# Tests for multiple regression based on simplicial depth

by Robin Wellmann, Christine H. Müller*
University of Kassel

August 4, 2008

**Abstract**

A general approach for developing distribution free tests for general linear models based on simplicial depth is applied to multiple regression. The tests are based on the asymptotic distribution of the simplicial regression depth, which depends only on the distribution law of the vector product of regressor variables. Based on this formula, the spectral decomposition and thus the asymptotic distribution is derived for multiple regression through the origin and multiple regression with Cauchy distributed explanatory variables. A simulation study suggests that the tests can be applied also to normal distributed explanatory variables. An application on multiple regression for shape analysis of fishes demonstrates the applicability of the new tests and in particular their outlier robustness.

**Keywords:** Degenerated U-statistic, distribution-free tests, multiple regression, outlier robustness, regression depth, simplicial depth, spectral decomposition, shape analysis.
**AMS Subject classification:** Primary 62G05, 62G10; secondary 62J05, 62J12, 62G20.

## 1 Introduction

Liu (1988, 1990) used the half space depth of Tukey (1975) to define simplicial depth of a multivariate location parameter $\theta \in \Theta = I\!R^q$ in a sample $z_1, ..., z_N \in I\!R^q$ as

$$d_S(\theta, (z_1, ..., z_N)) = \binom{N}{q+1}^{-1} \sum_{1 \le n_1 < n_2 < ... < n_{q+1} \le N} I\!I\{d(\theta, (z_{n_1}, ..., z_{n_{q+1}})) > 0\}, \qquad (1)$$

*Research supported by the SFB/TR TRR 30 Project D6

where $d$ is the half space depth of Tukey and $I\!I$ denotes the indicator funtion. This depth counts the simplices spanned by $q+1$ data points which are containing the parameter $\theta$. Since Tukey (1975), several other depth notions were introduced. Each of them can be used as depth $d$ in (1) leading to several different simplicial depth notions. Several depth notions can be obtained from the book of Mosler (2002) and the references in it. If $d$ is the regression depth of Rousseeuw and Hubert (1999), then $d_S$ is called simplicial regression depth. General concepts of depth were introduced and discussed by Zuo and Serfling (2000a,b) and Mizera (2002). Mizera (2002) in particular generalized the regression depth of Rousseuw and Hubert (1999) by basing it on quality functions instead of squared residuals. This approach makes it possible to define the depth of a parameter value with respect to given observations in various statistical models via general quality functions. Appropriate quality functions are in particular likelihood functions as studied by Mizera and Müller (2004) for the location - scale model and by Müller (2005) for generalized linear models.

Any concept of data depth can be used to generalize the notion of ranks and to derive distribution free tests by generalizing Wilcoxon's rank sum test. Nevertheless only few papers deal with tests based on data depth. Liu (1992) and Liu and Singh (1993) proposed distribution-free multivariate rank tests based on depth notions. While the asymptotic normality is derived for several depth notions for distributions on $I\!R^1$, it is shown only for the Mahalanobis depth for distributions on $I\!R^k$, $k > 1$. Hence it is unclear how to generalize the approach of Liu and Singh to other situations. More successful distribution free tests are provided by the concept of ranks and signs based on the multivariate Oja median (see Oja 1983). For an overview of this methods see Oja (1999). However this approach provides only tests for multivariate data and does not concern regression models. Bai and He (1999) derived the asymptotic distribution of the maximum regression depth estimator. However, this asymptotic distribution is given implicitly so that it is not convenient for testing. Tests for regression based on depth notions were only derived by Van Aelst et al. (2002), Müller (2005) and Wellmann et al. (2008). Van Aelst et al. (2002) even derived an exact test based on the regression depth of Rousseeuw and Hubert (1999) but did it only for linear regression.

Müller (2005) and Wellmann et al. (2008) used the fact that any simplicial depth is a U-statistic with kernel function

$$\psi_\theta(z_{n_1}, ..., z_{n_{q+1}}) = I\!I\{d(\theta, (z_{n_1}, ..., z_{n_{q+1}})) > 0\}.$$

For U-statistics the asymptotic distribution is known. However, the U-statistic is degenerated for most simplicial depth notions so that the spectral decomposition of the conditional expectation

$$\psi^2_\theta(z_1, z_2) := E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1})|Z_1 = z_1, Z_2 = z_2) - E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1})) \qquad (2)$$

is needed to derive the asymptotic distribution. But as soon as the spectral decomposition of (2) is known, asymptotic tests can be derived for any hypothesis of the form $H_0 : \theta \in \Theta_0$

where $\Theta_0$ is an arbitrary subset of the parameter space $\Theta$. These tests are based on the test statistic $T(z_1, \ldots, z_N) := \sup_{\theta \in \Theta_0} T_\theta(z_1, \ldots, z_N)$, where $T_\theta(z_1, \ldots, z_N)$ is defined as

$$T_\theta(z_1, \ldots, z_N) := N\left(d_S(\theta, (z_1, \ldots, z_N)) - \mu_\theta\right) \tag{3}$$

with $\mu_\theta = E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1}))$ (see Müller 2005 and Wellmann et al. 2008).

The spectral decomposition of (2) was derived by Müller (2005) for linear and quadratic regression by solving differential equations. Wellmann et al. (2008) extended this result to polynomial regression with polynomials of arbitrary degree by proving a general formula of (2) and then specifying the general formula for polynomial regression so that the spectral decomposition can be found by Fourier series representation.

The general formula can be specified also for multiple regression so that a spectral decomposition of (2) can be derived for this case as well. This is shown in this paper.

In Section 2, the general approach with this general formula is presented. In particular the assumptions for this general approach are given in this section. In Section 3 the general formula is specified for multiple regression through the origin. Based on the specified formula the spectral decomposition is derived, which is given by spherical functions and eigenvalues depending on Gegenbauer functions.

The asymptotic distribution for multiple regression with intercept, where the regressors have Cauchy distribution, is given in Section 4. This model is traced back to multiple regression through the origin by multiplying the regressors and the dependent variables with additional random variables $S_n$. The simulation study, which is presented at the end of Section 4 suggests, that the tests can be applied also to normal distributed explanatory variables.

Section 5 provides some applications on tests in multiple regression through the origin with two explanatory variables in the shape analysis of fishes. These examples in particular show that the new tests possess high outlier robustness. All proofs are given in Section 6.

# 2 The general case

We assume a statistical model for i.i.d. random variables $Z_1, ..., Z_N$ with values in $\mathcal{Z} \subset \mathbb{R}^p, p \geq 1$ and parameter space $\Theta = \mathbb{R}^q$. We choose functions $h : \mathcal{Z} \to \mathbb{R}$ and $v : \mathcal{Z} \to \mathbb{R}^q$ and call

| | | |
|---|---|---|
| $Y_n$ | $:= h(Z_n)$ | the dependent variable, |
| $X_n$ | $:= v(Z_n)$ | the regressor, and |
| $S_n(\theta)$ | $:= \text{sign}(Y_n - X_n^T\theta)$, | $\theta \in \mathbb{R}^q$, the sign of the residual. |

We assume that for all $\theta \in \Theta$:

$$\begin{aligned} &\bullet \quad P_\theta(S_1(\theta) = 1|X_1) \equiv \frac{1}{2} \text{ a.s.,} \\ &\bullet \quad P_\theta(S_1(\theta) = 0|X_1) \equiv 0 \text{ a.s., and} \\ &\bullet \quad P_\theta(X_1, \ldots, X_q \text{ are linearly dependent}) = 0. \end{aligned} \tag{4}$$

The last two conditions of (4) are easily satisfied for example by continuous distributions. Depending on the distribution of $Z_n$, the first condition can be satisfied by appropriate transformations $v$ and $h$. The first condition in particular implies that the true regression function is in the center of the data, which means that the median of the residuals is zero.

We denote random variables by capital letters and realizations by small letters. The depth of $\theta \in \Theta$ for observations $z = (z_1, ..., z_N)$ is given by

$$d_T(\theta, z) = \min_{u \neq 0} \#\{n : s_n(\theta)\, u^T v(z_n) \geq 0\}.$$

This depth coincides with the regression depth of Rousseeuw and Hubert (1999) and with Definition 2 from Wellmann et al. (2008), if the quality functions $\mathcal{G}_{z_n}(\theta) = -(h(z_n) - v(z_n)^T\theta)^2$ are used. It is a tangent depth in the sense of Mizera (2002).

This tangent depth has the disadvantage, that it provides a simplicial depth which attains rather high values in subspaces of the parameter space. This is in particular a disadvantage in testing if the aim is to reject the null hypothesis. To avoid this disadvantage, we introduce a modified version of the depth $d_T$, called harmonized depth. The harmonized depth of $\theta \in \Theta$ with respect to observations $z_1, ..., z_{q+1}$ is defined as

$$\psi_\theta(z_1, \ldots, z_{q+1}) = \begin{cases} d_T(\theta, (z_1, \ldots, z_{q+1})), & \text{if } s_n(\theta) \neq 0 \text{ for } n = 1, ..., q+1 \\ 0, & \text{otherwise,} \end{cases}$$

so that the simplicial depth is given by

$$d_S(\theta, z) = \binom{N}{q+1}^{-1} \sum_{1 \leq n_1 < n_2 < \ldots < n_{q+1} \leq N} \psi_\theta(z_{n_1}, ..., z_{n_{q+1}}).$$

Under the assumptions (4) we have

$$\mu_\theta = E_\theta(\psi_\theta(Z_1, \ldots, Z_{q+1})|Z_1 = z_1) = \frac{1}{2^q}$$

(see Wellmann et al. (2008)), so that $d_S(\theta, z)$ is a degenerated U-statistic. Hence the spectral decomposition of (2) is needed. This can be derived by the following Proposition 1 of Wellmann et al. (2008).

**Proposition 1** *Under the assumptions (4), the conditional expectation (2) satisfies*

$$\psi_\theta^2(z_1, z_2) = \frac{s_1(\theta)s_2(\theta)}{2^{q-1}} \left( P_\theta(x_1^T W \; x_2^T W < 0) - \frac{1}{2} \right),$$

*where* $W := X_3 \times ... \times X_{q+1}$ *is the vector product of* $X_3, \ldots, X_{q+1}$.

Recall that the vector product $w = x_3 \times \cdots \times x_{q+1}$ of $x_3, ..., x_{q+1} \in I\!R^q$ is the gradient of the linear function $x \mapsto \det(x_3, ..., x_{q+1}, x)$. For instance see Storch and Wiebe (1990, p 362 ff.). The vector $w$ is orhogonal to $x_3, ..., x_{q+1}$.

Because of this representation, only the spectral decomposition of the kernel $\mathcal{K}$, defined by

$$\mathcal{K}(x_1, x_2) := P_\theta(x_1^T W \; x_2^T W < 0) - \frac{1}{2}, \quad \text{for } x_1, x_2 \in I\!R^q \tag{5}$$

is needed. As soon as the spectral decomposition is given by

$$\mathcal{K}(x_1, x_2) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x_1)\varphi_j(x_2) \quad \text{in } I\!L_2\big(P^{X_1} \otimes P^{X_1}\big), \tag{6}$$

where $(\varphi_j)_{j=1}^\infty$ is an orthonormal system (ONS) in $I\!L_2\big(P^{X_1}\big)$ and $\lambda_1, \lambda_2, ... \in I\!R$, then the asymptotic distribution of the simplicial depth satisfies

$$N\big(d_S(\theta, (Z_1, \ldots, Z_N)) - \frac{1}{2^q}\big) \xrightarrow{\mathcal{L}} \sum_{l=1}^{\infty} \frac{(q+1)!}{(q-1)!2^q} \lambda_l \left( {U_l}^2 - 1 \right), \tag{7}$$

where $U_1, U_2, ...$ are i.i.d. random variables with $U_1 \sim \mathcal{N}(0, 1)$ (see e.g. Lee 1990, p. 79, 80, 90, Witting and Müller-Funk, p. 650). If the distribution of the vector product $W := X_3 \times ... \times X_{q+1}$ does not depend on $\theta$, which is the case for usual regressors, then the asymptotic distribution is independent of $\theta$.

Then any hypothesis of the form $H_0 : \theta \in \Theta_0$, where $\Theta_0$ is an arbitrary subset of the parameter space $\Theta$, can be tested by using the test statistic $T(z_1, \ldots, z_N) := \sup_{\theta \in \Theta_0} T_\theta(z_1, \ldots, z_N)$, where $T_\theta(z_1, \ldots, z_N)$ is defined by (3) with $\mu_\theta = \frac{1}{2^q}$. The null hypothesis $H_0$ is rejected if $T(z_1, \ldots, z_N)$ is less than the $\alpha$-quantile of the asymptotic distribution of $T_\theta(Z_1, \ldots, Z_N)$.

# 3 Multiple regression through the origin

Assuming a model for multiple regression through the origin,

$$Y_n = \theta_1 X_{n,1} + ... + \theta_q X_{n,q} + E_n = X_n^T \theta + E_n$$

we suppose that (4) holds and that there is an invertible matrix $A \in I\!R^{q,q}$, such that $\frac{1}{||A\,X_n||} A\,X_n$ is uniformly distributed on the unit sphere. This is in particular the case, if $X_n$ has a elliptical distribution like the multivariate normal distribution with mean zero. In order to derive the asymptotic distribution of the simplicial depth for this regression model, we have to simplify the kernel function $\mathcal{K}$ given by equation (5). By using that with $\frac{1}{||A\,X_3||} A\,X_3, ..., \frac{1}{||A\,X_{q+1}||} A\,X_{q+1}$ also the vector product is uniformly distributed on the unit sphere, we obtain the following proposition.

**Proposition 2** *For all $x_1, x_2 \in I\!R^q \backslash \{0\}$ we have*

$$\mathcal{K}(x_1, x_2) = \frac{1}{\pi} \arccos\left( < \frac{Ax_1}{||Ax_1||}, \frac{Ax_2}{||Ax_2||} > \right) - \frac{1}{2}.$$

The value $\mathcal{K}(x_1, x_2)$ depends only on the angle between $Ax_1$ and $Ax_2$. In Fenyö and Stolle (1983) it is shown, that we thus obtain the required eigenvalues, if we calculate some integrals, in which so called Gegenbauer functions occur. Therefor, let $T_{\mathcal{K}}$ defined by

$$T_{\mathcal{K}} : I\!L_2(P^{X_1}) \to I\!L_2(P^{X_1}) \text{ with } T_{\mathcal{K}} f(s) = \int \mathcal{K}(s,t) f(t)\, dP^{X_1}(t)$$

be the integral operator based on $\mathcal{K}(x_1, x_2)$. We obtain the following result:

**Proposition 3** *Let $S \subset I\!R^q$ be the unit sphere, where $q \geq 2$.*
*Let $K \in C(S \times S)$ be the function $K(s,t) := \frac{1}{\pi} \arccos(< s, t >) - \frac{1}{2}$ for all $s, t \in S$.*
*The values*

$$\begin{aligned} \lambda_0 &:= 0 \\ \lambda_p &:= -\frac{1}{2}\,\tau_q \left( \frac{\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{q}{2}+\frac{p}{2}\right)} \frac{\sin\left(\frac{p}{2}\pi\right)}{\pi} \right)^2 \text{ for } p \in I\!N \end{aligned}$$

*are the eigenvalues of the integral operator $T_K$, where $\tau_q = 2\frac{\pi^{\frac{q}{2}}}{\Gamma\left(\frac{q}{2}\right)}$ is the $q-1$-dimensional volume of the sphere. For $p \in I\!N$, the corresponding eigenfunctions with respect to the uniform measure $v$ on $S$ with $v(S) = \tau_q$ are the orthogonalized and normalized spherical functions $S_{p,1}^{(n)}, \ldots, S_{p,u_p}^{(n)}$ of degree $p$, where $n := q-2$. By Fenyö and Stolle (1983) we have $u_p = \frac{(p+n-1)!}{p!\,n!}(2p+n)$.*

Let $\left(S_{(p,k)}^{(q-2)}\right)_{(p,k)\in I}$ be the family of orthogonalized and normalized spherical functions from Proposition 3 with $I := \{(p,k) \in I\!N^2 : k \leq u_p\}$ and for $j \in I$ let $\varphi_j(x) := \sqrt{\tau_q}\, S_j^{(q-2)}\left(\frac{1}{||A\,x||} A\,x\right)$.

Because of $\frac{1}{||A\,X_1||}A\,X_1 \sim \ \frac{1}{\tau_q}\,v$, we obtain for all $i,j \in I$:

$$\begin{aligned}
\int \varphi_i\varphi_j \, d\,P^{X_1} &= \int \sqrt{\tau_q} S_i^{(q-2)}\big(\frac{1}{||A\,x||}A\,x\big)\sqrt{\tau_q} S_j^{(q-2)}\big(\frac{1}{||A\,x||}A\,x\big)\ P^{X_1}(d\,x) \\
&= \tau_q \int S_i^{(q-2)}\left(\frac{1}{||A\,X_1||}A\,X_1\right) S_j^{(q-2)}\left(\frac{1}{||A\,X_1||}A\,X_1\right) d\,P \\
&= \tau_q \int S_i^{(q-2)} S_j^{(q-2)} d\,P^{\frac{1}{||A\,X_1||}A\,X_1} \\
&= \frac{\tau_q}{\tau_q}\int S_i^{(q-2)} S_j^{(q-2)}\, d\,v.
\end{aligned}$$

Hence, $(\varphi_j)_{j\in I}$ is an ONS in $I\!L_2\big(P^{X_1}\big)$. From the previous propositions we conclude, that in $I\!L_2\big(P^{X_1}\otimes P^{X_1}\big)$ we have:

$$\begin{aligned}
\mathcal{K}(x_1,x_2) &= \frac{1}{\pi}\,\arccos\left(< \frac{A\,x_1}{||A\,x_1||}, \frac{A\,x_2}{||A\,x_2||} >\right) - \frac{1}{2} \\
&= \sum_{(p,k)\in I} \lambda_p S_{(p,k)}^{(q-2)}\left(\frac{1}{||A\,x_1||}A\,x_1\right) S_{(p,k)}^{(q-2)}\left(\frac{1}{||A\,x_2||}A\,x_2\right) \\
&= \sum_{(p,k)\in I} \frac{\lambda_p}{\tau_q}\sqrt{\tau_q} S_{(p,k)}^{(q-2)}\left(\frac{1}{||A\,x_1||}A\,x_1\right)\sqrt{\tau_q} S_{(p,k)}^{(q-2)}\left(\frac{1}{||A\,x_2||}A\,x_2\right) \\
&= \sum_{(p,k)\in I} \frac{\lambda_p}{\tau_q}\varphi_{(p,k)}(x_1)\varphi_{(p,k)}(x_2).
\end{aligned}$$

Hence with (7), we immediately get the next theorem:

**Theorem 1** *Suppose, that there is an invertible matrix $A \in I\!R^{q,q}$ with $q \geq 2$, such that $\frac{1}{||A\,X_n||}A\,X_n$ is uniformly distributed on the unit sphere and suppose that assumption (4) holds. Let $\lambda_1, \lambda_2, ...$ and $u_1, u_2, ...$ be as in the previous proposition.*

*Then there are i.i.d. random variables $U_1, U_2, ...$ with $U_p \sim \chi^2_{u_p}$ such that*

$$N\big(d_S(\theta,(Z_1,\ldots,Z_N)) - \frac{1}{2^q}\big) \xrightarrow{\mathcal{L}} \sum_{p=1}^{\infty} \frac{(q+1)!}{(q-1)!2^q}\frac{\lambda_p}{\tau_q}\,\big(U_p - u_p\big).$$

A simple possibility for estimating the quantiles is the generation of random numbers of the distribution. The quantiles given in Table 1 were calculated by computing 10000 random numbers of the distribution (only the first 150 summands). The calculation of the quantiles was repeated 500 times. The means of these quantiles are given in the table. The 99.5% confidence band is $\pm 0.01$ at most for each estimated quantile. The test statistic for multiple regression can be calculated similarly as for polynomial regression described in Wellmann et al. (2007). But here the calculation of the simplicial depth of a

given parameter is based on Lemma 1 in Wellmann et al. (2008) by checking if $s_{n_1}(\theta)x_{n_1}$ is a linear combination of $s_{n_2}(\theta)x_{n_2}, ..., s_{n_{q+1}}(\theta)x_{n_{q+1}}$ with negative coefficients.

Table 1: **Means of the simulated quantiles for multiple regression**

| $\alpha$-quantile | $q = 2$ | $q = 3$ | $q = 4$ |
|---|---|---|---|
| 0.5% | -2.607 | -1.845 | -1.222 |
| 1.0% | -2.189 | -1.566 | -1.044 |
| 2.0% | -1.771 | -1.284 | -0.863 |
| 2.5% | -1.635 | -1.192 | -0.805 |
| 5.0% | -1.216 | -0.905 | -0.619 |
| 10.0% | -0.795 | -0.612 | -0.426 |
| 20.0% | -0.368 | -0.310 | -0.224 |
| 30.0% | -0.127 | -0.126 | -0.099 |
| 40.0% | 0.048 | 0.008 | -0.006 |
| 50.0% | 0.183 | 0.116 | 0.072 |
| 60.0% | 0.293 | 0.209 | 0.140 |
| 70.0% | 0.388 | 0.292 | 0.203 |
| 80.0% | 0.473 | 0.373 | 0.265 |
| 90.0% | 0.554 | 0.456 | 0.331 |
| 95.0% | 0.600 | 0.504 | 0.373 |

# 4 Multiple regression with intercept

We derive the asymptotic distribution of the simplicial depth for different models of multiple regression with intercept as follows:

We define two different statistical models with different simplicial depths. We want to calculate the asymptotic distribution of the simplicial depth $d_S$ for a statistical model $(\mathcal{Z}^N, \mathcal{A}, \mathcal{P})$ with $\mathcal{P} = \{\otimes_{n=1}^N P_\theta : \theta \in \Theta\}$. We consider an other statistical model $(\tilde{\mathcal{Z}}^N, \tilde{\mathcal{A}}, \tilde{\mathcal{P}})$ with $\tilde{\mathcal{P}} = \{\otimes_{n=1}^N \tilde{P}_\theta : \theta \in \Theta\}$ and in this model, we define the simplicial depth $\tilde{d}_S$. Assume that there is a transformation $\varphi : \Theta \to \Theta$ of the parameters, such that $(\otimes_{n=1}^N P_\theta)^{d_S(\theta,\cdot)} = (\otimes_{n=1}^N \tilde{P}_{\varphi(\theta)})^{\tilde{d}_S(\varphi(\theta),\cdot)}$. If the asymptotic distribution of the simplicial depth in the second model does not depend on the unknown parameter, it follows that the asymptotic distribution is equal in both models.

To prove the next Lemma, we add to the random vectors $Z_n = (Y_n, T_n)$ a random variable $S_n$ so that the second model bases on random vectors $\tilde{Z}_n = (Y_n, T_n, S_n)$. In Section 6, we work out this idea.

**Lemma 1** *Let $(Y_1, T_1, E_1), ..., (Y_N, T_N, E_N)$ be i.i.d continuous distributed random vectors such that there is a $\theta \in \mathbb{R}^q$ with*

$$Y_n = \theta_0 + \theta_1 T_{n,1} + ... + \theta_{q-1} T_{n,q-1} + E_n = x(T_n)^T \theta + E_n,$$

*where $T_n = (T_{n,1}, ..., T_{n,q-1})$ and $x(T_n) = (1, T_{n,1}, ..., T_{n,q-1})$.*

*Suppose that*

- $P_\theta(Y_n - x(T_n)^T\theta > 0 | T_n) = \frac{1}{2}$
- $P_\theta(Y_n - x(T_n)^T\theta = 0 | T_n) = 0$
- $f^{T_n}(t) = \frac{\Gamma(\frac{q}{2})}{\sqrt{\pi^q |\Sigma|}} \frac{1}{(1 + t^T \Sigma^{-1} t)^{\frac{q}{2}}}$.

*That is, $T_n$ has a centered, multivariate Cauchy Distribution. Let $Z_n = (Y_n, T_n)$. Then the asymptotic distribution of the simplicial depth which is based on the dependent variable $Y_n$ and the regressor $X_n$ is equal to the distribution given in Theorem 1.*

A similar idea is used to show that the random vector $T_n$ does not need to be centered. The next Theorem generalizes Lemma 1:

**Theorem 2** *Let $(Y_1, T_1, E_1), ..., (Y_N, T_N, E_N)$ be i.i.d continuous distributed random vectors such that there is a $\theta \in \mathbb{R}^q$ with*

$$Y_n = \theta_0 + \theta_1 T_{n,1} + ... + \theta_{q-1} T_{n,q-1} + E_n = x(T_n)^T \theta + E_n,$$

*where* $T_n = (T_{n,1}, ..., T_{n,q-1})$ *and* $x(T_n) = (1, T_{n,1}, ..., T_{n,q-1})$.

*Suppose that*

- $P_\theta(Y_n - x(T_n)^T\theta > 0|T_n) = \frac{1}{2}$
- $P_\theta(Y_n - x(T_n)^T\theta = 0|T_n) = 0$
- $f^{T_n}(t) = \frac{\Gamma(\frac{q}{2})}{\sqrt{\pi^q|\Sigma|}} \frac{1}{(1+(t-\mu)^T\Sigma^{-1}(t-\mu))^{\frac{q}{2}}}$.

*That is,* $T_n$ *has a multivariate Cauchy Distribution. Let* $Z_n = (Y_n, T_n)$*. Then the asymptotic distribution of the simplicial depth which is based on the dependent variable* $Y_n$ *and the regressor* $X_n$ *is equal to the distribution given in Theorem 1.*

The assumption of Cauchy distributed regressors is a technical requirement resulting from the proofs. In a simulation study we checked how the simplicial depth test controls the alpha level for different sample sizes and different distributional assumptions for the explanatory variables. In the model

$$Y_n = \theta_0 + \theta_1 T_{n,1} + \theta_2 T_{n,2} + E_n$$

we tested the null hypothesis

$$H_0 : \theta = 0 \text{ against } H_1 : \theta \neq 0$$

to the asymptotic level $\alpha = 0.05$. The observations are simulated under the null hypothesis with $T_n \sim \mathcal{N}_2(0, I)$ or $T_n \sim \text{Cauchy}_2(0, I)$ respectively, where $I$ is the identity matrix. The probabilities to reject the null hypothesis, estimated from 10000 samples with sample size 50 or 100 respectively, are given in Table 2.

Table 2: **The estimated probability to reject the null hypothesis**

| N | Cauchy | Normal |
|---|---|---|
| 50 | 0.06 | 0.06 |
| 100 | 0.05 | 0.05 |

Thus, the test may also be applied to normal distributed explanatory variables. For a power comparison with other existing tests in the case of simple linear regression ($q = 2$) see Wellmann et al. (2008).

# 5 Application: Test for multiple regression through the origin

The North American Sunfish "pumpkinseed" (*Lepomis gibbosus*) was introduced to European waters about 100 years ago. Near Brighton, 162 specimens were collected in 2003 from the Tanyards fisheries pond. Nineteen landmarks (see Figure 1) were identified for each fish.

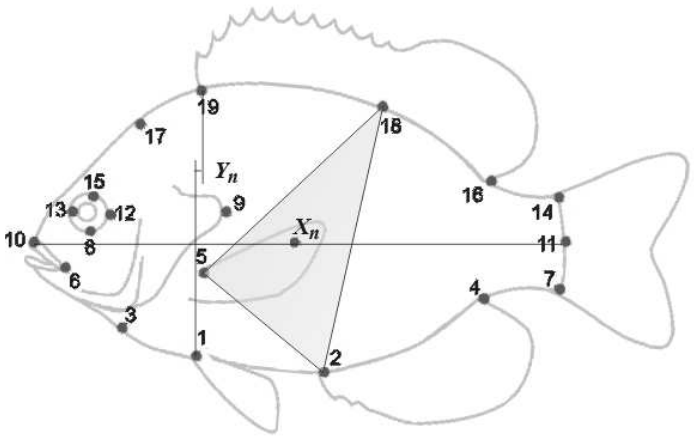


Figure 1: Landmarks

In this section, we want to find out relationships between the landmarks. We restrict ourselves on those relationships, that can be tested within the model for multiple regression through the origin (for other ones, see e.g. Tomeček et al. 2005 or Wellmann et al. 2007). We rotate, rescale and translate the fishes (the landmarks), such that landmark 10 (anterior tip of the upper jaw) is equal to $(-\frac{1}{2}, 0)^T$ and landmark 11 (caudal fin base) is equal to $(\frac{1}{2}, 0)^T$.

Let $\lambda_n^p = (\lambda_{n,1}^p, \lambda_{n,2}^p)^T \in I\!R^2$ be the landmark number $p$ of the $n$-th transformed fish. We choose 3 landmarks near the origin and define the center of the fish as a convex combination, for which the hypothesis, that it is centered cannot be rejected componentwise with the sign-test. We take the center of a fish to be

$$x_n = 0.34\, \lambda_n^{18} + 0.22\, \lambda_n^2 + 0.44\, \lambda_n^5.$$

Figure 1 shows, that the horizontal position of the anterior edge of the dorsal fin base $\lambda_{n,1}^{19}$ is nearly equal to the horizontal position of the anterior edge of the pelvic fin base $\lambda_{n,1}^1$. Indeed, the sign test for testing that

$$y_n = \lambda_{n,1}^{19} - \lambda_{n,1}^1$$

is centered provides the very high p-value 0.937. We call $y_n$ the fin base difference in this paper.

We test within the model for multiple regression through the origin ($q = 2$), how $Y_n$ depends on the center $X_n = (X_{n,1}, X_{n,2})^T$ of the fish. Therefore we choose a random sample that consists on 50 fishes. The original data are discrete, due to rounding errors. To make them continuous, we add a small uniformly distributed random number to each observation, such that we would obtain the original data by rounding.

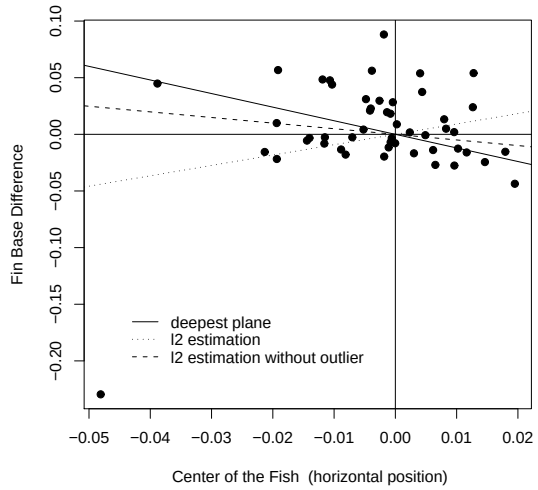


Figure 2: A deepest plane with $\theta_2 = 0$ and least squares fits at the $x_{n,1}$-axis

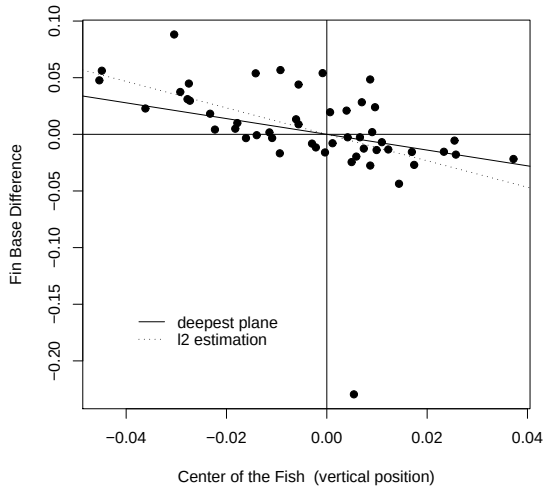


Figure 3: A deepest plane with $\theta_1 = 0$ and the least square fit at the $x_{n,2}$-axis

The parameter with maximum simplicial depth is $\widehat{\theta}_D := (-0.5411, -0.8660)^T$ and the least squares fit is $\widehat{\theta}_{l_2} = (0.9152, -1.1676)^T$. At first we test the hypothesis, that $X_{n,1}$ has no influence on $Y_n$, that is, $H_0 : \theta_1 = 0$. The test statistic depends on the depth of the deepest plane with $\theta_1 = 0$, given by the parameter $(0, -0.6952)^T$ (see Figure 3). The test statistic is 0.122, which is more than the 40% quantile of the asymptotic distribution and thus, we have no rejection (see Table 1). Hence, we may assume that $Y_n$ does not depend on the horizontal position of the center. Contrary to this result, the classical F-test rejects this hypothesis with respect to a significance level 5% (p-value = 0.028). This is due to the outlier in the left lower corner of Figure 2. The outlier strongly influences the first component of the least squares fit $\widehat{\theta}_{l_2}$, who's first component is positive (see the dashed line in Figure 2).

Without the outlier, the least squares fit is $\tilde{\theta}_{l_2} := (-0.4958, -0.9630)^T$ so that its first component is negative. Then the classical F-test would not reject the null-hypothesis with respect to a significance level 5%. Note that the least squares fit for the data without the outlier is close to the parameter $\widehat{\theta}_D$ with maximum simplicial depth.

On the other hand, the hypothesis that $X_{n,2}$ has no influence on $Y_n$, that is, $H_0 : \theta_2 = 0$ has to be rejected with respect to a significance level 2%, since the test statistic $-2.184$ is

near the 1% quantile of the asymptotic distribution. In particular, the deepest plane with $\theta_2 = 0$ given by the parameter $(-1.2, 0)^T$ gives not a good description of the data (see Figure 2). The classical F-test also rejects the null-hypothesis and provides a p-value of 0.0001. Indeed, the least square fit is strongly decreasing at the $x_{n,2}$-axis (see the dashed line in Figure 3).

We conclude that $Y_n$ depends on $X_{n,2}$, but not on $X_{n,1}$. As shown in Figure 3, the fin base difference becomes smaller if the center of the fish is shifted upwards. Roughly speaking, $\lambda_n^{19}$ shifts to the left and/or $\lambda_n^1$ shifts to the right, if the center is shifted upwards. This is possibly due to a curved vertebral column. If this interpretation is correct, then one could take into consideration a nonlinear transformation of the landmarks before further investigations, such that the vertebral columns of the transformed fishes can expected to be a straight line.

# 6 Proofs

See also Wellmann (2007) for details of the proofs.

**Proof of Proposition 2**
Let $x_1, x_2 \in \mathbb{R}^q \backslash \{0\}$.
For $j = 3, ..., q+1$ let $W_j := \frac{1}{||AX_j||} AX_j$, $U := \frac{W_3 \times ... \times W_{q+1}}{||W_3 \times ... \times W_{q+1}||}$
and for $j = 1, 2$ let

$$\begin{aligned} K^+(x_j) &:= \{w \in \mathbb{R}^q : (Ax_j)^T w \geq 0\}, \\ K^-(x_j) &:= \{w \in \mathbb{R}^q : (Ax_j)^T w \leq 0\}. \end{aligned}$$

Then we have

$$\begin{aligned} \mathcal{K}(x_1, x_2) + \frac{1}{2} &= P(x_1^T (X_3 \times ... \times X_{q+1})\, x_2^T (X_3 \times ... \times X_{q+1}) < 0) \\ &= P(\det(x_1, X_3, ..., X_{q+1}) \det(x_2, X_3, ..., X_{q+1}) < 0) \\ &= P(\det(A(x_1, X_3, ..., X_{q+1})) \det(A(x_2, X_3, ..., X_{q+1})) < 0) \\ &= P(\det(Ax_1, AX_3, ..., AX_{q+1}) \det(Ax_2, AX_3, ..., AX_{q+1}) < 0) \\ &= P(\det(Ax_1, W_3, ..., W_{q+1}) \det(Ax_2, W_3, ..., W_{q+1}) < 0) \\ &= P((Ax_1)^T U\, (Ax_2)^T U < 0) \\ &= P(U \in K^+(x_1) \cap K^-(x_2)) + P(U \in K^-(x_1) \cap K^+(x_2)) \\ &= P(U \in K^+(x_1) \cap K^-(x_2)) + P(-U \in K^+(x_1) \cap K^-(x_2)). \end{aligned}$$

Because of $-U \sim U$, it follows that

$$\mathcal{K}(x_1, x_2) = 2\, P(U \in K^+(x_1) \cap K^-(x_2)) - \frac{1}{2}.$$

Since $W_3, ..., W_{q+1}$ are uniformly distributed on the unit sphere, this is the case also for $U$. The proportion of the unit sphere, that is contained in $K^+(x_1) \cap K^-(x_2)$ is equal to the angle between $Ax_1$ and $Ax_2$, divided by $2\pi$.
Hence,

$$\begin{aligned}\mathcal{K}(x_1, x_2) &= 2\,\frac{\sphericalangle(Ax_1, Ax_2)}{2\pi} - \frac{1}{2} \\ &= \frac{1}{\pi}\arccos\left(< \frac{Ax_1}{||Ax_1||}, \frac{Ax_2}{||Ax_2||} >\right) - \frac{1}{2}. \quad \square\end{aligned}$$

**Proof of Proposition 3**
Since the the required Gegenbauer functions have different definitions for $q = 2$ and $q \geq 3$, both cases have to be handled separately. At first, we investigate the case $q \geq 3$.
For brevity let us write $\lambda := \frac{n}{2}$. For all $s, t \in S$ we have

$$\begin{aligned}K(s,t) &= \frac{1}{\pi}\arccos(< s, t >) - \frac{1}{2} \\ &= \frac{1}{\pi}\arccos(\cos(\sphericalangle(s,t))) - \frac{1}{2} \\ &= k(\cos(\sphericalangle(s,t))),\end{aligned}$$

where $k(\sigma) := \frac{1}{\pi}\arccos(\sigma) - \frac{1}{2} \in I\!L^2[-1, 1]$.
Since the kernel function only depends on $\cos(\sphericalangle(s,t))$, it follows by Fenyö and Stolle (1983, p.273), that $\{S_{p,l}^{(n)}\}$ is the complete system of eigenfunctions of $T_K$ with eigenvalues

$$\begin{aligned}\lambda_p &= \frac{4\,\pi^{\frac{n}{2}+1}}{(2p+n)\Gamma\left(\frac{n}{2}\right)} b_p c_p, \text{ for } p \in I\!N_0, \text{ where} \\ b_p &:= \frac{2^{n-1} p! \left(\frac{n}{2}+p\right)\Gamma\left(\frac{n}{2}\right)^2}{\pi\,\Gamma(n+p)}, \\ c_p &:= \int_{-1}^{1} k(\sigma) C_p^{\lambda}(\sigma)\left(1-\sigma^2\right)^{\frac{(n-1)}{2}} d\,\sigma.\end{aligned}$$

We denote by $C_p^{\lambda}$ the $n+2$-dimensional Gegenbauer function. Useful properties of this function are derived in Tricomi (1955).
Since $\lambda > 0$ we have

$$C_p^{\lambda}(x) = \frac{\prod_{j=0}^{p-1}(2\lambda + j)}{\prod_{j=0}^{p-1}(\lambda + \frac{1}{2} + j)} P_p^{\left(\lambda-\frac{1}{2}, \lambda-\frac{1}{2}\right)}(x),$$

where

$$P_p^{(\alpha,\beta)}(x) = \frac{1}{2^p}\sum_{k=0}^{p} \frac{\prod_{m=0}^{k-1}(p+\alpha-k+1+m) \cdot \prod_{m=0}^{p-k-1}(\beta+k+1+m)}{k!(p-k)!}(x-1)^{p-k}(x+1)^k$$

is a Jacobi polynomial. For instance, see Tricomi (1955, p.161 and p.178).
By the doubling formula of the Gamma function $\Gamma(2z) = \frac{2^{2z-1}}{\sqrt{\pi}}\Gamma(z)\Gamma(z+\frac{1}{2})$ we obtain:

$$\lambda_p = \pi^\lambda p\, \frac{\Gamma(\lambda)\Gamma\big(\frac{p}{2}\big)\Gamma\big(\frac{p}{2}+\frac{1}{2}\big)}{\Gamma\big(\lambda+\frac{p}{2}\big)\Gamma\big(\lambda+\frac{p}{2}+\frac{1}{2}\big)}\, c_p.$$

Because of $C_0^\lambda \equiv 1$ and $\arcsin(-x) = -\arcsin(x)$ we obtain

$$\begin{aligned}
c_0 &= \int_{-1}^{1} \frac{1}{\pi}\big(\arccos(x) - \frac{\pi}{2}\big)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x \\
&= -\int_{-1}^{1} \frac{1}{\pi}\arcsin(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x \\
&= -\int_{-1}^{0} \frac{1}{\pi}\arcsin(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x - \int_{0}^{1} \frac{1}{\pi}\arcsin(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x \\
&= -\int_{0}^{1} \frac{1}{\pi}\arcsin(-x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x - \int_{0}^{1} \frac{1}{\pi}\arcsin(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x \\
&= 0.
\end{aligned}$$

Hence, $\lambda_0 = 0$. It is well known (see for example http://functions.wolfram.com/Polynomials/GegenbauerC3/21/01/02/02/), that the function

$$F(x) := -\frac{2\,\lambda}{p\,(p+2\,\lambda)} C_{p-1}^{\lambda+1}(x)\big(1-x^2\big)^{\frac{1}{2}+\lambda}$$

has the derivative

$$F'(x) = C_p^\lambda(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}}.$$

This is needed to simplify $c_p$ for $p > 0$.
Let $p > 0$. Since $\arccos'(x) = -\big(1-x^2\big)^{-\frac{1}{2}}$ we obtain by integration by parts:

$$\begin{aligned}
c_p &= \int_{-1}^{1} \big(\frac{1}{\pi}\arccos(x) - \frac{1}{2}\big) C_p^\lambda(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x \\
&= \frac{1}{\pi}\int_{-1}^{1} \arccos(x) C_p^\lambda(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x - \frac{1}{2}\int_{-1}^{1} C_p^\lambda(x) C_0^\lambda(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x \\
&\overset{T.,\,p.179}{=} \frac{1}{\pi}\int_{-1}^{1} \arccos(x) C_p^\lambda(x)\big(1-x^2\big)^{\lambda-\frac{1}{2}} d\,x - 0 \\
&= \frac{1}{\pi}\Big([F(x)\arccos(x)]_{-1}^{1} - \int_{-1}^{1} \arccos'(x) F(x) d\,x\Big) \\
&= \frac{1}{\pi}\Big(0 - 0 + \int_{-1}^{1} \big(1-x^2\big)^{-\frac{1}{2}} F(x) d\,x\Big) \\
&= -\frac{1}{\pi}\int_{-1}^{1} \big(1-x^2\big)^{-\frac{1}{2}} \frac{2\ \lambda}{p\,(p+2\,\lambda)} C_{p-1}^{\lambda+1}(x)\big(1-x^2\big)^{\frac{1}{2}+\lambda} d\,x \\
&= -\frac{2\ \lambda}{\pi\, p\,(p+2\,\lambda)}\int_{-1}^{1} C_{p-1}^{\lambda+1}(x)\big(1-x^2\big)^{\lambda} d\,x.
\end{aligned}$$

The calculation of this integral is somewhat tedious, so we give only the result:

$$\int_{-1}^{1} C_{p-1}^{\lambda+1}(x)\big(1-x^2\big)^{\lambda} d\,x = \frac{\Gamma\big(\frac{p}{2}+\lambda+\frac{1}{2}\big)}{\Gamma\big(\frac{p}{2}+\lambda+1\big)} \frac{\Gamma\big(\frac{p}{2}\big)}{\Gamma\big(\frac{p}{2}+\frac{1}{2}\big)} \sin\big(\frac{p}{2}\pi\big)^2.$$

An other (rather ugly) expression for this integral can easily be obtained by the explicit representation of $C_{p-1}^{\lambda+1}$. Note, that $\lambda+1=\frac{q}{2}$. Putting together all steps, we obtain:

$$\begin{aligned}
\lambda_p &= \pi^{\lambda} p\, \frac{\Gamma(\lambda)\Gamma\big(\frac{p}{2}\big)\Gamma\big(\frac{p}{2}+\frac{1}{2}\big)}{\Gamma\big(\lambda+\frac{p}{2}\big)\Gamma\big(\lambda+\frac{p}{2}+\frac{1}{2}\big)} c_p \\
&= -\pi^{\lambda} p\, \frac{\Gamma(\lambda)\Gamma\big(\frac{p}{2}\big)\Gamma\big(\frac{p}{2}+\frac{1}{2}\big)}{\Gamma\big(\lambda+\frac{p}{2}\big)\Gamma\big(\lambda+\frac{p}{2}+\frac{1}{2}\big)} \frac{\lambda}{\pi\, p\, \big(\lambda+\frac{p}{2}\big)} \int_{-1}^{1} C_{p-1}^{\lambda+1}(x)\big(1-x^2\big)^{\lambda} d\,x \\
&= -\pi^{\lambda-1}\, \frac{\Gamma(\lambda+1)\Gamma\big(\frac{p}{2}\big)\Gamma\big(\frac{p}{2}+\frac{1}{2}\big)}{\Gamma\big(\lambda+\frac{p}{2}+1\big)\Gamma\big(\lambda+\frac{p}{2}+\frac{1}{2}\big)}\, \frac{\Gamma\big(\frac{p}{2}+\lambda+\frac{1}{2}\big)}{\Gamma\big(\frac{p}{2}+\lambda+1\big)} \frac{\Gamma\big(\frac{p}{2}\big)}{\Gamma\big(\frac{p}{2}+\frac{1}{2}\big)} \sin\big(\frac{p}{2}\pi\big)^2 \\
&= \frac{-\pi^{\lambda+1}}{\Gamma(\lambda+1)}\, \frac{\Gamma(\lambda+1)^2\Gamma\big(\frac{p}{2}\big)^2}{\Gamma\big(\lambda+\frac{p}{2}+1\big)^2}\, \frac{\sin\big(\frac{p}{2}\pi\big)^2}{\pi^2} \\
&= -\frac{1}{2} 2 \frac{\pi^{\frac{q}{2}}}{\Gamma\big(\frac{q}{2}\big)}\, \frac{\Gamma\big(\frac{q}{2}\big)^2\Gamma\big(\frac{p}{2}\big)^2}{\Gamma\big(\frac{q}{2}+\frac{p}{2}\big)^2}\, \frac{\sin\big(\frac{p}{2}\pi\big)^2}{\pi^2} \\
&= -\frac{1}{2}\tau_q \left(\frac{\Gamma\big(\frac{q}{2}\big)\Gamma\big(\frac{p}{2}\big)}{\Gamma\big(\frac{q}{2}+\frac{p}{2}\big)}\, \frac{\sin\big(\frac{p}{2}\pi\big)}{\pi}\right)^2.
\end{aligned}$$

Now let $q=2$. The eigenvalues for $q=2$ can be obtained by calculating the formula

$$\lambda_p = \int_0^{2\pi} \big(\frac{1}{\pi}\arccos(\cos(\sigma)) - \frac{1}{2}\big)\, \cos(p\,\sigma)\, d\,\sigma,$$

given in Fenyö and Stolle (1983). It's not difficult to show, that $\lambda_0=0$ and for $p\in I\!N$ we have

$$\begin{aligned}
\lambda_p &= \int_0^{\pi} \big(\frac{\sigma}{\pi} - \frac{1}{2}\big)\, \cos(p\,\sigma)\, d\,\sigma + \int_{\pi}^{2\pi} \big(\frac{2\pi-\sigma}{\pi} - \frac{1}{2}\big)\, \cos(p\,\sigma)\, d\,\sigma \\
&= -\frac{1}{2} 2\pi \frac{2(1-\cos(p\pi))}{p^2\pi^2} \\
&= -\frac{1}{2} 2\pi \left(\frac{2}{p}\frac{\sin(\frac{p}{2}\pi)}{\pi}\right)^2.
\end{aligned}$$

In order to validate the last equation, note that $\sin(\frac{p}{2}\pi)^2$ is just an indicator function. Hence, the proposition holds also for $q=2$. □

**Proof of Lemma 1**
We compare the simplicial depth in the statistical model for $Z_1,...,Z_N$ with a simplicial

depth for i.i.d. random variables $\tilde{Z}_1, ..., \tilde{Z}_N$, where $\tilde{Z}_n$ is obtained from $Z_n$ by appending an independent standard normal distributed random variable $S_n$. That is, $\tilde{Z}_n = (Z_n, S_n)$ and $\tilde{P}_\theta^{\tilde{Z}_n} := P_\theta^{Z_n} \otimes P_{\mathcal{N}(0,1)}$ is the distribution of $\tilde{Z}_n$. Take $\tilde{f}_\theta$ to be a density of $\tilde{P}_\theta^{\tilde{Z}_n}$.

Simplicial depth $\tilde{d}_S$ and tangent depth $\tilde{d}_T$ of $\theta$ with respect to the observations $\tilde{z}_n = (y_n, t_n, s_n)$ are based on the dependent variable $\tilde{h}(\tilde{z}_n) = s_n y_n$ and the regressor $\tilde{v}(\tilde{z}_n) = s_n x(t_n)$. Note, that the sign of the residual of observation $\tilde{z}_n = (z_n, s_n)$ is given by

$$\begin{aligned} \tilde{\text{sig}}_\theta(\tilde{z}_n) &= \text{sign}(s_n y_n - s_n x(t_n)^T \theta) \\ &= \text{sign}(s_n)\, \text{sig}_\theta(z_n). \end{aligned}$$

Since

$$\begin{aligned} \tilde{d}_T(\theta, \tilde{z}) &= \min_{u \neq 0} \#\{\text{sign}(s_n)\, \text{sig}_\theta(z_n) s_n u^T x(t_n) > 0\} \\ &= \min_{u \neq 0} \#\{\text{sig}_\theta(z_n) u^T x(t_n) > 0\} \\ &= d_T(\theta, z), \end{aligned}$$

tangent depths are equal in both models for $s_1, ..., s_N \neq 0$. This holds also for the harmonized depths and thus, also the simplicial depths coincide, that is, for all $\theta \in \Theta$ and all $\tilde{z}_n = (z_n, s_n) \in \mathcal{Z} \times I\!R$ with $s_n \neq 0$ for $n = 1, ..., N$, we have

$$d_S(\theta, z) = \tilde{d}_S(\theta, \tilde{z}).$$

Thus,

$$\begin{aligned} \otimes_{n=1}^N \tilde{P}_\theta^{\tilde{Z}_n}(\{\tilde{z} : \tilde{d}_S(\theta, \tilde{z}) < \lambda\}) &= (\otimes_{n=1}^N P_\theta^{Z_n}) \otimes (\otimes_{n=1}^N P_{\mathcal{N}(0,1)})(\{z : d_S(\theta, z) < \lambda\} \times I\!R^N) \\ &= \otimes_{n=1}^N P_\theta^{Z_n}(\{z : d_S(\theta, z) < \lambda\}) \end{aligned}$$

for all $\lambda > 0$, so that also the distributions of the simplicial depths are equal in both models.

It remains to show that $\tilde{Z}_1, ..., \tilde{Z}_N$ satisfy the assumptions of Theorem 1. Since the random variables are continuous distributed, conditional densities can be used to check that $\tilde{\text{sig}}_\theta(\tilde{Z}_n)$ is positive (negative) with probability $\frac{1}{2}$, given $\tilde{v}(\tilde{Z}_n) = S_n x(T_n)$.

The main part is to show that $K(\tilde{Z}_1) := \frac{1}{||A\tilde{v}(\tilde{Z}_1)||} A\tilde{v}(\tilde{Z}_1)$ with $A = \begin{pmatrix} 1 & 0 \\ 0 & \Sigma^{-\frac{1}{2}} \end{pmatrix}$ is uniformly distributed on the unit sphere $S$. The random variable $U(y_1, t_1, s_1) := \Sigma^{-\frac{1}{2}} t_1$ is multivariate Cauchy-distributed with density

$$\tilde{f}_\theta^U(u) = \frac{\Gamma(\frac{q}{2})}{\sqrt{\pi^q}} \frac{1}{(1 + u^T u)^{\frac{q}{2}}}$$

and for $\tilde{Z}_1 = (Y_1, T_1, S_1)$ we can write

$$K(\tilde{Z}_1) = \text{sign}(S_1) \frac{1}{\sqrt{1 + U(\tilde{Z}_1)^T U(\tilde{Z}_1)}} \begin{pmatrix} 1 \\ U(\tilde{Z}_1) \end{pmatrix}.$$

It suffices to show that

$$\frac{\mu(V)}{\mu(S)} = \int_V 1 \, d(\tilde{P}_\theta^{\tilde{Z}_1})^K$$

for each event $V \subset S \cap I\!R_{>0} \times I\!R^{q-1}$ and each event $V \subset S \cap I\!R_{<0} \times I\!R^{q-1}$ which is open in $S$, where $\mu$ is the uniform measure on $S$ with $\mu(S) = \tau_q$.

Consider the case $V \subset S \cap I\!R_{>0} \times I\!R^{q-1}$. Letting

$$U(V) := \{u \in I\!R^{q-1} : \frac{1}{\sqrt{1+u^T u}}(1, u_1, ..., u_{q-1})^T \in V\},$$

the function

$$\psi : U(V) \to V, \psi(u) := \frac{1}{\sqrt{1+u^T u}}((-1)^{i+1}, u_1, ..., u_{q-1})^T$$

is a local parametrization of $V$. Hence,

$$\mu(V) = \int_{U(V)} \sqrt{g\psi(u)} d\lambda^{q-1},$$

where the gram determinant $g\psi(u)$ is defined as

$$g\psi(u) = \det \begin{pmatrix} \sum_{j=1}^q \frac{\partial \psi_j}{\partial u_1}(u)\frac{\partial \psi_j}{\partial u_1}(u) & ... & \sum_{j=1}^q \frac{\partial \psi_j}{\partial u_1}(u)\frac{\partial \psi_j}{\partial u_{q-1}}(u) \\ \vdots & & \vdots \\ \sum_{j=1}^q \frac{\partial \psi_j}{\partial u_{q-1}}(u)\frac{\partial \psi_j}{\partial u_1}(u) & ... & \sum_{j=1}^q \frac{\partial \psi_j}{\partial u_{q-1}}(u)\frac{\partial \psi_j}{\partial u_{q-1}}(u) \end{pmatrix}.$$

It is tedious to check that

$$g\psi(u) = \frac{1}{(1+u^T u)^{2(q-1)}} \det((1+u^T u)I - uu^T),$$

where $I = (e_1, ..., e_{q-1})$ is the identity matrix. With $a_{1,j} := (1+u^T u)e_j$, and $a_{2,j} := -u_j u$ for $j = 1, ..., q-1$ we have

$$\det((1+u^T u)I - uu^T) = \det(a_{1,1} + a_{2,1}, ..., a_{1,q-1} + a_{2,q-1}).$$

Since the determinant is linear in each column and since the determinant of a matrix is 0, if two columns are linearly dependent, we obtain

$$\det((1+u^T u)I - uu^T) = (1+u^T u)^{q-2}$$

(see www.owlnet.rice.edu/∼fjones/chap3.pdf, Problem 3-41).

It follows that $g\psi(u) = \frac{1}{(1+u^T u)^q}$ and thus,

$$\mu(V) = \int_{U(V)} \sqrt{\frac{1}{(1+u^T u)^q}} \, d\lambda^{q-1} \quad \text{for } V \subset S \cap I\!R_{>0} \times I\!R^{q-1}. \quad (1)$$

Now let $V \subset S \cap I\!R_{>0} \times I\!R^{q-1}$ or $V \subset S \cap I\!R_{<0} \times I\!R^{q-1}$ be open in $S$. Let $i = 1$ or $i = 2$, such that $(-1)^i V \subset S \cap I\!R_{>0} \times I\!R^{q-1}$. For brevity we write $\tilde{P}_\theta := \tilde{P}_\theta^{\tilde{Z}_1}$.

With $\bar{T}(y_n, t_n, s_n) := t_n$ and $\bar{S}(y_n, t_n, s_n) := s_n$, we have

$$\begin{aligned}
\int_V 1 \, d\tilde{P}_\theta^K &= \tilde{P}_\theta(K \in V) \\
&= \tilde{P}_\theta(K \in V | \bar{S} > 0)\tilde{P}_\theta(\bar{S} > 0) + \tilde{P}_\theta(K \in V | \bar{S} < 0)\tilde{P}_\theta(\bar{S} < 0) \\
&= \tilde{P}_\theta(\frac{Ax(\bar{T})}{||Ax(\bar{T})||} \in V)\frac{1}{2} + \tilde{P}_\theta(-\frac{Ax(\bar{T})}{||Ax(\bar{T})||} \in V)\frac{1}{2} \\
&= \frac{1}{2}\tilde{P}_\theta(\frac{Ax(\bar{T})}{||Ax(\bar{T})||} \in (-1)^i V) \\
&= \frac{1}{2}\tilde{P}_\theta(\psi^{-1}\left(\frac{1}{\sqrt{1+U^TU}}\begin{pmatrix}1\\U\end{pmatrix}\right) \in \psi^{-1}((-1)^i V)) \\
&= \frac{1}{2}\tilde{P}_\theta(U \in \psi^{-1}((-1)^i V)) \\
&= \frac{1}{2}\int\limits_{\psi^{-1}((-1)^i V)} \tilde{f}_\theta^U(u) \, d\lambda^{q-1} \\
&= \frac{\Gamma(\frac{q}{2})}{2\sqrt{\pi^q}}\int\limits_{\psi^{-1}((-1)^i V)} \frac{1}{(1+u^Tu)^{\frac{q}{2}}} \, d\lambda^{q-1} \\
&\overset{(1)}{=} \frac{\mu((-1)^i V)}{\mu(S)} = \frac{\mu(V)}{\mu(S)}
\end{aligned}$$

It follows that the assumptions of Theorem 1 hold, so that $\tilde{d}_S$ has the asymptotic distribution, mentioned there. Since the distributions of the simplicial depths are equal, it follows that also $d_S$ has that asymptotic distribution. □

**Proof of Theorem 2**
We compare the simplicial depth in the statistical model for $Z_1, ..., Z_N \sim P_\theta$ with a simplicial depth of i.i.d. random variables $\tilde{Z}_1, ..., \tilde{Z}_N$, where $\tilde{Z}_n := (\tilde{Y}_n, \tilde{T}_n) := (Y_n, T_n - \mu)$ is obtained from $Z_n$ by shifting $T_n$. Let $\varphi(\theta) := (\theta_0 + \theta_1\mu_1 + ... + \theta_{q-1}\mu_{q-1}, \theta_1, ..., \theta_{q-1})$. The position of the true regression function $g_\theta$ relative to realizations $z_1, ..., z_N$ is equal to the position of $g_\theta(t - \mu) = g_{\varphi^{-1}(\theta)}(t)$ relative to the shifted observations $\tilde{z}_n = (y_n, t_n - \mu)$, so it is convenient to assume that $\tilde{Z}_n \sim (P_{\varphi^{-1}(\theta)})^{\widehat{Z}}$, where $\widehat{Z}(y_n, t_n) = (y_n, t_n - \mu)$. That is, the distribution of $\tilde{Z}_n$ is defined by $\tilde{P}_\theta := P_{\varphi^{-1}(\theta)}^{\widehat{Z}}$ for $\theta \in \Theta$.

Simplicial depth $\tilde{d}_S$ and tangent depth $\tilde{d}_T$ of $\theta$ with respect to the observations $\tilde{z}_n = (\tilde{y}_n, \tilde{t}_n)$ are based on the dependent variable $\tilde{y}_n$ and the regressor $x(\tilde{t}_n)$. We have to show that the distributions of the simplicial depths are equal in both models and that the second model satisfies the assumptions of the previous theorem.

The sign of the residual of observation $\tilde{z}_n = (y_n, t_n - \mu) = z_n - (0, \mu)$ with respect to parameter $\varphi(\theta)$ is given by

$$\begin{aligned} \tilde{\text{sig}}_{\varphi(\theta)}(\tilde{z}_n) &= \text{sign}(y_n - x(t_n - \mu)^T \varphi(\theta)) \\ &= \text{sign}(y_n - x(t_n)^T \theta) \\ &= \text{sig}_\theta(z_n). \end{aligned}$$

Since the function $F : I\!R[X] \to I\!R[X]$ with $p(X) \mapsto p(X + \mu)$ is bijective, it follows that

$$\begin{aligned} \tilde{d}_T(\varphi(\theta), \tilde{z}) &= \min_{u \neq 0} \#\{n : \text{sig}_{\varphi(\theta)}(\tilde{z}_n) u^T x(t_n - \mu) > 0\} \\ &= \min_{u \neq 0} \#\{n : \text{sig}_\theta(z_n) F(u)^T x(t_n - \mu) > 0\} \\ &= \min_{u \neq 0} \#\{n : \text{sig}_\theta(z_n) u^T x(t_n) > 0\} \\ &= d_T(\theta, z). \end{aligned}$$

This holds also for the harmonized depths and thus, for all $\theta \in \Theta$ and all $z_1, ..., z_N \in \mathcal{Z}$ we have

$$\tilde{d}_S(\varphi(\theta), \bar{Z}(z)) = d_S(\theta, z),$$

where $\bar{Z}(z) := ((y_1, t_1 - \mu), ..., (y_N, t_N - \mu))$. Since for all $\lambda > 0$ we have

$$\begin{aligned} \otimes_{n=1}^N \tilde{P}_{\varphi(\theta)}(\{\tilde{z} : \tilde{d}_S(\varphi(\theta), \tilde{z}) < \lambda\}) &= \otimes_{n=1}^N P_\theta^{\widehat{\bar{Z}}}(\{\tilde{z} : \tilde{d}_S(\varphi(\theta), \tilde{z}) < \lambda\}) \\ &= (\otimes_{n=1}^N P_\theta)^{\bar{Z}}(\{\tilde{z} : \tilde{d}_S(\varphi(\theta), \tilde{z}) < \lambda\}) \\ &= \otimes_{n=1}^N P_\theta(\{z : \tilde{d}_S(\varphi(\theta), \bar{Z}(z)) < \lambda\}) \\ &= \otimes_{n=1}^N P_\theta(\{z : d_S(\theta, z) < \lambda\}), \end{aligned}$$

it follows that $(\otimes_{n=1}^N P_\theta)^{d_S(\theta,\cdot)} = (\otimes_{n=1}^N \tilde{P}_{\varphi(\theta)})^{\tilde{d}_S(\varphi(\theta),\cdot)}$ for all $\theta \in \Theta$.
It remains to show that $\tilde{Z}_1, ..., \tilde{Z}_N$ satisfy the assumptions of the previous theorem.

At first we show, that for $s \in \{-1, 1\}$ and for all $\theta \in \Theta$ the conditional probability that $\tilde{\text{sig}}_\theta(\tilde{Z}_n)$ is positive (negative), given $v(\tilde{Z}_n) := x(\tilde{T}_n)$ is equal to $\frac{1}{2}$. For all $v' \in \text{Image}(v)$ we can write $v' = x(t' - \mu)$ with a $t' \in I\!R^{q-1}$. It follows that

$$\begin{aligned} \tilde{P}_\theta(\{\tilde{z}_n : \tilde{\text{sig}}_\theta(\tilde{z}_n) = s\} | v = v') &= P_{\varphi^{-1}(\theta)}^{\widehat{\tilde{Z}}}(\{\tilde{z}_n : \tilde{\text{sig}}_\theta(\tilde{z}_n) = s\} | v = v') \\ &= P_{\varphi^{-1}(\theta)}(\{z_n : \tilde{\text{sig}}_\theta(\widehat{\tilde{Z}}(z_n)) = s\} | v \circ \widehat{\tilde{Z}} = x(t' - \mu)) \\ &= P_{\varphi^{-1}(\theta)}(\{z_n : \text{sig}_{\varphi^{-1}(\theta)}(z_n) = s\} | v = x(t')) \\ &= \frac{1}{2}. \end{aligned}$$

To show that $\tilde{T}_n$ has a centered, multivariate Cauchy distribution, we have to calculate it's density $\tilde{f}_\theta^{\bar{T}}$, where $\tilde{f}_\theta$ is the density of $\tilde{P}_\theta$ and $\bar{T}(\tilde{y}_n, \tilde{t}_n) := \tilde{t}_n$. Take $f_\theta$ to be the density

of $P_\theta$. Then,

$$\begin{aligned}
\tilde{f}_\theta^{\bar{T}}(t_n) &= \int \tilde{f}_\theta(y_n, t_n) dy_n \\
&= \int f^{\widehat{Z}}_{\varphi^{-1}(\theta)}(y_n, t_n) dy_n \\
&= \int f_{\varphi^{-1}(\theta)}(y_n, t_n + \mu) dy_n \\
&= f^{T_n}(t_n + \mu) \\
&= \frac{\Gamma(\frac{q}{2})}{\sqrt{\pi^q |\Sigma|}} \frac{1}{(1 + t^T \Sigma^{-1} t)^{\frac{q}{2}}}
\end{aligned}$$

Hence, the assumptions of Lemma 1 hold, so that $\tilde{d}_S$ has the asymptotic distribution, mentioned there. Furthermore, it follows that also $d_S$ has this asymptotic distribution. □

# Acknowledgement

The shape analysis data are from a study funded by the Slovak Scientific Grant Agency, Project No. 1/9113/02. The data base for the example was created by Stanislav Katina. We thank him for permission to include it.

# References

[1] Bai, Z.-D. and He, X. (1999). Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *Ann. Statist.* **27**, 1616-1637.

[2] Fenyö, S. and Stolle, H.W. (1983). *Theorie und Praxis der linearen Integralgleichungen 2*. Birkhäuser Verlag, Basel.

[3] Lee, A.J. (1990). *U-Statistics. Theory and Practice*. Marcel Dekker, New York.

[4] Liu, R.Y. (1988). On a notion of simplicial depth. *Proc. Nat. Acad. Sci. USA* **85**, 1732-1734.

[5] Liu, R.Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18**, 405-414.

[6] Liu, R.Y. (1992). Data depth and multivariate rank tests. In $L_1$-*Statistical Analysis and Related Methods*, ed. Y. Dodge, North-Holland, Amsterdam, 279-294.

[7] Liu, R.Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* **88**, 252-260.

[8] Mizera, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* **30**, 1681-1736.

[9] Mizera, I. and Müller, Ch.H. (2004). Location-scale depth. *Journal of the American Statistical Association* **99**, 949-966. With discussion.

[10] Mosler, K. (2002). *Multivariate Dispersion, Central Regions and Depth. The Lift Zonoid Approach.* Lecture Notes in Statistics **165**, Springer, New York.

[11] Müller, Ch.H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis* **95**, 153-181.

[12] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probab. Letters* **1**, 327-332.

[13] Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scandinavian J. of Statistics* **26**, 319-343.

[14] Rousseeuw, P.J. and Hubert, M. (1999). Regression depth (with discussion). *J. Amer. Statist. Assoc.* **94**, 388-433.

[15] Storch, U. and Wiebe, H. (1990). *Lehrbuch der Mathematik, Band 2: Lineare Algebra.* Spektrum, Heidelberg.

[16] Tomeček, J., Kováč, V. and Katina, S. (2005). Ontogenetic variability in external morphology of native (Canadian) and nonnative (Slovak) populations of pumpkinseed (Lepomis gibbosus, Linnaeus 1758). *Journal of Applied Ichthyology.* **21**, 335–344.

[17] Tukey, J.W. (1975). Mathematics and the picturing of data. In *Proc. International Congress of Mathematicians*, Vancouver 1974, **2**, 523- 531.

[18] Tricomi, F.G. (1955). *Vorlesungen über Orthogonalreihen.* Springer, Berlin.

[19] Van Aelst, S., Rousseeuw, P.J., Hubert, M. and Struyf, A. (2002). The deepest regression method. *J. Multivariate Anal.* **81**, 138-166.

[20] Wellmann, R. (2007). On data depth with application to regression and tests. *Ph.D. thesis.* University of Kassel, Germany.

[21] Wellmann, R., Katina, S. and Müller, Ch.H. (2007). Calculation of simplicial depth estimators for polynomial regression with applications. *Computational Statistics and Data Analysis* **51**, 5025-5040.

[22] Wellmann, R., Harmand, P. and Müller, Ch.H. (2008). Distribution free tests for polynomial regression based on simplicial depth. To appear in *J. Multivariate Anal.*.

[23] Witting, H. and Müller-Funk, U. (1995). *Mathematische Statistik II.* Teubner, Stuttgart.

[24] Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *Ann. Statist.* **28**, 461-482.

[25] Zuo, Y. and Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *Ann. Statist.* **28**, 483-499.

wellmann@mathematik.uni-kassel.de
Department of Mathematics
University of Kassel
D-34109 Kassel
Germany