{2}} \\
&\leq \sqrt{128^2 \cdot \operatorname{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 \cdot \operatorname{Sp}^2\boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\boldsymbol{\Sigma}_{i'}} \\
&= 128\operatorname{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)^2 \cdot \operatorname{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_i\boldsymbol{T}\boldsymbol{\Sigma}_{i'}
\end{aligned}$$

Die letzten drei Behauptungen können mit denselben Umformungen bewiesen werden. □

Theorem 6.9. *B_1 und B_2 sind dimensionsstabil und konsistent.*

Beweis. Der Übersichtlichkeit halber werden die Kovarianzen und Varianzen getrennt abgeschätzt. Diejenigen Kovarianzen der quadratischen Formen aus $B_1^{(1)}$ und C_1, die nicht verschwinden, werden durch die Resultate aus Lemma 6.8 abgeschätzt. Man erinnere sich außerdem, dass $n_1, n_2 \geq 4$ sein muss.

$$\begin{aligned}
\operatorname{Cov}\left(B_1^{(1)}, C_1\right) &= \left(16\, n_1^2\left(n_1-1\right)^2\left(n_1-2\right)\left(n_1-3\right)n_2\left(n_2-1\right)\right)^{-1} \\
&\quad \cdot \sum_{k,l,s,t=1}^{n_1}\sum_{k',l'=1}^{n_1}\sum_{k''l''=1}^{n_2} \operatorname{Cov}\left(A^{(1)}_{kl}A^{(1)}_{st}, A^{(1)}_{k'l'}A^{(2)}_{k''l''}\right) \\
&\leq \tfrac{2n_1-5}{4n_1(n_1-1)} \cdot 128\operatorname{Sp}^3\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right) \cdot \operatorname{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_2 \\
&< 128\operatorname{Sp}^3\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right) \cdot \operatorname{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_2
\end{aligned}$$

Analog:

$$\operatorname{Cov}\left(B_1^{(2)}, C_1\right) \leq \operatorname{Sp}^3\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right) \cdot \operatorname{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1 < 128\operatorname{Sp}^3\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right) \cdot \operatorname{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1.$$

Weiterhin führen die Abschätzungen $\nu_i \leq 1$ und $\nu \leq 1$ zu $\operatorname{Var}\left(B_1^{(i)}\right) \leq 128\operatorname{Sp}^4\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)$ und $\operatorname{Var}\left(C_1\right) \leq 128\operatorname{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2$. Zusammen ergibt sich:

$$\begin{aligned}\mathrm{Var}\left(B_1\right) &= \sum_{i=1}^{2} n_i^{-4}\mathrm{Var}\left(B_1^{(i)}\right) + \sum_{i\neq i'}^{2} n_i^{-3}n_{i'}^{-1}\mathrm{Cov}\left(B_1^{(i)}, C_1\right) + \left(n_1 n_2\right)^{-2}\mathrm{Var}\left(C_1\right) \\ &\leq 128\left(\sum_{i=1}^{2} n_i^{-4}\mathrm{Sp}^4\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right) + \sum_{i=1, i\neq i'}^{2} n_i^{-3}n_{i'}^{-1}\mathrm{Sp}^3\left(\boldsymbol{T}\boldsymbol{\Sigma}_i\right)\cdot \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_{i'}\right. \\ &\quad \left. + n_1^{-2}n_2^{-2}\mathrm{Sp}^2\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)\cdot \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_2\right) \\ &\leq 128\,\mathrm{Sp}^4\left(\boldsymbol{T}\boldsymbol{\Sigma}\right)\end{aligned}$$

Aus der letzten Zeile folgt die Dimensionsstabilität. Die Konsistenz folgt, wenn man bedenkt, dass in der vorletzten Zeile die Vorfaktoren mit n_i, $n_{i'}$ verschwinden und die Spuren fest sind.

Nach demselben Schema kann die Dimensionsstabilität von B_2 untersucht werden. Die Kovarianzen schätzt man ab durch

$$\begin{aligned}\mathrm{Cov}\left(B_2^{(1)}, C_2\right) &= \left(16n_1^2\left(n_1-1\right)^2\left(n_1-2\right)\left(n_1-3\right)n_2\left(n_2-1\right)\right)^{-1} \\ &\quad \cdot \sum_{k,l,s,t=1}^{n_1}\sum_{k',l'=1}^{n_1}\sum_{k''l''=1}^{n_2}\mathrm{Cov}\left(A_{klst}^{(1)2}, A_{k'l'k''l''}^{2}\right) \\ &\leq \tfrac{2n_1-5}{4n_1(n_1-1)}\cdot 128\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^2\cdot \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2 \\ &< 128\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_1\right)^2\cdot \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2\end{aligned}$$

und analog $\mathrm{Cov}\left(B_2^{(2)}, C_2\right) \leq 128\,\mathrm{Sp}\left(\boldsymbol{T}\boldsymbol{\Sigma}_2\right)^2\cdot \mathrm{Sp}\boldsymbol{T}\boldsymbol{\Sigma}_1\boldsymbol{T}\boldsymbol{\Sigma}_2$. Auch hier gelten selbstverständlich $\nu_i \leq 1$ und $\nu \leq 1$, so dass nach der folgenden Ungleichungskette völlig identisch argumentiert werden kann:

$$\mathrm{Var}\left(B_2\right) = \sum_{i=1}^{2} n_i^{-4}\mathrm{Var}\left(B_2^{(i)}\right) + \sum_{i\neq i'}^{2} n_i^{-3}n_{i'}^{-1}\mathrm{Cov}\left(B_2^{(i)}, C_2\right) + n_1^{-2}n_2^{-2}\mathrm{Var}\left(C_2\right)$$

$$\begin{aligned}
&\leq 128 \left(\sum_{i=1}^{2} n_i^{-4} \mathrm{Sp}^2 \left(\boldsymbol{T}\boldsymbol{\Sigma}_i \right)^2 + \sum_{i=1, i \neq i'}^{2} n_i^{-3} n_{i'}^{-1} \mathrm{Sp}^2 \left(\boldsymbol{T}\boldsymbol{\Sigma}_i \right) \mathrm{Sp} \boldsymbol{T}\boldsymbol{\Sigma}_i \boldsymbol{T}\boldsymbol{\Sigma}_{i'} \right. \\
&\quad \left. + n_1^{-2} n_2^{-2} \mathrm{Sp}^2 \left(\boldsymbol{T}\boldsymbol{\Sigma}_1 \right) \cdot \mathrm{Sp} \boldsymbol{T}\boldsymbol{\Sigma}_2 \right) \\
&\leq 128 \, \mathrm{Sp}^2 \left(\boldsymbol{T}\boldsymbol{\Sigma} \right)^2
\end{aligned}$$

□

6.4 Zusammensetzen des Quotienten

Als nächstes sollen B_1 und B_2 zu den Schätzern $\hat{f}_0 = \frac{B_1}{\sum_{i=1}^{2} (n_i - 1)^{-1} n_i^{-2} B_2^{(i)}}$ und $\hat{f} = \frac{B_1}{B_2}$ zusammengesetzt werden.

Theorem 6.10. *$B_2^{(i)}$ ist fast sicher positiv.*

Beweis. Es reicht, nur ein $A_{klrs}^{(i)2}$ zu betrachten. Es ist $\boldsymbol{T} \left(\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il} \right)$ fast sicher nicht orthogonal zu jedem beliebigen, festen Vektor $\boldsymbol{x}$, weil andernfalls

$$P \left(\boldsymbol{x}' \boldsymbol{T} \left(\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il} \right) = 0 \right) > 0$$

wäre, was bei einer Normalverteilung der $\boldsymbol{Y}_{ik}$, $\boldsymbol{Y}_{il}$ mit $\mathrm{Sp}\boldsymbol{\Sigma}_i > 0$ nicht sein kann. Wegen der Unabhängigkeit zu $\boldsymbol{T} \left(\boldsymbol{Y}_{ir} - \boldsymbol{Y}_{is} \right)$ ist

$$\begin{aligned}
&P \left(\left(\boldsymbol{Y}_{ir} - \boldsymbol{Y}_{is} \right)' \boldsymbol{T} \left(\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il} \right) = 0 \right) \\
=&P \left(\boldsymbol{x}' \boldsymbol{T} \left(\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il} \right) = 0 | \boldsymbol{x} = \left(\boldsymbol{Y}_{ir} - \boldsymbol{Y}_{is} \right) \right) \\
=&P \left(\boldsymbol{x}' \boldsymbol{T} \left(\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il} \right) = 0 \right) = 0
\end{aligned}$$

Also muss $A_{klrs}^{(i)2}$ fast sicher positiv sein. □

Damit ist geklärt, dass $\hat{f}$ und $\hat{f}_0$ fast sicher existieren.

Zuletzt ist noch zu zeigen, dass $\hat{f}$ und $\hat{f}_0$ dimensionsstabil und konsistent sind. Die Konsistenz ergibt sich automatisch aus dem Satz von Slutsky. Dimensionsstabilität und Erwartungstreue kann man am einfachsten approximativ mittels (15) und (16) zeigen.

Theorem 6.11. *$\hat{f}$ und $\hat{f}_0$ sind approximativ dimensionsstabil.*

Beweis. Die approximative Erwartungstreue kann man mit der Approximation von Casella und Berger [11], Abschnitt 5.5.4 zeigen. Die Varianz von $\hat{f}$ ist mit der Taylorapproximation (16) und den Abschätzungen aus dem vorherigen Unterabschnitt folgendermaßen beschränkt:

$$\begin{aligned}
\operatorname{Var}\left(\hat{f}\right) &\approx \frac{E^2B_1}{E^2B_2}\left(\frac{\operatorname{Var}(B_1)}{E^2B_1} + \frac{\operatorname{Var}(B_2)}{E^2B_2} - 2\frac{\operatorname{Cov}(B_1,B_2)}{EB_1EB_2}\right) \\
&\leq \frac{E^2B_1}{E^2B_2}\left(\frac{\operatorname{Var}(B_1)}{E^2B_1} + \frac{\operatorname{Var}(B_2)}{E^2B_2} + 2\frac{\sqrt{\operatorname{Var}(B_1)\operatorname{Var}(B_2)}}{EB_1EB_2}\right) \\
&\leq f^2\left(128 + 128 + 2\cdot 128\right)
\end{aligned}$$

Für $\hat{f}_0$ geht die Rechnung analog. □

Damit ist folgende Statistik hergeleitet:

$$F_B = \frac{(\boldsymbol{Y}_{1\cdot} - \boldsymbol{Y}_{2\cdot})'\boldsymbol{T}(\boldsymbol{Y}_{1\cdot} - \boldsymbol{Y}_{2\cdot})}{\operatorname{Sp}\hat{\boldsymbol{\Sigma}}} \mathrel{\dot{\sim}} F_{\hat{f},\hat{f}_0}$$

7 Simulationen

7.1 Eine Stichprobe

In diesem Unterabschnitt soll hauptsächlich die Qualität der Approximationen (6) und (10) sowie der Einfluss des Einsetzens eines Freiheitsgradschätzers auf das Niveau der Statistik untersucht werden. Bei (6) kommt es nur auf die Kovarianzmatrix $n_1^{-1}\boldsymbol{\Sigma}_1 + n_2^{-1}\boldsymbol{\Sigma}_2$ des Vektors $\left(\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}\right)$ an, nicht auf $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, n_1 und n_2 im einzelnen. Es reicht also die Betrachtung einer Einstichprobenversion mit dem Modell (11) aus. Aber auch der Zweistichprobentest bei gleichen Kovarianzmatrizen und Stichprobenumfängen würde sich genauso verhalten wie die Einstichprobenstatistik (12) auf S. 27.

Um die Folgen geschätzter Freiheitsgrade zu untersuchen, wird in der Statistik (12) zum einen der naive Freiheitsgradschätzer mit den empirischen Kovarianzmatrizen und zum anderen der Freiheitsgradschätzer aus [31] eingesetzt. Die Statistik C_W auf S. 33 wird als bereits etablierte Statistik zum Vergleich hinzugefügt.

Möchte man nun die Kovarianzmatrizen finden, für die die Approximationen am schlechtesten sind, so reicht es, nur das Spektrum der Matrizen zu betrachten, denn da die simulierten Statistiken orthogonal invariant sind, kann ohne Einschränkung eine Diagonalgestalt angenommen werden. Wegen der Invarianz unter $(\mathbb{R}, \cdot)$ weiß man außerdem, dass eine Streckung der Kovarianzmatrix um einen positiven Faktor das Simulationsergebnis nicht verändern kann. Eine weitere zulässige Vereinfachung ist es, nur mit regulären Kovarianzmatrizen zu simulieren, denn sind Eigenwerte der Kovarianzmatrix Null, dann haben die Daten in dem zugehörigen Eigenraum gleichzeitig keine Varianz, die sich auf die quadratische Form auswirken könnte. Als Kovarianzmatrizen wurden deswegen $\boldsymbol{I}_d$, eine Compound-Symmetry-Struktur mit $\sigma = \sigma_V = \frac{1}{2}$ sowie eine autoregressive Struktur mit $\rho = 0,5$ und $\rho = 0,9$ gewählt. Damit sind die Fälle gleicher Eigenwerte, eines großen und mehrerer kleiner Eigenwerte sowie langsam bzw. stark ansteigender geordneter Eigenwerte abgedeckt.

In Abbildung 2 sind die simulierten Quantile der vier Statistiken mit

$\boldsymbol{T} = \boldsymbol{P}_d$ versetzt gegen die Dimensionalität aufgetragen. Jeder eingezeichnete Punkt entspricht einem Simulationsergebnis. Als Stichprobenumfänge wurden 10, 20, 40 und 60 gewählt, so dass bei demselben d jede Statistik 16 mal simuliert wurde. Um den Fehler durch die endliche Anzahl der Simulationsläufe $m = 10000$ mit dem Approximationsfehler der Statistik vergleichen zu können wurden gestrichelte Linien eingezeichnet, die ein 99%-Zufallsintervall für das $1-\alpha$-Quantil mit den Grenzen $1-\alpha-u_{0.995}\sqrt{m^{-1}\alpha\,(1-\alpha)}$ und $1-\alpha+u_{0.995}\sqrt{m^{-1}\alpha\,(1-\alpha)}$ angeben. Bei einer exakten Statistik ist also die Wahrscheinlichkeit 99%, dass das Simulationsergebnis zwischen diesen Grenzen liegt.

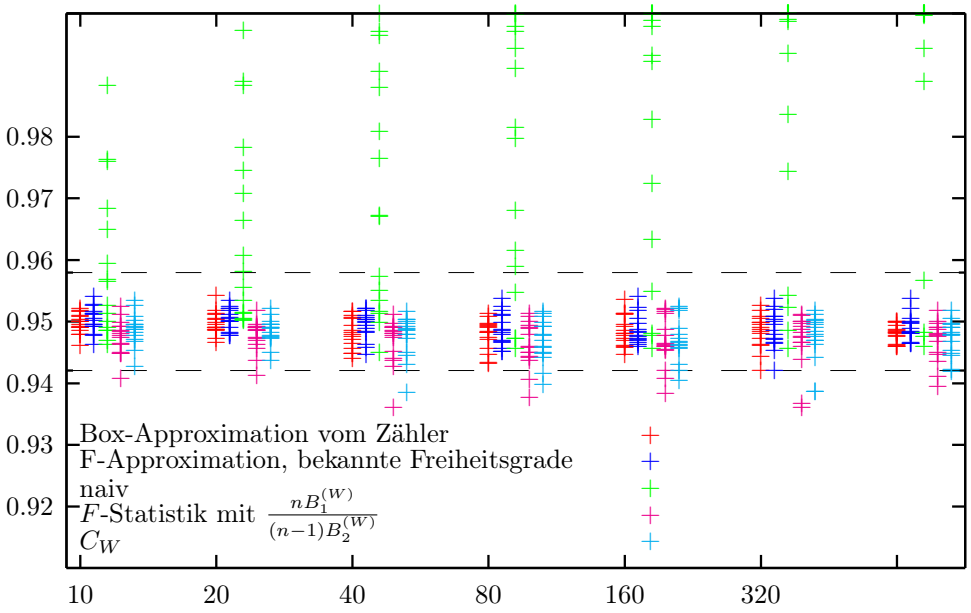

Abbildung 2: Simuliertes 95%-Quantil im Einstichprobenfall

Man sieht, dass sowohl die approximative Verteilung des Zählers als auch des ganzen Bruches im für das Testen üblichen 95%-Bereich sehr genau ist. Die Statistiken mit dimensionsstabilen Schätzern sind erwartungsgemäß etwas schlechter, als wenn die echten Freiheitsgrade eingesetzt worden wären. Zum Testen reicht die Genauigkeit trotzdem. Die empirische Kovarianzmatrix einzusetzen, ist eindeutig schlechter als dimensionsstabile Schätzer zu benutzen, wenn d sehr groß ist.

7.2 Zwei Stichproben

7.2.1 Ungleiche Kovarianzmatrizen und Stichprobenumfänge

Im Fall ungleicher Kovarianzmatrizen und Stichprobenumfänge kann man die Approximation in eine „Box-Komponente" und eine „Satterthwaite-Komponente" aufteilen. Die erste trägt den unterschiedlichen Eigenwerten der Kovarianzmatrix Rechnung, die zweite den unterschiedlichen Kovarianzmatrizen. In der Doppelsumme aus (9) wird das deutlich. Zu untersuchen ist hier also, ob die doppelte Approximation befriedigend ist oder ob die Fehler sich dergestalt addieren, dass die Statistik ihr Niveau zu sehr verfehlt. Deswegen werden für die Simulation die unterschiedlichen Kovarianzmatrizen mit einem Streckungsparameter und unterschiedlichen Stichprobenumfängen kombiniert.

$$\begin{aligned}
\boldsymbol{\Sigma}_1 &\in \left\{ \boldsymbol{I}_d,\ \frac{1}{2}\boldsymbol{I}_d + \frac{1}{2}\boldsymbol{J}_d,\ \left(0,5^{|j-j'|}\right)_{j,j'=1,\ldots,d} \right\} \\
\boldsymbol{\Sigma}_2 &\in \left\{ c\boldsymbol{I}_d,\ \frac{c}{2}\boldsymbol{I}_d + \frac{c}{2}\boldsymbol{J}_d,\ \left(c \cdot 0,5^{|j-j'|}\right)_{j,j'=1,\ldots,d} \middle| \ c \in \{1,2\} \right\} \\
n_1, n_2 &\in \{10,\ 20,\ 40,\ 60\}
\end{aligned}$$

Außerdem ist zu prüfen, wie sich die neuen Spurschätzer verhalten. Die Schätzer $B_1^{(i)}$, $B_2^{(i)}$ und C_1 verhalten sich identisch, auch wenn auf die Kovarianzmatrizen jeweils verschiedene orthogonale Transformationen angewendet werden. Es kommt also auch hier nur auf die Eigenwerte an. Der Term C_2 hingegen kann zusätzlich davon abhängen, wie die Eigenvektoren zu bestimmten Eigenwerten von $\boldsymbol{\Sigma}_1$ und $\boldsymbol{\Sigma}_2$ zueinander liegen, denn allgemein gilt nicht $\mathrm{Sp}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 = \mathrm{Sp}\boldsymbol{Q}'\boldsymbol{\Sigma}_1\boldsymbol{Q}\boldsymbol{\Sigma}_2$ mit $\boldsymbol{Q} \in O_d$. Darum werden mit den ansteigenden Eigenwerten $\lambda_1, \ldots, \lambda_d$ einer autoregressiven Kovarianzmatrix zusätzlich die Matrizen $\boldsymbol{\Sigma}_1 = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_d)$ und $\boldsymbol{\Sigma}_2 = \mathrm{diag}\,(\lambda_d, \ldots, \lambda_1)$ benutzt.

Um einen Eindruck davon zu gewinnen, wie schnell die Konsistenz der Freiheitsgradschätzer die Genauigkeit der Statistik verbessert, sind die Ergebnisse in der Abbildung 4 für ungünstig kleine Stichprobenumfänge –d. h. eine Stichprobe enthält nur 10 oder beide enthalten nur 20 unabhängige

Beobachtungen– und für große Stichprobenumfänge in der Abbildung 3 dargestellt. Die gestrichelten Ränder der Zufallsintervalls sind etwas breiter als in Abbildung 2, weil angesichts der größeren Anzahl an Parameterkombinationen jeweils nur mit 4000 Durchläufen simuliert wurde.

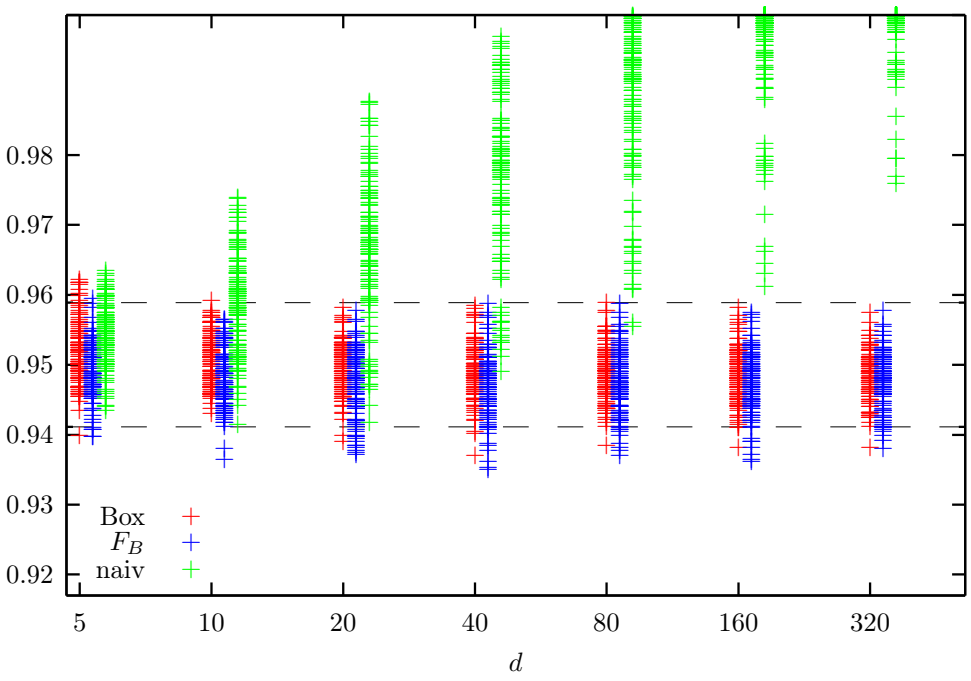

Abbildung 3: Simuliertes 95%-Quantil bei $n_1 \neq n_2$,$\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$,$n_1, n_2 > 800$

Die kombinierte F-Approximation mit bekannten Freiheitsgraden (10) stellt sich beim 95%-Quantil als sehr genau heraus. Die dimensionsstabilen Spurschätzer verschlechtern das Ergebnis kaum spürbar, so dass F_B als geeignete Teststatistik betrachtet werden kann. Das Einsetzen der empirischen Kovarianzmatrizen ergibt wiederum bei wachsender Dimensionalität eine zunehmend konservative Teststatistik.

Die Qualität der F-Approximation ist auch bei kleinen Stichprobenumfängen sehr genau. Wie im Einstichprobenfall wird die naive Schätzung der Spuren bei derselben Dimensionalität noch konservativer, wenn der Stichprobenumfang reduziert wird. Die F_B-Statistik hingegen wird leicht liberal. Der Grund dafür kann in der Verzerrung von $\hat{f}$ und $\hat{f}_0$ bei kleinen Stichprobenumfängen liegen. Wirklich erwartungstreu sind beide Schätzer nur asymptotisch,

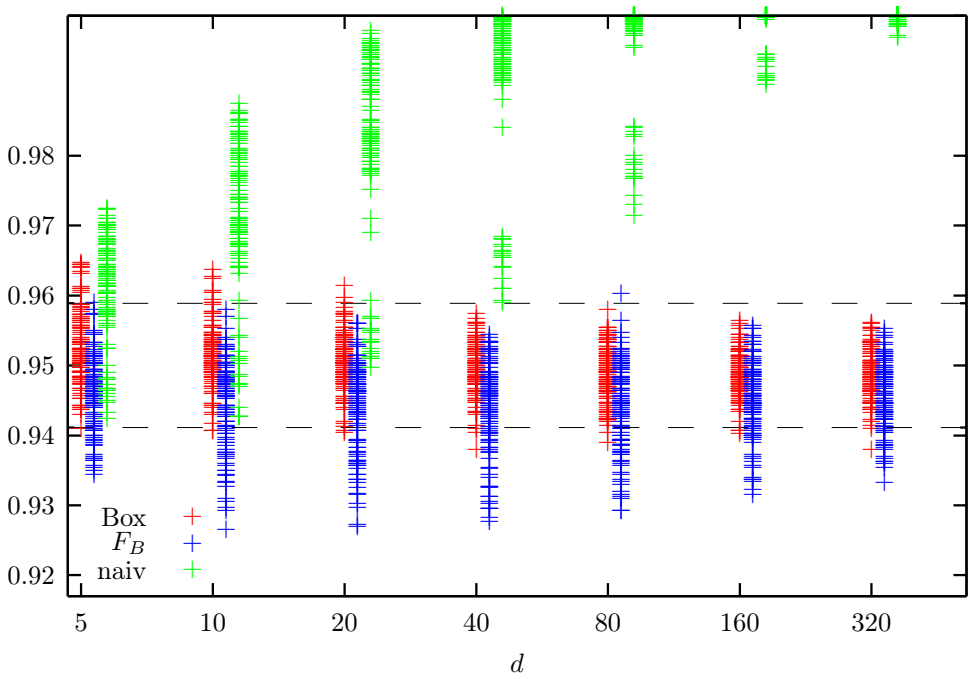

Abbildung 4: Simuliertes 95%-Quantil bei $n_1 \neq n_2$, $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, $n_1 n_2 \leq 800$

denn im Allgemeinen ist der Erwartungswert eines Quotienten nicht gleich dem Quotient der Erwartungswerte. Dieses Problem wurde für den Einstichprobenfall in der Arbeit [31] durch die Taylorapproximation des Erwartungswertes des Quotienten abgefangen. Dabei entstand der Freiheitsgradschätzer $\frac{nB_1^W}{(n-1)B_2^W}$, der –wie im vorigen Abschnitt gesehen– auch bei kleinen Stichprobenumfängen den Test nicht sichtbar verzerrt.

7.2.2 Gleiche Stichprobenumfänge oder Kovarianzmatrizen

Die Qualität der Approximation wird sich in diesen Fällen ähnlich darstellen wie im Einstichprobenfall. Die Approximation bei bekannten Freiheitsgraden muss also nicht nochmal untersucht werden. Stattdessen konkurriert hier die Statistik F_B mit C_A und der Statistik von Geisser und Greenhouse, in die die Untergrenze für den Freiheitsgrad eingesetzt wird. Bei gleichen Kovarianzmatrizen ist außerdem die Statistik (13) anwendbar. Deswegen wird in der Darstellung der Ergebnisse danach getrennt, ob Stichprobenumfänge (Abbildungen 5 und 6) oder Kovarianzmatrizen identisch sind. Simulationsergeb-

nisse, auf die beides zutrifft, sind in beiden Abbildungen eingezeichnet.

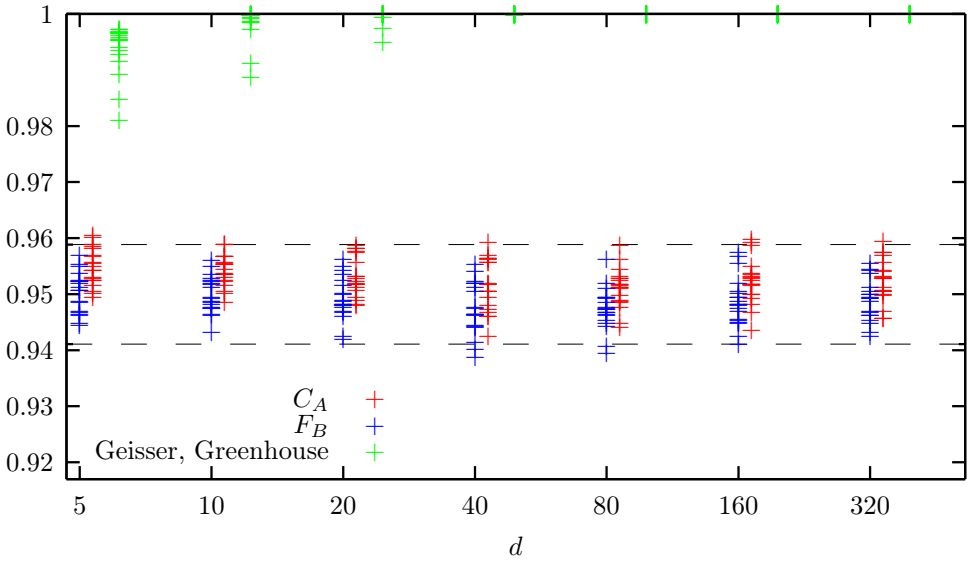

Abbildung 5: Simuliertes 95%-Quantil bei $n_1 = n_2 > 20$

Die Statistik von Geisser und Greenhouse ist extrem konservativ. Dies wird bei steigender Dimensionalität sogar noch schlimmer. Die Statistiken C_A und F_B halten beide ihr Niveau sehr gut ein.

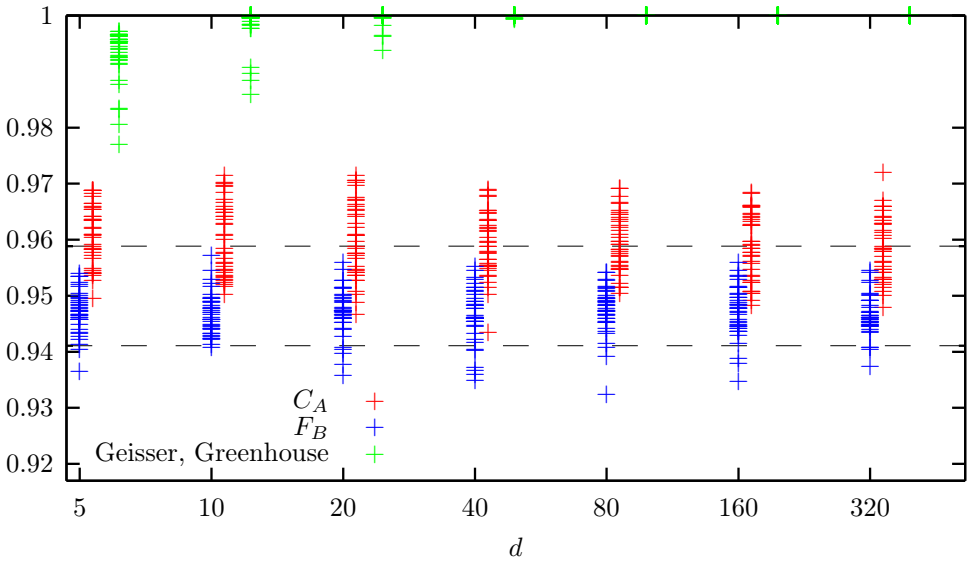

Abbildung 6: Simuliertes 95%-Quantil bei $n_1 = n_2 \leq 20$

Ähnlich wie bei ungleichen Kovarianzmatrizen und Stichprobenumfängen wird F_B etwas liberaler. C_A hingegen wird etwa in dem gleichen Maße konservativer.

In den Abbildungen 7 und 8 werden die Simulationsergebnisse für identische Kovarianzmatrizen präsentiert, die auch für die Statistik (13) ausgerechnet worden sind. Ist eine Teststatistik für gleiche Kovarianzen in dieser Situationen befriedigend genau, muss man für den praktischen Einsatz bedenken, dass bei ungleichen Stichprobenumfängen irrtümlich als identisch angenommene Kovarianzen zu einem extrem liberalen oder konservativen Test führen, abhängig davon, ob in der großen Stichprobe die Spur der Kovarianzmatrix eher groß bzw. eher klein ist. Ein vergleichbares Phänomen ist bei dem Zweistichproben-t-Test bekannt.

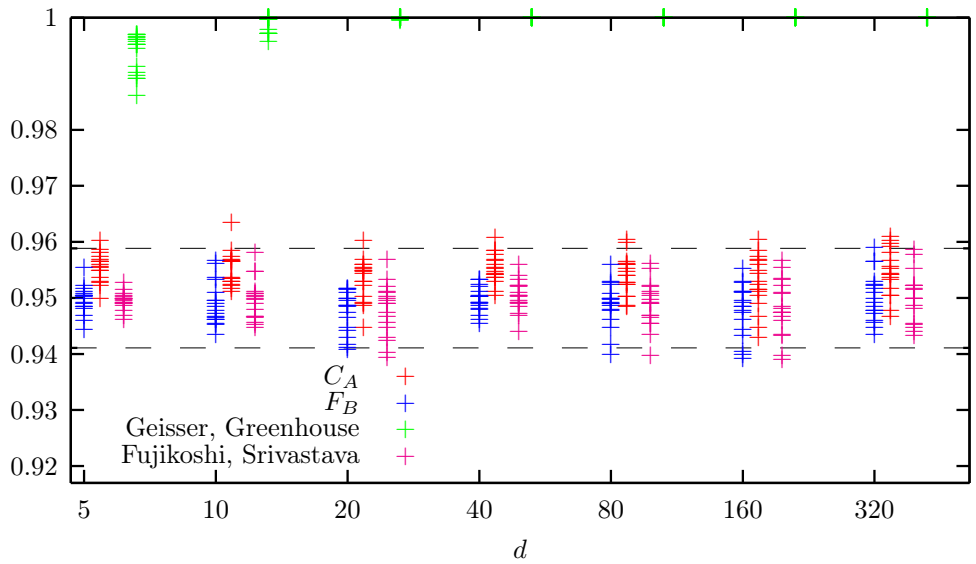

Abbildung 7: Simuliertes 95%-Quantil bei $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ und großen Stichprobenumfängen

Bis auf die Statistik von Geisser und Greenhouse treffen alle Statistiken bei großen Stichprobenumfängen ihr Niveau sehr gut.

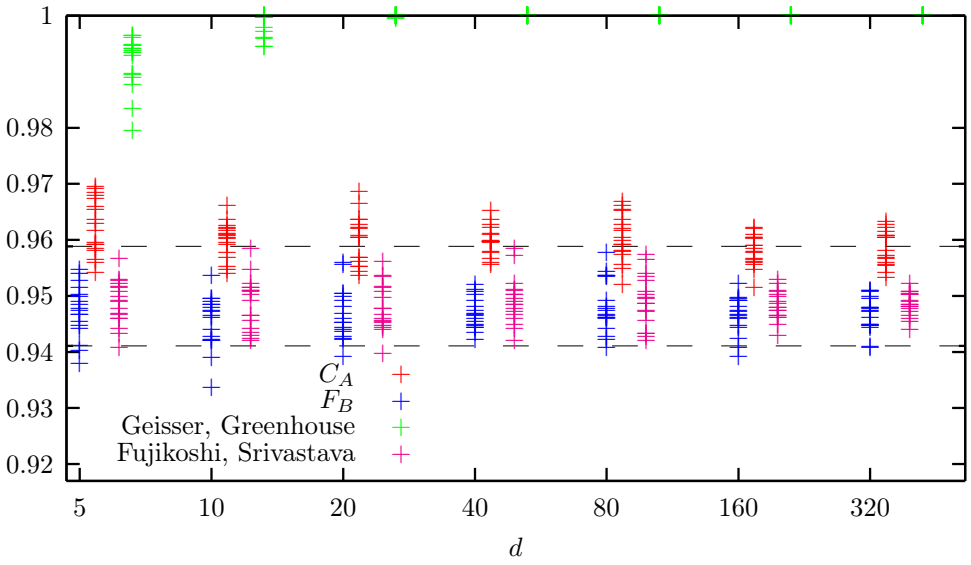

Abbildung 8: Simuliertes 95%-Quantil bei $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ und kleinen Stichprobenumfängen

Bei kleineren Stichprobenumfängen wird C_A wieder etwas konservativ. Der Effekt, dass demgegenüber F_B leicht liberal wird, stellt sich bei gleichen Kovarianzmatrizen nicht ein, die Statistik bleibt genau. Die Statistik (13) ist ähnlich genau. Da sie aber wegen des gepoolten Kovarianzmatrixschätzers nur in dieser Situation so gut sein kann, empfiehlt das für die Praxis nicht diese Statistik. Denn wenn unbekannt ist, dass die Kovarianzmatrizen identisch sind, ist nur F_B gefahrlos anwendbar.

7.3 Güte der Teststatistik

Die Power wird zusätzlich zu n_1,n_2, $\boldsymbol{T}\mathbf{\Sigma}_1$ und $\boldsymbol{T}\mathbf{\Sigma}_2$ durch die Gestalt des Alternativenvektors beeinflusst. Deswegen wäre eine erschöpfende Untersuchung, die alle denkbaren Fälle gut abdeckt, noch umfangreicher. Die Power wird allerdings nur durch die Approximation (10) beeinflusst, nicht durch die Spurschätzer. Da diese nur aus Differenzenvektoren von Beobachtungsvektoren derselben Stichprobe konstruiert werden, ändern sich ihre Schätzungen nicht, wenn ein Alternativenvektor auf die Beobachtungen einer oder beider Stichproben aufaddiert wird.

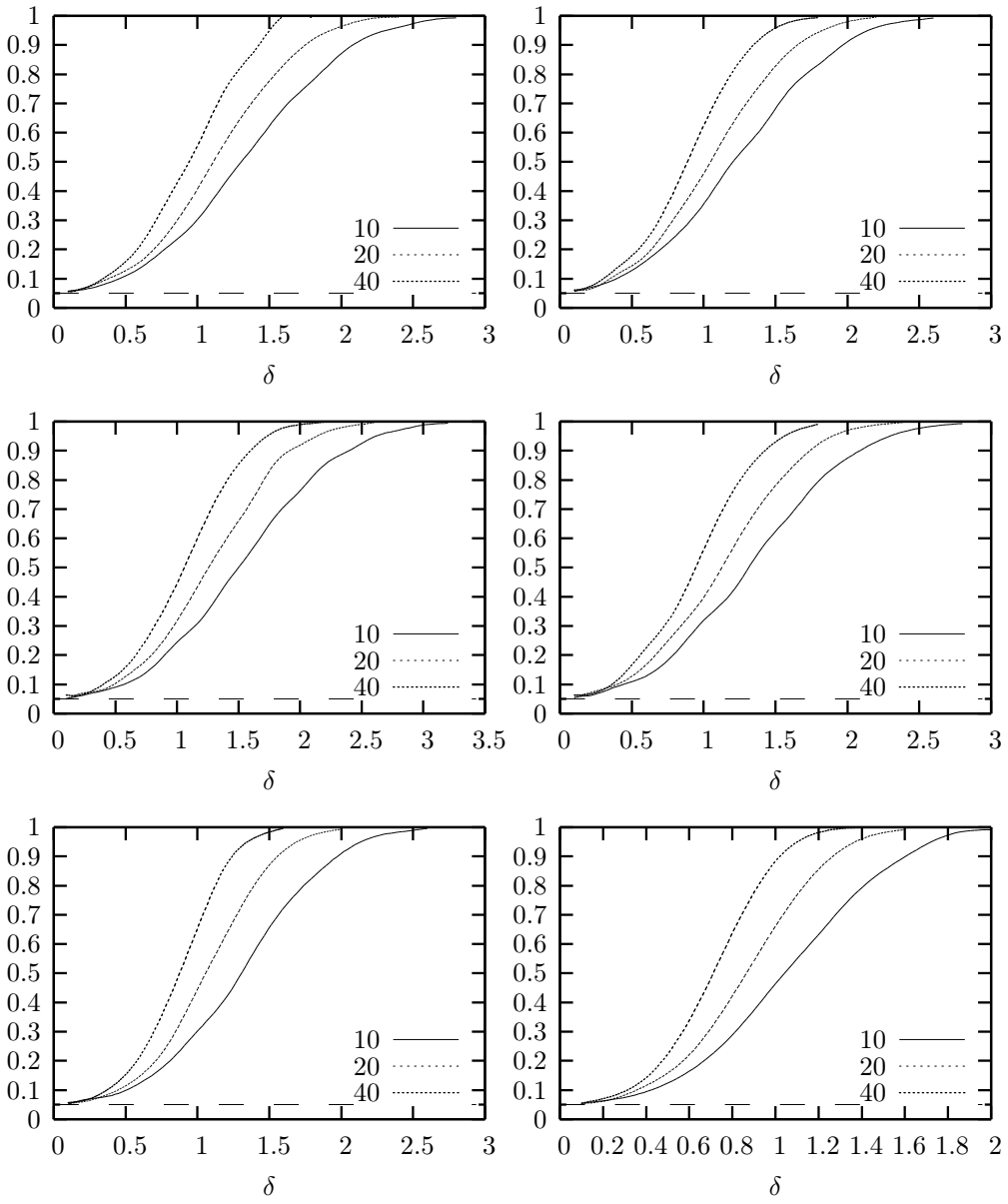

Abbildung 9: Power bei Trendalternative $\beta_j - \beta_{j'} = \delta d^{-1} |j - j'|$

Hier sei zunächst in Abbildung 9 der Fall einer Trendalternative dargestellt, d. h. es ist $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \left(\begin{array}{cccc} 0 & \frac{1}{d} & \ldots & \frac{d-1}{d} \end{array} \right)$ und es wird $H_0(B)$ getestet. Es fanden 5000 Simulationsdurchläufe statt. Die Dimensionalität war $d = 10, 20, 40$, und die Zufallszahlen wurden von oben nach unten mit den Kovarianzmatrizen $\boldsymbol{\Sigma}_1 = \left(0, 5^{|l-l'|}\right)_{l=1,\ldots,d}$ und $\boldsymbol{\Sigma}_2 = \frac{1}{2}\boldsymbol{I}_d + \frac{1}{2}\boldsymbol{J}_d$, $\boldsymbol{\Sigma}_1 = \left(0, 5^{|l-l'|}\right)_{l=1,\ldots,d}$ und $\boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$ sowie zuletzt $\boldsymbol{\Sigma}_1 = \frac{1}{2}\boldsymbol{I}_d + \frac{1}{2}\boldsymbol{J}_d$ und

$\boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$ erzeugt. In der linken Spalte sind $n_1 = n_2 = 10$ und in der rechten $n_1 = 10$ sowie $n_2 = 20$.

Es sind keine negativen Auffälligkeiten der Gütefunktion zu sehen. Die Power verbessert sich in den simulierten Beispielen, wenn die Dimensionalität steigt. Dies liegt aber an der speziellen Form der Alternative und darf nicht den Schluss nahelegen, dass Versuchsaufbauten am besten möglichst viele Messwiederholungen beinhalten sollten.

8 Makros

8.1 Benutzerschnittstelle

In diesem Abschnitt sollen die SAS-Makros dokumentiert werden. In [9] werden die Abkürzungen F1-LD-F1 und F1-LD-F2 für die nichtparametrische Analyse longitudinaler Daten mit einem Whole-Plot und einem bzw. zwei Sub-plot Faktoren verwendet. Um zu betonen, dass hier Makros für hochdimensionale Daten vorliegen, werden sie F1-HD-F1 und F1-HD-F2 genannt. Sie werden mit den SAS-Befehlen

```
%include ‘‘<Pfad zu F1-HD-F1.sas>F1-HD-F1.sas’’
%include ‘‘<Pfad zu F1-HD-F2.sas>F1-HD-F2.sas’’
```

geladen. Die Syntax ist

```
%f1_hd_f1(data=<SAS-Datensatz>, VAR=<gemessene Größe>,
  Subplot=<Zeitfaktor1>,Subject=<Versuchseinheiten>,
  Wholeplot=<Gruppe>);
```

und

```
%f1_hd_f2(data=<SAS-Datensatz>, VAR=<Größe>,
  Subplot1=<Zeitfaktor1>, Subplot2=<Zeitfaktor2>
  Subject=<Versuchseinheiten>, Wholeplot=<Gruppe>);
```

`SAS-Datensatz` ist der Bezeichner für einen SAS-Datensatz, der in demselben Format ist, wie er es für eine entsprechende Analyse mit `proc mixed` sein müsste. Der Datensatz enthält also die Variable `Größe`, die die einzelnen skalaren Messungen enthält, und zu deren richtiger faktorieller Zuordnung die Variablen `Versuchseinheiten`, `Gruppe` und `Zeitfaktor1` sowie eventuell `Zeitfaktor2`. Eine Vorsortierung des Datensatzes ist nicht notwendig.

Anwendungsfehler wie zu geringer Stichprobenumfang, „vergessene” Kombinationen von Faktorstufen, fehlende Varianz in beiden Stichproben oder mehr als zwei Stichproben im Datensatz führen zu Fehlermeldungen oder

Warnungen. Nicht geprüft wird, ob ein Messwert nur ein Punkt ist, das SAS-Symbol für einen fehlenden Wert. Dies müsste durch einen zeitraubenden Schleifendurchlauf überprüft werden.

Die Ausgabe erfolgt in demselben Format wie bei `proc mixed`. Zur Kontrolle werden erst die Namen des Datensatzes und der Variablen mit den Messwerten so ausgegeben, wie sie eingelesen worden waren. Anschließend wird die Codierung der Stichproben und die Anzahl der Faktorstufen der Sub-plot-Faktoren angezeigt. Die Tabelle mit den Testergebnissen listet die Werte der Teststatistik Q_F und von $\hat{f}$ und $\hat{f}_0$ sowie den p-Wert auf. Die Effekte werden in der Ausgabe entsprechend den jeweiligen Variablen im Datensatz benannt. Ein Beispiel für eine solche Tabelle steht in Abschnitt 9.

8.2 Rechnerische Details

Die Berechnung der Spurschätzer B_1 und B_2 wurde auf zwei numerisch nicht identischen Wegen realisiert. Bei der einen Methode wurde ein FORTRAN95-Modul geschrieben, das mit Zählschleifen erst die Differenzen zwischen den Beobachtungsvektoren und anschließend deren Skalarprodukte berechnet. Die andere Methode multipliziert die Skalarprodukte aus und bildet die Differenzen zum Schluss. Dies ist zwar für die numerische Stabilität (siehe [14]) nicht förderlich, trotzdem lieferte im direkten Vergleich die zweite Methode nie sichtbar unterschiedliche Ergebnisse in der Schätzung der Freiheitsgrade. Hinsichtlich Ausführungsgeschwindigkeit sind in der matrixorientierten SAS-Umgebung beide Methoden etwa gleich. Die zweite Methode hat den Vorteil besserer Portierbarkeit, da nicht auf betriebssystemabhängig compilierte Module und Systemaufrufe zugegriffen werden muss. Deswegen wurde sie in den Makros und Simulationen verwendet. Sie wird im Anhang A.2 erklärt.

9 Anwendungsbeispiel

Beispielhaft soll das Makro F1-HD-F2 auf die Daten der Schlafstudie aus [2] angewendet werden. Dort wurde bei jeweils zehn Männern und Frauen während drei Tagen in vierstündigem Abstand die Konzentration eines Enzyms gemessen. In dieser Zeit haben die Probanden in der ersten Nacht normal geschlafen, wurden in der zweiten Nacht am Schlaf gehindert und haben in der dritten Nacht einen Erholungsschlaf gehabt. Die Enzymkonzentrationen jedes Probanden sind in Abbildung 10 aufgeplottet. Die Beobachtungen bei männlichen Probanden sind mit blauen, die der weiblichen mit roten Markierungen versehen.

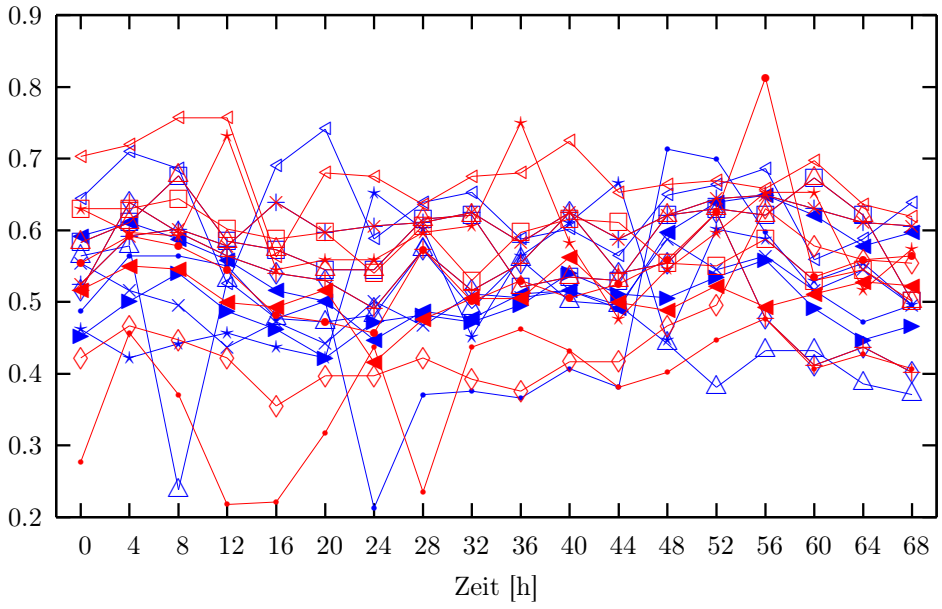

Abbildung 10: Verlauf der Enzymkonzentration bei den männlichen (blau) und den weiblichen Probanden (rot).

Den Ausgangsdaten ist unmittelbar nichts besonderes anzusehen, was für die eine oder andere Hypothese spräche. Werden Tag und Tageszeit als Subplot-Faktoren betrachtet und Geschlecht als Whole-plot-Faktor, unter den die Probanden verschachtelt sind, dann ergeben sich mit der Anwendung des neuen Tests folgende Resultate:

Effekt	Q_F	$\hat{f}$	$\hat{f}_0$	p-Wert
Geschlecht	0,0750	1,0000	17,481	0,7873
Tag	1,2414	2,0837	32,727	0,3034
Tageszeit	8,2746	4,9884	120,43	0,0000
Geschlecht × Tag	0,0094	2,0837	32,727	0,9921
Geschlecht × Tageszeit	1,3052	4,9884	120,43	0,2663
Tageszeit ×Tag	3,4021	9,0973	196,89	0,0006
Geschlecht × Tag × Tageszeit	0,6215	9,0973	196,89	0,7797

Es konnte nur eine Wechselwirkung zwischen Tageszeit und Tag gefunden werden. Der Verlauf der Enzymkonzentration an einem Tag unterscheidet sich also mit der Art des Schlafes.

10 Zusammenfassung und Ausblick

In dieser Arbeit ist es gelungen, einen allgemeinen Test für Zweistichproben-Split-Plot-Designs unter Normalverteilung zu konstruieren, der keinerlei Annahmen über die Kovarianzmatrizen voraussetzt, weder über die Gleichheit der Matrizen noch über deren Struktur, und der keine Einschränkungen bezüglich der Anzahl der Messwiederholungen macht. Lediglich ein Mindeststichprobenumfang von vier Versuchseinheiten ist zu gewährleisten, damit die Statistik definiert ist.

Neben den schwachen Voraussetzungen sind seine leichte rechentechnische Implementierbarkeit und seine schnelle Ausführbarkeit besondere praktische Vorteile des neuen Tests. Implementationen bisher existierender Varianzkomponentenverfahren benötigen für viele Messwiederholungen deutlich mehr Zeit, weil sie auf die iterative Berechnung der großen Kovarianzmatrizen angewiesen sind. Dies ist algorithmisch deutlich komplexer als die Skalarprodukte der neuen Statistik auszurechnen.

In umfangreichen Simulationen wurde beobachtet, dass der Test sein Niveau für mäßig große Stichprobenumfänge ab $n_1, n_2 > 20$ sehr gut und für geringere Umfänge in ausreichendem Maße einhält.

Zur Verbesserung kann man eine Taylorapproximation des Erwartungswertes der Freiheitsgradschätzer heranziehen, um bei sehr kleinen Stichprobenumfängen die Genauigkeit zu verbessern. Dies würde die exakte Berechnung der Varianz und der Kovarianz der Spurschätzer voraussetzen, was zwar möglich, doch sehr aufwendig ist.

Eine andere Verbesserungsmöglichkeit für kleine Stichprobenumfänge wäre, die Spurschätzer für $\mathrm{Sp}^2\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)$ und $\mathrm{Sp}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)^2$ aus [2] als Ausgangspunkt zu nehmen und das Ergebnis mit den neu gefundenen zu korrigieren. Damit könnte es möglich werden, nur noch in einer Stichprobe mindestens vier Versuchseinheiten zu verlangen, in der anderen mindestens zwei. Ob die resultierenden Freiheitsgradschätzer dann tatsächlich besser sind, müsste anschließend erforscht werden.

Weiterhin sollte sich der Test unkompliziert für mehrere Stichproben ausbauen lassen, wenn man bedenkt, dass die Differenz $\overline{\boldsymbol{Y}}_{1\cdot} - \overline{\boldsymbol{Y}}_{2\cdot}$ und die Sum-

men $n_1^{-1}\boldsymbol{\Sigma}_1 + n_2^{-1}\boldsymbol{\Sigma}_2$ nur von der speziellen Form der Matrix $\boldsymbol{P}_2$ herrühren, die der Gruppenstruktur Rechnung trägt. Bei mehreren Stichproben sollten sich gewichtete Summen der Kovarianzmatrizen bilden, die sich ebenfalls so zerlegen lassen, dass die neuen Spurschätzer anwendbar sind.

Auf die Robustheit des Test wurde in dieser Arbeit überhaupt nicht eingegangen, aber dafür kann ein Rahmen ausgearbeitet werden, der allgemeiner als die Normalverteilungsannahme ist. Die Erwartungstreue der Spurschätzer kann schon ohne Normalverteilung bewiesen werden. Schwieriger ist die Dimensionsstabilität.

Ein besonders herausforderndes Problem ist die Übertragung der Statistik auf Ränge. Bei komponentenweiser Rangvergabe entstünde eine hochdimensional besser anwendbare Version des nichtparametrischen, multivariaten Tests aus [10]. Bei Rangvergabe über alle Werte ergäbe sich ein nichtparametrischer Test für longitudinale Daten. Da die Ränge stets abhängig voneinander sind, die Erwartungstreue der Spurschätzer aber Unabhängigkeit der Beobachtungsvektoren voraussetzt, müssen entweder die Schätzer grundlegend modifiziert werden oder gezeigt werden, dass die Abhängigkeit keinen allzu schädlichen Einfluss auf das Endergebnis hat.

A Anhang

A.1 Benutzte Sätze

Theorem A.1. *(Rang der empirischen Kovarianzmatrix) Seien* $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ *unabhängig identisch multivariat normalverteilt mit Kovarianzmatrix* $\boldsymbol{V} = Cov(\boldsymbol{Y}_k)$*. Dann gilt für die empirische Kovarianzmatrix*

$$\hat{\boldsymbol{V}} = (n-1)^{-1} \sum_{k=1}^{n} \left(\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_\cdot\right) \left(\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_\cdot\right)'$$

fast sicher $\mathrm{rg}\left(\hat{\boldsymbol{V}}\right) = \min\left(\mathrm{rg}\boldsymbol{V}, n\right)$*.*

Beweis. Siehe [22] S. 82. □

Theorem A.2. *(Satz von* Craig *und* Sakamoto*) Sei* $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *, und seien* $\boldsymbol{A}$ *und* $\boldsymbol{B}$ *symmetrisch positiv semidefinit. Dann sind* $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$ *und* $\boldsymbol{Y}'\boldsymbol{B}\boldsymbol{Y}$ *unabhängig, falls* $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$ *ist.*

Theorem A.3. *(Repräsentationstheorem) Sei* $\boldsymbol{A}$ *symmetrisch und* $\boldsymbol{Y} \in \mathbb{R}^d$ *mit* $E\boldsymbol{Y} = \boldsymbol{0}$ *und* $Cov(\boldsymbol{Y}) = \boldsymbol{\Sigma}$ *mit* $\mathrm{rg}\boldsymbol{\Sigma} = r \leq d$ *und* $\boldsymbol{\Sigma}$*. Dann gibt es* $\boldsymbol{Z} \in \mathbb{R}^d$ *mit* $E\boldsymbol{Z} = \boldsymbol{0}$ *und* $Cov(\boldsymbol{Z}) = \boldsymbol{I}_d$*, so dass*

$$\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y} = \sum_{j=1}^{d} \lambda_j Z_i$$

gilt und λ_j *die Eigenwerte von* $\boldsymbol{A}\boldsymbol{\Sigma}$ *sind. Insbesondere ist* $Z_i \sim \chi_1^2$*, falls* $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$*.*

Theorem A.4. *(Satz von Slutsky)*

1. Die Folge von Zufallsvektoren $(\boldsymbol{Y}_k)_{k=1,\ldots,\infty}$ mit $\boldsymbol{Y}_k \in \mathbb{R}^d$ konvergiere stochstisch gegen $\boldsymbol{Y}_0 \in \mathbb{R}^d$ und f sei eine in $\boldsymbol{Y}_0$ stetige Abbildung. Dann konvergiert auch $f(\boldsymbol{Y}_k)$ stochastisch gegen $f(\boldsymbol{Y}_0)$.

2. Sei $\boldsymbol{Y}_k$ eine Folge von Zufallsvektoren, die in Verteilung gegen $\boldsymbol{Y}_0 \sim F(\boldsymbol{x})$ konvergiert und sei f eine F-fast überall stetige Abbildung. Dann konvergiert auch $f(\boldsymbol{Y}_k)$ in Verteilung gegen $f(\boldsymbol{Y}_0)$.

Theorem A.5. *(Satz von der Singulärwertzerlegung) Sei $\boldsymbol{A} \in \mathbb{R}^{u\times v}$. Dann gibt es $\boldsymbol{Q}_1 \in O_u$ und $\boldsymbol{Q}_2 \in O_v$, so dass $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{D}\boldsymbol{Q}_2$ und $\boldsymbol{D}$ eine Matrix nur mit Diagonaleinträgen ist.*

Beweis. Siehe [14], S. 147. □

A.2 Berechnung der Spurschätzer

A.2.1 Die Terme $B_1^{(1)}$, $B_1^{(2)}$ und C_1

Es sei an die Konvention erinnert, dass die Indices k, l, s und t paarweise verschieden sein sollen. Weiterhin sei vereinbart, dass $\boldsymbol{M}_i = \begin{pmatrix} \boldsymbol{Y}_{i1} & \dots & \boldsymbol{Y}_{in_i} \end{pmatrix}'$ die $n_i \times d$-Datenmatrix der Stichprobe i bezeichnen soll.

Ausmultiplizieren der $A_{kl}^{(i)}$ ergibt die Einträge der Matrix

$$\boldsymbol{K}_i = (\boldsymbol{Y}'_{it}\boldsymbol{Y}_{it} - 2\,\boldsymbol{Y}'_{it}\boldsymbol{Y}_{iu} + \boldsymbol{Y}'_{iu}\boldsymbol{Y}_{iu})_{t,u=1,\dots,n_i} = \left(A_{tu}^{(i)}\right)_{t,u=1,\dots,n_i},$$

die aus $\boldsymbol{M}_i\boldsymbol{M}'_i$ mit der Auswahl des Vektors $\boldsymbol{m}_i$ ihrer Diagonaleinträge (SAS-Funktion `vecdiag`) erzeugt werden kann:

$$\boldsymbol{K}_i = \boldsymbol{m}_i\mathbf{1}'_{n_i} - 2\,\boldsymbol{M}_i\boldsymbol{M}'_i + \mathbf{1}_{n_i}\boldsymbol{m}'_i$$

Als Grundlage für $B_1^{(i)}$ kann $\left(\mathbf{1}'_{n_i}\boldsymbol{K}_i\mathbf{1}_{n_i}\right)^2$ benutzt werden. In $\left(\mathbf{1}'_{n_i}\boldsymbol{K}_i\mathbf{1}_{n_i}\right)^2 = \sum_{t,u,v,w=1}^{n_i} A_{tu}^{(i)}A_{vw}^{(i)}$ sind nun die Fälle zu viel, wo einer der Indices mit einem anderen übereinstimmt. Man kann alle $A_{tu}^{(i)}A_{vw}^{(i)}$ mit $t = v$, $u = w$, $t = w$ oder $u = v$ abziehen, wenn man $4\,\mathbf{1}'_{n_i}\boldsymbol{K}_i\boldsymbol{K}_i\mathbf{1}_{n_i}$ subtrahiert. Dadurch hat man aber Terme der Art $A_{tu}^{(i)}A_{tu}^{(i)}$ zweimal zu viel abgezogen. Ihre Summe kann man durch $\boldsymbol{K}_i\#\boldsymbol{K}_i$, d. h. durch komponentenweise Matrixmultiplikation, darstellen. Die Berechnungsformel lautet also

$$B_1^{(i)} = \tfrac{1}{4n_i(n_i-1)(n_i-2)(n_i-3)}\left(\left(\mathbf{1}'_{n_i}\boldsymbol{K}_i\mathbf{1}_{n_i}\right)^2 - 4\,\mathbf{1}'_{n_i}\boldsymbol{K}_i\boldsymbol{K}_i\mathbf{1}_{n_i} + 2\,\boldsymbol{K}_i\#\boldsymbol{K}_i\right)$$

C_1 kann mit den empirischen Kovarianzmatrizen ausgerechnet werden. Unter Verwendung der Notation 7 kann man nämlich schreiben:

$$\boldsymbol{y}_i' \boldsymbol{T}_{n_i d} \left(\left((\boldsymbol{e}_k - \boldsymbol{e}_l)(\boldsymbol{e}_k - \boldsymbol{e}_l)' \right) \otimes \boldsymbol{I}_d \right) \boldsymbol{T}_{n_i d} \boldsymbol{y}_i = (\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il})' \boldsymbol{T} (\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il}) = A_{kl}^{(i)}$$

Wenn man $\sum_{k,l=1}^{n_i} (\boldsymbol{e}_k - \boldsymbol{e}_l)(\boldsymbol{e}_k - \boldsymbol{e}_l)' = 2\, n_i \boldsymbol{P}_{n_i}$ einsieht und sich an Gleichung (8) auf S.25 erinnert,

$$\begin{aligned} C_1 &= \tfrac{1}{4n_1(n_1-1)n_2(n_2-1)} \left(\sum_{k,l}^{n_1} A_{kl}^{(1)} \right) \left(\sum_{k'l'}^{n_2} A_{rs}^{(2)} \right) \\ &= \tfrac{1}{(n_1-1)(n_2-1)} \left(\boldsymbol{y}_i' \boldsymbol{T}_{n_i d}' (\boldsymbol{P}_{n_i} \otimes \boldsymbol{I}_d) \boldsymbol{T}_{n_i d} \boldsymbol{y}_i \right) \left(\boldsymbol{y}_i' \boldsymbol{T}_{n_i d}' (\boldsymbol{P}_{n_i} \otimes \boldsymbol{I}_d) \boldsymbol{T}_{n_i d} \boldsymbol{y}_i \right) \\ &= \mathrm{Sp}\hat{\boldsymbol{\Sigma}}_1 \cdot \mathrm{Sp}\hat{\boldsymbol{\Sigma}}_2 \end{aligned} \tag{20}$$

Man sieht dabei nebenbei auch, dass $\mathrm{Sp}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_i = \frac{1}{2n_i(n_i-1)} \sum_{k,l}^{n_i} A_{kl}^{(i)}$ gilt. Das Bestreben, $B_0^{(A)}$ aus Statistik (18) durch entsprechende quadratische Formen für ungleiche Kovarianzmatrizen und Stichprobenumfänge zu ersetzen, hat also zwangsläufig zur Spur der empirischen Kovarianzmatrix geführt. Diese wiederum hat die Ausnutzung des Satzes von Craig und Sakamoto möglich gemacht.

Der Berechnung von C_1 durch $\hat{\boldsymbol{\Sigma}}_1$ und $\hat{\boldsymbol{\Sigma}}_2$ ist die Berechnung durch

$$C_1 = \frac{1}{4n_1(n_1-1)n_2(n_2-1)} \left(\mathbf{1}_{n_1}' \boldsymbol{K}_1 \mathbf{1}_{n_1} \right) \left(\mathbf{1}_{n_2}' \boldsymbol{K}_2 \mathbf{1}_{n_2} \right)$$

vorzuziehen, da die $\boldsymbol{K}_i$ im Programm ohnehin ausgerechnet werden.

A.2.2 Die Terme $B_2^{(1)}$, $B_2^{(2)}$ und C_2

Die Bilinearformen $A_{klrs}^{(i)}$ können ausmultipliziert werden.

$$\begin{aligned} A^{(i)}_{klrs} &= (\boldsymbol{Y}_{ik} - \boldsymbol{Y}_{il})' (\boldsymbol{Y}_{ir} - \boldsymbol{Y}_{is}) \\ &= \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir} - \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is} - \boldsymbol{Y}'_{il}\boldsymbol{Y}_{ir} + \boldsymbol{Y}'_{il}\boldsymbol{Y}_{is} \end{aligned}$$

Wird dieser Ausdruck quadriert, ergeben sich genau 16 Terme der Form $\boldsymbol{Y}'_{it}\boldsymbol{Y}_{iu}\boldsymbol{Y}'_{iv}\boldsymbol{Y}_{iw}$. Terme, bei denen ein Index genau doppelt vorkommt –etwa $t = v$ oder $u = w$–, werden subtrahiert, während Terme mit verschiedenen Indices oder mit zwei Paaren identischer Indices addiert werden. Nun werden die Terme mit denselben Indexwiederholungen über alle Summationsindices hinweg zusammengefasst. Es wird dabei nur über verschiedene Indices k, l, r, s aufsummiert. Man könnte auch $k = l$ und $r = s$ zulassen, weil die Bilinearformen dann sowieso verschwinden, aber die Abzählung und die Zwischenergebnisse würden komplizierter werden.

In $4\,n_i\,(n_i-1)\,(n_i-2)\,(n_i-3)$ Summanden kommen die Indexpermutationen des Typs $(\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir})^2$ vor, von denen natürlich nur $n_i\,(n_i-1)$ nicht identisch sind. Von den Typen $\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is}$ und $\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{il}\boldsymbol{Y}_{ir}$ kommen insgesamt $8\,n_i\,(n_i-1)\,(n_i-2)\,(n_i-3)$ Summanden vor, von denen wiederum $n_i\,(n_i-1)\,(n_i-2)$ nicht identisch sind. Vom Typ $\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{il}\boldsymbol{Y}_{is}$ sind es $4\,n_i\,(n_i-1)\,(n_i-2)\,(n_i-3)$ Summanden.

Mit $\boldsymbol{A}^0_i := (\boldsymbol{M}_i\boldsymbol{M}'_i)\,\#\,(\boldsymbol{J}_{n_i} - \boldsymbol{I}_{n_i})$ ist die Summe aller $(\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir})^2$ somit

$$\begin{aligned} B^{(i)}_{21} &= 4 \sum_{k,l,r,s \text{ versch.}}^{n_i} (\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir})^2 = 4\,(n_i-2)\,(n_i-3) \sum_{k\neq r}^{n_i} (\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir})^2 \\ &= 4\,(n_i-2)\,(n_i-3)\,\mathbf{1}'_{n_i}\left(\boldsymbol{A}^0_i \# \boldsymbol{A}^0_i\right)\mathbf{1}_{n_i}. \end{aligned}$$

Als Summe aller Terme der Art $\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is}$ und $\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{il}\boldsymbol{Y}_{ir}$ geschrieben ist

$$
\begin{aligned}
B_{22}^{(i)} &= 4 \sum_{k,l,r,s \text{ versch.}}^{n_i} \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is} + 4 \sum_{k,l,r,s \text{ versch.}}^{n_i} \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{il}\boldsymbol{Y}_{ir} \\
&= 8 \sum_{k,l,r,s \text{ versch.}}^{n_i} \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is} \\
&= 8\,(n_i-3) \sum_{k\neq r\neq s\neq k}^{n_i} \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is} \\
&= 8\,(n_i-3) \left[\sum_{k\neq r,s}^{n_i} \boldsymbol{Y}'_{ir}\boldsymbol{Y}_{ik}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is} - \sum_{k\neq r}^{n_i} \boldsymbol{Y}'_{ir}\boldsymbol{Y}_{ik}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\right] \\
&= 8\,(n_i-3) \left[\mathbf{1}'_{n_i}\left(\boldsymbol{A}_i^0 \cdot \boldsymbol{A}_i^0\right)\mathbf{1}_{n_i} - \mathrm{Sp}\left(\boldsymbol{A}_i^0 \cdot \boldsymbol{A}_i^0\right)\right] \\
&= 8\,(n_i-3) \left[\mathbf{1}'_{n_i}\left(\boldsymbol{A}_i^0 \cdot \boldsymbol{A}_i^0\right)\mathbf{1}_{n_i} - \mathbf{1}'_{n_i}\left(\boldsymbol{A}_i^0 \# \boldsymbol{A}_i^0\right)\mathbf{1}_{n_i}\right] \\
&= 8\,(n_i-3)\,\mathbf{1}'_{n_i}\left(\left(\boldsymbol{A}_i^0 \cdot \boldsymbol{A}_i^0\right) \# \left(\boldsymbol{J}_{n_i} - \boldsymbol{I}_{n_i}\right)\right)\mathbf{1}_{n_i}
\end{aligned}
$$

Komplizierter ist die Summe der $\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{il}\boldsymbol{Y}_{is}$.

$$
\begin{aligned}
B_{23}^{(i)} &= \sum_{k,l,r,s \text{ versch.}}^{n_i} \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{il}\boldsymbol{Y}_{is} \\
&= \sum_{k\neq r,l\neq s}^{n_i} \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{il}\boldsymbol{Y}_{is} - 2\sum_{k\neq r}^{n_i} \left(\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\right)^2 - 4 \sum_{k\neq r\neq s\neq k}^{n_i} \boldsymbol{Y}'_{ik}\boldsymbol{Y}_{ir}\boldsymbol{Y}'_{ik}\boldsymbol{Y}_{is} \\
&= \left(\mathbf{1}'_{n_i}\boldsymbol{A}_i^0\mathbf{1}_{n_i}\right)^2 - 2\cdot\mathbf{1}'_{n_i}\left(\boldsymbol{A}_i^0 \# \boldsymbol{A}_i^0\right)\mathbf{1}_{n_i} - 4\cdot\mathbf{1}'_{n_i}\left(\left(\boldsymbol{A}_i^0 \cdot \boldsymbol{A}_i^0\right) \# \left(\boldsymbol{J}_{n_i} - \boldsymbol{I}_{n_i}\right)\right)\mathbf{1}_{n_i}
\end{aligned}
$$

$B_2^{(i)}$ berechnet sich damit durch

$$
B_2^{(i)} = \frac{B_{21}^{(i)} - B_{22}^{(i)} + B_{23}^{(i)}}{n_i\,(n_i-1)\,(n_i-2)\,(n_i-3)}
$$

C_2 könnte am einfachsten durch $\mathrm{Sp}\left(\hat{\boldsymbol{V}}_1\hat{\boldsymbol{V}}_2\right)$ berechnet werden vermöge der folgenden Umformungen:

$$
\begin{aligned}
C_2 &= \tfrac{1}{4n_1n_2(n_1-1)(n_2-1)} \sum_{i=1}^{n_1}\sum_{k=1}^{n_1}\sum_{j=1}^{n_2}\sum_{l=1}^{n_2} \left((\boldsymbol{X}_{1i} - \boldsymbol{X}_{1k})' (\boldsymbol{X}_{2j} - \boldsymbol{X}_{2l}) \right)^2 \\
&= \tfrac{1}{n_1n_2(n_1-1)(n_2-1)} \sum_{i=1}^{n_1}\sum_{k=1}^{n_1}\sum_{j=1}^{n_2}\sum_{l=1}^{n_2} \left(\boldsymbol{X}'_{2j}\boldsymbol{X}_{1i} - \boldsymbol{X}'_{2j}\boldsymbol{X}_{1k} \right) \left(\boldsymbol{X}'_{2j}\boldsymbol{X}_{1i} - \boldsymbol{X}'_{2l}\boldsymbol{X}_{1i} \right) \\
&= \tfrac{1}{n_1n_2(n_1-1)(n_2-1)} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \boldsymbol{X}'_{2j} \left(\boldsymbol{X}_{1i} - \overline{\boldsymbol{X}}_{1\cdot} \right) \left(\boldsymbol{X}_{2j} - \overline{\boldsymbol{X}}_{2\cdot} \right)' \boldsymbol{X}_{1i} \qquad (21) \\
&= \tfrac{1}{(n_1-1)(n_2-1)} \mathrm{Sp} \left[\boldsymbol{X}'_{2j} \left(\boldsymbol{X}_{1i} - \overline{\boldsymbol{X}}_{1\cdot} \right) \right]_{\substack{j=1,\ldots,n_2 \\ i=1,\ldots,n_1}} \left[\left(\boldsymbol{X}_{2j} - \overline{\boldsymbol{X}}_{2\cdot} \right)' \boldsymbol{X}_{1i} \right]_{\substack{j=1,\ldots,n_2 \\ i=1,\ldots,n_1}} \\
&= \tfrac{1}{(n_1-1)(n_2-1)} \mathrm{Sp} \boldsymbol{M}_2 \boldsymbol{M}'_1 \boldsymbol{P}_{n_1} \boldsymbol{M}_1 \boldsymbol{M}'_2 \boldsymbol{P}_{n_2} \\
&= \tfrac{1}{(n_1-1)(n_2-1)} \mathrm{Sp} \boldsymbol{M}'_1 \boldsymbol{P}_{n_1} \boldsymbol{M}_1 \boldsymbol{M}'_2 \boldsymbol{P}_{n_2} \boldsymbol{M}_2 \\
&= \mathrm{Sp} \hat{\boldsymbol{V}}_1 \hat{\boldsymbol{V}}_2
\end{aligned}
$$

Technisch ist es aber günstiger, nicht mit den empirischen Kovarianzmatrizen zu rechnen, sondern lieber den Ansatz für die Berechnung von B_2 zu übertragen. Anstatt mit $d \times d$-Matrizen wird auf diese Weise nur mit $n_1 \times n_2$-Matrizen gearbeitet, was im hochdimensionalen Fall viel Speicher spart. Analog zum Verfahren bei $B_2^{(i)}$ fasst C_{21} die Quadrate der Bilinearformen zusammen, C_{22} und C_{23} fassen die Terme zusammen, bei denen ein Index aus der ersten bzw. zweiten Stichprobe wiederholt wird und die subtrahiert werden, und C_{24} vereinigt die Terme mit vier verschiedenen Indices. Ausgangspunkt ist hierbei $\boldsymbol{A}_{12} = \boldsymbol{M}_1 \boldsymbol{M}'_2$, die Matrix der Skalarprodukte von Beobachtungsvektoren jeweils beider Stichproben.

$$
\begin{aligned}
C_{21} &= (n_1 - 1)(n_2 - 1) \sum_{k=1}^{n_1}\sum_{r=1}^{n_2} \left(\boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \right)^2 \\
&= (n_1 - 1)(n_2 - 1) \, \mathbf{1}'_{n_1} \left(\boldsymbol{A}_{12} \# \boldsymbol{A}_{12} \right) \mathbf{1}_{n_2} \\
C_{22} &= (n_1 - 1) \sum_{k=1}^{n_1}\sum_{r \neq s}^{n_2} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2s}
\end{aligned}
$$

$$= (n_1 - 1) \left(\sum_{k=1}^{n_1} \sum_{r,s=1}^{n_2} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2s} - \sum_{k=1}^{n_1} \sum_{r=1}^{n_2} \left(\boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \right)^2 \right)$$
$$= (n_1 - 1) \left(\mathbf{1}'_{n_2} \boldsymbol{A}'_{12} \cdot \boldsymbol{A}_{12} \mathbf{1}_{n_2} - \mathbf{1}'_{n_1} \left(\boldsymbol{A}_{12} \# \boldsymbol{A}_{12} \right) \mathbf{1}_{n_2} \right)$$

Vertauscht man in C_{22} die Rollen der Stichproben, erhält man C_{23}:

$$C_{23} = (n_2 - 1) \left(\mathbf{1}'_{n_1} \boldsymbol{A}_{12} \cdot \boldsymbol{A}'_{12} \mathbf{1}_{n_1} - \mathbf{1}'_{n_1} \left(\boldsymbol{A}_{12} \# \boldsymbol{A}_{12} \right) \mathbf{1}_{n_2} \right)$$

$$\begin{aligned} C_{24} &= \sum_{k \neq l}^{n_1} \sum_{s \neq r}^{n_2} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \boldsymbol{Y}'_{1l} \boldsymbol{Y}_{2s} \\ &= \sum_{k,l=1}^{n_1} \sum_{r,s=1} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \boldsymbol{Y}'_{1l} \boldsymbol{Y}_{2s} - \sum_{k}^{n_1} \sum_{r,s}^{n_2} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2s} \\ &\qquad - \sum_{k,l}^{n_1} \sum_{r}^{n_2} \boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \boldsymbol{Y}'_{1l} \boldsymbol{Y}_{2r} + \sum_{k}^{n_1} \sum_{r}^{n_2} \left(\boldsymbol{Y}'_{1k} \boldsymbol{Y}_{2r} \right)^2 \\ &= \left(\mathbf{1}'_{n_1} \boldsymbol{A}_{12} \mathbf{1}_{n_2} \right)^2 - \mathbf{1}'_{n_1} \boldsymbol{A}_{12} \cdot \boldsymbol{A}'_{12} \mathbf{1}_{n_1} - \mathbf{1}'_{n_2} \boldsymbol{A}'_{12} \boldsymbol{A}_{12} \mathbf{1}_{n_2} + \mathbf{1}'_{n_1} \left(\boldsymbol{A}_{12} \# \boldsymbol{A}_{12} \right) \mathbf{1}_{n_2} \end{aligned}$$

Mit diesen Termen erhält man C_2 durch

$$C_2 = \frac{(C_{21} - C_{22} - C_{23} + C_{24})}{n_1 (n_1 - 1) n_2 (n_2 - 1)}.$$

A.3 Makro F1-HD-F1

Hier ist der Quellcode des Makros F1-HD-F1. Das Makro F1-HD-F2 unterscheidet davon sich lediglich im Einlesen der Daten, in der Berechnung der Hypothesen und in der Ausgabe. Das zentrale Modul `box` ist bei beiden Makros identisch.

```
%macro f1_hd_f1(data=,
VAR=,
Subplot=,
Subject=,
Wholeplot=,);
```

Zunächst werden die Daten sortiert:

```
proc sort data=&data;
by &Wholeplot &Subject &Subplot;
run;
proc iml;
start box(X1,X2, Qf,ndf,ddf,p);
n1=nrow(X1);
n2=nrow(X2);
d=ncol(X1);
Auswahl1=J(n1,n1,1)-I(n1);
Auswahl2=J(n2,n2,1)-I(n2);
Pn1=I(n1)-J(n1,n1,1/n1);
Pn2=I(n2)-J(n2,n2,1/n2);
trV1=trace(Pn1*X1*X1‘*Pn1);
trV2=trace(Pn2*X2*X2‘*Pn2);
sumtr=trV1+trV2;
if sumtr=0 then print "Warning: No variance found. Testing is
  neither possible nor necessary.";
B0=trV1/((n1-1)*n1)+trV2/((n2-1)*n2);
A1=X1*X1‘;
A2=X2*X2‘;
A12=X1*X2‘;
```

Schätzer für $\mathrm{Sp}^2\mathbf{\Sigma}$:

```
K1=vecdiag(A1)*J(1,n1,1)+J(n1,1,1)*(vecdiag(A1))‘-2*A1;
B1_1=((sum(K1))**2-4*sum(K1*K1)+2*sum(K1#K1))
  /(4*n1*(n1-1)*(n1-2)*(n1-3));
K2=vecdiag(A2)*J(1,n2,1)+J(n2,1,1)*(vecdiag(A2))‘-2*A2;
B1_2=((sum(K2))**2-4*sum(K2*K2)+2*sum(K2#K2))
  /(4*n2*(n2-1)*(n2-2)*(n2-3));
C1=sum(K1)*sum(K2)/(4*n1*(n1-1)*n2*(n2-1));
B1=B1_1/(n1**2)+2*C1/(n1*n2)+B1_2/(n2**2);
```

Schätzer für $\mathrm{Sp}\boldsymbol{\Sigma}^2$:

```
ANull1=A1#Auswahl1;
B211=(n1-2)*(n1-3)*sum(ANull1#ANull1);
B212=2*(n1-3)*sum((ANull1*ANull1)#Auswahl1);
B213=((sum(ANull1))**2-2*sum(ANull1#ANull1)
   -4*sum((ANull1*ANull1)#Auswahl1));
B2_1=(B211-B212+B213)/(n1*(n1-1)*(n1-2)*(n1-3));
ANull2=A2#Auswahl2;
B221=4*(n2-2)*(n2-3)*sum(ANull2#ANull2);
B222=8*(n2-3)*sum((ANull2*ANull2)#Auswahl2);
B223=4*( (sum(ANull2))**2-2*sum(ANull2#ANull2)
   -4*sum((ANull2*ANull2)#Auswahl2));
B2_2=(B221-B222+B223)/(4*n2*(n2-1)*(n2-2)*(n2-3));
C21=(n1-1)*(n2-1)*sum(A12#A12);
C22=(n1-1)*(sum(A12‘*A12)-sum(A12#A12));
C23=(n2-1)*(sum(A12*A12‘)-sum(A12#A12));
C24=(sum(A12))**2-sum(A12*A12‘)-sum(A12‘*A12)+sum(A12#A12);
C2=(C21-C22-C23+C24)/(n1*(n1-1)*n2*(n2-1));
B2c=B2_1/(n1**2)+2*C2/(n1*n2)+B2_2/(n2**2);
B2d=B2_1/(n1**2*(n1-1))+B2_2/(n2**2*(n2-1));
```

Zusammensetzen der Freiheitsgrade und Ausrechnen der ANOVA-Typ-Statistik:

```
ndf=B1/B2c;
ddf=B1/B2d;
XQuer1=X1[:,];
XQuer2=X2[:,];
XQuer=XQuer1-XQuer2;
QF=XQuer*XQuer‘/B0;
p=1-probf(QF,ndf,ddf);
finish box;
print "F1-HD-F1 --- high-dimensional Split-Plot",
"Subject(Group) x Time",
"Random effect: Subject; Fixed effects: Group, Time";
print ’SAS-datafile-name: ’ "&data" ,
```

```
'Response variable: ' "&var" ;
USE &data;
```

Variable für die Stichproben einlesen und warnen, wenn nicht genau zwei Stichproben vorhanden sind:

```
READ ALL VAR{&Wholeplot} INTO FaktA;
lev_a=unique(FaktA);
a=ncol(lev_a); *es sollte a=2 sein;
if a^=2 then print "Warning: Statistic can only use two samples.";
```

Einlesen der Faktorstufen des Sub-plot-Faktors:

```
READ ALL VAR{&Subplot} INTO FaktB;
lev_b=unique(FaktB);
d=ncol(lev_b);
```

Hier werden die Nummern der Versuchseinheiten und Sup-plot-Faktorstufen eingelesen und für den Fall zu geringer Stichprobenumfänge das Programm mit einer Fehlermeldung abgebrochen:

```
READ ALL VAR{&Subject} INTO FaktC1 where(&Wholeplot=(lev_a[1]));
READ ALL VAR{&Subject} INTO FaktC2 where(&Wholeplot=(lev_a[2]));
lev_c1=unique(FaktC1);
lev_c2=unique(FaktC2);
n1=ncol(lev_c1);
n2=ncol(lev_c2);
call symput('ssn1',left(char(n1)));
call symput('ssn2',left(char(n2)));
%if %eval(&ssn1<4 | &ssn2<4) %then %do;
%put %str(Error: At least 4 individuals in each sample
   required.);
%abort;
%end;
if type(lev_a)="N" then print ((lev_a[1]//lev_a[2])||
```

```
   (n1//n2)) [rowname={"First","Second"} colname
   ={"Group Name","sample size"} label="Samples used"];
if type(lev_a)="C" then print ((lev_a[1]//lev_a[2])||
   (char(n1)//char(n2)))[rowname={"First","Second"} colname
   ={"Group Name","sample size"} label="Samples used"];
READ ALL VAR{&VAR} INTO werte1 where(&wholeplot=(lev_a[1]));
READ ALL VAR{&VAR} INTO werte2 where(&wholeplot=(lev_a[2]));
CLOSE &data;
```

Eine Warnung, wenn nicht in jeder Stichprobe exakt $n_i d$ Beobachtungen sind:

```
if (nrow(werte1)^=n1*d)|(nrow(werte2)^=n2*d) then print
   "Warning: Dataset contains missing or dublicate Values.";
print ({&Subplot}||(char(ncol(lev_b))))[colname
   ={"Factor name" "levels"} label="Subplot factor"];
```

Werte jeder Stichprobe in $n_i \times d$-Matrixform bringen:

```
X1mat = (shape(werte1,n1,d));
X2mat = (shape(werte2,n2,d));
Erg=J(3,4,0);
colnameErg={"Qf","ndf","ddf","Pr > F"};
rownameErg={"&Wholeplot" "&Subplot"}||concat("&Wholeplot",
   "*","&Subplot");
qf=0;
ndf=0;
ddf=0;
p=0;
```

Haupteffekt A – Hypothesenmatrix ist $d^{-1}\boldsymbol{J}_{d\times d}$:

```
X1matA=X1mat*J(d,1,1/d);
X2matA=X2mat*J(d,1,1/d);
run box(X1matA,X2matA, qf,ndf,ddf,p);
Erg[1,]=qf||ndf||ddf||p;
```

Haupteffekt B – der Operator „:" ist schneller als Multiplikation mit $\boldsymbol{P}_d$:

```
X1matB = X1mat-X1mat[,:]*J(1,d,1);
X2matB = X2mat[,:]*J(1,d,1)-X2mat;
run box(X1matB,X2matB, qf,ndf,ddf,p);
Erg[2,]=qf||ndf||ddf||p;
```

Wechselwirkung AB:

```
X2matAB=-X2matB;
run box(X1matB,X2matAB, qf,ndf,ddf,p);
Erg[3,]=qf||ndf||ddf||p;
```

Ausgabe der Ergebnistabelle:

```
print Erg[colname=colnameErg rowname=rownameErg label
   ='F-Approximation' format=6.4];
%mend f1_hd_f1;
```

Literatur

[1] Analysis of high-dimensional repeated measures designs: The one sample case. In: *Computational Statistics and Data Analysis* 53 (2008), Nr. 2, S. 416 – 427

[2] Ahmad, Muhammad R.: *Analysis of High Dimensional Repeated Measures Designs: The One- and Two-sample Test Statistics*, Georg-August-Universität zu Göttingen, Dissertation, 2008

[3] Aldworth, Jeremy ; Hoffman, Wherly P.: Split-plot model with covariate: a cautionary tale. In: *Amer. Statist.* 56 (2002), Nr. 4, S. 284–289

[4] Bai, Zhidong ; Saranadasa, Hewa: Effect of high dimension: by an example of a two sample problem. In: *Statist. Sinica* 6 (1996), Nr. 2, S. 311–329

[5] Box, George E. P.: 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. In: *Ann. Math. Statistics* 25 (1954), S. 290–302

[6] Box, George E. P.: 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. In: *Ann. Math. Statistics* 25 (1954), S. 484–498

[7] Boysen, Leif: *Analyse von intra-individuellen Effekten bei longitudinalen Daten*, Inst. für Mathematische Stochastik, Universität Göttingen, Diplomarbeit, 2002

[8] Brunner, Edgar ; Dette, Holger ; Munk, Axel: Box-type approximations in nonparametric factorial designs. In: *J. Amer. Statist. Assoc.* 92 (1997), Nr. 440, S. 1494–1502

[9] Brunner, Edgar ; Domhof, Sebastian ; Langer, Frank: *Nonparametric analysis of longitudinal data in factorial experiments*. New York : Wiley-Interscience [John Wiley & Sons], 2002 (Wiley Series in Probability and Statistics)

[10] Brunner, Edgar ; Munzel, Ullrich ; Puri, Madan L.: The multivariate nonparametric Behrens-Fisher problem. In: *J. Statist. Plann. Inference* 108 (2002), Nr. 1-2, S. 37–53. – C. R. Rao 80th birthday felicitation volume, Part II

[11] Casella, George ; Berger, Roger L.: *Statistical inference*. Pacific Grove, CA : Wadsworth & Brooks/Cole Advanced Books & Software, 1990 (The Wadsworth & Brooks/Cole Statistics/Probability Series)

[12] Dempster, Arthur P.: A high dimensional two sample significance test. In: *Ann. Math. Statist.* 29 (1958), S. 995–1010

[13] Dempster, Arthur P.: A significance test for the separation of two highly multivariate small samples. In: *Biometrics* 16 (1960), S. 41–50

[14] Deuflhard, Peter ; Hohmann, Andreas: *Numerische Mathematik. 1.* Third. Berlin : Walter de Gruyter & Co., 2002 (de Gruyter Lehrbuch. [de Gruyter Textbook]). – Eine algorithmisch orientierte Einführung. [An algorithmically oriented introduction]

[15] Fischer, Gerd: *Grundkurs Mathematik [Foundational Course in Mathematics]*. Bd. 17: *Lineare Algebra*. Fifth. Braunschweig : Friedr. Vieweg & Sohn, 1979. – In collaboration with Richard Schimpl

[16] Geisser, Seymour ; Greenhouse, Samuel W.: An extension of Box's results on the use of the F distribution in multivariate analysis. In: *Ann. Math. Statist.* 29 (1958), S. 885–891

[17] Greenhouse, Samuel W. ; Geisser, Seymour: On methods in the analysis of profile data. In: *Psychometrika* 24 (1959), S. 95–112

[18] Hotelling, H.: The generalization of Student's ratio. In: *Ann. Math. Statistics* 2 (1931), S. 360–378

[19] Krishnamoorthy, K. ; Yu, Jianqi: Modified Nel and van der Merwe test for the multivariate Behrens-Fisher problem. In: *Statist. Probab. Lett.* 66 (2004), Nr. 2, S. 161–169

[20] Lehmann, Erich L.: *Testing statistical hypotheses*. John Wiley & Sons Inc., 1959

[21] Mathai, A. M. ; Provost, Serge B.: *Statistics: Textbooks and Monographs*. Bd. 126: *Quadratic forms in random variables*. New York : Marcel Dekker Inc., 1992. – Theory and applications

[22] Muirhead, Robb J.: *Aspects of multivariate statistical theory*. New York : John Wiley & Sons Inc., 1982. – Wiley Series in Probability and Mathematical Statistics

[23] Niedokos, Edward: On mathematical models of split-plot design. In: *Ann. Univ. Mariae Curie-Sklodowska Sect. A* 18 (1964), S. 123–136 (1967)

[24] Potthoff, Richard F. ; Roy, S. N.: A generalized multivariate analysis of variance model useful especially for growth curve problems. In: *Biometrika* 51 (1964), S. 313–326

[25] Roebruck, P.: Canonical forms and tests of hypotheses. I. The general univariate mixed model. In: *Statist. Neerlandica* 36 (1982), Nr. 2, S. 63–74

[26] Satterthwaite, Franklin E.: Synthesis of variance. In: *Psychometrika* 6 (1941), S. 309–316

[27] Scheffé, Henry: *The analysis of variance*. New York : John Wiley & Sons Inc., 1999 (Wiley Classics Library). – Reprint of the 1959 original, A Wiley Publication in Mathematical Statistics

[28] Srivastava, Muni S. ; Du, Meng: A test for the mean vector with fewer observations than the dimension. In: *J. Multivariate Anal.* 99 (2008), Nr. 3, S. 386–402

[29] Srivastava, Muni S. ; Fujikoshi, Yasunori: Multivariate analysis of variance with fewer observations than the dimension. In: *J. Multivariate Anal.* 97 (2006), Nr. 9, S. 1927–1940

[30] Stępniak, Czesław: Simultaneous canonization of linear models. In: *Comm. Statist. Theory Methods* 36 (2007), Nr. 13-16, S. 2405–2412

[31] Werner, Carola: *Dimensionsstabile Approximation für Verteilungen von zufälligen quadratischen Formen im Repeated-Measures-Design*, Inst. für Mathematische Stochastik, Universität Göttingen, Diplomarbeit, 2002

[32] Zhang, Fuzhen: *Matrix theory.* New York : Springer-Verlag, 1999 (Universitext). – Basic results and techniques