Assessing the Similarity of Distributions – Finite Sample Performance of the Empirical Mallows Distance

By
Claudia Czado [†] and Axel Munk [*]

Abstract

The problem of assessing similarity of two cumulative distribution functions (c.d.f.'s) has been the topic of a previous paper by the authors (Munk & Czado (1995)). Here, we developed an asymptotic test based on a trimmed version of the Mallows distance (Mallows 1972) between two c.d.f.'s F and G. This allows to assess the similarity of two c.d.f.'s with respect to this distance at controlled type I error rate. In particular, this applies to bioequivalence testing within a purely nonparametric setting. In this paper, we investigate the finite sample behavior of this test. The effect of trimming and non equal sample size on the observed power and level is studied. Sample size driven recommendations for the choice of the trimming bounds are given in order to minimize the bias. Finally, assuming normality and homogeneous variances, we simulate the relative efficiency of the Mallows test to the (asymptotically optimal) standard equivalence t test, which reveals the Mallows test as a robust alternative to the standard equivalence t test.

Keywords: Scientific relevant difference, Mallows Distance, Model validation, Population bioequivalence, Goodness of fit.

1 Introduction

In many applications we are interested in showing that two independent samples arise from similar underlying populations F and G. As pointed out by Munk & Czado (1995) testing the hypothesis $H_0 : F = G$ versus $K_0 : F \neq G$ is not suitable for the assessment of similarity of populations, although this is common practice. Even when the observed p-value of a test for H_0 is large, this does not allow for a controlled error rate when assessing similarity. Therefore, Munk & Czado (1995) suggested interval hypotheses tests of the form

$$H : \Delta(F,G) > \Delta_0 \qquad \text{versus} \qquad K : \Delta(F,G) \leq \Delta_0, \tag{1}$$

where $\Delta(F,G)$ is an appropriate measure of discrepancy between F and G. In the above paper we recommended a trimmed version of the Mallows distance (Mallows, 1972) as such measure, since it

[†] Claudia Czado is Associate Professor for Statistics at York University, Canada. Parts of this paper were written while C. Czado was visiting the Sonderforschungsbereich 386 at the Institute for Statistics at the Ludwig-Maximilians-Universität in Munich. Their hospitality is greatly acknowledged. C. Czado was supported by research grant OGP0089858 of the Natural Sciences and Engineering Research Council of Canada.

[*] A. Munk is Assistant Professor at the Ruhr Universität Bochum, Fakultät und Institut für Mathematik, Universitätsstr. 150, D-44780 Bochum, Germany.

Address for correspondence: C. Czado, Department of Mathematics and Statistics, York University, 4700 Keele Street, North York, Ontario, Canada, M3J 1P3, email:czado@mathstat.yorku.ca.

depends on the entire distribution and not only on a difference in specific distributional characteristics such as the mean and/or variance. Observe, that testing (1) allows for the assessment of similarity of F and G at controlled error rate α. In order to obtain tests for the hypotheses in (1), we derived the asymptotic normal law for the corresponding trimmed empirical Mallows distance using quantile process theory (Munk & Czado 1995). For this a consistent estimate of the asymptotic variance was derived, which is not required when testing the classical hypothesis H_0 (see De Wet & Venter (1972) for the corresponding test in the one-sample case with F being a normal c.d.f.). In fact, this variance estimation turns out to be crucial for the performance of the test. Therefore, we report in Section 3.2 the results of a simulation study for the bias and mean squared error (MSE) of the empirical variance. From this study sample size driven recommendations for the choice of the trimming bound, in order to minimize bias and MSE, can be drawn.

In particular, our approach applies to bioequivalence testing of two formulations of an active ingredient without any parametric assumptions. This is denoted as population bioequivalence (Hauck & Anderson, 1992) and has been of particular interest in the recent literature (see also Schall (1995) for a test based on another measure of discrepancy). As recommended by the FDA (1992) and other drug agencies, it is sufficient to choose the sample sizes in bioequivalence studies rather small, say $n = m = 12$ when normally distributed errors can be assumed (see Chow & Liu (1992) for a description). In this situation the use of the two one-sided t test procedure is recommended, which will be denoted in the sequel as the standard test (FDA, 1992). Following these guidelines parametric analysis of bioequivalence should be performed in an additive two period crossover design because this allows for the control of the within subject variability. However, in a nonparametric setting, crossover designs bear significant difficulties for the interpretation of similarity in effects of interest. Therefore, Munk & Czado (1995) argued that a proper nonparametric analysis of bioequivalence studies should be based solely on two independent samples without repeated measurements for single individuals. When A. Munk was presenting these results at an invited talk in a session on nonparametric bioequivalence at the conference on 'Current Biometrical Issues on Longitudinal Observations in Medicine' in Düsseldorf (1995), the question of required sample sizes, when designing such an experiment turned out to be central in the subsequent discussion. Therefore, before the use of the Mallows equivalence test can be recommended to the working statistician, the behavior in small samples under various conditions on F and G has to be studied. In particular, we address through simulation the following questions:

Q1 How large have the sample sizes n and m to be chosen for the test to maintain its nominal level and to achieve reasonable power?

Q2 How does the test perform under normal distributions for F and G compared to the parametric standard test?

Four separate simulation studies have been designed to investigate the above questions. The first three studies assume symmetric (normal) samples and different trimming constants. In the first two studies, equal sample sizes ($n = m$) have been assumed, while unequal sample ($n \neq m$) sizes are investigated in the third study for large sample sizes ($m, n \geq 50$). The second study ($n, m \leq 25$) focuses on the small sample behavior. Here (nearly) equal sample sizes are strictly recommended. To answer Q2, the relative efficiency of the Mallows test and the (asymptotic optimal) standard bioequivalence test is simulated. Surprisingly, we find that the Mallows equivalence test is (finite) more efficient in most cases than the equivalence t test although being asymptotically less efficient. This finding can be explained by the fact that the power of the latter test tends uniformly to zero as the variance increases as criticized by many authors (cf. Müller-Cohrs (1990), Brown, Hwang & Munk (1995) among others).

Chow & Tse (1990) pointed out that especially in bioequivalence studies outliers may affect the data analysis drastically when using the standard test. In contrast, the Mallows equivalence test represents a robust tool to protect against outliers by the flexible choice of the trimming bound. In Section 3.4.2 it becomes apparent that the observed relative efficiency compared to the standard bioequivalence test is even slightly increased when trimming is increased. We found however, that this is caused by the increasing liberalness of the test as trimming is increased. We are therefore interested in answering the following

Q3 What is a reasonable choice of the trimming bound α?

Munk & Czado (1995) obtained precise requirements regarding the shape of the underlying c.d.f.'s F and G in order to guarantee the quality of asymptotic law for the empirical Mallows distance. In particular, strong peaked densities and heavy tailed distributions are to be expected to affect on this. Therefore, the following final question is addressed by a subsequent simulation study.

Q4 How does the test perform under nonsymmetric nonnormal distributions for F and G?

To evaluate the behavior of the equivalence test for nonsymmetric nonnormal populations, generalized logistic distributions (Czado, 1992) have been utilized because they exhibit a wide range of skewness and tail patterns.

The following recommendations summarize the observed finite sample behavior presented in Section 3. When both samples arise from a symmetric distribution, such as the normal distribution, a sample size of $m = 20$ and a sample size ratio of $1/2 \leq \frac{m}{n} \leq 2$ is sufficient to maintain its nominal level quite accurately, while guaranteeing reasonable power. In addition, a trimming of maximal 10% is preferable over a larger trimming when small samples are used, otherwise the test becomes too liberal.

Further, we find that the Mallows equivalence test turns out to be always liberal rather than conservative, especially when small sample sizes are present. In addition, the estimated asymptotic variance of the empirical Mallows distance was found to be positively biased in small samples (cf. Section 3.2). The bias of the actual level was always found to be maximal when the tolerance bound $\Delta_0 \leq 1$. This value may serve as a change point of accuracy. A larger tolerance bound implies a highly accurate approximation of the nominal level whereas a smaller tolerance bound leads to a liberal test.

If one of the samples comes from a distribution, which is highly skewed compared to the other one, sample sizes of $n, m \geq 50$ may become necessary as well as a larger trimming ($\alpha \geq 10\%$).

Following the above recommendations, the trimmed Mallows equivalence test can be used to assess equivalence of two populations within a purely nonparametric setting even when the underlying distributions are highly skewed. Although larger trimming will increase the efficiency in many cases, we still recommend only slight trimming which may be increased when sample size increases. Otherwise the actual level may become too liberal.

The paper is organized as follows. Section 2 introduces the trimmed Mallows distance and summarizes briefly its asymptotic properties necessary for the construction of the critical region of the equivalence test. Section 3 describes the simulation design and presents the finite sample results under the various designs described above. The paper closes with a discussion section.

2 The Equivalence Test Based on the Trimmed Mallows Distance

In this section, we first introduce the trimmed Mallows distance and briefly review its properties including asymptotic results. Finally, we give the critical region of the equivalence test to be studied. For additional details and the derivation of the asymptotic distribution the reader is referred to Munk & Czado (1995).

Throughout this paper, we have an i.i.d. sample $\{X_1, \cdots, X_m\}$ and $\{Y_1, \cdots, Y_n\}$ from F and G, respectively, available. Further, F and G are assumed to be continuous c.d.f.'s, with finite second moments. $F_m(x) = m^{-1} \sum_{i=1}^m 1_{(-\infty,x]}(X_i)$ denotes the empirical distribution function of the i.i.d. sample $X_1, \cdots, X_m \sim F$. The trimmed Mallows distance between F and G is now defined as

$$\Gamma_\alpha(F,G) \quad := \quad (1-2\alpha)^{-1} \left\{ \int_\alpha^{1-\alpha} |F^{-1}(u) - G^{-1}(u)|^2 du \right\}^{1/2}. \tag{2}$$

Here $\alpha \in [0, 1/2)$ denotes a trimming bound. If $\alpha = 0$, we write Γ for Γ_α. Note, that in this case, $\Gamma^2(F,G)$, measures the square area of difference between the quantile functions F^{-1} and G^{-1}.

In particular, Munk & Czado (1995) showed that a small Mallows distance between F and G implies similarity of the (trimmed) first two moments which reveals Γ_α as an appropriate measure of bioequivalence (see also Holder & Hsuan (1994)).

In the special case of symmetric location-scale families $F = H((x-\eta)/\tau)$ and $G = H((x-\mu)/\sigma)$ with $\sigma, \tau > 0$, we find that

$$\Gamma_\alpha(F,G) = \frac{1}{(1-2\alpha)^{1/2}} \left\{ (\mu-\eta)^2 + (\sigma-\tau)^2 (1 - 2\frac{u_{1-\alpha} h(u_{1-\alpha})}{1-2\alpha}) \right\}^{1/2}. \tag{3}$$

Here, h denotes the density of H and u_α the α-quantile of H, i.e. $H(u_\alpha) = \alpha$. This shows that trimming puts slightly more weight on the mean difference than on the difference in scales. This is to be expected since extreme tails are discarded when trimming is used. Figure 1 gives the corresponding Mallows distance contours in the case of H being the standard normal distribution when 0%, 10% and 20% trimming is used. The effect of trimming is to increase the Mallows distance by a factor of $(1-2\alpha)^{-1/2}$ in location scale models if F and G have equal variances, i.e. (3) reduces to

$$\Gamma_\alpha(F,G) \quad = \quad \frac{1}{(1-2\alpha)^{1/2}} |\mu - \eta|. \tag{4}$$

One problem commonly encountered in bioequivalence studies is the presence of outliers when a normal model, which is a special symmetric location scale family, is used. We mention, that outliers change the results drastically obtained by standard tests (Chow & Tse, 1990). Therefore, the choice of the trimming bound is of particular interest. Based on the simulation results presented in Section 3 recommendations will be given. This provides an excellent tool to robustify against outliers. Observe, that trimming changes (under the normal assumption with homogeneous variances) the bioequivalence criterion by a factor $(1-2\alpha)^{-1/2}$ which can simply be adjusted by rescaling the hypotheses (cf. Section 3.4.1).

Following Munk & Czado (1995) a consistent estimate for the trimmed Mallows distance is obtained, when F and G in (2) are replaced by their empirical counter parts F_m and G_n, respectively. In particular, we have for $n = m$ and $\alpha = 0$

$$\hat{\Gamma}^2 := \Gamma(F_n, G_n)^2 = \frac{1}{n} \sum_{i=1}^n (X_{(i)} - Y_{(i)})^2,$$

where $X_{(i)}$ and $Y_{(i)}$ denote the i-th order statistic of the i.i.d. sample arising from F and G, respectively. In the case of unequal sample sizes ($n \neq m$) and the presence of trimming ($\alpha > 0$) the expressions for $\hat{\Gamma}_\alpha := \Gamma_\alpha(F_m, G_n)$ in terms of the order statistics $X_{(i)}$ and $Y_{(i)}$ are more complex, but remain tractable. The exact expressions are given in the Appendix of Munk & Czado (1995).

In the above paper, we constructed an asymptotic test for the equivalence testing problem

$$H : \Gamma_\alpha(F,G) > \Delta_o \text{ versus } K : \Gamma_\alpha(F,G) \leq \Delta_o \tag{5}$$

by utilizing that

$$\mathcal{L}\left\{\left(\frac{nm}{n+m}\right)^{\frac{1}{2}} \left(\Gamma_\alpha^2(F_m,G_n) - \Gamma_\alpha^2(F,G)\right)\right\} \longrightarrow \mathcal{N}(0,\sigma_\alpha^2) \qquad \text{as} \qquad n,m \to \infty \tag{6}$$

is normal with mean 0 and variance

$$\begin{aligned}\sigma_\alpha^2 \;=\; & \frac{4}{(1-2\alpha)^4}\left\{\lambda\left[\int_0^{1-\alpha}\left\{\int_{\alpha\vee s}^{1-\alpha} h_{(F,G)}(t)dt\right\}^2 ds - \left(\int_0^{1-\alpha}\left\{\int_{\alpha\vee s}^{1-\alpha} h_{(F,G)}(t)dt\right\} ds\right)^2\right]\right.\\ & \left.+(1-\lambda)\left[\int_0^{1-\alpha}\left\{\int_{\alpha\vee s}^{1-\alpha} h_{(G,F)}(t)dt\right\}^2 ds - \left(\int_0^{1-\alpha}\left\{\int_{\alpha\vee s}^{1-\alpha} h_{(G,F)}(t)dt\right\} ds\right)^2\right]\right\},\end{aligned}$$

where $\lambda \in (0,1)$ is the limiting value of $\frac{n}{n+m}$ and

$$h_{(F,G)}(t) := \frac{F^{-1}(t) - G^{-1}(t)}{f \circ F^{-1}(t)}.$$

Here, regularity conditions A1-A4 of Munk & Czado (1995, p. 5) for the densities f of F and g of G have to be satisfied. Explicit expressions for a consistent estimate $\hat{\sigma}_\alpha^2 := \sigma_\alpha^2(F_m, G_n)$ of the asymptotic variance (see Appendix A of Munk & Czado (1995)) are available. The estimate we have used is obtained again by 'plugging in' the empirical c.d.f's in σ_α^2 and evaluating the resulting expression as a Riemannian sum. More general results for trimming bounds depending on sample size have been obtained as well. We found that the limit law remains true, when trimming is neglected asymptotically at a rate of at most $\alpha_n \sim n^{-1}\log\log n$. Hence, it is one aim of the paper to provide a sample size driven guide for a good choice of the trimming bounds (see Section 3.4).

The critical region of the equivalence test for H against K in (5) can now be expressed as

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2}} \frac{\hat{\Gamma}_\alpha^2 - \Delta_0^2}{\hat{\sigma}_\alpha} \leq z_\alpha \tag{7}$$

where z_α denotes the α-quantile of the standard normal distribution. Splus functions to compute $\hat{\Gamma}_\alpha$ and $\hat{\sigma}_\alpha^2$ have been written to perform this test and can be obtained from the authors on request.

3 Simulation

3.1 Simulation Design

To investigate the questions raised in the introduction, five different simulation studies were designed. To study the effect of trimming, several trimming sizes have been used. The aim of the first study is the investigation of bias and MSE of the variance estimate $\hat{\sigma}_\alpha^2$ under various conditions on F and G. Primary goal of the next three simulations is to evaluate the performance of the test (7) when normal c.d.f.'s are used for F and G, respectively. We investigated the case of homogeneous variances (remember, that this is the standard assumption in bioequivalence trials) as well as the inhomogeneous case. In the second study, larger sample sizes (≥ 50) have been used, while small sample sizes (≤ 25) are investigated in the third study. The fourth simulation is performed with unequal sample sizes ($n \neq m$), while still using normal distributions for F and G. The aim of the last simulation is to investigate the performance under nonnormal nonsymmetric populations. We were particularly interested in the effect of skewness and heavy tails of the distributions. because these situations are expected to decrease the accuracy of the approximation of the finite sample distribution for $\hat{\Gamma}_\alpha^2$ given by the limit law in (6) (cf. Munk & Czado (1995)).

In the first four studies, F is assumed to be standard normal $N(0,1)$, while G is a normal with mean μ and variance σ^2. For second and and fourth study, mean values of $\mu = .1, .3, .5, .9, 1, 1.1, 1.2$ and 1.3 and standard errors of $\sigma = .5, 1.$ and 1.5 have been chosen. For the third study, larger mean values $\mu = 1, 1.2, 1.4, 1.6, 2.0, 2.2, 2.4, 2.6$ have been selected to allow for adequate power.

Sample sizes $n = m = 50, 100, 200$ are used in the second study, $n = m = 10, 15, 20, 25$ for the third study and $n = 100$ and $\frac{m}{n} = .9, .75, .5, .3$ are assumed for the fourth one. In addition, the effect of 0%, 10% and 20% trimming has been studied in the second simulation, while 0% and 10% trimming have been used in the fourth study because the results of the first study show that trimming of 20% was too large. For the third study, sample sizes were too small to allow for a 10% trimming, therefore only 0% and 20% trimming were investigated. For each combination of sample size, mean and standard error values, two samples, one of size m from a standard normal population and one of size n from a normal population with mean μ and variance σ^2, have been generated and the corresponding equivalence test of H versus K using equivalence bound Δ_o and trimming constant α at significance level $\alpha_s = .05$ has been performed. Results from the first three studies have been based on 500 replications, while the fourth study was based on 250 replications. Here, $\Delta_o = .3, .5, .7, .9, 1$ and 1.2 was used in the second and fourth study, and $\Delta_o = 1.2, 1.4, 1.6, 1.8, 2.0$ and 2.2 in the third study.

As already mentioned in the case of normal samples with equal variances, the equivalence t test is a valid alternative to the Mallows equivalence test. If we define $\Delta_\alpha = (1-2\alpha)^{-1/2}\Delta_0$, the rejection region of the equivalence t test for

$$H_\alpha : |\mu - \eta| > \Delta_\alpha \text{ versus } K_\alpha : |\mu - \eta| \leq \Delta_\alpha \tag{8}$$

is given by (Schuirmann, 1987)

$$\left\{\Delta_\alpha \geq |\overline{X} - \overline{Y}| + t_{n+m-2,1-\alpha}\frac{S}{\sqrt{n+m-2}}\right\},$$

where $t_{f,\alpha}$ denotes the α-quantile of the central t distribution with f degrees of freedom. Further, $S^2 = (\frac{1}{n} + \frac{1}{m})\left[\sum_{i=1}^n (X_i - \overline{X})^2 + \sum_{i=1}^n (Y_i - \overline{Y})^2\right]$ denotes the pooled sample variance and $\overline{X}(\overline{Y})$ the average of the sample $X_1, \cdots, X_m, (Y_1, \cdots, Y_n)$. Note from (4), that the testing problem (5) for the trimmed Mallows distance under normal populations with equal variances σ^2 and mean differences $\mu - \eta$ turns into (8), when we parametrize the problem in $|\mu - \eta|$.

In a final simulation study, the properties of the Mallows equivalence test H versus K are investigated when the underlying distributions are skewed. For this study, we considered a family of generalized logistic distributions $\{F_\psi, \psi \in I\!R\}$ (Czado (1992)) with heavier right tail ($\psi < 1$) and lighter right tail ($\psi < 1$) than the logistic distribution ($\psi = 1$). In particular, the distribution functions F_ψ are given by

$$F_\psi(x) = \frac{\exp(h_\psi(x))}{1 + \exp(h_\psi(x))},$$

where

$$h_\psi(x) = \begin{cases} \frac{(x+1)^\psi - 1}{\psi} & \text{if } x > 0 \\ x & \text{otherwise} \end{cases} .$$

Figure 2 presents the corresponding c.d.f.'s and densities for several ψ values of the generalized logistic distribution family. This shows, that this family includes a wide range of skewed distributions. In particular, it includes highly peaked ($\psi \gg 1$) and heavy tailed ($\psi \ll 1$) distributions. A similar modification of the left tail or both tail modification is possible as well. The Mallows distance to the logistic c.d.f. increases rapidly as $\psi < 1$ becomes smaller and slower as $\psi > 1$ (see Figure 3). The effect of a larger trimming is to decrease the Mallows distance.

For our simulation, F is assumed to be the logistic c.d.f. and G a generalized logistic c.d.f. with one of the following values for $\psi = .1, .15, .25, .5, .75.9, 1, 1.25, 1.5, 2, 5, 10$. Sample sizes of $n = 100, 50, 25$ for F and m for G, such that $\frac{m}{n} = .96, .48$, were considered. In addition, 8% and 16% trimming were used and 250 equivalence tests with $\Delta_0 = .5, .75$ for $\psi > 1$ and $\Delta_0 = .75, 1, 1.25, 1.5$ for $\psi < 1$ were performed. It should be noted that for $n = 25(50)(100)$, a trimming of 8% (4%) (2%)is the smallest positive trimming possible corresponding to discarding the largest and smallest value in the sample.

Table 1 summarizes the different simulation designs used.

Study 1: Simulated Bias and MSE of Mallows Sample Variance under Normality					
Sample sizes and trimming bounds as in Study 2 and 4.					
Study 2: Large Normal Samples with Equal Sample Sizes					
F	G	n	$\frac{m}{n}$	α	Δ_o
$N(0,1)$	$N(\mu, \sigma^2)$	50	1	0	$\Delta_o = .3, .5, .7, .9, 1, 1.2$
	$\mu = .1, .3, .5, .9, 1$	100		.05	
	$= 1.1, 1.2, 1.3$	200		.1	
	$\sigma = .5, 1, 1.5$				
Study 3: Small Normal Samples with Equal Sample Sizes					
F	G	n	$\frac{m}{n}$	α	Δ_o
$N(0,1)$	$N(\mu, \sigma^2)$	10	1	0	$\Delta_o = 1.2, 1.4, 1.6, 1.8, 2.0, 2.2$
	$\mu = 1, 1.2, 1.4, 1.6, 1.8$	15		.1	
	$= 2., 2.2, 2.4, 2.6$	20			
	$\sigma = .5, 1, 1.5$	25			
Study 4: Normal Samples with Unequal Sample Sizes					
F	G	n	$\frac{m}{n}$	α	Δ_o
$N(0,1)$	$N(\mu, \sigma^2)$	50	.9	0	.7
	$\mu = .1, .3, .5, .9, 1$	100	.75	.05	.9
	$= 1.1, 1.2, 1.3$	200	.5		1
	$\sigma = .5, 1, 1.5$		.3		1.2
Study 5: Nonnormal Samples					
F	G	n	$\frac{m}{n}$	α	Δ_o
logistic	generalized logistic	25	.96	0.04	$\Delta_o = .5, .75(\psi > 1)$
$\psi = 1$	$\psi = .1, .15, .25, .5, .75, .9$	50	.48	.08	$\Delta_o = .75, 1, 1.25, 1.5(\psi < 1)$
	$= 1, 1.25, 1.5, 2, 5, 10$	100			

Table 1: Summary of Simulation Designs

All simulations were performed in S+ Version 3.2 on medium sized IBM RS6000 Unix workstations, where up to 500 replications for each setting were conducted, respectively. Observe, that the required CPU time is considerable, for example 5.2 hours (34 min) of CPU time was required for a single parameter setting with sample sizes $n = m = 200 (n = m = 20)$ and 500 replications.

3.2 Performance of the Mallows sample variance

To investigate the performance of the estimated asymptotic variance of the empirical Mallows distance, $\hat{\sigma}^2_\alpha$, its estimated standardized bias

$$SBIAS\left[\hat{\sigma}^2_\alpha\right] = E\left[\frac{\hat{\sigma}^2_\alpha}{\sigma^2_\alpha}\right]$$

and its mean squared error (MSE)

$$MSE\left[\hat{\sigma}^2_\alpha\right] = E\left[\hat{\sigma}^2_\alpha - \sigma^2_\alpha\right]^2$$

have been displayed in Tables 2 and 3.

			$\alpha = 0$			$\alpha = 0.05$			$\alpha = 0.1$		
			μ			μ			μ		
σ	n=m		.1	.3	.5	.1	.3	.5	.1	.3	.5
.5	20	SBIAS	1.54	1.22	1.09	2.73	2.09	1.42	3.44	1.77	1.43
		MSE	.30	.35	.70	.46	1.24	1.80	.72	1.25	3.41
	50	SBIAS	1.21	1.14	1.02	1.96	1.54	1.29	1.87	1.26	1.11
		MSE	.08	.13	.31	.13	.30	.61	.10	.27	.69
	100	SBIAS	1.15	1.08	1.04	1.37	1.15	1.01	1.49	1.12	1.02
		MSE	.04	.06	.15	.03	.08	.18	.04	.13	.32
	200	SBIAS	1.09	1.01	1.03	1.21	1.09	.99	1.24	1.05	.94
		MSE	.02	.03	.07	.01	.04	.09	.01	.06	.14
1	20	SBIAS	16.17	2.62	1.50	12.94	2.55	1.42	13.43	2.26	1.46
		MSE	.94	1.36	2.33	1.19	2.89	4.22	2.42	4.20	8.80
	50	SBIAS	6.92	1.63	1.21	5.51	1.37	1.09	5.25	1.39	1.12
		MSE	.15	.32	.96	.21	.49	1.25	.31	.98	2.06
	100	SBIAS	4.23	1.44	1.13	3.42	1.14	.98	3.17	1.13	1.01
		MSE	.05	.18	.36	.08	.21	.57	.11	.40	1.10
	200	SBIAS	2.46	1.12	1.08	2.00	1.02	.94	1.40	1.10	.96
		MSE	.01	.07	.18	.02	.11	.29	.03	.21	.52
1.5	20	SBIAS	2.68	2.16	1.61	4.36	2.58	1.59	5.63	2.57	1.62
		MSE	8.74	12.25	16.91	11.43	17.56	22.59	16.72	26.16	38.31
	50	SBIAS	1.74	1.42	1.25	2.19	1.31	1.04	2.68	1.52	1.19
		MSE	2.07	2.55	5.45	1.69	2.26	4.80	2.27	5.47	9.00
	100	SBIAS	1.40	1.30	1.15	1.67	1.19	1.04	1.83	1.19	1.06
		MSE	.65	1.37	2.28	.65	1.16	2.71	.73	1.80	5.04
	200	SBIAS	1.19	1.10	1.10	1.26	1.05	.97	1.40	1.10	.96
		MSE	.23	.52	1.02	.16	.54	1.35	.20	.98	2.35

Table 2: Standardized bias and MSE of the estimated asymptotic variance of the empirical Mallows distance $\hat{\sigma}^2_\alpha$ for equal sample sizes.

We are particularly interested in the behavior of $\hat{\sigma}_\alpha^2$ when the underlying true Mallows distance is small. Table 2 reports the standardized bias and the MSE of $\hat{\sigma}_\alpha^2$, when $\Gamma_\alpha(F,G)$ ranges between .1 and .71. With regard to the standardized bias, we observe that in small samples ($n = m = 20$) with small underlying true $\Gamma_\alpha(F,G)$ ($\mu = .1, .3, \sigma = 1$), $\hat{\sigma}_\alpha^2$ is extremely positively biased (SBIAS $\gg 1$). This has to be expected, since the limit law (6) fails to be valid when $\Gamma_\alpha(F,G)$ approaches zero. However, a doubling of the sampling size reduces the bias by roughly 50%. For the homogeneous variance case ($\sigma = 1$), $\Gamma_\alpha(F,G)$ increases with trimming thus explaining the reduction in bias when trimming is increased. For the nonhomogeneous variance case ($\sigma \neq 1$), the effect of trimming is more complicated, a larger trimming decreases (increases) $\Gamma_\alpha(F,G)$ for small (large) μ (cf. Figure 1). This is the reason why the standardized bias increases with trimming for small μ. The opposite effect was seen for large μ and $\sigma \neq 1$ (not presented). Finally, note that for unequal variances ($\sigma \neq 1$) the standardized bias is smaller when $\sigma < 1$ compared to $\sigma > 1$ despite the same underlying true Mallows distance. With regard to the MSE of $\hat{\sigma}_\alpha^2$, the MSE decreases rapidly as sample size increases. Further, the MSE increases with increasing $\Gamma_\alpha(F,G)$, i.e. as μ increases for fixed σ.

The standardized bias of the asymptotic variance $\hat{\sigma}_\alpha^2$ has been reported in Table 3 for unequal sample sizes where $n = 100$ and $m = 30, 50, 75, 90$. First, the absolute standardized bias reduces as m increases. Again, we see that the standardized bias is substantially overestimated when the true underlying Mallows distance $\Gamma_\alpha(F,G) \leq .1$ ($\sigma = 1, \mu = .1$). The effect of trimming is similar to the one observed in the equal sample size case. With regard to the MSE of $\hat{\sigma}_\alpha^2$, it also decreases as m increases. Further, the MSE and $SBIAS$ increases with increasing trimming bound α when the variances are very inhomogeneous.

In summary, the asymptotic variance estimator performs well when the true Mallows distance is not too small ($\Gamma_\alpha(F,G) > 1$). In this case, a 10% trimming is preferable over 0% trimming in order to minimize bias. For small $\Gamma_\alpha(F,G)(\leq 1)$, trimming is only preferable in the homogeneous variance case ($\sigma = 1$). In addition, higher sample sizes for $n = m$ are required. In the case of nonequal sample sizes the same conclusions can be drawn, in addition to require $1/2 \leq \frac{m}{n} \leq 2$ to reduce bias.

			$\alpha = 0$			$\alpha = 0.05$		
	n=100		μ			μ		
σ	m		.1	.3	.5	.1	.3	.5
.5	30	SBIAS	.68	.61	.60	1.00	.72	.56
		MSE	.06	.17	.43	.05	.16	.60
	50	SBIAS	.86	.78	.80	1.11	.83	.71
		MSE	.05	.09	.28	.04	.10	.35
	75	SBIAS	.98	1.01	.96	1.30	.99	.86
		MSE	.03	.10	.21	.03	.09	.20
	90	SBIAS	1.17	1.10	1.06	1.32	1.01	.96
		MSE	.05	.08	.17	.03	.07	.17
1	30	SBIAS	7.24	1.79	1.39	5.70	1.29	1.00
		MSE	.16	.56	1.06	.27	.53	1.22
	50	SBIAS	5.48	1.61	1.28	4.91	1.26	1.00
		MSE	.09	.27	.75	.19	.35	.88
	75	SBIAS	4.65	1.30	1.21	3.84	1.21	.98
		MSE	.07	.18	.47	.09	.29	.69
	90	SBIAS	4.51	1.52	1.25	3.14	1.20	1.02
		MSE	.06	.23	.53	.06	.28	.79
1.5	30	SBIAS	2.34	2.08	2.12	3.06	2.02	1.62
		MSE	1.99	3.96	12.05	2.38	5.10	8.24
	50	SBIAS	1.77	1.79	1.58	2.11	1.56	1.32
		MSE	.97	2.81	6.08	1.07	2.55	4.90
	75	SBIAS	1.44	1.36	1.20	1.82	1.42	1.14
		MSE	.68	1.33	3.10	.64	2.12	3.60
	90	SBIAS	1.30	1.32	1.11	1.59	1.19	1.11
		MSE	.58	1.52	1.95	.53	1.00	3.26

Table 3: Standardized bias and MSE of the estimated asymptotic variance of the Mallows distance $\hat{\sigma}^2_\alpha$ for unequal sample sizes where $n = 100$.

3.3 Performance Measures and their Graphical Display of the Power Study

After conducting the equivalence test for each parameter combination and replication, the observed power and significance level was calculated. In simulation studies 2-4, the observed power is a function of the mean μ and the standard error σ, of the sample sizes n and m, the trimming bounds α and finally of the equivalence bound Δ_o. For fixed σ, α and Δ_o the observed significance level and the power at the alternative $k = .75\Delta_o$ are displayed in Figures 4-5 and 7-10.

Finally, in the case of normal samples with equal sample size and equal variances ($\sigma = 1$), the Mallows equivalence test has been compared to the equivalence t test by simulating the observed relative efficiency, i.e. the ratio of required sample sizes to obtain the same power β at a fixed alternative k. For the testing problem (5), a trimming size α and an equivalence bound Δ_o have to be chosen, thus the observed relative efficiency will depend on α and Δ_o as well and may be expressed as

$$RE^{(Mallows:t\ test)}_{\alpha,\Delta_o}(\beta, k) = \frac{n^{Mallows}_{\alpha,\Delta_o}(\beta, k)}{n^{ttest}_{\alpha,\Delta_o}(\beta, k)},$$

where $n_{\alpha,\Delta_o}^{Mallows}$ is the sample size required by the Mallows test for (5) to achieve an observed power of β at alternative k, and $n_{\alpha,\Delta_o}^{t\ test}$ is the sample size required by the equivalence t test for the testing problem (8) to achieve by (8) an observed power of β at alternative $k = |\mu - \eta|(1-2\alpha)^{-1/2}$. A value of $RE_{\alpha,\Delta_o}^{Mallows:t\ test} < 1$ indicates that the Mallows test is more efficient than the equivalence t test. For example, a value of $RE_{\alpha,\Delta_o}^{Mallows:t\ test} = .5$ can be interpreted as the equivalence t test requiring double the sample size as the Mallows test to achieve the same power.

For the last simulation study, the considered underlying populations are distinguished by the single tail parameter ψ. Hence the power functions depend only on this parameter in addition to α and Δ_o.

3.4 Simulation Results

3.4.1 Results for Large Normal Samples of Equal Sample Size

As discussed in the previous section, results are summarized by observed significance level (Figure 4) and the power at the alternative $k = .75\Delta_0$ for the Mallows test (Figure 5) and at $k = (1-2\alpha)^{-1/2}.75\Delta_0$ for the t test (Figure 6).

From Figure 4, we see that the actual level of the Mallows test approximates the nominal level of significance $\alpha_s = 0.05$ better as the sample size increases. Further the accuracy of approximation increases as the tolerance bound Δ_0 increases. Deviation of the variances between F and G will decrease the approximation, in particular for smaller sample sizes as 50.

Large trimming should only be used when a tolerance bound $\Delta_0 \geq 1$ is to be tested or when the variances are nearly homogeneous, otherwise the test may become too liberal. In this case, even a larger sample size does not improve the situation sufficiently. Recall, however, that usually in bioequivalence studies homogeneity of the variances is justified, and therefore trimming is an excellent tool to protect against outliers. It is interesting, that the bound $\Delta_0 = 1$ serves as a 'change point' indicating the quality of the performance. For values larger than 1 the finite sample approximation is significantly superior compared to small values. This may be explained by the quadratic structure of the statistic $\hat{\Gamma}_\alpha^2$. In particular, when Δ_0 tends to zero, rather large sample sizes are required to keep the nominal level exactly. Otherwise, the actual level tends to be larger than the nominal level. This is supported by the fact that the asymptotic normal law of the empirical Mallows distance degenerates to a Dirac measure with mass at 0 (for more details see Munk & Czado (1995)). However, the test remains consistent although being liberal. Note, that the performance of the asymptotic variance estimate $\hat{\sigma}_\alpha^2$ is consistent with these observations. From Figure 4 precise required sample sizes for decreasing bounds Δ_0 can be drawn in order to adjust the actual level.

With regard to the observed power in the alternative $|\mu - \eta| = .75\Delta_0$ (see Figure 5), the power always increases as sample size increases. Further, as expected, a larger tolerance constant Δ_0 increases the power as well. In addition, the power is higher for smaller nonhomogeneous variances ($\sigma < 1$) than for larger nonhomogeneous variances ($\sigma > 1$).

From the above results for the observed significance level and the attained power, we recommend to use a slight trimming $\alpha \leq 0.05$. Only when the variances in both groups may be assumed as nearly homogeneous, a larger trimming may be suitable because otherwise the liberalness of the test may become too drastic.

In Figure 6 the observed power at corresponding alternatives under variance homogeneity ($\sigma = 1$) of the equivalence t test is shown. We find the surprising fact, that the power is lower in most cases for the equivalence t test compared to the Mallows test, particularly for moderate Δ_0 values. The specific choice of the trimming bound α does not influence this observation. These findings are somewhat curious because we have a proof that the t test is an asymptotic uniformly most powerful invariant test (invariance with respect to the group of translations). This paradox may be explained as follows: Although asymptotically optimal, the t test may become arbitrarily inefficient for finite sample size as the variance increases (Munk, 1993). Only, when the variance is rather small compared to the bound Δ_0, say $\sigma^2/\Delta_0 << 1$ the asymptotic optimality result allows an interpretation in realistic sample sizes, say $n, m \leq 200$. This behavior is reflected precisely in the observed relative efficiency results presented in Table 3. The numbers in parentheses represent the sample size required by the Mallows test to achieve power β at the alternative $k = .75\Delta_o$. The significance level of the tests was always $\alpha_s = .05$. Observe, that the efficiency may be increased by increasing the trimming bound, however, this is due to the rather large liberalness in this case (cf. Figure 4)

Equal Variances: $\sigma = 1$						
		Δ_0				
α	β	.5	.7	.9	1	1.2
.0	.6	***	***	.85 (134)	.77 (96)	.98 (82)
	.4	.96 (176)	.95 (114)	.84 (70)	.80 (50)	.81 (38)
.05	.6	***	**	.89 (144)	.88 (116)	.90 (74)
	.4	.82 (151)	.73 (87)	.78 (67)	.83 (55)	.65 (30)
.1	.6	***	*	.99 (160)	.91 (110)	.91 (74)
	.4	.76 (150)	.75 (90)	.91 (70)	.90 (51)	.71 (31)

Table 3: Observed relative efficiency of the Mallows equivalence test and the equivalence t test with $k = .75\Delta_o$ (* equivalence t test requires $n > 200$ to achieve power β, * Mallows test requires $n > 200$ to achieve power β, *** both tests require $n > 200$ to achieve power β)

3.4.2 Results for Small Normal Samples of Equal Sample Size

In bioequivalence testing assuming normal samples, extremely small sample sizes of usually 12 observations in each period and sequence of a crossover trial are commonly used. Recall, that a crossover design is not suitable for the analysis based on Mallows distance within a purely nonparametric setting. Therefore, it is of particular interest to investigate the behavior of the Mallows test for rather small sample sizes in a design with two independent groups. To achieve similar power as in the two-period crossover design (cf. Chow & Liu, 1992) sample sizes of $n = m = 24$ in each group are required. This corresponds to the total sample size in both periods of one sequence in a standard cross over experiment. The tolerance bounds are chosen accordingly to achieve reasonable power.

			Δ_0			
α	k	β	1.6	1.8	2.0	2.2
.0	$.75\Delta_o$	.4	.70 (18)	.70 (14)	.67 (11)	**
	$.6\Delta_o$	.7	.88 (21)	.85 (17)	.88 (13)	**
.1	$.75\Delta_o$	.4	*	.87 (18)	.95 (15)	**
	$.6\Delta_o$	.7	.88 (22)	.84 (17)	1.02 (16)	.90 (12)

Table 4: Observed relative efficiency of the Mallows equivalence test and the equivalence t test (* equivalence t test requires $n > 25$ to achieve power β, ** Mallows test requires $n < 10$ to achieve power β)

Comparing again the observed relative efficiency (Table 4), we see that the Mallows test is again slightly more efficient than the equivalence t test in the normal case, even for very small sample sizes as $n, m = 10$. Nevertheless, we have to take into account that Mallows test is always liberal (Figure 7), in particular when trimming is increased. In contrast, it can be shown that the equivalence t test never exceeds the nominal level (Munk, 1994). However, we find that the power of Mallows equivalence test (Figure 8) is surprisingly large for small samples compared to the asymptotically optimal equivalence t test. Therefore this test may serve as a powerful and robust nonparametric alternative as long as the sample size is not too small.

3.4.3 Results for Normal Samples of Unequal Sample Size

Since the results for the normal samples with equal sample size indicate that a trimming of less than 10% should be used, only two trimming constants $\alpha = 0$ and $\alpha = .05$ have been investigated.

Figures 9 and 10 give the observed significance level and the observed power for the Mallows test, respectively. Again the Mallows test is slightly liberal, especially when $\sigma = .5$ and $m = 30$. Surprisingly, the significance level is better maintained when the ratio of the sample sizes is reciprocal to that of the variances as in the converse case. As the sample sizes approach each other, observed power and accuarcy of the observed significance level is maximized. Therefore, we recommend to perform Mallows test only when $1/2 \leq \frac{m}{n} \leq 2$, which should be satisfied in most applications. In particular for very small samples, nearly equal sample sizes seem to be necessary.

3.4.4 Results for Nonnormal Samples

We will see that the trimmed Mallows distance behaves differently for the heavier right tail cases ($\psi < 1$) than for the lighter right tail cases ($\psi > 1$). Therefore results will be presented separately.

Since the shape of the c.d.f. G depends only on the single tail parameter ψ, the Mallows distance Γ_α is fully parametrized by ψ. Figure 11 gives the observed power curves for the lighter right tail cases ($\psi > 1$) when 8% trimming is used. Observe that now in each figure the bound Δ_0 is fixed whereas the true underlying Mallows distance is represented in the ordinate. The dotted vertical line separates the hypothesis from the alternative. The level of significance is $\alpha_s = 0.05$. The equivalence test maintains the level with high accuracy for all cases considered. Power, however, is decreased by about 50% as sample size is reduced from $n = 100$ to 50. Sample sizes of $n = 25$ are insufficiently large. The loss in power when changing from $\frac{m}{n} = .96$ to $\frac{m}{n} = .48$ is about 10-20%. It was also observed that there was no significant change in power when the higher trimming of 16% was used (not presented). For the heavier right tail cases ($\psi < 1$), the size $\alpha_s = .05$ is better maintained , when a trimming of 16% is used compared to 8% trimming (see Figure 12 and 13). A possible explanation might be the exponential growth of the Mallows distance as $\psi < 1$ decreases (see Figure 3). This growth is considerably reduced when trimming is increased. The loss in power for smaller sample sizes n is less dramatic in the heavier right tail cases ($\psi < 1$) compared to the lighter right tail cases ($\psi > 1$) resulting in strongly peaked distributions. Here, a sample size of $n = 25$ might be sufficient, especially if a larger equivalence bound Δ_0 can be tolerated. There is again a 10-20% loss in power when $\frac{m}{n} = .96$ is changed to $\frac{m}{n} = .48$.

In summary, this study indicates that the equivalence test performs adequately for broad classes of non normal distributions. Sampling sizes of $m, n \geq 25$ are recommended in order to guarantee sufficient accuracy of the finite sample approximation to the asymptotic normal law. A larger trimming might be necessary if one of the samples comes from a highly skewed distribution compared to the second one, in which case sample sizes $n, m \geq 50$ may become necessary.

4 Discussion

Our aim in this paper was the investigation of the finite sample behaviour of the Mallows equivalence test. The assessment of the precise hypothesis $H : \Gamma_\alpha(F,G) > \Delta_0$ instead of the simple hypothesis $H_0 : F = G$ bears the additional difficulty of estimating the variance of the empirical Mallows distance. Roughly speaking this is the price we have to pay for the additional gain in information provided by interval hypotheses of the type H. This is reflected by the rather liberal approximation of the nominal level when sample sizes are very small or the tolerance bound $\Delta_0 << 1$. Therefore, sample sizes of $n, m \geq 25$ should be available even under distributional assumptions close to a normal law. However, testing the classical hypothesis $H_0 : F = G$ leaves the experimenter always in the somewhat embarrassing situation how to specify the amount of similarity or difference between populations. In other words, since estimation of the variance of the empirical Mallows distance is *not* necessary under H_0 the outcomes of the corresponding tests are questionable.

When the densities are highly skewed or strongly peaked our simulations indicate that larger sample sizes may become necessary. Therefore, an initial nonparametric estimation of the densities may serve as a guard. Slight trimming has been found to be a very useful tool to protect against outliers. In particular, in bioequivalence assessment of the means the flexible choice of trimming weights is extremely helpful because the bioequivalence criterion remains invariant (after rescaling the hypotheses). Most surprisingly, our results show that Mallows equivalence test is (finite) more efficient than its parametric standard competitor which is asymptotically optimal under normality. Note that the Mallows test is a purely nonparametric test and therefore robust against distributional misspecifications. These results, however, should not be interpreted too optimistically because the liberality of Mallows test may be responsible for this superiority.

References

Brown, L.D., Hwang, J.T.G. and Munk, A. (1995). An unbiased test for the bioequivalence problem. *Dresdner Schriften zur Mathematischen Stochastik, ISSN 0946-4735* to appear in *The Annals of Statistics.*

Chow, S.C., J.P. Liu (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies.* Marcel Dekker, Inc..

Chow, S.C, Tse, S.K. (1990) Outlier detection in bioavailability/bioequivalence studies. *Statistics in Medicine* **9**, 549-58.

Czado, C. (1992), "On Link Selection in Generalized Linear Models" in L. Fahrmeir et al: Advances in GLIM and Statistical Modelling, Proceedings of the GLIM 92 conference and the 7th International Workshop on Statistical Modelling, Munich, 13-17 July 1992, Lecture Notes in Statistics, **Vol. 78**, Springer Verlag, 60-65.

Hauck, W.W. and Anderson, S. (1992), "Types of Bioequivalence and related statistical considerations," *Biostatistics Technical Reports* #19, *Department of Epidemiology and Biostatistics, University of California, San Francisco.*

Holder, D.J. and Hsuan, F. (1994), "Moment-based criteria for determining bioequivalence" *Biometrika,* **80**, 835-46.

Mallows, C.L. (1972). A note on asymptotic joint normality. *Annals Math. Stat.***43**, 508-15.

Munk, A. (1993). An improvement on commonly used tests in bioequivalence assessment. *Biometrics* **49**, 1225–31.

Munk, A. (1994). Reader Reaction Response – On a method of combining double t test and Anderson-Hauck test. *Biometrics* **50**, 884-86.

Munk, A. (1996). Equivalence and interval testing for Lehmann's alternative. *Journ. Amer. Statist. Ass.* **91**(435), 1-10.

Munk, A., Czado, C. (1995). Nonparametric Validation of Similar Distributions and Proving Goodness of Fit, *Dresdner Schriften zur Mathematischen Stochastik, ISSN 0946-4735.* Under revision for publication in The Journ. Royal Statist, Societ. B.

Müller–Cohrs, J. (1990). The power of the Anderson–Hauck test and the double t–test. *Biometrical Journal* **32**, 259–266.

Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics* **51**, 615–26.

Schuirmann, D. L. (1987). A comparison of the two one–sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacocinetics and Biopharmaceutics* **15**, 657–680.

de Wet, T., Venter, J.H. (1972). Asymptotic distributions of certain test criteria of normality. *S. Afr. Stat. Journ.* **6**, 135-49.

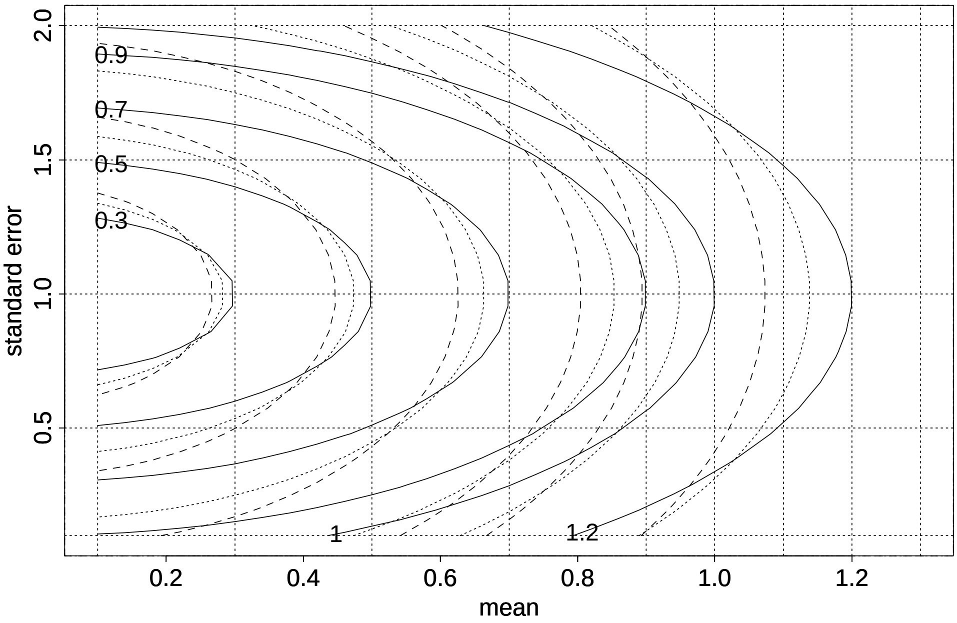

Figure 1: Mallow distance contours ($F \sim N(0,1)$, $G \sim N(\mu,\sigma)$) for no trimming (—), 10% trimming ($\cdot\cdot\cdot$) and 20% trimming ($- -$)

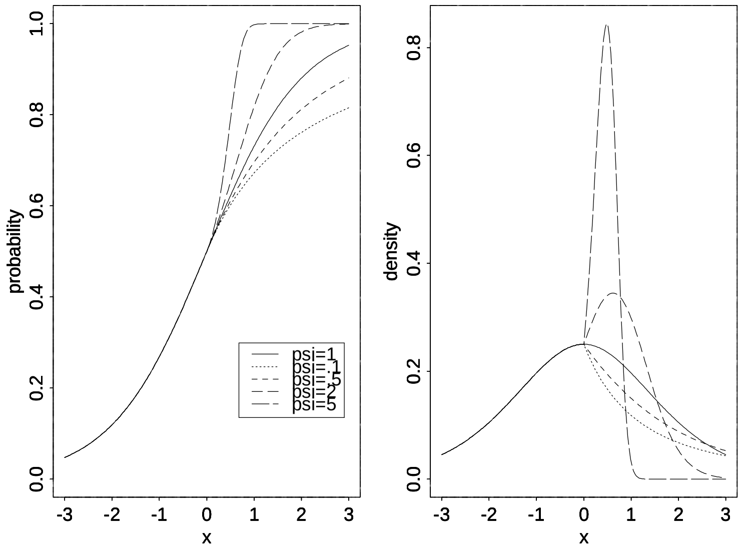

Figure 2: C.d.f. and density of the generalized logistic family.

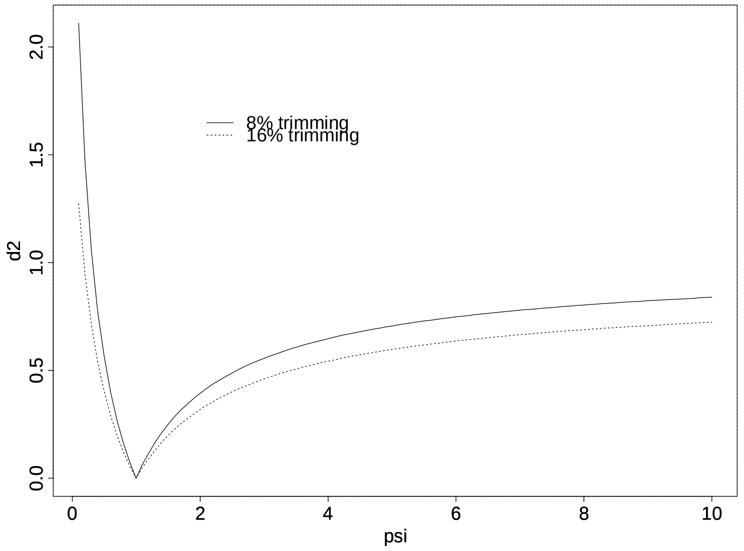

Figure 3: Mallows distance ($F \sim F_1,\ G \sim F_\psi$).

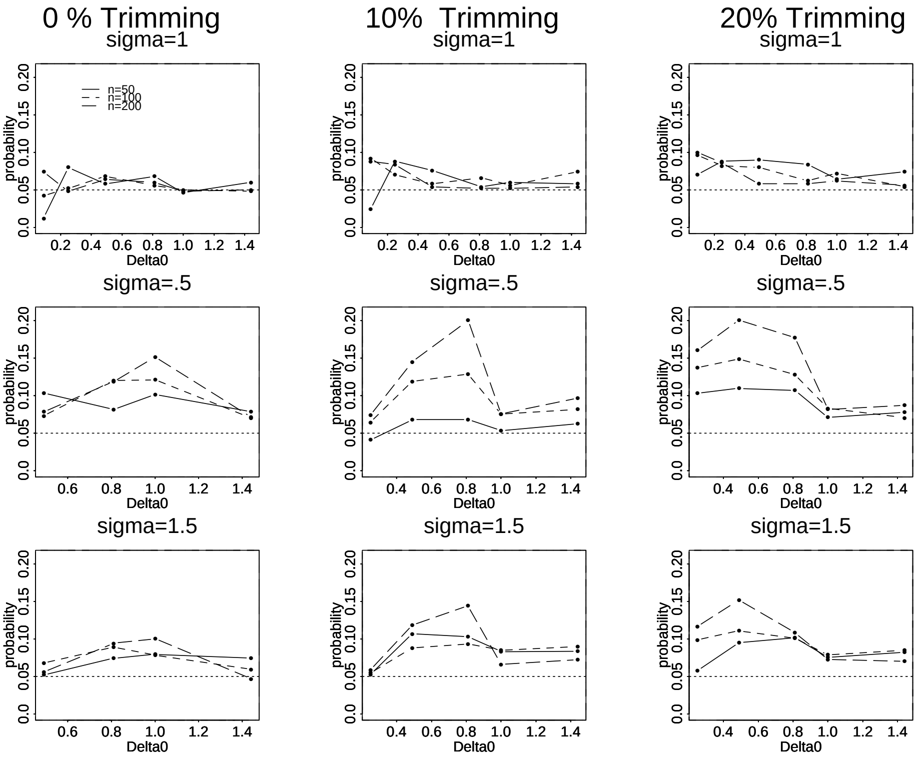

Figure 4: Observed significance level of the Mallow test for large $n = m$

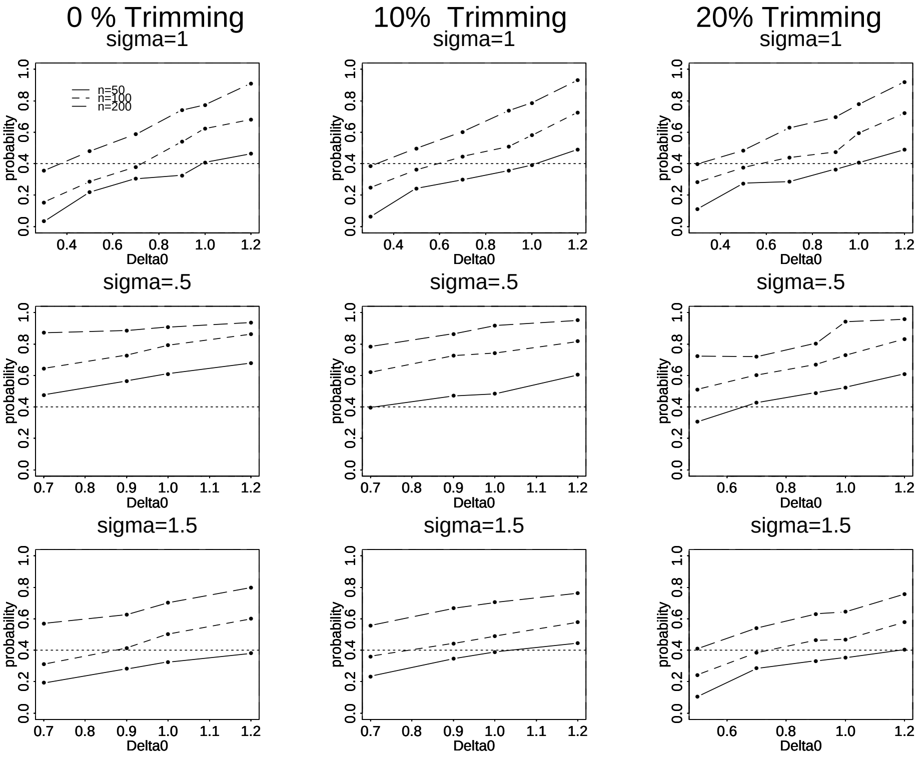

Figure 5: Observed power of the Mallow test at $.75\Delta_o$ large $n = m$

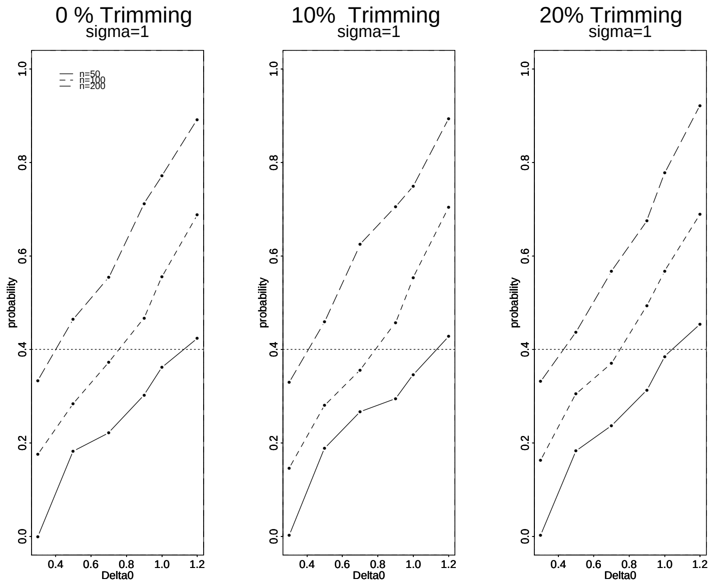

Figure 6: Observed power of the equivalence t test at $.75(1-2\alpha)^{-1/2}\Delta_o$ when $n = m$ and equal variances

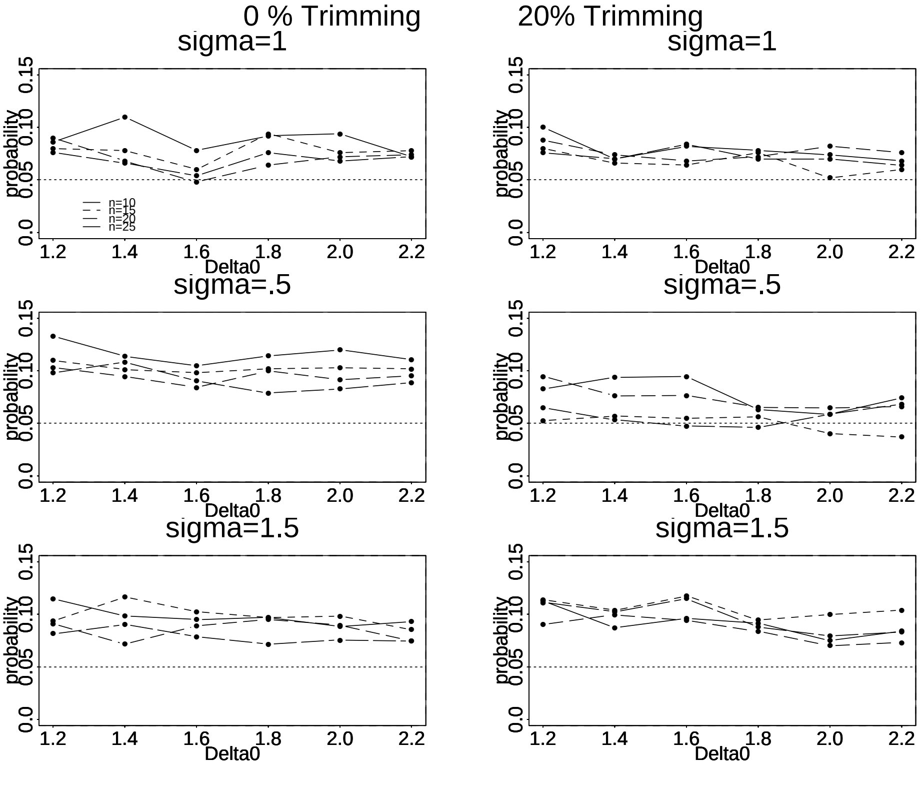

Figure 7: Observed significance level of the Mallow test for small $n = m$

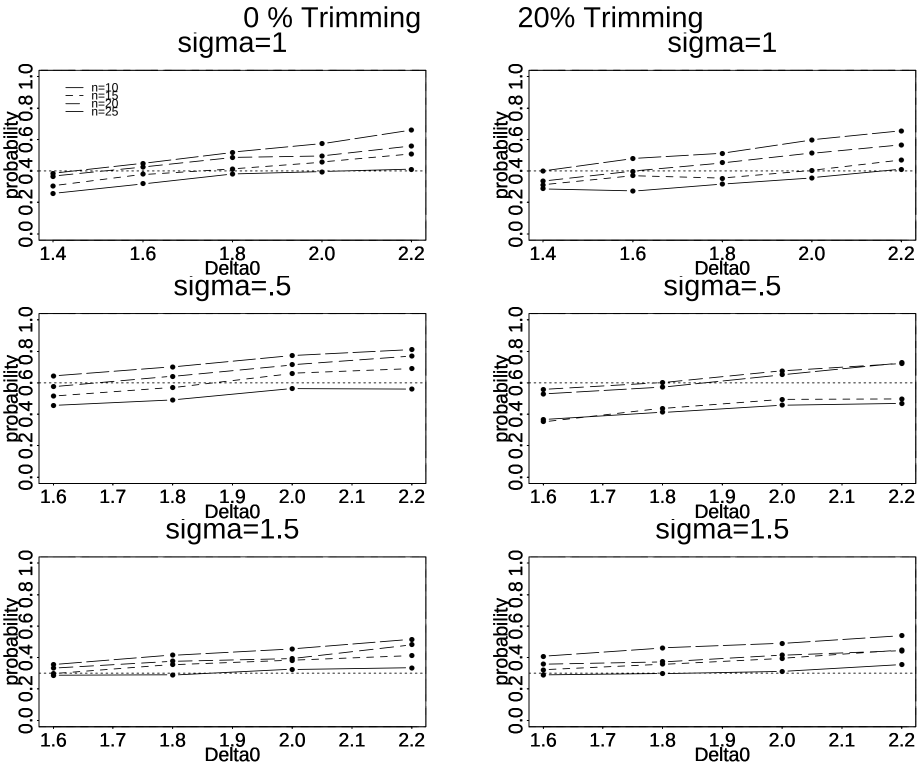

Figure 8: Observed power level of the Mallow test for small $n = m$

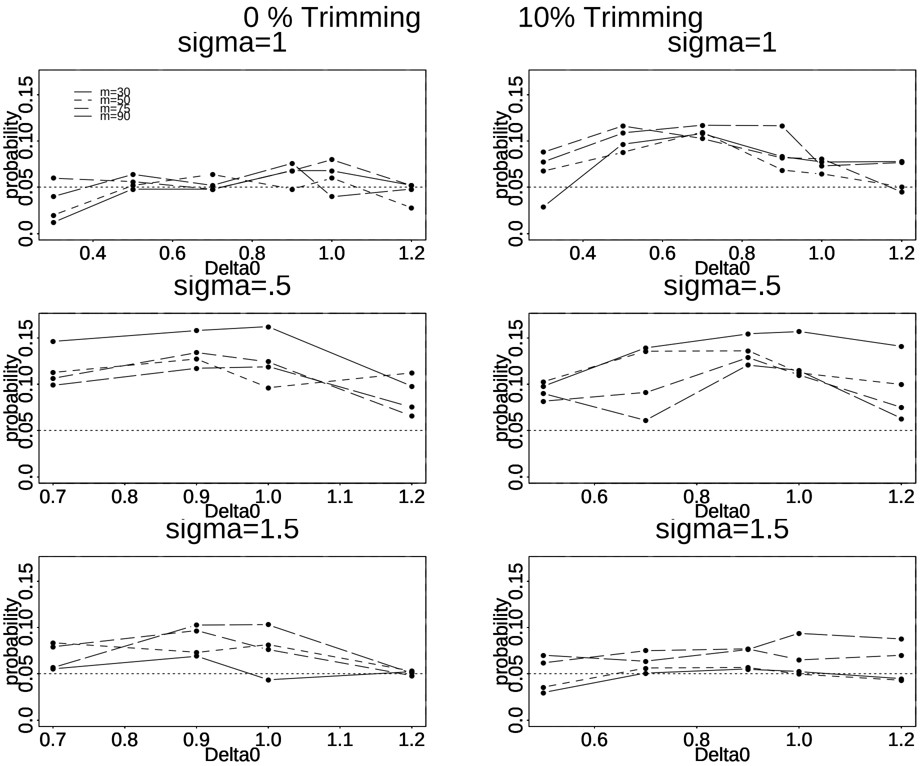

Figure 9: Observed significance level of the Mallow test when $n = 100$ and $m \neq n$

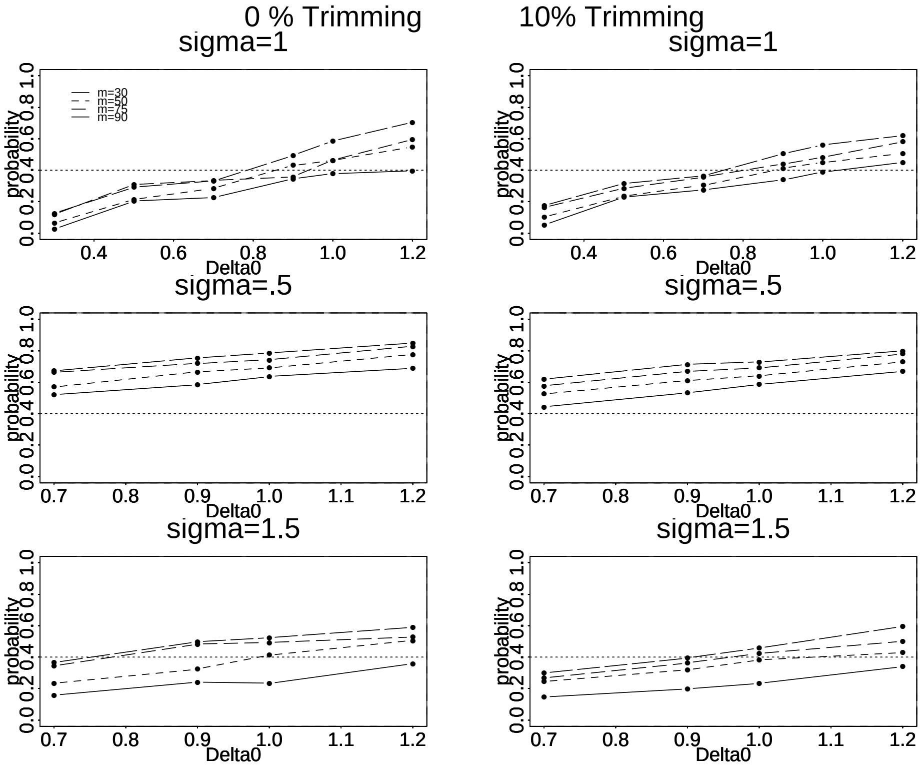

Figure 10: Observed power of the Mallow test at $.75\Delta_o$ when $n = 100$ and $n \neq m$

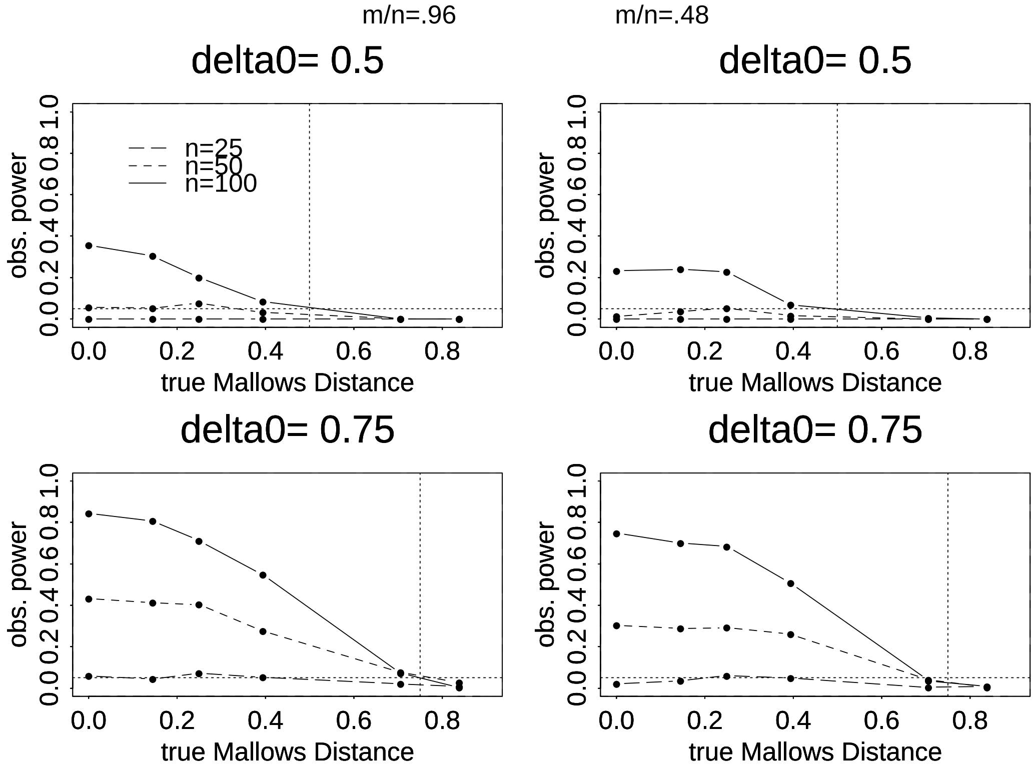

Figure 11: Observed power curves for $\psi > 1$ when $m/n = 0.96$ and $.48$ and 8% trimming is used

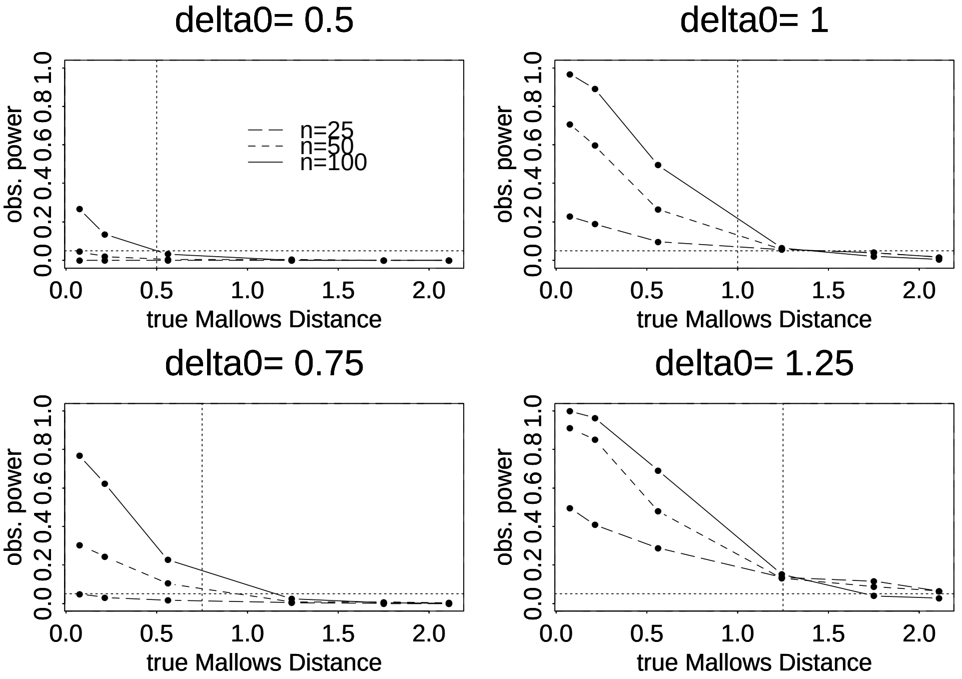

Figure 12: Observed power curves for $\psi < 1$ when $m/n = 0.96$ and 8% trimming is used

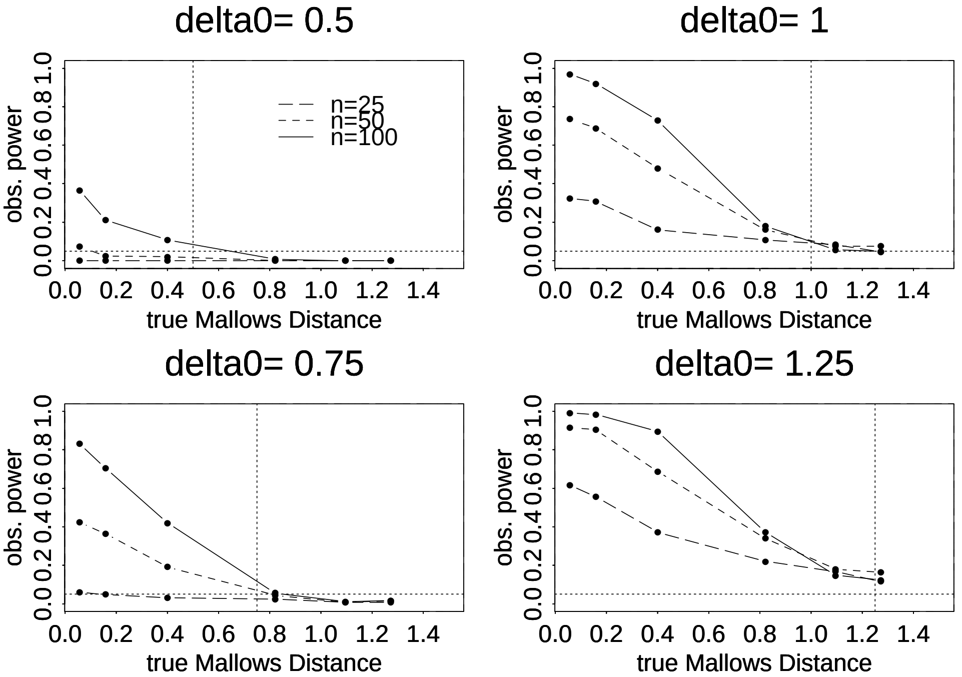

Figure 13: Observed power curves for $\psi < 1$ when $m/n = 0.96$ and 16% trimming is used