Nonparametric Multiple Comparisons in Repeated Measures Designs for Data with Ties

Ullrich Munzel[1]

Department of Medical Biometrics and Data Management
Merz + Co. GmbH & Co.
Germany

Ajit C. Tamhane

Department of Statistics
Northwestern University
USA

Summary

We consider some multiple comparison problems in repeated measures designs for data with ties, particularly ordinal data; the methods are also applicable to continuous data, with or without ties. A unified asymptotic theory of rank tests of Brunner, Puri and Sen (1995) and Akritas and Brunner (1997) is utilized to derive large sample multiple comparison procedures (MCP's).

First, we consider a single treatment and address the problem of comparing its time effects with respect to the baseline. Multiple sign tests and rank tests (and the corresponding simultaneous confidence intervals) are derived for this problem. Next, we consider two treatments and address the problem of testing for treatment×time interactions by comparing their time effects with respect to the baseline. Simulation studies are conducted to study the type I familywise error rates and powers of competing procedures under different distributional models. The data from a psychiatric study are analyzed using the above MCP's to answer the clinicians' questions.

Key Words: Rank statistics, Sign statistics, Midranks, Rank transform tests, Ordinal data, Joint ranking, Familywise error rate, Power.

1. Introduction

This work was motivated by the research questions addressed in a randomized parallel group trial conducted at the Department of Psychiatry, University of Göttingen to compare a new drug versus a placebo to cure panic disorder in psychiatric patients. A total of 30 patients were randomly assigned to the two treatment groups with 15 per group. Patients were assessed at the baseline and then were monitored on seven occasions over a period of 10 weeks. One of the efficacy variables was the clinical global impression (CGI) measured on a seven-point ordinal scale from 0 = best to 6 = worst. The data are shown in Table 1. Side-by-side box plots of the weekly data for the drug (solid circles) and placebo (open circles) groups over the 10-week period are shown in Figure 1. We see that the trend is flat in the placebo group, but is downward sloping in the drug group indicating improvement.

The following questions of interest to the clinicians will be addressed in the present paper:

[1] This work was supported in part by the Boehringer Ingelheim Fonds. The authors thank Professor Edgar Brunner for useful comments.

1. What is the earliest time point at which the drug shows a significant improvement w.r.t. the baseline?

2. Is there a placebo effect as measured by a significant improvement w.r.t. the baseline at any time point?

3. What is the earliest time point at which the drug shows a significantly higher improvement than the placebo w.r.t. the baseline?

To formally answer these questions, not only does one need to deal with ordinal data involving many ties, but also one needs to perform multiple comparisons over time points (in particular, comparisons with the baseline). The goal of the present paper is to derive the necessary procedures. These procedures will be of use to researchers in many disciplines, including medicine, psychometry and marketing, where repeated measures designs with ordinal data are common.

There are many nonparametric multiple comparison procedures (MCP's) for continuous data. In particular, the multiple sign test of Steel (1959) and the multiple signed rank test of Nemenyi (1963) discussed in Section 2.2, Chapter 9 of Hochberg and Tamhane (1987) deal with dependent data, which is the focus of the present paper. Nonparametric procedures for repeated measures designs with continuous data have been studied by Thompson (1991), Akritas and Arnold (1994) and Brunner and Denker (1994). In applications, measurement scales are essentially discrete because of limited gage accuracy; hence ties are common. In many applications, as in our example, an ordinal scale is used in which case ties are a rule rather than an exception. The above procedures are not suitable in the present context because of the prevalence of ties in ordinal data.

Most rank tests handle ties by assigning them midranks based on heuristic grounds. The concept of normalized distribution functions due to Ruymgaart (1980) leads to midranks in a natural way. Using this approach, a unified asymptotic theory of rank tests for continuous as well as discrete data has been developed by Brunner, Puri and Sun (1995) and Akritas and Brunner (1997). Brunner and Langer (1999, 2000) applied the results to longitudinal data. We utilize the results of these papers to derive large sample MCP's to answer the questions of interest. Some technical details are omitted for brevity; the interested reader may refer to the aforementioned papers. More than two treatments are not considered in this paper, so there are no multiple comparisons between treatments.

The organization of the paper is as follows. Section 2 defines the notation and basic assumptions. Section 3 considers the single treatment case (e.g., only the placebo or only the drug group) where multiple comparisons stem from the need to compare different time points with each other. We focus on the many-to-one comparisons with the baseline to answer questions 1 and 2 above. Multiple sign and rank tests (and the associated simultaneous confidence intervals) are derived for suitably defined effects. Section 4 considers two treatment groups. Multiple rank tests are given for treatment$\times$time interactions to answer question 3 above. Section 5 gives simulation results to study the type I error rates and powers of the proposed tests. In Section 6 we return to the example and analyze the data in Table 1 using the procedures proposed in earlier sections. Finally, a discussion of the resulting methods is given in Section 7.

2. Notation and Assumptions

Consider $a = 1$ or 2 treatments and $b+1 \geq 2$ repeated measures, which are assumed to be observations at $b+1$ successive *occasions*, beginning with occasion 0, called the *baseline*. The subjects are assumed to be drawn as a random sample from a homogeneous population. For $a = 2$, an independent samples design is assumed with subjects assigned at random to the treatments. Let X_{ijk} denote the observation at the jth occasion on the kth subject in the ith treatment group ($i = 1, 2, j = 0, 1, \ldots, b, k = 1, 2, \ldots, n_i$). The X_{ijk} are measured on at least ordinal scale. For asymptotic considerations, we assume that $n_1, n_2 \to \infty$ at the same rate so that the ratio n_1/n_2 is bounded away from zero and ∞.

The common distribution of $\boldsymbol{X}_{ik} = (X_{i0k}, X_{i1k}, \ldots, X_{ibk})'$ is assumed to be non-degenerate, but otherwise completely arbitrary. In order to account for ties and ordinal data, we define the marginal normalized cumulative distribution function (c.d.f.) of X_{ijk} as (Ruymgaart, 1980)

$$F_{ij}(x) = \frac{1}{2} \cdot [F_{ij}^{+}(x) + F_{ij}^{-}(x)], \quad i = 1, 2, \ j = 0, 1, \ldots, b,$$

where $F_{ij}^{+}(x)$ is the right-continuous and $F_{ij}^{-}(x)$ is the left-continuous version of the original marginal c.d.f. of X_{ijk}.

In Section 3 we consider a single treatment and drop the subscript i, letting $\boldsymbol{X}_k = (X_{0k}, X_{1k}, \ldots, X_{bk})'$ denote independent and identically distributed (i.i.d.) observation vectors on subjects $k = 1, 2, \ldots, n$ and $F_j(x)$ the corresponding normalized c.d.f.'s.

The comparisons of interest are formulated as multiple hypothesis testing problems in each case. The type I *familywise error rate (FWE)* (Hochberg and Tamhane, 1987) for a family of hypotheses is defined as

$$\text{FWE} = P\{\text{At least one true null hypothesis is rejected}\}. \tag{2.1}$$

A requirement for a multiple test procedure is that the FWE be strongly controlled, i.e., controlled under all possible configurations of the true hypotheses, at a specified level α.

3. A Single Treatment

In this case, we focus on many-to-one comparisons between the occasions $j = 1, \ldots, b$ with occasion 0, the baseline. The results can be readily extended to other comparisons, e.g., pairwise comparisons between all occasions or comparisons between successive occasions.

We define the *effect* of occasion j w.r.t. the baseline as

$$\theta_j = P(X_{0k} < X_{jk}) + \frac{1}{2}P(X_{0k} = X_{jk}) - \frac{1}{2}, \ \ j = 1, \ldots, b. \tag{3.1}$$

Depending on whether $\theta_j >, =$ or < 0, we can say that X_{jk} is *tendentiously* larger, comparable or smaller than X_{0k}. In Section 3.1 we derive multiple sign tests on the θ_j.

Although θ_j has a very simple interpretation, it only admits sign-type of tests. To admit rank-type tests we propose an alternative definition of the *effect* of occasion $j = 1, \ldots, b$ w.r.t. the baseline:

$$\psi_j = P(X_{0k} < X_{j\ell}) + \frac{1}{2}P(X_{0k} = X_{j\ell}) - \frac{1}{2} = \int F_0(x)dF_j(x) - \frac{1}{2}, \ \ k \neq \ell. \tag{3.2}$$

Note that by comparing two independent subjects, this measure reduces the comparison between occasion j with the baseline in terms of the respective marginal distributions. In Section 3.2 we derive multiple rank tests on the ψ_j.

Remark 1: Define $G_j(x) = 0.5\left[F_0(x) + F_j(x)\right]$. Then $\psi_j = E\left[G_j(X_{jk}) - G_j(X_{0k})\right]$. Thus G_j may be thought of as an unknown link function or transformation which results in a linear scale on which X_{jk} and X_{0k} can be compared. The above equation can also be expressed as

$$\psi_j = \frac{1}{2}E\left[F_0(X_{jk}) - F_j(X_{0k})\right], \quad j = 1, \ldots, b. \tag{3.3}$$

We use these alternative representations of ψ_j in the sequel. □

3.1 Multiple Sign Tests and Confidence Intervals

Consider the family of two-sided hypothesis testing problems:

$$H_{0j} : \theta_j = 0 \text{ vs. } H_{1j} : \theta_j \neq 0, \quad j = 1, \ldots, b. \tag{3.4}$$

Let $Y_{jk} = 1, 1/2$ or 0 according as $X_{0k} <, =$ or $> X_{jk}$. Then an unbiased estimate of θ_j is given by

$$\widehat{\theta}_j = \overline{Y}_{j\cdot} - \frac{1}{2} = \frac{1}{n}\sum_{k=1}^{n} Y_{jk} - \frac{1}{2} = \frac{1}{n}\left[N_j^+ + \frac{1}{2}N_j^0\right] - \frac{1}{2},$$

where $N_j^+ = \sharp\{k|X_{0k} < X_{jk}\}$ and $N_j^0 = \sharp\{k|X_{0k} = X_{jk}\}$. The standardized test statistics

$$Z_j = \frac{(\overline{Y}_{j\cdot} - 1/2)\sqrt{n}}{\sqrt{\widehat{\sigma}_{jj}}}, \quad j = 1, \ldots, b \tag{3.5}$$

are asymptotically standard normal under H_{0j} of (3.4) where $\widehat{\sigma}_{jj}$ is a consistent estimate of $\sigma_{jj} = \mathrm{Var}(Y_{jk})$. Let σ_{ij} denote $\mathrm{Cov}(Y_{ik}, Y_{jk})$ (which equals $\mathrm{Var}(Y_{jk})$ for $i = j$). Since the Y_{jk} for $k = 1, 2, \ldots, n$ are i.i.d., the σ_{ij} and correlations ρ_{ij} are consistently estimated by

$$\widehat{\sigma}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\left(Y_{ik} - \overline{Y}_{i\cdot}\right)\left(Y_{jk} - \overline{Y}_{j\cdot}\right) \text{ and } \widehat{\rho}_{ij} = \frac{\widehat{\sigma}_{ij}}{\sqrt{\widehat{\sigma}_{ii}\widehat{\sigma}_{jj}}}. \tag{3.6}$$

The testing family $\{(H_{0j}, Z_j), \ \ j = 1, \ldots, b\}$ is joint (Gabriel, 1969). By using the union-intersection method (see Hochberg and Tamhane, 1987, p. 28), a multiple test procedure that strongly controls the FWE asymptotically at level α for the family of hypotheses (3.4) is given by:

$$\text{Reject } H_{0j} : \theta_j = 0 \text{ if } |Z_j| > |z|_{b,\{\widehat{\rho}_{ij}\},\alpha}, \quad j = 1, \ldots, b,$$

where $|z|_{b,\{\widehat{\rho}_{ij}\},\alpha}$ is the two-sided upper α equicoordinate critical point of the b-variate standard normal distribution (with zero means and unit variances) and correlation matrix $\{\widehat{\rho}_{ij}\}$. The corresponding $100(1-\alpha)\%$ simultaneous confidence intervals (SCI's) on the θ_j are given by

$$\theta_j \in \left[\widehat{\theta}_j \pm |z|_{b,\{\widehat{\rho}_{ij}\},\alpha}\sqrt{\widehat{\sigma}_{jj}/n}\right], \quad j = 1, \ldots, b.$$

Extensions to one-sided hypotheses and SCI's are straightforward.

REMARK 2: The above method of deriving a multiple test procedure by first obtaining the test statistics Z_j that form a joint testing family $\{(H_{0j}, Z_j)\}$ and then using the union-intersection method to find the common critical point from the asymptotic multivariate normal (MVN) distribution of the Z_j is used in all of the problems discussed in the remainder of the paper. Therefore these technical details are not always mentioned. Also, note that it is possible to derive more powerful stepwise testing procedures if SCI's are not required. $\square$

REMARK 3: The statistic (3.5) is a paired t-statistic where each paired difference, $X_{jk} - X_{0k}$, is coded as 1, 1/2 or 0 as explained above. This suggests that, especially for small n, the joint distribution of the Z_j under the overall null hypothesis may be better approximated by a b-variate t-distribution. In fact, the critical point $|z|_{b,\{\widehat{\rho}_{ij}\},\alpha}$ gives a slightly anti-conservative test. So we suggest that it be replaced by the two-sided upper α equicoordinate critical point $|t|_{b,n-1,\{\widehat{\rho}_{ij}\},\alpha}$ from the b-variate t-distribution with $n-1$ degrees of freedom (d.f.). This critical point can be calculated by using the SAS-IML program of Genz and Bretz (1999) (available at `www.bioinf.uni-hannover.de`). It should be noted that the multivariate t-distribution assumes a common estimate of variance for all statistics, whereas here we have different estimates that are correlated. However, we do take this into account by not pooling the d.f. The superiority of the t- approximation in accurately controlling the FWE was confirmed in the simulation studies. The same type of approximation is used in all of the problems discussed in the remainder of the paper. $\square$

3.2 Multiple Rank Tests and Confidence Intervals

Consider the family of two-sided hypothesis testing problems:

$$H_{0j} : \psi_j = 0 \text{ vs. } H_{1j} : \psi_j \neq 0, \quad j = 1, \ldots, b. \tag{3.7}$$

A natural estimate of ψ_j is obtained as follows. Let $\widehat{F}_j$ denote the empirical normalized c.d.f.'s for $j = 0, 1, \ldots, b$. Then from (3.3) we have

$$\widehat{\psi}_j = \frac{1}{2n} \sum_{k=1}^{n} \left[\widehat{F}_0(X_{jk}) - \widehat{F}_j(X_{0k}) \right]. \tag{3.8}$$

Now,

$$\widehat{F}_0(X_{jk}) = \frac{1}{n} \left[R_{jk}^{(0j)} - R_{jk}^{(j)} \right] \text{ and } \widehat{F}_j(X_{0k}) = \frac{1}{n} \left[R_{0k}^{(0j)} - R_{0k}^{(0)} \right],$$

where $R_{0k}^{(0j)}$ and $R_{jk}^{(0j)}$ denote the midranks of X_{0k} and X_{jk} in the *joint* ranking of the baseline and the occasion j samples together, and $R_{0k}^{(0)}$ and $R_{jk}^{(j)}$ denote the midranks of X_{0k} and X_{jk} in their respective *internal* rankings, i.e., within the baseline and the occasion j samples separately. Let $\overline{R}_{0\cdot}^{(0j)} = n^{-1} \sum_{k=1}^{n} R_{0k}^{(0j)}$ and $\overline{R}_{j\cdot}^{(0j)} = n^{-1} \sum_{k=1}^{n} R_{jk}^{(0j)}$ be the sample means of the midranks. Then it is easy to verify that (3.8) reduces to

$$\widehat{\psi}_j = \frac{1}{2n} \left(\overline{R}_{j\cdot}^{(0j)} - \overline{R}_{0\cdot}^{(0j)} \right). \tag{3.9}$$

Notice that the internal ranks cancel out in the calculation of this estimate; however, they appear in the formulas for the estimates of the variances and covariances. The estimate $\widehat{\psi}_j$ is asymptotically unbiased and consistent (Brunner, Puri and Sun, 1995).

To obtain the asymptotic joint distribution of the $\widehat{\psi}_j$, define

$$Y_{jk}=F_0(X_{jk})-F_j(X_{0k}) \text{ and } \widehat{Y}_{jk}=\widehat{F}_0(X_{jk})-\widehat{F}_j(X_{0k})=\frac{1}{n}\left[R_{jk}^{(0j)}-R_{jk}^{(j)}-R_{0k}^{(0j)}+R_{0k}^{(0)}\right].$$

Then from (3.8) we have

$$\widehat{\psi}_j = \frac{1}{2}\overline{\widehat{Y}}_{j\cdot} = \frac{1}{2n}\sum_{k=1}^{n}\widehat{Y}_{jk}. \tag{3.10}$$

Note that $\overline{\widehat{Y}}_{j\cdot}$ is the sample mean of $\widehat{Y}_{jk}$, which are not independent. However, using methods similar to those in Brunner, Puri and Sun (1995), we can show that

$$\sqrt{n}\left\|\overline{\widehat{Y}}_{j\cdot}-\overline{Y}_{j\cdot}\right\|_2 \to 0, \quad j=1,\ldots,b$$

in the L_2-norm, where $\overline{Y}_{j\cdot}$ is the sample mean of the Y_{jk}, which are i.i.d. Since componentwise convergence implies multivariate convergence in finite dimensions, it follows that

$$\sqrt{n}\left\|\left(\overline{\widehat{Y}}_{1\cdot},\ldots,\overline{\widehat{Y}}_{b\cdot}\right)-\left(\overline{Y}_{1\cdot},\ldots,\overline{Y}_{b\cdot}\right)\right\|_2 \to 0.$$

By the multivariate central limit theorem (Gnedenko, 1962), $(\overline{Y}_{1\cdot},\ldots,\overline{Y}_{b\cdot})$ is asymptotically MVN. Hence $\left(\overline{\widehat{Y}}_{1\cdot},\ldots,\overline{\widehat{Y}}_{b\cdot}\right)$ has the same asymptotic MVN distribution with

$$\begin{aligned}
E\left(\overline{\widehat{Y}}_{j\cdot}\right) &\longrightarrow E\left(\overline{Y}_{j\cdot}\right)=2\psi_j, \quad j=1,\ldots,b,\\
n\mathrm{Cov}\left(\overline{\widehat{Y}}_{i\cdot},\overline{\widehat{Y}}_{j\cdot}\right) &\longrightarrow n\mathrm{Cov}\left(\overline{Y}_{i\cdot},\overline{Y}_{j\cdot}\right)=\mathrm{Cov}\left(Y_{ik},Y_{jk}\right)=\sigma_{ij}, \quad i,j=1,\ldots,b.
\end{aligned}$$

The σ_{ij} can be consistently estimated by the corresponding sample variances (for $i=j$) and covariances (for $i\neq j$) among the Y_{ik} and Y_{jk}. But since the latter are unobservable, we use the corresponding sample quantities, $\widehat{Y}_{ik}$ and $\widehat{Y}_{jk}$, respectively, resulting in the following estimates:

$$\begin{aligned}
\widehat{\sigma}_{ij} &= \frac{1}{n-1}\sum_{k=1}^{n}\left(\widehat{Y}_{ik}-\overline{\widehat{Y}}_{i\cdot}\right)\left(\widehat{Y}_{jk}-\overline{\widehat{Y}}_{j\cdot}\right)\\
&= \frac{1}{2n(n-1)}\sum_{k=1}^{n}\left\{R_{ik}^{(0i)}-R_{ik}^{(i)}-R_{0k}^{(0i)}+R_{0k}^{(0)}-\left(\overline{R}_{i\cdot}^{(0i)}-\overline{R}_{0\cdot}^{(0i)}\right)\right\}\\
&\qquad\times\left\{R_{jk}^{(0j)}-R_{jk}^{(j)}-R_{0k}^{(0j)}+R_{0k}^{(0)}-\left(\overline{R}_{j\cdot}^{(0j)}-\overline{R}_{0\cdot}^{(0j)}\right)\right\}.
\end{aligned} \tag{3.11}$$

These estimates can be shown to be consistent by using the techniques of Brunner, Puri and Sun (1995). Therefore,

$$Z_j=\frac{\widehat{\psi}_j\sqrt{4n}}{\sqrt{\widehat{\sigma}_{jj}}}=\frac{\left(\overline{R}_{j\cdot}^{(0j)}-\overline{R}_{0\cdot}^{(0j)}\right)\sqrt{2}}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left\{R_{jk}^{(0j)}-R_{jk}^{(j)}-R_{0k}^{(0j)}+R_{0k}^{(0)}-\left(\overline{R}_{j\cdot}^{(0j)}-\overline{R}_{0\cdot}^{(0j)}\right)\right\}^2}}, \quad j=1,\ldots,b,$$

are asymptotically standard normal under the respective hypotheses $H_{0j} : \psi_j = 0$. The correlation matrix of the Z_j can be consistently estimated by $\{\widehat{\rho}_{ij}\}$ in the usual manner. As before, a multiple test procedure that strongly controls the FWE asymptotically at level α for the family of hypothesis testing problems (3.7) is given by:

$$\text{Reject } H_{0j} : \psi_j = 0 \text{ if } |Z_j| > |z|_{b,\{\widehat{\rho}_{ij}\},\alpha}, \quad j = 1, \ldots, b.$$

The corresponding $100(1-\alpha)\%$ simultaneous two-sided confidence intervals on the ψ_j are given by

$$\psi_j \in \left[\widehat{\psi}_j \pm |z|_{b,\{\widehat{\rho}_{ij}\},\alpha}\sqrt{\widehat{\sigma}_{jj}/4n}\right], \quad j = 1, \ldots, b.$$

In analogy with Remark 2, we recommend the use of the multivariate t critical point $|t|_{b,n-1,\{\widehat{\rho}_{ij}\},\alpha}$ in place of $|z|_{b,\{\widehat{\rho}_{ij}\},\alpha}$.

4. Two Treatments

In this case, one of the questions of interest is whether the time effects are different for the two treatment groups. This can be formulated as a test of hypothesis of no treatment×time interaction. The multiple sign and rank tests of the previous section can be extended to the two treatment setting. For example, for the multiple sign test we can take the interaction effect for occasion j as $\theta_{1j} - \theta_{2j}$ where θ_{ij} is the effect of treatment i at time j w.r.t. the baseline as defined in (3.1). Similarly, for the multiple rank test we can take the interaction effect for occasion j as $\psi_{1j} - \psi_{2j}$ where ψ_{ij} is the effect of treatment i at time j w.r.t. the baseline as defined in (3.2). The estimates of these interaction effects and their asymptotic distributions can be derived in a straightforward manner. We omit the details for brevity.

The multiple rank test faces a difficulty arising because it ranks observations within each treatment separately (but jointly over the occasion j and baseline). This is a result of the fact that the measures ψ_{ij} are defined separately for each treatment $i = 1, 2$. An undesirable consequence of this separate ranking is that a large quantitative interaction (i.e., the time effects in both treatments are in the same direction, but are of different magnitudes) is likely to go undetected. As a simple example, suppose we have two observations from each treatment at the baseline and occasion 1: $\boldsymbol{X}_{11} = (1,3), \boldsymbol{X}_{12} = (2,4), \boldsymbol{X}_{21} = (5,15), \boldsymbol{X}_{22} = (6,16)$. Then $\overline{X}_{11\cdot} - \overline{X}_{10\cdot} = 2$ and $\overline{X}_{21\cdot} - \overline{X}_{20\cdot} = 10$. This is a quantitative interaction. Unfortunately, since the ranks are assigned separately for each treatment, we get identical joint ranks for the two subjects in each treatment group: $R^{(01)}_{101} = R^{(01)}_{201} = 1$, $R^{(01)}_{102} = R^{(01)}_{202} = 2$, $R^{(01)}_{111} = R^{(01)}_{211} = 3$, $R^{(01)}_{112} = R^{(01)}_{212} = 4$. Their internal ranks are also identical: $R^{(0)}_{101} = R^{(0)}_{201} = 1$, $R^{(0)}_{102} = R^{(0)}_{202} = 2$, $R^{(1)}_{111} = R^{(1)}_{211} = 1$, $R^{(1)}_{112} = R^{(1)}_{212} = 2$. As a result, $\widehat{\psi}_{11} - \widehat{\psi}_{21} = 0$ (the corresponding variance estimate is also zero) and the interaction will go undetected. The same difficulty occurs with the multiple sign test on $\theta_{1j} - \theta_{2j}$. However, other types of interactions (e.g., quantitative interactions with small time effects in both treatments or qualitative interactions) are detectable using these two tests.

To avoid the above problem, we need to assign ranks jointly over the two occasions as well as over the two treatments (called *overall ranking*). Rank statistics based on overall ranks result naturally if we define the *effect* of treatment i at occasion j w.r.t. the baseline as

$$\phi_{ij} = E[G_j(X_{ijk}) - G_j(X_{i0k})] = \int G_j(x)\,dF_{ij}(x) - \int G_j(x)\,dF_{i0}(x), \quad i = 1, 2, j = 1, 2, \ldots, b,$$

where $G_j = N^{-1}[n_1(F_{10} + F_{1j}) + n_2(F_{20} + F_{2j})]$ is a weighted average of all four distributions and $N = 2(n_1 + n_2)$ denotes the total number of all observations.

Consider the hypothesis testing problem:

$$H_{0j} : \phi_{1j} - \phi_{2j} = 0 \text{ vs. } H_{1j} : \phi_{1j} - \phi_{2j} \neq 0, \quad j = 1, \ldots, b. \tag{4.1}$$

Note that H_{0j} is implied by the stricter null hypothesis $H_{0j}^F : F_{1j} - F_{10} - F_{2j} + F_{20} = 0$. This latter null hypothesis was considered by Akritas and Brunner (1997). Their test is based on the same transformation G_j above.

A natural estimate of ϕ_{ij} is given by

$$\widehat{\phi}_{ij} = \int \widehat{G}_j\,d\widehat{F}_{ij} - \int \widehat{G}_j\,d\widehat{F}_{i0} = \frac{1}{N}\left(\overline{R}_{ij\cdot}^{(0j)} - \overline{R}_{i0\cdot}^{(0j)}\right),$$

where $R_{i\ell k}^{(0j)}$, $\ell = 0, j$ denotes the overall rank of $X_{i\ell k}^{(0j)}$ among all observations at time j and baseline in both groups. Using arguments similar to those in Akritas and Brunner (1997) it can be shown that $\sqrt{N}(\widehat{\phi}_{1j} - \widehat{\phi}_{2j})$ is asymptotically equivalent to $\sqrt{N}\left(\overline{Y}_{j,1j\cdot} - \overline{Y}_{j,10\cdot} - \overline{Y}_{j,2j\cdot} + \overline{Y}_{j,20\cdot}\right)$ under the null hypothesis H_{0j}^F, where $Y_{j,i\ell k} = G_j(X_{i\ell k})$ and $\overline{Y}_{j,i\ell\cdot} = n_i^{-1}\sum_{k=1}^{n_i} Y_{j,i\ell k}$, $\ell = 0, j$. For every nonempty subset $J \subseteq \{1, 2, \ldots, b\}$, the asymptotic multivariate normality of $\sqrt{N}(\widehat{\phi}_{1j} - \widehat{\phi}_{2j})$, $j \in J$ under $\bigcap_{j \in J} H_{0j}^F$ follows from the application of the multivariate central limit theorem.

If $\sigma_{jj'}$ denotes the asymptotic covariance between $\sqrt{N}(\widehat{\phi}_{1j} - \widehat{\phi}_{2j})$ and $\sqrt{N}(\widehat{\phi}_{1j'} - \widehat{\phi}_{2j'})$ then the abovementioned asymptotic equivalence implies that

$$\sigma_{jj'} = N(\sigma_{1,jj'}/n_1 + \sigma_{2,jj'}/n_2),$$

where $\sigma_{i,jj'} = \mathrm{Cov}(Y_{j,ijk} - Y_{j,i0k}, Y_{j',ij'k} - Y_{j',i0k})$. Again, the $Y_{j,i\ell k}$ are not observable and hence we use their empirical counterparts

$$\widehat{Y}_{j,i\ell k} = \widehat{G}_j(X_{i\ell k}) = \frac{1}{N}\left(\overline{R}_{i\ell k}^{(0j)} - \frac{1}{2}\right)$$

to obtain consistent estimates of the variances and covariances. Under $H_{0j}^F \bigcap H_{0j'}^F$ we obtain a consistent estimate of $\sigma_{jj'}$ by

$$\widehat{\sigma}_{jj'} = N(\widehat{\sigma}_{1,jj'}/n_1 + \widehat{\sigma}_{2,jj'}/n_2),$$

where

$$\begin{aligned}
\widehat{\sigma}_{i,jj'} &= \frac{N}{n_i - 1}\sum_{k=1}^{n_i}\left(\widehat{Y}_{j,ijk} - \widehat{Y}_{j,i0k} - \overline{\widehat{Y}}_{j,ij\cdot} + \overline{\widehat{Y}}_{j,i0\cdot}\right)\left(\widehat{Y}_{j',ij'k} - \widehat{Y}_{j',i0k} - \overline{\widehat{Y}}_{j',ij'\cdot} + \overline{\widehat{Y}}_{j',i0\cdot}\right) \\
&= \frac{1}{N(n_i - 1)}\sum_{k=1}^{n_i}\left(R_{ijk}^{(0j)} - R_{i0k}^{(0j)} - \overline{R}_{ij\cdot}^{(0j)} + \overline{R}_{i0\cdot}^{(0j)}\right)\left(R_{ij'k}^{(0j')} - R_{i0k}^{(0j')} - \overline{R}_{ij'\cdot}^{(0j')} + \overline{R}_{i0\cdot}^{(0j')}\right).
\end{aligned}$$

The test statistics
$$Z_j = \frac{\widehat{\phi}_{1j} - \widehat{\phi}_{2j}}{\sqrt{\widehat{\sigma}_{jj}}},\ j = 1, 2, \ldots, b$$
are asymptotically standard normal under H_{0j}^F. The correlation matrix of the Z_j is consistently estimated by $\{\widehat{\rho}_{jj'}\}$ in the usual manner. Multiple tests of the hypotheses (4.1) can be based on these statistics with strong FWE control as discussed before.

SCI's cannot be computed for $\phi_{1j} - \phi_{2j}$ because the variance and covariance estimates are consistent only under the corresponding null hypotheses $H_{0j}^F \bigcap H_{0j'}^F$. Moreover, note that the test statistics are based on the effects ϕ_{1j} and not consistent against alternatives in $H_{0j} \backslash H_{0j}^F$ (Brunner, Munzel and Puri, 1999). Solutions of this problem for the independent case (e.g. see Cohen et al., 2000) cannot be easily extended to repeated measures.

5. Simulations

We carried out extensive simulations to study the FWE and power properties of the multiple sign and rank tests for a single treatment proposed in Section 3. The normal theory multiple paired t-test with empirically estimated covariance matrix was included in the simulation study as a benchmark for comparison.

Two values of b were considered: $b = 3$ and $b = 7$. Four different $(b+1)$-variate distributions were simulated: 1) Multivariate normal (MVN), 2) rounded multivariate normal (RMVN) as an example of a discrete distribution, 3) multivariate lognormal (MVLN) as an example of a skewed distribution, 4) multivariate Cauchy (MVCAUCHY) as an example of a heavy-tailed distribution.

MVN observations with zero means and unit variances were generated with two different correlation structures: 1.) Autocorrelation structure having $\mathrm{Corr}(X_{ik}, X_{jk}) = \rho^{|i-j|}$ with autocorrelation coefficient $\rho = 0.5$, 2.) compound symmetry structure with $\mathrm{Corr}(X_{ik}, X_{jk}) = \rho = 0.5$ for all $i \neq j$. The latter structure arises from a random subject effects model. Specifically, if $X_{jk} = s_k + e_{jk}$, where the s_k are i.i.d. $N(0, \sigma_s^2)$ subject effects and the e_{jk} are i.i.d. $N(0, \sigma_e^2)$ random errors then the common correlation $\rho = \sigma_s^2/(\sigma_s^2 + \sigma_e^2)$. Because the simulation results were similar in both cases, only the results for the autocorrelated data are reported here.

Five different sample sizes were studied: $n = 10, 15, 20, 30$ and 50. A nominal α level of 5% was used throughout. The multivariate t critical point approximations were used for all tests. The estimated correlation matrix $\{\widehat{\rho}_{ij}\}$ among the test statistics was used to determine the multivariate t critical points for $b = 3$ using the Genz and Bretz (1999) algorithm. For $b = 7$, this algorithm is too slow. Therefore we used an approximation obtained by replacing the $\widehat{\rho}_{ij}$ by a common correlation $\overline{\widehat{\rho}}$, equal to their arithmetic average. This approximation is known to be slightly conservative in case of the normal distribution (Hochberg and Tamhane, 1987, p. 146). It was calculated using the SAS PROBMC procedure.

The FWE simulations were conducted under the overall null hypothesis $H_0 : F_0 = F_1 = \cdots = F_b$, i.e., when all marginal distributions are identical. From Gabriel (1969, Theorem 2), it follows that the FWE is maximized under this configuration if the multiple test procedure is a UI procedure based on a joint testing family, which is the case here. The simulation results are summarized in Table 2. We see that the multiple t-test controls the FWE quite accurately for MVN and RMVN data, but is overly conservative for MVLN and MVCAUCHY data. The multiple sign test controls the FWE reasonably well for $n \geq 30$, but for smaller n, its FWE values are highly variable ranging from as low as 2.6% to as high as 9.2%. The multiple rank test controls the FWE quite well in most cases for $n \geq 20$, but for smaller n, its FWE is as high as 6.9%.

The powers of the three tests were simulated for MVN and MVLN data for $n = 30$ under location alternatives. It should be noted that for location models there is a one-to-one relationship between the effects θ_j and ψ_j on the one hand and the location shift on the other hand. Location shifts were created by adding a quantity equal to $\delta j/b$ to the jth time point ($j = 0, 1, \ldots, b$) for selected values of $\delta > 0$. This creates a linear shift w.r.t. time. The resulting power curves for the three tests are shown in Figure 2 for MVN data and in Figure 3 for MVLN data.

Figure 2 shows that for MVN data the multiple rank test is only slightly less powerful than the multiple t-test, while the multiple sign test is markedly less powerful. For MVLN data, however, the rank test is markedly more powerful than the t-test, which is even less powerful than the sign test. Hence we can conclude that the possible loss of power of the rank test w.r.t. the t-test is only small for MVN data, whereas the possible gain can be quite large. The sign test is less powerful than the rank test in general.

For two treatments we compare the multiple interaction tests proposed in Section 4, i.e., the multiple sign test, the test based on different joint and internal ranks in each group and the test based on overall ranks as given in Section 4. A parametric multiple t-test procedure was also included in the simulations. This procedure was based on the interactions $(\overline{X}_{1j\cdot} - \overline{X}_{10\cdot}) - (\overline{X}_{2j\cdot} - \overline{X}_{20\cdot})$. The covariances of these interactions were estimated by the corresponding sample covariances. The critical points for all tests were taken from multivariate t- distributions with $n_1 + n_2 - 2$ d.f. and appropriately estimated correlation matrices.

The same distributions as for the single treament simulations were used to generate the data. Location shifts were created by adding $0.5 + j\delta/b$ to each observation from Group 1 and $j\delta/b$ to each observation from Group 2 (so that there is a constant group effect of 0.5, and hence there is no treatment$\times$time interaction) where $\delta = 1$ represented a low time effect and $\delta = b$ represented a high time effect. The simulated type I FWE's for these two situations are displayed in Table 3. The results show that the sign test and the joint rank test maintain α quite accurately even for small sample sizes ($n_i \geq 10$) for a low time effect, but are very conservative for a high time effect. The parametric t-test is slightly conservative for MVLN and very conservative for MVCAUCHY data. Only the overall rank test is robust against different distributions and low or high time effects.

Comparing the power of the overall rank test and the t-test shows that the t-test is slightly more efficient if the differences $X_{ijk} - X_{i0k}$ have a symmetrical distribution whereas the rank test is much more powerful if the distribution is heavy-tailed or skew. The resulting power curves are comparable to those in the single treatment situation and therefore are not displayed.

6. Example

Let us return to the questions stated in the Introduction and answer them using the proposed tests. For each test we use $\alpha = .05$. The sample size of 15 per group is slightly on the low side for asymptotics to work, but the rank tests used below perform satisfactorily as seen in the simulation study. (The highest FWE of the rank test in the single treatment case for $b = 7$ and $n = 15$ is 6.2% and the highest FWE of the overall rank test in the two treatment case for $b = 7$ and $n = 15$ is 5.1%. Both these are for MVCAUCHY data.)

To answer Question 1 in the Introduction we computed the rank test statistics (3.2) and their two-sided p-values using the multivariate t approximation. The results are shown in Table 4. We see that the drug shows a significant effect relative to the baseline beginning with week 3. Note that when comparing week 10 with baseline, the empirical distributions of the data for the two time points are disjoint. In that case the variance estimate of the rank statistic equals 0 and hence the Z-statistic cannot be computed. However, it is clear that the difference is highly significant in that case.

To answer Question 2 we apply the same test used to answer Question 1 to the placebo group. The results are shown in Table 4. We see that there is no significant placebo effect at any time.

To answer Question 3 we apply the multiple test for treatment$\times$time interaction based on overall ranks. The results are shown in Table 4. We see that the active drug has a significant effect relative to the placebo beginning week 3.

Note that for the given example more powerful tests could be derived by using stepwise methods. For example a closure test could be applied to compare the time points to baseline successively beginning from the last. This method is reasonable if a monotone time effect can be assumed. However, it does not offer the possibility to compute simultaneous confidence intervals and it is rather non-robust if the assumption is not fulfilled. For example a time dependent placebo effect may result in an umbrella alternative that could not be detected.

7. Discussion

In this paper we have given nonparametric MCP's to compare time effects w.r.t. the baseline for a single treatment or two treatments. Although the approach does not offer a general solution to deal with repeated measures (for different objectives see, e.g., Reiczigel, 1999 or Keselman et al., 2001), it can easily be extended to some other MCP's, e.g., multiple comparisons between successive time points.

Missing values are not considered. However, the results could be extended to values missing completely at random by using the methods of Brunner, Munzel and Puri (1999), who among others generalized the approach of Akritas and Brunner (1997). Other nonparametric approaches to missing values (e.g., see Davis, 1991) compare the multivariate cdf's or are based on linear forms of the dependent components.

All results derived throughout the paper are asymptotic and the MCP's cannot be used with small sample sizes. The accuracy of the approximation depends on the sample sizes and the number of time points. Moreover, the accuracy of the approximation is affected slightly by the number of ties. Exact nonparametric tests or resampling methods are known only for independent observations. Therefore, several authors use linear forms (e.g. see Davis, 1991) or summary statistics (e.g. Weinberg and Lagakos, 2001) to condense the repeated measures.

As in the parametric theory, however, summary statistics as well as linear forms or multivariate tests do not offer the opportunity to distinguish between the treatment effects, time effects and interactions. This implies that they could not be used, e.g., to answer the questions 1-3. in the Introduction. Moreover many of the standard summary statistics cannot be applied to ordinal data (AUC, maximum change). Thus, new methodology as proposed in this paper is required.

Note that all tests are based on the asymoptotic rank transform (ART) method. The ARTs are known to have heterogenous variances and covariances even if the original observations are homogenous. Consequently all proposed methods are appropriate for homogenous as well as for heterogenous situations.

References

Akritas, M. G. and Arnold, S. F., 1994: Fully nonparametric hypotheses for factorial designs I: multivariate repeated measures designs. *Journal of the American Statistical Association*, **89**, 336–343.

Akritas, M. G. and Brunner, E., 1997: A unified approach to rank tests in mixed models. *Journal of Statistical Planning and Inference*, **61**, 249–277.

Brunner, E. and Denker, M., 1994: Rank statistics under dependent observations and applications to factorial designs. *Journal of Statistical Planning and Inference*, **42**, 353–378.

Brunner, E. and Langer, F., 1999: *Nichtparametrische Analyse ordinaler Daten.* Oldenbourg, München.

Brunner, E. and Langer, F., 2000: Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biometrical Journal*, **42**, 663-675.

Brunner, E., Munzel, U. and Puri, M. L., 1999: Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis*, **70**, 286–317.

Brunner, E., Puri, M. L. and Sun, S., 1995: Nonparametric methods for stratified two-sample designs with application to multi-clinic trials. *Journal of the American Statistical Association*, **90**, 1004–1014.

Cohen, A., Sackrowitz, H.B. and Sackrowitz, M., 2000: Testing whether treatment is 'better' than control with ordered categorical data: an evaluation of new methodoloy. *Statistics in Medicine*, **19**, 2699-2712.

Davis, C. S., 1991: Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine*, **10**, 1959-1980.

Gabriel, K. R., 1969: Simultaneous test procedures - Some theory of multiple comparisons. *Annals of Mathematical Statistics*, **40**, 224–250.

Genz, A. and Bretz, F., 1999: Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, **63**, 361–378.

GNEDENKO, B. V., 1962: *The Theory of Probability*, 2nd Edition. Chelsea Publishing, New York.

HOCHBERG, Y. and TAMHANE, A. C., 1987: *Multiple Comparison Procedures*, Wiley, New York.

KESELMAN, H.J., ALGINA, L. and KOWALCHUK, R.K., 2001: The analysis of repeated measures designs: a review. *British Journal of Mathematical and Statistical Psychology*, **54**, 1-20.

NEMENYI, P., 1963: *Distribution-free multiple comparisons.* Unpublished doctoral dissertation, Princeton University, Princeton, NJ.

REICZIGEL, J., 1999: Analysis of experimental data with repeated measures. *Biometrics*, **55**, 1059-1063.

RUYMGAART, F. H., 1980: A unified approach to the asymptotic distribution theory of certain midrank statistics. In J.P. Raoult (ed.): *Statistique non Parametrique Asymptotique.* Lecture Notes on Mathematics, **821**. Springer, Berlin, 1–18.

STEEL, R. G. D., 1959: A multiple comparison sign test: Treatment vs. control. *Journal of the American Statistical Association*, **54**, 767–775.

THOMPSON, G. L., 1991: A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, **86**, 410–419.

WEINBERG, J.M. and LAGAKOS, S. W., 2001: Efficiency comparisons of rank and permutation tests based on summary statistics computed from repeated measures data. *Statistics in Medicine*, **15**, 705-731.

ULLRICH MUNZEL
Department of Medical Biometrics and Datamanagement
Merz + Co. GmbH & Co.
Eckenheimer Landstr. 100-104
D-60318 Frankfurt/Main
Germany

AJIT C. TAMHANE
Department of Statistics
Northwestern University
2006 Sheridan Road
Evanston, Illinois 60208
USA

Table 1: CGI Values of Panic Disorder Patients

Active Drug Group									Placebo Group								
	Week									Week							
Patient	0	1	2	3	4	6	8	10	Patient	0	1	2	3	4	6	8	10
1	4	5	5	4	3	1	1	1	1	5	5	5	4	3	5	5	5
2	5	5	4	3	2	1	1	1	2	5	4	5	5	5	5	6	5
3	4	3	3	1	1	2	1	2	3	4	4	3	3	4	4	5	5
4	4	3	2	1	0	1	1	2	4	5	5	5	5	4	5	4	5
5	5	4	4	3	2	2	2	2	5	4	4	4	4	3	3	3	2
6	5	5	5	5	5	3	3	2	6	5	5	4	4	4	4	4	4
7	6	6	5	6	4	5	5	3	7	3	3	3	2	3	1	1	1
8	4	4	4	3	3	2	2	1	8	4	4	4	4	4	4	4	4
9	4	4	4	3	4	4	2	1	9	3	3	4	4	3	1	2	2
10	5	2	2	2	4	4	2	2	10	4	4	4	4	5	4	4	4
11	4	5	4	3	2	2	1	2	11	5	4	5	4	4	4	5	5
12	5	5	5	4	4	3	2	1	12	4	4	4	4	4	3	4	4
13	5	5	5	2	1	4	1	1	13	3	3	3	2	1	4	4	5
14	5	4	4	3	2	2	1	1	14	5	5	5	4	4	5	4	5
15	5	5	2	2	2	3	1	2	15	4	4	4	4	4	4	4	3

Table 2: Empirical Type I Familywise Error Rates of Multiple t, Sign and Rank Tests ($\alpha = 5\%$)

Distribution	n	$b = 3$			$b = 7$		
		t-Test	Sign Test	Rank Test	t-Test	Sign Test	Rank Test
MVN	15	5.5	4.8	4.8	5.6	9.2	5.7
	20	4.8	5.1	4.6	5.1	3.9	5.2
	30	5.4	5.7	5.1	5.2	4.6	5.1
RMVN	15	5.6	5.7	5.5	5.6	6.1	6.0
	20	5.4	5.4	5.3	5.2	6.1	5.6
	30	5.0	5.1	5.2	5.1	5.5	5.4
MVLN	15	3.3	4.8	5.2	2.5	9.1	5.5
	20	3.0	5.5	4.8	2.4	3.9	5.4
	30	3.3	5.5	4.8	2.7	4.8	5.5
MVCAUCHY	15	2.0	4.7	5.7	1.2	9.3	6.2
	20	2.0	5.8	5.5	1.1	3.7	5.3
	30	2.1	5.9	5.5	1.3	4.9	5.4

Table 3: Empirical Type I Familywise Error Rates of the Multiple Interaction Tests for $b = 7$ Time Points ($\alpha = 5\%$)

Distribution	$n_1 = n_2$	Low Time Effect				High Time Effect			
		t-Test	Sign Test	Joint Rank Test	Overall Rank Test	t-Test	Sign Test	Joint Rank Test	Overall Rank Test
MVN	15	4.9	5.4	4.6	5.1	5.3	0.9	1.0	4.9
	20	4.8	5.1	4.4	4.7	4.8	0.4	0.9	4.6
	30	4.8	5.2	4.7	4.9	4.9	0.7	1.0	4.3
RMVN	15	4.7	4.9	4.8	4.7	4.9	0.7	1.0	4.3
	20	4.9	5.2	5.0	5.1	4.9	0.6	1.1	4.4
	30	5.2	5.0	4.9	5.0	4.9	1.1	1.4	4.6
MVLN	15	3.4	4.9	4.1	4.5	3.4	1.0	1.7	4.7
	20	3.9	5.4	4.5	4.5	3.9	1.3	1.9	4.7
	30	3.7	5.1	4.4	4.5	4.3	2.3	2.7	4.8
MVCAUCHY	15	1.2	4.8	5.1	5.5	1.4	3.5	4.2	5.1
	20	1.3	4.9	5.4	5.5	1.2	4.2	4.2	5.0
	30	1.2	5.0	5.5	5.6	1.2	4.1	4.9	4.9

Table 4: Rank Test Statistics and Associated p-Values for Answering Research Questions in Panic Disorder Psychiatric Study

Question	Week	1	2	3	4	6	8	10
1	Z	0.952	2.698	5.294	8.880	9.457	10.980	–
	p	0.869	0.087	<0.001	<0.001	<0.001	<0.001	<0.001
2	Z	1.390	0.646	2.084	2.281	1.808	0.866	0.196
	p	0.636	0.976	0.256	0.186	0.385	0.918	>0.999
3	Z	0.362	2.046	3.614	3.463	5.247	6.750	7.499
	p	0.998	0.231	0.007	0.010	<0.001	<0.001	<0.001

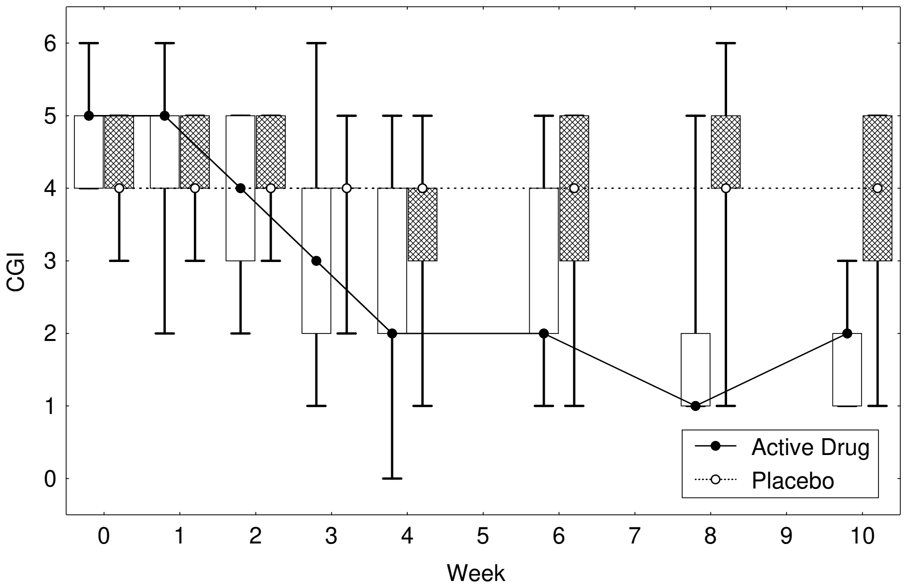

Figure 1: A Plot of the Panic Disorder Data

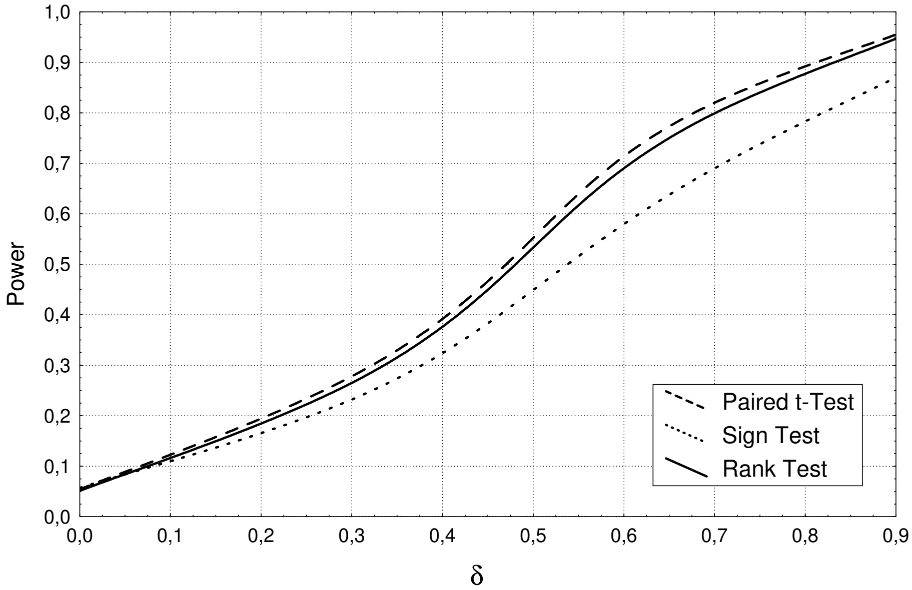

Figure 2: Power of the Multiple Sign Test, Multiple Rank Test and Multiple Paired t-Test for Multivariate Normal Data and $b + 1 = 4$ Time Points

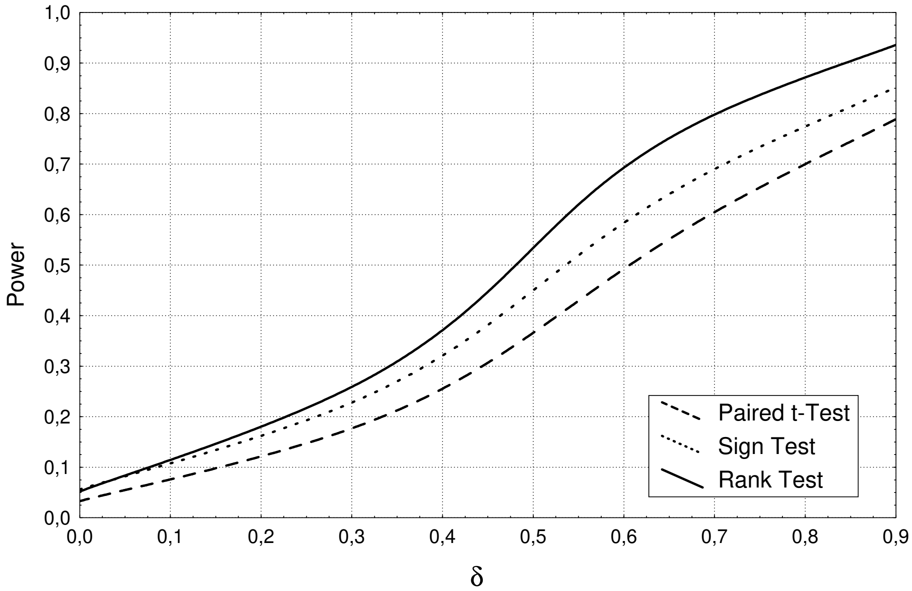

Figure 3: Power of the Multiple Sign Test, Multiple Rank Test and Multiple Paired t-Test for Multivariate Lognormal Data and $b + 1 = 4$ Time Points