Nonparametric efficient estimation of prediction error for incomplete data models

Dissertation zur Erlangung des Doktorgrades
der Mathematischen Fakultät
der Albert-Ludwigs-Universität Freiburg i. Brsg.

vorgelegt von
Thomas Gerds

November 2002

Dekan:	Prof. Dr. Rolf Schneider
1.Referent:	Prof. Dr. Martin Schumacher Institut für Medizinische Biometrie und Medizinische Informatik Albert-Ludwigs-Universität Freiburg Stefan-Meier-Str. 26 79104 Freiburg
2.Referent:	Prof. Jon A. Wellner Department of Statistics University of Washington Seattle, WA 98195 USA

Datum der Promotion: 06.03.2003

Institut für Mathematische Stochastik
Albert-Ludwigs-Universität Freiburg
Eckerstraße 1
D-79104 Freiburg im Breisgau

Contents

Chapter 1

Introduction

This thesis is on defining and estimating of measures of prediction error with incomplete data. We are in particular concerned with the very old question of how to assess and compare probability distributions. In a regression situation we accept as candidate predictions any specification of the conditional probability distribution of a dependent outcome variable given a vector of covariates. This shall include experts guesses or computer intensive methods that can not be considered regression models. Of particular interest are applications in the medical field where one needs to validate a prognosis or diagnosis of a random quantity, which could be any measurement or an event in time related to a particular disease. Since medical diagnoses and prognoses are special forms of predictions we simply talk about predictions or forecasts in this thesis.

For situations with completely observable values of the outcome variable and the covariates, ready made and widely accepted measures of prediction error exist. For instance, in the classical linear regression model the well-known summary statistic R^2 unifies multiple interpretations. Many extensions and generalizations of R^2 exists. At a time, however, typically only one of its interpretations can be preserved outside the linear model with normal errors (Kvalseth 1985). In section 2.1 we recall the definition of R^2, cite related quantities and illustrate the problems that occur when one tries to generalize such summary measures to incomplete data situations.

We are able to show that for continuous outcome variables R^2 can be too crude even in the classical linear model. As a sensitive alternative we suggest prediction error curves defined on the range of the outcome variable. These curves provide a global picture of the predictive accuracy of a given forecast, provide appealing diagnostic curves with multiple functionality, and are specifically useful if the distribution of the outcome variable is not well described by its expectation and variance. Moreover, prediction error curves can be consistently estimated for incomplete data situations, in particular for right censored event times. Applied to the classical linear model, our approach leads to the same conclusions as obtained with R^2 in most situations, and may even show a little bit more under particular circumstances. We explore these issues by means of worked examples in section

2.3.

The basic idea for our definition of prediction error curves is to replace classical squared residuals by squared differences of event indicators and corresponding predicted event probabilities. This idea was established and utilized by meteorologists for the assessment of weather forecasts (Brier 1950). To make things work, we have to assume that predictions are made in terms of predicted probabilities. This means that a forecaster specifies a conditional probability distribution for the outcome variable given the covariates. In terms of patients data this implies that a valid prediction includes a probability distribution of the relevant outcome variable for every patient. A forecast could be thought of as the decision of a statistician by using a regression model: a set of estimated or predicted conditional probabilities is of course part of the result of any statistical regression model. A valid forecast could also be the result of a model selection procedure, a classification scheme, or simply the guess of an expert based on patient specific characteristics. We assume that forecasts are provided externally, for instance based on a build data set. Estimation of the prediction error of a forecast will then use an independent test data set. In applications where predictions are derived from the same data that has to be used for validation, one would apply resampling methods to get rid of the apparent error problem (Efron 1978). This problem, however, is not addressed in this thesis.

We introduce to the special characteristics of our approach to the assessment of prediction error and, particularly, to the concept of predictions made in terms of probabilities in section 2.2. We define prediction error and related quantities as population parameters. A population parameter is defined as an maybe infinite dimensional parameter of the (unknown) underlying joint distribution function of the outcome variable and the covariates. In particular, this includes that we do not assume a parametric model for estimation. Instead we try to find measures of prediction error that are means, or quasi means, of the losses incurred by the single observations. Estimates of prediction error should be of nonparametric nature if we want to arrive at conclusions that are objective with respect to the models which are to be assessed. This is in agreement with commonly used estimators of mean squared error of prediction (MSEP), respectively R^2, in the classical linear model that converge to a well defined parameter in the nonparametric model of all possible distribution function. However, this is in contrast to the approach of Korn and Simon (1990), who, in context of right censored survival data propose to estimate the prediction error of the result of a regression model under the assumptions of the regression model. Our approach bases on and extends the work of Graf (1998b). However, the model used by Graf for right censored survival data is a semiparametric model, at least in the sense of Groeneboom and Wellner (1994, definition 1.1), which we adopt throughout this thesis. We provide nonparametric estimators of the measures of prediction error defined in Graf (1998b) and Graf, Schmoor, and Schumacher (1999).

In situations where a substantial set of values of the outcome variable is not observable (with probability one), parameters such as MSE are not identifiable from any model. Observations in such situations are called incomplete or coarsened. We provide a general definition of measures of prediction error that can be nonparametrically estimated from the observable (maybe incomplete) data. Statistical models for the probability distribution underlying the incomplete data are often called information loss models. This is due to the fact that the nonparametric information bound, in the sense of Bickel, Klaassen, Ritov, and Wellner (1993), for estimating a population parameter with incomplete data, is known to be smaller compared to the corresponding information bound in the underlying complete data model. In section 3.1 we recall the basic definitions needed for identifiability and optimal estimation of real valued and of function valued parameters in a nonparametric model. The function valued parameters we have in mind are the prediction error curves mentioned above.

In section 3.2 we study the incomplete data problem from an abstract viewpoint: how can population parameters of a random variable of interest be estimated if only a random transformation of the random variable is observable. We discuss identifiability and asymptotically efficient estimation in information loss models for the distribution of the observable random transformation. Throughout this thesis, asymptotic efficiency is understood in terms of first order approximations of estimators and general convolution and minimax theorems. We refer to Bickel, Klaassen, Ritov, and Wellner (1993) and Van der Vaart and Wellner (1996) for proofs of these theorems in case of Banach space valued parameters.

Our elaborations lead us to a general definition of inverse probability of censoring weighted estimators: complete observations are reweighted by the current probability of being observed. We proceed along the lines of Van der Vaart (1991) and derive nonparametric information bounds and efficient influence functions for such estimators via spectral decomposition of certain conditional expectation operators. The main cognition hereby is that the existence of asymptotically consistent, Gaussian regular estimators requires further assumptions on the origin of the observable random transformation. The problem is well-known as the non-identifiability of a competing risk.

Situations in which the outcome variable is subject to missing, grouping, or censoring can be summarized under the keyword coarsened data. Justifying its name, a coarsened observation is a random set that almost always includes the values of the unobservable variable, but the information is lost which element of the set the actual value is. In section 3.3 we show that under the coarsening at random assumption (Gill, Van der Laan, and Robins 1995) statistical inference for parameters of the unobservable variable can be performed by taking conditional expectations over the observed set of values. We prove some of the results of Gill, Van der Laan, and Robins (1995) in a setting that is suitable for our aims. In particular, we discuss the connection of the coarsening at random assumption and the conditional independence assumption in the special case of right censored

survival data with covariates.

In section 3.4 we apply the developed tools in the situation of multivariate right censored survival data. We obtain the efficient influence function in an explicit formula for estimation of real and also for function valued parameters.

Among the results of chapter 3 are representations for the parameters of interest in the incomplete data model that are integrated functionals of the so-called inverse probability of censoring function. In our main example, multivariate right censored event times, this function corresponds to the conditional distribution function of the censoring variable given the covariates. Therefore, chapter 4 is devoted to nonparametric estimation of such functionals. We start by studying a closely related problem: optimal nonparametric estimation of integral density derivatives (section 4.1). Borrowing the framework from the literature, see in particular Bickel and Ritov (1988) and Goldstein and Messer (1992), we find that undersmoothed plug-in kernel estimators are asymptotically efficient. In section 4.2 parallel results are obtained for nonparametric functionals that integrate a Hadamard differentiable functional of a conditional distribution function. The reason for doing this is that in the right censoring situation it is well-known, that the reciprocal of the Kaplan-Meier estimator (for the conditional censoring distribution) is a Hadamard differentiable functional. Asymptotically efficient estimators can be obtained by plugging-in a nonparametric estimator for the conditional distribution function, e.g. a symmetrized nearest neighbor type estimator (Stute 1986).

In chapter 5 we modify our abstract definition of prediction error. It is then flexible enough to provide identifiable population parameters under a general coarsening mechanism. In section 5.1 we also dicuss identifiability of the prediction error curves. In section 5.2 we briefly discuss optimal estimation for general coarsening at random situations; similar considerations can be found e.g. in Van der Vaart (1998, section 25.5.3).

In section 5.3 we apply the methods and formulas developed throughout chapters 3 and 4 to the estimation of measures of prediction error in the right censored survival model with covariates. A large number of papers address optimal estimation of functionals for censored survival data. Without covariates, asymptotically optimal estimators for functionals of a survival distribution in presence of independent right censoring can be obtained by plugging-in the Kaplan-Meier estimator (Wellner 1982; Gill 1983; Schick, Susarla, and Koul 1988, only to name a few). Estimation of general Kaplan-Meier integrals, where the integrated function is not necessarily of locally bounded variation has been studied in Stute and Wang (1993) and Stute (1995). Analogous results for situations with covariates can be found under the assumption that censoring is stochastically independent of the survival time and the covariates (Stute 1993; Stute 1996). Also in the paper of Graf, Schmoor, and Schumacher (1999) it is assumed that censoring is independent of the covariates. We show that this assumption leads to an incomplete data model that is of semiparametric nature, where a semiparametric model

is characterized by definition 1.1 of Groeneboom and Wellner (1994).

We generalize the work of Akritas (1994), as we are able to establish explicit expressions for efficient influence functions and information bounds for linear functionals of bivariate survival functions. Moreover, our results seem to be valid for a general class of nonparametric estimators for the involved conditional distribution functions; including the example treated in Hubbard, Van der Laan, and Robins (1998) where we can simplify the formula for the efficient influence function considerably. We arrive at explicit formulas describing the optimal limit distribution for estimation of measures of prediction error, including the measures defined by Graf (1998b), in the semiparametric model of Graf and also for a class of nonparametric models. Finally, we illustrate the use of prediction error curves for right censored survival data with a worked example in breast cancer.

Chapter 2

Measures of prediction error

A further requirement of a statistical model is that it allows a reasonable assessment of the uncertainty in the primary conclusions. Moreover this should be done without introducing unnecessary elaboration and complication. The arguments against complication are nowadays not so much to reduce the burden of computation but rather to make the path between the data and the conclusions more direct and transparent so that sensitivity to assumptions and data deficiencies is easier to assess.

D.R. Cox (1995)

Let $T : (\Omega, \Gamma, \mathcal{P}) \to (\mathbb{R}, \mathbb{B})$ be the dependent outcome variable in a regression problem, where $(\Omega, \Gamma, \mathcal{P})$ is a statistical experiment and $\mathbb{B}$ is the Borel σ-field on $\mathbb{R}$. Let Z be a k-dimensional vector of covariate random or design variables. We denote by $\mathcal{Q}$ the class of all joint probability distributions of (T, Z) that are dominated by some σ-finite measure μ. Suppose the aim is to quantify the predictive power of estimates of the conditional distribution function of T given Z.

We shall work with the following definition of predictions made in terms of probabilities. A prediction or forecast for the distribution of T based on covariate information is any specification of the conditional distribution function of T given Z. Throughout, we denote such predictions by the symbol π. Recall that $\pi : \mathbb{B} \times \mathbb{R}^k \mapsto [0, 1]$ being a conditional probability distribution implies that $\pi(\cdot \mid z)$ is a probability measure on the range of T for almost every z and that $\pi(B \mid \cdot)$ is a random variable for every $B \in \mathbb{B}$. Predictions made in terms of probabilities are part of the result of most of the commonly used regression models after fitting it to a data set. However, we want to explicitely accept other sources of predictions, such as computer intensive classification schemes or plainly guesses of experts.

We assume that forecast conditional probabilities π are established externally, e.g. by fitting a regression model to a build data set. Predictions are not compared to the underlying conditional distribution in terms of bias and standard

error, say; rather we want to assess the performance of predictions based on π in context with a test data set. Estimation of measures for the prediction error of π use an independent test data set. We take the role of the decision maker and shall develop devices suitable for the assessment and comparison of competing forecasts. It is important to note that our approach is related to the work of Dawid (Dawid 1984; Dawid 1985).

In terms of a given test data set with covariate values for n different individuals, say, a valid forecast π is a prognostic system that provides n, not necessarily different, probability distributions dependent on the covariate constellations. For instance, point predictions for T can be obtained by computing the first (conditional) moment of $\pi(\cdot \mid Z)$. The function $z \mapsto m(\pi, z) \equiv \int t\pi(\,\mathrm{d}t \mid z)$, is a forecast of the regression function $z \mapsto m(z) = \mathrm{E}\,(T \mid z)$.

Remark 2.1

- For better discrimination all symbols used for quantities that depend on the test data set are marked with a hat, and others, that depend on a build data set or are derived completely data-free, such as π or $m(\pi, Z)$, do not.

- If the first moment of a prediction made in terms of probabilities exists, then this kind of predictions contain at least the same information as the commonly used predictions of the most likely value of T, from which, on the other hand, it is clearly impossible to regain the complete probability distribution provided by a forecaster.

□

Basically, there are three kinds of applications for which we want to provide tools that preserve their applicability with incomplete data:

1. Comparison of the accuracy of predictions coming from different statistical regression models and other sources.

2. Selection of covariates that have predictive power and explain variation of the outcome of interest: on the one hand by adjusting a particular regression model, on the other hand without model assumptions in a nonparametric setup.

3. Visualize misspecification and overfitting of predicted probability distributions.

In section 2.1 we explain problems with the definition of prediction error outside the classical linear model and in particular for situations with incomplete observations. This gives the motivation for choosing a different but more general path that bases on the ideas of probability forecasting (section 2.2). We extend the concept of scoring rules that were originally invented for the assessment of weather forecasts (Brier 1950; Winkler 1967) to applications with prognostic

systems. In particular, we adapt and extent the ideas of Graf, Schmoor, and Schumacher (1999). The quadratic score which is also known as Brier score will be central to our approach, since it unifies several desirable properties (Savage 1971; Friedman 1983) and allows a decomposition into a calibration and a resolution term (Hand 1997). We keep the settings general, however, to cover measures of predictive accuracy obtained with other loss functions as well, such as e.g. summary measures for time-dependent versions of receiver operating characteristics (ROC) (Heagerty, Lumley, and Pepe 2000). Finally, in section 2.3 we introduce a concept called prediction error curves and illustrate it with worked examples. We introduce plug-in estimators of several parameters representing prediction error and investigate their distributional properties in the situation where (T,Z) are completely observable.

2.1 Prediction error beyond the linear model

The Brier score applied to time-to-event data is the best idea we have had in years at our institute.

M. Schumacher (2002)

Assessment of the accuracy of predictions for a dependent variable T in a regression problem and, closely related, assessment of the variance explained by a model is of practical importance. For the classical linear regression model most, if not all, commonly used measures of a model's predictive capability are based on the squared difference between observations of the outcome variable and point predictions specified by a model. For a (test) data set of size n, $\{(T_i, Z_i) : i = 1 \dots n\}$, point predictions for T can be obtained by evaluating an estimate $m(\pi)$ of the regression function $m(Z) = \mathrm{E}(T \mid Z)$ at the points of realization. The squared residuals are then given by $\{(T_i - m(\pi, Z_i))^2 : i = 1 \dots n\}$. Suppose there is a regression model involved for setting up $m(\pi)$ and the aim is to assess the predictive power of the model. If the model is established independently of the data that is used for validation we call the expected value of the squared residuals mean squared error of prediction (MSEP):

$$\mathrm{MSEP} = \mathrm{E}(T - m(\pi, Z))^2.$$

We emphasize that the expectation in the latter display is taken with respect to the unknown joint distribution of (T, Z) and not with respect to e.g. the model that was used to build $m(\pi)$.

In order to avoid the apparent error problem of estimation of MSEP (Efron 1978) one could split the data into a build or training data set and then validate the result with a test data set. If there is not a training and a test data set

available, one would try to make the 'test' data set as independent as possible of the model π by applying an appropriate resampling procedure (Efron 1986; Efron and Tibshirani 1986). Any estimator of MSEP is considered a measure of prediction error. The most prominent estimator is probably the statistic called residual sum of squares:

$$\text{RSS} = \frac{1}{n}\sum_{i=1}^{n}(T_i - m(\pi, Z_i))^2.$$

For a fixed $m(\pi)$ RSS converges almost surely to MSEP by the law of large numbers. The important point here is that RSS estimates MSEP consistently in the nonparametric model of all distribution functions of (T, Z).

The statistic RSS occurs, appropriately reweighted, in the crucial part of many goodness-of-fit statistics and model selection criteria, e.g. Mallows' criterion, the Akaike information and Schwarz's criterion. Probably the most frequently used estimator of explained variation involving the squared residuals is the coefficient of determination R^2. It is defined as one minus the ratio of the residual sum of squares of the model and the so-called total sum of squares which corresponds to the null model that ignores the covariates completely. The null model produces constant point predictions for T that are equal to the ordinary mean, denoted by $\bar{T}$, of the observations (in the build data set)

$$R^2 = \left(1 - \frac{\sum(T_i - m(\pi, Z_i))^2}{\sum(T_i - \bar{T})^2}\right).$$

If $m(\pi)$ is provided externally, i.e. based on an independent build data set, the statistic R^2 asymptotically consistently estimates the parameter

$$(1 - \text{Var}(m(\pi, Z))/\text{Var}(\bar{T})).$$

For nonparametric estimation of this parameter it is sufficient to consider estimation of MSEP for arbitrary estimates of the regression function, including the naive estimator $\bar{T}$. Note that the U-statistic $(1/(n-1)\sum(T_i - \bar{T})^2$ is a nonparametric estimate of $\text{Var}(T)$; but $m(\pi)$ estimates m consistently only if the model is valid. Thus, if the model is valid, R^2 asymptotically estimates one minus the parameter $\text{Var}(m(Z))/\text{Var}(T)$, which is known as Pearson's correlation coefficient.

In the classical linear model R^2 has multiple interpretations, as a measure of goodness-of-fit, with the classical interpretation of how consistent the data and the model are, as a measure of explained variation, that tells how valuable the covariates are for explaining variation in the outcome values, and also as a correlation coefficient. Apparently, all the nice attributes and multiple interpretations of R^2 can typically not be preserved at the same time outside the classical linear model. See Helland (1987), Kvalseth (1985) and Zengh and Agresti

(2000) for overviews on relevant statistics and common mistakes with the use and the interpretation of R^2 in context with generalized linear models. Doksum and Samarov (1995) investigate nonparametric methods for the assessment of explanatory power of covariates with $m(\pi)$ a nonparametric estimate of the regression function. For the assessment of error rates with binary data we refer to Efron (1978) who studies cross-validation and bootstrap strategies for unbiased estimation of MSEP. Measures for the precision of diagnosis and prognosis of binary outcome are ROC curves, where also time-dependent versions can be considered (Heagerty, Lumley, and Pepe 2000). Graf (1998a) reviews definitions of and common mistakes with measures of explained variation and prediction error proposed for use in survival analysis. Work in this area was done by Korn and Simon (1990), Van Houwelingen and Le Cessie (1990), Schemper and Stare (1996), Schemper and Henderson (2000) and Graf, Schmoor, and Schumacher (1999).

Typically, observations of survival times are right censored. This is a special case of a broad class of situations for which we intend to define prediction error. The problem that motivates us is that if a non neglectable set of values of the outcome variable is unobservable with probability one, then the asymptotic value of RSS is not identifiable from any model. Reweighting schemes can be used for construction of asymptotically unbiased estimates of MSE if there is strictly positive probability for observing all values in the range of (T, Z). However, commonly all candidate prognostic systems perform bad if this probability is low. Perhaps the most appealing way out of the dilemma is the loss function approach proposed by Korn and Simon (1990). However, the authors circumvent the problem of nonparametric estimation of MSE with incomplete data by assuming the model which they intend to assess. The resulting measures do neither converge to an uniquely defined population parameter for different models, nor do they provide convenient strategies for approaching the apparent error problem. In our view, the biggest problem with the approach of Korn and Simon (1990) is that prediction error computed for different models, a semiparametric Cox regression model and a fully parametric Cox regression model, say, are not directly comparable.

We close this section with a list of attributes that a good measure of predictive accuracy should possess. Here we are selecting and extending the points of Kvalseth (1985) for applications where prediction error has to be estimated from incomplete data. One could substitute R^2 or any other relevant statistic for the phrase 'prediction error' in the following table.

PE1 Prediction error must have an intuitively reasonable interpretation.

PE2 The potential range of prediction error has to be well defined with endpoints corresponding to perfect fit and complete lack of fit.

PE3 Positive and negative deviations from the observations should be weighted equally by prediction error.

PE4 Prediction error should be sufficiently general to be applicable to any type of model.

PE5 Prediction error should be such that its values for different models fitted to the same data set are directly comparable.

PE6 Prediction error should not be confined to any specific model-fitting technique, and should be independent of the model which is to be assessed.

PE7 Prediction error should be identifiable as a population parameter in situations with incomplete data.

2.2 Scoring probability forecasts and expected loss

I think most of us feel that if we could use explicitly such variables as , e.g., what people think prices or incomes are going to be, we would be able to establish relations that could be more accurate and have more explanatory value. But because the statistics on such variables are not very far developed, we do not take the formulation of theories in terms of these variables seriously enough.

Leonard J Savage (1971)

In this section we introduce the basic concepts of probability forecasting. Suppose that a number of experts have knowledge not generally available about the probability distribution of T, and the problem (of the decision maker) is one of eliciting personal probability distributions. The idea is to use scoring rules to encourage honesty (Dawid 1986). This means that a forecaster is motivated to reveal the probability distribution she believes is the most accurate for prediction of e.g. future events or of the most likely value of T.

Suppose for the moment that the forecast π of the distribution of T does not depend on Z. Let α be a deterministic function of T; the value of which has to be predicted by using the forecast probability distribution π. Write $\mathrm{E}\,\alpha(T)$ for the expectation of α with respect to the true distribution of T and let $\pi(\alpha) \equiv \int \alpha(t)\,\pi(\,\mathrm{d}t)$ be the prediction of $\alpha(T)$ based on π.

In the traditional approach α is an event indicator function of a set $A \in \mathbb{B}$: $\alpha(T) = 1\{T \in A\}$. A prediction or forecast for α based on π is then simply the probability of the event A under π: $\pi(\alpha) = \pi(A)$. Similarly, point predictions for future values of T are obtained as $\int t\pi(\,\mathrm{d}t)$. Let S be a score function or scoring rule, taking two arguments. The loss of π, denoted by $L(\pi)$, with respect to the function α is defined by:

$$L(\pi, S, \alpha) = \mathrm{E}\,\{S(\alpha(T), \pi(\alpha))\}.$$

The function S is called proper scoring rule when the loss is minimized if π agrees with the true underlying probability distribution of T. If the minimum is unique S is called strictly proper. Thus a (strictly) proper scoring rule motivates a forecaster to use the probability distribution she believes is most accurate for prediction of future values of α. In case of event indicator functions $L(\pi,S)$ is completely specified by the weighted sum of the two possible values:

$$L(\pi, S) = P(T \in A)\, S(1, \pi(A)) + P(T \notin A)\, S(0, \pi(A)).$$

The loss incurred by a single observation $T = t$ is $S(\alpha(t), \pi(\alpha))$. Given a dataset with values for T corresponding to different individuals of a homogeneous population the loss of π can be estimated by the average loss:

$$\hat{L}_n(\pi) = \frac{1}{n} \sum_{i=1}^{n} (S(\alpha(T_i), \pi(\alpha))).$$

It turns out that basically there are only two possibilities to construct proper scoring rules; either rate the differences $\{\alpha(t) - \pi(\alpha)\}$ or the ratio $\{\alpha(t)/\pi(\alpha)\}$ of the observations and the predictions. Savage (1971) shows that in the former case reasonable loss functions (loss has to be nonnegative, zero at zero and not zero everywhere) are of the form $S(\alpha(t), \pi(\alpha)) = M\{\alpha(t) - \pi(\alpha)\}^2$, where M is some positive constant. The choice $M = 1$ corresponds to the Brier or quadratic score (Brier 1950):

$$S_{BS}(\alpha(t), \pi(\alpha)) = (\alpha(t) - \pi(\alpha))^2.$$

Rating predictions via the ratio discrepancy is too quite restrictive: Savage (1971) proves that proper scoring rules in this case are of the form $S(\alpha(t), \pi(\alpha)) = M\{\alpha(t)/\pi(\alpha) - 1 - \log(\alpha(t)/\pi(\alpha))\}$, for some positive constant M. If $\alpha(T) = 1\{T \in A\}$ the logarithmic score is defined by

$$S_{LS}(\alpha(t), \pi(\alpha)) = \begin{cases} -log(\pi(\alpha)) & \text{if } \alpha(t) = 0, \\ -log(1 - \pi(\alpha)) & \text{if } \alpha(t) = 1. \end{cases}$$

Note that the usage of the logarithmic score for more general functions $\alpha(T)$ is limited to predictions π, such that $\pi(\alpha) > 0$. Another special case in which the logarithmic score is prominent is prediction of the density function $q = \mathrm{d}Q/\,\mathrm{d}\mu$, of Q with respect to a dominating σ-finite measure μ. Then $\alpha = q$, predictions are e.g. given by $\pi(\alpha) = \mathrm{d}\pi/\,\mathrm{d}\mu$, and the logarithmic score is defined by $S_{LS}(t, \pi(\alpha)) = -log\{(\,\mathrm{d}\pi/\,\mathrm{d}\mu)(t)\}$. The distance function associated with S_{LS} is the well-known Kullback-Leibler information (Dawid 1986). The logarithmic score and the Kullback-Leibler information are minimized if $\mathrm{d}\pi/\,\mathrm{d}\mu$ is the true density function. However, it is neither obvious how to define logarithmic score for arbitrary measurable functions of T nor how to generalize to forecasts that depend on covariates. Therefore, we are mainly concerned with the Brier score in

this thesis, which is a proper scoring rule and which is also effective in the sense of Friedman (1983).

Our intended definition of the prediction error of a forecast π takes into account that π may depend on covariates. We define an 'aspect' of the prediction error of π as a measurable function of T, for which accurate predictions are of practical relevance. For instance, in the field of medical diagnosis and prognosis prediction of events in the course of a disease are important. Thus, interesting aspects are event indicator functions of T. In a similar way one could be interested in predicting the conditional variance or the conditional quantiles of T and use correspondingly defined aspects. Clearly, a forecast for those aspects can be derived from a valid set of predictions made in terms of probabilities. From now on we use the notation

$$S(t, z) = S(\alpha(t), \pi_z(\alpha)),$$

where $\pi_Z(\alpha) = \int \alpha(t)\, \pi(\mathrm{d}t \mid Z)$ is the forecast of α based on π.

Definition 2.2 (Abstract prediction error) *Let S be a scoring rule, let α be a function of T which is uniformly integrable on $\mathcal{Q}$. The prediction error of π with respect to α and S is defined as the expected loss, via $\psi : \mathcal{Q} \to \mathbb{R}$:*

$$\psi(Q) = L(\pi, \alpha, S) = \int S(t, z)\, Q(\mathrm{d}t, \mathrm{d}z).$$

□

We may think of the loss of π with respect to the function α as an aspect of abstract prediction error of π. It is clear that given to two such aspect functions, α_1 and α_2, and two predictions, π_1 and π_2, the relation $L(\pi_1, \alpha_1, S) > L(\pi_2, \alpha_1, S)$ does not imply $L(\pi_1, \alpha_2, S) > L(\pi_2, \alpha_2, S)$. This shows that a sensible choice of aspects of prediction error for each application has to be made beforehand. That means the decision maker has to device a weights for all $\alpha \in \mathcal{H}$ for a reasonably chosen class of aspect functions $\mathcal{H}$.

Prediction error thus defined satisfies PE2 (c.f. section 2.1) if the scoring rule is proper. Savage (1971) shows that the quadratic score is also the only proper scoring rule that is symmetric and thus the only proper scoring rule that satisfies PE3. PE4 is satisfied for every aspect α such that $\pi_Z(\alpha)$ is defined and interpretable as a prediction of α. We take care of PE5 and PE6 by taking the expectation in definition 2.2 with respect to the underlying distribution of (T, Z) in the nonparametric model $\mathcal{Q}$. Any consistent estimator of $L(\pi, S, \alpha)$ will therefore be of nonparametric nature. PE7 will be dealt with in an analogous definition of prediction error appropriate for situations with incomplete data in chapter 5. At this point we only recall that the usual measures of prediction error are not identifiable from any model if a subset of the range of (T,Z) which is not a null set with respect to Q, is not observable with probability one. This fact motivated the abstract definition of prediction error given above. We will see in

chapter 5 that it is flexible enough to provide measures of prediction error that are identifiable as population parameters for situations with incomplete data.

Example 2.3 (MSEP)
Let $\alpha(t) = t$ be the identity function. The mean squared error of prediction of π is the expected Brier score corresponding to α:

$$\mathrm{E}\, S_{BS}(T, \pi_Z(\alpha)) = \mathrm{E}\,(T - m(\pi, Z))^2,$$

where $m(\pi, Z) = \int t\, \pi(\,\mathrm{d}t \mid Z)$.

□

Example 2.4 (Brier score)
Let A be a subset of the range of T and set $\alpha(t) = 1\{t \in A\}$. The expected Brier score of π for A is given by

$$\mathrm{E}\, S_{BS}(\alpha, \pi_Z(\alpha))) = \mathrm{E}\,(1\{T \in A\} - \pi(A \mid Z))^2.$$

□
It is well-known that MSEP can be represented as the sum of the variance and the squared bias of the prediction π. The decomposition into two terms, one for calibration and one for resolution can be obtained for the more general situation where α is any arbitrary measurable function, such that $\pi_Z(\alpha) = \int \alpha(t)\pi(\,\mathrm{d}t \mid Z)$ exists almost surely:

$$\mathrm{E}\,(\alpha(T) - \pi_Z(\alpha))^2 = \mathrm{E}\,(\alpha(T) - \mathrm{E}\,(\alpha(T) \mid Z))^2 + \mathrm{E}\,(\mathrm{E}\,(\alpha(T) \mid Z) - \pi_Z(\alpha))^2.$$

It should be clear that any estimate of an aspect of prediction error, such as defined in 2.2, can be used for comparison of different forecast probabilities. Choosing one such aspect we can also analyze the predictive power of a particular covariate in that direction. For instance, by using a regression model framework and then comparison of the forecast when the model is adjusted for the covariate against the forecast obtained when the covariate is omitted. This works for all kinds of covariates that are consistent with the regression model considered and the model can be adjusted for supplement covariates that have potential influence on the outcome variable. Testing the predictive power of a covariate without model assumptions requires a nonparametric estimate of the conditional distribution of T given Z. This estimate could be validated by using an estimates of the expected loss for the nonparametric prediction obtained, when subsequently including and excluding the candidate prognostic factor. For instance Doksum and Samarov (1995) provide interesting results in that direction for the complete data situation.

The next example shows that the Brier score can be used as a summary measure for ROC diagnostics.

Example 2.5 (ROC)
Suppose that T is a binary variable taking on the values 1 meaning 'diseased' and 0 meaning 'disease free'. A diagnostic test based on a continuous covariate Z and cutpoint ξ is a decision rule that says 'diseased' if $Z > \xi$, say, and 'disease free' otherwise. The ROC curve is the monoton increasing function that assigns to one minus the specificity at ξ of such a diagnostic test the corresponding value for the sensitivity:

$$\mathrm{ROC}(\xi) = \{(1 - P(Z \leq \xi \mid T = 0)), P(Z > \xi \mid T = 1)\}$$

Accordingly, an analogous function, called predictive value curve from now on, assigns to one minus the positive predictive value at ξ, which is $P(T = 1 \mid Z > \xi)$, the corresponding negative predictive value $P(T = 0 \mid Z \leq \xi)$. The power of the diagnostic test is high if commonly the four values, sensitivity, specificity, positive and negative predictive values are close to one. This observation is taken into account by most of the existing summary measures for the ROC curve, featuring the area under the ROC curve as the most prominent example.

We observe that a diagnostic test provides a prediction made in terms of probabilities that is specified by $\pi(1 \mid Z) = 1\{Z > \xi\}$, say. If $\alpha(T) = 1\{T = 1\}$ the expected Brier score of π (c.f. definition 2.2) reads

$$\begin{aligned} L(\pi, \alpha, S_{BS}) &= \mathrm{E}\left(1\{T = 0\} - 1\{Z \leq \xi\}\right)^2 + \mathrm{E}\left(1\{T = 1\} - 1\{Z > \xi\}\right)^2 \\ &= \{1 - P(Z > \xi \mid T = 1)\}\, P(T = 1) \\ &\quad + \{1 - P(Z \leq \xi \mid T = 0)\}\, P(T = 0). \end{aligned} \tag{2.1}$$

The very right hand side of the previous display is a weighted sum of one minus the sensitivity and one minus the specificity of π. If the prevalence $P(T = 1)$ is high more weight is put on the sensitivity and on the specificity otherwise. This shows that we can use the Brier score as a natural summary measure for the diagnostic power of diagnostic tests. At the same time the expected Brier score is a weighted sum of negative and positive predictive value, viz.

$$\begin{aligned} L(\pi, \alpha, S_{BS}) = \{1 - P(T = 1 \mid Z > \xi)\}\, P(Z > \xi) \\ + \{1 - P(T = 0 \mid Z \leq \xi)\}\, P(Z \leq \xi). \end{aligned}$$

A somewhat more general form of predictions made in terms of probabilities in the present situation is determined by the two probabilities $\pi(1 \mid Z > \xi)$, $\pi(1 \mid Z \leq \xi)$; in which case the expected Brier score for the event $T = 1$ is given by

$$\begin{aligned} L(\pi, \alpha, S_{BS}) = &(1 - \pi(1 \mid Z \leq \xi))^2\, P(Z \leq \xi, T = 1) \\ &+ (1 - \pi(1 \mid Z > \xi))^2\, P(Z > \xi, T = 1) \\ &+ (1 - \pi(0 \mid Z \leq \xi))^2\, P(Z \leq \xi, T = 0) \\ &+ (1 - \pi(0 \mid Z > \xi))^2\, P(Z > \xi, T = 0). \end{aligned} \tag{2.2}$$

Obviously, (2.2) generalizes (2.1). We note that ROC analysis is available only for a limited class of predictions made in terms of probabilities, respectively associated diagnostic tests, while the expected Brier score is applicable to any classification rule that produces a decision for binary outcome based on covariate information. For the assessment of the predictive power of one continuously distributed prognostic factor, however, ROC analysis can be generalized for continuous outcome, which can be generalized to the right censored survival situation (Heagerty, Lumley, and Pepe 2000).

□

2.3 Prediction error curves

Prediction error curves to be defined in this section provide a graphical representation of prediction error over the range of T. We consider score processes much in the spirit of Nolan (1992) with the main difference that we assume forecast probabilities fixed and established externally. In situations, where the distribution is not well described by its expectation and variance only, considering prediction error curves should provide a more detailed picture of the predictive accuracy of a forecast π. Moreover, the decision maker is given the opportunity to define weighting schemes accordingly to the problem at hand to arrive at a real valued summary measure of prediction error.

Definition 2.6 (Prediction error processes) *Let S be a scoring rule, and let $\mathcal{H}$ be a class of uniformly integrable functions of T. The prediction error curve with respect to S and $\mathcal{H}$ is defined as a functional on $\mathcal{Q}$ via $\psi : \mathcal{Q} \to l^\infty(\mathcal{H})$*

$$\psi(Q, \alpha) = L(\pi, \alpha, S) = \int S(t, z)\, Q(\mathrm{d}t, \mathrm{d}z)$$

where $l^\infty(\mathcal{H})$ is the set of bounded real functions on $\mathcal{H}$ with the uniform norm.

□

Suppose S is the quadratic score. Setting $\mathcal{H} = \{1\{T > t\} : t \in \mathbb{R}\}$ then yields the following prediction error curve on the range of T:

$$t \mapsto \mathrm{PEC}(t) \equiv \int \{1\{s > t\} - \pi((t, \infty) \mid z)\}^2\, Q(\mathrm{d}s, \mathrm{d}z). \tag{2.3}$$

(Alternatively, we could use the class of functions $\{1\{T \leq t\} : t \in \mathbb{R}\}$.)

Example 2.7 (Weighted prediction error)
A class of summary measures for the prediction error process is obtained by averaging, suitably weighted to meet the problem at hand (Matheson and Winkler

1976; Graf, Schmoor, and Schumacher 1999). Suppose ω is a σ-finite (probability) measure on the range of T, then a summary measure called weighted prediction error or integrated Brier score is given by

$$\begin{aligned}\mathrm{WPE} &\equiv \int \mathrm{PEC}(t)\,\omega(\,\mathrm{d}t) \\ &= \int \mathrm{E}\left\{\int \{1\{T > t\} - \pi((t,\infty) \mid Z)\}^2\right\}\omega(\,\mathrm{d}t) \\ &= \mathrm{E}\left\{\int \{1\{T > t\} - \pi((t,\infty) \mid Z)\}^2\,\omega(\,\mathrm{d}t)\right\}.\end{aligned}$$

In practical applications where for some reason accurate predictions of low values of T are more important than for high values ω would assign monotone decreasing weights. If no such preferences are present the choice would be either uniform weights (Graf, Schmoor, and Schumacher 1999) or weights according to the empirical distribution function of T (Schemper and Henderson 2000).

□

The rest of the section is used to define estimators for the parameters defined in definitions 2.2 and 2.6. Let $\{(T_i, Z_i) : i = 1 \ldots n\}$ be an *iid* sample of (T,Z). For a given forecast π determines n not necessarily different probabilities: $\{\pi(\cdot \mid Z_i) : i = 1 \ldots n\}$. The empirical measure is defined by

$$\hat{Q}_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1\{(T_i, Z_i) \in A\},$$

where $A \in \mathbb{B} \times \mathbb{B}^k$. A natural nonparametric estimator of PEC is obtained by pluggin in the empirical measure:

$$\widehat{\mathrm{PEC}}(t) = \frac{1}{n}\sum_{i=1}^{n} \{1\{T_i > t\} - \pi((t,\infty) \mid Z_i)\}^2. \tag{2.4}$$

$\widehat{\mathrm{PEC}}$ is a step function that has jumps at the points of realizations $\{T_1, \ldots, T_n\}$. By using the notation

$$S_{BS}(t; T, Z) = \{1\{T > t\} - \pi((t,\infty) \mid Z)\}^2$$

equation (2.3) reads

$$\mathrm{PEC}(t) = \int S_{BS}(t; s, z)\, Q(\,\mathrm{d}s,\, \mathrm{d}z).$$

Fix $t \in \mathbb{R}$. By the strong law of large numbers the estimator

$$\widehat{\mathrm{PEC}}(t) = \int S_{BS}(t; s, z)\, \hat{Q}_n(\,\mathrm{d}s,\, \mathrm{d}z) \tag{2.5}$$

is consistent for estimating PEC(t). Since $\mathrm{E}\,(S_{BS}(t))^2 < \infty$, we can apply the central limit theorem to obtain for every $t \in \mathbf{R}(T)$ that

$$\mathrm{PEC}(t) \Rightarrow \mathcal{N}(0, \Sigma),$$

where the asymptotic variance is given by $\Sigma = \mathrm{E}\,\{S_{BS}(t; T, Z)^2\} - \mathrm{PEC}(t)^2$. It is well known, that the plug-in estimator is an asymptotically efficient estimator and that Σ^{-1} is the nonparametric information bound for estimation of PEC(t), see BKRW. Similarly we obtain that the estimator

$$\widehat{\mathrm{WPE}} = \int\int \{1\{T > t\} - \pi((t, \infty) \mid Z)\}^2 \, \omega(\,\mathrm{d}t)\, \hat{Q}_n(\,\mathrm{d}s,\, \mathrm{d}z)$$

is an asymptotically efficient, Gaussian regular estimator for WPE.

Now consider estimation of the function valued estimator defined in 2.6. Due to measurablility problems weak convergence of estimators with values in $l^\infty(\mathcal{F})$, where $\mathcal{F}$ is a class of functions, has to be understood in terms of outer measure. Note, however, that this is not made visible here in the notation. We refer to Van der Vaart and Wellner (1996) for a comprehensive representation of the corresponding (empirical process) theory. In case of the function valued estimator

$$\{\widehat{\mathrm{PEC}}(t) : t \in \mathbb{R}\},$$

where $\widehat{\mathrm{PEC}}(t)$ is given in (2.5), it is comparably simple to arrive at a functional central limit theorem. Before we write down the limit distribution of the estimator. We briefly introduce some notions and notation of empirical process theory. For any Q-integrable function φ, $Q\varphi$ is linear functional notation for $\int \varphi \,\mathrm{d}Q$. A class $\mathcal{F} \subseteq \mathcal{L}_1(Q)$, where $\mathcal{L}_1(Q)$ is the set of integrable functions of U, is called Q-Glivenko-Cantelli class if almost surely

$$\sup_\varphi(|Q_n\varphi - Q\varphi|) \to 0.$$

Introduce by $\mathcal{G}_n = \sqrt{n}(\hat{Q}_n - Q)$ the empirical process associated with Q_n. A class $\mathcal{F} \subseteq \mathcal{L}_2(Q)$ is called Q-Donsker class if $\mathcal{G}_n$ converges weakly in $l^\infty(\mathcal{F})$:

$$\{\mathcal{G}_n\varphi : \varphi \in \mathcal{F}\} \Rightarrow \mathcal{G},$$

where $\mathcal{G}$ is a tight Borel measurable element in $l^\infty(\mathcal{F})$. $\mathcal{G}$ is sometimes called transformed Q-Brownian bridge because its covariance is similar to the covariance of the ordinary Brownian bridge:

$$\begin{aligned} \mathrm{E}\,(\mathcal{G}\varphi) &= 0 \\ \mathrm{E}\,(\mathcal{G}\varphi\, \mathcal{G}g) &= Q\varphi g - Q\varphi Qg. \end{aligned}$$

Whether a class $\mathcal{F}$ is Glivenko-Cantelli or Donsker depends merely on its size which is usually measured in terms of entropy or bracketing numbers. If $\mathcal{F}$ is Q-Donsker, the plug-in estimator $\psi(\mathcal{F},\hat{Q}_n)$ is an asymptotically efficient estimator

for $\psi(\mathcal{F},Q)$ (Van der Vaart and Wellner 1996, section 3.11.1). The empirical process plug-in estimator for the prediction error curves corresponds to the class of functions given by

$$\mathcal{F} = \{S_{BS}(t;T,Z) : t \in \mathbb{R}\}. \tag{2.6}$$

It is easy to show that the class $\mathcal{F}$ given in 2.6 is a so-called Vapnik-Červonenkis subgraph class (VC subgraph class) of functions. For, the class of half intervals $\{(t,\infty) : t \in \mathbb{R}\}$ is a VC class of sets. Now, any VC subgraph class is a Donsker class. The process $\{\widehat{\mathrm{PEC}}(t) : t \in \mathbb{R}\}$ converges outer weakly in $l^\infty(\mathcal{F})$ to a mean zero Brownian bridge process with asymptotic covariance matrix given by

$$\mathrm{E}\left(S_{BS}(t;T,Z), S_{BS}(s;T,Z) = \mathrm{E}\left\{S_{BS}(t;T,Z)\, S_{BS}(s;T,Z)\right\} - \mathrm{PEC}(t)\mathrm{PEC}(s).\right.$$

Our first example with real data illustrates the usefulness of the prediction error curves. It deals with a classical linear model and we can therefore compare R^2 and a quasi-R^2 statistic based on WPE. Estimation of prediction error curves is based on the plug-in estimators defined above.

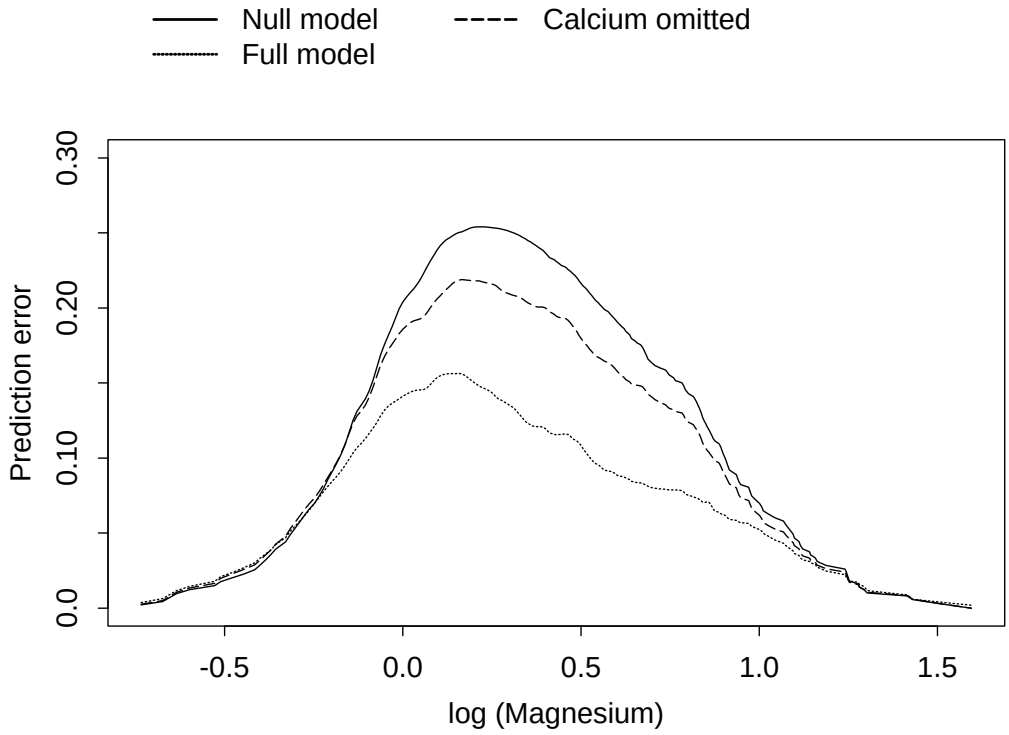

Figure 2.1: Prediction error curves for the full model, the null model and a reduced model omitting the factor calcium concentration.

Example 2.8 (Predicting magnesium concentration)
In the following illustrative data analysis the concentration of magnesium in the needles of trees is the dependent variable. The logarithm of the magnesium

concentration can be considered normally distributed and the standard linear regression model was the main tool for the statistical analysis Soil characteristics, trace elements and the age of the trees were considered as prognostic factors. The main effect was observed for the concentration of calcium in the needles. The linear model with all covariates is called full model in the following. The graph of the estimated prediction error curve for the full model should be compared to the corresponding graph for the null model that ignores covariate information completely and provides constant predictions for magnesium concentration (2.1). Now, if we omit the most effective factor calcium concentration from the linear model, the prediction error curve increases considerably (2.1). This reflects the predictive power of the factor calcium concentration. The values of R^2 for the null model, the full model and the reduced model are 0, 0.66 and 0.21, respectively. Using uniform weights, WPE yields 0.157, 0.099 and 0.139 for the null model, the full model and the reduced model, respectively. We can define a quasi-R^2 statistic by $1 - \text{WPE(full)}/\text{WPE(null)}$, say, and obtain the values 1, 0.37 and 0.11 for the null model, the full model and the reduced model, respectively. We may therefore conclude that the two approaches would lead to similar results, at least in this example.

□

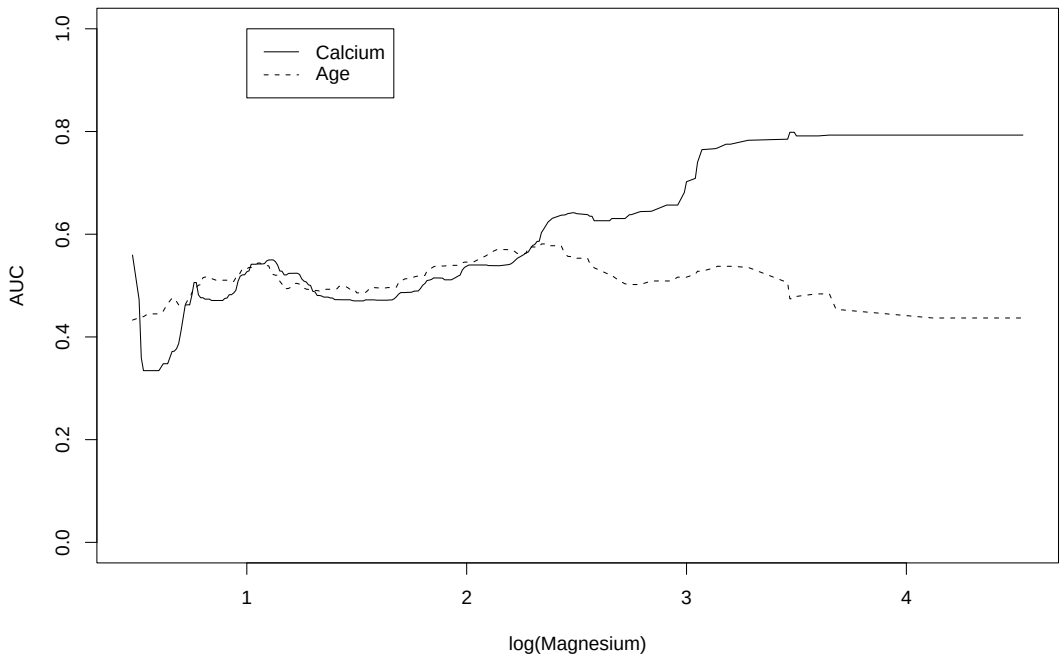

Figure 2.2: AUC process for the prognostic factors calcium concentration and age.

Example 2.9 (ROC **time-dependent)**
In the situation of example 2.5, let T now be a continuous random variable. We apply ROC diagnostics to the sequence of binary variables $D(t) = 1\{T > t\}$, using markers determined by a continuous covariate Z. We suggest to consider a summary measure of the ROC curves, for instance the area under the curve (AUC), as a function of the cutpoint t. This defines a process on the range of T:

$$\{\text{AUC}(t) : t \in \mathbb{R}\}.$$

Pluggin-in nonparametric estimators for the sensitivity function and the specificity function at each cutpoint t, yields an estimate of the AUC-process. This suggests another graphical tool that shows the predictive power of Z. Figure 2.2 shows the estimated AUC-process for the data of example 2.8 where again magnesium concentration is the dependent variable and we consider calcium concentration and age as prognostic factors. In contrast to the prediction error curves estimated in example 2.8, the AUC-process does not depend on the linear model fit; both types of diagnostic plots have in common that they may show the predictive power of a particular covariate as a function of the values of the outcome variable of interest.

□

Chapter 3

Information bounds for information loss models

At the end of the night this gives the influence function.

N.L. Hjort (1993)

This chapter deals with the correspondence of differentiability and the existence of regular estimators in incomplete data models. Some essentials of the theory of identifiability and asymptotical efficiency of regular estimators with values in $\mathbb{R}$ or l^∞ are recalled in section 3.1. We refer to Bickel, Klaassen, Ritov, and Wellner (1993) (BKRW from now on) for a detailed discussion of parameters with values in general Banach spaces.

In section 3.2 we derive conditions under which identifiability and differentiability of functionals is preserved if only a random transformation X of a random variable of interest U is observable. Information loss models are models for the distribution of X. In section 3.2 we consider information loss models that are indexed by the product of a model for the distribution of U and a model for the conditional distributions of X given U. We are therefore in the situation of section 5.5 of BKRW where the authors investigate models with composite parameter spaces. By the familiar embedding of dominated statistical models into $\mathcal{L}_2$-spaces we are also treating a special case of models indexed by a Hilbert space (see Van der Vaart (1991), Begun, Hall, Huang, and Wellner (1983) and BKRW for the general case). Statistical inference for parameters of the distribution function of U based on *iid* observations of a random map X was studied by Le Cam and Yang (1988) and Van der Vaart (1988). Their results include the preservation of differentiability in quadratic mean of differentiable functionals. We recall these results specialized for our aims. The elaborations lead almost automatically to a general characterization of the class of inverse probability of censoring weighting (IPCW) estimators.

In section 3.3 we study a class of information loss models called coarsening at

random (CAR) models. In section 3.4 we compute efficient influence functions and compare information bounds for different CAR submodels in the example of right censored event times with completely observable covariates.

3.1 Estimability and differentiability of functionals

Let $(\Omega, \Gamma, \mathcal{P})$ be a statistical experiment, i.e. a probability space, a σ-field and a set of probability measures. Let $U : (\Omega, \Gamma, \mathcal{P}) \to (E, \mathcal{E})$ be a random map with values in a Borel space: there exists a complete and separable metric space such that E is an element of the corresponding Borel σ-field and $\mathcal{E}$ is the intersection of the Borel σ-field with E. Let μ be a σ-finite measure on $\mathcal{E}$; we define

$$\mathcal{Q} = \{\text{all probability distributions on } \mathcal{E} \text{ that are dominated by } \mu\}.$$

Here and then we use linear functional notation and write $Qg \equiv \int g(u)\, Q(\mathrm{d}u)$ for the expectation of g under Q. For any $Q \in \mathcal{Q}$ we write $\mathcal{L}_2(E, \mathcal{E}, Q) \equiv \mathcal{L}_2(Q)$ for the Hilbert space of $\mathcal{E}$-measurable random variables that have finite variance with respect to Q. On $\mathcal{L}_2(Q)$ we have the usual inner product and norm

$$< \varphi, g >_Q = Q\varphi\, g, \qquad \|\varphi\|_Q^2 = < \varphi, \varphi >_Q .$$

The subspace of $\mathcal{L}_2(Q)$ that consists of mean zero variables is denoted by $\mathcal{L}_2^0(Q)$. We can view $\mathcal{Q}$ as a subset of $\mathcal{L}_2(\mu)$ via the embedding $Q \mapsto (\,\mathrm{d}Q/\,\mathrm{d}\mu)^{1/2}$ and for $Q \in \mathcal{Q}$ fixed as a subspace of $\mathcal{L}_2(Q)$ via

$$Q' \mapsto 2\left(\frac{(\,\mathrm{d}Q'/\,\mathrm{d}\mu)^{1/2}}{(\,\mathrm{d}Q/\,\mathrm{d}\mu)^{1/2}} - 1\right) 1\{(\,\mathrm{d}Q/\,\mathrm{d}\mu)^{1/2} > 0\}.$$

We present the parameters of our interest in a general form such that all the measures of prediction error defined in chapter 2 are included. For any class of functions $\mathcal{F} \subset \mathcal{L}_2(Q)$ a class of real valued linear functionals of the distribution function of U is defined by varying $\varphi \in \mathcal{F}$: $\psi(\varphi) : \mathcal{Q} \to \mathbb{R}$,

$$\psi(\varphi, Q) = Q\varphi. \tag{3.1}$$

Recall that $l^\infty(\mathcal{F})$ is the space of bounded functions with the uniform norm. The corresponding function valued parameter $\psi(\mathcal{F}) : \mathcal{Q} \to l^\infty(\mathcal{F})$ is defined by

$$\psi(\mathcal{F}, Q) = \{\varphi \mapsto Q\varphi\}. \tag{3.2}$$

For instance, φ will be the loss of a forecast conditional distribution π incurred by a single observation. Similarly, prediction error curves are a special case of the functionals defined in (3.2) (c.f. definitions 2.2 and 2.6).

Note that the symbol ψ refers to the process $\{\varphi \mapsto Q\varphi\}$ and is at the same time used for the values $\psi(\varphi, Q)$ of the process. Likewise, in this chapter, we write $\psi(\varphi)$ and $\psi(\mathcal{F})$ for $\psi(\varphi, Q)$, and $\psi(\mathcal{F}, Q)$, in cases where $Q \in \mathcal{Q}$ is fixed. The symbol ψ is used as a wildcard for a parameter on $\mathcal{Q}$ with values in either $\mathbb{R}$ or $l^\infty(\mathcal{F})$.

Let $U_1, \ldots, U_n$ be *iid* random variables with the same distribution as U. An estimator $\hat{\psi}_n$ of ψ is any map from the product space $E^n = E \times \cdots \times E$ to the range of ψ which is assumed to be a subset of either $\mathbb{R}$ or $l^\infty(\mathcal{F})$. Before certain requirements on the performance of estimators can be formulated, the parameter ψ has to be identifiable as a functional on $\mathcal{Q}$. We denote $\mathbf{R}(U)$ for the range of a random variable U; the following definition can be found in chapter 1 of Prakasa Rao (1983).

Definition 3.1 *The parameter ψ is estimable of degree n in $\mathcal{Q}$ if there exists a map $\phi : E^n \to \mathbf{R}(\psi)$ such that for every $Q \in \mathcal{Q}$*

$$\mathrm{E}\,(\phi(U_1, \ldots, U_n)) = \psi(Q).$$

ψ is called identifiable (or asymptotically estimable) in $\mathcal{Q}$ if the preceding display is satisfied in the limit as $n \to \infty$.

The parameters defined in equation (3.1) are always estimable of degree one in $\mathcal{Q}$. Indeed, a necessary and sufficient condition for the existence of an unbiased estimator of a functional of degree one is that it has a representation as an integral operator of the first kind, see Prakasa Rao (1983).

The estimator $\hat{\psi}_n$ is called asymptotically consistent if $(\hat{\psi}_n - \psi(Q)) = o_P(1)$ and asymptotically $\sqrt{n}$-consistent if $(\hat{\psi}_n - \psi(Q)) = O_P(\sqrt{n})$ (in the familiar o_P/O_P-notation). Note that $\psi(Q)$ is identifiable if and only if there exists an asymptotically consistent estimator.

The estimator $\hat{\psi}_n$ is called regular at $Q \in \mathcal{Q}$ if there exists a tight, Borel measurable random element $G \in \mathbf{R}(\psi)$ such that $\sqrt{n}(\hat{\psi}_n - \psi(Q))$ converges weakly to G. If ψ is real valued and G is mean-zero Gaussian, the limit law is readily characterized by the variance of G. In case of $\mathbf{R}(\psi) = l^\infty(\mathcal{F})$ weak convergence is understood in terms of outer expectations to avoid measurability problems (see e.g. Van der Vaart and Wellner (1996) for the corresponding theory). Without loss of generality the law of the limit process of regular estimators is determined by the coordinate projections on $l^\infty(\mathcal{F})$, c.f. the discussion below theorem 5.2.3 of BKRW. Furthermore, if G is a mean-zero Gaussian element of $l^\infty(\mathcal{F})$ the covariance matrix determines the limit law.

Regular consistent estimators can be (locally) approximated by a linear functional: if there exists a function $\tilde{\psi}_0(\varphi) \in \mathcal{L}_2^0(Q)$ such that

$$\sqrt{n}(\hat{\psi}_n(\varphi) - \psi(\varphi, Q)) = \sqrt{n}\,\hat{Q}_n(\tilde{\psi}_0(\varphi)) + o_P(1),$$

where $\hat{Q}_n\tilde{\psi}$ denotes expectation of $\tilde{\psi}$ with respect to the empirical measure corresponding to $U_1, \ldots, U_n$, then $\tilde{\psi}_0(\varphi)$ is called an influence function of the estimator $\hat{\psi}_n(\varphi)$. The latter display is also known as a first order Von Mises (1947) expansion of the functional ψ. The asymptotic covariance of the process $\varphi \mapsto \tilde{\psi}_0(\varphi; U)$ is given by

$$\mathrm{E}\, G(\varphi)\, G(g) = \int \tilde{\psi}_0(\varphi; u)\, \tilde{\psi}_0(g; u)\, Q(\,\mathrm{d}u).$$

Before we define asymptotical efficiency for a sequence of estimators we discuss differentiability of the functional ψ on $\mathcal{Q}$, and introduce a tangent space for $\mathcal{Q}$ and the efficient influence function for estimation of ψ.

Let q denote the Radon-Nikodym derivative of Q with respect to the dominating measure μ. A tangent space $\dot{\mathcal{Q}} \equiv \dot{\mathcal{Q}}_Q$ for $\mathcal{Q}$ at Q is defined relative to a class of submodels as the closure of the set of all score functions. A score function or tangent g is defined as the mean square or Hellinger derivative at q of a (one-dimensional) submodel $\{q_\epsilon : |\epsilon| \leq \epsilon_0\}$ passing through q:

$$\lim_{\epsilon \downarrow 0} \int \left[\frac{1}{\epsilon}(q_\epsilon^{\frac{1}{2}} - q^{\frac{1}{2}}) - \frac{1}{2}\, q^{\frac{1}{2}} \right]^2 g = 0. \tag{3.3}$$

The idea of the tangent space, which we always assume to be a closed linear subspace of $\mathcal{L}_2^0(Q)$, is that all densities in a neighborhood of the underlying density can be well approximated by elements of the tangent space. If the parameter ψ is sufficiently smooth, i.e. differentiable in an appropriate sense along all one-dimensional submodels of $\mathcal{Q}$ that are under consideration, then in a neighborhood of the value $\psi(Q)$, ψ can be approximated by its derivative evaluated at elements of the tangent space. Usually it is sufficient to consider only those paths along which the parameter of interest is differentiable for setting up the tangent space. However, since $\mathcal{Q}$ consists of all dominated distribution functions of U the tangent space of $\mathcal{Q}$ at Q is full (or saturated): $\dot{\mathcal{Q}} = \mathcal{L}_2^0(Q)$. It follows directly from equation (3.3) that tangents are in $\mathcal{L}_2^0(Q)$; to show that also $\dot{\mathcal{Q}} \supseteq \mathcal{L}_2^0(Q)$ one considers usually one-dimensional submodels of the type $\{Q(\,\mathrm{d}u)\,(1 + \epsilon g(u)) : \epsilon \leq \epsilon_0\}$ where g ranges over a dense subset of $\mathcal{L}_2^0(Q)$, e.g. the set of mean-zero bounded or infinitely often differentiable functions.

Definition 3.2 *The functional ψ is differentiable at $Q \in \mathcal{Q}$ relative to a collection of submodels (passing through Q) if there exists a continuous linear map $\dot{\psi} : \dot{\mathcal{Q}} \to \mathbf{R}(\psi)$ such that for every submodel with score function g*

$$\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon}[\psi(q_\epsilon) - \psi(q)] = \dot{\psi}_q(g) \equiv \dot{\psi}(g).$$

The functional ψ is called differentiable on $\mathcal{Q}$ if it is differentiable at every $Q \in \mathcal{Q}$. This notion of pathwise differentiability is precisely Hadamard differentiability of ψ tangentially to the tangent space $\dot{\mathcal{Q}}$, c.f. appendix A5 of BKRW. Note that in

the present chapter, in the notation we suppress the dependence of the derivative of ψ on q. Differentiability of a functional is under additional assumptions concerning weak convergence necessary and sufficient for the existence of efficient estimators in the sense of Definition 3.4 (see Van der Vaart (1991) and theorem 5.2.3 of BKRW).

The functionals defined in (3.1) and (3.2) are differentiable at every $Q \in \mathcal{Q}$; the following result (i) is due to Van der Vaart (1988), and the extension (ii) is given in example 5.3.8 of BKRW.

Proposition 3.3

(i) If φ is uniformly integrable in $\mathcal{Q}$, then $\psi(\varphi) : \mathcal{Q} \to \mathbb{R}$ given in (3.1) *is pathwise differentiable on $\mathcal{Q}$ with derivative $\dot{\psi}(\varphi) : \dot{\mathcal{Q}} \to \mathbb{R}$,*

$$\dot{\psi}(\varphi, g) = \int \{\varphi(u) - Q\varphi\}\, g(u)\, Q(\mathrm{d}u).$$

The efficient influence function for estimation of (3.1) *is given by $\tilde{\psi}(\varphi, U) = \varphi(U) - Q\varphi$.*

(ii) If the class $\mathcal{F}$ has a uniformly square integrable envelope function F, i.e. such that $\sup_Q(QF^2) < \infty$. Then $\psi(\mathcal{F},Q)$ as defined in (3.2) *is pathwise differentiable on $\mathcal{Q}$ with derivative given by $\dot{\psi} : \dot{\mathcal{Q}} \to l^\infty(\mathcal{F})$,*

$$\dot{\psi}(\mathcal{F}, g)(\varphi) = \int \tilde{\psi}(\varphi, u) g(u)\, Q(\mathrm{d}u).$$

The efficient influence function is given by $\{\varphi \mapsto \tilde{\psi}(\varphi) : \varphi \in \mathcal{F}\}$.

□

The following definition of efficiency of regular estimators is justified by general convolution and asymptotic optimality theorems (a complete representation can be found e.g. in section 5.2 of BKRW). These theorems state that the asymptotic distribution of efficient estimators is less dispersed at the true parameter than that of any other regular estimator and, for real valued parameters are part of traditional Hájek-le Cam theory.

Definition 3.4 *The estimator $\hat{\psi}_n$ of ψ is called asymptotically efficient at $Q \in \mathcal{Q}$ if $\sqrt{n}(\hat{\psi}_n - \psi(Q))$ converges weakly to a separable mean-zero Gaussian random element G of $\mathbf{R}(\psi)$, and the covariance matrix of G equals the inverse of the nonparametric information bound, i.e. is given by*

$$\mathrm{E}\, G(\varphi)\, G(g) = \int \tilde{\psi}(\varphi, u)\, \tilde{\psi}(g, u)\, Q(\mathrm{d}u),$$

where $\tilde{\psi}(\varphi)$ is the efficient influence function of the real parameter $\psi(\varphi,Q)$.

By the convolution theorem the Gaussian limit law of any other regular estimator is the law of the convolution of the limit of an efficient estimator with an independent tight Borel measurable element in $\mathbf{R}(\psi)$. The asymptotic optimality theorem shows that the asymptotic normal distribution of efficient estimators is the best one can get with respect to the class of bowl-shaped loss functions. In fact, the Gaussian distribution performs best with respect to these criteria.

The information bound for estimation of the parameter defined in (3.1) is given by

$$I_\psi^{-1}(\varphi) = Q\tilde{\psi}(\varphi) = Q\varphi^2 - Q\varphi Q\varphi.$$

The covariance structure of the process defined in (3.2) is characterized by the so-called inverse information covariance functional: $I_\psi^{-1} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$,

$$I_\psi^{-1}(\varphi, g) =< \tilde{\psi}(\varphi), \tilde{\psi}(g) >_Q = \int \tilde{\psi}(\varphi, u)\tilde{\psi}(g, u)\, Q(\mathrm{d}u).$$

In view of the convolution theorem the information bound is a measure for the lowest variance we can achieve among asymptotically unbiased estimators.

3.2 Differentiability of functionals and information loss

We have gradients and influence functions on two levels.

Van der Vaart (1991)

In the rest of this chapter we study identifiability and differentiability of functionals if only a random transformation of U is observable. Let X be a random element of a Borel space $(\mathcal{S},\Sigma)$ and $\mathcal{R}$ a dominated model for the conditional distribution of X given U: every $R \in \mathcal{R}$ is a stochastic kernel on $(E \times \Sigma)$, such that

- for every $u \in E$, $R(\cdot \mid u)$ is a probability measure on Σ that is dominated by a sigma finite measure η
- for every $A \in \Sigma$, $R(A \mid \cdot)$ is a $\sigma(U)$-measurable function on $(\Omega, \Gamma, \mathcal{P})$.

By the factorization lemma (see e.g. theorem 4.2.8 of Dudley (1989)) any real, $\sigma(U)$-measurable function is equal to a Borel measurable function on $(E, \mathcal{E})$. Abusing notation as usual we write $R(\cdot \mid u)$ for the function that corresponds to $R(\cdot \mid \omega)$. Since E and $\mathcal{S}$ are Borel spaces, the product of any $R \in \mathcal{R}$ with any $Q \in \mathcal{Q}$ determines a σ-finite measure on the product space $(E \times \mathcal{S}, \mathcal{E} \otimes \Sigma)$,

$$m_{Q,R}(\mathrm{d}x,\, \mathrm{d}u) \equiv R(\mathrm{d}x \mid u)\, Q(\mathrm{d}u).$$

Note that $m_{Q,R}(\cdot \times \mathcal{S})$ is equal to Q. (Appendix A2 of Last and Brandt (1995) provides a comprehensive representation of these facts). Thus, varying $Q \in \mathcal{Q}$ and $R \in \mathcal{R}$ gives rise to a model for the joint law of (U, X) and also determines a model for the marginal law of X, which is given by

$$\mathcal{W} \equiv \{W_{Q,R}(\,\mathrm{d}x) = \int R(\,\mathrm{d}x \mid u)\, Q(\mathrm{d}u) : R \in \mathcal{R}, Q \in \mathcal{Q}\}. \tag{3.4}$$

Note that $W_{Q,R}$ is equal to $m_{Q,R}(E \times \cdot)$ on Σ.

The Radon-Nikodym density of $R(\cdot, u)$ with respect to η on $\mathcal{S}$ will be denoted by $r(\cdot \mid u)$: for every $A \in \Sigma$

$$R(A \mid u) = \int_A r(x \mid u)\, \eta(\,\mathrm{d}x).$$

For $Q \in \mathcal{Q}$ fixed recall that $\dot{\mathcal{Q}} = \mathcal{L}_2^0(Q)$ is the tangent space of $\mathcal{Q}$ at Q. A tangent space $\dot{\mathcal{R}}_R \equiv \dot{\mathcal{R}}$ for $\mathcal{R}$ at R can be obtained as follows: for b bounded consider all submodels of $\mathcal{R}$ of the form

$$\{r_\epsilon(\cdot, u) : |\epsilon| \le \epsilon_0\}$$

such that b is the tangent or mean square derivative at $r(\cdot, u)$:

$$\lim_{\epsilon \downarrow 0} \int \left[\frac{1}{\epsilon}(r_\epsilon(\cdot, u)^{\frac{1}{2}} - r(\cdot \mid u)^{\frac{1}{2}}) - \frac{1}{2} b(\cdot, u)\, r(\cdot \mid u)^{\frac{1}{2}}\right]^2 = 0,$$

Since the bounded functions are dense in $\mathcal{L}_2^0(W_{Q,R})$ we have

$$\dot{\mathcal{R}} = \{b \in \mathcal{L}_2^0(m_{Q,R}) : \int b(x, u)\, R(dx \mid u) = 0 \text{ for almost all } u \in E\}.$$

A tangent space $\dot{\mathcal{W}}_{W_{Q,R}} \equiv \dot{\mathcal{W}}$ for $\mathcal{W}$ can now be obtained by using the results of Le Cam and Yang (1988). Consider a class of submodels of the form $\{W_{Q_\epsilon, R_\epsilon} : |\epsilon| \le \epsilon_0\}$ passing through $W_{Q,R}$, such that Q_ϵ has tangent g and R_ϵ has tangent b. Then define the score operator $\dot{l} : \dot{\mathcal{Q}} \times \dot{\mathcal{R}} \to \mathcal{L}_2^0(W_{Q,R})$ by

$$\dot{l}(g, b) = \mathrm{E}\,(g(U) + b(U, X) \mid X).$$

Differentiability in quadratic mean is preserved (see e.g. proposition A.5.5 of BKRW):

$$\lim_{\epsilon \downarrow 0} \int \left[\frac{1}{\epsilon}(\,\mathrm{d}W_{Q_\epsilon, R_\epsilon}^{\frac{1}{2}} - \,\mathrm{d}W_{Q,R}^{\frac{1}{2}}) - \frac{1}{2}\dot{l}(g, b)\,\mathrm{d}W_{Q,R}^{\frac{1}{2}}\right]^2 = 0.$$

Thus, the score operator $\dot{l}$ maps score functions from $\dot{\mathcal{Q}}$, respectively $\dot{\mathcal{R}}$, to score functions in $\dot{\mathcal{W}}$.

We define the parameter of interest in the information loss model $\mathcal{W}$ through the relation

$$\nu(W_{Q,R}) = \psi(Q). \tag{3.5}$$

Parallel to definition 3.1 we have the following

Definition 3.5 *Let ψ be estimable of degree n in $\mathcal{Q}$. The parameter ν defined in* (3.5) *is estimable of degree n in $\mathcal{W}$ if there exists a map $\phi : \mathcal{S}^n \to B$ such that for every $Q \in \mathcal{Q}$ and $R \in \mathcal{R}$*

$$\mathrm{E}\,(\phi(X_1, \ldots, X_n)) = \psi(Q),$$

where the expectation is with respect to $W_{Q,R} \in \mathcal{W}$. ν is called identifiable in $\mathcal{W}$ if the preceding display is valid in the limit.

To find conditions for the existence of asymptotically unbiased estimators based on incomplete observation is more involved. Assume that ψ is estimable of degree one in $\mathcal{Q}$. Estimability, respectively identifiability, with observations of X for all non-constant functionals on $\mathcal{Q}$ is dictated by the dependence structure of X and U. Clearly, if X is stochastically independent of U no reasonable functional is identifiable.

We introduce the indicator variable $\Delta \equiv 1\{X = U\}$ and assume that E is a subset of $\mathcal{S}$ guaranteeing that Δ is well-defined. We use the notations

$$W^{(1)}(\,\mathrm{d}x) = P(X \in \,\mathrm{d}x, \Delta = 1)$$

and

$$W^{(0)}(\,\mathrm{d}x) = P(X \in \,\mathrm{d}x, \Delta = 0)$$

as short-hand for

$$\int h(x) W^{(j)}(\,\mathrm{d}x) = \int h(x) P(X \in \,\mathrm{d}x, \Delta = \delta), \qquad \delta = 0, 1$$

for all bounded Borel functions h. The relation $W_{Q,R}(\,\mathrm{d}x) = W^{(1)}(\,\mathrm{d}x) + W^{(0)}(\,\mathrm{d}x)$ yields the following decomposition of the nonparametric tangent space

$$\mathcal{L}_2^0(W_{Q,R}) = \mathcal{L}_2^0(W^{(1)}) \oplus \mathcal{L}_2^0(W^{(0)});$$

yet every $h \in \mathcal{L}_2^0(W_{Q,R})$ can be split into $h = \Delta h + (1 - \Delta)h$.

For identifiability of general parameters only the complete observations, where $X = \{U\}$, are relevant. In addition, positive point mass of the kernel R at almost every $u \in E$ given $U = u$ is required as can be seen from the following lines:

$$\begin{aligned} W^{(1)}(\,\mathrm{d}x) &= P(X \in \,\mathrm{d}x, X = U) \\ &= P(U \in \,\mathrm{d}x, X = U) \\ &= P(X = x \mid U = u) P(U \in \,\mathrm{d}u) \\ &= R(x \mid x)\, Q(\,\mathrm{d}x). \end{aligned} \tag{3.6}$$

We define the inverse probability of censoring function by

$$d(x) = R(x \mid x);$$

thus almost surely $W^{(1)}(\mathrm{d}x) = d(x)\,Q(\mathrm{d}x)$. From equation (3.6) we claim that identifiability of a functional $\psi(Q)$ is only possible restricted to the set where the function d is strictly positive. For instance, if there is a set $A \in \mathcal{E}$ with $Q(A) > 0$, such that $d(u) = 0$ for all $u \in A$ then the parameter $\psi(Q) = \int u\,Q(\mathrm{d}u)$ is not identifiable in $\mathcal{W}$.

The following propostion provides sufficient conditions for identifiability of parameters that are estimable of degree one in Q.

Proposition 3.6
Let $\varphi \in \mathcal{L}_2(Q)$ and suppose that also d is identifiable in $\mathcal{W}$. A sufficient condition for identifiability of the parameter $\nu(W_{Q,R}) = \psi(Q) = Q\varphi$ in $\mathcal{W}$ is that

$$\int 1\{d(u) > 0\}\varphi(u)\,Q(\mathrm{d}u) = \int \varphi(u)\,Q(\mathrm{d}u). \tag{3.7}$$

and $\varphi \in \mathcal{L}_1(W^{(1)}/d)$:

$$\int \varphi(u)\frac{W^{(1)}(\mathrm{d}u)}{d(u)} < \infty. \tag{3.8}$$

Proof: Let $\hat{W}_n^{(1)}$ be the empirical measure corresponding to an *iid* sample $X_1, \ldots, X_n$. Since d is identifiable there exists an uniformly (in $\mathcal{Q}$), asymptotically consistent sequence of estimators $\hat{d}_n$. Then as $n \to \infty$, by the law of large numbers

$$\int 1\{\hat{d}_n(u) > 0\}\varphi(u)\frac{\hat{W}_n^{(1)}(\mathrm{d}u)}{\hat{d}_n(u)} \to \int 1\{d(u) > 0\}\varphi(u)\frac{W^{(1)}(\mathrm{d}u)}{d(u)}.$$

If (3.8) holds then the right hand side is finite and equals $Q\varphi$ by (3.6) and (3.7). □

Any estimator of the form

$$\int 1\{\hat{d}_n(u) > 0\}\varphi(u)\frac{W_n^{(1)}(\mathrm{d}u)}{\hat{d}_n(u)},$$

where $\hat{d}_n$ is an estimator of d and $W_n^{(1)}$ is the empirical distribution of the observations with $\Delta = 1$, is called inverse probability of censoring weighting (IPCW) estimator. The name is due to Robins and Rotnitzky (1992), famous examples are the Kaplan-Meier and the Horvitz-Thompson estimator. Note that in order to achieve efficiency of IPCW estimators the observations with $\Delta = 0$ have to be used for the estimator of d.

Example 3.7 (Right censoring)
Suppose U is positive real, an event time, say, and $X = (U \wedge C, \Delta = 1\{U \leq C\})$, where C is a censoring time. The inverse probability of censoring function is in this situation given by

$$d(u) = P(X = U \mid U = u) = P(U \wedge C = u \mid U = u) = P(C > u \mid U = u).$$

If C is independent of U, then $P(C > u \mid U = u) = P(C > u) \equiv G(u)$. Define $\tau = \inf_u\{P(U \wedge C \leq u) = 1\}$ and suppose $\tau < \infty$. Clearly, the parameter $\psi(Q) = \int u\, Q(\mathrm{d}u)$ is not identifiable if $P(U > \tau) > 0$. The survival function of the censoring function G can be consistently estimated on the interval $[0, \tau)$ by the so-called reverse Kaplan-Meier estimator. The question of almost sure convergence of the Kaplan-Meier estimator on $[0, \tau)$ is more delicate, see Stute and Wang (1993) for necessary and sufficient conditions for the case where the distributions of U and C do not jump in common, and Shorack and Wellner (1986, page 306) for the general case. From a practical viewpoint the results of Wang (1987) are important, see also Kosorok (2002) results for the bivariate survival estimator. According to proposition 3.6, so-called Kaplan-Meier integrals of the form $Q\varphi$ can be identified if

$$\int \varphi(u)\, Q(\mathrm{d}u) = \int_0^\tau \varphi(u)\, Q(\mathrm{d}u) = \int_0^\tau \varphi(u) \frac{W^{(1)}(\mathrm{d}u)}{G(u)} < \infty.$$

□

In the following paragraph we discuss the question when the indexed model $\mathcal{W}$ for X is nonparametric, or equivalently when is the tangent space full: $\dot{\mathcal{W}} = \mathcal{L}_2^0(W_{Q,R})$. Let $\dot{l}_1$ be the restriction of $\dot{l}$ on $\dot{\mathcal{Q}}$ and by $\dot{l}_2$ the restriction of $\dot{l}$ on $\dot{\mathcal{R}}$. The adjoint operator $\dot{l}_1^* : \mathbf{R}(\dot{l}_1) \to \dot{\mathcal{Q}}$ of $\dot{l}_1$ is uniquely determined by the equality

$$< \dot{l}_1 g, h >_Q = < g, \dot{l}_1^* h >_{W_{Q,R}}$$

for all functions $g \in \dot{\mathcal{Q}}$ and $h \in \dot{\mathcal{W}}$. It is clear that $\dot{l}_1^*$ is given by $\dot{l}_1^*(b) = \mathrm{E}\,(h(X) \mid U)$; for, if $g \in \dot{\mathcal{Q}}$ and $h \in \dot{\mathcal{W}}$, then

$$\begin{aligned} < \dot{l}_1 g, h >_{W_{Q,R}} &= \mathrm{E}\,\mathrm{E}\,(g(U) \mid X)\, h(X) \\ &= \mathrm{E}\, g(U)\, \mathrm{E}\,(h(X) \mid U) = < g, \dot{l}_1^* h >_Q . \end{aligned}$$

We also compute $\dot{l}_2^*$. For every $b \in \dot{\mathcal{R}}$ and every $h \in \dot{\mathcal{W}}$ we have

$$\begin{aligned} < \dot{l}_2\, b, h >_{W_{Q,R}} &= \mathrm{E}\,\mathrm{E}\,(b(U,X) \mid X)\, h(X) \\ &= \mathrm{E}\,\mathrm{E}\,(b(U,X)\, h(X) \mid X) \\ &= \mathrm{E}\,\mathrm{E}\,(b(U,X)\, \{h(X) - \mathrm{E}\,(h(X) \mid U)\} \mid U) \\ &\quad + \mathrm{E}\,\mathrm{E}\,(b(U,X) \mid U)\, \mathrm{E}\,(h(X) \mid U) \\ &= \mathrm{E}\, b(U,X)\ \{h(X) - \mathrm{E}\,(h(X) \mid U)\} \\ &= < b, \dot{l}_2^* h >_{m_{Q,R}}, \end{aligned}$$

since $\mathrm{E}\,(b(U,X) \mid U) = 0$. Thus, the operator $\dot{l}_2^*$ maps functions that are in $\mathcal{L}_2^0(W_{Q,R})$ onto the orthogonal complement of $\mathcal{L}_2^0(Q)$ in $\mathcal{L}_2^0(m_{Q,R})$. This shows that a necessary condition for a function in $\mathcal{L}_2^0(Q)$ to be in the range of $\dot{l}$ is that it is contained in the range of $\dot{l}_1$. This observation is important for obtaining the

efficient influence functions of parameters of the type $\nu(W_{Q,R}) = \psi(Q)$ in $\mathcal{W}$, see theorem 3.8. To find the efficient influence function one can project an initial influence function onto the orthogonal complement of the tangent space for the nuisance parameter, which is $\mathbf{R}(\dot{l}_2)$ in the present situation.

The following decomposition of the Hilbert space $\mathcal{L}_2^0(W_{Q,R})$ holds simply because $\dot{l}_1$ is a linear operator:

$$\begin{aligned}\mathcal{L}_2^0(W_{Q,R}) &= \{h : \mathrm{E}\,(h(X) \mid U) = 0\} \oplus \{h : \mathrm{E}\,(h(X) \mid U) = 0\}^\perp \\ &= \mathbf{N}(\dot{l}_1^*) \oplus \overline{\mathbf{R}(\dot{l}_1)}. \end{aligned} \tag{3.9}$$

Here $A^\perp$ denotes the orthogonal complement of a set A. If we do not impose restrictions for the conditional distribution of X given U, then obviously $\mathbf{N}(\dot{l}_1^*) \subseteq \mathbf{R}(\dot{l}_2)$ and it follows that the range of $\dot{l}$ is dense in $\mathcal{L}_2^0(W_{Q,R})$. However, in the examples of interest the inverse probability of censoring function is not identifiable unless a restricted model for the conditional distribution of X given U is assumed. This is known as the non-identifiability of a competing risk. A necessary condition for the existence of adaptive estimates of parameters of the distribution of U in presence of the nuisance parameter d is that the range of $\dot{l}_1$ is orthogonal to the range of $\dot{l}_2$ (compare section 5.5 of BKRW). We want to briefly discuss this problem here, and provide a solution in section 3.3. Suppose the ranges are not orthogonal, i.e. there is a function in the intersection: $h_0 \in \mathbf{R}(\dot{l}_1) \cap \mathbf{R}(\dot{l}_2)$. Then there are functions $b \in \dot{\mathcal{R}}$ and $g \in \dot{\mathcal{Q}}$ such that $h_0(x) = \mathrm{E}\,(b(U,X) \mid X)$ and $h_0(x) = \mathrm{E}\,(g(U) \mid X)$. The following equality holds if for all measurable functions b and g,

$$\mathrm{E}\,(b(U,X)\, g(U) \mid X) = \mathrm{E}\,(b(U,X) \mid X)\, \mathrm{E}\,(g(U) \mid X),$$

$$\begin{aligned}\|h_0\|_{W_{Q,R}} &= \mathrm{E}\, h_0(X)\, h_0(X) \\ &= \mathrm{E}\, \mathrm{E}\,(b(U,X)\, g(U) \mid X) \\ &= \mathrm{E}\, g(U)\, \mathrm{E}\,(b(U,X) \mid U) = 0. \end{aligned}$$

This shows that adaptive (IPCW) estimation is possible if (U,X) and U are conditionally independent given X, or if $\dot{\mathcal{R}}$ consists of measurable functions of X only. In section 3.3 we discuss these issues in detail.

For differentiable functionals efficiency of an estimator can be conveniently checked by comparison of the influence function of the estimator with the efficient influence function. The following special case of theorem 3.1 of Van der Vaart (1991) can be used to establish differentiability of identifiable parameters in the model $\mathcal{W}$. The original theorem treats general parameters with values in a Banach space and models indexed by arbitrary Hilbert spaces. Our version is basically corollary 5.5.1 of BKRW, specialized to functionals such as defined in (3.1) and (3.2), and specialized to the information loss model defined in (3.4). The theorem describes a method for finding the efficient influence function in models with composite parameterspaces: an initial influence function has to be projected onto

the orthogonal complement of the tangent space for the nuisance parameter. In the present situation the tangent space for the nuisance parameter is $\mathbf{R}(\dot{l}_2)$, and the (Hilbert space) projection thereon will be denoted by $\Pi_2(\cdot) \equiv \Pi(\cdot \mid \mathbf{R}(\dot{l}_2))$.

Theorem 3.8 (Van der Vaart 1991)
Let $\psi(Q)$ be an identifiable parameter in $\mathcal{Q}$ of the type given in (3.1) or (3.2). In both cases denote by $\dot{\psi}$ the derivative of ψ at $Q \in \mathcal{Q}$ and by $\dot{\psi}^*$ the adjoint map of $\dot{\psi}$.

(i) The functional $\nu(W_{Q,R}) = \psi(Q)$ is differentiable at $W_{Q,R}$ in the sense of definition 3.2 if for every $\varphi \in \mathcal{F}$

$$\mathbf{R}(\dot{\psi}^*) \subset \mathbf{R}(\dot{l}_1^*(1 - \Pi_2)). \tag{3.10}$$

The efficient influence functions $\tilde{\nu}(\varphi)$ for estimation of the real parameters $\nu(\varphi) = Q\varphi$ solve the system of equations

$$\begin{aligned} \tilde{\psi}(\varphi) &= \dot{l}_1^*(1 - \Pi_2)\tilde{\nu}(\varphi) \\ 0 &= \dot{l}_2^*\tilde{\nu}(\varphi). \end{aligned}$$

(ii) Assume that $\mathbf{R}(\dot{l}_1)$ is orthogonal to $\mathbf{R}(\dot{l}_2)$. Then the conclusion of (i) holds with $\dot{l}_1^*(1 - \Pi_2)$ replaced by $\dot{l}_1^*$.

□

It follows that the information bound for estimation of $\nu(W_{Q,R}) = Q\varphi$ is given by

$$I_\nu^{-1} = \mathrm{E}\,\tilde{\nu}(X)^2.$$

Since $\dot{l}_1^*$ is a (semi-) contraction, there exists a number $\theta \leq 1$ such that

$$\|\dot{l}_1^*(\tilde{\nu})\|_{W_{Q,R}} \leq \theta\, \|\tilde{\psi}\|_Q\ .$$

θ reflects the information loss, because $\|\tilde{\psi}\|_Q^2$ and $\|\tilde{\nu}\|_{W_{Q,R}}^2$ are the variances of efficient estimators in $\mathcal{Q}$ respectively in $\mathcal{W}$.

The inverse information covariance functional for estimation of the function valued parameter $\nu(W_{Q,R}) = \{\varphi \mapsto Q\varphi : \varphi \in \mathcal{F}\}$ is given by

$$I_\nu^{-1}(\varphi, g) = <\tilde{\nu}(\varphi), \tilde{\nu}(g)>_{W_{Q,R}} = \int \tilde{\nu}(\varphi, x)\, \tilde{\nu}(g, x) W_{Q,R}(\mathrm{d}x).$$

3.3 Coarsening at random models

Locally, CAR *is everything.*

Gill, Robins, van der Laan (1997)

In this section we consider a class of information loss models that fall under the keyword coarsening at random (CAR). These include grouped, missing and censored data models and were studied by Heitjan and Rubin (1991), Jacobsen and Keiding (1995) and Gill, Van der Laan, and Robins (1995). A coarsened observation is obtained by a random many-to-one mapping of an unobservable random variable U, and therefore can be identified with a set in the σ-field of U. The more values are assigned to the same value by the many-to-one mapping the less information is carried by the coarsened observation, and thus the greater is the observed set. Consequently, a missing value corresponds to case where the observed set is equal to the range of U. The other extreme is where one observes U respectively the singleton $\{U\}$.

Let $U : (\Omega, \Gamma, \mathcal{P}) \to (E, \mathcal{E})$ be an unobservable random map with values in a Borel space and $X : (\Omega, \Gamma, \mathcal{P}) \to (\mathcal{S}, \Sigma)$ a set-valued random map where $\mathcal{S}$ is a Borel subset of $\mathcal{E}$ and Σ the corresponding Borel σ-field. We assume that $\mathcal{E}$ contains all the singletons $\{u : u \in E\}$ and that $U \in X$ is measurable. For every $u \in E$ we define the following set of sets that contain u:

$$\mathcal{A}_u \equiv \{A \in \mathcal{S} : u \in A\}.$$

(To emphasize that X is now set-valued we use the letters $A, B, C \ldots$ for elements of $\mathcal{S}$. Sets in Σ, on the other hand, are denoted by letters $\mathcal{A}, \mathcal{B}, \mathcal{C} \ldots$.)

Define $\mathcal{Q}$ and $\mathcal{R}$ as in the previous section to be the models of all dominated probability distributions of U and all dominated stochastic kernels from E to $\mathcal{S}$, respectively. For any $Q \in \mathcal{Q}$ and any $R \in \mathcal{R}$ recall that $m_{Q,R}(\mathrm{d}u, \mathrm{d}A) = R(\mathrm{d}A \mid u)\, Q(\mathrm{d}u)$ is the induced law on the product space $(E \times \mathcal{S}, \mathcal{E} \otimes \Sigma)$. Since both, E and $\mathcal{S}$, are assumed Borel spaces, there exists also a stochastic kernel $V_{Q,R}$ from $\mathcal{S}$ to E, such that

$$m_{Q,R}(\mathrm{d}u, \mathrm{d}A) = V_{Q,R}(\mathrm{d}u \mid A)\, W_{Q,R}(\mathrm{d}A),$$

where $W_{Q,R}(\mathrm{d}A) = \int R(\mathrm{d}A \mid u)\, Q(\mathrm{d}u)$ is the induced marginal law on $(\mathcal{S}, \Sigma)$. (We refer to the appendix of Last and Brandt (1995) for an overview of these relations.)

Definition 3.9 *Assume that* $1\{U \in X\}$ *is measurable.* X *is called coarsening of* U *(or coarsening variable) if*

$$P(U \in X) = 1. \tag{3.11}$$

Note that (3.11) is equivalent to $P(\{\omega : X(\omega) \in \mathcal{A}_{U(\omega)}\}) = 1$, where $\mathcal{A}_U$ is a random set of sets. If we assume that X is a coarsening variable, then any model for the conditional distribution of X given U is a subset of

$$\mathcal{R}_C \equiv \{R \in \mathcal{R} : R(\,\mathrm{d}A \mid u) = 1\{u \in A\}R(\,\mathrm{d}A \mid u), \text{ for almost every } u\}.$$

We will use this fact repeatedly in the sequel. In contrast to the variable X of the previous section, observations of a coarsening variable X of U are always informative for statistical inference on parameters of the distribution of U, except perhaps for the case $X(\omega) = E$ which corresponds to a missing value. In some instances there may even be enough additional information to recover the value of U from the corresponding observation of X. For example, if the values of X are symmetric sets and it is known that U is the central point. In the interesting cases, however, such information is not available. In view of (3.11), U and X can not be stochastically independent, since observing $X(\omega)$ always tells that $U \in X(\omega)$.

To identify parameters of the distribution of U a model is required for the conditional distribution of U given X. Roughly speaking, assuming coarsening at random (CAR) for a coarsening variable corresponds to a uniform model for the conditional distributions of U given X, at least for the part which is continuous with respect to the marginal distribution of U. The definition of CAR by Gill, Van der Laan, and Robins (1995) is formulated in terms of the conditional distribution of X given U:

Definition 3.10 *A coarsening X of U is called coarsening at random (*CAR*) if for all $u, v \in E$ and every $A \in \mathcal{A}_u \cap \mathcal{A}_v$,*

$$R(dA \mid u) = R(dA \mid v). \tag{3.12}$$

(This is again shorthand for

$$\int h(A)R(dA \mid u) = \int h(A)R(dA \mid v),$$

for all bounded Borel functions h with support on $\{\mathcal{A}_u \cap \mathcal{A}_v\}$.)

To explain the implications of CAR, we consider for a fixed set $A \in \mathcal{S}$ the following Lebesgue decomposition of $V_{Q,R}(\cdot \mid A)$ with respect to Q:

$$V_{Q,R}(B \mid A) = 1\{Q(A) > 0\} \int_B \frac{1\{u \in A\}}{Q(A)}\, Q(\,\mathrm{d}u) + 1\{Q(A) = 0\}\, V^s_{Q,R}(B \mid A), \tag{3.13}$$

for $B \in E$. $V^s_{Q,R}(\cdot \mid A)$ is the singular part of $V_{Q,R}(\cdot \mid A)$ with respect to Q. For singletons $\{u\} \in \mathbf{R}(X)$ always the singular part of decomposition (3.13) is active. Moreover, singletons are the only sets for which it is obvious how to

use the coarsened information: if $X = \{u\}$ then, in view of (3.11), it follows immediately that $U = u$. In some applications of interest, however, there are sets in $\mathbf{R}(X)$ that are not singletons but have Q-probability zero.

For all sets in $\mathbf{R}(X)$ a specification of how to infer on parameters of U is required. The CAR assumption is one such specification and characterized by Radon-Nikodym derivatives of the form given in the continuous part of (3.13). Basically, the interpretation of CAR is that the information of the observation $X(\omega) = A$ is nothing else but the obvious $U \in A$. This is seemingly best characterized by a uniform density on the set A. In the special case where Q is discrete, all sets in $\mathcal{S}$ have positive Q-probability and CAR is be readily characterized by the relation

$$V_{Q,R}(u \mid A) = 1\{u \in A\}\frac{Q(u)}{Q(A)},$$

for every u and A, compare Gill, Van der Laan, and Robins (1995). If Q is continuous and the singular part is active only for singletons $A \in \{\{u\} : u \in E\}$, then (3.13) would readily be good as a definition for CAR. However, as already noted this is not true in general, and a great part of this section is used to find decompositions of the kind given in (3.13) for general CAR models.

Example 3.11 (MAR)
In the notation of this section, a situation where U is either observed or completely missing can be specified by setting $\mathcal{A}_u = \{\{u\}, E\}$ for all $u \in E$. Then the conditional distribution of X given $U = u$ is binary, and thus already specified by the probability $R(E \mid u) = \delta$, where $\delta \in [0, 1]$. If X satisfies CAR, then by (3.12) this probability is the same for all values of U. By taking complements we see that the function $R(\{u\} \mid u) = (1 - \delta)$ is also independent of u.

□

We write $\mathcal{R}_{\mathrm{CAR}} \subseteq \mathcal{R}_C$ for the submodel of all kernels that satisfy (3.11) and (3.12), $\dot{\mathcal{R}}_{\mathrm{CAR}}$ for the tangent space of $\mathcal{R}_{\mathrm{CAR}}$ and define the model for the marginal distribution of a coarsening at random variable by

$$\mathcal{W}_{\mathrm{CAR}} = \{W_{Q,R} : Q \in \mathcal{Q}, R \in \mathcal{R}_{\mathrm{CAR}}\}.$$

It is valuable to analyze what the definition of CAR implies for the conditional distribution of U given X. For this we recall that $\mathcal{R}_{\mathrm{CAR}}$ is assumed dominated by a σ-finite measure η on $\mathcal{S}$. We denote by $r(\cdot \mid u)$ the density that satisfies almost surely for every $\mathcal{A} \in \Sigma$

$$R(\mathcal{A} \mid u) = \int_{\mathcal{A}} r(A \mid u)\, \eta(\mathrm{d}A). \tag{3.14}$$

If $R \in \mathcal{R}_C$, the function $\tilde{r}(A \mid u) \equiv 1\{u \in A\}r(A \mid u)$ also satisfies the preceding display and in that case the integral in (3.14) is limited to $\mathcal{A} \cap \mathcal{A}_u$.

The following lemma relates the CAR assumption to any conditional distribution of U given X.

Lemma 3.12
Assume that $R \in \mathcal{R}_{\mathrm{CAR}}$. Let V be any stochastic kernel on $(\mathcal{E} \times \mathcal{S})$. For every $u \in E$, and every bounded Borel function h

$$\int 1\{V(A \mid A) > 0\} h(A) R(\mathrm{d}A \mid u) =$$
$$\int \left\{ 1\{V(A \mid A) > 0\} h(A) \frac{1\{u \in A\}}{V(A \mid A)} \int 1\{v \in A\} r(A \mid v)\, V(\mathrm{d}v \mid A) \right\} \eta(\mathrm{d}A).$$

Proof: Let h be a function such that $h(A) = 1\{V(A \mid A) > 0\}\, h(A)$ for every $A \in \mathcal{S}$. By (3.11) and since by definition $A \in \mathcal{A}_v$ for every $v \in A$ we have

$$\begin{aligned} &\int h(A) r(A \mid u)\, \eta(\mathrm{d}A) \\ &\quad = \int h(A)\, 1\{u \in A\}\, \frac{V(A \mid A)}{V(A \mid A)}\, r(A \mid u)\, \eta(\mathrm{d}A) \qquad (3.15) \\ &\quad = \int \left\{ h(A)\, \frac{1\{u \in A\}}{V(A \mid A)} \int 1\{v \in A\} r(A \mid v)\, V(\mathrm{d}v \mid A) \right\} \eta(\mathrm{d}A). \end{aligned}$$

□
We emphasize that lemma 3.12 holds for any conditional distribution of U given X, in particular for $V_{Q,R}$ and Q. With the help of lemma 3.12 we can show that if $R \in \mathcal{R}_{\mathrm{CAR}}$ there exists a version of the density function r given in (3.14) which depends on U only through the condition given in (3.11). Note that this relates the set-up of Gill, Van der Laan, and Robins (1995) to the treatment of CAR in section 25.5.3 of Van der Vaart (1998), where CAR is defined such that there exists a version of r that is a measurable function of X.

Lemma 3.13
Suppose $R \in \mathcal{R}_{\mathrm{CAR}}$ and that equation (3.14) is satisfied for a σ-finite measure η and a nonnegative function r. There exists a nonnegative function $\tilde{r} : \mathcal{S} \to \mathbb{R}$ such that for every $\mathcal{A} \in \Sigma$

$$\int_{\mathcal{A}} R(\mathrm{d}A \mid u) = \int_{\mathcal{A}} 1\{u \in A\}\, \tilde{r}(A)\, \eta(\mathrm{d}A).$$

Proof: By (3.11) $V_{Q,R}(A \mid A) = P(U \in A \mid X = A) = 1 > 0$ for almost every $A \in \mathcal{S}$. Using lemma 3.12 with $V = V_{Q,R}$ suggests the function

$$\tilde{r}(A) \equiv \int 1\{v \in A\}\, r(A \mid v)\, V_{Q,R}(\mathrm{d}v \mid A).$$

Substituting $V_{Q,R}$ for V and $1\{\mathcal{A}\}$ for h in Equation (3.15) shows that $\tilde{r}$ is a solution.

□

Example 3.14
Suppose $X = \phi(U, C)$ where C is a censoring variable and ϕ a measurable map. If C is stochastically independent of U, and X is a coarsening of U, then $\phi(U, C)$ satisfies CAR. Since X is a coarsening we have

$$P(\phi(U, C) \in \mathrm{d}A \mid U = u) = 1\{u \in A\}\, P(\phi(u, C) \in \mathrm{d}A),$$

thus $r(A \mid u)$ does not depend on $u \in A$. On the other hand, CAR does not imply stochastic independence in general as can be easily seen by considering $E = \{0, 1\}$ for the range of U and $\mathcal{S} = \{\{0\}, \{1\}, E\}$ for the range of X.

□
It is now possible to show that the tangent space of $\mathcal{W}_{\mathrm{CAR}}$ at $W_{Q,R}$ equals all square integrable mean-zero functions, i.e. is saturated in the language of BKRW, and that the score operators $\dot{l}_1$ and $\dot{l}_2$ (as defined in the previous section) map to orthogonal spaces. The reason is that by lemma 3.13 score function in $\dot{\mathcal{R}}_{\mathrm{CAR}}$ are functions of X only and thus are not projected by the score operator.

Lemma 3.15
Let $\dot{l}_1 : \dot{\mathcal{Q}} \to \dot{\mathcal{W}}_{\mathrm{CAR}}$ and $\dot{l}_2 : \dot{\mathcal{R}}_{\mathrm{CAR}} \to \dot{\mathcal{W}}_{\mathrm{CAR}}$ be the conditional expectation operators given X. Then

$$\dot{\mathcal{W}}_{\mathrm{CAR}} = \mathbf{R}(\dot{l}_1) \oplus \mathbf{R}(\dot{l}_2) = \mathcal{L}_2^0(W_{Q,R}).$$

Proof: By lemma 3.13 every score function $b \in \dot{\mathcal{R}}_{\mathrm{CAR}}$ satisfies

$$b(U, X) = 1\{U \in X\}\, \tilde{b}(X)$$

for some function $\tilde{b}$ of X. By (3.11) almost surely

$$(\dot{l}_2 b)(A) = \mathrm{E}\,(1\{U \in X\} \mid X = A)\, \tilde{b}(A) = \tilde{b}(A).$$

For any $g \in \dot{\mathcal{Q}}$ this yields

$$< \dot{l}_1 g, \dot{l}_2 b >= \mathrm{E}\left\{\mathrm{E}\,(g(U) \mid X)\, \tilde{b}(X)\right\} = \mathrm{E}\left\{g(U)\, \mathrm{E}\,(\tilde{b}(X) \mid U)\right\} = 0,$$

since $1\{u \in X\} = 1$ almost surely and since $b \in \dot{\mathcal{R}}_{\mathrm{CAR}}$

$$\mathrm{E}\,(\tilde{b}(X) \mid U) = \mathrm{E}\,(1\{U \in X\}\tilde{b}(X) \mid U) = \mathrm{E}\,(b(U, X) \mid U) = 0.$$

□
We have seen in the previous section how the complete observations, where $X = \{U\}$, can be used to identify functionals of the type $\nu(W_{Q,R}) = \psi(Q)$. Simply omitting the incomplete observations usually leads to inefficient estimators. The following theorem shows how under CAR the observations $X \neq \{U\}$ can be used for statistical inference on parameters of the distribution of U. It also shows that the continuous part of the Lebesgue decomposition of $V_{Q,R}(\cdot \mid A)$ with respect to Q has the form specified in (3.13).

Theorem 3.16
Assume $R \in \mathcal{R}_{\mathrm{CAR}}$ and suppose that equation (3.14) is satisfied. Then, for every $A \in \mathcal{S}$ with $Q(A) > 0$,

$$\mathrm{E}\,(g(U) \mid X = A) = \mathrm{E}\,(g(U) \mid U \in A). \tag{3.16}$$

Proof: Define $\tilde{r}$ as in lemma 3.13. By lemma 3.13 we have for every bounded Borel function h, that satisfies $h(A) = 1\{Q(A) > 0\}\, h(A)$ for every $A \in \mathcal{S}$, that

$$\begin{aligned}
\int h(A) W_{Q,R}(\,\mathrm{d}A) &= \int h(A) \int 1\{u \in E\}\, r(\,\mathrm{d}A \mid u)\, Q(\,\mathrm{d}u)\, \eta(\,\mathrm{d}A) \\
&= \int h(A) \int 1\{u \in A\}\, r(\,\mathrm{d}A \mid u)\, Q(\,\mathrm{d}u)\, \eta(\,\mathrm{d}A) \\
&= \int h(A) \int 1\{u \in A\}\, \tilde{r}(A) Q(\,\mathrm{d}u)\, \eta(\,\mathrm{d}A).
\end{aligned}$$

Let g be a bounded Borel function of U. Substituting Q for V in lemma 3.12 and by Fubini's theorem we have

$$\begin{aligned}
&\int h(A)\, \mathrm{E}\,(g(U) \mid X = A)\, W_{Q,R}(\,\mathrm{d}A) \\
&\quad = \iint h(A)\, g(u)\, V_{Q,R}(\,\mathrm{d}u \mid A)\, W_{Q,R}(\,\mathrm{d}A) \\
&\quad = \iint h(A)\, g(u)\, R(\,\mathrm{d}A \mid u)\, Q(\,\mathrm{d}u) \\
&\quad = \iint h(A)\, g(u) \frac{1\{u \in A\}}{Q(A)}\, \tilde{r}(A)\, \eta(\,\mathrm{d}A)\, Q(\,\mathrm{d}u) \\
&\quad = \iint h(A)\, g(u) \frac{1\{u \in A\}}{Q(A)}\, W_{Q,R}(\,\mathrm{d}A)\, \eta(\,\mathrm{d}A)\, Q(\,\mathrm{d}u) \\
&\quad = \int h(A) \int g(u) \frac{1\{u \in A\}}{Q(A)}\, Q(\,\mathrm{d}u)\, W_{Q,R}(\,\mathrm{d}A) \\
&\quad = \int h(A)\, \mathrm{E}\,(g(U) \mid U \in A)\, W_{Q,R}(\,\mathrm{d}A).
\end{aligned}$$

□
If $Q(A) > 0$ for all $A \in \mathbf{R}(X)$ that are not singletons then CAR is equivalent to (3.16). For, if $A \in \mathcal{A}_u \cap \mathcal{A}_v$, then $1\{u \in A\} = 1\{v \in A\}$ almost surely and (3.12) is satisfied. For real and continuously distributed U the assumption that singletons are the only Q-probability zero sets in the range of X seems not to be very restrictive: without loss, all problematic sets of Q-probability zero are unions of singletons $\{u_1, \ldots, u_n\} \in \mathcal{A}_{u_i}$ for $i = 1, \ldots, n$ and we can replace each element by a small interval that has positive probability under Q, $\{[u_i - \epsilon, u_i + \epsilon]\}$, say. However, if $X = (T, Z)$ takes values in $\mathbb{R} \times \mathbb{R}$, say, then all sets of the form $A \times \{z\} \in \mathcal{A}_{(t,z)}$ are not singletons and have zero Q-probability at least for continuous Z. This discussion leads to the following corollaries of Theorem 3.16.

Corollary 3.17
Suppose U is a (real) random variable and $Q(A) > 0$ for all $A \in \mathcal{S} \setminus \{\{u\} : u \in E\}$. The coarsening variable X satisfies CAR *if and only if for every $A \in \mathbf{R}(X)$ not a singleton*

$$\mathrm{E}\,(g(X) \mid X = A) = \mathrm{E}\,(g(U) \mid U \in A).$$

□

In the situation of corollary 3.17 the Lebesgue decomposition of $V_{Q,R}$ with respect to Q is given by

$$V_{Q,R}(B \mid A) = \int_B \frac{1\{u \in A\}}{Q(A)}\, Q(\mathrm{d}u) + 1\{A = \{u\}\}\, V^s_{Q,R}(\mathrm{d}u).$$

Corollary 3.18
Suppose $U = (T, Z)$, where T is real and Z is k-dimensional real. Assume that Z is always observed, i.e. $\mathcal{A}_{(t,z)} = \{(A, \{z\}) : t \in A\}$ and $X = (X_T, Z)$ for a coarsening variable X_T of T. If the function $z \mapsto Q((A, \{z\}) \mid Z = z)$ is strictly positive for every $(A, \{z\}) \in \mathcal{S}$, then the coarsening variable X satisfies CAR *if and only if for every $(A, \{z\}) \in \mathcal{S} \setminus \{(\{t\}, \{z\}) : t \in \mathbb{R}\}$*

$$\mathrm{E}\,(g(T, z) \mid X_T = A, Z = z) = \mathrm{E}\,(g(T, z) \mid T \in A, Z = z).$$

□

Corollary 3.18 is used in the next section for the analysis of right censored event times in the presence of completely observed covariates. There are also applications where $U = (T, Z)$ and both variables coarsened, e.g. double censoring or missing covariates with right censored survival data. Suppose Z is one-dimensional and $\mathcal{S}$ is a Borel subset of the product σ-field of U=(T,Z). Then CAR implies for every $A = (A_t, A_z)$

$$\begin{aligned} V_{Q,R}(B \mid A) =& 1\{Q(A) > 0\} \int_B \frac{1\{u \in A\}}{Q(A)}\, Q(\mathrm{d}u) \\ &+ 1\{Q_z(A_t) > 0\} \int_B \frac{1\{s \in A_t\}}{Q_z(A)}\, Q_z(\mathrm{d}s) \\ &+ 1\{Q_t(A_z) > 0\} \int_B \frac{1\{\xi \in A_z\}}{Q_z(A)}\, Q_t(\mathrm{d}\xi) \\ &+ 1\{A = (\{t\}, \{z\})\}\, V^s_{Q,R}(B, A), \end{aligned}$$

where Q_t and Q_z are the conditional distributions of Z given $T = t$ and of T given $Z = z$, respectively, and where we have assumed that $Q_t(A_z) > 0$ and $Q_z(A_t) > 0$. The singular part of the Lebesgue decomposition of $V_{Q,R}$ is then active only for complete observations in both components of U. Information bounds for missing covariates and right censored survival data were recently obtained by Nan (2001), who also corrected the bounds obtained by Robins, Rotnitzky, and

Zhao (1994) for the linear model with missing covariates.

Now we discuss identifiability and differentiability of parameters of the type $\nu(W_{Q,R}) = \psi(Q)$ along the lines of the previous section. Theorem 3.16 and its corollaries shows that under CAR the conditional expectation of functions of U given that $U \in X(\omega)$ can be used for inference on $\psi(Q)$ without introducing bias. If we assume that $Q(A) > 0$ for all $A \in \mathcal{S}$ that are not singletons, then the score operator has the following representation: $\dot{l}_1 : \dot{\mathcal{Q}} \to \dot{\mathcal{W}}_{\mathrm{CAR}}$,

$$\begin{aligned}(\dot{l}_1 g)(X) = \mathrm{E}\,(g(U) \mid X) &= \mathrm{E}\,(\Delta g(U) \mid X) + \mathrm{E}\,((1-\Delta)g(U) \mid U \in X) \\ &= \Delta g(X) + \frac{(1-\Delta)}{Q(X)} \int 1\{u \in X\} g(u) Q(du).\end{aligned}$$

Here we use the indicator of complete observations $\Delta = 1\{X = \{U\}\}$. Note that this formula is similar to the one obtained for the relevant score operator in the random censoring example, see BKRW and Van der Vaart (1991). Under CAR the range of $\dot{l}_2^*$ is a (dense) subset of the orthogonal complement of $\dot{\mathcal{Q}}$ (c.f. lemma 3.15), thus we can use (ii) of theorem 3.8 and establish that the efficient influence function for estimation of functionals such as defined in (3.1) and (3.2) are determined by the system of equations

$$\begin{aligned}\tilde{\psi}(\varphi) &= \dot{l}_1^* \tilde{\nu}(\varphi) \\ 0 &= < \tilde{\nu}(\varphi), h >_{W_{Q,R}} .\end{aligned}$$

for every $h \in \mathbf{R}(\dot{l}_2)$ and $\varphi \in \mathcal{F}$

The rest of this section is used to derive more detailed conditions for differentiability of functionals, where we try to extend the computations for univariate random censoring of BKRW section 6.6. examples 3, respectively Van der Vaart (1991, section 8). First note that by the spectral theorem (c.f. theorem 2.5.2 Davies (1995)) the self-adjoint operator $\dot{l}_1^* \dot{l}_1 : \dot{\mathcal{Q}} \to \dot{\mathcal{Q}}$ is unitarily equivalent to a multiplication operator $\mathcal{D}$, i.e. there exists a bounded function d that characterizes $\mathcal{D}$ through the following relation: for all $g \in \mathcal{L}_2^0(Q)$:

$$(\mathcal{D}g)(u) = d(u)\, g(u). \tag{3.17}$$

Then there exists a unitary operator $\mathcal{R}$, such that

$$\dot{l}_1^* \dot{l}_1 = \mathcal{R}^{-1} \mathcal{D}\, \mathcal{R}.$$

The polar decomposition of the bounded linear operator $\dot{l}_1$ yields existence of the square root $(\dot{l}_1^* \dot{l}_1)^{\frac{1}{2}}$ of $\dot{l}_1^* \dot{l}_1$ and also that $\mathbf{R}(\dot{l}_1^*) = \mathbf{R}(\overline{(\dot{l}_1^* \dot{l}_1)^{\frac{1}{2}}})$ (c.f. Van der Vaart (1991)). Moreover, if $\dot{l}_1^* h = (\dot{l}_1^* \dot{l}_1)^{\frac{1}{2}} g$ for some $h \in \overline{\mathbf{R}(\dot{l}_1)}$ and some $g \in \overline{\mathbf{R}(\dot{l}_1^*)}$, then (c.f. proposition A.1.6 of BKRW) $\| h \|_{W_{Q,R}} = \| g \|_Q$. By the spectral representation we have $\mathbf{R}(\dot{l}_1^*) = \mathbf{R}(\mathcal{R}^{-1} \mathcal{D}^{1/2} \mathcal{R})$. Because of the unitary equivalence

it is necessary and sufficient for a function g to be in $\mathbf{R}(\dot{l}_1^*)$ that $\mathcal{R}g \in \mathcal{L}_2(Q/d)$. This result can be used as follows to establish differentiability of ν. The influence function $\tilde{\psi}(\varphi)$ lies in the range of $\dot{l}_1$ if and only if

$$\int \mathcal{R}(\tilde{\psi}(\varphi))^2(u)\frac{Q(\,\mathrm{d}u)}{d(u)} < \infty.$$

By theorem 3.8 this is a sufficient condition for differentiability of ν. For a particular class of examples which includes random missing and random censoring (c.f. section 3.4) it can be seen that $d(u) = R(\{u\} \mid u) = P(\Delta = 1 \mid U = u)$ is the inverse probability of censoring function. In these examples it is also possible to explicitly determine the univariate operator $\mathcal{R}$ of the spectral representation.

Suppose for the moment that U is real. We introduce the martingal operator $\mathcal{L}_{uc} : \dot{\mathcal{Q}} \to \dot{\mathcal{W}}$ that is an important tool for establishing the referenced form of the spectral decomposition in an explicit way.

$$(\mathcal{L}_{uc}\, g)(x) = \int g\,\mathrm{d}M_{uc} \tag{3.18}$$

where M_{uc} is the martingal for the uncensored observations

$$M_{uc}(u) \equiv 1\{U \le u, \Delta = 1\} - \int_{-\infty}^{u} 1\{U > v\}\,\mathrm{E}\,(\Delta = 1 \mid U = v)\,\frac{Q(\,\mathrm{d}v)}{Q(v,\infty)}.$$

Using counting process theory (Last and Brandt 1995, example 1.6.2) it can be seen that the process $1\{U \le u, \Delta = 1\}$ is adapted to the filtration that is jointly generated by $1\{U \le u\}$, $\Delta 1\{U \le u\}$, and that the corresponding compensator is given as in the definition of M_{uc}. Observe also that M_{uc} is exactly analogous to the martingal of section 6.6 in BKRW for a survival time T in the presence of a right censoring variable C, where $\Delta = 1\{T \wedge C = T\}$: the counting process $1\{T \wedge C \le t, \Delta = 1\} = 1\{T \le t, \Delta = 1\}$ has compensator

$$\int_0^t 1\{T \wedge C > t\}\,\Lambda(\,\mathrm{d}s) = \int_0^t 1\{T > t, \Delta = 1\}\,\Lambda(\,\mathrm{d}s),$$

where Λ is the cumulative hazard function corresponding to the distribution function of T. For e.g. missing at random or (multivariate) right censoring, it can be shown that $(\dot{l}_1^*\dot{l}_1 g)(u) = (\mathcal{R}^*\,\mathcal{L}_{uc}^*\,\mathcal{L}_{uc}\mathcal{R}g)(u)$, where $\mathcal{R}$ is a unitary operator. It follows from martingal calculus and equation (3.6) that $(Dg)(u) = (\mathcal{L}_{uc}^*\,\mathcal{L}_{uc}g)(u) =$

$R(\{u\} \mid u)g(u)$:

$$
\begin{aligned}
< \mathcal{L}_{uc}^{*}\, \mathcal{L}_{uc}\, g, \varphi >_Q &=< \mathcal{L}_{uc}\, g, \mathcal{L}_{uc}\, \varphi >_{W_{Q,R}} \\
&= \mathrm{E} \int g(v)\varphi(v) 1\{U > v, \Delta = 1\} \frac{Q(\,\mathrm{d}v)}{Q(v,\infty)} \\
&= \mathrm{E} \int g(v)\varphi(v) R(\{v\} \mid v) Q(v,\infty) \frac{Q(\,\mathrm{d}v)}{Q(v,\infty)} \qquad (3.19) \\
&= \int g(v)\varphi(v) R(\{v\} \mid v)\, Q(\,\mathrm{d}v) \\
&=< R(\{v\} \mid v)\, g, \varphi >_Q .
\end{aligned}
$$

3.4 Right censored regression with completely observed covariates

Now we can flip the 'R' and the 'L' operators.

Jon Wellner (2001)

In this section we derive explicit formulas for information bounds in the random censoring model with covariates. We use spectral decompositions of the involved score operators to find explicit formulas for the efficient influence functions for estimation of parameters of the form $\nu(W_{Q,R}) = \psi(Q)$ along the lines of the previous section. The formulas are obtained by using modifications of the 'R' and 'L' operators such as defined by Ritov and Wellner (1988) (see also appendix A1 of BKRW), adaptively modified to work in situations with right censoring and covariates. Similar computations can be found in Nan (2001) for estimation of the finite-dimensional parameters in the Cox regression model with covariates subject to missing at random.

Suppose that the components of $U = (T, Z) : (\Omega, \Gamma, \mathcal{P}) \to (\mathbb{R}_+ \times \mathbb{R}^k)$ correspond to the dependent variable in a regression problem, a positive event time, say, and a k-dimensional vector of covariates, respectively. We consider a coarsening variable $X = (X_T, X_Z)$, where $X_T = 1\{T \le C\}\{T\} + 1\{T > C\}[C, \infty)$ for a positive censoring random variable C, and $X_Z = \{Z\}$, reflecting that Z is not coarsened. Examples where more than one component of a random vector are subject to coarsening are bivariate right censoring (Van der Laan 1996) or a right censored event time with missing covariates (Nan 2001). Obviously X is a coarsening of U, since $P(T \in X_T) = 1$. The more familiar way is to code the observed data X by the vector $(Y = T \wedge C, \Delta = 1\{T \le C\}, Z)$ which we shall also call X in what follows. We hope that this does not lead to confusions. Note that Δ is an indicator variable for complete observations in agreement with $\Delta = 1\{X = U\}$

used in the previous sections. Some additional notation is needed. Our model $\mathcal{Q}_1$ for the conditional distribution of T given Z is always a submodel of the set of all probability kernels on $\mathbb{B} \times \mathbb{R}^k$ that are dominated in the first argument. Every element of $\mathcal{Q}_1$ is determined by the conditional cumulative distribution function of T given Z, viz. $F(t \mid z) \equiv P(T \leq t \mid Z = z)$. We will see later, in chapter 4, that for uniformly consistent estimation in case of Z continuous, the elements of $\mathcal{Q}_1$ have to satisfy a certain smoothness condition in the second argument.

The model for the marginal distribution of Z is denoted by $\mathcal{Q}_2$ and consists of all dominated distributions $H(\mathrm{d}z) \equiv P(Z \in \mathrm{d}z)$ on $\mathbb{B}^k$. We assume CAR for the conditional distribution of X given U which was seen in the previous section to be implied by the more familiar assumption that C is conditionally independent of T given Z (c.f. example 3.14). Identifiability of the distribution of T (see e.g. Example 6.6.1 of BKRW) continues to hold under CAR (equations (3.20)-(3.22) below).

It is easy to see (c.f. lemma 3.13) that under CAR the conditional distribution of X given (T,Z) is completely specified by the conditional cumulative distribution function of C given Z, viz. $G(t \mid z) = P(C \leq t \mid Z = z)$. We assume a model $\mathcal{G}$ for the conditional distribution of C given Z which is a submodel of the set of all dominated probability kernels on $\mathbb{B} \times \mathbb{R}^k$.

As in the previous section we can establish a model for the distribution of X, indexed by $\mathcal{Q} = \mathcal{Q}_1 \times \mathcal{Q}_2$ and $\mathcal{G}$. For this we introduce the following conditional subdistribution functions

$$\begin{aligned} W^{(1)}(\mathrm{d}y \mid z) &= P(Y \in \mathrm{d}y, \Delta = 1 \mid Z = z) = (1 - G(y \mid z))\, F(\mathrm{d}y \mid z) \\ W^{(0)}(\mathrm{d}y \mid z) &= P(Y \in \mathrm{d}y, \Delta = 0 \mid Z = z) = (1 - F(y \mid z))\, G(\mathrm{d}y \mid z), \end{aligned} \tag{3.20}$$

where the last step follows by CAR. Then any distribution W of X can be decomposed as

$$\begin{aligned} W(\mathrm{d}y, \delta, \mathrm{d}z) &= W(\mathrm{d}y, \delta \mid z)\, H(\mathrm{d}z) \\ &= \{W^{(1)}(\mathrm{d}y \mid z)\, H(\mathrm{d}z)\}^{\delta} + \{W^{(0)}(\mathrm{d}y \mid z)\, H(\mathrm{d}z)\}^{(1-\delta)}. \end{aligned}$$

This shows that W is uniquely determined by F_z, G_z and H, and hence we have a model $\mathcal{W}$ for W that is indexed by $\mathcal{Q}_1 \times \mathcal{Q}_2 \times \mathcal{G}$:

$$\mathcal{W}_{\mathrm{CAR}} \equiv \{W_{F_z,H,G_z} : F_z \in \mathcal{Q}_1, H \in \mathcal{Q}_2, G_z \in \mathcal{G}\}.$$

The conditional distributions of T and C given Z can be identified by the well-known product integral formulas: if for almost every z $t \mapsto F(t \mid z)$ and $t \mapsto G(t \mid z)$ are continuous functions, which we assume for simplicity, then

$$(1 - F(t \mid z)) = \exp\left\{ - \int_0^t \frac{W^{(1)}(\mathrm{d}s \mid z)}{(1 - W(s \mid z))} \right\} \tag{3.21}$$

and

$$(1 - G(t \mid z)) = \exp\left\{ - \int_0^t \frac{W^{(0)}(\,\mathrm{d}s \mid z)}{(1 - W(s \mid z))} \right\}. \tag{3.22}$$

Tangent spaces corresponding to $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{G}$ are respectively given by

$$\begin{aligned} \dot{\mathcal{Q}}_1 &= \{a \in \mathcal{L}_2^0(F \times H) : \mathrm{E}\,(a(T,Z) \mid Z) = 0\} \\ \dot{\mathcal{Q}}_2 &= \mathcal{L}_2^0(H) \\ \dot{\mathcal{G}} &= \{c \in \mathcal{L}_2^0(G \times H) : \mathrm{E}\,(c(C,Z) \mid Z) = 0\}. \end{aligned}$$

The relevant score operators for estimation of parameters of the joint distribution of (T, Z) (c.f. section 3.2) are denoted by $\dot{l}_{11}, \dot{l}_{12}, \dot{l}_2$ and defined on $\dot{\mathcal{Q}}_1, \dot{\mathcal{Q}}_2, \dot{\mathcal{G}}$, respectively by:

$$\begin{aligned} (\dot{l}_{11}\, a)(y,\delta,z) &= \mathrm{E}\,(a(T,Z) \mid Y > y, \Delta = \delta, Z = z) \\ &= \delta\, a(y,z) + \frac{(1-\delta)}{(1 - F(y \mid z))} \int_y^\infty a(s,z)\, F(\,\mathrm{d}s \mid z) \\ (\dot{l}_{12}\, b)(y,\delta,z) &= \mathrm{E}\,(b(Z) \mid Y = y, \Delta = \delta, Z = z) \\ &= b(z) \\ (\dot{l}_2\, c)(y,\delta,z) &= \mathrm{E}\,(c(C,Z) \mid Y > y, \Delta = \delta, Z = z) \\ &= (1-\delta)\, c(y,z) + \frac{\delta}{(1 - G(y \mid z))} \int_y^\infty c(s,z)\, G(\,\mathrm{d}s \mid z). \end{aligned}$$

As a consequence of lemma 3.15 we find that the tangent space of $\mathcal{W}_{\mathrm{CAR}}$ at W is full:

$$\dot{\mathcal{W}}_{\mathrm{CAR}} = \mathbf{R}(\dot{l}_{11}) \oplus \mathbf{R}(\dot{l}_{12}) \oplus \mathbf{R}(\dot{l}_2) = \mathcal{L}_2^0(W).$$

Furthermore, we denote

$$\begin{aligned} \dot{\mathcal{W}}_1 &= \mathcal{L}_2^0(W_z) = \mathcal{L}_2^0(W_z^{(1)}) \oplus \mathcal{L}_2^0(W_z^{(2)}) \\ \dot{\mathcal{W}}_2 &= \mathcal{L}_2^0(H), \end{aligned}$$

where $\mathcal{L}_2^0(W_z^{(j)}) = \{h : \mathrm{E}\,(h(Y,j,Z) \mid Z) = 0,\ j = 1,2\}$; giving $\dot{\mathcal{W}}_{\mathrm{CAR}} = \dot{\mathcal{W}}_1 \times \dot{\mathcal{W}}_2$.

The following functionals are of particular interest; see chapter 2 for examples. For $\varphi \in \mathcal{F} \subseteq \mathcal{L}_2(F_z \times H)$ we define $\psi(\varphi, F_z \times H)$ now as a parameter of the information loss model $\mathcal{W}$: $\nu(\varphi) : \mathcal{W} \to \mathbb{R}$,

$$\nu(\varphi, W_{F_z,H,G_z}) = \psi(\varphi, F \times H) = \int \varphi(t,z) F(\,\mathrm{d}t \mid z) H(\,\mathrm{d}z). \tag{3.23}$$

The corresponding function valued parameter is given by $\nu(\mathcal{F}) : \mathcal{W} \to l^\infty(\mathcal{F})$

$$\nu(\mathcal{F}, W_{F,G,H})(\varphi) = \psi(\varphi, F \times H). \tag{3.24}$$

Our calculations of the efficient influence function for estimation of ν in $\mathcal{W}$ base on the following corollary of theorem 3.8. In view of proposition 3.3, it clearly applies to the functionals defined in (3.23) and (3.24).

Corollary 3.19
Suppose ψ is differentiable on $\mathcal{Q} = \mathcal{Q}_1 \times \mathcal{Q}_2$ with efficient influence function $\tilde{\psi} = \tilde{\psi}_{11} + \tilde{\psi}_{12}$, where

$$\tilde{\psi}_{11}(t,z) \equiv \tilde{\psi}(t,z) - \mathrm{E}\,(\tilde{\psi} \mid Z = z) \in \dot{\mathcal{Q}}_1$$

and

$$\tilde{\psi}_{12}(z) \equiv \mathrm{E}\,(\tilde{\psi} \mid Z = z) \in \dot{\mathcal{Q}}_2.$$

The parameter ν defined by $\nu(W_{F_z,H,G_z}) = \psi(F \times H)$ is differentiable at W in $\mathcal{W}$, if and only if there exist a function $\tilde{\nu} \in \dot{\mathcal{W}}$ that satisfies the following system of equations:

$$\dot{l}^*_{11}(\tilde{\nu}) = \tilde{\psi}_{11}, \qquad \dot{l}^*_{12}(\tilde{\nu}) = \tilde{\psi}_{12} \qquad \textit{and} \qquad \dot{l}^*_2(\tilde{\nu}) = 0.$$

□

The 'R' operator corresponding to the conditional distribution of T given Z is defined by $\mathcal{R}_{F_z} : \dot{\mathcal{Q}}_1 \to \dot{\mathcal{Q}}_1$,

$$\begin{aligned}(\mathcal{R}_{F_z}\, a)(t,z) &= a(t,z) - \mathrm{E}\,(a(T,z) \mid T > t, Z = z) \\ &= a(t,z) - \frac{1}{(1 - F(t \mid z))} \int_t^\infty a(s,z)\, F(\,\mathrm{d}s \mid z).\end{aligned}$$

It was shown by Ritov and Wellner (1988) that $\mathcal{R}_{F_z}$ is a unitary, bounded operator whose inverse $\mathcal{R}^{-1}_{F_z}$ equals the adjoint operator $\mathcal{R}^*_{F_z}$ on $\dot{\mathcal{Q}}_1$:

$$(\mathcal{R}^*_{F_z} a)(t,z) = a(t,z) - \int_0^t a(s,z)\, \frac{F(\,\mathrm{d}s \mid z)}{(1 - F(s \mid z))}.$$

Correspondingly, let $\mathcal{R}^{-1}_{G_z}$ and $\mathcal{R}^*_{G_z}$ denote the inverse and adjoint operators of $\mathcal{R}_{G_z}$, then analogous relations holds. The 'L' operator corresponding to the conditional distribution of T given $Z = z$ is defined by $\mathcal{L}_{F_z} : \dot{\mathcal{Q}}_1 \to \dot{\mathcal{W}}$,

$$(\mathcal{L}_{F_z}\, a)(Y,\Delta,z) = \Delta\, a(Y,z) + \int_Y^\infty a(s,z) \frac{F(\,\mathrm{d}s \mid z)}{(1 - F(s \mid z))}$$

and similarly we shall define $\mathcal{L}_{G_z} : \dot{\mathcal{G}} \to \dot{\mathcal{W}}$ by

$$(\mathcal{L}_{G_z}\, c)(Y,\Delta,z) = (1 - \Delta)\, c(Y,z) + \int_Y^\infty c(s,z) \frac{G(\,\mathrm{d}s \mid z)}{(1 - G(s \mid z))}.$$

Without right censoring, i.e. if $C \equiv \infty$, the 'L' and the 'R' operators are inverses and adjoints of each other on the set of mean zero square integrable functions (Ritov and Wellner 1988). However, these relations do not maintain to hold for censored data. We show that $\dot{l}_{11} = \mathcal{L}_{F_z} \mathcal{R}_{F_z}$ and $\dot{l}_2 = \mathcal{L}_{G_z} \mathcal{R}_{G_z}$. Let $a \in \dot{\mathcal{Q}}_1$, then

$$\begin{aligned}(\mathcal{L}_{F_z} \mathcal{R}_{F_z} a)(y, \delta, z) = \delta\, a(y,z) & \\ - \frac{\delta}{(1 - F(y \mid z))} \int_y^\infty a(s,z)\, F(\mathrm{d}s \mid z) & \\ - \int_0^y a(s,z)\, \frac{F(\mathrm{d}s \mid z)}{(1 - F(y \mid z))} & \\ + \int_0^y \int_s^\infty a(u,z)\, F(\mathrm{d}s \mid z)\, \frac{F(\mathrm{d}s \mid z)}{(1 - F(y \mid z))^2}. &\end{aligned} \tag{3.25}$$

Fubini's theorem and the fact that the relation

$$\int_a^b \frac{F(\mathrm{d}s)}{\{1 - F(s)\}^2} = \frac{1}{\{1 - F(b)\}} - \frac{1}{\{1 - F(a)\}}$$

holds for every distribution function F and any $a \leq b$ can be used to show that

$$\begin{aligned}\int_0^y \int_s^\infty a(u,z)\, F(\mathrm{d}s \mid z)\, \frac{F(\mathrm{d}s \mid z)}{(1 - F(y \mid z))^2} & \\ = \int_0^y \left[\frac{1}{(1 - F(u \mid z))} - 1\right] a(u,z)\, F(\mathrm{d}u \mid z) & \\ + \left[\frac{1}{(1 - F(y \mid z))} - 1\right] \int_y^\infty a(u,z)\, F(\mathrm{d}u \mid z) & \\ = \int_0^y a(s,z) \frac{F(\mathrm{d}s \mid z)}{(1 - F(s \mid z))} & \\ + \frac{1}{(1 - F(y \mid z))} \int_y^\infty a(s,z)\, F(\mathrm{d}s \mid z). &\end{aligned}$$

Substituting the last expression for the last term on the right hand side of (3.25) yields $\dot{l}_{11} = \mathcal{L}_{F_z} \mathcal{R}_{F_z}$. A similar computation shows that $\dot{l}_2 = \mathcal{L}_{G_z} \mathcal{R}_{G_z}$. We can

also compute $\mathcal{L}^*_{F_z}$ directly: for every $h \in \dot{\mathcal{W}}_1$ and $a \in \dot{\mathcal{Q}}_1$,

$$
\begin{aligned}
< \mathcal{L}_{F_z} a, h >_W &= \sum_\delta \int \left\{ \delta a(y,z) - \int_0^y a(s,z) \frac{W^{(1)}(\,\mathrm{d}s \mid z)}{(1 - W(s \mid z))} \right\} h(y,\delta,z)\, W(\,\mathrm{d}y, \delta \mid z) \\
&= \int h(y,1,z)\, a(y,z)\, W^{(1)}(\,\mathrm{d}y \mid z) \\
&\quad - \int \frac{a(y,z)}{(1 - W(y \mid z))} \sum_\delta \int_y^\infty h(s,\delta,z)\, W(\,\mathrm{d}s, \delta \mid z)\, W^{(1)}(\,\mathrm{d}y \mid z) \\
&= \int \left\{ h(y,1,z) - \frac{1}{(1 - W(y \mid z))} \sum_\delta \int_y^\infty h(s,\delta,z)\, W(\,\mathrm{d}s \mid z) \right\} \\
&\quad \times\ a(y,z)\, W^{(1)}(\,\mathrm{d}y \mid z) \\
&\equiv \int \{(\mathcal{R}_1 h)(y,\delta,z)\}\ a(y,z)\, W^{(1)}(\,\mathrm{d}y \mid z) \\
&= \int \left\{ (1 - G(y \mid z))\, h(y,1,z) - \frac{1}{(1 - F(y \mid z))} \sum_\delta \int_y^\infty h(s,\delta,z)\, W(\,\mathrm{d}s \mid z) \right\} \\
&\quad \times\ a(y,z) F(\,\mathrm{d}y \mid z) \\
&= \quad < a, \mathcal{L}^*_{F_z} h >_F,
\end{aligned}
$$

where we have implicitly defined the operator $\mathcal{R}_1 : \dot{\mathcal{W}} \to \mathcal{L}_2^0(W^{(1)})$. $\mathcal{R}_1$ satisfies for every h and almost every $z \in \mathbf{R}(Z)$ the equation

$$
\begin{aligned}
(\mathcal{L}^*_{F_z} h)(y,z) &= (1 - G(y \mid z))\, (\mathcal{R}_1 h)(y,z) \\
&= (1 - G(y \mid z))\, \{h(y,1,z) - \mathrm{E}\,(h(Y,\Delta,Z) \mid Y > y, Z = z)\}\,.
\end{aligned}
$$

Thus,

$$
(\mathcal{R}_1 h)(y,z) = 1\{(1 - G(y \mid z)) > 0\}\, \{h(y,1,z) - \mathrm{E}\,(h(Y,\Delta,Z) \mid Y > y, Z = z)\}\,.
$$

Analogously, we define $\mathcal{R}_2 : \dot{\mathcal{W}} \to \mathcal{L}_2^0(W^{(0)})$ by

$$
(\mathcal{R}_2 h)(y,z) = 1\{(1 - F(y \mid z)) > 0\}\, \{h(y,0,z) - \mathrm{E}\,(h(Y,\Delta,Z) \mid Y > y, Z = z)\}\,.
$$

The operator $\mathcal{L}^*_{G_z} : \dot{\mathcal{W}} \to \dot{\mathcal{G}}$ is determined by the almost sure identity $\mathcal{L}^*_{G_z} = (1 - F_z)\, \mathcal{R}_2$.

We have established the following representations for the (adjoint) score operators of corollary 3.19:

$$
\begin{aligned}
\dot{l}^*_{11} &= (\mathcal{L}_{F_z} \mathcal{R}_{F_z})^* = \mathcal{R}^*_{F_z} \mathcal{L}^*_{F_z} = \mathcal{R}^*_{F_z} (1 - G_z) \mathcal{R}_1 \\
\dot{l}^*_{12} &= \mathrm{E}\,(\cdot \mid Z) \\
\dot{l}^*_{2} &= (\mathcal{L}_{G_z} \mathcal{R}_{G_z})^* = \mathcal{R}^*_{G_z} \mathcal{L}^*_{G_z} = \mathcal{R}^*_{G_z} (1 - F_z) \mathcal{R}_2.
\end{aligned}
$$

Note that these equations represent the spectral decomposition of the information operators. We are now able to compute the efficient influence function for parameters of the form $\nu(W_{F,G,H}) = \psi(F \times H)$. In view of corollary 3.19 we have to find a function $h \in \mathcal{L}_2^0(W)$ that satisfies the following three equations (3.26)-(3.28). The first equation is equivalent to $\dot{l}_{11}^*(h) = \tilde{\psi}_{11}$ (because $\mathcal{R}_{F_z}^* = \mathcal{R}_{F_z}^{-1}$) and holds restricted to the set $\{t : (1 - G_z(t)) > 0\}$:

$$\begin{aligned}(\mathcal{R}_1 h)(y,z) &= \frac{1}{(1-G(y \mid z))}(\mathcal{R}_{F_z}\tilde{\psi}_{11})(y,z) \\ &= \frac{1}{(1-G(y \mid z))}\Big\{\tilde{\psi}(y,z) - \mathrm{E}\,(\tilde{\psi}(y,z) \mid Z = z)\Big\} \\ &\quad - \frac{1}{(1-W(y \mid z))}\int_y^\infty \Big\{\tilde{\psi}(y,z) - \mathrm{E}\,(\tilde{\psi}(y,z) \mid Z = z)\Big\}\, F(\,\mathrm{d}s \mid z), \\ &= \frac{\tilde{\psi}(y,z)}{(1-G(y \mid z))} - \frac{1}{(1-W(y \mid z))}\int_y^\infty \tilde{\psi}(y,z)\, F(\,\mathrm{d}s \mid z).\end{aligned} \tag{3.26}$$

The second equation is equivalent to $\dot{l}_{12}^*(h) = \tilde{\psi}_{12}$:

$$\mathrm{E}\,(h(Y,Z) \mid Z = z) = \mathrm{E}\,(\tilde{\psi}(Y,Z) \mid Z = z). \tag{3.27}$$

And finally the third equation is equivalent to $\dot{l}_2^*(h) = 0$ (because $\mathcal{R}_{G_z}^* = \mathcal{R}_{G_z}^{-1}$) and holds restricted to the set $\{t : (1 - F(t)) > 0\}$:

$$(\mathcal{R}_2 h)(y,z) = \frac{1}{(1-F(y \mid z))}\mathcal{R}_{G_z}(0) = 0. \tag{3.28}$$

A solution to this system of equations represents the efficient influence function for estimation of $\nu = \psi$ in the information loss model. If $\psi(F \times H)$ is differentiable with efficient influence function $\tilde{\psi}$, then we obtain the efficient influence function of ν as $\tilde{\nu} : \mathbb{R}^+ \times \{0,1\} \times \mathbb{R}^k \to \mathbb{R}$:

$$\begin{aligned}\tilde{\nu}(y,\delta,z) = \delta\, &\frac{\tilde{\psi}(y,z)}{(1-G(y \mid z))} \\ &+ \frac{(1-\delta)}{(1-W(y \mid z))}\int_y^\infty \tilde{\psi}(s,z)\, F(\,\mathrm{d}s \mid z) \\ &- \int C_2(y \wedge s \mid z)\, \tilde{\psi}(s,z)\, F(\,\mathrm{d}s \mid z),\end{aligned} \tag{3.29}$$

where

$$C_2(y \mid z) = \int_0^y \frac{W^{(0)}(\,\mathrm{d}s \mid z)}{(1-W(s \mid z))^2} = \int_0^y \frac{1}{(1-W(s \mid z))}\frac{G(\,\mathrm{d}s \mid z)}{(1-G(s \mid z))} \tag{3.30}$$

is the well-known asymptotic variance of the Nelson-Aalen estimator for the conditional cumulative hazard function of the censoring variable.

Example 3.20 (Marginal survival function)
Let $\nu(W_{F_z,H,G_z}) = \psi(F \times H)(t) = \int \int_0^t F(\,\mathrm{d}s \mid z) H(\,\mathrm{d}z) = F(t)$, then it is well-known that the efficient influence function in $\mathcal{Q}$ is given by $\tilde{\psi}(t) = 1\{T \leq t\} - F(t)$. Substituting into (3.29) yields, after several applications of integration by part, the efficient influence function of the Kaplan-Meier estimator:

$$\tilde{\nu}(y, \delta, z) = (1 - F(t)) \left\{ \delta \frac{1\{s \leq t\}}{(1 - W(s))} - \int_0^{t \wedge s} \frac{W^{(1)}(\,\mathrm{d}s)}{(1 - W(s))^2} \right\}. \tag{3.31}$$

Here $(1 - W(t)) = \int\int_0^t W(\,\mathrm{d}s \mid z)\, H(\,\mathrm{d}z)$ and $W^{(1)}$ is the marginal distribution of ΔY. The information bound for estimation of the (marginal) survival function at a fixed time is given by

$$\mathrm{E}\left\{\tilde{\nu}(Y, \Delta, Z)^2\right\} = (1 - F(t))^2 \int_0^{t \wedge s} \frac{W^{(1)}(\,\mathrm{d}s)}{(1 - W(s))^2}.$$

□

We shall provide an alternative path for establishing the system of equations (3.26)-(3.28). The starting point is the following decomposition of a function $h \in \dot{\mathcal{W}}$ (compare Nan (2001)):

$$h(Y, \Delta, Z) = (\mathcal{L}_{F_z} \mathcal{R}_1 h)(Y, \Delta, Z) + \mathrm{E}\,(h \mid Z) + (\mathcal{L}_{G_z} \mathcal{R}_2 h)(Y, \Delta, Z).$$

We can recover (3.26)-(3.28) by 'flipping' operators and by noting the relations $\mathcal{L}^*_{F_z} \mathcal{L}_{F_z} = (1 - G_z)$ and $\mathcal{L}^*_{G_z} \mathcal{L}_{G_z} = (1 - F_z)$, which can be obtained as a special case of (3.19). Let $(a, b, c) \in \dot{\mathcal{Q}}_1 \times \dot{\mathcal{Q}}_2 \times \mathcal{G}$ and let $h \in \dot{\mathcal{W}}$, then

$$\begin{aligned}
&< h, \dot{l}(a, c, b) >_W \\
&=< \mathcal{L}_{F_z} \mathcal{R}_1 h + \mathrm{E}\,(h \mid Z) + \mathcal{L}_{G_z} \mathcal{R}_2 h, \mathcal{L}_{F_z} \mathcal{R}_{F_z} a + b + \mathcal{L}_{G_z} \mathcal{R}_{G_z} c >_W \\
&=< \mathcal{L}_{F_z} \mathcal{R}_1 h, \mathcal{L}_{F_z} \mathcal{R}_{F_z} a >_W + < \mathrm{E}\,(h \mid Z), b >_W + < \mathcal{L}_{G_z} \mathcal{R}_2 h, \mathcal{L}_{G_z} \mathcal{R}_{G_z} c >_W \\
&=< \mathcal{L}^*_{F_z} \mathcal{L}_{F_z} \mathcal{R}_1 h, \mathcal{R}_{F_z} a >_{F_z} + < \mathrm{E}\,(h \mid Z), b >_H + < \mathcal{L}^*_{G_z} \mathcal{L}_{G_z} \mathcal{R}_2 h, \mathcal{R}_{G_z} c >_{G_z} \\
&=< (1 - G_z) \mathcal{R}_1 h, \mathcal{R}_{F_z} a >_{F_z} + < \mathrm{E}\,(h \mid Z), b >_H + < (1 - F_z) \mathcal{R}_2 h, \mathcal{R}_{G_z} c >_{G_z} \\
&=< \mathcal{R}^*_{F_z} (1 - G_z) \mathcal{R}_1 h, a >_{F_z} + < \mathrm{E}\,(h \mid Z), b >_H + < \mathcal{R}^*_{G_z} (1 - F_z) \mathcal{R}_2 h, c >_{G_z}.
\end{aligned}$$

Since the last preceding display holds for every choice $(a, b, c) \in \mathcal{Q}_1 \times \mathcal{Q}_2 \times \mathcal{G}$ we can (in view of corollary 3.19) substitute $\tilde{\psi}_{11}$ for a, $\tilde{\psi}_{12}$ for b and 0 for c which then yields (3.26)-(3.28).

Next we show that the function defined in (3.29) solves (3.26)-(3.28). By

Fubini's theorem we have

$$\begin{aligned}
\mathrm{E}\,(\tilde\nu(Y,\Delta,Z)\mid Z=z) &= \int \tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)\\
&\quad+\int\frac{1}{(1-W(y\mid z))}\int_y^\infty \tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)\,W^{(0)}(\,\mathrm{d}y\mid z)\\
&\quad-\int\int_0^s\frac{W^{(0)}(\,\mathrm{d}u\mid z)}{(1-W(u\mid z))}\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)\\
&=\int\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)\\
&=\mathrm{E}\,(\tilde\psi(T,Z)\mid Z=z);
\end{aligned}$$

this shows (3.27). The conditional expectations given $Y>y$ and $Z=z$ applied to each of the three terms of the right hand side of (3.29) can be represented by the following system of equations:

$$\mathrm{E}\left\{\frac{\Delta\,\tilde\psi(Y,Z)}{(1-G(Y\mid Z))}\mid Y>y,Z=z\right\}=\frac{1}{(1-W(y\mid z))}\int_y^\infty\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z),$$

$$\begin{aligned}
&\mathrm{E}\left\{\frac{(1-\Delta)}{(1-W(Y\mid Z))}\int_Y^\infty\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)\mid Y>y,Z=z\right\}\\
&\qquad=\frac{1}{(1-W(y\mid z))}\int_y^\infty\int_s^\infty\tilde\psi(u,z)\,F(\,\mathrm{d}u\mid z)\frac{W^{(0)}(\,\mathrm{d}s\mid z)}{(1-W(s\mid z))}\\
&\qquad=\frac{1}{(1-W(y\mid z))}\int_y^\infty\int_y^u\frac{W^{(0)}(\,\mathrm{d}s\mid z)}{(1-W(s\mid z))}\,\tilde\psi(u,z)\,F(\,\mathrm{d}u\mid z)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}\left(\int C_2(Y\wedge s\mid z)\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)\mid Y>y,Z=z\right)\\
&\qquad=\frac{1}{(1-W(y\mid z))}\int_y^\infty\int\tilde\psi(u,z)\,F(\,\mathrm{d}u\mid z)\,W(\,\mathrm{d}s\mid z)\\
&\qquad=\int C_2(y\wedge s\mid z)\,\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)\\
&\qquad\quad+\frac{1}{(1-W(y\mid z))}\int_y^\infty\int_y^s\frac{W^{(0)}(\,\mathrm{d}u\mid z)}{(1-W(u\mid z))}\,\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z).
\end{aligned}$$

Substituting these formulas into the definition of $\mathcal{R}_1$ yields (3.26):

$$\begin{aligned}
(\mathcal{R}_1\tilde\nu)(y,z)&=\tilde\nu(y,1,z)-\mathrm{E}\,(\tilde\nu(Y,\Delta,Z)\mid Y>y,Z=z)\\
&=\frac{\tilde\psi(y,z)}{(1-G(y\mid z))}-\frac{1}{(1-W(y\mid z))}\int_y^\infty\tilde\psi(s,z)\,F(\,\mathrm{d}s\mid z)
\end{aligned}$$

Similar computations can be used to check (3.28):

$$\begin{aligned}(\mathcal{R}_2\tilde{\nu})(y,z) &= \tilde{\nu}(y,0,z) - \mathrm{E}\,(\tilde{\nu}(Y,\Delta,Z) \mid Y > y, Z = z) \\ &= \frac{1}{(1-W(y \mid z))}\int_y^\infty \tilde{\psi}(s,z)\, F(\,\mathrm{d}s \mid z) \\ &\quad - \frac{1}{(1-W(y \mid z))}\int_y^\infty \tilde{\psi}(s,z)\, F(\,\mathrm{d}s \mid z) \\ &= 0.\end{aligned}$$

Thus, by corollary 3.19 the efficient influence function of $\nu(W_{F_z,H,G_z})$ is given in (3.29).

The efficient influence functions for estimating functionals of the type defined in (3.23) and (3.24) are now obtained by substituting $\tilde{\psi} = \tilde{\psi}_{11} + \tilde{\psi}_{12}$ into (3.29). For instance, the efficient influence function for estimation of the functional defined in (3.23) equals (compare proposition 3.3)

$$\tilde{\psi}(\varphi) = \varphi - \mathrm{E}\,(\varphi \mid Z) + \mathrm{E}\,(\varphi \mid Z)Q\varphi$$

Substituting into (3.29)

$$\begin{aligned}\tilde{\nu}(\varphi; y, \delta, z) = \delta\,\frac{\varphi(y,z)}{(1-G(y \mid z))} + \frac{(1-\delta)}{(1-W(y \mid z))}\int_y^\infty \varphi(s,z)\, F(\,\mathrm{d}s \mid z) \\ - \int C_2(y \wedge s \mid z)\varphi(s,z)\, F(\,\mathrm{d}s \mid z) - \mathrm{E}\,(\varphi(T,Z)). \quad (3.32)\end{aligned}$$

Consequently, the efficient influence function of the function valued parameter defined in (3.24) is given by $\tilde{\nu}(\mathcal{F}; y, \delta, z)(\varphi) = \{\varphi \mapsto \tilde{\nu}(\varphi; y, \delta, z)\}$.

The information bound I_ν for estimation of $\nu(W_{F_z,H,G_z}) = \int \varphi\,\mathrm{d}(F_z \times H)$ is given by the inverse of the variance of $\tilde{\nu}(Y,\Delta,Z)$:

$$I_\nu^{-1}(\varphi) = \mathrm{E}\,\{\tilde{\nu}(\varphi; Y, \Delta, Z)^2\}.$$

For estimation of a function valued parameter such as defined in (3.24), the inverse information covariance function (BKRW equation 5.2.23) becomes

$$I_\nu^{-1}(\varphi_1, \varphi_2) = \mathrm{E}\,\{\tilde{\nu}(\varphi_1; Y, \Delta, Z)\,\tilde{\nu}(\varphi_2; Y, \Delta, Z)\} - \mathrm{E}\,\tilde{\nu}(\varphi_1; Y, \Delta, Z)\,\mathrm{E}\,\tilde{\nu}(\varphi_2; Y, \Delta, Z).$$

Using relations (3.20)-(3.22) we can identify the functionals of our interest as parameters of the model $\mathcal{W}$. For instance, if $\varphi \in \mathcal{L}_2(W^{(1)}/G_z)$, then ν is identifiable, by proposition 3.6:

$$\begin{aligned}\nu(\varphi, W_{F,G,H}) &= \int \varphi(t,z)\frac{W^{(1)}(\,\mathrm{d}t,\,\mathrm{d}z)}{(1-G(t \mid z))} \\ &= \int \varphi(t,z)\exp\{\int_0^t \frac{W^{(0)}(\,\mathrm{d}s \mid z)}{(1-W(s \mid z))}\}\, W(\,\mathrm{d}t, 1,\,\mathrm{d}z)\end{aligned} \quad (3.33)$$

This suggests an IPCW estimator that re-weights all uncensored observations by the inverse of an estimate of the conditional survival function of the censoring variable. Nonparametric consistent estimation of the conditional distribution function G_z, however, requires some extra conditions: if the vector of covariates is discrete, the stratified Kaplan-Meier estimator for the censoring distribution can be used to estimate the conditional distribution function G_z. Asymptotically the stratified Kaplan-Meier estimator is unbiased. Otherwise, if Z has also continuous components, nonparametric, asymptotically unbiased estimates for G_z in nonparametric models can be achieved e.g. by window convolution smoothing methods (c.f. chapter 4). In terms of a real data set these methods would artificially define strata by assuming that the distribution functions of individuals with close covariate values are close with respect to an appropriate distance measure defined on the $\mathbf{R}(Z)$. In chapter 4 we investigate nonparametric estimation of general parameters that integrate a Hadamard differentiable functional of a conditional distribution function given a continuous covariate. The results will then be applied to functionals of the form (3.33), i.e. to integrals of the Hadamard differentiable functional

$$\frac{1}{(1 - G(t \mid z))} = \exp\left\{ \int_0^t \frac{W^{(0)}(\,\mathrm{d}s \mid z)}{(1 - W(s- \mid z))} \right\} ..$$

On the other hand, one specific semiparametric model is of particular interest. If censoring is stochastically independent of the covariates, then re-weighting with the Kaplan-Meier estimator for the marginal censoring distribution yields asymptotically consistent estimators (Stute 1993). The model is semiparametric in the sense of Groeneboom and Wellner (1994, definition 1.1) because the induced tangent space of the model for X under independent censoring is not full.

Chapter 4

Nonparametric functional estimation

In this chapter we investigate methods for constructing efficient estimators of parameters that arise by integrating Hadamard differentiable functionals of a smooth function. The results obtained here will be used in chapter 5 for showing asymptotic efficiency of certain inverse probability of censoring weighting estimators.

The empirical distribution is well-known to be an asymptotically efficient estimator in nonparametric models. Moreover, it is known that asymptotic optimality can be preserved for plug-in estimators of functionals that are Hadamard or compactly differentiable transformations of the underlying probability distribution (for details see e.g. Van der Vaart (1988) or Gill (1989)). We are concerned with estimation of linear functionals of nonparametric functionals of a conditional distribution function; the motivation comes from certain representations of (prediction error) functionals in information loss models. How such functionals can be estimated, can be learned from the large literature on a closely related problem: estimation of integral functionals of density derivatives. This problem has been investigated by many authors, at hand its most famous example, the integral of a squared density, which occurs for instance in the asymptotics of the Hodges-Lehmann estimator and the Wilcoxon statistic, see e.g. Dmitriev and Tarasenko (1974) or Schweder (1975) for pioneering work. Other motivation for studying functionals of this type comes from practical problems such as estimation of integrated mean squared error of kernel density estimates in context with nonparametric bandwidth selection (Hall and Marron 1987). For estimation of nonparametric functionals of a regression function that are of practical importance see e.g. Doksum and Samarov (1995).

Convergence rates for estimating high dimensional parameters, such as a density function or a conditional distribution function are typically slower than $\sqrt{n}$. Stone (1980) has established the optimal rate of convergence for a general class of nonparametric estimators of density and regression functions uniformly over Sobolev classes of functions. Averaging smooth functionals defined on a function

space, however, can recover the rate $\sqrt{n}$ provided that the underlying function satisfies an appropriate Hölder condition. Indeed, it was shown by Ritov and Bickel (1990) (see also Birgé and Massart (1995) and Donoho and Liu (1991)) that for the existence of optimal nonparametric estimators of integrated density derivatives a certain amount of smoothness of the underlying density function is needed. For instance, no locally uniformly consistent estimator converges at any positive rate uniformly in the class of bounded density functions on the line. Similar minimal smoothness conditions have to be assumed for estimation of nonparametric functionals of regression derivatives and for nonparametric estimation in the white noise model (Brown and Low 1996; Nussbaum 1996; Efromovich and Samarov 1996).

For integral functionals of density, different authors favor different methods and models for the underlying density, they commonly achieve the rate $\sqrt{n}$ with their estimators (Schweder 1975; Bickel and Ritov 1988; Hall and Marron 1987; Laurent 1996).

Typically, the optimal rates of convergence and thus also the optimal smoothing parameters depend on how smooth the underlying function is. But in practical applications the degree of smoothness of the underlying density or regression function is typically unknown. It is therefore important to establish adaptive estimates that have the same asymptotic performance as the rate-optimal estimator obtained when the degree of smoothness is known (Hall and Johnstone 1992; Efromovich and Low 1996; Efromovich and Samarov 2000).

Another requirement for the existence of Gaussian regular, efficient estimators is that the functional can be locally approximated by a smooth linear functional. We adopt the classification of Goldstein and Messer (1992) to distinguish between functionals that are smooth and others that have atomic components. In our setting, integrating smooth and particular atomic functionals leads to smooth and regularly estimable functionals. It turns out that plugging-in an undersmoothed kernel type estimator for the density, respectively for the regression function, is efficient. A kernel type estimator is called undersmoothed if the bandwidth is smaller than the optimal bandwidth that corresponds to the optimal rate of convergence, see e.g Goldstein and Khas'minskii (1996) for results in that direction.

Our main result will be an application of a version of the Delta method using smoothed empirical distributions as plug-in estimators (section 4.1). We also make use of results from empirical process theory (Van der Vaart 1994; Rost 2000) and extend the results obtained to general functionals and function valued parameters. In section 4.2 we transport the results to the regression problem. Also here undersmoothed multivariate kernel and symmetrized nearest neighbor type estimators can be shown to be efficient.

4.1 Efficient estimation of integral functionals of a density

Suppose U is a real random variable with law Q and Radon-Nikodym density q with respect to Lebesgue measure on $(\mathbb{R}, \mathcal{B})$. In view of the results of Ritov and Bickel (1990) we consider only models for Q that are included in the model of all densities that satisfy the following (almost sure) Hölder condition for some $\alpha > 1/4$: for some essentially bounded function $g \in \mathcal{L}_2$,

$$\mathcal{Q}_0^\alpha = \{Q : |q(u+v) - q(u)| \leq g(u)\,|v|^\alpha\}. \tag{4.1}$$

Lemma 1 of Bickel and Ritov (1988) then yields that

$$\sup |q(u) : u \in \mathbb{R}, Q \in \mathcal{Q}_0^\alpha| < \infty;$$

whence $\mathcal{Q}_0^\alpha \subseteq \mathcal{L}_\infty \cap \mathcal{L}_1$. We endow $\mathcal{Q}_0^\alpha$ with the supremum norm and denote the closure by $\mathcal{Q}^\alpha = \overline{(\mathcal{Q}_0^\alpha, \|\cdot\|_\infty)}$. Moreover, in this section, a particular sequence of density estimators $\hat{q}_n$ will be always connected to a choice of a submodel $\mathcal{Q} \subseteq \mathcal{Q}^\alpha$, such that convergence of $\hat{q}_n$ to q is uniform (for $q \in \mathcal{Q}$). We also assume that $\mathcal{Q}$ includes all infinitely often differentiable densities on the real line, which readily implies that $\mathcal{Q}$ is a nonparametric model.

We start by motivating functionals of interest and candidate estimators. The empirical measure $\hat{Q}_n$ corresponding to *iid* copies $U_1, \ldots, U_n$ of U assigns uniform mass to each observation: for every measurable set B and denoting Dirac measure at zero by δ_0:

$$\hat{Q}_n(B) = \frac{1}{n}\sum_{i=1}^n \delta_0(U_i - B),$$

where $u - B \equiv \{y : \exists\, b \in B :\, y = u - b\}$.

Let $\mathcal{F} \subset \mathcal{L}_2 \subset \mathcal{L}_2(Q)$ and recall the functional of Q defined for every $\varphi \in \mathcal{F}$ by

$$\psi(Q) = \psi(\varphi, Q) = \int \varphi(u)\, Q(\,\mathrm{d}u) = \int \varphi(u)\, q(u)\,\mathrm{d}u. \tag{4.2}$$

It is well-known that the plug-in estimator

$$\psi(\hat{Q}_n) = \int \varphi(u)\hat{Q}_n(\,\mathrm{d}u)$$

is an asymptotically efficient estimator of $\psi(Q)$. Similarly, the process $\varphi \mapsto Q\varphi$ can be efficiently estimated by $\varphi \mapsto \hat{Q}_n\varphi$ (c.f. chapter 3). It is known that efficiency is preserved for the plug-in estimator under appropriately differentiable transformations of Q (see e.g. Van der Vaart (1988)). For instance, the parameter

$$\psi(Q) = \int \varphi(u)\phi(Q)\, Q(\,\mathrm{d}u),$$

where ϕ is Hadamard differentiable, can be efficiently estimated by

$$\psi(\hat{Q}_n) = \int \varphi(u)\phi(\hat{Q}_n)\,\hat{Q}_n(\,\mathrm{d}u).$$

We are interested in analogous results for linear functionals where ϕ is a function of q instead of a function of Q. A so-called 'direct' estimator plugs-in a (nonparametric) estimator of the density. Alternatively, one could use a preliminary estimate of q and then consider estimation of the quasi U-statistic which results after substituting the estimate for the unknown density. In the latter case one estimates a linear approximation of the influence function of the functional of interest.

Smoothed empirical measures can be obtained by convolution of $\hat{Q}_n$ with a sequence of signed, random measures of uniformly bounded variation that converge weakly in probability to Dirac measure at zero (see e.g. Winter (1973)).

We consider a particular subclass of smoothed versions of $\hat{Q}_n$ that is obtained by kernel or window convolution. Throughout the rest of this and the next section let K be a symmetric kernel density function of bounded variation, with mean zero and such that $\int |K(u)|\,\mathrm{d}u \leq M < \infty$ for some constant M. Define also a (data dependend) sequence of bandwidths $a = a(n)$ such that $a(n) \to 0$ as $n \to \infty$. We use the notation $K_a(u) = (1/a)K(\frac{u}{a})$ in what follows. Let $\hat{Q}_{K_a}$ be the probability measure defined on the range of U that has density with respect to Lebesgue measure given by

$$\hat{q}_{K_a}(u) = \int K_a(u-v)\,\hat{Q}_n(\mathrm{d}v) = (K_a \star \hat{Q}_n)(u).$$

Then $\hat{q}_{K_a}$ is the well-known Rosenblatt-Parzen estimator. $\hat{Q}_{K_a}$ is a smoothed empirical measure which converges weakly to Dirac measure at zero, including the cases with nonnegative kernel functions and randomly chosen bandwidth, for instance by cross-validation.

The following result of real analysis is important for proving consistency of kernel density estimators (for a proof see e.g. theorem 2.1.1 of Prakasa Rao (1983)).

Theorem 4.1
Let K be a kernel function such as specified above. If $a(n) \to 0$ and $g : \mathbb{R}^k \to \mathbb{R}$ is integrable and continuous, then

$$\lim_{a \downarrow 0}(g \star K_a)(u) = \lim_{a \downarrow 0} \int g(v)\,K_a(u-v)\;\mathrm{d}v = g(u).$$

If the function g is uniformly continuous then convergence in the last preceding display is also uniform in u.

□

Now we introduce functionals of the type that we want to estimate. Suppose that $\phi : \mathcal{Q} \subseteq \mathcal{Q}^\alpha \to \mathcal{L}_1 \cap \mathcal{L}_\infty$ is Hadamard differentiable at $q \in \mathcal{Q}$; that is, there exists a continuous linear map $\dot{\phi}_q : \mathcal{Q}^\alpha \to \mathcal{L}_1 \cap \mathcal{L}_\infty$ such that

$$\left\| \frac{\phi(q + \epsilon g) - \phi(q)}{\epsilon} - \dot{\phi}_q(g) \right\|_\infty \to 0 \tag{4.3}$$

for any $\epsilon \to 0$ uniformly for g in every compact subset of $\mathcal{Q}$. Note that in contrast to chapter 3, we now make the dependence of the derivative on q visible in the notation.

For statistical applications a weaker form of Hadamard differentiability is useful (Gill 1989): ϕ is called Hadamard differentiable tangentially to a subset $D \subseteq \mathcal{Q}^\alpha$ if (4.3) holds with g replaced by any sequence g_n that converges to $g \in D$.

A natural choice for D is the tangent space $\dot{\mathcal{Q}}$ at $q \in \mathcal{Q}$. In the present situation, the tangent space for $\mathcal{Q}^\alpha$ at q equals $\mathcal{L}_2^0(Q)$, because the continuous functions are dense in $\mathcal{L}_2(Q)$, see example 2.1.4 of Pfanzagl (with the assistance of W. Wefelmeyer) (1985). Any model whose tangent space is a proper subset of the nonparametric tangent space could be called semiparametric (Groeneboom and Wellner 1994, definition 1.1). In view of this, and our assumption that $\mathcal{Q}$ includes all infinitely often differentiable densities we are dealing here with nonparametric models. Hadamard differentiability tangentially to $\dot{\mathcal{Q}}$ is sometimes called pathwise differentiability. Note also that the set of local variations such as used by Goldstein and Messer (1992), in a very similar situation, is closely related to the tangent space.

Let $\mathcal{F} \subset \mathcal{L}_2$ and define $\psi : \mathcal{Q} \times \mathcal{F} \to \mathbb{R}$

$$\psi(\varphi, q) = \int \varphi(u)\, \phi(q)(u)\, \mathrm{d}u, \tag{4.4}$$

where $\mathcal{Q} \subseteq \mathcal{Q}^\alpha$. Since $\varphi \in \mathcal{F}$ is fixed most of the time we also write $\psi(q)$ for $\psi(\varphi, q)$. The most extensively studied example is perhaps the integrated square of a density (see also example 4.9 below):

$$\psi(q) = \int q^2(u)\, \mathrm{d}u \int q(u)\, Q(\mathrm{d}u).$$

Define the plug-in estimator for (4.4) by

$$\hat{\psi}_n \equiv \psi(\hat{q}_n) = \int \varphi(u)\, \phi(\hat{q}_n)(u)\, du, \tag{4.5}$$

where $\hat{q}_n$ is any nonparametric estimator of q. Using the Rosenblatt-Parzen estimator, the question of when the plug-in estimator is efficient reduces to the question of how to choose the bandwidth in a given model $\mathcal{Q}$ – a rather delicate problem.

For our purposes it is important to have that the derivative $\dot{\phi}_q$ at $g \in \mathcal{Q}$ admits a certain integral representation. Since $\mathcal{Q} \subseteq \mathcal{Q}^\alpha$ we have that $g \mapsto \dot{\phi}_q(g)(u)$ is a bounded functional from (a subset of) the continuous functions to $\mathbb{R}$. We can therefore use the Riesz representation theorem for linear functionals on $C_0(\mathbb{R})$ (Rudin 1987, theorem 6.19) and establish the almost sure identity

$$\dot{\phi}_q(g)(u) = \int g(v)\, \mu_q(u, \mathrm{d}v), \tag{4.6}$$

where $\mu_q(u, \cdot)$ is a regular Borel measure on $\mathbb{B}$ depending on q and u. Clearly boundedness of the function $u \mapsto \dot{\phi}_q(g)(u)$ implies that $u \mapsto \int g(v)\, \mu_q(u, \mathrm{d}v)$ is bounded. Following Goldstein and Messer (1992) we call the functional $\dot{\phi}_q$ smooth if the representing measure has a Lebesgue density $\tilde{\phi}_q(u, \cdot)$. And ϕ is called smooth on $\mathcal{Q}$ if it has smooth derivatives at all $q \in \mathcal{Q}$. On the other hand, if the representing measure μ_q has discrete components the functional $\dot{\phi}_q$ is called atomic.

Typically, smooth functionals are regularly estimable and atomic functionals are not. If $\dot{\phi}_q$ is smooth at q and assuming that the density of the representing measure exists pointwise and is included in the tangent space $\dot{\mathcal{Q}}$ of $\mathcal{Q}$ at q, then $\tilde{\phi}_q(u, \cdot)$ is called the efficient influence function for estimation of the real functional $\phi(q)(u)$. If ϕ is atomic it is typically not estimable at the $\sqrt{n}$ convergence rate, hence there exists no efficient influence function.

However, even if ϕ is atomic it is often possible to construct a sequence of smooth functionals whose (efficient) influence functions approximates μ_q.

Definition 4.2 *Let $\phi : \mathcal{Q} \to \mathcal{L}_1 \cap \mathcal{L}_\infty$ be Hadamard differentiable tangentially to $\dot{\mathcal{Q}} \cap \mathcal{Q}$. We call $\tilde{\phi}_q$ a generalized influence function for estimation of ϕ if there exists a sequence of smooth functionals ϕ_ϵ with influence function $\tilde{\phi}_{q,\epsilon}$ such that*

$$\dot{\phi}_{q,\epsilon}(g)(u) = \int g(v)\, \tilde{\phi}_{q,\epsilon}(u, v)\, \mathrm{d}v \to \int g(v)\, \tilde{\phi}_q(u, v)\, \mathrm{d}v$$

as $\epsilon \to 0$.

If ϕ is atomic typically the generalized influence is a distribution. For (heuristicall) illustration we consider the identity operator $\phi(q) = q$. The Gateaux derivative of the identity at q in direction $g \in \mathcal{Q}$ is given by

$$\dot{\phi}_q(g) = \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \phi(q + \epsilon g) = g.$$

The functional $\dot{\phi}_q$ is atomic, for it can be represented by a Dirac measure:

$$\dot{\phi}_q(g)(u) = \int g(u - v)\, \delta_0(\mathrm{d}v).$$

Define the sequence

$$K_\epsilon(x) = \begin{cases} \frac{1}{2\epsilon} & \text{for } |x| \le \epsilon \\ 0 & \text{else.} \end{cases}$$

Then $\phi_\epsilon = K_\epsilon \star Q$ is a smooth functional that converges to ϕ (by theorem 4.1) as $\epsilon \to 0$ with influence function given by

$$\tilde{\phi}_{q,\epsilon}(U) = K_\epsilon(U) - \phi_\epsilon.$$

Clearly $\tilde{\phi}_{q,\epsilon}$ converges to $\delta_U - q(U)$ as $\epsilon \to 0$. The generalized influence function of $q \mapsto q$ is thus given by $\tilde{\phi}_q(u,v) = \delta_0(u-v) - q(u)$. Note that this construction is related to how Hampel (1974) defined influence functions for functionals defined on distribution functions.

We will now show that functionals of the type (4.4) are smooth and regularly estimable on suitably defined models, if $\dot{\phi}_q$ is either smooth or if the generalized influence function $\tilde{\phi}_q$(u,v) is of the form $\tilde{\phi}_q(u)\,\delta_0(u-v)$.

The following theorem is truly an application of the Delta method.

Theorem 4.3
Let $\mathcal{Q} \subseteq \mathcal{Q}^\alpha$, let $\hat{Q}_n^*$ be a sequence of smoothed empirical distributions with Lebesgue density $\hat{q}_n$ such that for every function $f \in \mathcal{L}_2(Q)$

$$\sqrt{n}(\hat{Q}_n^* - Q)(f) = \sqrt{n}(\hat{Q}_n - Q)(f) + o_P(1).$$

Let $\phi : (\mathcal{Q}, \|\cdot\|_\infty) \to (\mathcal{L}_1 \cap \mathcal{L}_\infty, \|\cdot\|_\infty)$ be Hadamard differentiable tangentially to the tangent space $\dot{\mathcal{Q}}$ at q such that the derivative $\dot{\phi}_q$

(i) is smooth at q with influence function $\tilde{\phi}_q(u,v)$. Then the efficient influence function for estimation of ψ in the model $\mathcal{Q}$ is given by

$$\tilde{\psi}(U) = \int \varphi(U)\,\tilde{\phi}_q(u,U)\ \mathrm{d}u - \int\int \varphi(u)\,\tilde{\phi}_q(u,v)\ \mathrm{d}u\,Q(\,\mathrm{d}v).$$

(ii) is atomic with generalized influence function $\tilde{\phi}_q(u)\,\delta_0(u-v)$. Then the efficient influence function for estimation of ψ in the model $\mathcal{Q}$ is given by

$$\tilde{\psi}(U) = \varphi(U)\,\tilde{\phi}_q(U) - \int \varphi(u)\,\tilde{\phi}_q(u)\,Q(\,\mathrm{d}u).$$

In both cases, if in addition $\mathrm{Var}(\tilde{\psi}(U)) < \infty$, the plug-in estimator (4.5) is asymptotically efficient for estimation of the functional $\psi(q)$ defined in (4.4). Moreover,

$$\sqrt{n}(\hat{\psi}_n - \psi) \Rightarrow \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \mathrm{E}\,(\tilde{\psi}^2)$ *equals the nonparametric information bound for estimation of* (4.4)*:*

$$\sigma^2 = \mathrm{E} \iint \varphi(u)\varphi(v)\,\mu_q(U,\,\mathrm{d}u)\,\mu_q(U,\,\mathrm{d}v) - \left\{\mathrm{E} \int \varphi(u)\,\mu_q(U,\,\mathrm{d}u)\right\}^2,$$

where $\mu_q(u,\,\mathrm{d}v)$ *has either Lebesgue density* $\tilde{\phi}_q(u,v)$ *or is represented by the distribution* $\tilde{\phi}_q(u)\,\delta_0(u-v)$*.*

Remark 4.4
From the form of the (efficient) influence function of ψ it becomes clear that we can not regularly estimate the functional (4.4) if $\dot{\phi}_q$ is atomic and $\mu_q(u,\,\mathrm{d}v) = \tilde{\phi}_q(u,v)\delta_x(\,\mathrm{d}v)$ for some value $x \neq u$, since then ψ is atomic.
□

Proof:

We first establish the linear expansion. Since

$$\sqrt{n}\,\|\phi(\hat{q}_n) - \phi(q) - \sqrt{n}\,\dot{\phi}_q(\hat{q}_n - q)\,\|_\infty \to 0,$$

by Hadamard differentiability of ϕ we can write

$$\begin{aligned}\sqrt{n}(\hat{\psi}_n - \psi) &= \sqrt{n}\int \varphi(u)\,\{\phi(\hat{q}_n) - \phi(q)\}\,\mathrm{d}u \\ &= \sqrt{n}\int \varphi(u)\,\dot{\phi}_q(\hat{q}_n - q)(u)\,\mathrm{d}u + o_P(1).\end{aligned} \tag{4.7}$$

(i) if $\mu_q(u,\,\mathrm{d}v) = \tilde{\phi}_q(u,v)\,\mathrm{d}v$ the right hand side of (4.7) equals

$$\iint \varphi(u)\,\tilde{\phi}_q(u,v)\,\mathrm{d}u\,\sqrt{n}\,\{\hat{q}_n(v) - q(v)\}\,\mathrm{d}v$$
$$= \iint \varphi(u)\,\tilde{\phi}_q(u,v)\,\mathrm{d}u\,\sqrt{n}\,\{\hat{Q}_n^* - Q\}(\,\mathrm{d}v);$$

(ii) if $\mu_q(u,\,\mathrm{d}v) = \tilde{\phi}_q(u)\delta_0(u-v)$ the right hand side equals

$$\int \varphi(u)\,\tilde{\phi}_q(u)\,\sqrt{n}\,\{\hat{q}_n(u) - q(u)\}\,\mathrm{d}u$$
$$= \int \varphi(u)\,\tilde{\phi}_q(u)\,\sqrt{n}\,\sqrt{n}\,\{\hat{Q}_n^* - Q\}(\,\mathrm{d}u)$$

In both cases the functions under the outermost integral are square integrable since $u \mapsto \tilde{\phi}_q(u,\cdot)$ is bounded. Thus, the claim of the theorem follows from the assumption that $\sqrt{n}(\hat{Q}_n^* - Q)f$ is asymptotically equivalent to $\sqrt{n}(\hat{Q}_n - Q)f$ for $f \in \mathcal{L}_2(Q)$ and from the following lemma.
□

Lemma 4.5 (Efficient influence function)
Under the conditions of theorem 4.3, the functional $\psi(q)$ defined in (4.4) is Hadamard differentiable tangentially to $\dot{\mathcal{Q}}$ with derivative at q in direction $g \in \dot{\mathcal{Q}}$ given by

$$\dot{\psi}_q(g) = \int \varphi(u)\, \dot{\phi}_q(g)(u)\, \mathrm{d}u.$$

Dependent of whether ϕ is (i) smooth or (ii) atomic, the efficient influence function for estimation of ψ is as given in theorem. 4.3.

Proof: Let q_ϵ be a path in $\mathcal{Q}$ with tangent $g \in \dot{\mathcal{Q}}$, i.e.

$$g = \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \log(q_\epsilon).$$

Then $\dot{\psi}_q(g)$ is as given in the lemma, since

$$\begin{aligned}
\psi(q) - \psi(q_\epsilon) - \dot{\psi}_q(g) &= \int \varphi(u) \left\{ \phi(q)(u) - \phi(q_\epsilon)(u) - \dot{\phi}_q(g)(u) \right\} \mathrm{d}u \\
&\leq \sup_u \left| \phi(q)(u) - \phi(q_\epsilon)(u) - \dot{\phi}_q(g)(u) \right| \int \varphi(u)\, \mathrm{d}u \\
&= o_P(1), \qquad \text{as } \epsilon \to 0.
\end{aligned}$$

Under (i) of the theorem, we can expand $\dot{\psi}_q$ to

$$\int \varphi(u) \int \tilde{\phi}_q(u,v) g(v)\, \mathrm{d}v\, \mathrm{d}u = \iint \varphi(u)\, \tilde{\phi}_q(u,v)\, \mathrm{d}u\, g(v)\, \mathrm{d}v.$$

Thus, $\int \varphi(u)\, \tilde{\phi}_q(u,U)\, \mathrm{d}u$ is an influence function. Similarly, under (ii) we can expand $\dot{\psi}_q$ to

$$\int \varphi(u) \int \tilde{\phi}_q(u) g(v)\, \delta_0(u-v)\, \mathrm{d}v\, \mathrm{d}u = \int \varphi(u)\, \tilde{\phi}_q(u)\, g(u)\, \mathrm{d}u,$$

which shows that $\varphi(U)\, \tilde{\phi}_q(U)$ is an influence function. The respective efficient influence functions are now obtained by projecting the influence functions onto the tangent space $\dot{\mathcal{Q}}$.
□

It is natural to decompose a given nonparametric density estimator $\hat{q}_n$ into its bias and variance term, respectively given by $\mathrm{E}\,(\hat{q}_n) - q$ and $\hat{q}_n - \mathrm{E}\,(\hat{q}_n)$. If the function φ in the definition of the functional (4.4) is uniformly continuous and bounded then the kernel method satisfies the conditions of theorem 4.3.

Theorem 4.6 (Undersmoothed kernels)
Let $q \in \mathcal{Q} \subseteq \mathcal{Q}^\alpha$, let $f \in \mathcal{L}_2 \cap \mathcal{C}_0^\alpha$ for some $\alpha > 0$, let K_a a kernel density function such as specified above with bandwidth $a(n)$. If $\sqrt{n}\, a(n) \to 0$ then

$$\int f(u) \left\{ \hat{q}_{K_a}(u) - q(u) \right\} \mathrm{d}u = \int f(u)\, (\hat{Q}_n - Q)(\mathrm{d}u) + o_P(1/\sqrt{n}).$$

Proof: First note that $\mathrm{E}\,(\hat{q}_{K_a}) = K_a \star q$. Since $\sqrt{n}\,a(n) \to 0$ and q is uniformly continuous it follows from theorem 4.1 that

$$\sqrt{n}\sup_u\{\mathrm{E}\,(\hat{q}_{K_a})(u) - q(u)\} = \sqrt{n}\sup_u\{(K_a \star q)(u) - q(u)\} \to 0,$$

whence

$$\int f(u)\{\mathrm{E}\,(\hat{q}_{K_a})(u) - q(u)\}\,\mathrm{d}u = o_P(1/\sqrt{n}).$$

By Fubini's theorem and symmetry of K_a

$$\begin{aligned}\sqrt{n}\int f(u)\{\hat{q}_{K_a}(u) - \mathrm{E}\,(\hat{q}_{K_a})(u)\}\,\mathrm{d}u \\ &= \iint f(u)\,K_a(u-v)\,\mathrm{d}u\,\sqrt{n}\{\hat{Q}_n(\,\mathrm{d}v) - Q(\,\mathrm{d}v)\} \\ &= \int\{(f \star K_a)(v) - f(v)\}\,\sqrt{n}\{\hat{Q}_n(\,\mathrm{d}v) - Q(\,\mathrm{d}v)\} \\ &\quad + \int f(v)\,\sqrt{n}\{\hat{Q}_n(\,\mathrm{d}v) - Q(\,\mathrm{d}v)\}.\end{aligned}$$

The first term on the right hand side of the last preceding display is smaller or equal to

$$\sup_u\{(f \star K_a)(u) - f(u)\}\int\sqrt{n}\{\hat{Q}_n(\,\mathrm{d}u) - Q(\,\mathrm{d}u)\} = o_P(1),$$

by theorem 4.1, since f is assumed uniformly continuous. Combination of the bias and variance term completes the proof.
□

Remark 4.7

- $a(n) = o_P(1/\sqrt{n})$ is smaller than the optimal bandwidth as defined in (Stone 1980). See also Goldstein and Messer (1992) or Goldstein and Khas'minskii (1996) for undersmoothed kernel estimators of general functionals that depend on the (unknown) degree of smoothness.

- In practical applications it is necessary to 'estimate' the smoothing parameter from the data. For instance, the function $f \star K_a$ in the preceding theorem depends on the bandwidth and it is therefore necessary to allow that $a(n)$ is a random variable. If $\|f \star K_a - f\|_Q$ converges in probability to 0 and the class $\{f_a = f \star K_a : a \in [0,1]\}$ is a Donsker class, then theorem 19.24 of Van der Vaart (1998) yields

$$\sqrt{n}(\hat{Q}_n - Q)f_a \Rightarrow \mathcal{N}(0, \|f\|_Q^2).$$

□

Theorem 4.6 shows that the conditions of theorem 4.3 are satisfied for the undersmoothed kernel density estimator, either if $\dot{\phi}_q$ is smooth and

$$u \mapsto \int \varphi(v)\, \tilde{\dot{\phi}}_q(u,v)\, \mathrm{d}u$$

is integrable and uniformly continuous, or if $\dot{\phi}_q$ is atomic and then

$$u \mapsto \varphi(u) \tilde{\dot{\phi}}_q(u)$$

is integrable and uniformly continuous.

The results of (Van der Vaart 1994) imply the conditions of theorem 4.3 also for discontinuous functions φ. Moreover, it seems possible to deduce a functional limit theorem for the process $\{\psi(\varphi, q) : \varphi \in \mathcal{F}\}$.

Theorem 4.8 (Van der Vaart (1994))
Let λ_n be a signed random measure of bounded variation that converges weakly to Dirac measure at 0, such that $\hat{Q}_n^* = \lambda \star \hat{Q}_n$ has a density function $\hat{q}_n$ with respect to Lebesgue measure. Let $\mathcal{F}$ be a Q-Donsker class that is closed under translation. If both terms,

$$\left(\iint \{f(x+y) - f(x)\}\, \lambda_n(\mathrm{d}y)\, Q(\mathrm{d}u) \right)^2$$

and

$$\sqrt{n} \left| \iint \{f(x+y) - f(x)\}\, \lambda_n(\mathrm{d}y)\, Q(\mathrm{d}u) \right|,$$

converges to 0 in outer probability uniformly in $f \in \mathcal{F}$, then the smoothed empirical process $\sqrt{n}(\hat{Q}_n^* - Q)$ converges weakly in distribution in $l^\infty(\mathcal{F})$ to a tight Brownian bridge process.

□

The problem with theorem 4.8 is that for applications one has to show that the class of all translates of all functions in a Donsker class of interest maintains the Donsker propertie. It is however not clear if the class of translates of a single Q-square integrable function is again Q-square integrable. On the other hand, if the class of translates of a single monotone function has measurable envelope function, then it is a Vapnik-Červonenkis class of index 2 (Van der Vaart and Wellner 1996, lemma 2.6.16). This indicates that theorem 4.8 considerably enlarges the class of functions, where theorem 4.3 applies. However, in view of our previous comment, it seems valuable to study the results of Rost (2000) who is able to drop that $\mathcal{F}$ is translation invariant and still arrives at functional

limit theorems for the smoothed empirical process. Moreover, the results of the latter authors could be used to prove functional limit theorems for processes of the type $\{\psi(\varphi, q) : \varphi \in \mathcal{F}\}$. Thinking of functional central limit theorems for prediction error curves (c.f. chapter 2) we note that the class of half intervals in $\mathbb{R}$ is translation invariant and a Vapnik-Červonenkis class, so that theorem 4.8 applies and it should be possible to establish confidence bands. However, this is not achieved in the framework of this thesis.

Example 4.9 (Integral of a squared density)
Suppose $\mathcal{Q} \subseteq \mathcal{Q}^\alpha$ and set $\psi(q) = \int q(u)^2 \, \mathrm{d}u$, i.e. $\varphi = 1$ and $\phi : \mathcal{Q} \to \mathcal{Q}^\alpha$ is given by $\phi(q) = q^2$. ϕ is Hadamard differentiable with derivative at q in direction g given by $\dot{\phi}_q(g) = 2gq$. The function $\dot{\phi}_q(\cdot)$ is bounded and uniformly continuous; and evaluated at u can be represented by the generalized influence function

$$\tilde{\phi}_q(u, U) = 2\, q(u)\, \delta_0(u - U) :$$

Observe that $\tilde{\phi}_q(u, v)$ is the limit of $2 \int K_\epsilon(u - v)\, q(u) \, \mathrm{d}v$ which is an influence function for estimating the functional $\phi_\epsilon(Q) = \left(\int K_\epsilon(u - v)\, Q(\mathrm{d}v)\right)^2$. The efficient influence function and the information bound for estimating ψ can now be obtained by substituting φ and $\tilde{\phi}_q$ into the formulas of theorem 4.3; they are respectively given by:

$$\begin{aligned}\tilde{\psi}(U) &= \varphi(U)\, \phi(q)(U) + \iint \varphi(u)\, \tilde{\phi}_q(u, U)\, Q(\mathrm{d}u) \\ &= 2 \left\{ q(U) - \int q(u)^2 \, \mathrm{d}u \right\}\end{aligned}$$

and

$$\begin{aligned}\sigma^2 &= 4 \left\{ \mathrm{E}\, q^2(U) - 2\, \mathrm{E}\, q(U) \int q^2(u) \, \mathrm{d}u + \left(\int q(u)^2 \, \mathrm{d}u \right)^2 \right\} \\ &= 4 \left\{ \int q^3(u) \, \mathrm{d}u - \left(\int q(u)^2 \, \mathrm{d}u \right)^2 \right\}.\end{aligned}$$

It follows readily from theorem 4.6 that by using a non-optimal bandwidth and considering first order approximation only, the plug-in Rosenblatt-Parzen estimator is efficient.

A number of papers address optimal estimation of the integral of the square of a density assuming the degree of smoothness of the underlying density to be known, see for instance Bickel and Ritov (1988), Hall and Marron (1987) and Laurent (1996). See Hall, Hu, and Marron (1995), Efromovich and Samarov (2000) and the references therein for methods that adaptively select the smoothing parameter.
□

4.2 Efficient estimation of nonlinear functionals of conditional distribution and regression functions

In this section we analyze nonparametric estimators of integral functionals of conditional distribution functions. Apparently, the representations obtained for parameters of interest in the right censored regression setting are of this form (c.f. equation (3.33)). We will make use of the results obtained in the current section later, in chapter 5, to derive the asymptotics of inverse probability of censoring weighting estimators. Since nonparametric density estimation is closely related to nonparametric estimation of the regression function we can effectively use the methods of the previous section.

Recall the regression problem of chapter 2 and suppose that $U = (T, Z)$ is a random vector with values in $(\mathbb{R} \times \mathbb{R}^k)$ and joint distribution Q. Again we denote F_z for the conditional distribution function of T given Z and H for the marginal distribution of Z. The relation $Q = F_z \times H$ shows that any model for Q can be indexed by models for F_z and H. Since nonparametric regression estimation is in some sense equivalent to nonparametric density estimation we shall assume the same amount of smoothness for the underlying conditional distribution function as in the previous section: we assume submodels $\mathcal{Q} = \mathcal{Q}_1 \times \mathcal{Q}_2$ of $\mathcal{Q}^\alpha = \mathcal{Q}_1^\alpha \times \mathcal{Q}_2$, where for some $\alpha > 1/4$ and some essentially bounded function $g \in \mathcal{L}_2(H)$

$$\mathcal{Q}_1^\alpha = \{F_z : |F(t \mid z + \xi) - F(t \mid z)| \leq g(z)|z|^\alpha\},$$
$$\mathcal{Q}_2 = \{H : H \text{ is a continuous probability distribution on } \mathbb{R}^k\}.$$

We assume that $\mathcal{Q}_1$ is nonparametric in the sense that it includes all conditional distributions that are infinitely often differentiable in the conditioning argument.

Note that we do not assume a density for Z. If the vector Z is discret, or has discrete components then a stratified empirical distribution could be used for estimating F_z. Although this case is not treated in an explicit way in this section, for practical purposes it seems valuable to note that nearest neighbor type smoothing methods would produce estimates that are equal to the stratified empirical distribution if the sample size gets large.

Tangent spaces for $\mathcal{Q}_1$ at F_z and for $\mathcal{Q}_2$ at H could be approached as in the previous section. Since the infinitely often differentiable functions are dense in $\mathcal{L}_2$ (see e.g. Levit (1978) for a similar argument), we have that

$$\dot{\mathcal{Q}}_1 = \{a \in \mathcal{L}_2^0(F_z \times H) : \mathrm{E}\,(a(T, Z) \mid Z) = 0\}$$

and

$$\dot{\mathcal{Q}}_2 = \mathcal{L}_2^0(H).$$

We then use $\dot{\mathcal{Q}} = \dot{\mathcal{Q}}_1 \times \dot{\mathcal{Q}}_2 = \mathcal{L}_2(F_z \times H)$ as a tangent space for $\mathcal{Q}$.

For $\varphi \in \mathcal{F}$ define the parameter $\psi : \mathcal{Q} \to \mathbb{R}$ by

$$\psi(\varphi, Q) = \int \varphi(t,z)\, \phi(F_z)(t,z)\, Q(\mathrm{d}t,\, \mathrm{d}z). \tag{4.8}$$

Assume that $\phi : (\mathcal{Q}_1, \|\cdot\|_\infty) \to (\mathcal{L}_\infty(Q), \|\cdot\|_\infty)$ is Hadamard differentiable tangentially to the tangent space $\dot{\mathcal{Q}}$ at Q. Then, for any $\epsilon \to 0$ and uniformly for G_z in compact subsets of $\mathcal{Q}_1$,

$$\left\| \frac{\phi(F_z + \epsilon G_z) - \phi(F_z)}{\epsilon} - \dot{\phi}_q(G_z) \right\|_\infty \to 0. \tag{4.9}$$

Parallel to the previous section we need an integral representation of the derivative of the functional ϕ. For simplicity we assume that $k = 1$. In view of applications treated in section 5.3, we concentrate on the case where $\dot{\phi}_F(G_z)$ is atomic in the conditioning argument, and assume the following almost sure representation

$$\dot{\phi}_F(G_z)(t,z) = \int \tilde{\phi}_F(t,z,s)\, G(\mathrm{d}s \mid z),$$

where $s \mapsto \tilde{\phi}_F(t,z,s)$ an integrable function. The function $(t,z) \mapsto \tilde{\phi}_F(t,z,s)$ is essentially bounded since $(t,z) \mapsto \dot{\phi}_F(t,z)$ is. In view of definition 4.2, $\tilde{\phi}_F$ is a generalized influence function of the functional ϕ.

Suppose we are given a nonparametric estimator of the conditional distribution function F_z, e.g. a Nadaraya-Watson type estimator:

$$\hat{F}_{1,n}(t \mid z) = \frac{\int_0^t K_a(z-\xi)\, \hat{Q}_n(\mathrm{d}t,\, \mathrm{d}\xi)}{\int K_a(z-\xi)\, \hat{H}_n(\mathrm{d}\xi)}.$$

Here $\hat{H}_n$ denotes the empirical measure corresponding to $Z_1, \ldots, Z_n$ and K_a is a kernel density function with bandwidth $a(n)$ such as specified in the previous section. Alternatively, a symmetrized nearest neighbor type estimator can be used, such as was introduced by Yang (1981) and studied by Stute (1984, 1986):

$$\hat{F}_{2,n}(t \mid z) = \int_0^t K_a(\hat{H}_n(z) - \hat{H}_n(\xi))\, \hat{Q}_n(\mathrm{d}t,\, \mathrm{d}\xi).$$

One advantage of $\hat{F}_{2,n}$ is that for showing asymptotic normality of the estimator at a fixed covariate value, the distribution of Z need not have a density but smoothness of the function $z \mapsto F(\mathrm{d}t \mid H(z))$ in a neighborhood of z is sufficient (Stute 1984). A comparison of the bias of the estimators $\hat{F}_{1,n}$ and $\hat{F}_{2,n}$ and their mean square error performance can be found in Carroll and Härdle (1989). Other nonparametric methods for estimating conditional distributions are possible, such as ordinary nearest neighbor, spline smoothing or orthogonal

series, see e.g. Prakasa Rao (1983) for an overview. Some of these methods lead to estimators that are special cases of the class proposed by Stone (1977):

$$\hat{F}_{B_n}(t \mid z) = \int_0^t B_n(z,\xi)\, \hat{Q}_n(\,\mathrm{d}s,\, \mathrm{d}\xi), \tag{4.10}$$

where $B_n(z,\xi) = B_n(z,\xi,\hat{H}_n)$ is a random sequence of weights that depend on the Z-sample only. This class of estimators was generalized by Beran (1981) to estimation of a conditional distribution function with randomly censored survival data (see also chapter 5.3). The general class includes the normalized symmetrized nearest neighbor type estimator proposed by Stute (1986), viz.

$$\hat{F}_{3,n}(t \mid z) = \frac{\int_0^t K_a(\hat{H}_n(z) - \hat{H}_n(\xi))\, \hat{Q}_n(\,\mathrm{d}t,\, \mathrm{d}\xi)}{\int K_a(\hat{H}_n(z) - \hat{H}_n(\xi))\, \hat{H}_n(\,\mathrm{d}\xi)}.$$

Here is the analogue to theorem 4.3.

Theorem 4.10
Let $\mathcal{Q}_1 \subseteq \mathcal{Q}_1^\alpha$, let B_n be a sequence of random weights determining a nonparametric estimator $\hat{F}_{B_n}$ of F_z such that the following two conditions hold for every $f \in \mathcal{L}_\infty(Q)$ as $n \to \infty$:

$$\sup_{F_z \in \mathcal{Q}_1} \sup_z \int f(t,z)\, \{\hat{F}_{B_n}(\,\mathrm{d}t \mid z) - F(\,\mathrm{d}t \mid z)\} \to 0$$

$$\sup_{F_z \in \mathcal{Q}_1} \sup_z \sqrt{n} \int f(t,z)\, \{B_n(z,\xi)\, Q(\,\mathrm{d}t,\, \mathrm{d}\xi) - F(\,\mathrm{d}t \mid z)\} \to 0.$$

Let $\phi : (\mathcal{Q}_1, \|\cdot\|_\infty) \to (\mathcal{L}_\infty(Q), \|\cdot\|_\infty)$ be Hadamard differentiable tangentially to $\dot{\mathcal{Q}}_1$ such that the derivative ϕ_F at F_z in direction $G_z \in \mathcal{Q}_1$ can be represented as

$$\dot{\phi}_F(G_z)(t,z) = \int \tilde{\phi}_F(t,z,s)\, G(\,\mathrm{d}s \mid z),$$

where $\tilde{\phi}_F \in \dot{\mathcal{Q}}_1$. Then the plug-in estimator

$$\hat{\psi}_n \equiv \int \varphi(t,z)\, \phi(\hat{F}_{B_n})\, \hat{Q}_n(\,\mathrm{d}t,\, \mathrm{d}z)$$

efficiently estimates the functional ψ defined in (4.8) in the model $\mathcal{Q}$. Moreover,

$$\sqrt{n}(\hat{\psi}_n - \psi) \Rightarrow \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \mathrm{E}\, \tilde{\psi}(T,Z)^2$ is the infomation bound and

$$\tilde{\psi}(T,Z) = \varphi(T,Z)\, \phi(F_z)(T,Z) - \nu(\varphi,Q) + \int \varphi(t,Z)\, \tilde{\phi}_F(t,Z,T)\, F(\,\mathrm{d}t \mid Z) \tag{4.11}$$

the efficient influence function of ψ.

Proof: Expand the centered and scaled estimator as follows

$$\begin{aligned}\sqrt{n}(\hat{\psi}_n - \psi) = \sqrt{n} \int \varphi(t,z)\,\{\phi(\hat{F}_{B_n}) - \phi(F_z)\}(t,z)\,\hat{Q}_n(\,\mathrm{d}t,\,\mathrm{d}z) \\ + \int \varphi(t,z)\,\phi(F_z)(t,z)\,\sqrt{n}(\hat{Q}_n - Q)(\,\mathrm{d}t,\,\mathrm{d}z).\end{aligned} \tag{4.12}$$

The second term on the right hand side of (4.12) is a quasi U-statistic to which the central limit theorem applies directly.
By Hadamard differentiability of ϕ the first term on the right hand side of (4.12) equals

$$\begin{aligned}&\sqrt{n} \int \varphi(t,z)\,\dot{\phi}_F(\hat{F}_{B_n} - F_z)(t,z)\,\hat{Q}_n(\,\mathrm{d}t,\,\mathrm{d}z) + o_P(1) \\ &= \sqrt{n} \iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s)\,\{\hat{F}_{B_n}(\,\mathrm{d}s \mid z) - F(\,\mathrm{d}s \mid z)\}\,\hat{Q}_n(\,\mathrm{d}t,\,\mathrm{d}z) + o_P(1) \\ &= \sqrt{n} \iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s)\,\{\hat{F}_{B_n}(\,\mathrm{d}s \mid z) - F(\,\mathrm{d}s \mid z)\}\,\{\hat{Q}_n(\,\mathrm{d}t,\,\mathrm{d}z) - Q(\,\mathrm{d}t,\,\mathrm{d}z)\} \\ &\quad + \sqrt{n} \iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s)\,\{B_n(z,\xi)\,Q(\,\mathrm{d}s,\,\mathrm{d}\xi) - F(\,\mathrm{d}s \mid z)\}\,Q(\,\mathrm{d}t,\,\mathrm{d}z) \\ &\quad + \sqrt{n} \iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s)\,B_n(z,\xi)\,Q(\,\mathrm{d}t,\,\mathrm{d}z) \\ &\qquad\qquad \times \{\hat{Q}_n(\,\mathrm{d}s,\,\mathrm{d}\xi) - Q(\,\mathrm{d}s,\,\mathrm{d}\xi)\} + o_P(1).\end{aligned} \tag{4.13}$$

The first term on the very right hand side of (4.13) is bounded above by

$$\sup_{t,z} \left\{ \int \tilde{\phi}_F(t,z,s)\,\{\hat{F}_n(\,\mathrm{d}s \mid z) - F(\,\mathrm{d}s \mid z)\} \right\} \sqrt{n} \int \varphi(t,z)\,\{\hat{Q}_n(\,\mathrm{d}t,\,\mathrm{d}z) - Q(\,\mathrm{d}t,\,\mathrm{d}z)\}. \tag{4.14}$$

Since $\varphi \in \mathcal{L}_2(Q)$, the integral term is asymptotically tight by the central limit theorem, thus by the first condition on B_n the term in (4.14) is $o_P(1)$. The second term on the very right hand side of (4.13) is bounded above by

$$\begin{aligned}\sup_{t,z} \sqrt{n} \int \tilde{\phi}_F(t,z,s)\,\{B_n(z,\xi)\,Q(\,\mathrm{d}s,\,\mathrm{d}\xi) - F(\,\mathrm{d}s \mid z)\} \\ \times \int \varphi(t,z)\,\{\hat{Q}_n(\,\mathrm{d}t,\,\mathrm{d}z) - Q(\,\mathrm{d}t,\,\mathrm{d}z)\}.\end{aligned}$$

By the second condition on B_n this term in the last preceding display is $o_P(1)$. Finally, by similar arguments, the third term on the very right hand side of (4.13)

can be expanded to

$$\begin{aligned}
&\sqrt{n}\iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s)\{B_n(z,\xi)\,Q(\,\mathrm{d}t,\,\mathrm{d}z) - F(\,\mathrm{d}t \mid \xi)\} \\
&\qquad\qquad \times \{\hat{Q}_n(\,\mathrm{d}s,\,\mathrm{d}\xi) - Q(\,\mathrm{d}s,\,\mathrm{d}\xi)\} \\
&\quad + \iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s) F(\,\mathrm{d}t \mid \xi)\,\sqrt{n}\{\hat{Q}_n(\,\mathrm{d}s,\,\mathrm{d}\xi) - Q(\,\mathrm{d}s,\,\mathrm{d}\xi)\} \\
&= \iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s) F(\,\mathrm{d}t \mid \xi)\,\sqrt{n}\{\hat{Q}_n(\,\mathrm{d}s,\,\mathrm{d}\xi) - Q(\,\mathrm{d}s,\,\mathrm{d}\xi)\} + o_P(1) \\
&= \iint \varphi(t,z)\,\tilde{\phi}_F(t,z,s) F(\,\mathrm{d}t \mid \xi)\,\sqrt{n}\hat{Q}_n(\,\mathrm{d}s,\,\mathrm{d}\xi) + o_P(1),
\end{aligned}$$

where the last step follows since $\tilde{\phi}_F(t,z,\cdot) \in \dot{\mathcal{Q}}_1$, i.e. $\tilde{\phi}_F(t,z,T)$ has conditional expectation given Z equal to zero almost surely. Combination of the last term with the second term on the right hand side of (4.12) yields the expansion

$$\begin{aligned}
\sqrt{n}(\hat{\psi}_n - \psi) = \sqrt{n}\int \Bigg[&\varphi(t,z)\ \{\phi(F_z)(t,z)\ - \nu(\varphi,Q)\} \\
&+ \int \varphi(s,z)\,\tilde{\phi}_F(s,z,t)\,F(\,\mathrm{d}s \mid z)\Bigg]\ \hat{Q}_n(\,\mathrm{d}t,\,\mathrm{d}z) \\
&+ o_P(1).
\end{aligned}$$

Thus, the influence function is as given in the theorem.
□

Remark 4.11

Instead of integrating with respect to the empirical distribution function one could asymptotically equivalently use a smoothed empirical distribution (4.8). However, we prefer the estimator given in theorem 4.10, in view of the special characteristics of the parameters in this thesis; that is, we prefer estimates for measures of prediction error that are quasi means of independent random variables.

Example 4.12 (Integral of a squared regression function)
The regression function is defined as the first moment of the conditional distribution of T given Z: $m(z) = \mathrm{E}\,(T \mid Z = z)$. A familiar example of the functional in (4.8) arises by setting

$$\phi(F_z)(t,z) = m(z)^2 = \left\{\int F(\,\mathrm{d}s \mid z)\right\}^2$$

and $\varphi(t,z) = \varphi(z)$:

$$\psi(\varphi, m) = \int \varphi(z)\,m(z)^2\,H(\,\mathrm{d}z).$$

The derivative of ϕ at F_z can be represented by the generalized influence function

$$\tilde{\phi}_F(T, Z) = 2\left\{T\, m(Z) - m^2(Z)\right\}.$$

The efficient influence function is thus obtained by substituting into (4.11):

$$\begin{aligned}\tilde{\psi}(T, Z) &= \varphi(Z)\, m(Z)^2 - \psi(\varphi, m) + 2\left\{T\, m(Z) - m^2(Z)\right\} \\ &= \left\{2\, T\, m(Z) - m^2(Z)\right\} - \psi(\varphi, m).\end{aligned}$$

This functional occurs in the asymptotics of nonparametric coefficients of determination (Doksum and Samarov 1995) and is important for the choice of the smoothing parameters in nonparametric regression.

Chapter 5

Efficient estimation of prediction error with incomplete data

This chapter deals with efficient nonparametric functional estimation based on incomplete data. In particular, we provide efficient estimators for parameters that represent the prediction error of a regression model such as discussed in chapter 2.

In section 5.1 we define measures of prediction error that are identifiable as population parameters in general incomplete data situations. A necessary condition for estimation of such parameters is that the inverse probability of censoring function is itself identifiable from the coarsened data and bounded away from zero on a substantial part of the range of the outcome variable (section 5.2). We suggest to use prediction error curves (see chapter 2) restricted to the identifiable range of the outcome. In addition, weighted summary measures of the restricted prediction error curve can serve as real valued measures of prediction error. By using the results of chapters 3 and 4 we can establish efficient IPCW estimates in the example of right censored survival data in presence of completely observed covariates (section 5.3).

5.1 Definition of prediction error for coarsened data

For convenience of the reader we shall briefly recall the relevant notation introduced in the earlier chapters. Let (T, Z) be the dependent variable and a vector of covariates with values in $(\mathbb{R} \times \mathbb{R}^k)$; let X be a coarsening of (T, Z), which is a random map that takes values in a Borel subset of the σ-field of the range of (T,Z) and $P((T, Z) \in X) = 1$.

Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ be nonparametric models for the conditional distribution F_z of T given Z and the marginal law H of Z, respectively, such that $\mathcal{Q} = \mathcal{Q}_1 \times \mathcal{Q}_2$ consists of all dominated probability distributions on $(\mathbb{R}, \mathbb{R}^k)$. The conditional distribution of X given (T, Z) is denoted by R. We assume that X is a coarsening

at random: $R \in \mathcal{R} \subseteq \mathcal{R}_{\mathrm{CAR}}$ (see section 3.3). For fixed $Q \in \mathcal{Q}$ and $R \in \mathcal{R}$ the induced distribution of X is determined by:

$$\int h(x)\, W_{Q,R}(\,\mathrm{d}x) = \iint h(x)\, R(\,\mathrm{d}x \mid u)\, Q(\,\mathrm{d}u) = \int h(x)\, r(x \mid u)\, \eta(\,\mathrm{d}x)\, Q(\,\mathrm{d}u)$$

where r is defined in (3.14), η dominates $\mathcal{R}$ and the equality holds for bounded Borel functions h, say. This gives rise to a model for the distribution of X:

$$\mathcal{W} \subseteq \mathcal{W}_{\mathrm{CAR}} = \{W_{Q,R} = R \times Q :\, Q \in \mathcal{Q}, R \in \mathcal{R}_{\mathrm{CAR}}\}.$$

Recall that by lemma 3.15 the tangent space of of $\mathcal{W}_{\mathrm{CAR}}$ at $W_{Q,R}$ equals $\mathcal{L}_2^0(W_{Q,R})$. We assume throughout this section that $\mathcal{W}$ is nonparametric, in the sense (of Groeneboom and Wellner (1994)) that is fulfilled if and only if

$$\dot{\mathcal{W}} = \mathcal{L}_2^0(W_{Q,R}).$$

The indicator of complete observations is assumed to be a deterministic function of X, i.e. observing X includes knowledge on whether the observation is coarsened or not. Let $W^{(1)} = W^{(1)}_{Q,R}$ and $W^{(0)} = W^{(0)}_{Q,R}$ denote the subdistributions of X corresponding to the events $\Delta = \Delta(X) = 1$ and $\Delta = \Delta(X) = 0$, respectively. Throughout this chapter we use simplified notation for values of X that are singletons, i.e. we write $X = (T, Z)$ instead of $X = (\{T\}, \{Z\})$ and $R((t \mid z) \mid (t, z))$ for the conditional probability of the event $X = (\{t\}, \{z\})$ given $T = t$ and $Z = z$.

A central concept is the inverse probability of censoring function which is now given by

$$d(t, z) = d(R)(t, z) = P(\Delta = 1 \mid T = t, Z = z) = R((t \mid z), (t, z)).$$

We view d as a functional on $\mathcal{R}$ with values in $(\mathbb{R}, \mathbb{R}^k)$ and assume that $d(R)$ is identifiable on $\mathcal{W}$. Then there exists a sequence of estimators $\hat{d}_n$ that is uniformly consistent for $d(R)$ (compare definition 3.5). Note that we suppress the dependence of d on R in the notation whenever possible in the sequel.

By using the relation $W^{(1)}(\,\mathrm{d}x) = R(x \mid x)\, Q(\,\mathrm{d}x) = d(x)\, Q(\,\mathrm{d}x)$, and according to proposition 3.6 we have that the parameter $\nu(W_{Q,R}) = \psi(Q) = Q\varphi$ is identifiable in $\mathcal{W}$ if

$$\int \varphi(t, z)\, Q(\,\mathrm{d}t,\, \mathrm{d}z) = \int \varphi(x)\, 1\{d(x) > 0\}\, \frac{W^{(1)}(\,\mathrm{d}x)}{d(x)} < \infty.$$

It is now obvious that the mean squared error,

$$\mathrm{MSE} = \int (s - \int t\, \pi(\,\mathrm{d}t \mid z))^2\, Q(\,\mathrm{d}s,\, \mathrm{d}z),$$

is identifiable only if $d(T, Z) > 0$ almost surely with respect to Q. This condition fails, for instance, when a positive event time is the outcome of interest and one has to deal with administrative end of study.

We want to explore what measures of prediction error are identifiable if $d(t, Z) > 0$ almost surely, restricted to t in a proper subset of the range of T.

Definition 5.1 (Abstract prediction error for incomplete data) *Suppose d is identifiable on $\mathcal{W}$. Let S be a scoring rule, and α a deterministic function of T, and π a forecast conditional probability. Set*

$$S(t, z) \equiv S(\alpha(t), \pi_z(\alpha)),$$

where $\pi_Z(\alpha) = \int \alpha(t)\, \pi(\mathrm{d}t, Z)$. Let S, α be such that he prediction error of π with respect to S and α is well-defined via $\nu(\alpha, S) : \mathcal{W} \mapsto \mathbb{R}$:

$$\nu(W_{Q,R}) = \int S(y, z) \frac{W^{(1)}(\mathrm{d}y, \mathrm{d}z)}{d(y, z)} < \infty.$$

□

We have to find aspects of prediction error that are supported on the set where $d > 0$. Here is a special case, meeting in particular the conditions of our main example, where the vector of covariates is always observed. Suppose $M(Z)$ is almost surely greater than zero and define

$$\mathcal{I}_M = \{t : d(t, Z) > M(Z) > 0 \text{ a.s.}\},$$

According to proposition 3.6, if d is identifiable, restricting the prediction error curves defined in chapter 2 to $\mathcal{I}_M$ provides a flexible measure of prediction error.

Definition 5.2 (Prediction error curves for incomplete data) *Let S=S_{BS} be the Brier score; the prediction error curve $t \mapsto \mathrm{PEC}(t)$ is defined on $\mathcal{I}_M$ by*

$$\mathrm{PEC}(t) = \int \{1\{y > t\} - \pi_z((t, \infty))\}^2 \frac{W^{(1)}(\mathrm{d}y, \mathrm{d}z)}{d(y, z)}.$$

Note that the present definition of prediction error requires also that the forecast prediction π is defined on $\mathcal{I}_M$; otherwise we have to intersect $\mathcal{I}_M$ with the set where π yields valid predictions.

A class of real valued summary measures of the prediction error of π is obtained by cumulating the prediction error curve suitably weighted to meet the problem at hand. Suppose ω is such a weighting scheme with support on $\mathcal{I}_M$,

then define the weighted prediction error in accordance with chapter 2 by

$$\begin{aligned}\mathrm{WPE} &\equiv \int \mathrm{PEC}(t)\,\omega(\,\mathrm{d}t)\\ &= \int\int (1\{y>t\} - \pi_z((t,\infty)))^2 \frac{W^{(1)}(\,\mathrm{d}y,\,\mathrm{d}z)}{d(y,z)}\,\omega(\,\mathrm{d}t)\\ &= \int\int (1\{y>t\} - \pi_z((t,\infty)))^2\,\omega(\,\mathrm{d}t)\frac{W^{(1)}(\,\mathrm{d}y,\,\mathrm{d}z)}{d(y,z)}.\end{aligned}$$

The very right hand side of the preceding display shows that WPE fits also in the abstract form of definition 5.1 (by setting $S = \int S_{BS}\,\mathrm{d}\omega$). In case of uniform weights Graf, Schmoor, and Schumacher (1999) called WPE integrated Brier score.

5.2 Efficient estimation of inverse probability of censoring functionals

In this section we shall (heuristically) discuss how to establish efficient estimators of smooth parameters for general CAR situations. Similar methods have been used by Hubbard, Van der Laan, and Robins (1998) and can be found in Van der Vaart (1998, section 25.5.3). Throughout this section we use the notation

$$S(t,z) = S(\alpha(t), \pi_z(\alpha)),$$

for a deterministic function α of T and for a forecast conditional probability π.

Recall that the efficient influence function of the parameter $\psi(Q) = QS$ is given by $S(T,Z) - \psi(Q)$. We want to find the efficient influence function of $\nu(W_{Q,R}) = \psi(Q)$. Suppose first that R is known which implies that the inverse probability of censoring function is known. Then the tangent set for the model $\mathcal{W}$ for the distribution of X consists only of the range of the score operator $\dot{l}_1$ defined in section 3.2. In this case the efficient influence function for estimation of the parameter

$$\nu(W_{Q,R}) = \psi(Q) = \int S(t,z)\,Q(\,\mathrm{d}t,\,\mathrm{d}z)$$

is given by

$$\begin{aligned}\tilde{\nu}(X) &= \frac{\Delta(X)}{d(T,Z)}S(T,Z) - \int\left\{\frac{\Delta(X)}{d(T,Z)}S(T,Z)\right\}\\ &= \frac{\Delta(X)}{d(T,Z)}S(T,Z) - \nu(W_{Q,R}).\end{aligned} \tag{5.1}$$

Since R is known we have $\dot{l}_2^* = 0$ and thus it is sufficient to show that $\tilde{\nu}$ is in the range of $\dot{l}_1^*$:

$$\begin{aligned}\dot{l}_1^*(\tilde{\nu}(X)) &= \mathrm{E}\left(\tilde{\nu}(X) \mid T, Z\right) - \nu(W_{Q,R}) \\ &= \frac{\mathrm{E}\left(\Delta(X) \mid T, Z\right)}{d(T,Z)} S(T,Z) - \nu(W_{Q,R}) \\ &= S(T,Z) - \psi(Q).\end{aligned}$$

In applications, however, R is typically unknown and has to be estimated from the data. Hence, identifiability of the function d is a necessary condition for identifiability of $\nu = \psi$ in presence of the nuisance parameter d. The function $\tilde{\nu}$ defined in (5.1) remains an influence function if R is not known but typically looses the attribute efficient.

It is well-known that the efficient influence function is obtained as the projection of an influence function onto the orthogonal complement of the tangent space for the nuisance parameter (c.f. theorem 3.8). However, it is in general hard to explicitly compute the relevant projection (Van der Vaart 1998, lemma 25.41). For multivariate right censored data Hubbard, Van der Laan, and Robins (1998, section 3.1) have directly computed the projection of an initial influence function. In section 3.4 we have derived the spectral decomposition of the involved information operator by tools from functional analysis, and established the efficient influence function for functionals that include the one analyzed by Hubbard, Van der Laan, and Robins (1998), see in particular the results of section 5.3.

In the rest of this section we want to guess the functional form of the efficient influence function in the general coarsened data situation. Suppose that d is identifiable on $\mathcal{W}$ and that there exists a Hadamard differentiable functional (tangentially to $\dot{\mathcal{W}}$) $\phi : \mathcal{W} \to \mathbb{R} \times \mathbb{R}^k$ such that

$$\phi(W_{Q,R})(t,z) = \frac{1}{d(R)(t,z)}.$$

Then the functional

$$\nu(W_{Q,R}) = \int S(x)\, \Delta(x) \phi(W_{Q,R})(x)\, W_{Q,R}(\,\mathrm{d}x)$$

is precisely of the form treated in section 4.2 (we see this by setting $\varphi(x) = \Delta(x)\, S(x)$) in theorem 4.10. Assume that the efficient influence function for estimation of $\phi(W_{Q,R})$ is given by $\tilde{\phi}_{W_{Q,R}}(T,Z)$, and that there exists an efficient estimator $\hat{\phi}_n$ of $\phi(W_{Q,R})$. In view of the results of Van der Vaart (1988), see in particular (Van der Vaart 1998, theorem 25.48) we conjecture that the plug-in estimator

$$\hat{\nu}_n = \int \varphi(x)\, \Delta(x) \hat{\phi}_n(x)\, \hat{W}_n(\,\mathrm{d}x)$$

is efficient for $\nu(W_{Q,R})$. Here $\hat{W}_n$ is the empirical distribution of an *iid* sample $X_1, \ldots, X_n$. By comparison with the influence function found in theorem 4.10 for a similar estimator we can guess the functional form of the efficient influence function:

$$\tilde{\nu}(X) = \frac{\Delta(X)\, S(T,Z)}{d(T,Z)} - \nu(W_{Q,R}) + \int S(x)\, \tilde{\phi}_{W_{Q,R}}(x) \Delta(x)\, W_{Q,R}(\,\mathrm{d}x).$$

The function in the preceding display is the projection of the initial influence function given in (5.1) onto the orthogonal complement of the tangent space for the nuisance parameter d. We have thus expressed the asymptotic distribution of IPCW estimators as a function of the efficient (generalized) influence function for estimation of the inverse probability of censoring function. See section 4.2 for an example where $\tilde{\phi}_F$ is a generalized function (distribution).

5.3 Prediction error for right censored event times

In the right censoring situation with completely observable covariates of section 3.4 we use the methods developed in chapter 4 to show that certain plug-in estimators are asymptotically efficient. A nonparametric model and a semiparametric model and correspondingly defined efficient estimators are discussed, namely, under conditionally independent censoring given the covariates a plug-in estimator for the conditional censoring distribution (Akritas 1994), and, under independent censoring of the covariates and the event time, the estimator of Stute (1996) and Graf, Schmoor, and Schumacher (1999).

We shall briefly recall the notation of section 3.4. Set $X = (Y, \Delta, Z)$, where $Y = T \wedge C$ and $\Delta = 1\{Y = T\}$. We let F_z, G_z and H denote the conditional distribution functions of T given Z and C given Z and the marginal distribution of Z, respectively. In what follows we denote $W \equiv W_{F_z,H,G_z}$ for the distribution of X, i.e. supressing the dependence on F_z,G_z and H. Furthermore, we write W_z for the conditional distribution functions of (Y, Δ) given Z, and define by $W_z^{(1)}$ and $W_z^{(2)}$ the conditional sub-distribution functions of ΔY and $(1-\Delta)Y$ given Z, respectively. Throughout this section we assume that X satisfies CAR, alternatively, that C is conditionally independent of T (see example 3.14). The density of W is given by

$$W(\,\mathrm{d}y, \delta, \,\mathrm{d}z) = \{(1 - G(y \mid z))\, F(\,\mathrm{d}y \mid z)\, H(\,\mathrm{d}z)\}^{\delta} \times \{(1 - F(y \mid z))\, G(\,\mathrm{d}y \mid z)\, H(\,\mathrm{d}z)\}^{(1-\delta)}$$

Let $\mathcal{W}_{\mathrm{CAR}}$ be the model for the distribution of X corresponding to the collection of all density functions so obtained.

The inverse probability of censoring function is now $d(t,z) = (1 - G(t \mid z))$ and the function d is identifiable via the relation

$$(1 - G(t \mid z)) = \exp\left\{ - \int_0^t \frac{W^{(0)}(\,\mathrm{d}s \mid z)}{(1 - W(s \mid z))} \right\}.$$

We will consider estimation of $(1 - G(t \mid z))$ on $[0, \tau(z) - M(z)]$, where

$$\tau(z) \equiv \inf_t \{P(Y \leq t \mid z) = 1\},$$

and for a deterministic function $z \mapsto M(z)$ that is bounded away from zero. It is important to emphasize that we do not assume that G is well-behaved near $\tau(z)$, with respect to Q. Rather we constrain the functions $S(t,z)$ occurring in our definition of prediction error to be almost surely zero on $[\tau(z) - M(z); \infty]$. Then the inverse probability of censoring function $1/(1 - G(t \mid z))$ is essentially bounded with respect to the measure $S(t,z)\, W^{(1)}(\,\mathrm{d}t,\, \mathrm{d}z)$.

Straightforward computation (parallel to the univariate situation described e.g. in BKRW (page 374)) yields that $(1 - G_z)$ is Hadamard differentiable with (generalized) influence function given by

$$-(1 - G(t \mid Z)) \left\{ \frac{(1 - \Delta)\, 1\{Y \leq t\}}{(1 - F(Y \mid Z))} - C_2(Y \wedge t, \mid Z) \right\}$$

where

$$C_2(y \mid z) = \int_0^y \frac{W^{(0)}(\,\mathrm{d}s \mid z)}{(1 - W(s \mid z))^2} = \int_0^y \frac{1}{(1 - W(s \mid z))} \frac{G(\,\mathrm{d}s \mid z)}{(1 - G(s- \mid z))} \qquad (5.2)$$

and $(1 - W(s \mid z)) = (1 - W^{(1)}(s \mid z) - W^{(0)}(s \mid z))$. Note that the conditional distribution function G_z is typically not nonparametrically estimable at rate $\sqrt{n}$. By interchanging the roles of T and C and setting $\varphi(y,z) = 1\{y \leq t\}$ in (3.32) we obtain the (generalized) influence function of $(1 - G(t \mid z))$ as given above (after several applications of integration by part).

To achieve $\sqrt{n}$ convergence rate of nonparametric estimators of integrated functionals of G_z (see chapter 4) a Hölder condition of level $\alpha > 1/4$ of the underlying conditional distribution function is needed (Ritov and Bickel 1990). Therefore, let $\mathcal{W} = \mathcal{W}_1 \times \mathcal{W}_2 \subseteq \mathcal{W}_{\mathrm{CAR}}$, where for $\alpha > 1/4$ and some essentially bounded function g,

$$\mathcal{W}_1 \subseteq \{W_z = W_z^{(1)} + W_z^{(2)} : |W_z^{\delta}(t \mid z + \xi) - W_z^{\delta}(t \mid z)| \leq g(z)|z|^{\alpha}, \delta = 1, 2\}$$

and

$$\mathcal{W}_2 \subseteq \{H : H \text{ is a continuous probability distribution on } \mathbb{R}^k\}.$$

We assume that $\mathcal{W}_1$ includes all conditional distribution functions that are infinitely often differentiable in the conditioning argument. The corresponding tangent spaces for $\mathcal{W}_1$ at W_z and $\mathcal{W}_2$ at H are given respectively by

$$\dot{\mathcal{W}}_1 = \mathcal{L}_2^0(W_z) = \mathcal{L}_2^0(W_z^{(1)}) \oplus \mathcal{L}_2^0(W_z^{(2)})$$
$$\dot{\mathcal{W}}_2 = \mathcal{L}_2^0(H).$$

In what follows we constrain the function $t \mapsto S(t,z)$ to be zero for $t > \tau(z) - M(z)$ and a strictly positive function $z \mapsto M(z)$. Then the parameter ϕ defined by $\phi : \mathcal{W}_1 \to \mathcal{L}_\infty(S\,W^{(1)})$

$$\phi(W_z)(t,z) \equiv \frac{1}{(1-G(t \mid z))} = \exp\left\{\int_0^t \frac{W^{(0)}(\,\mathrm{d}s \mid z)}{(1-W(s \mid z))}\right\}$$

is bounded. ϕ is also Hadamard differentiable tangentially to $\dot{\mathcal{W}}_1$ with generalized influence function

$$\tilde{\phi}_{W_z}(Y,\Delta,Z,t) = \frac{1}{(1-G(t \mid Z))}\left\{\frac{(1-\Delta)\,1\{Y \le t\}}{(1-W(Y \mid Z))} - C_2(Y \wedge t \mid Z)\right\}. \quad (5.3)$$

We can apply proposition 3.6 to show that measures of prediction error such as defined in definition 5.1 are identifiable:

$$\nu(S,\alpha,W) = \int S(t,z)\,1\{\phi(W_z)(t,z) > 0\}\,\phi(W_z)(t,z)\,W(\,\mathrm{d}t, 1, \,\mathrm{d}z). \quad (5.4)$$

This is due to the fact that under CAR

$$W(\,\mathrm{d}t, 1, \,\mathrm{d}z) = (1-G(t \mid z))\,F(\,\mathrm{d}t \mid z)\,H(\,\mathrm{d}z)$$

(compare section 3.4).

We want to apply theorem 4.10; therefore, in the following paragraph, we derive nonparametric estimators for W_z when Z is one-dimensional and continuous. The Beran estimate (Beran 1981) generalizes the class of estimators proposed by Stone (1977) to the right censored situation (compare section 4.2). Several instances have been studied e.g. by Dabrowska (1987).

Let $X_1, \ldots, X_n$ be an *iid* sample and $\hat{H}_n$ the empirical distribution corresponding to the observed covariates $Z_1, \ldots, Z_n$. Then define random weights depending on the sample of the covariates only by $B_n(z, Z_i) = B_n(z, Z_i, \hat{H}_n)$. A class of smooth estimators for the conditional sub-distribution functions can now be obtained by convolution of B_n with the empirical distribution function, viz.

$$\hat{W}_{B_n}^{(1)}(t \mid z) = \sum_{\delta=0,1} \int_0^t \delta\,B_n(z,\xi)\,\hat{W}_n(\,\mathrm{d}s, \delta, \,\mathrm{d}\xi)$$

and

$$\hat{W}^{(2)}_{B_n}(t \mid z) = \sum_{\delta=0,1} \int_0^t (1-\delta)\, B_n(z,\xi)\, \hat{W}_n(\,\mathrm{d}s, \delta, \,\mathrm{d}\xi).$$

Substituting $\hat{W}^{(2)}_{B_n}$ for $W^{(2)}_z$, and the estimate

$$(1 - \hat{W}_{B_n}(t \mid z)) = (1 - \hat{W}^{(1)}_{B_n}(t \mid z) + \hat{W}^{(2)}_{B_n}(t \mid z))$$

for $(1 - W(t \mid z))$ yields the following class of estimators for $(1 - G(t \mid z))$:

$$(1 - \hat{G}_{B_n}(t \mid z)) = \prod_{s \le t} \left\{ 1 - \frac{\hat{W}^{(2)}_{B_n}(\,\mathrm{d}s \mid z)}{(1 - \hat{W}_{B_n}(s \mid z))} \right\}. \tag{5.5}$$

For instance, a symmetrized nearest neighbor type estimator (Stute 1986) is obtained by setting

$$B_n(z) = \int K_a(\hat{H}_n(z) - \hat{H}_n(\xi)),$$

where K_a is a kernel function such as defined in chapter 4 and $a(n)$ is a data dependent sequence of bandwidth. Alternatively, one could use a Nadaraya-Watson type estimator. Akritas (1994) uses Stute's results connected to a rectangular kernel function for estimation of the bivariate survival function with univariate censoring.

The following theorem is suitable for parameters that represent prediction error such as defined in definition 5.1.

Theorem 5.3
Let B_n be a sequence of weights determining estimators $\hat{W}^{(\delta)}_{B_n}$ $(\delta = 1, 2)$ such that for every $f \in \mathcal{L}_2(Q)$

$$\sup_{W_z \in \mathcal{W}_1} \sup_z \sum_{\delta=1,2} \int f(y,z)\, \{\hat{W}^{(\delta)}_{B_n}(\,\mathrm{d}y \mid z) - W^{(\delta)}(\,\mathrm{d}y \mid z)\} \to 0$$

and

$$\sup_{W_z \in \mathcal{W}_1} \sup_z \sqrt{n} \sum_{\delta=1,2} \int f(y,z)\, \{B_n(z,\xi)\, W^{(\delta)}(\,\mathrm{d}y, \,\mathrm{d}\xi) - W(\,\mathrm{d}y \mid z)\} \to 0.$$

Consider a deterministic transformation α of T such that $S(t,z) \equiv S(\alpha(t), \pi_z(\alpha(T)))$ is zero on $[0, \tau(z) - M(z)]$ for almost every z and a given π. Then the IPCW *estimator*

$$\hat{\nu}_n = \int S(t,z)\, \phi(\hat{W}_{B_n})(t,z)\, \hat{W}_n(\,\mathrm{d}t, 1, \,\mathrm{d}z)$$

is asymptotically efficient for

$$\nu(W_{F_z,H,G_z}) = \psi(F_z \times H) = \int S(t,z)\, F(\,\mathrm{d}t \mid z)\, H(\,\mathrm{d}z) \tag{5.6}$$

in the model $\mathcal{W} = \mathcal{W}_1 \times \mathcal{W}_2$. *The efficient influence function for estimation of* $\nu(W)$ *is given by*

$$\begin{aligned}\tilde{\nu}(Y,\Delta,Z) = \frac{\Delta\, S(Y,Z)}{(1-G(Y \mid Z))} - \nu(W) \\ + \frac{(1-\Delta)}{(1-W(Y \mid Z))} \int_Y^\infty S(s,Z)\, F(\,\mathrm{d}s \mid Z) \\ - \int C_2(Y \wedge s \mid Z) S(s,Z)\, F(\,\mathrm{d}s \mid Z),\end{aligned}$$

where C_2 *is given in* (5.2)*; the information bound is given by* $\mathrm{E}\,(\tilde{\nu}(Y,\Delta,Z)^2)$.

Proof: We draw the correspondences

$$\begin{aligned} T &\leftrightarrow (Y,\Delta) \qquad & F_z &\leftrightarrow W_z \\ Q &\leftrightarrow W \qquad & \varphi(Y,\Delta,Z) &\leftrightarrow \Delta\, S(Y,Z), \end{aligned}$$

and then apply theorem 4.10: for any Hadamard differentiable functional ϕ with (generalized) influence function $\tilde{\phi}_{W_z}$ the following expansion holds

$$\begin{aligned}\sqrt{n}(\hat{\nu}_n - \nu) = \sqrt{n} \sum_{\delta=1,2} \int \Big[\delta\, S(y,z)\, \{\phi(W_z)(y,z) - \nu(W)\} \\ + \int S(t,z)\, \tilde{\phi}_{W_z}(t,\delta,z,y)\, W^{(1)}(\,\mathrm{d}t \mid z) \Big]\, \hat{W}_n(\,\mathrm{d}y,\delta,\,\mathrm{d}z) \\ + o_P(1).\end{aligned}$$

We have seen above that $\phi : \mathcal{W}_1 \to \mathcal{L}_\infty(SW)$ defined by $\phi(W_z) = (1-G_z)^{-1}$ is Hadamard differentiable tangentially to $\dot{\mathcal{W}}_1$ with generalized) influence function given in (5.3). Substituting yields

$$\begin{aligned}\sqrt{n}(\hat{\nu}_n - \nu) = \sqrt{n} \sum_{\delta=1,2} \int S(y,z)\, \frac{\hat{W}_n(\,\mathrm{d}y,\delta,\,\mathrm{d}z)}{(1-G(y \mid z))} - \sqrt{n}\,\nu(W) \\ + \sqrt{n} \sum_{\delta=1,2} \iint S(t,z) \left\{ \frac{(1-\delta)\, 1\{y \le t\}}{(1-W(y \mid z))} - C_2(y \wedge t \mid z) \right\} \\ \times \frac{W^{(1)}(\,\mathrm{d}t \mid z)}{(1-G(t \mid z))} \hat{W}_n(\,\mathrm{d}y,\delta,\,\mathrm{d}z) + o_P(1).\end{aligned}$$

Thus, the function $\tilde{\nu}$ is an influence function of ν in $\mathcal{W}$. It follows from equation (3.29) that it is the efficient influence function.
□

Remark 5.4

- If the vector of covariates consists of discrete random variables only, i.e. the probability of observing $Z = z$ is strictly positive for all $z \in \mathbf{R}(Z)$, then a stratified version of the Kaplan-Meier estimator for the censoring distribution can be used for estimation of $G(t \mid z)$. In this case the efficient influence function of the parameter $(1 - G(t \mid z))^{-1}$ is as given in (5.3) and it is obvious that the corresponding plug-in estimator is also efficient.

- Several authors address special cases of our theorem, see in particular Akritas (1994) for estimation of the bivariate survival function and Hubbard, Van der Laan, and Robins (1998) for the marginal survival function in presence of covariates. The bivariate survival function is obtained by setting $S(t, z) = 1\{t \geq s, z \geq \xi\}$, the marginal survival function by setting $S(t, z) = 1\{t \geq s\}$. The efficient influence functions for these two examples are obtained in a simplified form in corollary 5.5 below.

- If K_a is twice continuously differentiable, $a(n) \to 0$, and $na(n)^5 \to \infty$ then the plug-in symmetrized nearest neighbor estimator satisfies the conditions of the theorem (Stute 1986).

So far we have focused on inverse probability of censoring weighted estimators for the parameter $\nu(W)$. Here is another expansion which suggests an alternative estimator:

$$\begin{aligned}
\psi(Q) &= \int S(t,z)\, F(\,\mathrm{d}t \mid z)\, H(\,\mathrm{d}z) \\
&= \int S(t,z)\,(1 - F(t \mid z))\, \frac{Q(\,\mathrm{d}t,\,\mathrm{d}z)}{(1 - F(t \mid z))} \\
&= \int S(t,z)\,(1 - F(t \mid z))\, \frac{W^{(1)}(\,\mathrm{d}t,\,\mathrm{d}z)}{(1 - W(t \mid z))} \\
&= \int S(t,z)\, \exp\left\{ - \int_0^t \frac{W^{(1)}(\,\mathrm{d}s \mid z)}{(1 - W(s \mid z))} \right\} \frac{W^{(1)}(\,\mathrm{d}t,\,\mathrm{d}z)}{(1 - W(t \mid z))} \\
&= \nu(W).
\end{aligned}$$

The plug-in estimator using this second representation of ν is asymptotically equivalent to the IPCW estimator introduced in theorem 5.3. In particular, the estimator of Akritas (1994) for the bivariate survival function is asymptotically equivalent to the corresponding IPCW estimator. If S is of bounded variation it is possible to find an easier form for the efficient influence function of ν. Also the asymptotic variance simplifies considerably. The following result generalizes the formulas of Gill (1983) and Schick, Susarla, and Koul (1988) to the multivariate case.

Corollary 5.5
Suppose $t \mapsto S(t,z)$ is of bounded variation and equals zero almost surely for all $t > \tau(z) - M(z)$ for some strictly positive function $z \mapsto M(z)$. The efficient influence function for estimation of the parameter (5.6) equals

$$\begin{aligned}\tilde{\nu}(Y,\Delta,Z) = \int S(s,Z)\,F(\mathrm{d}s \mid Z) - \nu(W) \\ - \frac{\Delta}{(1-W(Y \mid Z))}\int_Y^\infty (1-F(s \mid Z))\,S(\mathrm{d}s,Z) \\ + \int_0^Y \int_s^\infty (1-F(u \mid Z))\,S(\mathrm{d}u,Z)\,\mathcal{C}_1(\mathrm{d}s \mid Z), \quad (5.7)\end{aligned}$$

where

$$\mathcal{C}_1(y \mid z) = \int_0^y \frac{W^{(1)}(\mathrm{d}s \mid z)}{(1-W(s \mid z))^2} = \int \frac{1}{(1-W(s \mid z))}\,\frac{F(\mathrm{d}s \mid z)}{(1-F(s \mid z))}$$

The information bound for estimation of ν is obtained as the inverse of

$$\begin{aligned}\mathrm{E}\,(\tilde{\nu}^2) = \int \left\{\int S(s,z)F(\mathrm{d}s \mid z)\right\}^2 H(\mathrm{d}z) - \nu(W)^2 \\ + \iint \left\{\int_s^\infty (1-F(u \mid Z))\,S(\mathrm{d}u,Z)\right\}^2 \mathcal{C}_1(\mathrm{d}s \mid z)\,H(\mathrm{d}z).\end{aligned}$$

Proof: The simplified form of $\tilde{\nu}$ follows from the relation

$$C_2(s \mid z) = \frac{1}{(1-W(s \mid z))} - 1 - \mathcal{C}_1(s \mid z)$$

and several applications of integration by parts. Note that the first two terms of the representation (5.7) are asymptotically independent of the last two terms. The variance of the last two terms generalizes the variance formula obtained by Schick, Susarla, and Koul (1988).
□

Example 5.6 (Estimation of the expected Brier score)
Fix some value $t^* \in \mathbf{R}(T)$ such that $(1-G(t^* \mid Z)) > M(Z) > 0$ almost surely. Let π be a set of forecast conditional probabilities, let $S(y,z) = S_{BS}(t^*;y,z)$ be the Brier score at t^*:

$$\begin{aligned}S_{BS}(t^*;y,z) &= \{1\{y > t^*\} - \pi((t^*,\infty) \mid z)\}^2 \\ &= 1\{y > t^*\}\{1 - 2\,\pi((t^*,\infty) \mid z)\} + \pi((t^*,\infty) \mid z)^2.\end{aligned}$$

First we check that the expected Brier score at t^* is identifiable in $\mathcal{W}$:

$$\begin{aligned}\nu(t^*, S_{BS}, W) &= \int\int S_{BS}(t^*; y, z) F(\,\mathrm{d}y \mid z)\, H(\,\mathrm{d}z) \\ &= \int (1 - F(t^* \mid z))\{1 - 2\,\pi((t^*, \infty) \mid z)\}\, H(\,\mathrm{d}z) \\ &\quad + \int \pi((t^*, \infty) \mid z)^2\, H(\,\mathrm{d}z).\end{aligned}$$

Since $(1 - G_z(t^*)) > M(z)$ almost surely, $1 - F(t^* \mid Z)$ is identifiable in $\mathcal{W}$ (see example 3.7).

Clearly the function $y \mapsto S_{BS}(t^*; y, z)$ is of bounded variation:

$$\int (1 - F(y \mid z))\, S_{BS}(t^*;\, \mathrm{d}y, z) = (1 - F(t^* \mid z))\,(1 - 2\,\pi((t^*, \infty) \mid z)).$$

Thus, we can apply corollary 5.5 and obtain that the plug-in IPCW estimator for estimation of the expected Brier score in $\mathcal{W}$ satisfies

$$\sqrt{n}(\hat{\nu}_n(t^*, S_{BS}) - \nu(t^*, S_{BS}, W)) \Rightarrow \mathcal{N}(0, \sigma^2(t^*, S_{BS})).$$

Here the estimator is given by

$$\begin{aligned}\hat{\nu}_n(t^*, S_{BS}) &= \int S_{BS}(t^*; y, z)\, \phi(\hat{W}_{B_n})(y, z)\, \hat{W}_n(\,\mathrm{d}y, 1, \,\mathrm{d}z), \\ &= \int 1\{y > t^*\}(1 - 2\,\pi((t^*, \infty) \mid z))\frac{\hat{W}_n(\,\mathrm{d}y, 1, \,\mathrm{d}z)}{(1 - \hat{G}_{B_n}(y \mid z))} \\ &\quad + \int \pi((t^*, \infty) \mid z)^2\, \hat{H}_n(\,\mathrm{d}z),\end{aligned}$$

$\hat{H}_n$ is the empirical distribution of the covariates only and $\phi(\hat{W}_{B_n}) = (1 - \hat{G}_{B_n})^{-1}$ is a nonparametric estimator of $(1 - G_z)^{-1}$, and the inverse information bound is given by

$$\begin{aligned}\sigma^2(t^*, S_{BS}) = &\int \left\{\pi((t^*, \infty) \mid z)^4 + (1 - F(t^* \mid z))^2\,(1 - 2\pi((t^*, \infty) \mid z))^2\right\} H(\,\mathrm{d}z) \\ &- \nu(t^*, S_{BS}, W)^2 \\ &+ \int (1 - F(t^* \mid z))\,(1 - 2\pi((t^*, \infty) \mid z)) 1\{s < t^*\} \frac{W^{(1)}(\,\mathrm{d}s, \,\mathrm{d}z)}{(1 - W(s \mid z))^2}.\end{aligned}$$

Note that the first term of $\sigma^2(t^*, S_{BS})$ is of similar form as the functional in example (4.12). The asymptotic variance can be asymptotically consistently estimated by

$$\begin{aligned}\hat{\sigma}_n^2(t^*, S_{BS}) = &\int \left\{\pi((t^*, \infty) \mid z)^4 + \hat{F}_{B_n}(t^* \mid z)^2\,(1 - 2\pi((t^*, \infty) \mid z))^2\right\} \hat{H}_n(\,\mathrm{d}z) \\ &- \hat{\nu}_n(t^*, S_{BS})^2 \\ &+ \int (1 - \hat{F}_{B_n}(t^* \mid z))\,(1 - 2\pi((t^*, \infty) \mid z)) 1\{s < t^*\} \frac{\hat{W}_{B_n}(\,\mathrm{d}s, 1, \,\mathrm{d}z)}{(1 - \hat{W}_{B_n}(s \mid z))^2},\end{aligned}$$

where $\hat{F}_{B_n}$ and $\hat{W}_{B_n}$ are nonparametric estimators for F_z and W_z, respectively. □

Example 5.7 (Estimation of WPE**: expected integrated Brier score)** Let ω be a finite measure which is zero on the complement of

$$\mathcal{I}_M = \{t : (1 - G(t \mid Z)) > M(Z) > 0 \text{ a.s.}\}.$$

We want to derive the asymptotics of the IPCW estimator for the parameter WPE, which is defined as the expected value of the scoring rule $S_\omega(y, z) = \int S_{BS}(t; y, z)\, \omega(\mathrm{d}t)$:

$$\begin{aligned}
\text{WPE} &= \iint \{1\{y > t\} - \pi((t, \infty) \mid z)\}^2 \, \omega(\mathrm{d}t)\, F(\mathrm{d}y \mid z)\, H(\mathrm{d}z) \\
&= \iint \pi((t, \infty) \mid z)^2 \, \omega(\mathrm{d}t)\, H(\mathrm{d}z) \\
&\quad + \iint (1 - F(t \mid z))\, (1 - 2\,\pi((t, \infty) \mid z))\, \omega(\mathrm{d}t)\, H(\mathrm{d}z).
\end{aligned}$$

Note that the function

$$y \mapsto S_\omega(y, z) = \int \pi((t, \infty) \mid z)^2 \, \omega(\mathrm{d}t) + \int_0^y (1 - 2\,\pi((t, \infty) \mid z))\, \omega(\mathrm{d}t)$$

is also of bounded variation such that

$$\int (1 - F(y \mid z))\, S_\omega(\mathrm{d}y, z) = \int (1 - F(y \mid z))(1 - 2\,\pi((y, \infty) \mid z))\, \omega(\mathrm{d}y).$$

Thus, we may again apply corollary 5.5 to show that the plug-in IPCW estimator for estimation of WPE in $\mathcal{W}$ satisfies

$$\sqrt{n}(\hat{\nu}_n(S_\omega) - \nu(S_\omega, W)) \Rightarrow \mathcal{N}(0, \sigma^2(S_\omega)),$$

where now

$$\begin{aligned}
\hat{\nu}_n(S_\omega) &= \int S_\omega(y, z)\, \phi(\hat{W}_{B_n})(y, z)\, \hat{W}_n(\mathrm{d}y, 1, \mathrm{d}z), \\
\nu(S_\omega, W) &= \int S_\omega(y, z)\, F(\mathrm{d}y \mid z)\, H(\mathrm{d}z),
\end{aligned}$$

and the inverse information bound is obtained by substituting the corresponding terms into the variance formula of corollary 5.5:

$$\begin{aligned}
\sigma^2(S_\omega) = &\int \left\{ \int \pi((t, \infty) \mid z)^2 + (1 - F(t \mid z))(1 - 2\,\pi((t, \infty) \mid z))\, \omega(\mathrm{d}t) \right\}^2 H(\mathrm{d}z) \\
&- \nu(S_\omega, W)^2 \\
&+ \int \left\{ \int_t^\infty (1 - F(s \mid z))\, (1 - 2\,\pi((s, \infty) \mid z)\, \omega(\mathrm{d}s) \right\}^2 \frac{W^{(1)}(\mathrm{d}t, \mathrm{d}z)}{(1 - W(t \mid z))}.
\end{aligned}$$

As in our previous example the asymptotic variance can be consistently estimated using a nonparametric estimator $\hat{W}_{B_n}$ for plug-in estimation of $(1 - F_z)$ and $(1 - W_z)$, the empirical distribution corresponding to the uncensored observations for $W^{(1)}(\mathrm{d}s, \mathrm{d}z)$ and the empirical distribution corresponding to the Z-sample for H.
□

In the following paragraph we investigate efficient estimation in the semiparametric situation where the censoring variable is stochastically independent of the vector (T,Z). This condition was used by Graf (1998b) and Graf, Schmoor, and Schumacher (1999), and can be seen to be implied by the assumptions of Stute (1993, Stute (1996). However, commonly these authors do not investigate asymptotic efficiency of the proposed estimators. Efficiency of the plug-in estimator with the (marginal) Kaplan-Meier estimator for the censoring distribution is obtained below as a corollary of theorem 5.3.

The induced model for the distribution of X is now obtained by varying $F_z \times H$ over $\mathcal{Q}$ and G over all marginal probability distributions of C. The corresponding tangent space is clearly a proper subspace of the nonparametric tangent space $\mathcal{L}_2^0(W)$. To see this, consider the decomposition $\mathcal{L}_2^0(W) = \mathcal{L}_2^0(W^{(1)}) \oplus \mathcal{L}_2^0(W^{(0)})$ and note that if C is independent of (T, Z) then

$$W^{(0)}(\mathrm{d}t, \mathrm{d}z) = (1 - F(t \mid z))\, G(\mathrm{d}t)\, H(\mathrm{d}z).$$

Clearly $\mathcal{L}_2^0(W^{(1)}) \oplus \mathcal{L}_2^0((1 - F(t \mid z))\, G(\mathrm{d}t)\, H(\mathrm{d}z))$ is a proper subset of $\mathcal{L}_2^0(W)$. This shows that the models considered e.g. by Stute (1996) and Graf, Schmoor, and Schumacher (1999) are semiparametric in the sense of Groeneboom and Wellner (1994, definition 1.1).

In the present situation the censoring survival function (1-G) is identifiable via the relations

$$(1 - W(t)) = \int (1 - F(t \mid z))\, H(\mathrm{d}z)\, (1 - G(t)),$$

and

$$(1 - G(t)) = \exp\left\{ - \iint_0^t \frac{W^{(0)}(\mathrm{d}t, \mathrm{d}z)}{(1 - W(t))} \right\}.$$

We obtain the following representation for $\nu(W_{F_z,H,G_z})$ in the independent censoring model:

$$\nu(W) = \int S(t,z) \exp\left\{ \int_0^t \frac{W^{(0)}(\mathrm{d}s \mid z)}{(1 - W(s \mid z))} \right\} W(\mathrm{d}t, 1, \mathrm{d}z). \tag{5.8}$$

If $(1-\hat{G}_n)$ is the reverse Kaplan-Meier estimator then $1/(1-\hat{G}_n)$ equals the weight that is assigned to each uncensored observation by the Kaplan-Meier estimator

for the marginal survival function of T. This justifies the name Kaplan-Meier integrals for functionals of the type given in (5.8).

We can replace the conditional by the marginal survival function of the censoring variable in all the computations that led to the efficient influence function in section 3.4. In the present situation the efficient influence function for estimation of $\nu(W)$ is thus obtained as

$$\begin{aligned}\tilde{\nu}(y,\delta,z) = \delta\, &\frac{S(y,z)}{(1-G(y))} \\ &+ \frac{(1-\delta)}{(1-W(y))}\int_y^\infty S(s,z)\, F(\,\mathrm{d}s \mid z)\, H(\,\mathrm{d}z) \\ &- \int C_2(y\wedge s) S(s,z)\, F(\,\mathrm{d}s \mid z)\, H(\,\mathrm{d}z) \\ &- \mathrm{E}\,(S(T,Z)),\end{aligned} \tag{5.9}$$

where

$$C_2(y) = \int_0^y \frac{1}{(1-W(s))}\frac{G(\,\mathrm{d}s)}{(1-G(s))}$$

is the asymptotic variance function of the Nelson-Aalen estimator for the (marginal) cumulative hazard function of the censoring variable. The proof of the following corollary follows either from theorem 5.3, noting that the marginal Kaplan-Meier estimator for the censoring distribution satisfies the convergence conditions of the theorem, or directly from the Delta method and its known relation to efficient estimation (Van der Vaart 1998, section 25.7).

Corollary 5.8
Suppose C is independent of (T,Z). Let $(1-\hat{G}_n)$ be the (marginal) Kaplan-Meier estimator for the censoring distribution. The plug-in estimator

$$\hat{\nu}_n = \int S(t,z)\,\frac{\hat{W}_n(\,\mathrm{d}t, 1, \,\mathrm{d}z)}{(1-\hat{G}_n(t))}$$

is efficient for estimation of

$$\nu(W) = \int S(t,z)\,\frac{W^{(1)}(\,\mathrm{d}t, \,\mathrm{d}z)}{(1-G(t))}.$$

The general form of the efficient influence function is given by (5.9).

□

By minor modifications of the examples 5.6 and 5.7 we obtain the asymptotic distribution for the estimators of Graf, Schmoor, and Schumacher (1999). The

estimator for the expected Brier score at t^* is given by

$$\int_0^{t^*} S_{BS}(t^*; y, z)\, \frac{\hat{W}_n(\,\mathrm{d}y, 1, \,\mathrm{d}z)}{(1-\hat{G}_n(y-))} + \sum_\delta \int_{t^*}^\infty S_{BS}(t^*; y, z)\, \frac{\hat{W}_n(\,\mathrm{d}y, \delta, \,\mathrm{d}z)}{(1-\hat{G}_n(t^*))}$$
$$= \int_0^{t^*} \pi((t^*, \infty) \mid z)^2\, \frac{\hat{W}_n(\,\mathrm{d}y, 1, \,\mathrm{d}z)}{(1-\hat{G}_n(y-))} + \sum_\delta \int_{t^*}^\infty \{1-\pi((t^*, \infty) \mid z)\}^2\, \frac{\hat{W}_n(\,\mathrm{d}y, \delta, \,\mathrm{d}z)}{(1-\hat{G}_n(t^*))}$$

where $(1-\hat{G}_n)$ is the Kaplan-Meier estimator for the censoring distribution. The estimator is Gaussian regular and asymptotically efficient with variance given by

$$\sigma^2(t^*, S_{BS}) = \sum_\delta \int \left\{ \delta\, \frac{S_{BS}(t^*; y, z)}{(1-G(y))} + \frac{(1-\delta)}{(1-W(y))} \int_y^\infty S_{BS}(t^*; s, \xi)\, Q(\,\mathrm{d}s, \,\mathrm{d}\xi) \right.$$
$$\left. - \int C_2(y \wedge s) S_{BS}(t^*; s, \xi)\, Q(\,\mathrm{d}s, \,\mathrm{d}\xi) - \nu(W, S_{BS}) \right\}^2 W(\,\mathrm{d}y, \delta, \,\mathrm{d}z).$$

The following worked example illustrates the use of prediction error curves for right censored event times; the curves should be compared to example 2.8.

Example 5.9 (Prediction of event-free survival in breast cancer)
We compute prediction error curves for various predictions (made in terms of forecast conditional probabilities) for event-free survival (first occurrence of either locoregional or distant recurrence, contralateral tumor, secondary tumor or death) in breast cancer with a build and a test data set. The build data set we consider origins from a prospective, controlled multicenter clinical trial on the treatment of node positive breast cancer conducted by the German Breast Cancer Study Group; it will be referred to as GBSG-2-study in the sequel. During six years, 720 patients were recruited of whom about two thirds were randomized. Complete data on the prognostic factors considered were available in 686 patient who form the population of this study. After a median follow-up of about 5 years 299 events for event-free survival were observed. The probability of event-free survival after 5 years was estimated as 50%. The data of this study are available from http://www.blackwellpublishers.co.uk/rss/. The database of the second study consists of all patients with primary, previously untreated breast cancer who were operated between 1982 and 1987 in the Department of Gynecology of the University of Freiburg and who fulfilled some retrospectively defined inclusion criteria e.g. standardized treatment. This left 139 patients out of 218 originally investigated; this study will be referred to as the Freiburg-DNA study in the sequel. Median follow-up was 6.9 years and 76 events with respect to event-free-survival were observed. The probability of being event free after 5 years was estimated as 50%. In both studies, information on the following prognostic factors is available in all patients (see Schumacher, Holländer, and Schwarzer (2001) for details).

Besides age at diagnosis, number of positive lymph nodes and size of the primary tumor, a grading score, as well as estrogen- and progesterone receptor was recorded. Estrogen and progesterone values above 20 fmol/mg cytosol protein were considered positive, negative otherwise. We introduce the following prognostic classification schemes, each determining a set of forecast conditional probabilities π. From the data of the GBSG-2 study, various prognostic classification schemes were derived by means of Cox regression models:

- marginal Kaplan-Meier estimator
- a full Cox model with all six predictors
- a selected Cox model obtained by backward elimination (selection level 5%) containing tumor grade, number of lymph nodes and progesterone receptor
- a misspecified Cox model omitting the number of lymph nodes as the most important predictor from the selected Cox model.

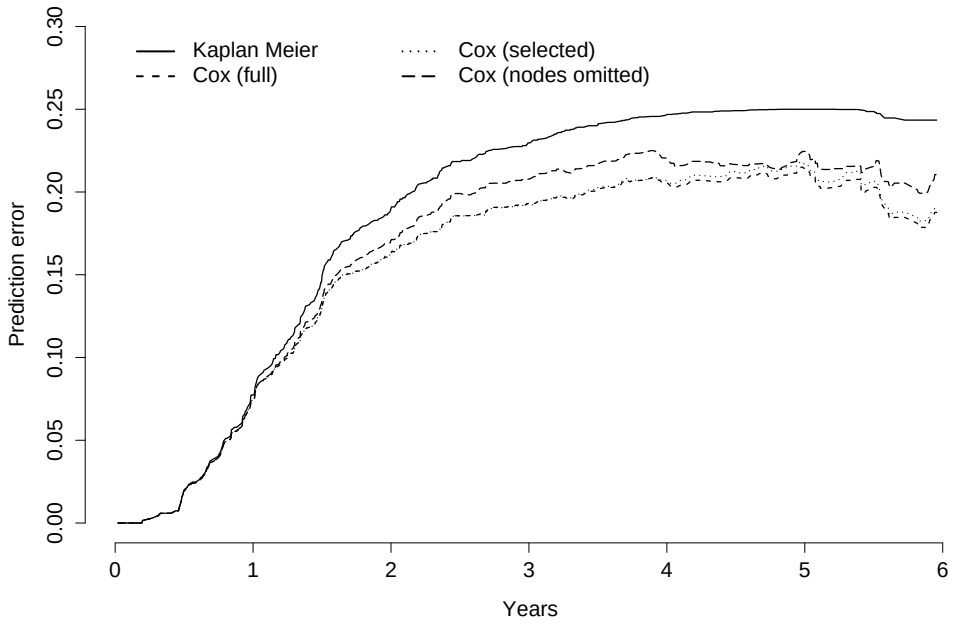

Figure 5.1: Estimated prediction error curves for Cox models with varying covariate settings. The Kaplan-Meier curve yields a benchmark value (null model). GBSG-2 data used for model fit and Freiburg-DNA study for estimation of Brier-Score.

Year	1	2	3	4	5	6
Kaplan Meier	0.023	0.083	0.127	0.155	0.174	0.186
Cox (full)	0.023	0.075	0.110	0.133	0.148	0.157
Cox (selected)	0.023	0.075	0.110	0.133	0.149	0.158
Cox (nodes omitted)	0.023	0.077	0.116	0.141	0.157	0.166

Table 5.1: Estimates of WPE, respectively integrated Brier score of prediction error curves from time zero until years 1 to 6.

Figure 5.1 displays the prediction error curves corresponding to predictions coming from various Cox regression models build in the GBSG-2-study, estimated with the Freiburg-DNA study. The prediction error curves of the full Cox model with all six predictors and of the selected Cox model containing tumor grade, number of lymph nodes and progesterone receptor are nearly identical reflecting the fact that the three other covariates do not exhibit strong effects in the presence of the three selected covariates. If, however, the number of lymph nodes as the most important predictor (relative risk equal to 2 for 4 - 9 and equal to 3.7 for more than 10 nodes) is omitted from the selected Cox model, the prediction error of this misspecified model is grossly inflated and comes closer to that obtained for the naive prediction with the pooled Kaplan-Meier estimate ignoring all covariate information. This reflects that at time zero individual vital status of all patients is equal to one and hence equal to any valid survival distribution function at that time. By symmetry, the expected prediction error curve would decrease to zero for large times. Note that only the first fraction of the true curves can be estimated for right-censored data. Hence the graphs of the curves have to stop at the maximal follow-up time at latest. Using uniform weights on the intervals $[0, i]$, for $i = 1, 2, 3, 4, 5, 6$ years, we have computed WPE for the prediction error curves of figure 5.1 (see table 5.1).

Chapter 6

Discussion

Predictions made in terms of probabilities

At a first glance, it might appear unnatural to use predicted probabilities instead of point predictions. We argue that point predictions can always be obtained from a given forecast conditional probability distribution, namely as the conditional first moment. This conditional first moment is, when it exists, an estimate or prediction of the regression function. Taking advantage of this fact, we find well-known measures for the accuracy of point predictions as special cases of our general definition of prediction error. The following reasoning for assessment of predictions by means of predicted probabilities comes from the theory of elicitation of personal probabilities, see e.g. Savage (1971). A scoring rule is a tool for comparison of observations and predictions. In our regression setting a scoring rule is called proper if it is minimized when the true conditional distribution of the outcome variable is used for establishing predictions. Moreover, assessment of forecast probabilities with a proper scoring rule in terms of expected loss is, to some extend, equivalent to direct assessment of point predictions; the underlying principle is known as encouraging honesty: a forecaster would always quote the probability distribution in which she believes and which she would use to obtain accurate point prediction of the outcome variable. We may conclude that predicted probabilities hold at least the same information as is provided by point predictions and that predictions originating from a forecaster's personal probabilities can be appropriately assessed by using the score function approach.

Confidence intervals and efficient tests

In this thesis we have shown asymptotic efficiency for certain plug-in inverse probability of censoring weighted estimators. In our main example, right censored survival data in presence of covariates, we have generalized results of Gill (1983), Schick, Susarla, and Koul (1988), Akritas (1994), Stute (1996) and Hubbard, Van der Laan, and Robins (1998). However, as far as we know, there is currently no statistical software (SAS, Splus and R) available for computation

of nonparametric estimates of the conditional survival function given a continuous covariate. The main problem here is clearly how to adaptively estimate the smoothing parameters. A task of future research is the construction and implementation of adaptive nonparametric estimates for situations where a nonparametric plug-in estimator of a conditional distribution function is needed.

For applications where two or more predictions have to be compared, the construction of confidence intervals for the estimates of prediction error is important. In view of the explicit formulas obtained for the variances of the estimators proposed in this thesis, see in particular examples 5.6 and 5.7, asymptotically consistent estimation of confidence intervals can be obtained by implementing a nonparametric estimator for the conditional distribution function. A little more involved are confidence bands for function valued parameters, such as the prediction error curves suggested here. One could try to use the results of Van der Vaart (1994) and Rost (2000) to obtain a functional central limit theorem for the estimators of these curves.

It would also be desirable to have statistical tests for the following types of hypothesis:

- The (weighted) prediction error of π_1 is equal to the prediction error of π_2.
- The predictive power of the covariate Z_1 is equal to that of covariate Z_2.

Again the results of this thesis can be used as a basis for the construction of asymptotically efficient tests, see Van der Vaart (1998, section 25.6).

Bootstrap and cross-validation

Throughout this thesis we worked in a build-test data setup – which is clearly not available in most applications. The apparent error problem occurs when the prediction error is estimated with the same dataset which was used for establishing the prediction (Efron 1978). In future research we should be able to take advantage of the well-known correspondence between 'working' bootstrap and differentiable parameters (Gill and der Vaart 1993; Van der Vaart 1998; Wellner 1992). In addition, resampling methods should be applicable for estimation of confidence intervals and the construction of efficient tests. These issues need careful analysis. Of particular importance is a comparison of the small sample performance of bootstrapped confidence intervals, say, and confidence intervals motivated by asymptotic representations.

Sensitivity analysis for CAR models

Our treatment of incomplete data in terms of abstract coarsening at random variables is good for situations where one random variable is coarsened and another random variable is completely observable. There are other examples of incomplete data models where measures of prediction error are needed.

A further aim of future research is a sensitivity analysis of the CAR assumption. A different model could be used (not CAR) for the dependence of the censoring (coarsening) mechanism and the unobservable variables, and then correspondingly defined estimators could be compared to the estimators defined in the CAR-model. An application would be the detection of a competing risk in survival analysis.

Symbol	Regression setting
T	Dependent variable in a regression problem
Z	Vector of covariates
C	Censoring variable
Y	Minimum of T and C
Δ	Indicator of complete observations
X	(Y, Δ, Z)
F_z	Conditional distribution function of T given Z
G_z	Conditional distribution function of C given Z
H	Marginal distribution function of Z
W_{F_z,H,G_z}	Induced distribution of X
W_z	Conditional distribution of X given Z
$\mathcal{W}_1$	Model for the conditional distribution of X given Z
$\mathcal{W}_2$	Model for the marginal distribution of Z
π	Conditional distribution function (predictions made in terms of predicted probabilities)
m	Regression function
$m(\pi)$	Estimated regression function based on π
S	Scoring rule
S_{BS}	Brier score
S_{LS}	Logarithmic score
α	Measurable function on $\mathbf{R}(T)$
$\mathcal{H}$	Class of measurable functions on $\mathbf{R}(T)$
R^2	Measure of explained variation
MSE	Mean squared error
MSEP	Mean squared error of prediction
RSS	Residual sum of squares
PEC	Prediction error curve
WPE	Weighted prediction error
ROC	Receiver operating characteristic
AUC	Area under the ROC-curve

Symbol	Incomplete data setting
U	Unobservable random map
X	Observable random map (coarsening variable)
$(E, \mathcal{E})$	Range and corresponding Borel σ-field of U
Q	Distribution of U
q	Density of Q with respect to μ
$(\mathcal{S}, \Sigma)$	Range and corresponding Borel σ-field of X

R	Conditional probability distribution of X given Z
r	Conditional density of R
$V_{Q,R}$	Induced conditional distribution of U given X
$W = W_{Q,R}$	Induced distribution of X
Δ	Indicator of complete observations
d	Inverse probability of censoring function
IPCW	Inverse probability of censoring weighting
$W^{(1)}, W^{(0)}$	Distributions of ΔX and $(1-\Delta)X$
$\mathcal{Q}$	Dominated model for the probability distribution of U
$\dot{\mathcal{Q}}$	Tangent space of $\mathcal{Q}$
$\mathcal{R}$	Model for the conditional distribution of X given U
$\dot{\mathcal{R}}$	Tangent space of $\mathcal{R}$
CAR	Coarsening at random
MAR	Missing at random
$\mathcal{R}_{\text{CAR}}$	Subset of $\mathcal{R}$ that satisfies CAR
$\mathcal{W}$	Model for W indexed by $\mathcal{Q} \times \mathcal{R}$
$\dot{\mathcal{W}}$	Tangent space for $\mathcal{W}$
$\mathcal{W}_{\text{CAR}}$	Model for W indexed by $\mathcal{Q} \times \mathcal{R}_{\text{CAR}}$
$\dot{l}, \dot{l}_1, \dot{l}_2, \dot{l}_{11}, \dot{l}_{12}$	Score operators
$\dot{l}^*$	Adjoint of the (score) operator $\dot{l}$
$\dot{l}_1^* \dot{l}_1$	Information operator
$(1 - \Pi_2)$	Hilbert space projection onto the orthogonal complement of the tangent space for the nuisance parameter $(\mathbf{R}(\dot{l}_2))$
$\psi, \dot{\psi}, \tilde{\psi}$	Parameter, score function and influence functions corresponding to U
I_ψ^{-1}	Information bound for estimation of ν
$\nu, \dot{\nu}, \tilde{\nu}$	Parameter, score function and influence functions corresponding to X
I_ν^{-1}	Information bound for estimation of ν
$\mathcal{L}_{F_z}, \mathcal{R}_{F_z}$	'L' and 'R' operators of F_z
$\mathcal{L}_{G_z}, \mathcal{R}_{G_z}$	'L' and 'R' operators of G_z

Density and distribution function estimation

Symbol	
$\mathcal{Q}^\alpha$	Model of density functions that satisfy a Hölder condition of degree α
$\hat{q}_n$	Nonparametric density estimator

$\hat{q}_{K_a}$	Kernel density estimator
$\mathcal{Q}_1^\alpha$	Model of conditional distributions that satisfy a Hölder condition of degree α in the second argument
$\mathcal{Q}_2$	Model for the marginal distribution of Z
B_n	a random sequence of weights that depend on the Z-sample only
$\hat{F}_{B_n}$	Nonparametric estimator of F_z
$\hat{G}_{B_n}$	Nonparametric estimator of G_z
$\hat{W}_{B_n}$	Nonparametric estimator of W_z
$\hat{Q}_n^*$	Smoothed empirical distribution
$a(n)$	Data dependent bandwidth
K_a	Kernel density function
$K_a \star f$	Convolution of functions
$K_a \star Q$	Convolution of a function and a measure

Miscellaneous symbols

Symbol	
$(\Omega, \Gamma, \mathcal{P})$	Statistical experiment
n	Sample size
iid	Independent and identical distributed
$\mathbf{R}(\cdot)$	Range of an operator
$\mathbf{N}(\cdot)$	Nullspace of an operator
$\mathrm{E}\,(X)$	Expectation of a random variable X
$Q\varphi$	Integral of φ with respect to Q
$\mathrm{Var}(X)$	Variance of a random variable X
$A^\perp$	Orthogonal complement of A
$(\mathbb{R}, \mathbb{B})$	Real numbers and Borel σ-field.
$\mathcal{F}$	Class of (square integrable) functions
$\mathcal{L}_p(Q)$	Space of p-integrable functions with respect to Q
$\mathcal{L}_p^0(Q)$	Subset of mean zero functions in $\mathcal{L}_p(Q)$
$< \cdot, \cdot >_Q$, $\|\cdot\|_Q$	Inner product and norm on $\mathcal{L}_2(Q)$
$\mathcal{L}_\infty(Q)$	Space of essentially bounded functions with respect to Q
$\mathcal{L}_p, \mathcal{L}_\infty$	$\mathcal{L}_p$-spaces with respect to Lebesgue measure
$l^\infty(\mathcal{F})$	Bounded real valued functions on $\mathcal{F}$ with supremum norm
$\mathcal{C}_0$	Continuous functions on $\mathbb{R}$
$\mathcal{C}_0^\alpha$	Hölder continuous functions of degree α

$\wedge$	Minimum
$\rightarrow$	Convergence
$\mapsto$	Function assignment
$\Rightarrow$	Weak convergence (also for nonmeasurable estimators)

BKRW	Bickel, Klaassen, Ritov, and Wellner (1993)

Bibliography

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics 22*, 1299–1327.

Begun, J. M., W. J. Hall, W.-M. Huang, and J. A. Wellner (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics 11*, 432–452.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins.

Bickel, P. J. and Y. Ritov (1988). Estimating integrated squared density derivates. *Sankhyā A 50*, 381–393.

Birgé, L. and P. Massart (1995). Estimation of integral functionals of a density. *The Annals of Statistics 23*(1), 11–29.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

Brown, L. D. and M. G. Low (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics 24*, 2384–2398.

Carroll, R. J. and W. Härdle (1989). Symmetrized nearest neighbor regression estimates. *Statistics and Probability Letters 7*, 315–318.

Dabrowska, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics 14*, 181–197.

Davies, E. B. (1995). *Spectral Theory and Differential Operators*. Cambridge University Press.

Dawid, A. P. (1984). Present position and potential developments: some personal views. *Journal of the Royal Statistical Society A 147*, 278–292.

Dawid, A. P. (1985). Calibration-based empirical probability. *The Annals of Statistics 13*, 1251–1273.

Dawid, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences (9 vols. plus Supplement)*, Volume 7, pp. 210–218. John Wiley, New York.

Dmitriev, Y. G. and F. P. Tarasenko (1974). On a class of non-parametric estimates of non-linear functionals of a density. *Theory of Probability and its Applications 19*, 390–394.

Doksum, K. and A. Samarov (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics 23*, 1443–1473.

Donoho, D. L. and R. C. Liu (1991). Geometrizing rates of convergence, II. *The Annals of Statistics 19*, 633–667.

Dudley, R. M. (1989). *Real Analysis and Probability*. Chapman & Hall.

Efromovich, S. and M. Low (1996). On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *The Annals of Statistics 2*, 682–686.

Efromovich, S. and A. Samarov (1996). Asymptotic equivalence of nonparametric regression and white noise model has its limits. *Statistics and Probability Letters 28*, 143–145.

Efromovich, S. and A. Samarov (2000). Adaptive estimation of the integral of squared regression derivatives. *Scandinavian Journal of Statistics 27*, 335–351.

Efron, B. (1978). Regression and ANOVA with zero-one data: Measures of residual variation. *Journal of the American Statistical Association 73*, 113–121.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association 81*, 461–470.

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science 1*, 54–75.

Friedman, D. (1983). Effective scoring rules for probabilistic forecasts. *Management Science 29*, 447–454.

Gill, R. D. (1983). Large sample behaviour of the product limit estimator on the whole line. *The Annals of Statistics 11*, 44–58.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von mises method (part1). *Scandinavian Journal of Statistics*, 97–128.

Gill, R. D. and A. W. V. der Vaart (1993). Non- and semi-parametric maximum likelihood estimators and the von mises method (part2). *Scandinavian Journal of Statistics 20*, 271–288.

Gill, R. D., M. J. Van der Laan, and J. M. Robins (1995). Coarsening at random: Characterizations, conjectures and counter-examples. In D. Y.

Lin and T. R. Fleming (Eds.), *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294. Springer Lecture Notes in Statistics.

Goldstein, L. and R. Z. Khas'minskii (1996). On efficient estimation of smooth functionals. *Theory of Probability and its Applications 40*, 151–158.

Goldstein, L. and K. Messer (1992). Optimal plug-in estimators for nonparametric functional estimation. *The Annals of Statistics 20*, 1306–1328.

Graf, E. (1998a). Explained variation measures in survival analysis. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*, Volume 2, pp. 1441–1443. John Wiley, Chichester.

Graf, E. (1998b). *Maßzahlen für erklärte Varianz in der Analyse von Überlebenszeiten*. Ph. D. thesis, Albert-Ludwig Universität Freiburg.

Graf, E., C. Schmoor, and M. Schumacher (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine 18*, 2529–2545.

Groeneboom, P. and J. A. Wellner (1994). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

Hall, P., T. C. Hu, and J. S. Marron (1995). Improved variable window kernel estimates of probability densities. *The Annals of Statistics 23*, 1–10.

Hall, P. and I. Johnstone (1992). Empirical functionals and efficient smoothing parameter selection. *Journal of the Royal Statistical Society B 54*, 475–509.

Hall, P. and J. S. Marron (1987). Estimation of integrated squared density derivatives. *Statistics and Probability Letters 6*, 109–115.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*, 1179–1186.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. John Wiley, Chichester.

Heagerty, P. J., T. Lumley, and M. S. Pepe (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics 56*, 337–344.

Heitjan, D. F. and D. B. Rubin (1991). Ignorability and coarse data. *The Annals of Statistics 19*, 2244–2253.

Helland, I. S. (1987). On the interpretation and use of R^2 in regression analysis. *Biometrics 43*, 61–69.

Hubbard, A. E., M. J. Van der Laan, and J. M. Robins (1998). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In *Proceedings of the IMA, Epidemiological Section*, Institut of Mathematical Applications, Minneapolis, Minnesota. Proceedings of the IMA, Epidemiological Section.

Jacobsen, M. and N. Keiding (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *The Annals of Statistics 23*, 774–786.

Korn, E. L. and R. Simon (1990). Measures of explained variation for survival data. *Statistics in Medicine 9*, 487–503.

Kosorok, M. R. (2002). On global consistency of a bivariate survival estimator under univariate censoring. *Statistics and Probability Letters 56*, 439–446.

Kvalseth, T. O. (1985). Cautionary note about R^2. *The American Statistician 39*, 279–285.

Last, G. and A. Brandt (1995). *Marked Point Processes on the Real Line. The Dynamic Approach.* Probability and Its Applications. New York, NY: Springer-Verlag.

Laurent, B. (1996). Efficient estimation of integral functionals of a density. *The Annals of statistics 24*, 659–681.

Le Cam, L. and G. L. Yang (1988). On the preservation of local asymptotic normality under information loss. *The Annals of Statistics 16*, 483– 520.

Levit, B. Y. (1978). Asymptotically efficient estimation of nonlinear functionals. *Problems of Information Transmission 14*, 65–72.

Matheson, J. and Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*, 1087–1096.

Nan, B. (2001). *Information Bounds and Efficient Estimates for Two-Phase Designs with Lifetime Data.* Ph. D. thesis, University of Washington.

Nolan, D. (1992). Functional limit theorems for probability forecasts. In *Probability in Banach spaces 8*, Volume 30 of *Proceedings of the eighth international conference*, pp. 430–450. Birkhäuser.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics 24*, 2399–2430.

Pfanzagl (with the assistance of W. Wefelmeyer), J. (1985). *Contributions to a General Asymptotic Statistical Theory*, Volume 13. New York: Springer-Verlag. Lecture Notes in Statistics.

Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation.* Probability and Mathematical Statistics. Academic Press Orlando.

Ritov, Y. and P. J. Bickel (1990). Achieving information bounds in non and semiparametric models. *The Annals of Statistics 18*, 925–938.

Ritov, Y. and J. A. Wellner (1988). Censoring, martingales and the Cox model. In N. U. Prabhu (Ed.), *Statistical Inference from Stochastic Processes*, pp. 191–219.

Robins, J. M. and A. Rotnitzky (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In N. P. Jewell, K. Dietz, and V. T. Farewell (Eds.), *AIDS Epidemiology, Methodological Issues*, pp. 297–331. Birkhäuser, Boston.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Rost, D. (2000). Limit theorems for Smoothed Empiricial Processes. In *Giné, Evarist (ed.) et al., High dimensional probability II. Progress in Probability, Birkhäuser. Prog. Probab. 47, 107-113.*

Rudin, W. (1987). *Real and Complex Analysis* (3rd ed.). McGraw-Hill.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*, 783–801.

Schemper, M. and R. Henderson (2000). Predictive accuracy and explained variation in Cox Regression. *Biometrics 56*, 249–255.

Schemper, M. and J. Stare (1996). Explained variation in survival analysis. *Statistics in Medicine 15*, 1999–2012.

Schick, A., V. Susarla, and H. Koul (1988). Efficient estimation of functionals with censored data. *Statistics and Decision 6*, 349–360.

Schumacher, M., N. Holländer, and G. Schwarzer (2001). *Prognostic Factor Studies*, pp. 331–378. New York: Marcel Dekker: Crowley, J ed. Handbook of Statistics in Clinical Oncology.

Schweder, T. (1975). Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian Journal of Statistics 2*, 113–126.

Shorack, G. R. and J. A. Wellner (1986). *Empirical Processes with Applications to Statistics*. John Wiley, New York.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics 5*, 595–620.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics 8*, 1348–1360.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *The Annals of Statistics 12*, 917–926.

Stute, W. (1986). Conditional empirical processes. *The Annals of Statistics 14*, 638–647.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis 45*, 89–103.

Stute, W. (1995). The central limit theorem under random censorship. *The Annals of Statistics 23*, 422–439.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics 23*, 461–71.

Stute, W. and J.-L. Wang (1993). The strong law under random censorship. *The Annals of Statistics 21*, 1591–1607.

Van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics 24*, 596–627.

Van der Vaart, A. (1994). Weak convergence of smoothed empirical processes. *Scandinavian Journal of Statistics 21*, 501–504.

Van der Vaart, A. W. (1988). *Statistical Estimation in Large Parameter Spaces*, Volume CWI Tract 44. Amsterdam: Centrum voor Wiskunde en Informatica.

Van der Vaart, A. W. (1991). On differentiable functionals. *The Annals of Statistics 19*, 178–204.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer series in statistics. Springer.

Van Houwelingen, J. C. and S. Le Cessie (1990). Predictive value of statistical models. *Statistics in Medicine 9*, 1303–1325.

Von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functionals. *Annals of Mathematical Statistics 18*, 309–348.

Wang, J. (1987). A note on the uniform consistency of the Kaplan-Meier estimator. *The Annals of Statistics 15*, 1313–1316.

Wellner, J. A. (1982). Asymptotic optimality of the product limit estimator. *The Annals of Statistics 10*, 595–602.

Wellner, J. A. (1992). Bootstrap limit theorems: A partial survey. In A. K. M. E. Saleh (Ed.), *Nonparametric Statistics and Related Topics*, pp. 313–329.

Winkler, R. L. (1967). The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association 62*, 1105–1120.

Winter, B. B. (1973). Strong uniform consistency of integrals of density estimators. *Canadian Journal of Statistics 1*, 247–253.

Yang, S.-S. (1981). Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *Journal of the American Statistical Association 76*, 658–662.

Zengh, B. and A. Agresti (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine 19*, 1771–81.