Ein SAS-Macro zur multivariaten nichtparametrischen Analyse bei gleichgerichteten Alternativen

Thomas Bregenzer
Institut für Mathematik und Datenverarbeitung in der Medizin, Universitätskrankenhaus Eppendorf, Hamburg

Zusammenfassung

Zur Analyse multivariater Probleme bei gleichgerichteten Alternativen eignen sich besonders (nichtparametrische) Summentests, die sich gegenüber den klassischen Verfahren (z.B. Hotellings T^2) durch höhere Power auszeichen. Dabei bereiten auch unvollständige Beobachtungsvektoren keine Schwierigkeiten, ebensowenig wie gemischte Vektoren aus metrischen und ordinalen kategoriellen Daten. Ein Macro in SAS/IML ermöglicht eine bequeme Anwendung verschiedener Summentests in parametrischer und nichtparametrischer Form.

1 Einleitung

Werden zur Beurteilung eines Behandlungseffekts mehrere gleichbedeutende Zielvariablen herangezogen, so kann oft nur dann von einer eindeutigen Überlegenheit eines Verfahrens gesprochen werden, wenn sich bei allen relevanten Zielgrößen der gewünschte Effekt in der gewünschten Richtung nachweisen läßt. Univariate statistische Methoden führen so zu einer Vielzahl von p–Werten, wohingegen ein einzelner p–Wert zur Beurteilung eines Behandlungseffekts von Vorteil wäre. Testverfahren zum Aufdecken dieser gleichgerichteten Alternativen im Fall zweier unverbundener Stichproben wie z.B. die Summenstatistiken nach O'Brien [12] und Wei–Lachin[8, 13], sind derzeit Gegenstand intensiver Diskussionen.

Unter Voraussetzung normalverteilter Daten mit gleichen Varianzen in den Gruppen ist der klassische T^2–Test von Hotelling zum Test auf beliebige Abweichungen von der Nullhypothese *„in keiner der Zielgrößen ein Behandlungseffekt“* gleichmäßig bester unverfälschter Test; zum Aufdecken gleichgerichteter Alternativen ist er aber gerade wegen seiner Eigenschaften als Omnibus–Test nicht geeignet. Hierzu lassen sich mit Summenstatistiken, gebildet durch gewichtetes oder ungewichtetes Aufsummieren der Mittelwertsdifferenzen über die Zielvariablen, Tests mit größerer Power finden [4].

2 Summenstatistiken von O'Brien–Typ

Sei $X_{ij}^{(k)}, i = 1,2; j = 1,\ldots,n_i; k = 1,\ldots,K$ die k–te gemessene Größe bei Patient j in Gruppe i, wobei einige der $X_{ij}^{(k)}$ fehlen können (*missing at random*). Die Vektoren $\mathbf{X}_{ij}{}' = (X_{ij}^{(1)},\ldots,X_{ij}^{(K)})$ sollen folgende Bedingungen erfüllen:

1. Die $\mathbf{X}_{ij}$ sind unabhängige Zufallsvektoren
2. $E(\mathbf{X}_{ij}) = \boldsymbol{\mu}_i = (\mu_{i1},\ldots,\mu_{iK})'$
3. $cov(\mathbf{X}_{ij}) = \boldsymbol{\Sigma}_i$ (nicht singulär), $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 =: \boldsymbol{\Sigma} = (\sigma_{kl})_{k,l=1,\ldots,K}$

$N := n_1 + n_2$, und n_{ik} bezeichne die Anzahl vorhandener Werte in Gruppe i für die k–te Variable. Die Summen bzw. Mittelwerte $X_{i\cdot}^{(k)} := \sum_j X_{ij}^{(k)}$, $\overline{X}_{i\cdot}^{(k)} := \frac{1}{n_{ik}} \sum_j X_{ij}^{(k)}$ werden dabei stets aus den *vorhandenen* Beobachtungen gebildet.

2.1 Konsistente Schätzer von $\boldsymbol{\mu}_i$ und $\boldsymbol{\Sigma}_i$

Ein konsistenter Schätzer für $\boldsymbol{\mu}_i$ ergibt sich komponentenweise aus $\overline{\boldsymbol{X}}_{i\cdot}$, wobei jede Komponente aus den n_{ik} vorhandenen Beobachtungen geschätzt wird. Die Kovarianz–Matrix $\boldsymbol{\Sigma}_i$ kann geschätzt werden durch $\widehat{\boldsymbol{\Sigma}}_i := (\widehat{\sigma}_{ikl})_{k,l}$ mit

$$\widehat{\sigma}_{ikl} = \sum_{j=1}^{n_i} (X_{ij}^{(k)} - \overline{X}_{i\cdot}^{(k)})(X_{ij}^{(l)} - \overline{X}_{i\cdot}^{(l)})/(n_{ikl} - 2 + 1_{\{k=l\}}),$$

der gepoolte Schätzer von $\boldsymbol{\Sigma}$ ist $\widehat{\boldsymbol{\Sigma}} = (\widehat{\sigma}_{\cdot kl})_{k,l}$ mit

$$\widehat{\sigma}_{\cdot kl} = \frac{\sum_{i=1}^{2} (n_{ikl} - 1)\widehat{s}_{ikl}}{\sum n_{ikl} - 2}.$$

Die Kovarianz–Matrix des Mittelwertvektors $\overline{\boldsymbol{X}}_{i\cdot}$ ist $cov(\overline{\boldsymbol{X}}_{i\cdot}) = \left(cov(\overline{X}_{i\cdot}^{(k)}, \overline{X}_{i\cdot}^{(l)})\right)_{k,l} =: (\omega_{ikl})_{k,l}$ mit $\omega_{ikl} = \frac{n_{ikl}}{n_{ik}n_{il}}\sigma_{ikl} = \frac{n_{ikl}^2}{n_{ik}n_{il}} cov^*(\overline{X}_{i\cdot}^{(k)}, \overline{X}_{i\cdot}^{(l)})$ wobei $cov^*(\cdot)$ die *bedingte* Kovarianz ist, basierend auf den n_{ikl} Beobachtungen mit vorhandenen Daten für beide Variablen. Der Faktor $\frac{n_{ikl}^2}{n_{ik}n_{il}}$ spielt dabei die Rolle eines *Korrektur–Faktors bei fehlenden Werten* mit der Eigenschaft, daß die so bestimmte „bedingte" Kovarianz–Matrix positiv semidefinit ist. Der Korrekturfaktor kann Werte zwischen 0 und 1 annehmen: 1, wenn die Daten vollständig sind, und 0, wenn jedem beobachteten Wert ein fehlender „gegenüberliegt". In letzterem Fall werden die Variablen also als *bedingt unabhängig* betrachtet. Die so unter Ausnutzung aller verfügbaren Daten konstruierten $\overline{\boldsymbol{X}}_{i\cdot}$ und $\widehat{\boldsymbol{\Sigma}}_i$ sind konsistente Schätzer für $\boldsymbol{\mu}_i$ und $\boldsymbol{\Sigma}_i$.

2.2 Asymptotische Verteilung

Die Mittelwerts–Differenzvektoren der beiden Behandlungsgruppen, $(\overline{X}_{1\cdot}^{(k)} - \overline{X}_{2\cdot}^{(k)})_{k=1,\ldots,K}$, sind ein konsistenter Schätzer für den Effekt $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Indem man zusätzlich zu den Vektoren der Beobachtungen $X_{ij}^{(k)}$ die Vektoren der Indikatorvariabeln, die ja beschreiben, ob ein Wert vorhanden ist oder fehlt, als Zufallsvariablen, stochastisch unabhängig von den $X_{ij}^{(k)}$, einführt, kann ein multivariater Zentraler Grenzwertsatz angewandt werden um die asymptotische Normalität der standardisierten Mittelwertsvektoren nachzuweisen:

$$\sqrt{N}(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \xrightarrow{\mathcal{D}} \mathbf{U} \sim N_K(\mathbf{0}, \boldsymbol{\Gamma})$$

mit $\boldsymbol{\Gamma} = (\gamma_{kl})_{k,l}$ und

$$\gamma_{kk} = N\sigma_{\cdot kk}\left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}}\right), \quad k = 1, \ldots, K \quad ; \quad \gamma_{kl} = N\sigma_{\cdot kl}\left(\frac{n_{1kl}}{n_{1k}n_{1l}} + \frac{n_{2kl}}{n_{2k}n_{2l}}\right), \quad k \neq l.$$

$\boldsymbol{\Gamma}$ kann konsistent aus $\widehat{\boldsymbol{\Sigma}}$ geschätzt werden.

3 Multivariate Teststatistiken für unvollständige Daten

3.1 Parametrische Tests für gleichgerichtete Alternativen

O'Brien schlägt zwei parametrische Verfahren zum Test von $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = c\mathbf{J}$ vor, dabei ist c ein reellwertiger Skalar und $\mathbf{J}$ ein Einservektor mit K Komponenten.[1] Die modifizierten OLS– und GLS–Statistiken — ihre Namen beziehen sich auf *ordinary least squares* und *generalized least squares*– Techniken —

[1] O'Brien setzt standardisierte Variablen voraus; im allgemeinen Fall wird also $\mathbf{J}$ durch den Vektor mit den Inversen der Standardabweichungen zu ersetzen sein.

sind nun

$$\boxed{T_{OLS} = \frac{\mathbf{J}'(\overline{\boldsymbol{X}}_{1\cdot} - \overline{\boldsymbol{X}}_{2\cdot})}{(\mathbf{J}'\widehat{\boldsymbol{\Gamma}}\mathbf{J})^{1/2}} \quad \text{und} \quad T_{GLS} = \frac{\mathbf{J}'\widehat{\boldsymbol{\Gamma}}^{-1}(\overline{\boldsymbol{X}}_{1\cdot} - \overline{\boldsymbol{X}}_{2\cdot})}{(\mathbf{J}'\widehat{\boldsymbol{\Gamma}}^{-1}\mathbf{J})^{1/2}},}$$

wobei $\overline{\boldsymbol{X}}_{1\cdot}$, $\overline{\boldsymbol{X}}_{2\cdot}$ und $\widehat{\boldsymbol{\Gamma}}$ aus den vorhandenen Beobachtungen bestimmt werden, wie im vorigen Abschnitt gezeigt.

Beide Statistiken sind asymptotisch standardnormalverteilt unter H_0.

3.2 Eine allgemeine lineare Rangstatistik im unverbundenen a–Stichproben–Fall

O'Brien schlägt auch einen nichtparametrischen Test vor, der sich allerdings schon in ähnlicher Form bei Koziol et al 1981 [6] findet. Hier wird nun ein allgemeines nichtparametrisches Verfahren für den Fall von a unabhängigen Stichproben angegeben, zunächst bei vollständigen Daten:

3.2.1 Modell und Effekte

$\mathbf{X} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(K)})', i = 1, \ldots, a; j = 1, \ldots, n_i$ bezeichne die nach $\mathbf{F}_i$ verteilten unabhängigen Zufallsvektoren mit K Komponenten in a Gruppen. Die Randverteilungen $F_i^{(k+)} := P(X_{ij}^{(k)} \leq x)$ werden nicht notwendig als stetig vorausgesetzt. Im weiteren Verlauf wird allerdings eine „normalisierte" Version der Verteilungsfunktionen verwendet: Bezeichnet $F_i^{(k+)}$ wie oben gezeigt die (üblicherweise gebräuchliche) rechtsstetige Version der Verteilungsfunktion und $F_i^{(k-)} := P(X_{ij}^{(k)} < x)$ die linksstetige Version, so wird mit

$$F_i^{(k)}(x) := \frac{1}{2}\left[F_i^{(k+)} + F_i^{(k-)}\right]$$

die gemittelte („normalisierte") Version der Verteilungsfunktion bezeichnet. Die Verwendung dieser Verteilungsfunktion hat den Vorteil, daß auch Bindungen und sogar ordinale kategorielle Daten zugelassen werden können. Ausgeschlossen sollen lediglich Einpunktverteilungen sein. Die empirischen Verteilungsfunktionen

$$\widehat{F}_i^{(k)} := \frac{1}{n_i}\sum_{j=1}^{n_i} c(x - X_{ij}^{(k)})$$

seien ebenfalls mit normalisierten Versionen der Zählfunktion c gebildet. Die ($a \cdot K$–Komponenten–) Vektoren der Verteilungsfunktionen werden mit

$$\mathbf{F} = (F_i^{(k)})'_{\substack{i=1,\ldots,a \\ k=1,\ldots,K}} \text{ bzw. } \widehat{\mathbf{F}} = (\widehat{F}_i^{(k)})'_{\substack{i=1,\ldots,a \\ k=1,\ldots,K}}$$

bezeichnet.

Das gewichtete Mittel der Verteilungsfunktionen ist

$$\begin{aligned} H^{(k)} &= \sum_{i=1}^{a} \frac{n_i}{N} F_i^{(k)} \\ \text{bzw.} & \\ \widehat{H}^{(k)} &= \sum_{i=1}^{a} \frac{n_i}{N} \widehat{F}_i^{(k)} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i} c(x - X_{ij}^{(k)}) \end{aligned}$$

mit $N = \sum_{i=1}^{a} n_i$.

3.2.2 Die nichtparametrischen Effekte

Der Erwartungswert der Zufallsvariablen X_{ij} (auf den Index k für die Variablennummer im oben betrachteten Vektor soll zunächst verzichtet werden) $E(X_{ij}) = \int x\, dF_i(x)$ läßt sich als Spezialfall eines „verallgemeinerten Erwartungswerts“ mit beliebiger Integrandenfunktion h (eben für $h = id$) $E_h(X_{ij}) := \int h(x)\, dF_i(x)$ auffassen. Im rein nichtparametrischen Kontext stehen neben der Identität nicht viele „sinnvolle“ Funktionen für h zur Auswahl, im Grunde nur die Verteilungsfunktionen der zugrundeliegenden Daten. Eine Möglichkeit wäre $\int F_{i'}(x)\, dF_i(x) \quad i \neq i'$, die Proversionswahrscheinlichkeiten $P(X_{ij} \leq X_{i'j})$ also. Da aber nicht a priori eine Gruppe vor den anderen ausgezeichnet sein muß, erscheint es sinnvoll, das gewichtete Mittel der Verteilungsfunktionen zu verwenden, womit sich, vgl. [3], die (i–ten) *relativen (Behandlungs–)Effekte*

$$p_i := \int H(x)\, dF_i(x)$$

als gemittelte Version der Proversionswahrscheinlichkeiten ergeben. Läßt man zusätzlich noch eine Scorefunktion J zu, so erhält man die (i–ten) allgemeinen relativen Effekte

$$p_i(J) := \int J[H(x)]\, dF_i(x).$$

Im multivariaten Kontext, also bei K–komponentigen Zufallsvektoren $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(K)})'; i = 1, \ldots, a, j = 1, \ldots, n_i$ mit Verteilungsfunktionen $\mathbf{F}_i$ und Randverteilungen $F_i^{(k)}$ betrachtet man nun also die *Vektoren* (bzw. deren Schätzer) der relativen Effekte

$$\begin{aligned} p_i^{(k)}(J^{(k)}) &:= \int J^{(k)}[H^{(k)}(x)]\, dF_i^{(k)}(x) \\ \widehat{p}_i^{(k)}(J^{(k)}) &:= \int J^{(k)}[\widehat{H}^{(k)}(x)]\, d\widehat{F}_i^{(k)}(x). \end{aligned}$$

Dabei setzten sich die Schätzer aus den durch variablenweise Rangvergabe gebildeten *Rängen* ($R_{ij}^{(k)}$) (bzw. Rangscores) zusammen, denn

$$J^{(k)}(\widehat{H}^{(k)}(X_{ij}^{(k)})) \stackrel{!}{=} J^{(k)}\left(\frac{R_{ij}^{(k)} - \frac{1}{2}}{N}\right).$$

Da das in Abschnitt 5 beschriebene SAS–Macro nur Wilcoxon–Scores verwendet, wird im Abschnitt 3.3 über direktionale Tests auf die Scorefunktionen $J^{(k)}$ nicht weiter eingegangen und $J^{(k)} = id$ vereinbart.

3.2.3 Hypothesen im nichtparametrischen Modell

Im nichtparametrischen Modell werden die Verteilungsfunktionen analog zur ANOVA additiv zerlegt, so z.B. im 2–faktoriellen Modell die Verteilungsfunktionen F_{ij} der Zufallsgrößen X_{ijm} in

$$F_{ij} = M + A_i + B_j + C_{ij}$$

mit den Nebenbedingungen: $\sum_i A_i = \sum_j B_j = \sum_i C_{ij} = \sum_j C_{ij} = 0$ (wobei 0 hier die konstante Nullfunktion bezeichnet). Mit den wie oben gebildeten Mitteln der Verteilungsfunktionen

$$\overline{F}_{i\cdot} = \frac{1}{b}\sum_j F_{ij}, \quad \overline{F}_{\cdot j} = \frac{1}{a}\sum_i F_{ij}, \quad \overline{F}_{\cdot\cdot} = \frac{1}{ab}\sum_i\sum_j F_{ij}$$

läßt sich nun so z.B. die rein nichtparametrische Hypothese „kein Effekt von A“ formulieren:

$$H_0^{(A)} : A_i = \overline{F}_{i\cdot} - \overline{F}_{\cdot\cdot} = 0 \ \ \forall\ i = 1, \ldots, a$$

Allgemeine Hypothesen lassen sich nun mittels einer Kontrastmatrix $\mathbf{C}$ formulieren als $H_0(C) : \mathbf{CF} = \mathbf{0}$, z.B. ist im hier besprochenen Einfaktor–Modell im Fall $K = 1$ (univariat) $\mathbf{C} = \mathbf{P}_a := \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$, im multivariaten Fall $\mathbf{C} = \mathbf{P}_a \otimes \mathbf{1}_K$.

3.2.4 Asymptotik und Teststatistiken

Durch Anwendung geeigneter Grenzwertsätze [2] läßt sich die multivariate Normalität von $\widehat{\mathbf{p}}(J) =: \widehat{\mathbf{p}}$ unter $H_0(C) : \mathbf{CF} = \mathbf{0}$ zeigen, sofern folgende Bedingungen erfüllt sind:

(A1) $\min_{\{i=1,\ldots,a\}} n_i \to \infty$,

(A2) $0 < \lambda_0 \leq n_i/N \leq 1 - \lambda_0 < 1$

(A3) $J^{(k)} : [0,1] \to I\!R$ sind beschränkte Scorefunktionen.

$N = \sum_{i=1}^{a} n_i$ ist die Anzahl der Beobachtungs*vektoren*.

Für die asymptotischen Betrachtungen ist es sinnvoll, zunächst nicht $\widehat{\mathbf{p}}$ zu betrachten, sondern eine aus pro Variable *unabhängigen* Komponenten zusammengesetzte *Vergleichsstatistik*, der *asymptotischen Rangtransformation* $\overline{\boldsymbol{Y}}_{\cdot}$ mit Komponenten

$$\int J^{(k)}[H^{(k)}]\, d\widehat{F}_i^{(k)}, i = 1, \ldots, a; k = 1, \ldots, K$$

und $Y_{ij}^{(k)} := J^{(k)}(H^{(k)}(X_{ij}^{(k)}))$, von der sich zeigen läßt, daß unter $H_0 : \mathbf{CF} = \mathbf{0}$ die Statistiken $\mathbf{C}\widehat{\mathbf{p}}$ und $\mathbf{C}\overline{\boldsymbol{Y}}_{\cdot}$ asymptotisch äquivalent sind. Grenzwertaussagen für $\mathbf{C}\overline{\boldsymbol{Y}}_{\cdot}$ gelten somit auch für $\mathbf{C}\widehat{\mathbf{p}}$.

Die Covarianzmatrix von $\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}$ ist wegen der Unabhängigkeit der $\overline{Y}_{i\cdot}^{(k)}$ und $\overline{Y}_{i'\cdot}^{(k)}$ für $i \neq i'$, die durch die asymptotische Rangtransformation aus den Ursprungsdaten erhalten blieb, eine Kroneckersumme

$$\mathrm{Cov}(\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}) =: \mathbf{V} = \bigoplus_{i=1}^{a} \mathbf{V}_i \quad ,$$

wobei $\mathbf{V}_i$ in der klassischen Weise geschätzt wird.

Mit der zusätzlichen Voraussetzung

(B1) Sei $\lambda_{\min}$ der kleinste Eigenwert der Covarianzmatrix $\mathbf{V}$ von $\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}$. Dann gelte mit einer Konstanten λ_0: $|\lambda_{\min}| \geq \lambda_0 > 0$.

läßt sich zeigen

THEOREM 3.1 (Multivariate NV der Rang(score–)statistik) *Unter den Bedingungen (A1) bis (A3) und (B1) ist* $\sqrt{N}\mathbf{C}\widehat{\mathbf{p}}$ *unter* $H_0 : \mathbf{CF} = \mathbf{0}$ *multivariat normalverteilt mit Erwartungswertvektor* $\mathbf{0}$ *und Covarianzmatrix* $\mathbf{V}$.

3.3 Nichtparametrische Tests bei gleichgerichteten Alternativen

Die im vorigen Abschnitt gezeigte multivariate Normalität (unter Hypothese) der allgemeinen linearen Rangstatistik $\mathbf{C}\widehat{\mathbf{p}}$ läßt ich zur Konstruktion geeigneter nichtparametrischer Tests bei gleichgerichteten Alternativen heranziehen. Dazu sei zunächst nur der Fall von vollständige Daten betrachtet; eine Erweiterung auf Daten mit Fehlstellen findet sich am Ende des Abschnitts.

Dem Test der Nullhypothese $F_1^{(k)} = F_2^{(k)} = \ldots = F_a^{(k)}; k = 1, \ldots, K$ entspricht in Matrixnotation

$$\mathbf{K}_a\mathbf{F} = \mathbf{0}$$

mit der $(a \cdot K \times a \cdot K)$–Matrix $\mathbf{K}_a := \mathbf{P}_a \otimes \mathbf{I}_K$.

Bei Tests auf strukturierte Alternativen im Sinne von *patterned alternatives*, vgl. [1], mit einer $(K \times a \cdot K)$–Gewichtsmatrix $\mathbf{W}$ können prinzipiell für jede Variable eigene Gewichtsvektoren $\mathbf{w}^{(k)}$ für die (Behandlungs–)

Gruppen festgelegt werden (wenn man beispielsweise beim Vergleich zwischen Behandlungs– und Placebogruppe bei einer Variablen einen Anstieg und bei einer anderen einen Abfall erwartet); $\mathbf{W}$ hat dann die Form

$$\mathbf{W} = \begin{pmatrix} w_1^{(1)} & 0 & \dots & 0 & \dots & w_a^{(1)} & 0 & \dots & 0 \\ 0 & w_1^{(2)} & 0 & \dots & \dots & 0 & w_a^{(2)} & 0 & \dots \\ \vdots & & & \vdots & \vdots & \vdots & & & \vdots \\ 0 & \dots & 0 & w_1^{(K)} & \dots & 0 & \dots & 0 & w_a^{(K)} \end{pmatrix} \quad .$$

Oft wird allerdings auch die Situation $\mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(K)} = \mathbf{w}$ vorliegen; dann kann $\mathbf{W}$ einfach als $\mathbf{W} = \mathbf{w}' \otimes \mathbf{I}_K$ geschrieben werden, und eine a–priori erwartete Ordnung der $F_1^{(k)}, \ldots, F_a^{(k)}$ kann durch einen a–komponentigen Gewichtsvektor $\mathbf{w}$ repräsentiert werden. Rechnet man beispielsweise mit einem Anstieg der outcome–Variaben von Dosisstufe I über II nach III und bei IV wieder mit einem Abfall, so läßt sich dies mit dem Gewichtsvektor $(1, 2, 4, 3)'$ charakterisieren. [1]

Desweiteren kann nun auch für die K Variablen im multivariaten Problem ein beliebiger Gewichtsvektor $\mathbf{v} = (v_1, \ldots, v_K)'$ angegeben werden, wie er bei den Summenstatistiken, vgl. [12, 13] oder [10] im parametrischen Kontext, Verwendung findet, so z.B.

- $\mathbf{v} = \mathbf{1}_K$ bei O'Briens OLS–Test (die Standardisierung durch Division mit den Standardabweichungen der Variablen kann hier entfallen, da die Variabeln durch die variablenprobenweise Rangvergabe schon „standardisiert" sind.) oder
- $\mathbf{v} = \widehat{\mathbf{\Sigma}}_R^{(-1)} \mathbf{1}_K$ mit einem Schätzer $\widehat{\mathbf{\Sigma}}_R^{(-1)}$ der Covarianzmatrix der (rangtransformierten) Variablen oder
- $\mathbf{v} = d(\mathbf{\Lambda})$ mit einer eindeutigen vektorwertigen Funktion d der Produktsummenmatrix $\mathbf{\Lambda}$ der Daten, vgl. [9, 10].

Mit Gewichtsmatrix $\mathbf{W}$ für die strukturierte Alternativen und Gewichtsvektor $\mathbf{v}$ für die Bildung der Summenstatistik kann nun eine lineare Rangstatistik

$$T_N := \sqrt{N} \mathbf{v}' \mathbf{W}' \mathbf{K}_a \widehat{\mathbf{p}} \tag{3.1}$$

gebildet werden, die sich als

$$T_N = \sqrt{N} \sum_{k=1}^{K} v_k \sum_{i=1}^{a} (w_i^{(k)} - \overline{w}^{(k)}) \overline{R}_{i\cdot}^{(k)} \tag{3.2}$$

schreiben läßt. Unter Hypothese ist nun T wieder asymptotisch normalverteilt mit Erwartungswert 0 und Varianz

$$\sigma^2 = \mathbf{v}' \mathbf{W}' \mathbf{K}_a \mathbf{V} \mathbf{K}_a \mathbf{W} \mathbf{v} \quad .$$

Die Covarianzmatrix $\mathbf{V}$ kann wieder konsistent aus den Rängen geschätzt werden. Dabei ergibt sich aber bei hochdimensionalen Versuchen das Problem, daß sehr viele Parameter zu schätzen sind. Ist $\mathbf{W} = \mathbf{w} \otimes \mathbf{I}_K$ (was in der Praxis oft der Fall sein dürfte [2]) ist die Schätzung der gesamten Covarianzmatrix bei der Verwendung der Summenstatistik nicht nötig, da nur die Varianz der linearen Rangstatistik bestimmt werden muß, und die läßt sich dann schreiben als

$$T_N \quad = \quad \sqrt{N} \sum_{k=1}^{K} v_k \sum_{i=1}^{a} (w_i - \overline{w}) \overline{R}_{i\cdot}^{(k)}$$

[1] Da die Gewichtsvektoren im Prinzip völlig frei wählbar sind, wurde bei der Bezeichnung der Alternativen der Begriff „strukturiert" gewählt; „geordnete Alternativen" wäre auch möglich gewesen, bezeichnet aber konventionell eher einen etwas eingeschränkteren Sachverhalt.

[2] d.h. für jede Variable ist — bezüglich der Stichproben — der gleiche Gewichtsvektor anzuwenden; dies ist der „klassische" Fall in anscheinend allen vorliegenden Arbeiten zum Thema „Summenstatistiken", z.B. bei [9, 12, 13]

$$= \sqrt{N}\sum_{i=1}^{a}(w_i-\overline{w})\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{K}v_k R_{ij}^{(k)}$$
$$= \sqrt{N}\sum_{i=1}^{a}(w_i-\overline{w})\frac{1}{n_i}\sum_{j=1}^{n_i}R_{ij}^{(\mathbf{v})}$$

mit $R_{ij}^{(\mathbf{v})} := \sum_{k=1}^{K} v_k R_{ij}^{(k)}$. Dann ist

$$\widehat{\sigma}^2 = \sum_{i=1}^{a}(w_i-\overline{w})^2\widehat{\sigma}_i^2 \quad , \tag{3.3}$$

und

$$\widehat{\sigma}_i^2 = \frac{N}{n_i(n_i-1)}\sum_{j=1}^{n_i}(R_{ij}^{(\mathbf{v})}-\overline{R}_{i\cdot}^{(\mathbf{v})})^2 \quad . \tag{3.4}$$

Zusammenhang mit O'Briens Testverfahren Die von O'Brien in [12] empfohlene nichtparametrische Variante seiner OLS–Statistik zum Test auf gleichgerichtete Unterschiede zwischen den Stichproben beruht, übertragen auf die Notation dieses Beitrags, auf der standardisierten Mittelwertsdifferenz der pro unabhängiger Beobachtung (j) gebildeten Rangsumme (über die K Variablen bei variablenweiser Rangvergabe), also (mit den $\overline{R}_{i\cdot}^{(\mathbf{v})}$ aus dem vorigen Absatz) auf der Teststatistik

$$T_{OB} = \overline{R}_{1\cdot}^{(\mathbf{v})} - \overline{R}_{2\cdot}^{(\mathbf{v})} \quad .$$

Der Vorschlag O'Briens, mit den Rangsummen $\overline{R}_{i\cdot}^{(\mathbf{v})}$ einen zwei–Stichproben–t-Test auszuführen, entspricht also genau dem Fall $a=2, \mathbf{v}=\mathbf{1}_K, \mathbf{W}=\mathbf{w}'\otimes\mathbf{I}_K, \mathbf{w}=(1,-1)'$ im hier besprochenen Test zu

$$H_0 : (\mathbf{K}_a)\mathbf{F}=\mathbf{0} \quad \Leftrightarrow \quad H_0 : F_1^{(k)} = F_2^{(k)}, k=1,\ldots,K$$
$$\text{vs.}$$
$$H_{11} : F_1^{(k)} > F_2^{(k)}, k=1,\ldots,K \quad \text{bzw.} \quad H_{12} : F_1^{(k)} < F_2^{(k)}, k=1,\ldots,K \quad ,$$

denn

$$\overline{R}_{1\cdot}^{(\mathbf{v})} - \overline{R}_{2\cdot}^{(\mathbf{v})} \stackrel{!}{=} N\mathbf{v}'\mathbf{W}'\mathbf{K}_a\widehat{\mathbf{p}} \quad ,$$

$\mathbf{K}_a = \mathbf{P}_a\otimes\mathbf{I}_K$, und $\widehat{\mathbf{p}}$ setzt sich aus den variablenweise vergebenen Rängen (bzw. Scores) zusammen:

$$\widehat{\mathbf{p}} = \left(\frac{1}{n_i}\sum_{j=1}^{n_i} J\left[\frac{1}{N}\left(R_{ij}^{(k)}-\frac{1}{2}\right)\right]\right)_{i=1,\ldots,a;k=1,\ldots,K} \quad ,$$

wobei hier J als die Identität zu wählen ist. Da die Rangvergabe pro Variable schon standardisierte Größen bildet, wird auf die bei dem parametrischen OLS–Test erforderliche Transformation der Originaldaten (Subtraktion des Gesamtmittels der Variablen und nachfolgend Division durch die Standardabweichungen) verzichtet. Analog zum parametrischen GLS–Test läßt sich hier ein Test mit $\mathbf{v}=\widehat{\mathbf{\Sigma}}_R^{(-1)}\mathbf{1}_K$ mit einem Schätzer $\widehat{\mathbf{\Sigma}}_R^{(-1)}$ der Rangkorrelationsmatrix der (rangtransformierten) Variablen konstruieren, der von O'Brien im nichtparametrischen Kontext nicht vorgeschlagen wurde.

Es ist noch zu erwähnen, daß die Teststatistik (3.1) für beliebig verteilte Zufallsvariablen anwendbar ist, also auch auf ordinale kategorielle Daten und somit auf gemischte Variablenvektoren mit sowohl ordinalen als auch metrischen Daten. Darüberhinaus können beliebige Gewichtsvektoren bzw. -Matrizen sowohl für die K Variablen als auch für die a Gruppen abgegeben werden, und wenn bei allen Variablen derselbe Gewichtsvektor für die Gruppen Verwendung findet (also wie bei O'Brien), kann die Zahl der abhängigen Variablen (K) beliebig groß sein und die Zahl der unabhängigen Beobachtungen (N) überschreiten, da die Schätzung der gesamten Covarianzmatrix nicht nötig ist und durch die einfache Schätzung der Varianz der linearen Rangstatistik nach (3.3) ersetzt werden kann.

Erweiterung bei unvollständigen Daten Bei unvollständigen Daten kann das vorgestellte Verfahren analog zum Vorgehen beim parametrischen Fall erweitert werden. Die empirische Verteilungsfunktion für die k–te Variable ist nun mit n_{ik} als Anzahl der für Variable k in Gruppe i vorhandenen Beobachtungen

$$\widehat{F}_i^{(k)} := \frac{1}{n_{ik}} \sum_{\text{beob. } j} c(x - X_{ij}^{(k)}) \quad .$$

Die gemittelten Verteilungsfunktionen $H^{(k)}$ sind nun

$$\begin{aligned} H^{(k)} &= \sum_{i=1}^{a} \frac{n_{ik}}{N_k} F_i^{(k)} \\ \text{bzw.} & \\ \widehat{H}^{(k)} &= \sum_{i=1}^{a} \frac{n_{ik}}{N_k} \widehat{F}_i^{(k)} = \frac{1}{N_k} \sum_{i=1}^{a} \sum_{\text{beob. } j} c(x - X_{ij}^{(k)}) \end{aligned}$$

mit $N_k = \sum_{i=1}^{a} n_{ik}$. Damit werden wieder die nichtparametrischen Effekte definiert und im asymptotischen Verhalten untersucht. Daher müssen die Bedingungen (A1) und (A2) aus 3.2.4 nur an die unterschiedlichen Anzahlen von Beobachtungen pro Variable angepaßt werden, es gelte also für alle $k = 1, \ldots, K$:

(A1') $\min_{\{i=1,\ldots,a\}} n_{ik} \to \infty$,

(A2') $0 < \lambda_k \leq n_{ik}/N_k \leq 1 - \lambda_k < 1$

Die Resultate der Asymptotik oben bleiben so unverändert bestehen, Nur die Definition der Vergleichsstatistik muß angepaßt werden, denn nun ist mit $Y_{ij}^{(k)} := J^{(k)}(H^{(k)}(X_{ij}^{(k)}))$

$$\overline{Y}_{i\cdot}^{(k)} := \frac{1}{n_{ik}} \sum_{\text{beob. } j} Y_{ij}^{(k)} \quad ,$$

und damit wird $\overline{\boldsymbol{Y}}_{\cdot} := \left(\overline{Y}_{i\cdot}^{(k)}\right)_{i=1,\ldots,a;k=1,\ldots,K}$ gebildet. $N_k \widehat{H}^{(k)}(X_{ij}^{(k)})$ ist nun der Rang von $X_{ij}^{(k)}$ unter den N_k *beobachteten* Werten.

Die asymptotische multivariate Normalität von $\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}$ ergibt sich wie oben mit der vorangestellten Bedingung (B1); und da unter $H_0 : \mathbf{CF} = \mathbf{0}$ $\sqrt{N}\,\mathbf{C}\widehat{\mathbf{p}}$ und $\sqrt{N}\,\mathbf{C}\overline{\boldsymbol{Y}}_{\cdot}$ asymptotisch äquivalent sind, erhält man dadurch auch die asymptotische Normalität der Rangstatistik $\sqrt{N}\,\mathbf{C}\widehat{\mathbf{p}}$:

THEOREM 3.2 *Unter den Bedingungen (A1'), (A2'), (A3) und (B1) ist* $\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}$ *asymptotisch multivariat normalverteilt.*

Die Schätzung der Covarianzmatrix $\mathbf{V}$ von $\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}$ muß nun auch auf die unvollständigen Daten angepaßt werden. Zur Schätzung von $\mathrm{Cov}(\sqrt{N}Y_{ij}^{(k)}, \sqrt{N}Y_{ij}^{(l)})$ können n_{ikl} vorhandene Beobachtungen verwendet werden:

$$\widehat{\mathrm{Cov}}(Y_{ij}^{(k)}, Y_{ij}^{(l)}) = \frac{1}{n_{ikl} - 1} \sum_{\text{beob. } j} \left(Y_{ij}^{(k)} - \overline{Y}_{i\cdot}^{(k)}\right) \left(Y_{ij}^{(l)} - \overline{Y}_{i\cdot}^{(l)}\right) \quad ,$$

gebildet aus den unbeobachtbaren asymptotischen Rangtransformationen, wird aus den vorhandenen Daten geschätzt durch

$$\widehat{\widehat{\mathrm{Cov}}}(Y_{ij}^{(k)}, Y_{ij}^{(l)}) = \frac{1}{n_{ikl} - 1} \sum_{\text{beob. } j} \left(\widehat{Y}_{ij}^{(k)} - \widehat{\overline{Y}}_{i\cdot}^{(k)}\right) \left(\widehat{Y}_{ij}^{(l)} - \widehat{\overline{Y}}_{i\cdot}^{(l)}\right) \quad ;$$

und die hier eingehenden Werte sind Ränge bzw. Scores, denn für die $\widehat{Y}_{ij}^{(k)}$ gilt:

$$\widehat{Y}_{ij}^{(k)} = J^{(k)}(\widehat{H}^{(k)}(X_{ij}^{(k)})) \stackrel{!}{=} J^{(k)}\left(\frac{R_{ij}^{(k)} - \frac{1}{2}}{N_k}\right)$$

mit $N_k = \sum_{i=1}^{a} n_{ik}$. Wie im Abschnitt bei den parametrischen Verfahren dargestellt, ergibt sich auch hier bei der Berechnung bzw. den Schätzern von $\mathrm{Cov}(\widehat{\overline{Y}}_{i\cdot}^{(k)}, \widehat{\overline{Y}}_{i\cdot}^{(l)})$ ein Korrekturfaktor, sodaß die analoge Gleichung nun

$$\widehat{\mathrm{Cov}}(\sqrt{N}\,\widehat{\overline{Y}}_{i\cdot}^{(k)}, \sqrt{N}\,\widehat{\overline{Y}}_{i\cdot}^{(l)}) =: \eta_i^{(k,l)} = \frac{n_{ikl}}{n_{ik}n_{il}} \frac{N}{n_{ikl}-1} \sum_{\text{beob. } j} \left(\widehat{Y}_{ij}^{(k)} - \widehat{\overline{Y}}_{i\cdot}^{(k)}\right)\left(\widehat{Y}_{ij}^{(l)} - \widehat{\overline{Y}}_{i\cdot}^{(l)}\right)$$

lautet, die im Fall von vollständigen Daten wieder die bekannte Form hat. Damit ist jedes Element von $\mathbf{V} = \mathrm{Cov}(\sqrt{N}\,\overline{\mathbf{Y}}_{\cdot}) = \bigoplus_{i=1}^{a} \mathbf{V}_i$ mit $\eta_i^{(k,l)}$ konsistent geschätzt.

Mit einer Gewichtsmatrix $\mathbf{W}$ für die strukturierten Alternativen und einem Gewichtsvektor $\mathbf{v}$ für die Bildung der Summenstatistik kann nun wieder eine lineare Rangstatistik

$$T_N := \sqrt{N}\mathbf{v}'\mathbf{W}'\mathbf{K}_a\widehat{\mathbf{p}}$$

gebildet werden (mit $\mathbf{K}_a := \mathbf{P}_a \otimes \mathbf{I}_K$ und $\mathbf{P}_a := \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$). Unter Hypothese ist T_N dann asymptotisch normalverteilt mit Erwartungswert 0 und Varianz

$$\sigma^2 = \mathbf{v}'\mathbf{W}'\mathbf{K}_a\mathbf{V}\mathbf{K}_a\mathbf{W}\mathbf{v} \quad .$$

$\mathbf{V}$ wird wie eben angegeben aus den Rängen bzw. Scores der vorhandenen Daten geschätzt.

Die oben angegebene vereinfachte Varianzschätzung durch Schätzung der Varianz der *linearen Rangstatistik* statt der gesamten Covarianzmatrix $\mathbf{V}$ ist bei unvollständigen Daten nicht anwendbar, denn nun lassen sich die Summationen über die Variablen ($\sum_{k=1}^{K}$) und über die Beobachtungseinheiten ($\sum_{j=1}^{n_{ik}}$) nicht mehr vertauschen. Daher entfällt bei unvollständigen Daten auch die angenehme Eigenschaft der genannten Summenstatistiken, im Prinzip beliebig hochdimensionale Probleme bei relativ geringem Stichprobenumfang bearbeiten zu können: Durch die Notwendigkeit der Schätzung der gesamten Covarianzmatrix wird die Anzahl der Variablen durch die Anzahl der unabhängigen Beobachtungen begrenzt.

Die allgemeine Form der Summenstatistik im a–Stichprobenproblem zum Test von $H_0 : \mathbf{F}_1 = \ldots = \mathbf{F}_a$ gegen eine durch die Gewichtsmatrix $\mathbf{W}$ definierte strukturierte Alternative ist also

$$T_N := \sqrt{N}\frac{\mathbf{v}'\mathbf{W}'\mathbf{K}_a\widehat{\mathbf{p}}}{\sqrt{\mathbf{v}'\mathbf{W}'\mathbf{K}_a\widehat{\mathbf{V}}\mathbf{K}_a\mathbf{W}\mathbf{v}}} \xrightarrow{\mathcal{D}} U \sim N(0,1) \quad . \tag{3.5}$$

Die Schätzer $\widehat{\mathbf{p}}$ und $\widehat{\mathbf{V}}$ werden wie in oben angegeben aus den vorhandenen Daten geschätzt. Ist für alle Variablen der gleiche Gewichtsvektor $\mathbf{w}$ anwendbar zur Beschreibung der Alternative (z.B. $\mathbf{w} = (-1,1)'$ im klassischen 2-Stichprobenproblem bei [12] oder [13]), so ist zum Test von

$$H_0 : \mathbf{F}_1 = \ldots = \mathbf{F}_a \quad \text{vs.} \quad H_1 : \mathbf{F}_1 \geq \ldots \geq \mathbf{F}_a$$

(das $\geq$ ist hier komponentenweise zu verstehen) vereinfacht $\mathbf{W} = \mathbf{w}' \otimes \mathbf{I}_K$ zu verwenden, wobei $\mathbf{w}$ aus der aufsteigenden Folge von Gewichten $w_1 \leq \ldots \leq w_a$ gebildet wird. Liegen die Daten zudem noch vollständig vor, so vereinfacht sich (3.5) zu

$$\boxed{T_N = \frac{\sum_{i=1}^{a}(w_i - \overline{w})\frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij}^{(\mathbf{v})}}{\sqrt{\sum_{i=1}^{a}(w_i-\overline{w})^2\widehat{\sigma}_i^2}} \xrightarrow{\mathcal{D}} U \sim N(0,1)}$$

mit $\widehat{\sigma}_i^2 = \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}(R_{ij}^{(\mathbf{v})} - \overline{R}_{i\cdot}^{(\mathbf{v})})^2$ und $R_{ij}^{(\mathbf{v})} := \sum_{k=1}^{K} v_k R_{ij}^{(k)}$.

Trotz der Tatsache, daß die Grenzwertsätze nur eine asymptotische Normalverteilung liefern, kann man natürlich versuchen, die finiten Verteilungen annähernd durch eine geeignete t–Verteilung zu bestimmen. Dabei ergaben Simulationsstudien, daß eine t–Verteilungs–Approximation mit $N-2$ Freiheitsgraden sehr brauchbare Ergebnisse liefert (eine exakte t–Verteilung ergibt sich auch bei einer bestimmten Klasse von Summenstatistiken im Fall normalverteilter Zufallsvariablen, vgl. [9]).

3.4 Hotelling's T^2 bei unvollständigen Daten

Nun läßt sich auch eine modifizierte T^2 Statistik für unvollständige Daten angeben:

$$\boxed{T^2 := (\overline{\boldsymbol{X}}_{1\cdot} - \overline{\boldsymbol{X}}_{2\cdot})'\widehat{\boldsymbol{\Gamma}}(\overline{\boldsymbol{X}}_{1\cdot} - \overline{\boldsymbol{X}}_{2\cdot}) \quad \text{mit} \quad \widehat{\boldsymbol{\Gamma}} = \widehat{cov}(\overline{\boldsymbol{X}}_{1\cdot} - \overline{\boldsymbol{X}}_{2\cdot})}$$

wobei $\widehat{\boldsymbol{\Gamma}}$ aus allen vorhandenen Daten geschätzt wird (wie in Abschnitt 2.2. beschrieben).

T^2 ist asymptotisch χ^2 verteilt mit K Freiheitsgraden.

Eine nichtparametrische Version (Notation wie im vorangegangenen Abschnitt) von Hotellings T^2–Test, analog zu Koziol et al, 1981, ist nun

$$\boxed{T_R^2 := \widehat{\mathbf{p}}'\widehat{\boldsymbol{\Sigma}}_R\widehat{\mathbf{p}} \quad \text{mit} \quad \widehat{\boldsymbol{\Sigma}}_R = \widehat{cov}(\widehat{\mathbf{p}}).}$$

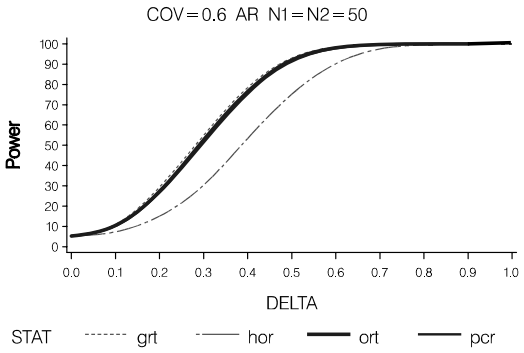
$\boldsymbol{\Sigma}_R$ und $\widehat{\mathbf{p}}$ werden wieder aus den vorhandenen Daten geschätzt.

Auch T_R^2 ist asymptotisch χ^2 verteilt mit K Freiheitsgraden.

4 Ergebnisse von Simulationen

Untersuchungen über asymptotische Optimalitätseigenschaften lassen allerdings kaum Rückschlüsse auf das Verhalten der Tests bei kleinen und mittleren Stichprobenumfängen zu. Durch Simulationen kann gezeigt werden, daß asymptotisch in gewisser Hinsicht optimale Tests im finiten Fall zu kaum mehr brauchbaren Ergebnissen führen (Antikonservativität, Güteverluste, vgl. auch [7, 11]). In die Simulationen fanden die parametrischen und nichtparametrischen Formen der O'Brien–Statistiken Eingang, ebenso wie Hotellings T^2 und, ohne daß hier näher darauf eingegasngen wurde, die Standardized–Sum– und Principal–Component–Statistik von Läuter [9]. Zu letzteren sei nur erwähnt, daß die Rangversion der Standardized–Sum–Statistik identisch ist mit der Rangversion von O'Briens OLS mit t–Verteilungs–Approximation. Insbesondere ergab sich

- Die GLS–Formen der Summenstatistiken sollten keine Verwendung finden, da die Antikonservativität auch noch bei mittleren Stichprobenumfängen recht deutlich ist. Auch der zu erwartende Vorteil von GLS gegenüber OLS stellt sich i.a. eher als marginal heraus.
- Die parametrische und nichtparametrische OLS–Form ist bei kleinen Stichprobenumfängen noch etwas antikonservativ, wobei die Anwendung einer t–Verteilungs–Approximation (statt Normalverteilung) diesem Effekt entgegenwirkt, sodaß dann die Rangversion auch bei kleinen Fallzahlen (schon bei 5 pro Gruppe) gut anwendbar ist.
- Uneingeschränkt anwendbar (im Sinne von Einhaltung des Niveaus) ist bei vollständigen Daten eine OLS–Rangvariante, die ohne Schätzung der Covarianzmatrix auskommt, denn bei vollständigen Daten ist es ausreichend, die Varianz der linearen Rangstatistik (3.1) zu bestimmen. Der resultierende Test entspricht dann einem unverbundenen 2–Stichproben–t–Test, angewandt auf die pro unabhängiger Beobachtungseinheit gebildeten Rangsummen. Mit t–Verteilungs–Approximation erhält man so eine OLS–Rangvariante, die das Niveau auch bei kleinen Stichprobenumfängen einhält.
- Bei unvollständigen Daten wird in jedem Fall eine für missings adjustierte Schätzung der gesamten Covarianzmatrizen notwendig; bei Anwendung der OLS–Rangversion mit t–Verteilungs–Approximation ist schon bei $n_1 = n_2 = 5$ keine Antikonservativität feststellbar.
- Die OLS–Rangvarianten zeigen in allen untersuchten Situationen eine Power, die den das Niveau ebenfalls einhaltenden anderen direktionalen Tests (z.B. Läuter's PC–Test [9]) mindestens vergleichbar ist. Beim Vergleich zu Hotellings T^2 zeigt sich ein umso deutlicherer Vorteil der OLS–(Rang)versionen, je höher der Anteil an informativen Variablen ist; erst bei etwa gleichviel informativen wie nicht–informativen Variablen beginnt sich dieser Vorteil umzukehren — dies dürfte dann allerdings i.a. auch kein für die Anwendung von Summenstatistiken relevantes direktionales Testproblem mehr sein.
- Bei unvollständigen Daten ist der Powerverlust bei allen betrachteten Varianten der Summenstatistiken vergleichsweise gering.



Für die praktische Anwendung sollte daher die nichtparametrische Variante von O'Briens OLS das Verfahren der Wahl sein; bei vollständigen Daten in der Fassung, bei der eine t–Verteilung mit $N-2$ Freiheitsgraden zugrundegelegt wird und nur die Varianz der linearen Rangstatistik und nicht (gruppenweise) die gesamte Covarianzmatrix geschätzt werden muß — in diesem Fall kann die Zahl der abhängigen Variablen beliebig groß sein (und die der unabhängigen Beobachtungseinheiten übertreffen), bei unvollständigen Daten mittels für missings adjustierter Schätzung der Covarianzmatrizen und wieder t–Verteilungs–Approximation.

Der Powervorteil der Summentests bei gleichgerichteten Alternativen zeigt sich gegenüber Hotellings T^2 bei allen verwendeten Varianten. In nebenstehender Abbildung sind die Ergebnisse von je 10000 Durchläufen einer Simulation mit 5 normierten normalverteilten Zufallsvariablen mit autoregressiver Struktur der Covarianzmatrix ($\rho = 0.6$) dargestellt, dabei bezeichnet Delta die Einheiten in Richtung der Winkelhalbierenden des 1. Quadranten (Tests: grt=GLS, hor=Hotelling, ort=OLS, pcr=Principal–Component; alle in Rangversion und bei Summentests mit t–Verteilungs–Approximation).

5 Das SAS–Macro

Das SAS/IML–Macro OB erlaubt die Durchführung von nichtparametrischen Summentests im unverbundenen Zweistichprobenfall zum Aufdecken gleichgerichteter Alternativen mit einigen der im Abschnitt 3.3 genannten Gewichtsvektoren. So werden die Rangversionen von OLS und GLS, jeweils mit t–Verteilungs–Approximation, ausgegeben, ebenso die nichtparametrische Version von Läuters Principal–Component–Tests. Daneben findet sich noch die Rangversion von Hotelling's T^2. Die Daten bedürfen dazu einer Aufbereitung, wie man sie

	G	N	N	N	N	N	L	L	L	L	L
P	R	R	R	R	R	R	E	E	E	E	E
A	U	P	P	P	P	P	U	U	U	U	U
T	P	_	_	_	_	_	_	_	_	_	_
N	P	T	T	T	T	T	T	T	T	T	T
R	E	3	4	5	6	7	3	4	5	6	7
001	P	2	2	0	2	2	8.3	10.4	16.9	16.80	13.1167
002	K	0	0	0	0	0	12.6	9.5	8.6	13.24	13.2400
003	P	1	1	0	0	0	9.0	8.2	9.0	6.40	9.3500
004	K	0	0	0	0	0	8.6	7.7	6.1	8.00	7.7000
005	K	0	0	0	0	0	6.5	6.1	5.1	7.70	8.0000

...

Abbildung 1: Orginaldaten (Ausschnitt)

von repeated–Analysen in PROC GLM kennt: Dazu enthält eine dataset–Variable die Namen der abhängigen Variablen, eine andere deren Werte. Dies ist leicht mit PROC TRANSPOSE zu erreichen. Als Beispiel sollen die Daten einer randomisierten Studie zur Wirksamkeit einer Antibiotikaprophylaxe bei Thoraxoperationen (vgl. [5]) dienen: 100 Patienten ohne Antibiotikum–Prophylaxe vor der Operation stehen 100 mit Prophylaxe gegenüber; bei allen wird im postoperativen Verlauf beobachtet, inwieweit sich im Röntenbild nachweisbare

```
            PATNR    GRUPPE    _NAME_      COL1
...
            001      P         NRP_T5     0.0000
            001      P         NRP_T6     2.0000
            001      P         NRP_T7     2.0000
            ...
            001      P         LEU_T5    16.9000
            001      P         LEU_T6    16.8000
            001      P         LEU_T7    13.1167
            ...
            002      K         NRP_T5     0.0000
            002      K         NRP_T6     0.0000
            002      K         NRP_T7     0.0000
...
```

Abbildung 2: Daten nach Anwendung von PROC TRANSPOSE (Ausschnitt)

Infiltrate (Variablen `nrp_t3, nrp_t4...`) entwickeln (welche dann bewertet werden und somit als ordinale kategorielle Daten aufzufassen sind; die Skala reicht von 0 „kein Nachweis“ bis 5 „schwer: Transparenzminderung in mehr als halber Thoraxhöhe mit mindestens zentraler Überdeckung der Lungenzeichnung und Aufhebung der Transparenz“). Daneben werden als Indikator für entzündliche Prozesse die Leukozyten–Zahlen (Variablen `leu_t3, leu_t4...`) bestimmt (zu den gleichen Zeitpunkten wie die Anfertigung der Röntgenaufnahmen). Von einem Erfolg der Prophylaxe kann also dann gesprochen werden, wenn *zu allen relevanten Zeitpunkten* sowohl die Infiltratentwicklung als auch die Leukozyten–Zahlen in der Prophylaxe–Gruppe niedriger sind als in der Kontrollgruppe. Einen Ausschnitt aus den Originaldaten gibt Abb. 1; nach PROC TRANSPOSE sind die Daten in der Form von Abb. 2 und somit in einer für die Anwendung des Macros OB geeigneten Form.

Der Macroaufruf erfolgt mit

```
%ob(data=tt, ds=temp, subj=patnr, vars=_name_, values=col1, class=gruppe, by=);
```

Dabei ist `tt` der Name der SAS–Datei, mit `ds=` kann optional ein Namensbestandteil für temporäre Dateien im WORK–Verzeichnis angegeben werden (`temp` ist default–Vorgabe, d.h. während des Macroablaufs werden Dateien `tempa`, `tempb` u.s.w. angelegt und nach Beendigung automatisch gelöscht), `subj=` gibt die Variable für die unabhängigen Beobachtungseinheiten an (hier die Patientennummer `patnr`), in der bei `vars=` angegebenen Variable finden sich die Namen der in die Analyse eingehenden Variablen (hier `_name_`) und in der nach `values=` angegebenen Variablen deren Werte (hier `col1`; letztere sind die Standardnamen nach Anwendung von PROC TRANSPOSE). Mit `class=` wird schließlich die Gruppierungsvariable und mit `by=` eine mögliche BY–Variable angegeben. Hierbei braucht die Datei, anders als sonst bei SAS üblich, nicht vorher sortiert zu werden.

Die Ergebnisse in Abb. 3 weisen auf gleichgerichtete Differenzen zwischen der Prophylaxe– und der Kontrollgruppe hin, die sich in Hotellings Test nicht deutlich zeigen. Univariate nichtparametrische Analysen geben ebenfalls einen Hinweis auf gleichgerichtete Alternativen, werden jedoch hier aus Platzgründen nicht dargestellt. Man beachte, daß hier ein gemischter Variablenvektor aus metrischen und ordinalen kategoriellen Daten in die Analyse eingeht.

Ausblick Das Macro OB deckt noch nicht in vollem Umfang die Möglichkeiten ab, welche die allgemeine lineare Rangstatistik (3.1) bietet; insbesondere die Analyse mit verschiedenen Gewichtsvektoren für die *Gruppen* und die Analyse im a–Stichprobenfall (mit $a > 2$) ist als Erweiterung in Arbeit. Das Macro ist auf Anfrage beim Autor erhältlich.

Adresse des Autors

Thomas Bregenzer
UKE/IMDM
Martinistr. 52
20246 Hamburg

Tel. 040 4717-2111
FAX 040 4717-4882
email: tbregen@bregenzer.uke.uni-hamburg.de

```
            MULTIVARIATE ANALYSIS OF DIRECTIONAL ALTERNATIVES

                              Class variable:
                                   GRUPPE

                      n1=       100 ,     n2=       100


                            Variables in analysis:
   LEU_T3 LEU_T4 LEU_T5 LEU_T6 LEU_T7 NRP_T3 NRP_T4 NRP_T5 NRP_T6 NRP_T7

                 NONPARAMETRIC TEST-STATISTICS and P-VALUES


   Hotelling's T^2:  1.0978396
           p-value:  0.3657571


 Tests for directional alternatives
 (with t-distribution approximation)


   O'Brien's    OLS-statistic:  2.0504508
                      p-value:   0.041638

   O'Brien's    GLS-statistic:  1.7908272
                      p-value:  0.0748486

   standardized sum statistic:  2.0504508
                      p-value:   0.041638

   principal comp.  statistic:  2.0498747
                      p-value:  0.0416948
```

Abbildung 3: Output von Macro OB (gekürzt)

Literatur

[1] M.G. Akritas and E. Brunner. Rank tests for patterned alternatives in factorial designs with interactions. Research Developments in Probability and Statistics, a Festschrift on the Occasion of the 65^{th} birthday of Madan L. Puri, VSP-International Science Publishers, Utrecht, The Netherlands., 1996.

[2] T. Bregenzer. Nichtparametrische Multivariate Summenstatistiken. Dissertation, Universität zu Köln, in Vorbereitung, 1997.

[3] E. Brunner and M.L. Puri. A class of rank–score tests in factorial designs. *preprint*, 1996.

[4] D. Follmann. Multivariate tests for multiple endpoints. Report, National Heart, Lung, and Blood Institute, Bethesda, 1993.

[5] D. Frey. Single-Shot–Antibiotika–Prophylaxe in der Thoraxchirurgie — Klinische Untersuchung zur Effizienz mit besonderer Berücksichtigung der Auswirkungen auf das Merkmal „Infiltrat“ im Thorax–Röntgenbild. Habilitationsschrift (in Vorbereitung), Medizinische Fakultät der Humboldt–Universität zu Berlin, Virchow–Klinikum, 1997.

[6] J.A. Koziol, D.A. Maxwell, M. Fukushima, M.E.M. Colmerauer, and Y.H. Pilch. A distribution–free test for tumor–growth curve analyses with application to an animal tumor immunotherapy experiment. *Biometrics*, 37:383–390, 1981.

[7] S. Kropf and J. Läuter. Comparing independent samples of highdimensional observation vectors. In *COMPSTAT, Proceedings in Computational Statistics*, pages 286 – 291, 1994.

[8] J.M. Lachin. Some large–sample distribution–free estimators and tests for multivariate partially incomplete data from two populations. *Statistics in Medicine*, 11:1151–1170, 1992.

[9] J. Läuter. Stabilisierte Tests zur multivariaten Auswertung medizinischer Studien. In *Neue Paradigmen in medizinischer Informatik, Biometrie und Epidemiologie, 39. Jahrestagung der GMDS*, pages 381–387, 1995.

[10] J. Läuter. New multivariate tests for data with an inherent structure. *Biom. J.*, 38:5–23, 1996.

[11] W. Lehmacher, G. Wassmer, and P. Reitmeier. Comment on: On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, 50:581–583, 1994.

[12] P.C. O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40:1079–1087, 1984.

[13] L.J. Wei and J.M. Lachin. Two–sample asymptotically distribution free tests for incomplete multivariate observations. *JASA*, 79:653–661, 1984.