A MULTIVARIATE BUCKLEY-JAMES ESTIMATOR

U. Pötter

Ruhr-Universität Bochum, Universitätsstr., D-44780 Bochum, Germany
e-mail: ulrich.poetter@ruhr-uni-bochum.de

ABSTRACT

Buckley and James (1979) extended the least-squares estimator to cover the case of censored dependent variables. I consider a generalisation of their estimator to the multivariate case based on a non-parametric estimator of the joint distribution of the residuals.

Keywords: censored data; multivariate regression

1 INTRODUCTION

Buckley and James (1979) introduced a regression technique suitable for censored dependent variables. Their estimator uses the least-squares estimating equations and an updating mechanism based on a non-parametric estimator of the residual distribution to deal with the censoring. The procedure is attractive because the use of the least-squares technique allows for an easy interpretation of results and the use of residual analysis, while the updating scheme is general enough to accommodate various forms of censoring and grouping. Consequently, many generalisations of the basic technique have been proposed.

In this paper I explore a possible extension to multivariate dependent variables. Related work, especially that of Lin and Wei (1992), Lee, Wei and Ying (1993), Pan and Kooperberg (1999), and Hornsteiner and collaborators in a series of papers (1996, 1997, 1998), is mainly inspired by the literature on generalised estimating equations. It concentrates on the estimation of the marginal effects of covariates on each of the dependent variables. Accordingly, the least-squares estimating equations are modified to accommodate the multivariate character of the dependent variables. Less emphasis is put on the updating scheme that deals with the censoring problem. The authors suggest to use non-parametric estimators of the marginal

distributions of the residuals only. I propose to incorporate the multivariate information from the residual distribution into the updating scheme.

In the next section I introduce Buckley and James' approach to regression estimation with censored observations and, in section 3, indicate why it works. Next I consider the multivariate case. Generalisations of the missing information principle are treated in section 5. This leads to a multivariate extension of Buckley and James' approach that uses the multivariate information also for the updating scheme. In the final section I examine the performance of the estimator through examples.

2 BUCKLEY-JAMES ESTIMATORS

Suppose that conditionally on some covariates x, the random variable Y follows a linear regression

$$Y = x\beta + \epsilon \,, \tag{1}$$

where x is a $1 \times p$ vector of covariates including a constant, β is a $p \times 1$ vector of unknown regression coefficients, and ϵ is a random variable with mean zero and finite variance. If Y is the logarithm of a positive random variable representing a duration or time to an event, this model is sometimes called accelerated failure time model (Cox and Oakes 1984, chap. 5.2).

In many applications only censored observations from Y are available. More precisely, suppose that the observations are given by the censored variable Z and censoring indicator δ:

$$Z := \min(C, Y) \,,\ \delta := I[C \geq Y] \,,$$

where $I[.]$ is the indicator function and the censoring variable C is (conditionally) independent of Y. The observations are n independent and identically distributed realizations from (x, ϵ, C). The $n \times (p+2)$ data matrix is given by $(z_i, \delta_i, x_i)_{i=1,\dots,n}$.

In the absence of censoring one can estimate β by minimising the least-squares criterion

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2 = n \int e^2 \, d\hat{F}_n(e) = \sum_{i=1}^{n} \int (y - x_i\beta)^2 \, d\tilde{F}_{ni}(y) \,, \tag{2}$$

where $\hat{F}_n(e)$ is the empirical distribution function of the residuals $e_i = y_i - x_i\beta$, and $\tilde{F}_{ni}(y) = I[y_i < y]$ is the empirical distribution of just one observation y_i.

Miller (1976) and Leurgans (1987), using the second and third representation respectively, proposed replacing the empirical distributions by versions appropriate for censored data. Instead of taking the least-squares criterium (2) as their starting point, Buckley and James (1979) suggested to modify the least-squares estimating equations

$$\sum_{i=1}^{n} x_i'(y_i - x_i\hat{\beta}) = 0 \quad \text{or} \quad \sum_{i=1}^{n} x_i' y_i = \left(\sum_{i=1}^{n} x_i' x_i\right)\hat{\beta}\,. \tag{3}$$

In the presence of censoring they proposed to replace the censored observations Z by the conditional expectation of Y given the observed (censored) data Z and the covariates:

$$Y^* = \mathbb{E}_\beta(Y \mid z, \delta, x) = \delta z + (1-\delta)\mathbb{E}_\beta(Y \mid Y \geq z, x)\,. \tag{4}$$

Note the dependence of the conditional expectation on the unknown parameter β. Replacing Y in expression (3) by its conditional expectation gives

$$\frac{1}{n}\sum_i x_i' \mathbb{E}_{\hat{\beta}}(Y \mid z_i, \delta_i, x_i) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i' x_i\right)\hat{\beta}\,. \tag{5}$$

In other words, the Buckley-James estimator $\hat{\beta}$ solves the normal score function for β when the expectation on the left hand side is computed using $\hat{\beta}$.

Using the model formula (1) and a fixed β, an empirical version of the conditional expectation can be evaluated:

$$\begin{aligned}
\hat{\mathbb{E}}_\beta(Y \mid z_i, \delta_i, x_i) &=: \hat{y}_i(\beta) \qquad (6)\\
&= \delta_i z_i + (1-\delta_i)\hat{\mathbb{E}}_\beta(Y \mid Y_i \geq z_i, x_i)\\
&= \delta_i z_i + (1-\delta_i)\left(x_i\beta + \frac{\int_{e_i}^{\infty} e\, d\hat{F}_\beta(e)}{\hat{S}_\beta(e_i)}\right)\\
&= \delta_i z_i + (1-\delta_i)\left(\sum_{k=i}^{n} v_{ik}(\beta)(z_k - x_k\beta) + x_i\beta\right)
\end{aligned}$$

where $\hat{F}_\beta$ is an estimator of the distribution function of the residuals (e.g. the Kaplan-Meier estimator), $\hat{S}_\beta$ is the estimated survivor function $1 - \hat{F}_\beta$, and I have put

$$v_{ik}(\beta) = \begin{cases} \dfrac{w_k(\beta)}{\hat{S}_\beta(e_i)} & \text{if } e_i < e_k \\ 0 & \text{otherwise} \end{cases}$$

and

$$w_k(\beta) = \hat{P}_\beta(\epsilon = e_k)\,,$$

so that $w_i(\beta)$ is the height of the jump of the estimated distribution at the i-th residual.[1] A solution $\hat{\beta}$ of the estimating equation (3) therefore satisfies:

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i' x_i\right)^{-1}\left(\sum_{i=1}^{n} \delta_i x_i' z_i + \sum_{i=1}^{n}(1-\delta_i)x_i'\hat{y}_i(\hat{\beta})\right)\,. \tag{7}$$

[1]For ease of notation it is assumed here that the observations are ordered according to the magnitude of the corresponding residuals.

This leads to a straightforward iterative procedure for the computation of $\hat{\beta}$:

1. Assign starting values $\hat{\beta}^0$.
2. Compute $\hat{y}_i(\hat{\beta}^j)$ according to (6) using the Kaplan-Meier procedure as an estimator for the distribution of the residuals.
3. Compute $\hat{\beta}^{j+1}$ using the right hand side from (7).
4. Go back to step 2 unless some convergence criterion is met.

To be numerically effective, this simple iterative strategy needs elaboration. Following the steps of the algorithm, the basic choices are:

1. Starting values may be obtained using the least-squares estimator treating all observations as uncensored. This was suggested by Buckley and James (1979). Other choices, e.g. using only uncensored observations, are of course possible, but do not seem to have a decisive influence on the procedure.
2. The Kaplan-Meier estimator is not uniquely defined on the whole real line if the largest residual is censored. Buckley and James suggest to always treat the largest residual as uncensored. This will lead to an underestimation of the regression constant, but should scarcely affect the other regression estimators. Other choices are discussed by Efron (1988), while Lai and Ying (1991) propose to smooth the risk sets.
4. The iteration may not converge to a unique value. This is due to the fact that the right hand side of (7) is a piecewise linear function in β. Changing β does not change the weights $v_{ik}(\beta)$ unless the ranks of the residuals change. Therefore, the iterations may oscillate between several values $\hat{\beta}$. The discontinuity of (7) hampers the analytic treatment of the estimator. Moreover, the number of limiting values in finite samples is not predictable, but may potentially be rather large (Currie, 1996). Fortunately, the phenomenon seems to be of practical interest only in rather small samples, in situations where the effect of covariates is small, or when the convergence criterion is very strict (Wu/Zubovic, 1995).[2]

[2]Wu and Zubovic (1995) suggested to use the arithmetic mean of all limit values of the algorithm as estimator. This suggestion may be useful in situations where a unique estimator is required (e.g. simulations, using the procedure as building block for more complicated models, etc.). Otherwise, the different values of the limiting cycle of estimators are often very close and it may suffice to report just one of them.

3 SCORE FUNCTIONS AND CENSORING

To appreciate why the Buckley-James procedure is a "good" generalisation of estimating equations to censored variables it is helpful to consider it from a more general point of view. Especially the relation between score functions with and without censoring is revealing. Write $\dot{\ell}(\beta) = \dot{\ell}(\beta; Y, x) = x'(Y - x\beta)$ for the score function from the normal linear regression model (1). The expectation satisfies

$$\mathbb{E}_\beta(\dot{\ell}(\beta; Y, x)) = 0\,. \tag{8}$$

Moreover, the root $\hat{\beta}$ of the empirical version of the expectation (8),

$$\frac{1}{n}\sum_i \dot{\ell}(\hat{\beta}; y_i, x_i) = 0\,,$$

is the maximum likelihood estimator. Even if the distribution is not normal — so that the root of the score function need no longer be a maximum likelihood estimator — $\hat{\beta}$ is consistent and often highly efficient. In the presence of censoring, the censored normal score function $\dot{\ell}^*$ can be expressed as

$$\dot{\ell}^*(\beta; Z, \delta, x) = \mathbb{E}(\dot{\ell}(\beta; Y, x) \mid Z, \delta, x)\,, \tag{9}$$

the conditional expectation of the score function with complete observations given the incomplete observations (see e.g. Ibragimov/Has'minskii, 1981, chap. I.7). This relation between score functions for complete and incomplete observations makes the score function an attractive starting point for the construction of estimators.

It remains to consider the computation of the conditional expectation. From the perspective of the normal linear regression model one might try to use the normal distribution. This was proposed by Schmee and Hahn (1979) and Aitkin (1981). However, one can only expect the good properties of the estimators even outside the normal distribution to extend to censored data situations if the conditional expectation is computed from a non-parametric estimator. In the case of right censored observations, the Kaplan-Meier estimator, being a non-parametric maximum likelihood estimator solving a self-consistency equation, seems to be an appropriate choice. In fact, Lai and Ying (1994), following Ritov (1990) and Severini and Wong (1992), provide a general argument for the use of self-consistent estimators in the computation of conditional expectations for censored and truncated observations.[3] To outline the reasoning it is best to regard the estimation problem as one involving both β and the distribution of ϵ, F, as unknown parameters. Here, β is the parameter of interest and F is treated as a nuisance parameter. In such a context, one may consider the score function corresponding to the profile likelihood. The profile log-likelihood is derived from the log-likelihood $\ell(\beta, F)$ by replacing F with an estimator $\hat{F}_\beta$ treating β as known. It is thus a function of β only. Symbolically, then, one may write

$$\frac{d}{d\beta}\ell(\beta, \hat{F}_\beta) = \frac{\partial}{\partial\beta}\ell(\beta, F)|_{(\beta,\hat{F}_\beta)} + \frac{\partial}{\partial F}\ell(\beta, F)|_{(\beta,\hat{F}_\beta)}\frac{\partial}{\partial\beta}\hat{F}_\beta \tag{10}$$

for its score function. If $\hat{F}_\beta$ is of maximum likelihood type, the sample mean of the second term vanishes. One needs only to consider the score function for β that would result if F was known.

This holds for all unbiased estimating equations for F_β. But an estimator $\hat{F}_\beta$ that maximises the likelihood $\ell(\beta, F)$ in F for β fixed automatically provides an estimator of the least favourable submodel $\beta \mapsto (\beta, \hat{F}_\beta)$ for the estimation of β and therefore (10) approximates the efficient score function, making efficient estimation of β feasible. Thus one would like to use an estimator $\hat{F}_\beta$ that simultaneously solves an estimating equation and maximises a non-parametric likelihood.

Taking $F_\beta(u) - I[Y - x\beta \le u]$ as a score function for F_β in the uncensored case, one is led via the projection of scores (9) to an estimator of F_β that satisfies the corresponding self-consistency equation, namely

$$0 = \mathbb{E}_n\dot{\ell}^*(\hat{F}_\beta) = \mathbb{E}_n(\mathbb{E}_{\hat{F}_\beta|Z,\delta,x}(\dot{\ell}(\hat{F}_\beta)|Z,\delta,x)) = \hat{F}_\beta(u) - \frac{1}{n}\sum_{i=1}^n \hat{F}_\beta(u|z_i, \delta_i, x_i)\,. \tag{11}$$

But the estimator $\hat{F}_\beta$ that solves the self-consistency equations and maximises the non-parametric likelihood is the Kaplan-Meier estimator. On the other hand, considering $\mathbb{E}(\partial\ell(\beta, \hat{F}_\beta; Y, x)/\partial\beta \mid Z, \delta, x)$ as the profile score function in the presence of censoring, one is led to the estimating equations

$$\begin{aligned} 0 &= \mathbb{E}_n\mathbb{E}_{\hat{F}_{\hat{\beta}}}(\dot{\ell}(\hat{\beta}; Y, x)|Z, \delta, x) \\ &= \frac{1}{n}\sum_{i=1}^n x_i'(x_i\hat{\beta} - \mathbb{E}_{\hat{F}_{\hat{\beta}}}(Y|z_i, \delta_i x_i)) \\ &= \frac{1}{n}\sum_{i=1}^n x_i'(x_i\hat{\beta} - \hat{y}_i(\hat{\beta}))\,, \end{aligned}$$

leading back to (5). From this perspective, then, both the choice of the normal score function $\dot{\ell}(\beta) = x'(Y - x\beta)$ as a starting point and the use of the Kaplan-Meier estimator are the appropriate extension of an estimating equation technique to censored data.

[3] The argument extents to estimating equations that are not derived from likelihood functions. See Bickel et al. (1998, chap. 7.7) for a general discussion of the construction of estimators along these lines.

4 MULTIVARIATE EXTENSIONS

To consider the multivariate situation I write $\boldsymbol{Y} = (Y_1, \ldots, Y_k)'$ for the column vector of k dependent variables. The covariates are given by a $k \times kp$ matrix $\boldsymbol{x}$ where the j-th row corresponds to the p covariates x_j of the j-th dependent variable Y_j with zeros padded in the appropriate places. The regression coefficients are given by a column vector $\boldsymbol{\beta}$ of dimension $kp \times 1$. The multivariate linear model can then be presented as

$$\boldsymbol{Y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} x_1 & \mathbf{0} & \ldots & \ldots & \mathbf{0} \\ \mathbf{0} & x_2 & \mathbf{0} & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & x_k \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \boldsymbol{\epsilon} \tag{12}$$

with residual vector $\boldsymbol{\epsilon}$. The mean of the residuals is $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and the covariances are given by $\mathbb{E}(\boldsymbol{\epsilon\epsilon}') = \boldsymbol{\Omega}$. As before, the covariate vectors x_j are assumed to contain a constant. Note that in the case of equal effects $\beta_1 = \beta_2 = \ldots = \beta_k$ the $\boldsymbol{x}$ matrix can be reduced to a $k \times p$ matrix.

Now suppose that the data are censored by a k-dimensional variable $\boldsymbol{C} = (C_1, \ldots, C_k)'$. Instead of $\boldsymbol{Y}$ only the vectors $\boldsymbol{Z} = (Z_1, \ldots, Z_k)' = (\min(Y_1, C_1), \ldots, \min(Y_k, C_k))' = \min(\boldsymbol{Y}, \boldsymbol{C})$ and $\boldsymbol{\delta} = (I[C_1 \geq Y_1], \ldots, I[C_k \geq Y_k])' = I[\boldsymbol{C} \geq \boldsymbol{Y}]$ are observed. Note that here and in the sequel minima, indicator functions, and (in-)equalities are interpreted componentwise.

To render the conditional distribution of $\boldsymbol{Y}$ identifiable from the censored version $(\boldsymbol{Z}, \boldsymbol{\delta})$ I will assume that the censoring vector $\boldsymbol{C}$ and the vector $\boldsymbol{Y}$ are (conditionally on $\boldsymbol{x}$) independent. Moreover, the support of $\boldsymbol{Y}$ is assumed to be contained in the support of $\boldsymbol{C}$.[4]

Using this model, Lin and Wei (1992), Lee, Wei and Ying (1993), and Hornsteiner and collaborators (1996, 1997, 1998) proposed extensions to the one-dimensional Buckley-James estimator. In these papers, a solution to an equation similar to (7) is used. Both Lin and Wei (1992), and Lee, Wei and Ying (1993) use k least-squares estimating equations disregarding possible correlations. Hornsteiner et al. (1996, 1997, 1998) and Pan and Kooperberg (1999) use a working correlation matrix $V(\alpha)$ (of dimension $k \times k$) in a generalised least-squares estimating equation

$$\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{V}(\hat{\alpha})^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{13}$$

[4] Some of the censoring patterns of interest in event history analysis, e.g. censoring of the recurrence times in a semi-Markov process by a fixed observation interval, are not easily represented in this setup. Reference to an underlying process would be necessary to line up censorings and durations according to their timing on a common time scale. See Dabrowska and Lee (1996), Li and Lagakos (1997), and Tsai and Crowley (1998) for some discussion.

in an attempt to gain efficiency. To deal with the censoring, all these proposals use an updating scheme parallel to the one-dimensional case, namely the conditional expectations

$$\begin{aligned} Y_j^{**} &= \mathbb{E}_{\beta_j}(Y_j \mid z_j, \delta_j, x_j) \\ &= \delta_j z_j + (1 - \delta_j)\mathbb{E}_{\beta_j}(Y_j \mid Y_j \geq z_j, x_j), \quad j \in \{1, \ldots, k\} \end{aligned} \tag{14}$$

from the j-th model equation. This leads to the correct mean structure while using only the marginal distributions of the residuals. The conditional expectations are then computed from the marginal Kaplan-Meier estimators of the distribution of the residuals. While Lin and Wei and Lee, Wei and Ying simply use the marginal Kaplan-Meier estimators, Hornsteiner (1998) also considers pooled and weighted versions to increase efficiency in certain situations. In addition to the contributions of Y_j^{**} in the updating scheme, both Hornsteiner et al. and Pan and Kooperberg (1999) also base their estimating equations for α on the values of Y_j^{**}. As Hornsteiner (1998, p. 49) notes, this approach is approximately valid only if the amount of censoring is small.

5 THE MISSING INFORMATION PRINCIPLE AND NON-PARAMETRIC ESTIMATION OF CENSORED MULTIVARIATE OBSERVATIONS

Starting with a score function $\dot{\ell}(\boldsymbol{\beta})$ derived from a likelihood $\ell(\boldsymbol{\beta})$, the missing information principle suggests to use the conditional expectation of $\dot{\ell}(\boldsymbol{\beta})$ based on all the available information, not just the information from the marginal distributions. Thus one may consider the conditional expectations

$$\begin{aligned} Y_j^{*} &= \mathbb{E}_{\boldsymbol{\beta}}(Y_j \mid \boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{x}) \\ &= \delta_j z_j + (1 - \delta_j)\mathbb{E}_{\boldsymbol{\beta}}(Y_j \mid Y_j \geq z_j, (\boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{x})), \quad j \in \{1, \ldots, k\} \end{aligned} \tag{15}$$

instead of (14). This conditional expectation is based on all the information on $\boldsymbol{Y}$ available from the data while (14) uses only the information from the distribution in the j-th dimension. Extending the argument from section 3 one would expect (15) to give an appropriate generalisation of one-dimensional censored regression if it was possible to exhibit a self-consistent estimator of the multivariate distribution of the censored residuals. Also, from a more practical point of view, it seems advantageous to use as much information as possible in dealing with the censoring process without imposing strong extraneous assumptions. If the degree of censoring is high and if there is considerable correlation within $\boldsymbol{Y}$ or $\boldsymbol{C}$, one might expect (15) to perform better than (14).

In the context of multivariate proportional hazards models this approach was implicitly suggested by Prentice and Hsu (1997) and Cai and Prentice

(1995). On the other hand, this extension has not been discussed in the context of the Buckley-James approach. This is not by accident: in the computation of $\mathbb{E}(\boldsymbol{Y}\,|\,\boldsymbol{Z},\boldsymbol{\delta},\boldsymbol{x})$ one would need a non-parametric estimator of the joint distribution of $\boldsymbol{\epsilon}$ from censored data that additionally should solve a self-consistency equation, maximise a non-parametric likelihood, and, for practical reasons, should allow for easy computation of conditional expectations along half-lines or orthants.

In dimension 2 or higher, there is no unique self-consistent non-parametric maximum likelihood estimator (NPMLE) of the distribution function of $\boldsymbol{\epsilon}$. In fact, the EM type argument leading to (11) will not even result in a consistent estimate. To fix ideas, consider the two-dimensional problem, $k = 2$, disregarding covariates for the moment. Suppose one observes $(z_1, z_2, 0, 1)$, censored in the first component, but exactly observed in the second. This says that the underlying tuple (y_1, y_2) is located on the ray $\{(y_1, y_2)|y_1 > z_1, y_2 = z_2\}$ parallel to the first axis. But if the distribution of (Y_1, Y_2) is absolutely continuous, the probability of obtaining another uncensored observation lying on this ray is 0.

Without uncensored observations on the ray there is no empirical support for the computation of the distribution function along this ray. To compute a self-consistent estimator, one needs an expression for $\Pr((Y_1, Y_2) \leq (u_1, u_2)|(Y_1, Y_2) \in \{(y_1, y_2)|y_1 > z_1, y_2 = z_2\})$, the last term in the self-consistency equation (11) based on a current estimate of the joint distribution. If there are no uncensored observations on the ray, the conditioning event has probability 0 for all sensible starting estimates. Therefore the conditional probability can be defined arbitrarily. But updates of the estimator based on the self-consistency equation will not change due to probability mass transferred from the censored observation to uncensored observations, thus leading to inconsistent estimators.

In response to these difficulties several alternative estimators of the joint distribution of multivariate censored observations have been developed. Pruitt (1993) describes six estimators, summarises their known properties, and compares their small sample behaviour in a limited Monte Carlo experiment. Further comparisons are contained in van der Laan (1997). Some of these estimators are based on a decomposition of the joint distribution into conditional times marginal distributions. The approaches then proceed using the one-dimensional Kaplan-Meier estimator. But the resulting estimators will generally depend on the ordering of the decomposition. Other approaches use smoothing techniques for singly censored observations, thus depending on the choice of a smoothing parameter. The proposals of Dabrowska (1988, 1989) and Prentice and Cai (1992) use special representations of the multivariate survivor function, both representations giving rise to explicit estimators of the distribution function. Gill (1992) provides

a lucid introduction to these methods, and both are discussed in Pruitt's 1993 article. Though computationally attractive, both estimators are neither solutions to some self-consistency equation nor are they of maximum likelihood type.

All these approaches may yield negative mass for the increments of the estimated distribution function (Pruitt 1991). This property is especially disturbing when one is interested in computing conditional expectations $\mathbb{E}_{\hat{F}}(Y_1|Y_1 > z_1, Y_2 = y_2)$ which may result in values $\leq z_1$ for these estimators. Moreover, the implied computation of conditional expectations used in (15) are indetermined in general and cannot directly be used in a generalisation of the Buckley-James procedure.

In contrast, there is an essentially unique non-parametric estimator for discrete censored data maximising a likelihood. It was first considered by Campbel (1981a,b). This let van der Laan (1995, 1996, 1997) to consider a non-parametric MLE based on discretised censored observations. In the two-dimensional case, let $D = D_1 \times D_2$ be a rectangle covering the observations so that $(z_{1i}, z_{2i}) \in D$ for all observations. Partition the side of D_1 into q_1 intervals of equal length, $i_{1,l}, l = 1, \ldots, q_1$. Partition D_2 into q_2 intervals $i_{2,l}, l = 1, \ldots, q_2$, also of equal length. This partitions D into $q_1 q_2$ congruent rectangular boxes $i_{1,l} \times i_{2,m}$. Now coarsen the observations as follows: if the observation is uncensored $((\delta_1, \delta_2) = (1, 1))$ or censored in both dimensions $((\delta_1, \delta_2) = (0, 0))$, keep the data as $(z_1, z_2, \delta_1, \delta_2)$. If the observation is censored in only one dimension $((\delta_1, \delta_2) = (0, 1)$ or $(\delta_1, \delta_2) = (1, 0))$, replace the uncensored dimension by the interval it falls into. That is, if the observation is censored in the first dimension, $(z_1, z_2, 0, 1)$, replace z_2 with the interval $i_{2,l}$ to which z_2 belongs. The corresponding (y_1, y_2) are therefore assumed to lie in the strip $\{(y_1, y_2)|y_1 > z_1, y_2 \in i_{2l}\}$. Moreover, the strip is restricted to the domain D, $\{(y_1, y_2)|y_1 > z_1, y_2 \in i_{2l}\} \cap D$. Similarly, observations only censored in the second dimension, $(z_1, z_2, 1, 0)$, are grouped into $(i_{1,l}, z_2, 1, 0) \cap D$.

In figure (1) five observations are depicted. The filled circles represent uncensored observations while the hollow ones represent singly and doubly censored observations. Feasible values of (y_1, y_2) in the case of singly censored observations lie on the rays indicated by solid lines, while values corresponding to the doubly censored observation lie in the orthant indicated by the brocken line. The box around the figure indicates the domain D which is partitioned by intervals of equal length along its two sides. The resulting grid is shown by light lines. The coarsening of the observations does not change the uncensored or doubly censored observations. However, the values of (y_1, y_2) corresponding to the two singly censored observations are now assumed to lie in the shaded strips. While the rays do not contain any uncensored observations, the strip corresponding to the observation censored in the second dimension now contains an uncensored observation. For the

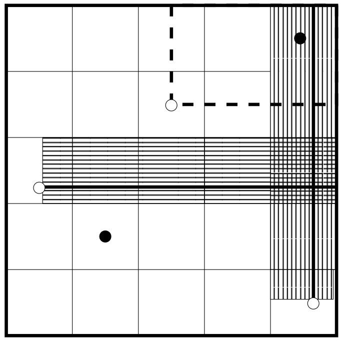

Figure 1: Coarsening censored observations

reduced data the self-consistency equations contain the term $\Pr((Y_1, Y_2) \leq (u_1, u_2)|(Y_1, Y_2) \in \{(y_1, y_2)|y_1 > z_1, y_2 \in i_{2,l(z_2)}\})$ for observations singly censored in the first dimension, where $l(z_2) = \{l \in \{1, \ldots, k\}|z_2 \in i_{2l}\}$. In general, there will be uncensored observations in the strips corresponding to the conditioning event. Thus, changes in the mass attributed to the uncensored observations will be reflected in the updating scheme for the singly censored observations. One may hope that this recaptures the properties of self-consistent estimators in the discrete multivariate and the one-dimensional case, albeit at the cost of throwing away some data.

In fact, van der Laan (1996) showed that the self-consistent MLE based on the reduced data is uniformly consistent and asymptotically normal[5]. To achieve asymptotic efficiency of the reduced data MLE, he shows that the length of the coarsening intervals i in the two-dimensional case have to shrink

[5] In his simulations and the proofs van der Laan uses a slightly more complicated method of data reduction than the one proposed above. It involves a simultaneous coarsening of the censoring variables $\boldsymbol{C}$ in addition to the coarsening of the uncensored dimensions. If $\boldsymbol{Y}$ is independent of $\boldsymbol{C}$ this is no longer true for the coarsened data version, since $\Pr(Y_1 \in i_{1,l}, \delta_1 = 1) = \Pr(Y_1 \in i_{1,l}, C_1 \geq Y_1) = \int_{i_{1,l}} 1 - G_1(u_-)dF_1(u)$, where F_1 and G_1 are the (marginal) distributions of Y_1 and C_1, respectively. Thus the likelihood no longer factors into a term only containing F and another only depending on the censoring distribution G. Van der Laan's proposal retains the orthogonality between $\boldsymbol{C}$ and $\boldsymbol{Y}$ and thus allows asymptotic arguments based on a sequence of identical models. From a practical point of view and considering that the independence of the censoring scheme cannot be ascertained from the observations one may as well assume that the non-parametric likelihood in the coarsened model factors. One should then bear in mind that different models for the original and coarsened experiment are used, and that one changes models when changing the coarsening grid.

to 0 at a rate slower than $n^{-1/18}$ (van der Laan 1996, Theorem 5.1). This does not provide much guidance for sample sizes practically encountered. His simulations (1997) suggest that a small interval length of 0.02 for the square $[0, 1] \times [0, 1]$ and $n = 200$ works well. Our limited experience indicates that in order to attain stable estimates of conditional expectations for the use in Buckley-James iterations it is expedient to use rather larger coarsening intervals.

The procedure is easily generalised to k dimensions. All observations with $0 < \sum_{l=1}^{k} \delta_l < k$ are coarsened to a lattice in $D = D_1 \times \ldots \times D_k$ induced by a partition of the D_j into intervals of equal length. This will ensure that the conditioning events in the self-consistency equations will have positive k-dimensional contents. The estimation procedure for the non-parametric self-consistent MLE of the reduced k-dimensional data can be summarised as follows:

1. Choose a region $D = D_1 \times \ldots \times D_k$. I use $D_l =]\min_i z_{li} - \sigma, \max_i z_{li} + \sigma]$. Note that the choice $\sigma = 0$ will exclude observations that are either right censored in this component at the maximum, or are uncensored in this component at the minimum of the observations.

2. Choose the number q_l of intervals i_l for each dimension l. Partition each side D_l into q_l intervals $i_{l,1}, \ldots, i_{l,q_l}$. I use left open and right closed intervals. Partition D accordingly in $\prod_{l=1}^{k} q_l$ boxes $i_{1,m_1} \times \ldots \times i_{k,m_k}$.

3. Choose starting values. The NPMLE is discrete. It suffices to specify point masses for $\widehat{\Pr}(\boldsymbol{Y} = \boldsymbol{y})$. We choose to put mass $1/n$ on all uncensored observations. The mass of $1/n$ of censored observations is equally spread over the strips implied by the censoring pattern of that observation. To all uncensored observations in the strip and to all intersections of the strip with other strips or with the boundary of D the appropriate part of $1/n$ is added. This will produce a superset of the support points of the NPMLE. Pruitt (1993), Betensky and Finkelstein (1999), and Prentice (1999) discuss the exact determination of the support points of the NPMLE in the two-dimensional case, but the formulation does not easily generalise to higher dimensions.

4. Iterate the self-consistency equations: For each support point $\boldsymbol{y}$ compute the new value $\widehat{\Pr}^{j+1}(\boldsymbol{Y} = \boldsymbol{y})$ as the mean of the conditional probabilities given the observed information, $1/n \sum_i \widehat{\Pr}^{j}(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{Z}_i, \boldsymbol{\delta}_i)$, where the probability of the conditioning event is the sum over the probabilities $\widehat{\Pr}^{j}(\boldsymbol{Y} = \boldsymbol{y})$ lying in the strip determined by $(\boldsymbol{Z}, \boldsymbol{\delta})_i$.

5. Stop the iteration using some convergence criterion. I use the maximum of $|\widehat{\Pr}^{j+1}(\boldsymbol{Y} = \boldsymbol{y}) - \widehat{\Pr}^{j}(\boldsymbol{Y} = \boldsymbol{y})|$ over all support points as

convergence criterion.

This EM algorithm generally converges very slowly. Especially the mass of points not in the support of the MLE, but given positive mass by our determination of starting values, decreases only slowly to 0. Prentice (1999) and Betensky and Finkelstein (1999) proposed to use a direct constraint maximisation algorithm based on the likelihood function. But the approach will fail if the maximum of the likelihood is not unique. This happens if there are strips (or orthants) corresponding to censored observations that intersect D without intersecting other strips or uncensored observations. Region of non-uniqueness can be ascertained in the two-dimensional case, though the procedure is quite tedious. Excluding these region from the maximisation problem would make direct maximisation algorithms very appealing. Unfortunately, we did not find a feasible formulation for the regions of non-uniqueness in the k-dimensional case. In contrast to the direct maximisation approaches the EM algorithm is not hampered by the possible non-uniqueness of the NPMLE. It simply does not change estimates in the regions of non-uniqueness. Since the estimator is to be used repeatedly based on changing data in the Buckley-James procedure, it seems appropriate to use the slow but reliable EM algorithm.

6 THE MULTIVARIATE BUCKLEY-JAMES ESTIMATOR

With a NPMLE for the distribution of multivariate censored data at hand, an algorithm for the computation of multivariate regression estimators in the model (12) using the Buckley-James approach can be described as follows:

1. Compute starting values for $\boldsymbol{\beta}$. I use the least-squares estimator treating all observations as uncensored.
2. For the $j+1$-th iteration, compute the NPMLE of the residuals based on the data $(\boldsymbol{z}_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}^j, \boldsymbol{\delta}_i)$.
3. Compute new values of the dependent variable as $\boldsymbol{Y}^{*(j+1)} = \hat{\boldsymbol{y}}(\hat{\boldsymbol{\beta}}^j)$ according to (15). The conditional expectations of the censored residuals $\boldsymbol{e}_i$ are evaluated as the weighted means of the residuals $\boldsymbol{e}_k$. The estimates from step 2 are used as weights and the summation is over the regions determined by the censoring pattern.
4. Compute new regression coefficients $\hat{\boldsymbol{\beta}}^{j+1}$ using a least squares regression of $\boldsymbol{Y}^{*(j+1)}$ on $\boldsymbol{x}$.
5. Go back to step 2 unless some convergence criterion is met. I use the maximum of $|\hat{\beta}_m^{j+1} - \hat{\beta}_m^j| / \max(|\hat{\beta}_m^{j+1}|, 1)$, where m indexes the elements of $\boldsymbol{\beta}$.

The distinctive feature of the estimator is the use of the joint distribution of the residuals to compute expected values in step 3. To illustrate the effectiveness of the computations, I generate $n = 300$ bivariate normal observations with $Y_1 \sim N(0,1), Y_2 \sim N(0,1)$ and $\text{corr}(Y_1, Y_2) = 0.8$. These are censored in the second dimension only by $C_2 \sim N(2.4, 1)$. Figure (2) compares the estimated expected values of the censored observations (circles) based on the joint distribution (diamonds) with those based on the marginal distribution only (crosses). The estimates based on the joint distribution are clearly better in mimicking the underlying distribution than are the estimates based on the marginal distribution only. Figure

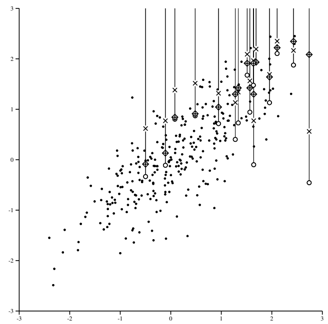

Figure 2: Conditional expectations: Correlation 0.8

(3) compares the two approaches in the case of independent components $Y_1 \sim N(0,1), Y_2 \sim N(0,1)$, once again with $n = 300$ and $C_2 \sim N(2.4, 1)$. In this situation the estimates based on the joint distribution may be thought to fare less well. While the joint distribution cannot supply any additional information over the marginal distribution, the estimator based on the joint distribution looses information due to the coarsening. In this (and the previous) example I partitioned the first dimension into 10 intervals. It seems apparent from figure (3) that the estimates based on the joint distribution do not suffer strongly from the coarsening.

As an example for the effect of joint versus marginal estimation on the regression coefficients I use data from Wei, Lin and Weissfeld (1989, table 1). The data give natural logarithm of the number of days, z_{li}, to virus positivity in the l-th serum sample of the i-th patient, $l = 1, 2, 3; i = 1, \ldots, 36$. There are thus three time dimensions. Patients were treated with ribavirin. There are three treatment groups: placebo, low dose, and high dose. This covariate information is coded in two dummy variables indicating low dose

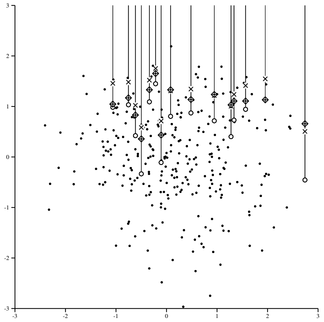

Figure 3: Conditional expectations: Correlation 0

group and high dose group, respectively. There are six observations with missing values in one of the z_{li}. These were excluded from the analysis. Table (1) compares the estimated regression coefficients from a model using a marginal Kaplan-Meier estimator with the proposed method using the joint distribution estimator. The latter was computed using a coarsening to five intervals of equal length in each of the three dimensions. The procedure converged after four Buckley-James iterations in each of which the computation of the NPMLE took four to five iterations. The resulting estimated coefficients are all slightly smaller than the coefficients from the marginal estimator.

Table 1: Dependent variable: natural logarithm of days to virus positivity

	marginal	joint
Constant 1	1.893	1.893
Constant 2	2.170	2.166
Constant 3	2.179	2.149
low dose 1	0.692	0.674
high dose 1	0.542	0.530
low dose 2	0.168	0.128
high dose 2	0.028	0.021
low dose 3	0.596	0.530
high dose 3	0.252	0.229

7 DISCUSSION

The suggested multivariate Buckley-James estimator seems to be a feasible alternative to approaches based on the marginal distribution of the residuals. I have tried it with real and simulated datasets with up to 4000 observations and up to 10 dimensions. The most time consuming part of its computation is the estimation of the joint distribution of the residuals, which may often take 20 to 30 iterations. It would therefore be of interest to develop reliable direct maximisation procedures for the NPMLE.

An obvious obstacle to the use of the estimator is the lack of a variance estimator for regression coefficients. This is due to the fact that there is no variance expression for the NPMLE. Nevertheless, it might be possible to obtain variance estimators from a numerical approximation of the score function.

References

Aitkin, M. (1981) A note on the regression analysis of censored data. *Technometrics*, **23**, 161–163.

Betensky, R.A., Finkelstein, D. M. (1999) A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statist. Med.*, **18**, 3089–3100.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., Wellner, J. (1998) *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, Berlin

Buckley, J., James, I. (1979) Linear regression with censored data. *Biometrika*, **66**, 429–436.

Cai, J., Prentice, R. L. (1995) Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, **82**, 151–164.

Campbell, G. (1981a) Nonparametric bivariate estimation with randomly censored data. *Biometrika*, **68**, 417–422.

Campbell, G. (1981b) Asymptotic properties of several nonparametric multivariate distribution function estimators under random censorship. In: *Survival Analysis*, (eds. J. Crowley, R. A. Johnson), Inst. of Math. Stat. Lecture Notes, Monograph Series, 2, Columbus, 243–256.

Cox, D. R., Oakes, D. (1984) *Analysis of Survival Data*. Chapman & Hall, London.

Currie, I. D. (1996) A note on Buckley–James estimators for censored data. *Biometrika*, **83**, 912–915.

Dabrowska, D. M. (1988) Kaplan–Meier estimate on the plane. *Ann. Statist.*, **16**, 1475–1489.

Dabrowska, D. M. (1989) Kaplan–Meier estimate on the plane: Weak convergence, LIL, and the bootstrap. *J. Multivariate Analysis*, **29**, 308–325.

Dabrowska, D. M., Lee, W. (1996) Nonparametric estimation of transition probabilities in a two-stage duration model. *Nonparametric Statist.*, **7**, 75–103.

Efron, B. (1988) Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.*, **83**, 414–425.

Gill, R. D. (1992) Multivariate survival analysis. *Theory Probab. Appl.*, **37**, 18–31, 284–301.

Gill, R. D., van der Laan, M., Wellner, J. A. (1995) Inefficient estimators of the bivariate survival function. *Annales de l'I.H.P., Probabilités et Statistiques*, **31** 545–597.

Hornsteiner, U., Hamerle, A. (1996) *A combined GEE/Buckley-James method for estimating an accelerated failure time model of multivariate failure times.* Discussion Paper 47, Sfb386, München.

Hornsteiner, U., Hamerle, A., Michels, P. (1997) *Parametric vs. nonparametric treatment of unobserved heterogeneity in multivariate failure times.* Discussion Paper 80, Sfb386, München.

Hornsteiner, U. (1998) *Statistische Analyse multivariater Ereignisdaten mit Anwendungen in der Werbewirkungsforschung und in der Kardiologie.* Dissertation, Regensburg.

Ibragimov, I. A., Has'minskii, R. Z. (1981) *Statistical Estimation. Asymptotic Theory.* Springer, Berlin.

Lai, T. L., Ying, Z. (1991) Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Statist.*, **19**, 1370–1402.

Lai, T. L., Ying, Z. (1994) A missing information principle and M-estimators in regression analysis with censored and truncated data. *Ann. Statist*, **22**, 1222–1255.

Lee, E. W., Wei, L. J., Ying, Z. (1993) Linear regression analysis for highly stratified failure time data. *J. Am. Statist. Assoc.*, **88**, 557–565.

Leurgans, S. (1987) Linear models, random censoring and synthetic data. *Biometrika*, **74**, 301–309.

Li, Q. H., Lagakos, S. W. (1997) Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Statist. Med.*, **16**, 925–940.

Lin, J. S., Wei, L. J. (1992) Linear regression analysis for multivariate failure time observations. *J. Am. Statist. Assoc.*, **87**, 1091–1097.

Miller, R. G. (1976) Least squares regression with censored data. *Biometrika*, **63**, 449–464.

Pan, W., Kooperberg, C. (1999) Linear regression for bivariate censored data via multiple imputation. *Statist. Med.*, **18**, 3111–3121.

Prentice, R. L. (1999) On non-parametric maximum likelihood estimation of the bivariate survivor function. *Statist. Med.*, **18**, 2517–2527.

Prentice, R. L., Cai, J. (1992) Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, **79**, 495–512.

Prentice, R. L., Hsu, L. (1997) Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika*, **84**, 349–363.

Pruitt, R. C. (1991) On negative mass assigned by the bivariate Kaplan–Meier estimator. *Ann. Statist.*, **19**, 443–453.

Pruitt, R. C. (1993) Small sample comparison of six bivariate survival curve estimators. *J. Statist. Comp. Simul.*, **45**, 147–167

Ritov, Y. (1990) Estimation in a linear regression model with censored data. *Ann. Statist.*, **18**, 303–328.

Schmee, J., Hahn, G. J. (1979) A simple method for regression analysis with censored data. *Technometrics*, **21**, 417–432.

Severini, T. A., Wong, H. (1992) Profile likelihood and conditionally parametric models. *Ann. Statist.*, **20**, 1768–1802.

Tsai, W.-Y., Crowley, J. (1998) A note on nonparametric estimators of the bivariate survival function under univariate censoring. *Biometrika*, **85**, 573–580.

van der Laan, M. J. (1995) *Efficient and Inefficient Estimation in Semiparametric Models.* CWI Tracts, Amsterdam.

van der Laan, M. J. (1996) Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.*, **24**, 596–627.

van der Laan, M. J. (1997) Nonparametric estimators of the bivariate survival function under random censoring. *Statistica Neerl.*, **51**, 178–200.

Wei, L. J., Lin, D. Y., Weissfeld, L. (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Statist. Assoc.*, **84**, 1065–1073.

Wu, C.-S. P., Zubovic, Y. (1995) A large-scale Monte Carlo study of the Buckley-James estimator with censored data. *J. Statist. Comp. Sim.*, **51**, 97–119.