1

Introduction

1.1 Motivation, Bibliographic History, and an Overview of the book.

In most empirical studies, the full (equivalently, complete) data X on certain subjects are censored (equivalently, missing or coarsened). That is, the data X that one would wish to collect are incompletely observed for a (possibly improper) subset of the study subjects; instead, only a random function (equivalently, a random coarsening) Y of X is observed. Furthermore, over the past decades, data from epidemiological, biostatistical, and econometric studies have become increasingly high-dimensional as longitudinal designs that collect data on many time-varying covariate processes at frequent intervals have become commonplace. Scientific interest, however, often focuses on a low-dimensional functional μ of the distribution F_X of the full data – say, as an example, the medians of the treatment-arm specific distributions of time to tumor recurrence in a cancer clinical trial in which recurrence times are right censored by lost -to-follow-up. In such a trial, X is often high dimesional because the study protocol specifies comprehensive laboratory and clinical measurements be taken monthly. In such settings, the use of non-or semiparametric models for F_X that do not model the components of F_X that are of little scientific interest have become commonplace, so as to insure that misspecification of the functional form of a parametric model for the entire distribution F_X does not induce biased estimates of μ. The methodology described in this book was developed to meet the analytic challenges posed by high dimensional censored

data in which a low-dimensional functional μ of the distribution F_X is the parameter of scientific interest.

We provide a general estimating function methodology for locally semiparametric efficient (LSE) estimation of smooth parameters (i.e., parameters estimable at rate $\sqrt{n}$) μ of large non- or semiparametric models for the law F_X of very high dimensional data X based on the observed data Y, when the data are coarsened at random (CAR). The data are said to be CAR if the coarsening mechanism (i.e., the conditional density g of the conditional distribution $G(\cdot \mid X)$ of the observed data Y given the full data X) is only a function of the observed data Y (Heitjan and Rubin, 1991, Jacobsen and Keiding, 1995, and Gill, van der Laan, and Robins, 1997), in which case we refer to the coarsening mechanism $G(\cdot \mid X)$ as a CAR mechanism.

When the data are CAR, the likelihood for Y factors into a part that depends only on the distribution of X and a part that equals the CAR mechanism g. As a consequence, all methods of estimation that obey the likelihood principle (including Bayesian methods with F_X and g apriori independent, and the methods of parametric maximum likelihood, nonparametric maximum likelihood and maximum regularized likelihood) must ignore the CAR mechanism g and thus provide the same inference regardless of whether g is completely known, completely unknown, or known to lie in a (variation independent) parametric or semiparametric model. Because of a historical preference for methods that obey the likelihood principle, a CAR mechanism is referred to as ignorable (Rubin, 1976). However Robins and Ritov (1997) showed that, with high dimensional coarsened at random data, any method of estimation must perform poorly in realistic-sized samples if the CAR mechanism is completely unknown. It follows that, even when the CAR mechanism is completely known or known to follow a lower dimensional model, methods that obey the likelihood principle and thus ignore the coarsening mechanism may perform poorly in realistic-sized samples; in contrast, our generalized estimating functions depend on the CAR mechanism (or a model-based estimate thereof) and thus violate the likelihood principle, yet yield estimators that perform well in the moderate-sized samples occurring in practice. Thus the slogan "With high dimensional censored data, one cannot ignore an ignorable coarsening mechanism."

The general estimating function approach to estimation of parameters of a full-data model based on censored data with known or correctly modelled CAR mechanism was originally introduced by Robins and Rotnitzky (1992), drawing on advances in the efficiency theory of semiparametric models due to Bickel, Klaassen, Ritov, and Wellner (1993), Newey (1990), van der Vaart (1988, 1991), among others. Robins and Rotnitzky restricted their investigation to data structures for which the full data X has a positive probability of being completely observed. As one example, Robins and Rotnitzky (1992) considered the data strucuture in which $X = \{T, (X(t); 0 \leq t \leq T)\}$ is a high-dimensional multivariate time-