Journal of Statistical Planning and Inference 140 (2010) 226–245

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimal robust influence functions in semiparametric regression

R. Hable[a,*], P. Ruckdeschel[b], H. Rieder[a]

[a] University of Bayreuth, Department of Mathematics, 95440 Bayreuth, Germany
[b] Fraunhofer-Institut für Techno-und Wirtschaftsmathematik, 67663 Kaiserslautern, Germany

ARTICLE INFO

Article history:
Received 4 May 2007
Received in revised form
12 November 2008
Accepted 8 July 2009
Available online 16 July 2009

MSC:
62F12
62F35
62G35
62N02

Keywords:
Robust statistics
Semiparametric model
Influence function
Neighborhoods
Mean square error
Cox regression

ABSTRACT

Robust statistics allows the distribution of the observations to be any member of a suitable neighborhood about an ideal model distribution. In this paper, the ideal models are semi-parametric with finite-dimensional parameter of interest and a possibly infinite-dimensional nuisance parameter.

In the asymptotic setup of shrinking neighborhoods, we derive and study the Hampel-type problem and the minmax MSE-problem. We show that, for all common types of neighborhood systems, the optimal influence function $\tilde{\psi}$ can be approximated by the optimal influence functions $\tilde{\psi}_n$ for certain parametric models.

For general semiparametric regression models, we determine $(\tilde{\psi}_n)_{n\in\mathbb{N}}$ in case of error-in-variables and in case of error-free-variables.

Finally, the results are applied to Cox regression where we compare our approach to that of Bednarski [1993. Robust estimation in Cox's regression model. Scand. J. Statist. 20, 213–225] in a small simulation study and on a real data set.

1. Introduction

1.1. Motivation

Although in most situations, a model distribution P may serve as a reasonable description for the bulk of the data from an experiment, the real distribution will only approximately be captured by the model. So, robust statistics only assumes that the real distribution lies in a suitable neighborhood about the "ideal" distribution P. Most results in semiparametrics being asymptotic, we employ the asymptotic setup of shrinking neighborhoods. See Bickel (1981), Rieder (1994) and, for a motivation of the shrinking rate $1/\sqrt{n}$, Ruckdeschel (2006).

On arbitrarily small such neighborhoods, bias may get out of control, all the more so if the procedure is efficient at the ideal model. Thus efficiency has to be complemented somehow with a robustness criterion to obtain a reasonable criterion for robust optimality. For the important class of asymptotically linear estimators, two optimality problems have been considered: minimizing the (trace of the asymptotic co)variance subject to a uniform bias bound on the whole neighborhood about P (Hampel-problem) and, more or less equivalent, minimizing the maximal (asymptotic) mean square error on the neighborhood about P (MSE-problem).

* Corresponding author.
E-mail addresses: Robert.Hable@uni-bayreuth.de (R. Hable), peter.ruckdeschel@itwm.fraunhofer.de (P. Ruckdeschel), Helmut.Rieder@uni-bayreuth.de (H. Rieder).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2009.07.010

The starting point of our investigations is Cox regression which is certainly the most common model in survival analysis. For this model, Cox (1972) proposes the so-called Cox partial likelihood estimator. But this estimator is not robust. This is e.g. shown by Samuels (1978), Struthers and Kalbfleisch (1986) and Bednarski (1989). Asymptotically linear estimators for Cox regression which are robust in the sense of Fréchet differentiability are derived in Bednarski (1993) and Sasieni (1993) (cf. also Bednarski and Nowak, 2003). Optimality of these estimators with respect to the statistical criteria mentioned above (Hampel-/MSE-problem), however, has not been investigated by these authors.

If the ideal model is a parametric model, it is well-known how to solve the MSE-/Hampel-problem (cf. Rieder, 1994). Motivated by Cox regression, we consider the MSE-problem in case of semiparametric models where the ideal distribution is only known up to a finite-dimensional parameter and a possibly infinite-dimensional nuisance parameter. In these models, there are three typical situations:

The solution of the parametric model (with nuisance parameter known) is already orthogonal to the tangent space of the nuisance component and, thus, it is the solution for the semiparametric model (with nuisance parameter unknown). This is the case of robust adaptivity. Symmetric adaptive location may serve as an example (cf. Stein, 1956; Bickel, 1982; Stabla, 2005). If, secondly, the tangent space of the semiparametric model equals an L_2-space with respect to some less informative σ-algebra (as it is e.g. in mixture models), an explicit solution can be derived (cf. Shen, 1995; Fischer, 2006). In the third situation, as in Cox regression, where none of the previous assumptions are satisfied, Shen (1995) shows that, for the Hampel-problem and contamination-type neighborhoods, the optimal influence function $\tilde{\psi}$ can at least be approximated by the optimal influence functions $\tilde{\psi}_n$ of certain parametric models.

Firstly, in Section 2, we extend this result to all common types of neighborhood systems and to the MSE-problems as well.

Secondly, in Section 3, the theory of linear robust regression of Rieder (1994, Chapter 7) is extended to a theory of general robust regression. A sequence of approximations $\tilde{\psi}_n$, $n \in \mathbb{N}$, is determined for contamination-type neighborhoods in case of errors-in-variables (Section 3.1) and in case of error-free-variables (Section 3.2).

Thirdly, the results are applied to Cox regression in Section 4 where we compare our approach to that of Bednarski (1993) by means of a small simulation study and a real data set.

1.2. Setup

A typical semiparametric model is a set of probability measures

$$\mathscr{P} = \{ P_{\theta,H} \mid \theta \in \Theta, H \in \mathscr{H} \}$$

on a measurable space $(\Xi, \mathscr{B})$ where Θ is an open subset of $\mathbb{R}^k$ and $\mathscr{H}$ is any infinite-dimensional set so that the parametric submodel $\{P_{\theta,H} | \theta \in \Theta\}$ is L_2-differentiable for every $H \in \mathscr{H}$. For $\theta \in \Theta$ and $H \in \mathscr{H}$, let $\Lambda_{\theta,H}$ be the score function of the parameter θ in $P_{\theta,H}$. Furthermore, let

$$\partial_2 \mathscr{P}_{\theta,H} \subset \left\{ g \in L_2^1(P_{\theta,H}) \,\middle|\, \mathbb{E}_{P_{\theta,H}} g = 0 \right\}$$

be a tangent set of $\mathscr{P}$ at $P_{\theta,H}$. For these basic concepts of semiparametric models, see e.g. Bickel et al. (1993) or van der Vaart (1998).

The class of asymptotically linear estimators consists of all sequences of estimators $(S_n)_{n \in \mathbb{N}}$ such that

$$S_n : (\Xi^n, \mathscr{B}^{\otimes n}) \longrightarrow (\mathbb{R}^k, \mathbb{B}^{\otimes k})$$

and

$$\sqrt{n}(S_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta,H}(x_i) + \mathrm{o}_{P_{\theta,H}^{\otimes n}}(n^0) \tag{1}$$

Here, $\psi_{\theta,H}$ is called (semiparametric) influence function and, according to Rieder (2000, Chapter 2), may be any element of

$$\Psi_{\theta,H} := \left\{ \psi \in L_2^k(P_{\theta,H}) \,\middle|\, \begin{array}{c} \mathbb{E}_{P_{\theta,H}} \psi = 0, \quad \mathbb{E}_{P_{\theta,H}} \psi \Lambda'_{\theta,H} = \mathrm{Id}_{k \times k}, \\ \mathbb{E}_{P_{\theta,H}} g\psi = 0 \ \forall g \in \partial_2 \mathscr{P}_{\theta,H} \end{array} \right\}$$

The influence function of an asymptotically linear estimator indicates its sensitivity to deviations from the ideal model. For the basic types of neighborhood systems used in robust statistics, the bias of an asymptotically linear estimator with influence function $\psi_{\theta,H}$ is defined by

$$\omega_{\theta,H;*}(\psi_{\theta,H}) = \sup \left\{ \left| \mathbb{E}_{P_{\theta,H}} \ \zeta \psi_{\theta,H} \right\| \zeta \in \mathscr{G}_*(\theta,H) \right\}$$

where

$$\mathscr{G}_*(\theta,H) \subset \left\{ \zeta \in L_\infty^1(P_{\theta,H}) \,\middle|\, \mathbb{E}_{P_{\theta,H}} \zeta = 0 \right\}$$