CMPO Working Paper Series No. 03/082

What you always wanted to know about censoring but never dared to ask

Parameter estimation for censored random vectors

Wendelin Schnedler

CMPO, University of Bristol

July 2003

Abstract

This article considers a wide class of censoring problems and presents a construction rule for an objective function. This objective function generalises the ordinary likelihood as well as particular "likelihoods" used for estimation in several censoring models. Under regularity conditions the maximiser of this generalised likelihood has all the properties of a maximum likelihood estimator: it is consistent and the respective root-n estimator is asymptotically efficient and normally distributed.

Keywords: censored variables, m-estimation, multivariate methods, random censoring, generalised likelihood

JEL Classification: C130, C240

Acknowledgements

I am indebted to Joseph Lafranchi, who initiated my research on this topic by confronting me with a censoring problem; as well as to Ingolf Dittmann, Bernd Fitzenberger, Winfried Pohlmeier, Uwe Sunde and seminar participants at Bristol University for helpful comments.

Address for Correspondence
Department of Economics
University of Bristol
12 Priory Road
Bristol
BS8 1TN
W.Schnedler@bristol.ac.uk

CMPO is funded by the Leverhulme Trust.

1 Introduction

Often, the value of a variable of economic interest can only be observed under particular circumstances – the variable is censored. Ignoring censoring may lead to inconsistent estimators. The seminal example is from Tobin (1958): Because household expenditure is only observed when it is positive, ordinary least squares estimators for the relationship between household expenditure and income are downwardly biased. To circumvent this problem, Tobin suggested the maximiser of a particular objective function as an estimator. Every observational unit contributes to this objective function and the contribution can take two forms: If the value is observed, the contribution is the density evaluated at the observed value. If the value is not observed, the contribution is the probability of not observing the value. So, the objective function is neither a density nor a probability function evaluated at the observed value; rather, it is a hybrid of both. Hence, it is not a likelihood function.

Since Tobin's discovery, a plethora of censoring problems has been estimated using objective functions which hybridise density and probability contributions. In fact, respective estimators have long found their way into econometrics textbooks. However, as Davidson and MacKinnon correctly point out (1993, p. 539), there is something "fishy" about such objective functions; despite the fact that they are usually called likelihood functions, their maximiser does not necessarily feature the properties of a maximum likelihood estimator. It might not even be consistent.

If the censoring problem is of the simple nature that characterises the original Tobin model, that is if the variable cannot be observed below a fixed threshold and if errors are normally distributed, there is no reason to worry. Under these conditions, Amemiya (1973) proves that the maximiser of the objective function suggested by Tobin has the properties of a maximum likelihood estimator. But many censoring problems do not fall into this category and hence it is not clear whether the maximiser of the objective function features the desired properties. While initially authors wandered about this deficiency and justified their objective functions (for example by monte carlo simulation, see Nelson 1977), more recent applications are less cautious (e.g. Attanasio 2000). A possible way to ensure the desired properties is to use standard results about M-estimators (Amemiya 1985, Newey and McFadden 1996). But taking this route requires the constant re-invention of the wheel: For every censoring problem and the respective objective function very similar conditions need to be checked. Possibly this is the reason why

M-estimator results are rarely evoked to justify objective functions in the censoring context. Finding out whether the maximiser of the objective function has certain properties is only the second step when estimating with censored data. Before getting to this stage, one has to construct an objective function. While there seems to be a lot of working knowledge, intuition, and experience involved in this process, there is no explicit rule how to derive such a function. So, we are left with two interrelated problems: How do we find an objective function for a given censoring problem? How do we ensure that its maximiser has desirable properties? Both questions will be addressed subsequently.

This article provides a rule how to construct an objective function for a very general class of censoring problems. Under regularity conditions, the maximiser of this objective function features important properties of a maximum likelihood estimator. Namely, it is consistent, invariant to monotone transformations, and root-n times the estimator is asymptotically normal and efficient. If there is no censoring, this estimator becomes the ordinary maximum likelihood estimator. Thus, the estimation method is in a sense a generalisation of maximum-likelihood estimation.

The next section lays out a formal description of the class of censoring problems considered and explains how to construct a generalised likelihood for a given censoring problem. In section 3, this construction method is applied to particular problems, which were addressed in the literature. In most of the cases, the respective "likelihood"-functions are monotone transformations of the generalised likelihood. This implies that the respective maximiser has maximum-likelihood properties. For some cases, comparing "likelihood" and generalised likelihood represents a simple complementary way of proving their properties. For other cases, it is the first time that these properties are proven. Section 4 gives the regularity conditions under which the maximiser of the generalised likelihood has the desired properties. The proofs are based on standard results for M-estimators and very similar to the respective proofs for maximum-likelihood estimators. Section 5 embeds the problem of censoring in a regression context. Finally, section 6 concludes.

2 Constructing the objective function

Consider the following example. An employer reimburses the moving costs for workers, who come from a different city. To keep costs low, the employer uses the following rule: the worker has to obtain two quotes and the cheaper

one is reimbursed. Suppose for simplicity, that there are only two moving companies and that the parameters of economic interest are the means of the price offers by these two companies. However, the employer only keeps track of the reimbursed costs and the name of the selected moving company. The average reimbursed costs when a particular company was chosen underestimates the mean price offer made by this company. The offer of a company is simply more likely to be observed when it is lower. Is there any way to consistently estimate this mean?

This example is a special case of the more general problem, how to estimate a p-dimensional parameter $\theta \in \mathbb{R}^p$, which governs a continuously distributed random vector $Y = (Y_1, \ldots, Y_q)$ with a joint density function $f(y, \theta)$ and realisation $y = (y_1, \ldots, y_q)$ when some components of y cannot be observed sometimes. To advance on this issue, we assume that the econometrician knows at least the conditions under which the components are not observable. Suppose further that any such condition can be expressed in terms of components of y. In other words, the observability of the j-th component of the realisation of Y depends on the random vector Y itself. Formally, $\mathcal{V}_j$ denotes the set of realisations of Y such that y_j is observable whereas $\bar{\mathcal{V}}_j$ is the complement, that is the set of realisations y of Y for which y_j is not observable. The set $\mathcal{V}_j$ is called *visibility set*.

In the example, Y is the random variable describing the moving costs, $\theta = (\mu_1, \mu_2)$ are the means of the price offer distribution, and $y = (y_1, y_2)$ are the actual price offers submitted by the two moving companies. The price offer by the first moving company y_1 is observed whenever y_1 is smaller than y_2. Conversely, y_2 is observed when it is smaller than y_1. So, the visibility set for y_1 is $\mathcal{V}_1 = \{y_1, y_2 | y_1 < y_2\}$ and that for y_2 is $\mathcal{V}_2 = \{y_1, y_2 | y_2 \leq y_1\}$.

Is it restrictive to assume that the conditions for observing a component of Y can be expressed in terms of realisations of Y? Not really, because we are not limited to a particular random vector Y. So if –for example– one component is censored below a random threshold, which is distributed with a particular distribution, we can add a component to Y, which follows this distribution and use it to write down the visibility set.

Based on the visibility set, we can define a random variable, which describes whether a particular component is visible or not. Consider the function $V_j(y)$,

which takes on the value one if component j is visible and zero else:

$$V_j(y) := \begin{cases} 1 \text{ if } y \in \mathcal{V}_j \\ 0 \text{ if } y \notin \mathcal{V}_j. \end{cases}$$

Then, $V_j := V_j(Y)$ is a random variable indicating visibility. For this random variable to be well defined, we must be able to compute the probability of observing the j-th component. Hence, we make the following assumption.

Assumption 1. *For all j, the visibility set $\mathcal{V}_j$ is (Lebesgue-)measurable.*

This implies that the vector of random variables $V := (V_1, \ldots, V_q)$ is also well defined. Each realisation $v = (v_1, \ldots, v_q)$ of this vector may be regarded as a *visibility state* s. A state s is thus characterised by a vector $v^s = (v_1^s, \ldots, v_q^s)$, where the j-th component indicates whether the respective variable can be observed in this state or not. Because any of the q components in the vector v^s can take on two values, there are exactly 2^q states. These states can be numbered $s = 0, \ldots, 2^q - 1$, where the label $s = 0$ is reserved for that state in which no component is visible: $v_0 = (0, \ldots, 0)$. A particular state s realises if and only if the associated visibility v^s occurs; according to the definition of visibility, this is the case if and only if the realisation y of Y is in the respective visibility sets:

$$y \in \underbrace{\bigcap_{\{j|v_j^s=1\}} \mathcal{V}_j \cap \bigcap_{\{j|v_j^s=0\}} \bar{\mathcal{V}}_j}_{=:\mathcal{V}^s}.$$

The *state set* $\mathcal{V}^s$ condenses the restrictions placed on the realisation y in state s by the fact that certain variables are visible in this state and others are not. Together, all state sets $\{\mathcal{V}^s\}_{s=0,\ldots,2^q-1}$ form a disjoint decomposition of the $\mathbb{R}^q$ (proof see appendix). This result is important, as it will later ensure that each realisation contributes to the objective function only via one state. The probability of a particular state s is equal to:

$$P(S = s) = P(y \in \mathcal{V}^s) = \int_{\mathcal{V}^s} f(y, \theta) dy,$$

where S is a random variable describing the visibility state before it is realised.

In the moving company example, there are four states of visibility: neither variable is observable ($s = 0$), only y_1 is observable ($s = 1$), only y_2 is observable ($s = 2$), or both are observable ($s = 3$) – see figure 1. However, the

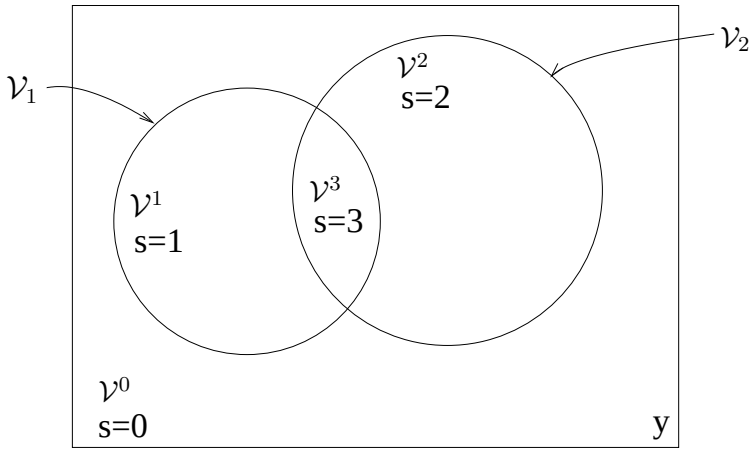

Figure 1: Relation between states, state sets, and visibility sets

If the realisation y of Y is in the visibility set $\mathcal{V}_1$, its first component y_1 is observable; if it is in the visibility set $\mathcal{V}_2$, the second can be observed. This induces four states: In state $s = 0$, no component can be observed and the realisation must be in the state set $\mathcal{V}^0$. In state $s = 1$, only the first component is visible and $y \in \mathcal{V}^1$. For the state $s = 2$ only the second and for $s = 3$ both components are visible while y is in the sets $\mathcal{V}^2$ and $\mathcal{V}^3$.

sets $\mathcal{V}^0$ and $\mathcal{V}^3$ are empty and the respective probability is zero: There is always exactly one price offer, which can be observed. If Y is jointly normal distributed with mean θ and variance-covariance matrix Σ, the probability for $s = 1$ is $P(S = 1) = P(y \in \mathcal{V}^1) = P(y_1 < y_2) = \Phi(y_2|\theta, \Sigma)$, where $\Phi(\cdot|\theta, \Sigma)$ is the cumulative density of the normal distribution.

Because estimators have to be defined in terms of observables and because some components of y are sometimes not observed, an operator which extracts the observable components of y in a given state s is very useful for the notation of estimators.

Definition 1. *Denote by ν^s an operator which extracts the visible components of y in state s:*

$$\begin{aligned} \nu^s : \mathbb{N} \times \mathbb{R}^q &\longrightarrow \mathbb{R}^{l(s)} \subseteq \mathbb{R}^q \\ (s, y) &\longmapsto (y_{j_1}, y_{j_2}, \ldots, y_{j_{l(s)}}), \end{aligned}$$

where $j_1, \ldots, j_{l(s)} \in \{j | v_j^s = 1\}$ and $l(s)$ is is the number of observable components in state s. Define $\bar{\nu}^s$ to be an operator which extracts the unobservables

components:

$$\begin{aligned} \bar{\nu}^s : I\!N \times I\!R^q &\longrightarrow I\!R^{q-l(s)} \subseteq I\!R^q \\ (s, y) &\longmapsto (y_{j_1}, y_{j_2}, \ldots, y_{j_{q-l(s)}}), \end{aligned}$$

where $j_1, \ldots, j_{q-l(s)} \in \{j | v_j^s = 0\}$.

To see how the operators work, reconsider the moving company example: $\nu^1 y = y_1$ because the observable component in state $s = 1$ is y_1, while $\nu^2 y = y_2$ in state $s = 2$ when the second provider submitted the lower bid ($y_2 < y_1$). Likewise the unobservable component in state $s = 1$ can be obtained by the invisibility operator $\bar{\nu}^1 y = y_2$ and similarly the unobservable component in the state $s = 2$ can be extracted: $\bar{\nu}^1 y = y_2$.

Let $i = 1, \ldots, n$ be the index of n observational units which are randomly sampled from Y; a particular realisation of this random sample is denoted by $y_{i\cdot} = (y_{i1}, \ldots, y_{iq})$ and leads to a state s_i. Then, the visibility operator allows the following succinct representation of the data which is available for estimation:

$$(s_i, \nu^{s_i} y_{i\cdot})_{i=1,\ldots,n}.$$

In words: the econometrician knows which variables are observable and the values for those variables.

The conditional distribution given a particular state s is characterised by the respective conditional density function. This function can be obtained by integrating over all components which are unobservable in that state ($\bar{\nu}^s y$) and dividing by the probability of the state to occur:

$$f_s(\nu^s y, \theta) = \frac{1}{P(S = s)} \int\limits_{\mathcal{V}^s} f(y, \theta) d(\bar{\nu}^s y).$$

The actual observations for a given state s are adequately described by this conditional density. Hence, the available data for a given state s, which are $\{\nu^{s_i} y_i | s_i = s\}$, can be used to obtain an ordinary maximum likelihood estimator of θ. Of course, it might be problematic to identify θ using the –possibly few– observable components of y in state s. Moreover, the described technique may be carried out for various observed states s and therefore lead to a multitude of estimators for θ. Ideally, we want to combine the information of different states to better identify θ, to increase efficiency, and to obtain a single estimator which incorporates all available information.

Based on the introduced notation, we now devise such an estimator. To derive the contribution of a particular observational unit, we start out with the density $f(y, \theta)$. The respective realisation y leads to a visibility state s. Some components of y are not observable in this state, so we integrate them out. But the unobserved components cannot have any value, they are restricted by the fact that we are in state s. Hence, we do not integrate over the whole domain but limit integration to the visibility set $\mathcal{V}^s$. Overall, we get:

$$\tilde{f}_s(\nu^s y, \theta) := \int_{\mathcal{V}^s} f(y, \theta) d(\bar{\nu}^s y),$$

where the integration is simply ignored if there are no unobserved components $\bar{\nu}^s y$. This function is closely related to the conditional density:

$$\tilde{f}_s(\nu^s y, \theta) = P(S = s) f_s(\nu^s y, \theta)$$

But it is simpler to compute because the probability of the state s need not be calculated. Note that $\tilde{f}_s(\nu^s y, \theta)$ is not a density function as integrating over the remaining variables yields $P(S = s)$ rather than one. Next, consider an objective function to which each observation $(s_i, \nu^{s_i} y_{i\cdot})$ contributes by $\tilde{f}_{s_i}(\nu^{s_i} y_{i\cdot}, \theta)$. Then, define the following estimator:

$$\theta_n := \text{argmax}_\theta \prod_{i=1}^{n} \tilde{f}_{s_i}(\nu^{s_i} y_{i\cdot}, \theta) \tag{1}$$

Note, that the objective function is not an ordinary likelihood function because it is not the product of density functions. Consequently, the criticism of Davidson and MacKinnon (1993) applies and the maximiser does not necessarily have the usual properties of a maximum likelihood estimator. Later, we determine when the maximiser in (1) is consistent and root-n times the estimator is asymptotically normally distributed and efficient. If there is no censoring, the objective function in (1) is an ordinary likelihood function. Because most properties of a likelihood estimator are preserved for the maximiser in (1), the respective objective function is called *generalised likelihood*.

3 Application

The developed framework covers a large range of censoring problems. Accordingly, maximising the generalised likelihood provides estimators with desirable properties for these censoring problems. This section reconsiders some censoring problems, derives the generalised likelihood, and compares

it with "likelihood functions" that were used for the respective problem by other authors.

Recall the simple tobit model in which a (one-dimensional) realisation y_1 cannot be observed when it is below zero. This model has two states $s_i = 0$ and $s_i = 1$. In the state $s_i = 0$, the variable is not observable since $y_1 < 0$; respectively $\nu^1 y = \bar{\nu}^0 y = y_1$. Thus, the visibility set of the variable is $\mathcal{V}_1 = \{y_1 | y_1 \geq 0\}$, so that the state sets are $\mathcal{V}^0 = \bar{\mathcal{V}}_1 = \{y_1 | y_1 < 0\}$. and $\mathcal{V}^1 = \mathcal{V}_1$. Given a normally distributed y_1 with mean μ and variance σ^2, the calculated contribution for state $s = 0$ is:

$$\tilde{f}_0(y_1, \theta) = \int_{\mathcal{V}^0} f(y, \theta) d\bar{\nu}^0 y = P(y_1 < 0) = \Phi(\mu; \sigma^2),$$

where $\theta = (\mu, \sigma^2)$. If y_1 is observable ($s = 1$), the formula for the contribution yields:

$$\tilde{f}_1(y_1, \theta) = f(y_1, \theta) = \phi(y_1 \{\mu, \sigma^2),$$

where $\phi(\cdot|\mu, \sigma^2)$ is the normal density. Hence, the objective function from equation (1) becomes:

$$\prod_{\{i|s_i=0\}} \Phi(0|\mu, \sigma^2) \prod_{\{i|s_i=1\}} \phi(y_1|\mu, \sigma^2)$$

This, however, is exactly the objective function which is usually used to obtain the tobit estimator. Consequently, this estimator is a special case of estimator (1) and all properties which are valid for this estimator are also valid for the tobit estimator. In the case of the tobit estimator, this might not be very exciting since Amemiya (1973) has already derived its properties. However, for other estimators, the properties of which have not been proven, the method is more useful.

As an example take the tobit type II model as introduced by Amemiya (1984). In this model, there are two components $y = (y_1, y_2)$ which are normally distributed around the means μ_1 and μ_2 with variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \text{ so that } \theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}).$$

The realisation y_1 is never observable but y_2 is observable whenever $y_1 > 0$. So, there are two states $s = 0$ and $s = 1$ and the visibility set for y_2 is $\mathcal{V}_1 = \{(y_1, y_2) | y_1 > 0\}$, while the state sets are $\mathcal{V}^0 = \bar{\mathcal{V}}_1 = \{(y_1, y_2) | y_1 \leq 0\}$

and $\mathcal{V}^1 = \mathcal{V}_1 = \{(y_1, y_2)|y_1 > 0\}$. The respective contribution for state $s = 0$ is:

$$\tilde{f}_0(y_2, \theta) = \int_{\mathcal{V}^0} f(y, \theta) d\bar{\nu}^0 y = P(y_1 \cdot 0) = \Phi(0|\mu_1, \sigma_{11})$$

while the state $s = 1$ contributes:

$$\tilde{f}_1(y_2, \theta) = \int_{\mathcal{V}^1} f(y, \theta) d\bar{\nu}^1 y = \phi(y_2|y_1 > 0, \mu_1, \mu_2, \Sigma) P(y_1 > 0).$$

So again, the generalised likelihood coincides with the standard objective function given by Amemiya (1984):

$$\prod_{\{i|s_i=0\}} \Phi(0|\mu_1, \sigma_{11}) \prod_{\{i|s_i=1\}} \phi(y_2|y_1 > 0, \mu_1, \mu_2, \Sigma) P(y_1 > 0).$$

However, previously the properties of the maximiser of this objective function were not known. Identifying this maximiser with the estimator defined by (1) enables us to state that the tobit type II estimator is consistent and root-n asymptotically normally distributed. It can be shown that the objective function for the tobit models of type III to V according to Amemiya' s classification (1984) are also special cases of the objective function leading to (1). Hence, the respective maximisers all have desirable properties (under regularity conditions).

A different censoring problem was analysed by Nelson (1977). In this model, there are again two realisations y_1 and y_2 from normally distributed random variables and the same parameters as in the tobit type II model. This time the second component operates as an unobservable censoring threshold. That means y_1 is observable whenever it is above y_2, the visibility set for y_1 is $\mathcal{V}_1 = \{y_1 > y_2\}$. Nelson (1977) proposes the following "likelihood function":

$$\prod_{\{i|s_i=0\}} \Phi\left(\frac{\mu_2 - \mu_1}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}\right) \prod_{\{i|s_i=1\}} \int_{-\infty}^{y_1} \phi(y_1, y_2|\mu_1, \mu_2, \Sigma) dy_2.$$

He neither provides a proof nor a reference why the maximiser of his "likelihood function" is –for example– consistent, but he conducts a small simulation study which suggests that the maximiser has this property. The properties can be formally affirmed if the proposed "likelihood function" coincides with the general likelihood function from (1). To check this, compute the contribution for the state $s = 0$ in which y_1 is not observable is:

$$\tilde{f}_0(y_1, \theta) = \int_{-\infty}^{\infty} \int_{y_1}^{\infty} \phi(y_1, y_2) dy_2 dy_1 = \Phi\left(\frac{\mu_2 - \mu_1}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}\right).$$

The contribution for the state where y_1 can be observed ($s = 1$) is:

$$\tilde{f}_1(y_1, \theta) = \int\limits_{\mathcal{V}_0 = \bar{\mathcal{V}}_1} f(y, \theta) d\bar{\nu}^0 y = \int\limits_{-\infty}^{y_1} \phi(y_1, y_2 | \mu_1, \mu_2, \Sigma) dy_2.$$

This implies that Nelsons objective function is indeed a generalised likelihood function and that the estimator is consistent and has the respective other properties.

However, not all objective functions used for censoring models coincide with the generalised likelihood. Attanasio (2000) proposed a model for inertia when the stock of a durable commodity is adjusted. Observations are censored in a very sophisticated way because sometimes the initial stock and sometimes the final stock is not observable. If we construct state sets according to the proposed rule, they are different from the "groups" by which observations contribute in Attanasio's setting. Attanasio does not explicitly relate the groups to theoretical realisations of random variables. It is therefore difficult to assess whether any realisation contributes to the objective function and does so in a unique way. Overall, the properties of Attanasio's estimator cannot be deduced from the properties of estimator (1) and still remain to be explored.

The last censoring problem shows that there is by no-means a unique approach to construct objective functions from densities for estimation purposes. Likewise, the term "likelihood" is ambiguous when it comes to censoring problems. The virtue of the approach taken here is that it allows a general treatment of these problems and leads to an estimator with known properties.

4 Properties of the maximiser

In this section, we analyse the properties of the estimator defined by (1). The first part deals with consistency, the second part with asymptotic normality, and the third part with efficiency.

4.1 Consistency

In this section, consistency is proven by using a standard result on the consistency of M-estimators. As $\hat{\theta}_n$ is the maximiser of any monotone transformation of the objective function in (1), one can alternatively work with the

following objective function:

$$Q_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log \tilde{f}_{s_i}(v^{s_i} y_{i\cdot}, \theta) \tag{2}$$

and use the machinery of M-estimation to determine the properties of the maximiser and in particular to check whether it is consistent. To do so we employ the following standard result (see e.g. Amemiya 1985 or Newey and McFadden 1996):

Theorem 1 (Consistency of M-estimators). *If there are measurable functions $Q_n(\theta)$ and a non-stochastic function $Q_0(\theta)$ such that (i) $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$, (ii) $Q_0(\theta)$ is continuous, (iii) $Q_0(\theta)$ is uniquely maximised at θ_0, and (iv) the parameter space is compact, then $\hat{\theta}_n$ is consistent for θ_0: $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

The rest of this section will be devoted to find primitive conditions for (i) to (iii) to hold.

The objective function needs to converge to the non-stochastic function $Q_0(\theta)$. Before we are able to find the maximiser of the limit of the objective function, we must ensure that this function exists and is finite. Thus, we assume that the expected value of the logarithm of the density exists:

$$\mathrm{E}|\log f(Y, \theta)| < \infty. \tag{FIN}$$

From this condition on the general density, we can conclude the finiteness of the expectation of the contributions of state s.

Lemma 1. *From* (FIN) *follows:*

$$E_{Y|S}\left|\log(\tilde{f}_s(\nu^s Y), \theta)\right| < \infty.$$

Proof. By the mean-value theorem for integrals, we can rewrite the contribution of state s:

$$\tilde{f}^s(\nu^s y) = \int_{\mathcal{V}^s} f(h(\nu^s y, \bar{\nu}^s \tilde{y}), \theta) d(\bar{\nu}^s y) = f(h(\nu^s y, \zeta)) \text{ for some } \zeta \in \mathcal{V}^s, \tag{3}$$

where h is the appropriate permutation of the values such that y_1 is the first argument of $f(\cdot)$ and y_n is the last. Now, take condition (FIN) and rewrite

it.

$$\begin{aligned}\infty &> \mathrm{E}\left|\log\left\{f(Y,\theta)\right\}\right| \\ &= \mathrm{E}_S\left[\mathrm{E}_{Y|S}\left|\log\left\{f(Y,\theta)\right\}\right|\right] \\ &= \mathrm{E}_S\left[\mathrm{E}_{\bar{\nu}^s Y,\nu^s Y|S}\left|\log\left\{f(Y,\theta)\right\}\right|\right] \\ &= \mathrm{E}_S\left[\mathrm{E}_{\bar{\nu}^s Y|S}\left[\mathrm{E}_{\nu^s Y|\bar{\nu}^s Y,S}\left|\log\left\{f(Y,\theta)\right\}\right|\right]\right].\end{aligned}$$

This implies:

$$\begin{aligned}&\forall(\bar{\nu}^S y):\ \mathrm{E}_{(\nu^S Y)|(\bar{\nu}^S Y),S}\left|\log\left\{f(h(\nu^S Y,(\bar{\nu}^S Y)),\theta)\right\}\right| < \infty \\ \Rightarrow\ &\mathrm{E}_{(\nu^s Y)|\zeta,S}\left|\log\left\{f(h(\nu^s Y,\zeta),\theta)\right\}\right| < \infty.\end{aligned}$$

Together with (3), we get:

$$\infty > \mathrm{E}_{(\nu^s Y)|\zeta,S}\left|\log\left\{f(h(\nu^s Y,\zeta),\theta)\right\}\right| = \mathrm{E}_{(\nu^s Y)|S}\left|\log\left\{\tilde{f}_s(\nu^s Y,\theta)\right\}\right|.$$

□

Using the finiteness and the law of the large numbers, we can determine the probability limit of the objective function when the number of observations tends to infinity.

Proposition 1 (Convergence). *Given condition* (FIN)*, $Q_n(\theta)$ converges uniformly in probability to*

$$Q_0(\theta) = \sum_s \int_{\mathcal{Y}^s} \log\left\{\tilde{f}_s(\nu^s y,\theta)\right\}\ \tilde{f}_s(\nu^s y,\theta_0)\, d(\nu^s y). \tag{4}$$

Proof. Since observational units are drawn independently, $Q_n(\theta)$ is the mean of independent random variables. The law of the large numbers applies, and the mean converges in probability to its expected value

$$\begin{aligned}&\mathrm{E}_{Y|\theta_0}\left[\log\left\{\tilde{f}_S(\nu^S Y,\theta)\right\}\right] \\ =\ &\mathrm{E}_{S|\theta_0}\left[\mathrm{E}_{Y|S,\theta_0}\left[\log\left\{\tilde{f}_S(\nu^S Y,\theta)\right\}\right]\right],\ \text{which is finite by (FIN)} \\ =\ &\sum_{s=0}^{2^q-1} P(S=s,\theta_0)\int_{\mathcal{Y}^s}\log\left\{\tilde{f}_s(\nu^s y,\theta)\right\} f_s(\nu^s y,\theta_0) d(\nu^s y) \\ =\ &\sum_{s=0}^{2^q-1}\int_{\mathcal{Y}^s}\log\left\{\tilde{f}_s(\nu^s y,\theta)\right\}\tilde{f}_s(\nu^s y,\theta_0) d(\nu^s y).\end{aligned}$$

□

Next, we have to ensure continuity of the limiting objective function $Q_0(\theta)$ so that stochastic convergence of the argument leads to stochastic convergence of the values of the function.

Proposition 2 (Continuity). *If*

$$f(y,\theta) \text{ is continuous in } \theta \quad \text{(CON)},$$

$Q_0(\theta)$ *is continuous in* θ.

Proof. If $f(y,\theta)$ is continuous in θ, $\tilde{f}_s(\nu^s y,\theta) = \int_{\mathcal{Y}^s} f(y,\theta)d(\bar{\nu}^s y)$ is continuous in θ, and so is $Q_0(\theta)$. □

Like in the case of maximum likelihood estimation, it must be possible to extract the desired information about parameters from the observations. Two *different* parameter values which generate the *same* observations cannot be distinguished. In other words it must be possible to identify the parameter from the observations. We define the statistical model to be *identified* if and only if

$$\forall \theta \neq \theta' \, \exists s : \, P(S=s) > 0 : \, f_s(\nu^s y,\theta) \neq f_s(\nu^s y,\theta'). \qquad \text{(ID)}$$

In other words, there must exist at least one state under which differences in the parameter translate into differences in the conditional density. Similarly, to the identification condition in maximum likelihood estimation, this condition may be difficult to verify. The next result proves that the parameter is indeed uniquely determined when the condition can be verified. The proof is very similar to the proof of the uniqueness of the maximiser of the limiting objective function when working with ordinary likelihoods.

Proposition 3 (Unique maximiser). *Under* (ID) *and* (FIN), $Q_0(\cdot)$ *is uniquely maximised at the true parameter* θ_0.

Proof. Consider the difference between the limiting objective function evaluated at the true parameter, $Q_0(\theta_0)$, and at a different parameter $Q_0(\theta)$:

$$\begin{aligned} Q_0(\theta_0) - Q_0(\theta) &= \mathrm{E}_{Y|\theta_0}\left[\log\left\{\tilde{f}_s(\nu^s y,\theta_0)\right\} - \log\left\{\tilde{f}_s(\nu^s y,\theta)\right\}\right] \\ &= \mathrm{E}_{Y|\theta_0}\left[-\log\left\{\frac{\tilde{f}_s(\nu^s y,\theta)}{\tilde{f}_s(\nu^s y,\theta_0)}\right\}\right] \\ &> \mathrm{E}_{S|\theta_0}\left[-\log\left\{\mathrm{E}_{Y|S,\theta_0}\left[\frac{\tilde{f}_s(\nu^s y,\theta)}{\tilde{f}_s(\nu^s y,\theta_0)}\right]\right\}\right], \end{aligned} \qquad (5)$$

where the last inequality follows from the strict version of Jensen's inequality for non-constant random variables. By (ID), the expected value is indeed

taken over a non-constant random variable. As $\mathrm{E}_{Y|S}\left[\frac{\tilde{f}_s(\nu^s y,\theta)}{\tilde{f}_s(\nu^s y,\theta_0)}\right] = 1$, we get $Q_0(\theta_0) - Q_0(\theta) > 0$, and θ_0 is the unique maximum. □

The consistency of the estimator (1) can now be deduced from the conditions (FIN), (CON), and (ID) which together with the propositions 1 to 3 imply that the requirements of the standard consistency theorem are valid.

Theorem 2 (Consistency of the estimator). *If there are measurable functions $Q_n(\theta)$,* (FIN)*,* (CON)*,* (ID) *hold, and the parameter space is compact, then $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

It is simple to construct an alternative to estimator (1) by replacing $\tilde{f}_s(\cdot)$ by the conditional density contributions $f_s(\cdot)$ where the "density" for state $s = 0$ is defined as $f_0(y,\theta) \equiv P(S = 0,\theta)$. Multiplying these conditional density contributions, we get the *state conditional likelihood function* Q_n^{SCL}; the maximiser will be referred to as *state conditional likelihood estimator* or as *SCL-estimator*. All propositions and proofs in this section can be adapted to the state conditional likelihood estimator, so that it is also consistent under (FIN), (CON), (ID). In addition to the evaluation of an integral, which is also necessary to obtain the generalised likelihood, the state conditional likelihood requires the computation of the probability of all states. If there is no closed form for the respective probabilities, this will increase the computational effort substantially. To distinguish the two estimators the estimator resulting from the generalised maximum likelihood, $\hat{\theta}_n$ is called *GL-estimator.*

4.2 Asymptotic normality

Another property of maximum likelihood estimators is their asymptotic normality and efficiency. In this section, conditions are derived under which (1) has these properties. More precisely, we assume that the objective function in (1) has an interior maximum and examine the solution to the first-order condition which results from maximising the objective function. Again, the conditions resemble the respective conditions for maximum likelihood estimators. This is no coincidence, since the proof is based on a standard result for M-estimators (Theorem 4.1.3 in Amemiya 1985 where assumption B is replaced using Theorem 4.1.5):

Proposition 4 (Asymptotic normality of M-estimators). *If (i) $\hat{\theta}_n$, the maximiser of $Q_n(\cdot)$, is consistent for θ_0, (ii) θ_0 lies in the interior of the parameter space Θ, (iii) Q_n is twice continuously differentiable in an open and convex neighbourhood $\mathcal{N}$ of θ_0, (iv) $\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0} \xrightarrow{d} N(0,J)$, (v)*

$\nabla_{\theta\theta} Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{P} H(\theta_0)$ *with* $H(\theta)$ *finite, non-singular, and continuous at* θ_0, *then* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, H^{-1}JH^{-1})$.

Under the assumptions of theorem 2, condition (i) is valid. Condition (ii) ensures that the maximum is not a corner solution and hence that the first derivative of $Q_0(\cdot)$ disappears at θ_0. Subsequently, conditions (iii) to (v) should be replaced by primitive conditions on the density $f(y, \theta)$.

We begin by assuming:

$$f(y, \theta) \text{ is twice continuously differentiable at } \theta_0. \tag{DIFF}$$

Denote the operator which yields the first derivative of a vector-valued function by ∇_θ and use the convention that this operator turns a real-valued component of the function into a $(1, p)$-vector, where the first value is the derivative with respect to θ_1, the second with respect to θ_2 and so forth. Likewise $\nabla_{\theta\theta}$ is the operator which gives the second derivative of a real-valued function, the Jacobian matrix. We want the differentiation operators to be exchangeable with integration which is for example fulfilled if the area over which is integrated does not depend on θ:

$$\nabla_\theta \int f(y, \theta) dy = \int \nabla_\theta f(y, \theta) dy \qquad \nabla_{\theta\theta} \int f(y, \theta) dy = \int \nabla_{\theta\theta} f(y, \theta) dy, \tag{EID}$$

where we require the equalities to hold only evaluated in the neighbourhood of $\theta = \theta_0$. We can use the exchangebility to compute the first and second derivative of $\tilde{f}^s(\nu^s y, \theta)$ with respect to θ at θ_0:

$$\begin{aligned} \nabla_\theta \tilde{f}^s(\nu^s y, \theta)\Big|_{\theta=\theta_0} &= \int_{\mathcal{V}^s} \nabla_\theta f(y, \theta) d(\bar{\nu}^s y)\Big|_{\theta=\theta_0} \\ \nabla_{\theta\theta} \tilde{f}^s(\nu^s y, \theta) d(\bar{\nu}^s y)\Big|_{\theta=\theta_0} &= \int_{\mathcal{V}^s} \nabla_{\theta\theta} f(y, \theta) d(\bar{\nu}^s y)\Big|_{\theta=\theta_0} \end{aligned} \tag{6}$$

Next, define:

$$J(\theta) := \mathrm{E}_{Y|\theta_0} \left[\frac{\nabla_\theta \tilde{f}_s(\nu^s Y, \theta)' \nabla_\theta \tilde{f}_s(\nu^s Y, \theta)}{\tilde{f}_s(\nu^s Y, \theta) \tilde{f}_s(\nu^s Y, \theta)} \right], \tag{7}$$

where expectations are taken with respect to the true parameter θ_0 and where the prime denotes the transpose of a vector or matrix. Later, it will be proven that $J(\theta)$ is the second derivative of the limit function $Q_0(\theta)$ and

thus deserves the letter "J" indicating that it is a Jacobian matrix. Since, we want this second derivative to exist and to be finite at its maximiser θ_0, we suppose that

$$J := J(\theta_0) \text{ exists and is finite.} \tag{EX}$$

Under the defined conditions, it is now possible to calculate the distribution of the first derivative of the objective function:

Proposition 5 (Asymptotic normality of the first derivative). *Under* (DIFF), (FIN), (EX), (EID), *and if* J *defined in* (EX) *is non-singular, then*

$$\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0} \xrightarrow{d} N(0, J).$$

Proof. $\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0}$ can be rewritten as the sum of i.i.d. random variables:

$$\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\frac{\nabla_\theta \tilde{f}_s(\nu^{s_i} y_{i\cdot}, \theta)|_{\theta=\theta_0}}{\tilde{f}_s(\nu^{s_i} y_{i\cdot}, \theta)|_{\theta=\theta_0}}}_{=:Q^i_\nabla}, \tag{8}$$

and by the central-limit theorem its distribution converges to a normal distribution with mean $\mathrm{E}(\mathrm{Q}^{\mathrm{i}}_\nabla)$ and variance-covariance matrix $\mathrm{COV}\left[Q^i_\nabla\right]$. The existence of $\mathrm{E}(\mathrm{Q}^{\mathrm{i}}_\nabla)$ is assured by (EX) and Jensen's inequality, its value is:

$$\begin{aligned} \mathrm{E}(Q^i_\nabla) = &\quad \sum_s P(S=s,\theta_0) \int \frac{\nabla_\theta \tilde{f}_s(\nu^s y,\theta)|_{\theta=\theta_0}}{\tilde{f}_s(\nu^s y,\theta_0)} \cdot \frac{\tilde{f}_s(\nu^s y,\theta_0)}{P(S=s,\theta_0)} d(\nu^s y) \\ \overset{\text{(EID)}}{=} &\quad \sum_s P(S=s,\theta) \cdot \nabla_\theta \underbrace{\int \frac{\tilde{f}_s(\nu^s y,\theta)}{P(S=s,\theta)} d(\nu^s y)}_{=1}\Bigg|_{\theta=\theta_0} = 0. \end{aligned} \tag{9}$$

As the expected value is the zero vector, $\mathrm{COV}\left[Q^i_\nabla\right] = \mathrm{E}\left[(Q^i_\nabla)' Q^i_\nabla\right]$. By plugging in Q^i_∇ one immediately gets $\mathrm{COV}\left[Q^i_\nabla\right] = J$, the existence of which is ensured by (EX). □

Assumptions (DIFF), (EX), and (EID) do not only enable us to compute the first but also the limit of the second derivative when it is evaluated at the maximiser:

Proposition 6 (Convergence of the second derivative). *Given* (DIFF), (EX), (EID), *and* $\hat{\theta}_n \to \theta_0$, *it follows that* $\nabla_{\theta\theta} Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{P} -J$, *where* J *is positive definite.*

Proof. The second derivative of the objective function with respect to θ is:

$$\nabla_{\theta\theta} Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\tilde{f}_s(\nu^s y,\theta)\nabla_{\theta\theta}\tilde{f}_s(\nu^s y,\theta) - \nabla_\theta \tilde{f}_s(\nu^s y,\theta)'\nabla_\theta \tilde{f}_s(\nu^s y,\theta)}{\tilde{f}_s(\nu^s y,\theta)^2}}_{=:Q^i_{\nabla\nabla}}. \tag{10}$$

By the law of large numbers $Q^i_{\nabla\nabla}$ approaches its expected value:

$$\begin{aligned}
\mathrm{E}(Q^i_{\nabla\nabla}(\theta)) = &\quad \mathrm{E}\left(\frac{\nabla_{\theta\theta}\tilde{f}_s(\nu^s y,\theta)}{\tilde{f}_s(\nu^s y,\theta)}\right) - \mathrm{E}\left(\frac{\nabla_\theta \tilde{f}_s(\nu^s y,\theta)'\nabla_\theta \tilde{f}_s(\nu^s y,\theta)}{\tilde{f}_s(\nu^s y,\theta)^2}\right) \\
= &\quad \sum_s P(S=s,\theta_0)\int \nabla_{\theta\theta}\tilde{f}_s(\nu^s y,\theta)d(\nu^s yy)\cdot\frac{1}{P(S=s,\theta_0)} - J(\theta) \\
\overset{\text{(EID)}}{=} &\quad \sum_s P(S=s,\theta_0)\underbrace{\nabla_{\theta\theta}\int f(\nu^s y,\theta)d(\nu^s yy)}_{=0} - J(\theta) = -J(\theta).
\end{aligned} \tag{11}$$

As this expected value is continuous in θ around θ_0, we can use Theorem 4.1.5 in Amemiya (1985) to conclude that from $\hat{\theta}_n \xrightarrow{p} \theta_0$ it follows that $\mathrm{E}[Q^i_{\nabla\nabla}(\hat{\theta}_n)] \xrightarrow{p} \mathrm{E}\left[Q^i_{\nabla\nabla}(\theta_0)\right]$. So overall, we get $\nabla_{\theta\theta}Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{p} \mathrm{E}\left[Q^i_{\nabla\nabla}(\theta_0)\right] = -J$. As $-J$ is the second derivative of the objective function evaluated at a unique and interior maximum, it must be negative definite. So, J must be positive definite. □

Using theorem 2 and propositions 4 to 6, we can state:

Theorem 3 (Asymptotic normality of the GL-estimator). *If* (ID), (EX), (FIN), (EID), *and* (DIFF) *hold, and* θ_0 *is in the interior of the compact parameter space* Θ, *then* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, J^{-1})$.

Again, the results for the state conditional estimator $\hat{\theta}^{\mathrm{SCL}}_n$ can be derived by replacing the contributions $\tilde{f}_s(\cdot)$ by the conditional density contributions $f_s(\cdot)$ and the generalised likelihood Q_n by the state conditional likelihood Q^{SCL}_n in propositions 4 to 6. This yields the following result.

Corollary 1 (Asymptotic normality of the SCL-estimator). *If* (ID), (EX), (FIN), (EID), *and* (DIFF) *hold, and* θ_0 *is in the interior of the compact parameter space* Θ, *then* $\sqrt{n}(\hat{\theta}^{SCL}_n - \theta_0) \xrightarrow{d} N(0, (\Sigma^{SCL})^{-1})$, *where*

$$\Sigma^{SCL} := E_{Y|\theta_0}\left[\frac{\nabla_\theta f_s(\nu^s Y,\theta)'\nabla_\theta f_s(\nu^s Y,\theta)}{f_s(\nu^s Y,\theta)f_s(\nu^s Y,\theta)}\right]. \tag{12}$$

The asymptotic variance-covariance matrix of the derivative of the state dependent likelihood function Σ^{SCL} is identical to the variance-covariance matrix for the generalised likelihood J, if the probabilities of the various states do not depend on the unknown parameter: $\nabla_\theta P(S = s, \theta) = 0 \Rightarrow \Sigma^{\text{SCL}} = J$. However, generally the two matrices will not be identical. When they are different, the respective root-n estimators may not be equally asymptotically efficient. But which of the estimators has the smaller asymptotic variance-covariance matrix?

4.3 Asymptotic Efficiency

To show that root-n times the GL-estimator is asymptotically efficient, we proceed in two steps. First, we determine the Cramer-Rao lower bound for the class of censoring problems under consideration. This yields the "smallest" variance-covariance matrix which can be attained using the available (censored) information. Second, we observe that the asymptotic variance-covariance matrix of the root-n GL-estimator coincides with this lower bound. Hence, the root-n GL-estimator must be asymptotically efficient.

Proposition 7 (Cramer-Rao lower bound). *In the censoring problem described above and given that* (ID), (EX), (FIN), (EID), *and* (DIFF) *hold, the asymptotical variance-covariance matrix* $\lim_{n\to\infty} COV[\sqrt{n}T]$ *of any asymptotically unbiased estimator* $\sqrt{n}T$ *for* θ_0 *is larger or equal to* J *according to the Löwner ordering:*

$$\forall x : \lim_{n\to\infty} x' COV\left[\sqrt{n}T\right] x \geq x' J^{-1} x.$$

Proof. Denote the observable sample by $\nu^s y := (\nu^{s_1} y_{1.}, \ldots, \nu^{s_n} y_{n.})$ Let $\sqrt{n}T(\cdot)$ be an asymptotically unbiased estimator for the true parameter: $\mathrm{E}\left[T(\nu^S Y)\right] = \theta_0$, for $n \to \infty$. Then, write out the expected value using the independence of the observations:

$$\begin{aligned}
\theta_0 &= \lim_{n\to\infty} \mathrm{E}_{Y|\theta_0}\left[T(\nu^S Y)\right] \\
&= \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} P(S = s_1, \theta_0) \cdots \sum_{s_n=0}^{2^q-1} P(S = s_n, \theta_0) \cdot \\
&\qquad \cdot \int_{\mathcal{V}^{s_1}} \cdots \int_{\mathcal{V}^{s_n}} T(\nu^{s_1} y_1, \ldots, \nu^{s_n} y_n) \prod_{i=1}^{n} f_{s_i}(\nu^{s_i} y_i, \theta_0)\, d(\nu^{s_1} y_1) \cdots d(\nu^{s_n} y_n).
\end{aligned}$$

Using relationship $\tilde{f}_s(\nu^s y, \theta) = P(S = s) f_s(\nu^s y, \theta)$, this simplifies to:

$$\theta|_{\theta=\theta_0} = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{\mathcal{V}^{s_1}} \cdots \int_{\mathcal{V}^{s_n}} T(\nu^s y) \prod_{i=1}^{n} \tilde{f}_{s_i}(\nu^{s_i} y_i, \theta_0)\, d(\nu^{s_1} y_1) \cdots d(\nu^{s_n} y_n). \tag{13}$$

Now, take the derivative with respect to θ on both sides:

$$I = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{\mathcal{V}^{s_1}} \cdots \int_{\mathcal{V}^{s_n}} T(\nu^s y) \nabla_\theta \prod_{i=1}^{n} \tilde{f}_{s_i}(\nu^{s_i} y_i, \theta_0)\, d(\nu^{s_1} y_1) \cdots d(\nu^{s_n} y_n). \tag{14}$$

Next, consider the following sophisticated expression for a zero matrix:

$$\theta_0\, \nabla_\theta I = \theta_0\, \nabla_\theta \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{\mathcal{V}^{s_1}} \cdots \int_{\mathcal{V}^{s_n}} \nabla_\theta \prod_{i=1}^{n} \tilde{f}_{s_i}(\nu^{s_i} y_i, \theta_0)\, d(\nu^{s_1} y_1) \cdots d(\nu^{s_n} y_n). \tag{15}$$

Subtracting this sophisticated zero from the right-hand side in (14) yields:

$$I = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{\mathcal{V}^{s_1}} \cdots \int_{\mathcal{V}^{s_n}} (T(\nu^s y) - \theta_0)\, \nabla_\theta \tilde{f}_{s_\cdot}(\nu^s y, \theta_0)\, d(\nu^{s_1} y_1) \cdots d(\nu^{s_n} y_n),$$

where $\tilde{f}_{s_\cdot}(\nu^s y, \theta_0) := \prod_{i=1}^{n} \tilde{f}_{s_i}(\nu^{s_i} y_i, \theta_0)$. By multiplying with and dividing by

$$\sqrt{n} f_{s_\cdot}(\nu^s y, \theta_0) := \prod_{i=1}^{n} f_{s_i}(\nu^{s_i} y_i, \theta_0),$$

we get:

$$I = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{\mathcal{V}^{s_1}} \cdots \int_{\mathcal{V}^{s_n}} \sqrt{n}\,(T(\nu^s y) - \theta_0)\, \frac{\nabla_\theta \tilde{f}_{s_\cdot}(\nu^s y, \theta_0)}{\sqrt{n} f_{s_\cdot}(\nu^s y, \theta_0)} f_{s_\cdot}(\nu^s y, \theta_0) d(\nu^s y)$$
$$= \lim_{n\to\infty} \mathrm{E}\left[\mathcal{T}\mathcal{W}\right], \tag{16}$$

where $\mathcal{T} = \sqrt{n}\,(T(\nu^s y) - \theta_0)$ and $\mathcal{W} = \frac{\left(\nabla_\theta \tilde{f}_{s_\cdot}(\nu^s y, \theta_0)\right)'}{\sqrt{n} f_{s_\cdot}(\nu^s y, \theta_0)} = \frac{\left(\nabla_\theta \log\left\{\tilde{f}_{s_\cdot}(\nu^s y, \theta_0)\right\}\right)'}{\sqrt{n}}$.
Next, write the complete asymptotic variance-covariance matrix of $(\mathcal{T}, \mathcal{W})$.

$$\lim_{n\to\infty} \left(\mathrm{E}\left[\begin{pmatrix} \mathcal{T}\mathcal{T}' & \mathcal{T}\mathcal{W}' \\ \mathcal{W}\mathcal{T}' & \mathcal{W}\mathcal{W}' \end{pmatrix} \right] - \begin{pmatrix} \mathrm{E}\left[\mathcal{T}\right] \mathrm{E}\left[\mathcal{T}\right]' & \mathrm{E}\left[\mathcal{T}\right] \mathrm{E}\left[\mathcal{W}\right]' \\ \mathrm{E}\left[\mathcal{W}\right] \mathrm{E}\left[\mathcal{T}\right]' & \mathrm{E}\left[\mathcal{W}\right] \mathrm{E}\left[\mathcal{W}\right]' \end{pmatrix} \right).$$

Recall that $\sqrt{n}T$ is an asymptotically unbiased estimator such that $\lim_{n\to\infty} \mathrm{E}\left[\mathcal{T}\right]$ is a zero vector of length q. By writing out $\mathrm{E}\left[\mathcal{W}\right]$ and exchanging the order of integration and differentiation, it can be shown that $\mathrm{E}\left[\mathcal{W}\right]$ is also a zero vector of length q. Thus, the subtracted matrix cancels and the asymptotical variance-covariance matrix is:

$$\lim_{n\to\infty} \mathrm{COV}\left[\mathcal{T}\mathcal{W}\right] = \lim_{n\to\infty} \mathrm{E}\left[\begin{pmatrix} \mathcal{T}\mathcal{T}' & \mathcal{T}\mathcal{W}' \\ \mathcal{W}\mathcal{T}' & \mathcal{W}\mathcal{W}' \end{pmatrix}\right].$$

Next, note that $\mathrm{E}\left[\mathcal{T}\mathcal{T}'\right] = \mathrm{COV}\left[\mathcal{T}\right] = \mathrm{COV}\left[\sqrt{n}(T-\theta_0)\right] = \mathrm{COV}\left[\sqrt{n}T\right]$, while

$$\mathrm{E}\left[\mathcal{W}\mathcal{W}'\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\nabla_\theta \log\left\{\tilde{f}_{s_i}(\nu^s y, \theta_0)\right\}' \nabla_\theta \log\left\{\tilde{f}_{s_i}(\nu^s y, \theta_0)\right\}\right] = J.$$

and $\mathrm{E}\left[\mathcal{T}\mathcal{W}'\right] = I$. So overall, we get:

$$\lim_{n\to\infty} \mathrm{COV}\left[\mathcal{T}\mathcal{W}\right] = \lim_{n\to\infty} \begin{pmatrix} \mathrm{COV}\left[\sqrt{n}T\right] & I \\ I & J \end{pmatrix}. \tag{17}$$

Being a variance-covariance matrix, this expression must be positive semi-definite, so in particular

$$\forall a \quad (a', -a'J^{-1}) \lim_{n\to\infty} \begin{pmatrix} \mathrm{COV}\left[\sqrt{n}T\right] & I \\ I & J \end{pmatrix} \begin{pmatrix} a \\ -J^{-1}a \end{pmatrix} \geq 0.$$

If we multiply out this inequality, we get the result:

$$\lim_{n\to\infty} \forall a \quad a'\left(\mathrm{COV}\left[\sqrt{n}T\right] - J^{-1}\right) a \geq 0.$$

□

Proposition 7 gives us the lower bound on the variance-covariance matrix. Because this bound is asymptotically attained by the root-n estimator, we can conclude immediately:

Corollary 2 (Asymptotic efficiency). *In the censoring problem described above and given that* (ID)*,* (EX)*,* (FIN)*,* (EID)*, and* (DIFF) *hold, the root-n estimator* $\sqrt{n}\hat{\theta}_n$ *is asymptotically efficient.*

The GL-estimator is thus superior to the SCL-estimator in the sense that its root-n estimator has a lower asymptotical variance-covariance matrix. In other words, the GL-estimator makes better use of the available information.

5 A remark on censored regression

In many applications, observational units will differ by observable characteristics X_i which have an effect on the distribution of Y. To allow for this in our modelling framework, we suppose that the formerly fixed θ is an individual parameter which results from the interplay of observable characteristics X_i with a fixed parameter β: $\theta_i = g(\beta, X_i)$.

Since observational units are drawn randomly, the observable characteristics X_i can be modelled by a random variable. We assume this random vector to be continuously distributed,[1] so that the joint density of Y and X can be decomposed: $f_{Y,X}\left(y, g(\beta, x)\right) = f_{Y|X}\left(y|g(\beta, x)\right) \cdot f_X(x)$. Accordingly, the contribution of a particular state s becomes:

$$\tilde{f}_s(\nu^s y, \beta, x) = \underbrace{\int_{\mathcal{V}^s} f(y|\beta, x) d(\bar{v}_s y)}_{=:\tilde{f}(\nu^s y|\beta, x)} \cdot f_X(x),$$

and thus the logarithmitised objective function is:

$$Q_n(\theta) = \sum_{i=1}^{n} \log\left(\tilde{f}_s(v^{s_i} y_{i\cdot}|\beta, x_i)\right) + \sum_{i=1}^{n} \log\left(f_X(x_i)\right). \tag{18}$$

As the last term does not change in β, it can be ignored when maximising. So we are left with an objective function which closely resembles the objective function from formula (2) which we analysed in the preceding sections. All conditions, proofs, and theorems can be adapted to this new objective function by replacing $f(y, \theta)$ by $f(y|\beta, x)$ and $\tilde{f}_s(v^s y, \theta)$ by $\tilde{f}_s(v^s y|\beta, x)$, and requiring the respective statement to hold for all x. Additionally, one needs $\int \log(f_X(x)) f_X(x) dx < \infty$, to ensure finiteness of the limiting objective function. The identifiability condition becomes:

$$\forall \beta \neq \beta' \, \exists s, \mathcal{X} : \, P_X(x \in \mathcal{X}) > 0 : \, \tilde{f}_s(\nu^s y, \beta, x) \neq \tilde{f}_s(\nu^s y, \beta', x). \tag{ID'}$$

For the linear case $g(X, \beta) = X\beta$ and given (ID), this is fulfilled if X has full rank with positive probability.

On the asymptotic normality result the introduction of X has no effect: all proofs are based on derivatives of the objective function with respect to the parameter. Since the second sum in the objective function is independent of the parameter β, it cancels when taking derivatives.

[1] Extension to the discrete case is straightforward.

6 Conclusion

Parameters in censoring models are often estimated by maximising an objective function which resembles a likelihood. In fact, it is often misleadingly called a likelihood function. It is then taken for granted that the maximiser of this function has the usual properties of a maximum likelihood estimator. Obviously, labelling the objective function "likelihood" is not sufficient for the maximiser to have the properties, instead the function must fulfil certain criteria.

In this article, it has been shown how to construct an objective function, the generalised likelihood, such that the maximiser has the desired properties of a maximum likelihood estimator under regularity conditions which are very similar to the usual regularity conditions. The generalised likelihood estimator can be applied to a wide range of censoring problems; in fact, the class of censoring problems includes most problems considered in the literature.

The estimator reduces to the ordinary maximum likelihood estimator when there is no censoring. Similar to the maximum likelihood estimator, it is vulnerable to a mis-specification of the density function and quantile-regression is a sensible alternative when the density is unknown (for an overview see Fitzenberger 1997). While quantile regression is less demanding with respect to the density, it is no alternative for some censoring problems: simply, because it cannot be applied when the respective quantile is not observed.

The estimator proposed here coincides with many classical estimators for censoring problems such as the tobit type I estimator. Checking whether the objective function of an estimator is a monotone transformation of the generalised likelihood function is a simple method to prove that the estimator has the desired properties. Hence, one can justify the use of many objective functions which –up to now– were based on rules of thumb and intuition. In this sense, the article can be viewed as an extension of the well-known result of Amemiya (1973) to more general distributions than the normal distribution, to more involved censoring conditions and to multivariate random variables.

Beyond proving the properties of existing estimators, maximising the generalised likelihood offers the same convenience and properties, which are known from ordinary maximum likelihood estimation, for various censoring settings. It provides a clear rule how to derive estimators in such settings, so that there is no need to rely on intuition or folk theorems.

References

Amemiya, T. (1973): "Regression Analysis when the Dependent Variable is Truncated Normal," *Econometrica*, 41(6), 997–1016.

——— (1984): "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3–61.

——— (1985): *Advanced Econometrics*. Basil Blackwell, Oxford.

Attanasio, O. P. (2000): "Consumer Durables and Inertial Behaviour: Estimation and Aggregation of (S,S) Rules for Automobile Purchases," *Review of Economic Studies*, 67, 667–696.

Davidson, R., and J. G. MacKinnon (1993): *Estimation and Inference in Econometrics*. Oxford University Press, Oxford.

Fitzenberger, B. (1997): "A Guide to Censored Quantile Regression," in *Handbook of Statistics, Vol 15: Robust Inference*, ed. by G. Maddala, and C. Rao, pp. 405–435. Elsevier, Amsterdam, Reprint.

Nelson, F. D. (1977): "Censored Regression Models with Unobserved Stochastic Censoring Thresholds," *Journal of Econometrics*, 6(3), 309–328.

Newey, W. K., and D. McFadden (1996): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. 4, pp. 2111–2241. North-Holland.

Tobin, J. (1958): "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36.

Appendix: Collection of states decomposes $\mathbb{R}$

Lemma 2. *$\{\mathcal{V}^s\}_{s\geq 0}$ is a disjoint decomposition of $\mathbb{R}^q$.*

Proof. Part 1: $\bigcup_s \mathcal{V}^s = \mathbb{R}^q$
Take any $y \in \mathbb{R}^q$. Then, the respective visibility is $(v_1, \ldots, v_q)$. Call the state corresponding to this visibility s. For this visibility realisation, it must hold that $y \in \mathcal{V}_j$ if y_j observable ($v_j^s = 1$) and $y \in \bar{\mathcal{V}}_j$ if y_j not observable ($v_j^s = 0$). Thus, $y \in \bigcap_{\{j|v_j^s=1\}} \mathcal{V}_j \cap \bigcap_{\{j|v_j^s=0\}} \bar{\mathcal{V}}_j$, which by definition is equivalent to $y \in \mathcal{V}^s$. So, $y \in \bigcup_s \mathcal{V}^s$ and $\bigcup_s \mathcal{V}^s \supseteq \mathbb{R}^q$. Because each $\mathcal{V}^s \subseteq \mathbb{R}^q$, it follows

that $\bigcup_s \mathcal{V}^s \subseteq I\!R^q$. Overall, $\bigcup_s \mathcal{V}^s = I\!R^q$.

Part 2: $\mathcal{V}^s \cap \mathcal{V}^{s'} = \emptyset$ for $s \neq s'$.
If states differ ($s \neq s'$), it follows that there exists a component k which is visible in one state but not in the other ($v_k^s \neq v_k^{s'}$). Without loss of generality, let s be the state where it is visible, then $\mathcal{V}^s \subseteq \mathcal{V}_k$ and $\mathcal{V}^{s'} \subseteq \bar{\mathcal{V}}_k$ by the definition of the visibility set. Hence, $(\mathcal{V}^s \cap \mathcal{V}^{s'}) \subseteq (\mathcal{V}_k \cap \bar{\mathcal{V}}_k)$. But the latter is by construction the empty set: $\mathcal{V}_k \cap \bar{\mathcal{V}}_k = \emptyset$. Thus, the visibility sets must be disjoint: $(\mathcal{V}^s \cap \mathcal{V}^{s'}) = \emptyset$. □