Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5025 – 5040
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Calculation of simplicial depth estimators for polynomial regression with applications

R. Wellmann[a,*], S. Katina[b,1], Ch.H. Müller[a]

[a] *Department of Mathematics and Computer Sciences, University of Kassel, 34132 Kassel, Germany*
[b] *Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and CS, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia*

Available online 9 November 2006

Abstract

A fast algorithm for calculating the simplicial depth of a single parameter vector of a polynomial regression model is derived. Additionally, an algorithm for calculating the parameter vectors with maximum simplicial depth within an affine subspace of the parameter space or a polyhedron is presented. Since the maximum simplicial depth estimator is not unique, l_1 and l_2 methods are used to make the estimator unique. This estimator is compared with other estimators in examples of linear and quadratic regression. Furthermore, it is shown how the maximum simplicial depth can be used to derive distribution-free asymptotic α-level tests for testing hypotheses in polynomial regression models. The tests are applied on a problem of shape analysis where it is tested how the relative head length of the fish species *Lepomis gibbosus* depends on the size of these fishes. It is also tested whether the dependency can be described by the same polynomial regression function within different populations.

Keywords: Polynomial regression; Simplicial depth; Maximum depth estimator; Distribution free tests; One-sample tests; Two-sample tests; Shape analysis

1. Introduction

For generalizing the median to higher dimensional settings, a variety of different maximum depth estimators have been introduced. They extend, for example, the half space depth of Tukey (1975) or the simplicial depth of Liu (1988, 1990).

The half space depth of Tukey can be characterized as the number of observations that needs to be removed, so that there is a better parameter for all remaining observations. This characterization can be extended to various statistical models. It was transferred to regression by Rousseeuw and Hubert (1999) and studied in a more general context by Mizera (2002), who gave general definitions of such depth notions. He called them 'global depth', 'local depth' and 'tangent depth', gave sufficient conditions for their equality, and investigated the breakdown point of the maximum depth estimator. Further depth notions can be found in the book of Mosler (2002).

* Corresponding author. Tel.: +49 0561 804 4348; fax: +49 0561 804 4646.
E-mail addresses: wellmann@mathematik.uni-kassel.de (R. Wellmann), katina@fmph.uniba.sk (S. Katina), cmueller@mathematik.uni-kassel.de (Ch.H. Müller).
[1] Supported by project No. 436 SLK 17/3/05 (DFG), the VEGA Grants No. 1/3023/06 and 1/2341/05 of Slovak Grant Agency.

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.10.015

Data depth is mainly used for outlier robust estimation. Only few papers deal with tests based on data depth. Van Aelst et al. (2002) derived an exact test based on the regression depth of Rousseeuw and Hubert (1999) but did it only for linear regression. Müller (2005) derived tests for linear and quadratic regression by using the simplicial regression depth, which is the U-statistic with the regression depth as the kernel function. The advantage of the simplicial regression depth is that its asymptotic distribution is known if the spectral decomposition of the reduced normalized kernel function is found. The decomposition was found in Müller (2005) for the linear and quadratic regression model, while in Wellmann et al. (2006), this approach was extended to polynomial regression of arbitrary degree and to multiple regression through the origin. Hence, asymptotic tests are available via simplicial regression depth for the mentioned regression models. Moreover, these tests are distribution free.

For the applicability of these tests, it is important to have fast algorithms for calculating the depth. In this paper we propose an algorithm to calculate the simplicial regression depth of a single parameter vector of a polynomial regression model of arbitrary degree. Moreover, we present an algorithm for calculating the maximum simplicial regression depth. This is not only done within the whole parameter space but also within an affine subspace of the parameter space or within a polyhedron of the parameter space. Such subsets of the parameter space are of interest in hypotheses testing so that these algorithms are in particular useful for testing hypotheses in polynomial regression models.

But the algorithms are also useful for calculating outlier robust estimators based on the maximum simplicial regression depth. Since the maximum simplicial depth estimator, which is the parameter which maximizes the simplicial depth, is not unique in the general case, we propose two methods based on l_1 and l_2 minimization to obtain unique estimators.

In Section 2, the theoretical background of the simplicial regression depth is provided. To improve the tests, the regression depth of Rousseeuw and Hubert (1999) which was used in Müller (2005) to define the simplicial regression depth is modified to a harmonized depth. Based on this harmonized regression depth, the simplicial regression depth and the maximum simplicial depth estimator for polynomial regression is defined in Section 2. Section 3 provides the algorithms for calculating the simplicial depth and the maximum simplicial depth estimator for polynomial regression. It also shows how to find a unique maximum simplicial depth estimator. In Section 4 the distribution-free tests based on simplicial depth for one-sample problems given by Wellmann et al. (2006) are provided and extended to general two-sample tests for polynomial regression. Sections 5 and 6 provide applications. In Section 5 the maximum simplicial depth estimator is compared with other estimators by means of the Hertzsprung-Russell data. The new distribution free tests are applied on some hypotheses appearing in a problem of shape analysis of the fish species *Lepomis gibbosus* in Section 6. There it is also shown that hypotheses which are difficult to test in classical theory can be tested with the new tests.

2. The simplicial regression depth

We assume that the bivariate random variables $Z_1, \ldots, Z_N$ are independent and identically distributed throughout the paper. The variables Z_n have values in $\mathscr{Z} \subset \mathbb{R}^2$. We assume that there is a known family of probability measures $\mathscr{P} = \{P_\theta^{(Z_1,\ldots,Z_N)} : \theta \in \Theta\}$ with $\Theta = \mathbb{R}^q$. For given observations $z_1, \ldots, z_N \in \mathscr{Z}$, we always write $z = (z_1, \ldots, z_N)$ and $z_n = (y_n, t_n)$.

For brevity, we define the sign of an observation $z_n \in \mathscr{Z}$ as

$$\mathrm{sig}_\theta(z_n) := \mathrm{sign}(y_n - x(t_n)^{\mathrm{T}}\theta),$$

where $x(t_n) := (1, t_n, \ldots, t_n^{q-1})^{\mathrm{T}}$.

We say that the observations are alternating, if there is a permutation π with $t_{\pi(1)} < \cdots < t_{\pi(N)}$, such that the sequence $\mathrm{sig}_\theta(z_{\pi(1)}), \ldots, \mathrm{sig}_\theta(z_{\pi(N)})$ is alternating.

Our definition of the simplicial regression depth will be based on the following depth function. This depth function is only needed to define the simplicial regression depth, so that we can restrict the definition on $q + 1$ observations. For a general definition see Wellman et al. (2006).