Tatra Mt. Math. Publ. **XX** (2006), 1–10

TATRA MOUNTAINS
Mathematical Publications

SIMPLICIAL DEPTH ESTIMATORS AND TESTS IN EXAMPLES FROM SHAPE ANALYSIS

S. Katina — R. Wellmann — Ch.H. Müller

ABSTRACT. In this paper we present the maximum simplicial depth estimator and compare it to the ordinary least square estimator in examples from $2D$ and $3D$ shape analysis focusing on bivariate and multivariate allometrical problems from zoology and biological anthropology. We compare two types of estimators derived under different subsets of parametric space on the basis of the linear regression model, $\theta = (\theta_1, \theta_2)^T \in \mathbb{R}^2$ and $\theta = (\theta_1, \theta_2, \theta_3)^T \in \mathbb{R}^3$, where $\theta_3 = 0$. We also discuss monotonically decreasing linear regression models in special situations. In applications where outliers in x- or y-axis direction occur in the data and residuals from ordinary least-square linear regression model are not normally distributed, we recommend the use of the maximum simplicial depth estimators.

1. Introduction

Allometry is the linear or linearized characterization of the dependence of shape on size. It is frequently used to describe average trends of shape change during postnatal ontogeny (growth and development) and also to describe growth trajectories where this postnatal ontogeny can contribute considerably to adult inter-specific differences [3]. The bivariate concept of allometry focuses on linear regression models of shape and size. In these models, the size measure is usually the natural logarithm of centroid size ($\ln(CS)$, $CS = \sqrt{(\sum_{i=1}^{k} \|\mathbf{x}_i - \overline{\mathbf{x}}_c\|_2)}$, where $\overline{\mathbf{x}}_c$ is the centroid of the $k \times d$ configuration matrix $\mathbf{X}$ with the rows $\mathbf{x}_i$, $d = 1, 2$, and $\|\cdot\|_2$ is L_2 Hilbert - Schmidt norm), and the dependent variable

2000 Mathematics Subject Classification: Primary 62P10; Secondary 62J05, 62H25, 65K05.
Keywords: simplicial depth, maximum depth estimator, distribution-free tests, one-sample tests, two-sample tests, shape analysis, allometry.
This research was supported by *DFG* project No. 436 SLK 17/3/05, the VEGA grants No. 1/3023/06 and 1/2341/05 of Slovak Grant Agency, by *EU FP6* Marie Curie Actions grant *MRTN-CT*-2005-019564 (EVAN). For data acquisition and pre-processing we thank Vladimír Kováč, Markus Bernhard and Philipp Gunz.

is a vector of Procrustes shape coordinates. Procrustes coordinates are derived from $\mathbf{X}$ using generalized Procrustes analysis, which minimizes the sum of square distances between homologous landmarks by translating, rotating and rescaling them to the best (least-squares) fit. The multivariate approach focuses on PCA in size-and-shape space, which is best constructed as a PCA of the Procrustes coordinates augmented by $\ln(CS)$, where mutually uncorrelated PC scores are a projection of a high-dimensional space onto a few-dimensional space and summarize most of the variance present in the data. Our first data set is a bivariate example and the second example a multivariate one. Using a linear regression model, we can test the following hypotheses: (1) for one ontogenetic trajectory: function is linear or polynomial (with the degree 2 or more, but we concentrate here on degree 2), (2) for two or more ontogenetic trajectories: (A) the same intercept and slope – the same ontogenetic trajectories; (B) the same slope but different intercept – parallel trajectories; and (C) different intercept and slope – different trajectories. The term "ontogenetic scaling" refers to differences in length of growth trajectories.

Problems arise when residuals of linear regression model are not normally distributed and some outliers occur in the data. Then one has to use alternative statistical methods. In such situations, we recommend the use of maximum simplicial depth estimators instead of ordinary least-square (OLS) estimators.

2. Mathematical background

Consider a bivariate data set $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. The *halfplane location depth* of an arbitrary point $\theta \in \mathbb{R}^2$ relative to X is defined as

$$ldepth\,(\theta, X) = \min_{H} \#\left\{i : \mathbf{x}_i \in H\right\}, i = 1, ..., n,$$

where H ranges over all closed halfplanes of which the boundary line passes through θ. This depth concept was transferred to regression by e.g. [5], [7], and we discuss it in the following paragraphs.

We assume that the bivariate random variables $Z_1, ..., Z_N$ are independent and identically distributed, that the variables Z_n have values in $\mathcal{Z} \subset \mathbb{R}^2$, and that there is a known family of probability measures $\mathbb{P} = \{P_\theta^Z : \theta \in \Theta\}$ with $\Theta = \mathbb{R}^q$. For given observations $z := (z_1, ..., z_N) \in \mathcal{Z}^N$, we always write $z_n = (y_n, t_n)$, $n = 1, 2, ..., N$. Let x be the function $x : \mathbb{R} \to \mathbb{R}^q, x(t) = (1, t, ..., t^{q-1})^T$. We model the relationship between y_n and t_n by the linear regression model given by

$$y_n = x\,(t_n)^T\,\theta + \varepsilon_n, \tag{1}$$

where $\theta = (\theta_1, ..., \theta_q)^T \in \mathbb{R}^q$. For given observations $z_1, ..., z_N \in \mathcal{Z}$ let $\mathrm{Dom}(z)$ be the set of all those domains with constant depth. A *maximal simplicial depth estimator* for given observations $z_1, ..., z_N \in \mathcal{Z}$ with respect to a subset $\mathbb{K} \subset \mathbb{R}^q$ is defined to be a parameter $\widehat{\theta}_S \in \overline{\arg\max_{\theta \in \mathbb{K}} d_S(\theta, z)}$ [5], [7], where the simplicial depth d_S of θ within z is defined as [6]

$$d_S(\theta, z) := \binom{N}{q+1}^{-1} \#\left\{ \{n_1, ..., n_{q+1}\} \subset \{1, ..., N\} : d_H\left(\theta, \left(z_{n_1}, ..., z_{n_{q+1}}\right)\right) > 0 \right\},$$

harmonized depth is the indicator function $d_H(\theta, z) = I\!I_{S(z)}(\theta)$ and $S(z) \in \mathrm{Dom}(z)$ is a bounded domain. This means that the simplicial depth is the fraction of simplices that contain the parameter θ (for details see [6]).

The maximum simplicial depth estimator $\widehat{\theta}_S$ is not unique. If $\mathbb{K}$ is an affine subspace of $\mathbb{R}^q$ or of a polyhedron, then the closure of the set of all parameters $\theta \in \mathbb{K}$ that maximize $d_S(\cdot, z)|_{\mathbb{K}}$ is a union of polytopes. Let $\mathbb{P}$ be the set of these polytopes. We calculate the vertices $\mathrm{ext}(P) = \left\{\theta_{P,1}, ..., \theta_{P,N_p}\right\}$ of each polytope $P \in \mathbb{P}$, where $\mathrm{conv}(\mathrm{ext}(P))$ is the set of all convex combinations of vertices from P. If we assume that the true probability measure belongs to $\{P_\theta : \theta \in \mathbb{K}\}$, then we can choose a best deepest parameter $\widehat{\theta}_{BD}$ based on L_1 and L_2 minimization from the set of $\widehat{\theta}_S \in \bigcup_{P \in \mathbb{P}} P$.

In order to compare the maximum simplicial depth estimator with some other estimate, we use the OLS-estimator $(\widehat{\theta}_{l_2})$. In the examples, we used the following S-PLUS functions implemented in the basic S-PLUS 6.2 package [5]: `lm` for the OLS-estimator, `anova` for the F-test for the full linear regression model versus the submodel, and `chisq.gof` for the Chi-square goodness-of-fit test of a hypothesized normal distribution versus the OLS residual distribution (the normality was rejected in all linear regression models used in the examples; all p-values < 0.001). S-PLUS programs for simplicial depth estimators and the best of all deepest parameters come from [6].

Assuming the linear regression model (1) we find (1) $\theta = (\theta_1, \theta_2)^T \in \mathbb{R}^2$ and (2) $\theta = (\theta_1, \theta_2, \theta_3)^T \in \mathbb{R}^3$. For the *one-sample problem*, we use a distribution–free, asymptotic α-level test for testing $H_0 : \theta \in \Theta_0$, where $\Theta_0 = \left\{\theta \in \mathbb{R}^3; \theta_3 = 0\right\}$, which can be rejected at significance level $\frac{\alpha}{2}$. We write the test statistic [7] as

$$T(z) = N\left(\sup_{\theta \in \Theta_0} d_S(\theta, z) - \frac{1}{2^q}\right),$$

where the *maximal simplicial depth* is given by $\sup_{\theta \in \Theta_0} d_S(\theta, z)$ for testing $H_0 : \theta \in \Theta_0$.

Let $\mathbb{K}$ be a subset of the parameters space $\Theta \in \mathbb{R}^q$. In the *two-sample problem*, we use a distribution-free, asymptotic α-level test for testing the null hypothesis

that independent observations from two populations can be described by the same polynomial regression function (1) with a parameter in $\mathbb{K}$. For $i = 1, 2$, take θ^i to be the unknown, true parameter for the observations $z^i := (z_{i,1}, ..., z_{i,N_i}) \in \mathcal{Z}^{N_i}$ from the i-th sample. For each population, we make the same assumptions as for the one-sample test. We do not reject the hypothesis $H_0 : \theta^1 = \theta^2 \in \mathbb{K}$ at significance level α, if there is a $\theta \in \mathbb{K}$, such that neither the hypothesis that θ is the true parameter for the first population, nor the hypothesis that θ is the true parameter for the second population, can be rejected at significance level $\frac{\alpha}{2}$. The test statistic is given [6] by

$$T(z^1, z^2) := \max_{\theta \in \mathbb{K}} \Phi_\theta(z^1, z^2),$$

where $\Phi_\theta(z^1, z^2) := \min\left(N_1(d_S(\theta, z^1) - \frac{1}{2^q}), N_2(d_S(\theta, z^2) - \frac{1}{2^q})\right)$, and we reject $H_0 : \theta^1 = \theta^2 \in \mathbb{K}$, if $T(z^1, z^2)$ is less than the $\frac{\alpha}{2}$-quantile of the distribution.

3. Examples

Example 1: $2D$ **shape analysis in zoology**

In the first application we wanted to find out how the relative head length of the North American sunfish pumpkinseed (*Lepomis gibbosus*) in different Canadian and introduced European populations depends on their size during growth.

In Canada, 85 specimens were collected in 2003 from the *Otonabee River* (**oto**) and 117 specimens from the *Looncall Lake* (**loon**). A total of 162 specimens were taken from *Tanyards fisheries pond* near Brighton, England (**eng**). From previous study of the external morphology of pumpkinseeds from an ontogenetical point of view [4], it is known that the smallest pumpkinseed (predominantly juveniles) differ significantly from the largest pumpkinseed (predominantly adults) in all populations studied.

Let t_n be *standard length*, defined as the distance between the anterior tip of the upper jaw and the caudal fin base, and let y_n be *relative head length* defined as the head length divided by t_n [4].

In the linear regression model (1) with $q = 2$, the best $\widehat{\theta}$ in the sense of L_1 and L_2 minimization in the domain $\mathrm{conv}\,(\mathrm{ext}\,(P))$ is $\widehat{\theta}_{BD}$: for **oto** this vector is $(0.2580, -0.0007)^T$, for **eng** $(0.2460, -0.0007)^T$ and for **loon** $(0.2593, -0.0008)^T$. All these estimators are on the domain boundary with 3 vertices.

In the linear regression model (1) with $q = 3$ and $\theta_3 = 0$, the best $\widehat{\theta}$ in the sense of L_1 and L_2 minimization in the domain $\mathrm{conv}\,(\mathrm{ext}\,(P))$ is $\widehat{\theta}_{BD}$ as follows: for **oto** $(0.2578, -0.0006, 0)^T$ (not on the boundary, 4 vertices), and

loon $(0.2563, -0.0007, 0)^T$ and **eng** $(0.251, -0.0007, 0)^T$ (on the boundary, 4 vertices). It is interesting to see the comparison of the linear regression model (1) for $q = 2$ versus $q = 3$ and $\theta_3 = 0$. In **oto** we have very similar estimators (but not exactly the same – the deepest convex hulls are slightly different), but for **loon** and **eng** we have different estimators (the deepest convex hulls are different).

Assuming a linear regression model (1) with parameter $\theta = (\theta_1, \theta_2, \theta_3)^T \in \mathbb{R}^3$, we want to test the hypothesis that the true function is linear, i.e. $H_0 : \theta \in \Theta_0$, where $\Theta_0 = \{\theta \in \mathbb{R}^3; \theta_3 = 0\}$. For **oto**, **loon** and **eng**, the maximal simplicial depth $\sup_{\theta \in \Theta_0} d_S(\theta, z)$ is 0.129, 0.128 and 0.120, test statistics $T(z)$ are 0.370, 0.359, -0.769, and $N = 85, 117, 162$, respectively. If the significance level is 10%, then we can reject only the null hypothesis about the **eng** population, according to [6].

Although the data are not normally distributed, we apply an F-test, yields p-values of 0.794, 0.031, and < 0.0001. So we reject null hypotheses about linearity for the **loon** and **eng** populations. The **loon** decision, which is contrary to the results of the maximal simplicial depth estimate, is due to outliers.

For the population **eng**, we reject the null hypothesis that the true function is linear, i.e. $H_0 : \theta \in \Theta_0$, where $\Theta_0 = \{\theta \in \mathbb{R}^3; \theta_3 = 0\}$ at significant level 10% according to [6]. In the linear regression model (1) with $q = 3$ and $\theta_3 = 0$, the best theta in the sense of L_1 and L_2 minimization in the domain $\mathrm{conv}\,(\mathrm{ext}\,(P))$ is $\widehat{\theta}_{BD} = (0.251, -0.0007, 0)^T$, on the domain boundary with 4 vertices. In the linear regression model (1) with $q = 3$ and $\theta_3 \neq 0$, the best theta in the sense of L_1 criterion is $\widehat{\theta}_{BD} = (0.2667, -0.0017, 0.00001)^T$, on the domain boundary with 8 vertices. We see that relative head length decreases during growth, but when juvenile fishes become adults, the relative head length increases (**Fig.1**). This is biologically puzzling, so we test the hypothesis that the true regression function is monotonically decreasing: the hypothesis that the derivative of the true linear regression function g_θ is negative for big fishes also. Cutting at standard length 107mm, we test $H_0 : g'_\theta(107) \leq 0$ and $g''_\theta \geq 0$, which is equivalent to $H_0 : \theta \in \Theta_0$ with $\Theta_0 := \left\{\widetilde{\theta} \in \mathbb{R}^3 : \widetilde{\theta}_2 + 214\widetilde{\theta}_3 \leq 0 \text{ and } \widetilde{\theta}_3 \geq 0\right\}$. Note it is not so easy to test this hypothesis in the classical way. The test statistic is $T(z) = 0.261$, which is more than the 60% quantile of the asymptotic distribution. We may thus assume that the true linear regression function is monotonically decreasing. The deepest region has 11 vertices (**Fig.1**). An example of some deepest parameter theta for monotonically decreasing functions is $\widehat{\theta}_S = (0.2599, -0.0013, 0.000006)^T$.

Now we investigate whether the growth changes of relative head length of the populations **oto** and **loon** can be described by the same regression line. The test statistic for this two-sample problem is $T(z^1, z^2) = -0.049$, if $q = 2$,

and $T(z^1, z^2) = -0.022$, if $q = 3$. Both are more than the 30% quantile of the asymptotic distribution [6] (**Fig.2**). Hence in both cases with respect to a significance level of 5% there is no rejection. We may assume that the regression lines are equal and described by one line. In the example, the estimate of the deepest parameter theta of such a line is $\widehat{\theta}_S = (0.2627, -0.0008)^T$, if $q = 2$, and $\widehat{\theta}_S = (0.2627, -0.0008, 0)^T$, if $q = 3$ with $\theta_3 = 0$, which maximizes $\Phi_\theta(z^1, z^2)$. Indeed, the deepest lines for **oto** and **loon** are rather similar and the lines for $q = 2$ and $q = 3$ are the same. It is also seen that the deepest regions with $q = 2$ are smaller than those with $q = 3$. Although OLS residuals, either for the full regression model or for the submodel, do not have a normal distribution, we apply the F-test, yields p-value=0.428, so that the F-test and depth test give the same result. The null hypothesis estimator is $\widehat{\theta}_{l_2} = (0.2571, -0.0007)^T$. But the true OLS function for **loon** is quadratic and for the **oto** it is linear, so the F-test is not valid.

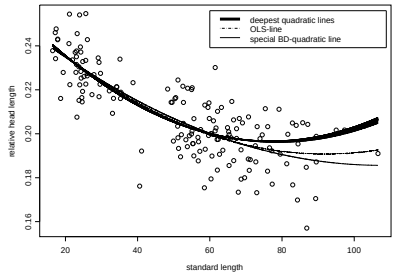

Figure 1: Deepest quadratic lines and the OLS quadratic line (**eng**)

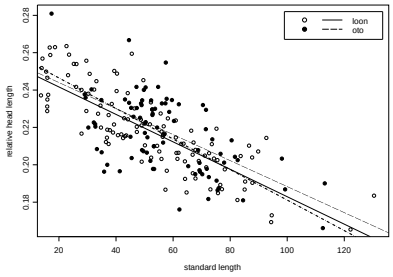

Figure 2: Deepest lines and the H_0 line ($q = 3$, **oto** and **loon**)

Example 2: $3D$ **shape analysis in biological anthropology**

In the second application we re-use part of a Vienna $3D$ data set of 372 crania that has already been the source of several dissertations using semilandmarks: data from 32 landmark points and 7 ridge curves totalling 161 semilandmarks. In this example, we use 6 landmarks and 28 semilandmarks of the midplane, from the three hominid species, subadult and adult male crania of 27 bonobos (*Pan paniscus*), 26 chimpanzees (*Pan troglodytes*) and 68 humans (*Homo sapiens*) [2]. Procrustes coordinates are rotated to the midplane of the average shape. Semilandmarks are allowed to slide along ridge curves to minimize bending energy [1]. Then each of the 121 specimens is unwarped to the pooled average shape [2].

There are clear effects of species and centroid size on the landmark configuration [3].

Let t_n be PC1 scores and y_n PC2 scores (together explaining 60.97% of variability, Pearson's correlation coefficient of $\ln(CS)$ and PC1 scores is 0.9995). For chimpanzee and bonobo males, in the linear regression model (1) with $q = 2, 3$, we can see a difference between $\widehat{\theta}_S$ and $\widehat{\theta}_{l_2}$ that is due to the outliers (**Fig.3**). For bonobos, $\widehat{\theta}_S = (-2.6871, 0.3929, 0)^T$, $\widehat{\theta}_S = (-2.6871, 0.3929)^T$ (domain boundary with 3, resp. 5 vertices) and $\widehat{\theta}_{l_2} = (-1.9845, 0.2914)^T$. For chimpanzees, $\widehat{\theta}_S = (-2.2673, 0.3343, 0)^T$, $\widehat{\theta}_S = (-1.7566, 0.2610)^T$ (domain boundary with 3, resp. 4 vertices) and $\widehat{\theta}_{l_2} = (-1.8704, 0.2772)^T$. For humans, all estimates are very similar, where $\widehat{\theta}_S = (-1.2585, 0.1666, 0)^T$, $\widehat{\theta}_S = (-1.3205, 0.1753)^T$ (domain boundary with 3, resp. 12 vertices) and $\widehat{\theta}_{l_2} = (-1.2575, 0.1665)^T$. In the sense of L_2 minimization, for chimpanzees and bonobos we can see that $\widehat{\theta}_{BD}$ is better if $q = 3$ than if $q = 2$; for humans there is a smaller difference.

Assuming a linear regression model (1) with the parameter $\theta = (\theta_1, \theta_2, \theta_3)^T \in \mathbb{R}^3$, we want to test the hypothesis that the true function is linear, i.e. $H_0 : \theta \in \Theta_0$ where $\Theta_0 = \{\theta \in \mathbb{R}^3; \theta_3 = 0\}$. For bonobos, chimpanzees and humans the maximal simplicial depth $\sup_{\theta \in \Theta_0} d_S(\theta, z) = 0.117$, 0.116 and 0.131, the test statistic $T(z) = -0.224$, -0.266 and 0.415 and $N = 27, 26, 68$, respectively. If the significance level is 10%, then we can not reject all null hypotheses about linearity according to [6].

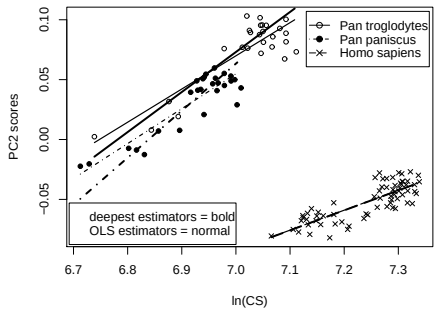

Figure 3: BD and OLS estimators ($q = 3$, *Pan sp.* and *Homo sapiens*)

Although the data are not normally distributed, we apply the F-test to compare OLS models for $q = 2$ and $q = 3$, yields p-values of 0.635, 0.121, 0.379, so we do not reject the null hypothesis about linearity for each hominid male population.

If we compare growth trajectories of bonobo and chimpanzee males, we find $T(z^1, z^2) = -1.101$, if $q = 2$, and $T(z^1, z^2) = -0.744$, if $q = 3$ in both cases more than the 7% quantile of the asymptotic distribution [6]. We may assume that the regression lines are equal and described by one line, where the example of the estimate of the deepest parameter of such a line is $\widehat{\theta}_S = (-2.4277, 0.3558, 0)^T$ (trajectories differ only in the length). If we compare growth trajectories of bonobo and human males and chimpanzee and human males, we find $T(z^1, z^2) = -6.5091$ and -7, if $q = 2$, and $T(z^1, z^2) = -3.375$ and -3.5, if $q = 3$, which is in both cases less than the 1% quantile of the asymptotic distribution [6] (**Fig.3**) and we can reject both null hypotheses – the trajectories have different ontogenetic directions.

Although the data are not normally distributed, we apply the F-test to compare the full nested model and the submodel, which provides p-values of 0.003, < 0.0001, < 0.0001. So, contrary to the simplicial depth estimators for growth trajectories of chimpanzee and bonobo males, we see strong rejection at $\alpha = 0.05$ (difference in intercepts). For the other two tests we have the same result as from the simplicial depth estimators.

4. Conclusions

In this paper we compared the maximum simplicial depth estimators to OLS estimators in examples from $2D$ and $3D$ shape analysis focusing on bivariate and multivariate allometrical problems in zoology and biological anthropology. We compared the behaviour of the two types of estimators in different subsets of parametric space on the basis of linear regression models in practical situations. We also discussed the monotonically decreasing linear regression model when nonmonotonic models make no biological sense. Whenever outliers in the x- or y-axis directions occur and residuals from the OLS linear regression model are not normally distributed, we recommend use of the maximum simplicial depth estimators. We also recommend analysis of not only the global growth of biological organisms but also its separate organismal regions. When the variability of the first two PCs does not exceed 90%, it is appropriate to use PC3 as well – to evaluate growth trajectories in $3D$ space, as can be seen in [1], [2].

REFERENCES

[1] GUNZ, PH.—MITTEROECKER, PH.—BOOKSTEIN, F.L.: Semilandmarks in three dimensions. In. Slice, D. (ed) *Modern Morphometrics in Physical Anthropology* (Kluwer, New York, 2005).

[2] KATINA, S., BOOKSTEIN, F.L., GUNZ, P., SCHAEFER, K.: *Was it worth digitizing all those curves? A worked example from craniofacial primatology. American Journal of Physical Anthropology.* **132**, **S44** (2007), 140

[3] MITTEROECKER, PH.—GUNZ, PH.—BOOKSTEIN, F.L.: Heterochrony and geometric morphometrics: a comparison of cranial growth in *Pan paniscus* versus *Pan troglodytes*. *Evolution and Development.* **7**, **3** (2005), 244–258

[5] MÜLLER, CH.H.: Depth estimators and tests based on the likelihood principle with applications to regression. *Journal of Multivariate Analysis.* **95** (2005), 153–181

[4] TOMEČEK, J.—KOVÁČ, V.—KATINA, S.: Ontogenetic variability in external morphology of native (Canadian) and nonnative (Slovak) populations of pumpkinseed (Lepomis gibbosus, Linnaeus 1758). *Journal of Applied Ichthyology.* **21** (2005), 335–344

[5] VENABLES, W.N.—RIPLEY, B.D.: *Modern Applied Statistics with S-PLUS.* (Springer, New York, 2002)

[6] WELLMANN, R.—KATINA, S.—MÜLLER, CH.H.: Calculation of simplicial depth estimators for polynomial regression with application to tests in shape analysis. *Computational Statistics & Data Analysis.* (accepted)

[7] WELLMANN, R.—HARMAND, P.—MÜLLER, CH.H.: Distribution free tests for polynomial and multiple regression based on simplicial depth. (submitted)

Department of Applied Mathematics and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University
Mlynská dolina
842 48 *Bratislava*
Slovakia
E-mail: katina@fmph.uniba.sk

Department of Anthropology
Faculty of Life Sciences
University of Vienna
Althanstrasse 14
A − 1090 *Vienna*
Austria
E-mail: stanislav.katina@univie.ac.at

Fachbereich Mathematik/Informatik
Universität Kassel
Heinrich-Plett-Str. 40
D − 34132 *Kassel*
Germany
E-mail: wellmann@mathematik.uni-kassel.de
cmueller@mathematik.uni-kassel.de