ECONOMETRIC INSTITUTE

ESTIMATION OF CONTINUOUS MODELS ON THE BASIS OF SET-VALUED OBSERVATIONS

B.M.S. VAN PRAAG AND J.P. HOP

REPORT 8705/A

ERASMUS UNIVERSITY ROTTERDAM - P.O. BOX 1738 - 3000 DR ROTTERDAM - THE NETHERLANDS

ESTIMATION OF CONTINUOUS MODELS ON THE BASIS OF SET-VALUED OBSERVATIONS

ABSTRACT

by B.M.S. van Praag and J.P. Hop.

In many empirical applications the phenomenon (Y,X) in which we are interested and for which we assume a model $Y=f(X;\theta)+\varepsilon$, cannot be observed exactly. Then we can only say that $(Y,X) \in A \subset R^{m+n}$ where A is a point set. Probit, Tobit or discrete choice models are examples.

In those cases we say that the observations are set-valued and that the latent phenomenon is observed through a filter. For the model $Y=BX+\varepsilon$ we present a general ML-method for estimating B and $E(\varepsilon\varepsilon')$ which bypasses the well-known problem of the computation of multi-dimensional integrals.

Examples are given for the case where we only observe $\max(Y_1,Y_2,Y_3)$ or $\max(Y_1,\ldots,Y_6)$.

February 15, 1987.

ESTIMATION OF CONTINUOUS MODELS ON THE BASIS OF SET-VALUED OBSERVATIONS.

1. INTRODUCTION.

One of the fascinating developments in econometrics, but also in other empirical sciences, is the emergence of Probit-type estimation problems. Basically, such problems arise if we have a latent model, say, $Y' = B'\overset{''}{X} + \varepsilon'$ with $(Y,X) \in R^{m+n}$ a random vector, $\varepsilon \sim N(0,\Sigma)$, $\overset{''}{X}$ a matrix, the columns of which are subvectors of X, where Y or X or both cannot always be exactly observed, but where we have to be satisfied with non-exact observations of the type $(Y,X) \in A \subset R^{m+n}$, where A is a point set in R^{m+n}. We call the observations set-valued. In that case estimation of B and Σ becomes a problem.

The problem is not so much a philosophical problem as it is fairly easy to construct the sample's likelihood that has to be maximized with respect to B and Σ. The problem is a computational one as, except for very small dimensions, the likelihood requires the computation of multi-dimensional integrals, that is still too much of an effort even for modern computers. Maddala (1983) and Amemiya (1981,1985) present surveys of the rapidly increasing literature. Most models are ad hoc approaches to a specific problem.

Gouriéroux et al. (1984) take a more general approach in introducing generalized or simulated residuals, but they use this concept for test construction only, and not for estimation purposes.

In this paper we shall try to develop a general approach to the estimation of models on the basis of set-valued observations and to implement that approach by a feasible computation method that does not break down on the integration problem. In Section 2 we consider the common structure of some problems. In Section 3 we derive two important theorems on consistency and efficiency of the ML-method in case of set-valued observations. In Section 4 we outline our computation method, while in Section 5 we consider a few examples on a large simulated data set. Section 6 concludes. The theory is an asymptotic theory, i.e., for large data sets only. At several points in this paper we interchange limiting operations like integration and differentiation. When doing so, we silently assume that the requirements for that procedure are satisfied. (see e.g. Billingsley (1979), section 16). In order to economize on space we do not spell out those standard conditions.

2. SET-VALUED OBSERVATIONS.

Let us assume a probability space $(R^{m+n}, \mathcal{C}, P)$ where $\mathcal{C}$ is the σ-field of Borel sets and P a probability measure. Consider now a partition $\mathcal{A}$ on R^{m+n}, i.e., it consists of disjoint sets such that for any two sets A and $A' \in \mathcal{A}$ holds

$$A \cap A' = \Phi \text{ and } \bigcup_{A \in \mathcal{A}} A = R^{m+n}$$

We denote the σ-field generated by the sets in $\mathcal{A}$ by $\sigma(\mathcal{A})$ and we have $\sigma(\mathcal{A}) \subset \mathcal{C}$. Notice that the sets in $\mathcal{A}$ are <u>atoms</u> of $\sigma(\mathcal{A})$, i.e., proper subsets of $A \in \mathcal{A}$ are no longer $\sigma(\mathcal{A})$-measurable. The P-measure induced on $\sigma(\mathcal{A})$ is denoted by $P_{\mathcal{A}}$.
Let Z = (Y,X) be a random vector and let $E(Z|\mathcal{A})$ be the conditional expectation of Z on $\sigma(\mathcal{A})$, i.e.

(1) $$\int_C Z dP = \int_C E(Z|\mathcal{A}) dP \quad \text{whenever } C \in \sigma(\mathcal{A}).$$

(see Billingsley (1979), p.395), then we call $\tilde{Z} = E(Z|\mathcal{A})$ the $\mathcal{A}$-filtered observation of Z.
$E(Z|\mathcal{A})$ is random and P-measurable on the subfield $\sigma(\mathcal{A})$. Finally we shall consider random vectors $\phi(\tilde{Z})$; they are also P -measurable on $\sigma(\mathcal{A})$ but not on $\mathcal{C}$. They are transformed $\mathcal{A}$-filtered observations.
The problems we have in mind are examples that reveal their common structure when we use the concept of $\mathcal{A}$-filtering.
Let us illustrate this setting by some examples.

a. <u>The Probit model.</u>
The latent vector is $(Y,X) \in R^{1+n}$. The partition $\mathcal{A}$ consists of sets $A_1 = \{Y \leq 0, X=x\}$ and $A_2 = \{Y > 0, X=x\}$. Notice that the partition is an uncountable class of sets.

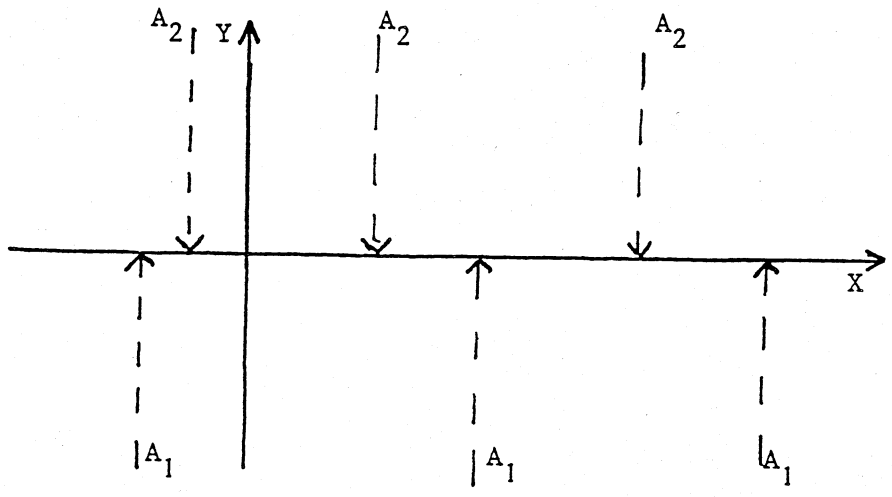

Figure 1. Probit-partition.

Our observations are $\mathcal{A}$-filtered. In figure 1 the partition is sketched.

b. The Tobit-model.

The latent vector is $\{Y,X\} \in R^{1+n}$. The partition $\mathcal{A}$ consists of sets $A_1=\{Y\leq 0, X=x\}$ and $A_2=\{Y=y, X=x\}$ if $Y>0$.

Tobit observations are $\mathcal{A}$ - filtered. $\mathcal{A}$ is sketched in Figure 2.

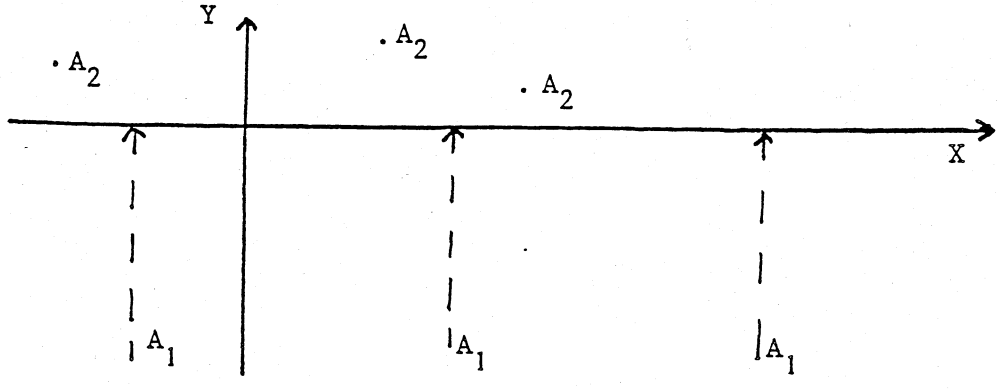

Figure 2. Tobit-partition.

c. The two-equation Heckman-model (1976).

The observations are of the type $(Y_1,Y_2,X)\epsilon R^{2+n}$.

If $Y_1\leq 0$, we do not know Y_2 and, if $Y_1>0$, we observe Y_2. We may think of Y_1 as the response-willingness and Y_2 as the response itself. The partition consists of the sets

$$A_1=\{Y_1\leq 0, X=x\} \text{ and } A_2=\{Y_1>0, Y_2=y_2, X=x\}.$$

After projection on the (Y_1,Y_2)- space we sketch the Heckman partition in figure 3.

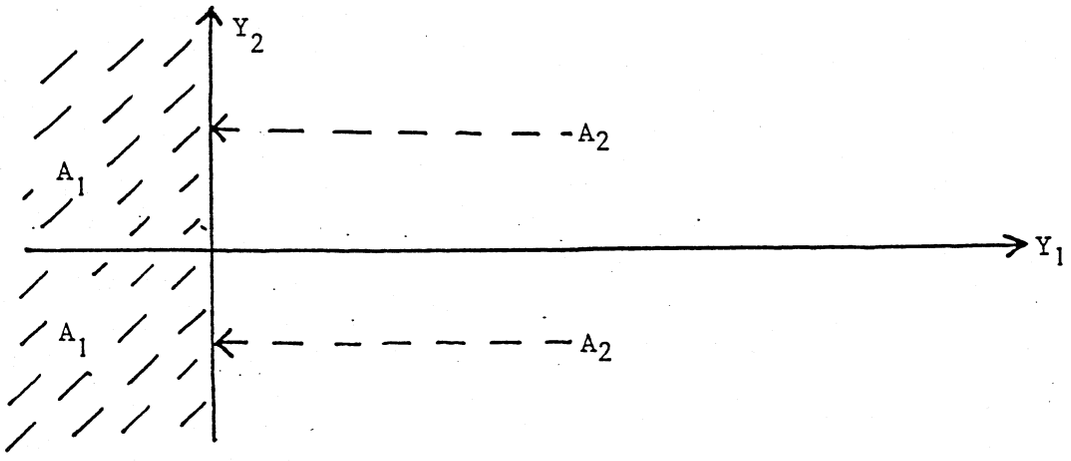

Figure 3. The Heckman-partition.

d. The discrete choice-model (Domencich and McFadden (1975))

The latent vector is $(Y_1,Y_2,X)\epsilon R^{2+n}$. Y_1 stands for the attractiveness of commodity 1 and Y_2 is the attractiveness of commodity 2.

If $Y_1>Y_2$ and $Y_1>0$, we observe the purchase Y_1 of the first commodity;

if $Y_2>Y_1$ and $Y_2>0$, we observe the purchase of the second commodity.

If Y_1 and $Y_2<0$, we observe that nothing is bought. The partition consisting of three types of sets is sketched in Figure 4.

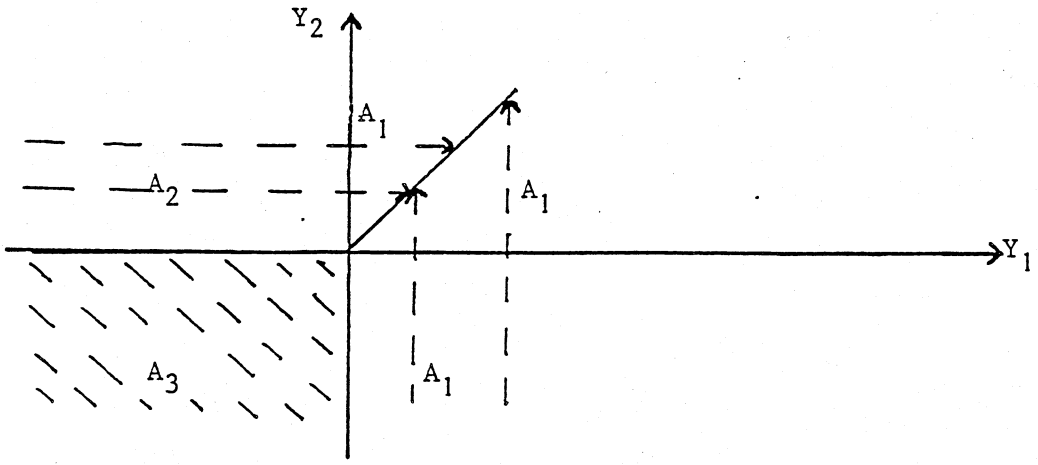

Figure 4. Discrete choice partition.

e. The disequilibrium model (Fair and Jaffee (1972)).

Let Y_1 stand for market demand and Y_2 for market supply. The market-turnover $S = \min(Y_1, Y_2)$ is observed or, if $S \leq 0$, there is no transaction.

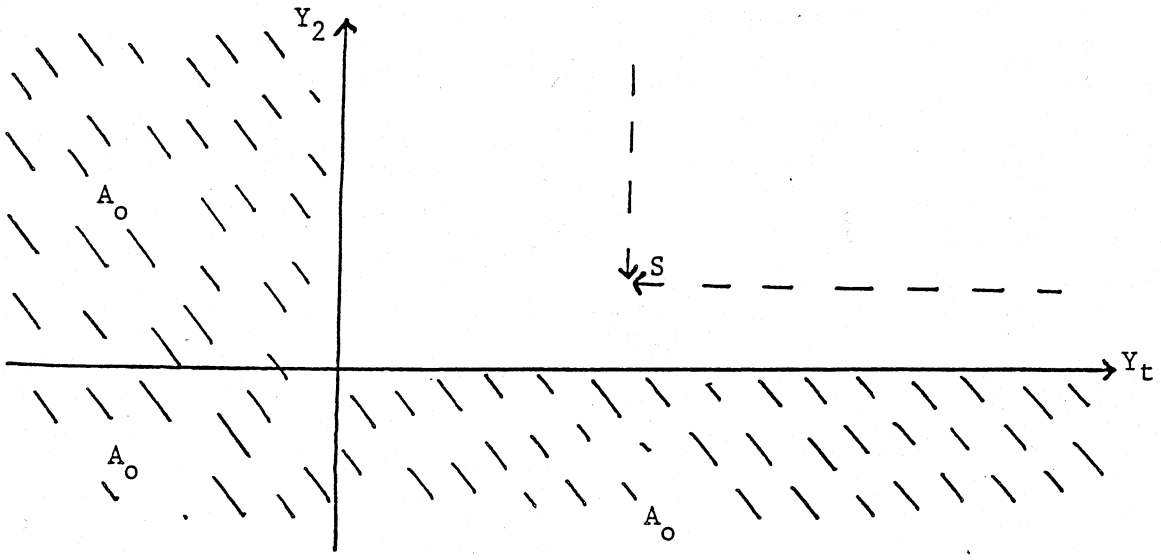

Figure 5. Disequilibrium-market partition.

The $\mathcal{A}$-partition consists of sets of the type
$\{(Y_1,S)|Y_1\geq S\}\cup\{(S,Y_2)|Y_2\geq S\}$ if $S>0$ and the set $\{Y_1\leq 0\}\cup\{Y_2\leq 0\}$.

f. Exact observations.

Obviously in the case of exact observation we have $\mathcal{A} = \mathcal{C}$.

In general the class $\mathcal{A}$ consists of the finest distinguishible events. They form the atoms of $\sigma(\mathcal{A})$.

If a measure P on $\mathcal{C}$ depends on a parameter vector θ, it is obvious that filtering of $P(\theta)$ through $\mathcal{A}$ yields a measure $P_{\mathcal{A}}(\theta)$ on $\sigma(\mathcal{A})$. Let us assume $P(\theta)$ is functionally specified to belong to a specific class, i.e. $N(\mu,\sigma^2)$ on R, then the crucial question for estimation purposes is whether the mapping $P(\theta) \to P_{\mathcal{A}}(\theta)$ is one-to-one or not. If there is only one θ that can generate $P(\theta)$, θ may be identified from $\mathcal{A}$-filtered observations; if the mapping is not one-to-one such an identification is impossible. So we define a filter $\mathcal{A}$ to be an identification-preserving filter (i.p.f.), iff there is a one-one relation between $P(\theta)$ and $P_{\mathcal{A}}(\theta)$.

We give two examples of filters that are not i.p.f.

a. Consider the case of $(R,\mathcal{C},P(\theta))$ with $P(\theta) \sim N(\mu,\sigma^2)$ that is filtered by the probit filter, i.e., $\mathcal{A} = \{A_1,A_2\}$ with $A_1 = (-\infty,0]$, $A_2 = (0,\infty)$. Let $P(A_1) = p(\mu,\sigma^2)$ and $P(A_2) = 1-p(\mu,\sigma^2)$.
If $p(\mu,\sigma^2)$ is known, it does not define two unkowns μ and σ^2. Hence, $\mathcal{A}$ is not an i.p.f. If σ^2 is known, it is i.p.f. with respect to μ.
If $\mathcal{A} = \{A_1,A_2,A_3\}$ it is posible to determine μ and σ^2 uniquely.

b. Assume that X is $N(\mu,1)$, and that we observe $|X|$. The filter is $\{x\in R|x=\pm a, a\geq 0\}$. Then μ and $(-\mu)$ will yield the same $P(\mu)$. Hence observing $|X|$ instead of X we can find only $|\mu|$. Hence we know that one of the "roots" μ and $(-\mu)$ is a "false root".

In what follows we consider only identification-preserving filters.

3. ML-ESTIMATION FROM SET-VALUED OBSERVATIONS.

Let us assume $Z \epsilon R^{m+n}$ and let us assume Z has a mass density-function $f(z;\theta_0)$. Moreover let us assume we have a random sample $\{Z_t\}_{t=1}^{T}$ of exact observations. We define the log-likelihood

$$(2) \qquad \hat{L}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \ln f(Z_t;\theta)$$

The ML-estimator $\hat{\theta}$ is the solution of the normal equation*)

$$(3) \qquad \frac{\partial}{\partial \theta} \hat{L}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \ln f(Z_t;\theta)}{\partial \theta} = 0.$$

It is well-known that $\hat{\theta}$ is a consistent estimator of the true θ_0, i.e.. $\lim_{T\to\infty} P(|\hat{\theta}-\theta_0|>\varepsilon) = 0$ for all $\varepsilon>0$ (see e.g. Amemiya (1985), p.116), if some regularity conditions are fulfilled.

Consider now the population analogue of the likelihood (2)

$$(4) \qquad L(\theta) = E[\ln f(Z;\theta)]$$

and the population analogue of the normal equations (3)

$$(5) \qquad \frac{\partial}{\partial \theta} L(\theta) = E[\frac{\partial}{\partial \theta} \ln f(Z;\theta)]$$

Then θ_0 is the solution of $\frac{\partial}{\partial \theta} L(\theta) = 0$.

---

*) We do not explicitly use the fact that θ is a p-vector, but (3) is actually a set of p simultaneous equations.

Let us now consider how this changes if we have set-valued observations. Let us consider first a countable patition $\mathcal{a}$, with

$$(6) \qquad P(A;\theta) = \int_A f(z;\theta)dz = E(I_A;\theta)$$

where the indicator function $I_A(Z)$ is one if $Z\epsilon A$ and zero otherwise.

The sample is given by the sequence of mutually independent random events

$\{A_t\}_{t=1}^T$ with likelihood $\prod_{t=1}^T P(A_t;\theta)$.

The log-likelihood is

$$(7) \qquad \hat{L}_{\mathcal{a}}(\theta) = \frac{1}{T}\sum_{t=1}^T \ln P(A_t;\theta)$$

with expectation $L_{\mathcal{a}}(\theta)$.

The normal equations read

$$\frac{\partial}{\partial\theta}\hat{L}_{\mathcal{a}}(\theta) = \frac{1}{T}\sum_{t=1}^T \frac{1}{P(A_t;\theta)}\int_{A_t}\frac{\partial}{\partial\theta}f(z;\theta)dz$$

$$(8) \qquad = \frac{1}{T}\sum_{t=1}^T \frac{1}{P(A_t;\theta)}\int_{A_t}\left(\frac{\partial}{\partial\theta}\ln f(z;\theta)\right)f(z;\theta)dz$$

$$= \frac{1}{T}\sum_{t=1}^T E\left(\frac{\partial}{\partial\theta}\ln f(Z;\theta)\,\middle|\,Z\epsilon A_t\right)$$

where $E(.|Z\epsilon A_t)$ stands for the conditional expectation, given $Z\epsilon A_t$. Writing for that conditional expectation W it is obvious that the sample

$\{A_t\}_{t=1}^T$ corresponds to a set of realizations $\{W_t\}_{t=1}^T$.

Actually (8) may be rewritten

$$(9) \qquad \frac{\partial}{\partial\theta}\hat{L}_{\mathcal{a}}(\theta) = \frac{1}{T}\sum_{t=1}^T W_t(\theta)$$

$W(\theta)$ is a random P-vector that is $P_{\mathcal{a}}$-measurable. W is constant on the atoms of $\sigma(\mathcal{a})$, i.e., constant on $A \in \mathcal{a}$.
For the population analogues it implies that

$$(10) \qquad \frac{\partial}{\partial\theta} L_{\mathcal{a}}(\theta) = E\left[\frac{\partial}{\partial\theta} \hat{L}_{\mathcal{a}}(\theta)\right] = E_{\mathcal{a}}(W(\theta))$$

where $E_{\mathcal{a}}$ stands for the expectation with respect to $P_{\mathcal{a}}$.
As $\sigma(\mathcal{a}) \subset \mathcal{E}$ it follows from Th. 34.4 in Billingsley (p.398) that

$$(11) \qquad E_{\mathcal{a}}(W) = E_{\mathcal{a}}\left[E\left[\frac{\partial}{\partial\theta} \ln f(Z;\theta) \,\middle|\, Z \in A\right]\right] = E\left[\frac{\partial}{\partial\theta} \ln f(Z;\theta)\right]$$

Combining (5), (10) and (11) we have

$$(12) \qquad \frac{\partial}{\partial\theta} L_{\mathcal{a}}(\theta) \equiv \frac{\partial}{\partial\theta} L(\theta) \quad \text{for all } \theta.$$

It follows that $\lim_{T\to\infty} P\left[\left|\frac{\partial}{\partial\theta}\hat{L}_{\mathcal{a}}(\theta) - \frac{\partial}{\partial\theta}\hat{L}(\theta)\right| > \varepsilon\right] = 0$ for all $\varepsilon > 0$.

It follows also that, if $\theta_{\mathcal{a},0}$ and θ_0 are the unique solutions of the normal equation systems

$$\frac{\partial}{\partial\theta} L_{\mathcal{a}}(\theta) = 0 \quad \text{and} \quad \frac{\partial}{\partial\theta} L(\theta) = 0,$$

then

$$(13) \qquad \theta_{\mathcal{a},0} = \theta_0.$$

If $\hat{L}(\theta)$ and $\hat{L}_{\mathcal{a}}(\theta)$ are continuous in θ and the normal equations have a unique solution, then $\lim_{T\to\infty} P\left[\left|\hat{\theta}_{\mathcal{a},0} - \hat{\theta}_0\right| > \varepsilon\right] = 0$ for all $\varepsilon > 0$.

Let us now consider the case of a non-denumerable $\mathcal{a}$. The subsets of $\mathcal{a}$ may be labeled by means of a q-dimensional vector $\tilde{Z}(A)$ which is constant on A, such that there is a one-one-relationship between $A \in \mathcal{a}$ and $\tilde{Z}(A)$.

As $\tilde{Z}$ is $P_{\mathcal{A}}$-measurable, a P-measure on R^q is induced such that

$$P[\tilde{Z}\epsilon B]=\int_B \tilde{f}(\tilde{z})d\tilde{z}$$

where $\tilde{f}(\tilde{z})$ is the density-function of $\tilde{Z}$.

If A is the pre-image of B under $\tilde{Z}$, we have $P(\tilde{Z}\epsilon B)=P_{\mathcal{A}}(A)=P(A)=\int_A f(z)dz=E(I_A)$ for all $A\epsilon\sigma(\mathcal{A})$.

In this case we may represent the sample by a sequence $\{A_t\}_{t=1}^T$ or a sequence $\{\tilde{Z}_t\}_{t=1}^T$ of independent realizations of the random vector $\tilde{Z}$.

Its log-likelihood is $\hat{L}_{\mathcal{A}}(\theta) = \frac{1}{T}\sum_{t=1}^{T} \ln \tilde{f}(\tilde{Z}_t;\theta)$

Using the same reasoning as in the discrete case we get

$$(15) \qquad \frac{\partial}{\partial\theta}\hat{L}_{\mathcal{A}}(\theta) = \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{\partial}{\partial\theta}\ln f(Z;\theta)\,\middle|\,\tilde{Z}=\tilde{Z}_t\right)$$

The expectation of the $\mathcal{A}$-filtered version of the normal equations equals the expected normal equations under exact observation

$$(16) \qquad E\left[\frac{\partial}{\partial\theta}\hat{L}_{\mathcal{A}}(\theta)\right]=\frac{\partial}{\partial\theta}L_{\mathcal{A}}(\theta) \equiv \frac{\partial}{\partial\theta}L(\theta)$$

The ML-estimators have the same asymptotic expectation,

$$(17) \qquad \theta_{\mathcal{A},0}=\theta_0$$

and they are stochastically converging, i.e.,

$$(18) \qquad \lim_{T\to\infty} P[|\hat{\theta}_{\mathcal{A},0}-\hat{\theta}_0|>\varepsilon]=0 \text{ for all } \varepsilon>0$$

Now let us compare the asymptotic variances, i.e. the variance of $\sqrt{T}(\hat{\theta}_0-\theta_0)$ or $\sqrt{T}(\hat{\theta}_{\mathcal{A},0}-\theta_0)$. According to standard ML-theory (see e.g. Rao (1973)) we have for a solution $\hat{\theta}_{\mathcal{A},0}$ of $\frac{\partial}{\partial\theta}\hat{L}_{\mathcal{A}}(\theta)=0$

(19) $\quad \text{avar}(\hat{\theta}_{a,0}) = -(E(\frac{\partial}{\partial\theta\partial\theta'}\hat{L}_a(\theta_0))^{-1}$

where we notice that the Hessian is symmetric and avar(.) stands for the asymptotic covariance matrix.
There holds also under ML assumptions, that

(20) $\quad -E(\frac{\partial^2}{\partial\theta\partial\theta'}\hat{L}_a(\theta_0)) = T.E[\frac{\partial}{\partial\theta}\hat{L}_a(\theta_0)\frac{\partial}{\partial\theta'}\hat{L}_a(\theta_0)].$

Then, it is easy to show that

(21) $\quad \text{var}(\frac{\partial}{\partial\theta}\hat{L}_a(\theta)) \leq \text{var}(\frac{\partial}{\partial\theta}\hat{L}(\theta))$

in the order of positive (semi-)definite matrices.

Consider a specific W_a and the corresponding $W = \frac{\partial}{\partial\theta}\ln f(Z;\theta)$.

Notice that W_a is $\sigma(a)$-measurable while W is $\mathcal{C}$-measurable.
The well-known variance decomposition gives

(22) $\quad \text{var}(W) = \text{var}(W_a) + E\,(\text{var}(W(Z)|Z\epsilon A))$

where $\text{var}(W(Z)|Z\epsilon A)$ is the conditional variance-covariance matrix of W(Z) given $Z\epsilon A$. The first term is the "between" and the second the "within"-contribution.
It follows that $\text{var}(W_a) \leq \text{var}(W)$ and consequently

(23) $\quad \text{var}(\frac{\partial}{\partial\theta}\hat{L}_a(\theta)) = \frac{1}{T}\text{var}(W_a) \leq \frac{1}{T}\text{var}(W) = \text{var}(\frac{\partial}{\partial\theta}\hat{L}(\theta))$

Combination of (23) with (19) and (20) yields

(24) $\quad \text{avar}\,(\hat{\theta}_{a,0}) \geq \text{avar}\,(\hat{\theta}_0)$

The result implies that the coarser the subfield $\sigma(a)$, generated by a is, the more variance about the true value θ_0 the ML-estimator $\hat{\theta}_{a,0}$ will exhibit.
Finally, let us remark that by using the standard Central Limit arguments it

may be shown that $\hat{\theta}_{a,0}$ will be asymptotically normal.
In this section we actually came to a unified theory of ML-estimation for exact and a-filtered observations. In the next section we shall see that this approach will be particularly helpful for practical estimation purposes.

4. OPERATIONALIZATION.

Let us return to expression (15). We are looking for roots of the equation

$$\frac{\partial}{\partial\theta}\hat{L}_{\mathcal{Q}}(\theta)=0.$$

With respect to the existence or uniqueness of such roots, there is no problem. As the ML-roots are asymptotically equal to those under exact observation, it follows that we may consider the equivalent "exact observation"-problem. If the latter yields a unique ML-estimator, so does the $\mathcal{Q}$-filtered version. Although there is philosophically no problem, there is a practical problem. How can we calculate the conditional expectations

$$E\left(\frac{\partial}{\partial\theta}\ln f(Z;\theta)\,\middle|\,\tilde{Z}=\tilde{Z}_t\right) \text{ and } E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln f(Z;\theta)\,\middle|\,\tilde{Z}=\tilde{Z}_t\right)$$

in (15) or (19)?

Here we meet the formidable stumbling-block. Those conditional expectations are integrals over sometimes weird sets $A\in\mathcal{Q}$ with $A\subset R^{m+n}$.
Let us consider a different way, where integrals are assessed by Monte Carlo simulation and more precisely by "importance sampling" (see Van Dijk (1984)).
Let us look more closely at (8).
Let us assume we know the continuous density of Z on R^{m+n} and we try to evaluate the integral

$$(25)\qquad E(g(Z)\,|\,Z\in A)=\frac{1}{P(A)}\int_A g(z)f(z)dz$$

where we assume for convenience $P(A)>0$.
Now both the direct evaluation of the integral (25) and of

$$(26)\qquad P(A)=\int_A f(z)dz$$

is a problem: direct Monte-Carlo drawing from an arbitrary density f(z) is mostly impossible.
Assume for simplicity of exposition that A is a bounded interval $[a,b]\subset R$.

Then we may draw k times from an homogeneous distribution H(a,b) on [a,b] yielding k values $Z_1,\ldots Z_k$ and P(A) may be evaluated by

$$(27) \qquad P(A) \approx \frac{1}{k} \sum_{i=1}^{k} f(Z_i)$$

Similarly we have

$$(28) \qquad \int_A g(z)f(z)dz \approx \frac{1}{k} \sum_{i=1}^{k} g(Z_i)f(Z_i)$$

and it follows in this case that

$$(29) \qquad E(g(Z)|Z\epsilon A) \approx \frac{\sum_{i=1}^{k} g(Z_i)f(Z_i)}{\sum_{i=1}^{k} f(Z_i)}$$

where $\{Z_i\}_{i=1}^{k}$ are k independent drawings from H(a,b).
In a similar way if A is a 2-dimensional bounded block in R^2, we may draw the 2-vector Z from a two-dimensional uniform distribution on A and do the same procedure. If g(Z) is a vector function, say $g=(g_1,\ldots,g_p)$, then we have

$$(30) \qquad \begin{aligned} E(g_1(Z)|Z\epsilon A) &\approx \sum_{i=1}^{k} g_1(Z_i)f(Z_i) \Big/ \sum_{i=1}^{k} f(Z_i) \\ &\vdots \\ E(g_p(Z)|Z\epsilon A) &\approx \sum_{i=1}^{k} g_p(Z_i)f(Z_i) \Big/ \sum_{i=1}^{k} f(Z_i) \end{aligned}$$

We notice that the assumption P(A)>0 is irrelevant. For example, if f is a continuous density in R^2 and A the line segment $\{Z_1=0, 0\leq Z_2\leq 1\}$, clearly A is a null-set but nevertheless the conditional expectation of g(Z) on A may be defined and it may still be evaluated by (29).
Actually A is not necessarily one bounded interval or a union of disjoint intervals. Let for instance $A=\{(Z_1,Z_2)|Z_1^2+Z_2^2=1\}$, i.e., the boundary of a circle with radius 1, then we have a one-one correspondence between the points on the boundary and the interval $[0,2\pi]$. It follows that we can draw ξ from a

uniform distribution on $[0,2\pi]$ and set $Z_1=\cos(\xi)$, $Z_2=\sin(\xi)$.

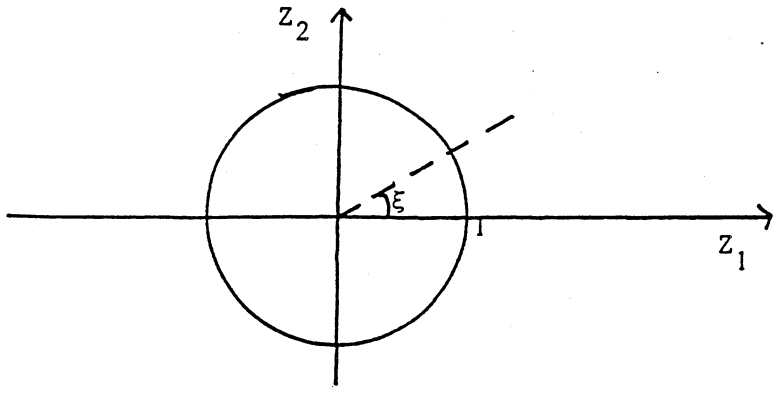

Figure 6. Drawing from the boundary of a circle.

Similarly if $A=\{(Z_1,Z_2)\,|\,Z_1^2+Z_2^2\leq 1\}$, we can draw from a uniform distribution on a rectangle with South-West corner (0,0) and North-East corner $(1,2\pi)$. Obviously there are sets A for which we cannot specify a uniform distribution. For instance $A = (0,\infty)$. Then we may specify an arbitrary drawing distribution h(z) on A.
We have

$$(31)\qquad \int_A g(z)f(z)dz=\int_A g(z)\,\frac{f(z)}{h(z)}\,h(z)dz$$

If we draw from a density h(z) on A we have

$$(32)\qquad E(g(Z)\,|\,Z\epsilon A)\approx \sum_{i=1}^{k} g(Z_i)f(Z_i)h^{-1}(Z_i)\Big/\sum_{i=1}^{k} f(Z_i)h^{-1}(Z_i)$$

In practice we take for h(z) if A is of the type $(0,\infty)$ an exponential density and if A is of the type$(-\infty,\infty)$ a normal density or products of those, if A is topologically more-dimensional.

In a similar way me may evaluate

$$(33) \qquad \mathrm{var}(g(Z)\,|\,Z\epsilon A) = E(g(Z)-E(g(Z)\,|\,Z\epsilon A))^2\,|\,Z\epsilon A) =$$

$$= E(g^2(Z)\,|\,Z\epsilon A)-[E(g(Z)\,|\,Z\epsilon A)]^2$$

Although any drawing density h(z) will do, nevertheless some are better than others. Actually we should look for a tractable drawing density h(z) such that $f(z)/h(z)\simeq 1$, i.e., that h(z) looks similar to f(z). (see Van Dijk (1984)).
In view of (27) and (28), expression (32) may be considered as the ratio of two sample moments. It is well-known that under fairly general conditions such functions stochastically converge to a normally distributed variable and that their sample variance, if it is finite, is of the order 1/k. (see e.g. Cramèr (1951)).
Hence (15) may be replaced by the approximation

$$(34) \qquad \frac{\partial}{\partial\theta}\hat{\hat{L}}\,(\theta) = \frac{1}{T}\sum_{t=1}^{T} W_t(\theta)$$

where each W_t has a variance $\frac{1}{k}\sigma_t^2$. Notice that σ_t^2 is a random variable as it depends on A_t. However, if we assume (almost surely) $\sigma_t^2<M<\infty$ for all $A\epsilon\mathcal{A}$, it is obvious that (34) converges stochastically to (8) or (15), and hence its ML-solution $\hat{\hat{\theta}}_{\mathcal{A},0}$ to θ_0. In this context we do not attempt to give formal proofs, as they will employ the technical standard arguments and would take a lot of space. Rather we like to consider some examples to see how the method works.

5. SOME EXAMPLES.

In this section we shall consider three examples of the general method outlined above. Our playground will be three data sets of 2,000 observations each that we have created by simulation.
First we specify the general setting:
Let us assume $Z=(Y,X)\in R^{m+n}$, where X are the exogenous and Y the endogenous variables. We assume a law

$$Y' = B'\overset{''}{X} + \varepsilon'$$

Where Y is an m-vector, B an $\overset{''}{n}$-vector of coefficients to be estimated, $\overset{''}{X}$ a $(\overset{''}{n}\times m)$- matrix the columns of which are subvectors of X or constants, and ε a random m-vector.
We shall assume ε to be $N(0,\Sigma)$-distributed, and we assume that X and ε are mutually independent. It follows that the density of an observation is

$$f(y,x)= n(y'-B'\overset{''}{x}).g(x)$$

where n(.) stands for the normal density.

Our objective will be to estimate B and Σ by $(\hat{B}_{\mathcal{Q},0},\hat{\Sigma}_{\mathcal{Q},0})$ when the observations are $\mathcal{Q}$-filtered. In order to test the method we have to know B and Σ. Then we may compare the estimates with the true values.
So we begin to specify B, Σ and the density g(x) and we create by Monte-Carlo simulation a data set of 2,000 observations.

We take X to be $N(\mu_X,\Sigma_X)$-distributed.

1. Our first example will be the Tobit-mode of observation. Clearly, Tobit-estimation is no problem, but it clarifies for a simple case how things work.

We assume $\mu_X=(0,0,0)'$ and a symmetric matrix

$$\Sigma_X=\begin{vmatrix} 1 & & \\ 0 & 0.50 & \\ 0.50 & 0.25 & 0.75 \end{vmatrix}$$

We specify m=1 and $Y_t = -2X_{1t}+3X_{2t}+X_{3t}+1$, while ε_t is N(0,1.5)-distributed. Under exact observation the problem is easy. The log-likelihood is

$$(35) \qquad \hat{L}(B,\sigma_\varepsilon^{-2}) = \frac{1}{T}\sum_{t=1}^{T}\left[\ln(n(Y'_t - B'X_t;\sigma_\varepsilon^2))+\ln(g(X_t))\right] \;^{*)}$$

$$= +\frac{1}{2}\ln(\sigma_\varepsilon^{-2}) - \frac{1}{2T\sigma_\varepsilon^2}\sum_{t=1}^{T}(Y'_t - B'X_t)^2 + \text{constant.}$$

where Y_t is a scalar and X_t a 4-vector, when we include the constant.

As the last part does not depend on B or σ_ε^2 we have

$$(36) \qquad \frac{\partial}{\partial B}\hat{L}(B,\sigma_\varepsilon^{-2}) = \frac{1}{T\sigma_\varepsilon^2}\sum_{t=1}^{T}X_t(Y'_t - B'X_t)'$$

$$(37) \qquad \frac{\partial}{\partial(\sigma_\varepsilon^{-2})}\hat{L}(B,\sigma_\varepsilon^{-2}) = \frac{1}{2T}\sum_{t=1}^{T}\left[\sigma_\varepsilon^2 - (Y'_t - B'X_t)^2\right]$$

Let us define $Y'_t - B'X_t = u'_t$ where u'_t is just the t^{th} residual. So we may rewrite (36) and (37) as

$$(36a) \qquad \frac{\partial}{\partial B}\hat{L} = \frac{1}{T\sigma_\varepsilon^2}\sum_{t=1}^{T}X_t u_t$$

$$(37a) \qquad \frac{\partial}{\partial(\sigma_\varepsilon^{-2})}\hat{L} = \frac{1}{2T}\sum_{t=1}^{T}(\sigma_\varepsilon^2 - u_t u'_t)$$

*) Although Y_t is a scalar we sometimes write Y'_t in the formulae in view of the generalization to m-vectors Y hereafter.

In the exact case those equations are easily solved for $\hat{B}_0$ and $\hat{\sigma}^2_{\varepsilon,0}$.

By OLS-regression we found the estimation result

$$Y = \underset{(.03)}{-2.01}\ X_1 + \underset{(.04)}{2.93}\ X_2 + \underset{(.04)}{1.05}\ X_3 + \underset{(.03)}{1.00} \qquad \sigma^2_\varepsilon = 1.53 \quad R^2 = .85$$

In the case of Tobit-observation an $\mathcal{A}$-filter is applied. (See figure 2).

If $Y_t > 0$ we may exactly observe (Y_t, X_t).
If $Y_t \leq 0$ we know only that the observation is on the vertical half-line A_1.
In the case of non-exact observation we replace the separate terms for those observations in (36a) by

$$(38) \qquad \frac{1}{T\sigma_\varepsilon^2}\ X_t E(u_t \mid Y_t \leq 0)$$

and in (37a) by

$$(39) \qquad \frac{1}{2T}\sigma_\varepsilon^2 - \frac{1}{2T} E[u_t u_t' \mid Y_t \leq 0]$$

Consider for instance

$$(40) \qquad E(Y_t \mid Y_t \leq 0) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0} y \exp\left(-\frac{1}{2\sigma^2}(y' - B'X_t)^2\right) dy$$

This is a rather complicated function in B. It follows that if in (36) we have to replace some "exact observation" terms by conditional expectations we cannot find an explicit solution like in the OLS-case, but (36) has to be solved by iteration until a satisfactory solution has been found.

The iteration process is now structured as follows:

1. Start with an initial guess for B and σ_ε^2 .
2. Calculate the conditional expectations in (38) and (39) in a way to be described below.
3. Calculate the values of (36), (37) and check whether they are nearly zero in terms of a tolerance limit. If so, the equation system has been solved.
4. If not, we calculate the first-order derivative of (36) with respect to B in order to find the second step in the Newton-Raphson process.
5. After having made (36) equal to zero for $B^{(1)}$ we calculate the residuals $u'^{(1)} = Y' - B'^{(1)}X_t$ and $\hat{\sigma}_\varepsilon^2$ as

$$(41) \qquad \hat{\sigma}_\varepsilon^2 = \frac{1}{T}\sum_{t=1}^{T} E[u_t^{(1)}u'^{(1)}_t) \mid Y_t \epsilon A_t]$$

where A_t is either a point $Y_t \epsilon R$ or the interval $[-\infty,0]$.

6. Start again with the new estimates.

We notice that this process is not completely straightforward as we do not change B and σ_ε^2 simultaneously, but solve (38) for $\hat{B}^{(1)}$ given $\hat{\sigma}_\varepsilon^{(0)}$, then correct $\hat{\sigma}_\varepsilon^{(0)}$ into $\hat{\sigma}_\varepsilon^{(1)}$ and restart the search for $\hat{B}^{(2)}$ and so on. However, this procedure of successive search may be generalized to m>1, while simultaneous minimization in more dimensions would give difficulties with respect to the positive-definiteness of Σ_ε.

Let us now consider the calculations in detail.

If $Y_t \leq 0$ it implies that $\varepsilon_t \leq -B'X_t$. Consider now the conditional distribution function of ε_t. It is given by

$$(42) \qquad F(\varepsilon \mid \varepsilon \leq -B'X_t) = \begin{cases} \dfrac{1}{N(-B'X_t;0,\sigma_\varepsilon^2)} \displaystyle\int_{-\infty}^{\varepsilon} n(\tau;0,\sigma_\varepsilon^2)d\tau & \text{if } \varepsilon \leq -B'X_t \\ 1 & \text{if } \varepsilon > -B'X_t \end{cases}$$

In this case we do not need importance sampling as we can draw directly from the distribution. We draw an η from a uniform distribution on [0,1] and

calculate the corresponding ε from the inverse of (42), i.e., we solve $F(\varepsilon|.)=\eta$ (see Figure 7).

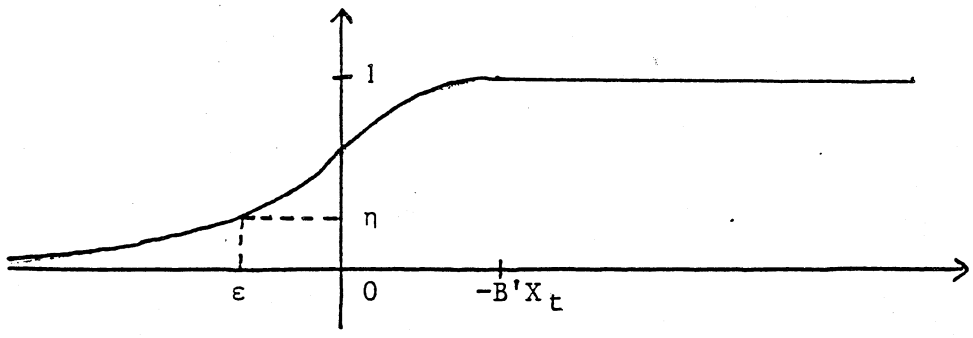

Figure 7. Drawing from a truncated normal distibution.

The corresponding drawing for Y'_t is $(B'X_t+\varepsilon')$. Let us assume we have k drawings $\{Y_{ti}\}_{i=1}^{k}$ for one X_t, then we may define the corresponding residuals u_{ti} and

$$X_t E(u_t|Y_t\leq 0)\approx \frac{1}{k} X_t \sum_{i=1}^{k} u_{ti} \tag{43}$$

and

$$E(u_t u'_t|Y_t\leq 0) \approx \frac{1}{k} \sum_{i=1}^{k} (u_{ti}u'_{ti})^2 \tag{44}$$

Notice that (43) and (44) are simultaneously assessed.
In practice k may be taken rather small, say k=4, if T is large.
So (36) and (37) may be evaluated.
In case of a non-exact observation differentiation of (38) with respect to B yields a (4x4)-matrix, consisting of two parts.

The latter part is $-\frac{1}{T\sigma_\varepsilon^2} X_t X_t'$. The first part is ignoring the factor $(1/T\sigma_\varepsilon^2)$

$$(45) \qquad \frac{\partial}{\partial B'}[X_t E(Y_t|Y_t\leq 0)] = X_t \frac{\partial}{\partial B'}\left[\frac{1}{P(Y_t\leq 0)}\int_{-\infty}^{0} \frac{y\cdot\exp(-\frac{1}{2\sigma^2}(y-B'X_t)^2)}{\sigma\sqrt{2\pi}}\,dy\right]$$

$$= X_t E(Y_t|Y_t\leq 0)\cdot\left(\frac{-1}{[P(Y_t\leq 0)]}\frac{\partial P}{\partial B'}\right) +$$

$$+ \frac{1}{P(Y_t\leq 0)} X_t \int_{-\infty}^{0} \frac{\partial}{\partial B'}\left[\frac{y\cdot\exp(-\frac{1}{2\sigma^2}(y-B'X_t)^2}{\sigma\sqrt{2\pi}}\right]dy.$$

We notice that

$$(46) \qquad \frac{1}{[P(Y_t\leq 0)]}\frac{\partial P(Y_t\leq 0)}{\partial B'} = \frac{\partial \ln P(Y_t\leq 0)}{\partial B'}$$

Applying (8) and (36) we have

$$\frac{\partial \ln P(Y_t\leq 0)}{\partial B'} = \frac{1}{\sigma_\varepsilon^2} E(u'_t|Y_t\leq 0)X_t'$$

The first part of (45) may be written

$$(47) \qquad -X_t \frac{1}{\sigma_\varepsilon^2} E(u_t|Y_t\leq 0)E(Y'_t|Y_t\leq 0)X_t'$$

In a similar way the second part of (45) may be written as

$$(48) \qquad X_t \frac{1}{\sigma_\varepsilon^2} E(u_t\ Y'_t | Y_t \leq 0) X'_t$$

Combination gives now the (n×n)-matrix

$$(49) \qquad \frac{\partial^2 \hat{L}_\alpha}{\partial B \partial B'} = \frac{1}{T\sigma_\varepsilon^2}\left[-\sum_{t=1}^{T} (X_t X'_t) + \frac{1}{\sigma_\varepsilon^2} \sum_{t=1}^{T} X_t \operatorname{var}(u_t | Y_t \leq 0) X'_t \right]$$

In the case of exact observation u_t and Y_t are exactly observed, and consequently the second term vanishes as the covariance of two non-random variables is zero.

Using this iterative procedure for which we require in this case 4 adjustments of σ_ε^2 and about 6 iterations per fixed σ_ε^2, i.e., 25 iterations in total we end up with the estimates of B and of σ_ε^2. The latter is found by solving (39) for σ_ε^2 yielding

$$(50) \qquad \hat{\sigma}_\varepsilon^2 = \frac{1}{T} \sum_{t=1}^{T} E((Y_t - B'X_t)^2 | Y_t \leq 0).$$

In Table 1 we present the results of OLS-regression on exact observations with standard deviations, the corresponding results of classical Tobit-estimation and the estimation results according to our procedure. In the last case we assess the standard deviations of $\hat{B}$ by taking the negative inverse of (49).

An R^2 is easily defined as

$$R^2 = 1 - \frac{\sum_{t=1}^{T} E((Y_t - B'X_t)^2 | Y_t \leq 0)}{\sum_{t=1}^{T} [E(Y_t^2 | Y_t \leq 0) - \{E(Y_t | Y_t \leq 0)\}^2]}$$

Notice that also for the computation of the denominator we need an estimate of B and σ^2.
Actually R^2, thus defined, is identical to the R^2 advocated by Gouriéroux et al. (1984).

Notice that $-\left(\frac{\partial^2 L}{\partial B \partial B'}\right)^{-1} = \sigma_\varepsilon^2 \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t'\right)^{-1}$

is the covariance matrix under exact observation.

TABLE 1. COMPARISON OF ESTIMATION RESULTS (N=2000)

	B_0	B_1	B_2	B_3	σ_ε^2	R^2
OLS' 1)	1.00	-2.01	2.93	1.05	1.53	.85
(st.error)	(.03)	(.03)	(.04)	(.04)		
Classical Tobit	.98	-2.00	2.98	1.05	1.21	.83
	(.04)	(.04)	(.06)	(.05)		
New Tobit 2)	1.01	-1.98	2.93	1.04	1.45	.86
	(.03)	(.04)	(.05)	(.05)		

1) Estimated on the complete Y-vector.
2) The program took 1 minutes and 26 seconds on a VAX 8650 for 4 rounds and a total of 25 iterations. For a usual mainframe the time has to be multiplied by a factor between 1/15 and 1/20.

Obviously, as said before, the Tobit-case is not a spectacular case, as the classical method is just as powerful. However, it is apparent how this method may be generalized to more difficult cases. Actually the observations may be filtered by any partition on R. We may replace $Y_t \leq 0$ by $Y_t \in A_t$ including exact observation if A_t is a point.

Let us now turn to a more-dimensional example.
More specifically, we assume that we study a random phenomenom $(Y,X) \in R^{m+n}$. We do observe X exactly but we are only able to observe one Y-component per observation, viz., the greatest of the m Y-components.

It behaves according to the system

$$Y' = B'\tilde{X} + \varepsilon'$$

where Y is an m-vector, X_i a column-vector of length n_i chosen from X,

$$\tilde{X} = \begin{array}{c} \\ \hat{n}_1 \\ \\ \\ n_m \checkmark \end{array} \begin{array}{c} \overset{<}{} \\ \left[\begin{array}{c} X_1 \\ 0 \\ \vdots \\ 0 \end{array}\right. \end{array} \begin{array}{c} \overset{m}{} \\ \begin{array}{c} 0 \\ X_2 \\ \\ \end{array} \end{array} \begin{array}{c} \overset{>}{} \\ \left.\begin{array}{c} 0 \\ \\ \\ X_m \end{array}\right] \end{array}$$

$\tilde{X}$ is a matrix of dimension $(\sum_{i=1}^{m} n_i) \times m$, B' a row-vector of length $\sum_{i=1}^{m} n_i = \tilde{n}$

and ε a random error-vector of length m that is $N(0,\Sigma_\varepsilon)$-distributed. The parameters B and Σ_ε have to be estimated.

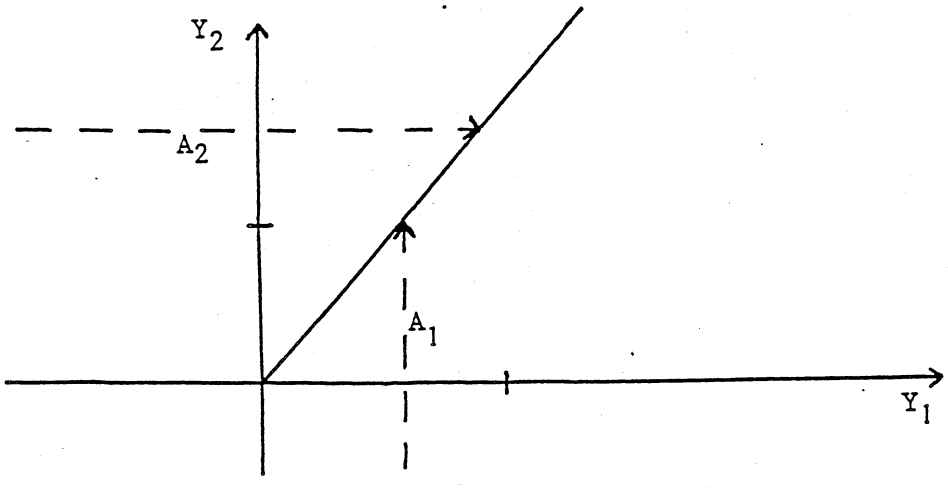

Figure 8. A partition according to maxima.

From Figure 8. it is clear for m=2 that $(Y_1,Y_2)\epsilon A_1$ if Y_1 is observed and in A_2 if Y_2 is observed.

If we have $Y_t\epsilon A_t, X_t=x_t$, it follows that the residual $Y_t'-B'X_t = u_t'$ is $N(0,\Sigma_\varepsilon)$ -distributed. Let us now define the set $\tilde{A}_t=\{u_t\epsilon R^m | u'_t = Y'_t-B'\overset{"}{X}_t, Y\epsilon A_t\}$. The set $\tilde{A}_t$ is $A_{t\odot}$ translated over a vector $B'\overset{"}{X}_t$. Then the conditional density of u_t given that it is in $\tilde{A}_t$ and $X_t=x_t$ is

$$(51)\qquad f(\tilde{u})=\begin{cases} C.n(\tilde{u};0,\Sigma_\varepsilon) & \text{if } u_t\epsilon\tilde{A}_t \\ 0 & \text{if } u_t\notin\tilde{A}_t \end{cases}$$

where C is a normalization constant. Notice that in the case considered in Figure 8. either Y_{1t} or Y_{2t} is exactly known. So in (51) either u_{1t} or u_{2t} is known.

Let us now assume a drawing distribution h(z) on $\tilde{A}_t$ that has support $\tilde{A}_t$ or a set $B\supset\tilde{A}_t$.

Then we may evaluate by means of (32) the m-vector

$$(52)\qquad E(u_t|Y_t\epsilon A_t)\approx \frac{\sum_{i=1}^{k} u_i f(u_i)h^{-1}(u_i)}{\sum_{i=1}^{k} f(u_i)h^{-1}(u_i)}$$

and similarly we find for the conditional second-order moments

$$(53)\qquad E(u_tu_t'|Y_t\epsilon A_t)\approx \frac{\sum_{i=1}^{k} u_iu_i' f(u_i)h^{-1}(u_i)}{\sum_{i=1}^{k} f(u_i)h^{-1}(u_i)}$$

From (52) we may derive $E(Y_t|Y_t\epsilon A_t) = E(u_t|Y_t\epsilon A_t)+ B'\overset{"}{X}_t$.

Then the combination of (52) and (53) yields the conditional covariance matrix of u_t and hence of Y_t.

Notice that in (52) and (53) the unknown constant C cancels out.

Consider now the normal equations in this filtered case. We have the straightforward generalization of (36) and (37) yielding

$$(54) \qquad \frac{\partial \hat{L}\,(B,\Sigma_\varepsilon^{-1})}{\partial B} = \frac{1}{T}\sum_{t=1}^{T}[\overset{''}{X}_t\Sigma_\varepsilon^{-1}E(Y_t|Y_t\epsilon A_t)-(\overset{''}{X}_t\Sigma_\varepsilon^{-1}\overset{''}{X}_t')B] = 0$$

$$(55) \qquad \frac{\partial \hat{L}\,(B,\Sigma_\varepsilon^{-1})}{\partial \Sigma_\varepsilon^{-1}} = \frac{1}{2T}\sum_{t=1}^{T}[\Sigma_\varepsilon - E(u_t u_t'|u_t\epsilon\tilde{A}_t)] = 0$$

By means of (52) and (53) we are able to solve (54) and (55) by iteration.

We start with $\Sigma_\varepsilon^{(o)}=I$ and $B^{(0)}=0$ and solve (54) for $B_0^{(1)}$. Similarly to (49) we may formulate the Hessian, needed for the Newton-Raphson process as a $(\overset{''}{n}\times\overset{''}{n})$-matrix.

In the general case the Hessian, as calculated in (49), changes into

$$(56) \qquad \frac{\partial^2 \hat{L}_a}{\partial B\partial B'} = \frac{1}{T}\left[-\sum_{t=1}^{T}\overset{''}{X}_t\Sigma_\varepsilon^{-1}\overset{''}{X}'_t + \sum_{t=1}^{T}\overset{''}{X}_t\Sigma_\varepsilon^{-1}\mathrm{var}(u_t|Y_t\epsilon A_t)\Sigma_\varepsilon^{-1}\overset{''}{X}'_t\right]$$

After the first round we calculate $\Sigma_\varepsilon^{(1)}$ from (52) and repeat the solution of (51) with respect to $B^{(1)}$, until the process has converged. If it converges, we know that it will converge to the ML-solution for the exact-observation case.

Finally we may calculate as a measure of goodness-of-fit

$$(57) \qquad R^2 = 1 - \frac{\sum_{t=1}^{T} E((Y_t' - B'\overset{''}{X}_t)\Sigma_\varepsilon^{-1}(Y_t' - B'\overset{''}{X}_t)'|Y_t\epsilon A_t)}{\sum_{t=1}^{T}\left[E(Y_t'\Sigma_\varepsilon^{-1}Y_t|Y_t\epsilon A_t) - (\bar{Y}'\Sigma_\varepsilon^{-1}\bar{Y})\right]}$$

where $$\bar{Y} = \frac{1}{T}\sum_{t=1}^{T}E(Y_t|Y_t\epsilon A_t)$$

This procedure has been performed on a simulated data set with N=2000.
We assume m=3 and n_i=4 for i=1,2,3 with X_0=1 for all t.
We assume for (X_1,X_2,X_3)

$$\mu_x = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \Sigma_x = \begin{bmatrix} 1 & & \\ 0 & 0.5 & \\ 0.5 & 0.25 & 0.5 \end{bmatrix}$$

$$E(\varepsilon) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } E(\varepsilon\varepsilon')=\Sigma_\varepsilon = \begin{bmatrix} 1 & & \\ 0.5 & 2 & \\ 0 & 1 & 1.5 \end{bmatrix}$$

finally *) $$\tilde{B} = \begin{bmatrix} 2 & 2 & 3 & -1 \\ 1 & 1 & -1 & 0 \\ 0 & 0 & .5 & -3 \end{bmatrix}$$

*) In our formulae we denote $B=vec(\tilde{B}')$.

The filter that we observe is the maximum of (Y_1,Y_2,Y_3), i.e. we know which of the three is maximal and what is its value.
Let us assume $Y_1=\max(Y_1,Y_2,Y_3)=y_1$. Then $A_1=\{Y_1=y_1, Y_2\leq y_1, Y_3\leq y_1\}$

The simulation is done as follows. We consider $u_t'=Y'_t-B'\overset{''}{X}_t$. It is trivariate normal $N(0,\Sigma_\varepsilon)$.
We have for the density

(58) $$n(u_1,u_2,u_3)=n(u_1).n(u_2|u_1).n(u_3|u_1,u_2).$$

Those conditional densities are all normal, where we use the well-known

formulae for conditional expectations and variances for the normal distribution; that is,

Let (X_1,X_2) be normal with expectation(μ_1,μ_2) and $\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$

then

$$E(X_1|X_2=x_2)=\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$$

and

$$var(X_1|X_2=x_2)=\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Using the formulae (58) and (52) we draw $\tilde{Y}$ successively. We start with u_{1t} which in this case equals the first component of $Y'_t-B'X''_t$, then we draw u_{2t} from a normal distribution truncated on $(-\infty,u_{1t})$ where the expectation and variance are calculated given $u_1=u_{1t}$.

Then we draw u_{3t}, given (u_{1t},u_{2t}) in the same way from the truncated density $n(u_3|u_{1t},u_{2t})$.

The process converges in about 22 Σ-adjustments with per stage about 10 iterations with respect to B. For each observation we draw on average 10 times.
Notice that we actually increase the number of drawings per observation after each stage to improve the accuracy. For one iteration we need 2kN (this case maximally 40000) random drawings. We re-utilize the same sequence of random drawings in subsequent stages, which implies that the drawing of the random numbers on [0,1] can be done once before the iteration process starts.

Notice that in this case $h(z) \equiv n(z)$, implying that in (28)
$f(z)h^{-1}(z)\equiv 1$.
So also here in fact we do not use importance sampling.

In Table 2 we present the estimation results for successive rounds and in the last part we give for comparison the estimation results based on the exact observations.

We see that the procedure yields very satisfactory results.

As a last example we tried a six-equation model where we observe the largest of $(Y_1,Y_2,Y_3,Y_4,Y_5,Y_6)$, The matrix Σ_ε was taken to be diagonal. The results are presented in Table 3.

TABLE 2. ESTIMATION RESULTS OF A THREE DIMENSIONAL EXAMPLE

	B_o	B_1	B_2	B_3	Σ_ε		
round 1							
1st.equation	1.94	2.05	3.09	-1.04	.97		
2nd. equation	1.17	1.09	-.72	.20	.09	1.35	
3rd.equation	-.16	.04	.70	-3.28	.02	.02	1.14
round 8							
1st. equation	2.06	1.97	2.89	-.96	.92		
2nd. equation	.91	1.03	-.99	.33	.38	2.18	
3rd. equation	-.21	-.02	.68	-3.21	.16	.31	1.44
round 15							
1st. equation	2.10	1.97	2.87	-.98	.92		
2nd. equation	1.03	1.09	-.82	.21	.53	2.15	
3rd. equation	-.08	.01	.68	-3.07	.20	.46	1.52
round 22 (final)							
1st. equation	2.13	1.96	2.82	-0.97	.92		
(st.error)	(.03)	(.05)	(.06)	(.07)			
2nd. equation	1.15	1.10	-.70	.16	.61	2.04	
	(.06)	(.08)	(.10)	(.13)			
3rd. equation	-.06	.02	.68	-3.05	.21	.49	1.52
	(.07)	(.08)	(.10)	(.13)			
Pseudo R^2 = .81							
Exact observations							
1st. equation	2.02	2.01	3.01	-1.01	.96		
	(.02)	(.04)	(.04)	(.06)			
2nd. equation	1.02	1.04	-.99	.04	.50	2.02	
	(.03)	(.05)	(.06)	(.09)			
3rd. equation	.03	-.00	.50	-2.94	.02	.96	1.42
	(.03)	(.05)	(.05)	(.07)			
Pseudo R^2 = .82							

Remark: the program ran for 1 hour 20 minutes on a VAX 8650, for 22 rounds with on average 10 iterations per round.

TABLE 3. ESTIMATION RESULTS OF A SIX DIMENSIONAL EXAMPLE

		B_o	B_1	B_2	B_3	Σ_ε(diag)	
1st equation	1.14 (1)	-1.94 (-2)	2.89 (3)	.94 (1)		.97	(1)
(st. error)	(.04)	(.04)	(.05)	(.04)			
2nd equation	-.37 (-½)	1.08 (1)	1.45 (1½)	-.84 (-1)		1.01	(1)
(st.error)	(.05)	(.05)	(.08)	(.06)			
3rd equation	.61 (1/3)	.92 (1)	-1.24 (-1½)	.45 (½)		.87	(1)
(st.error)	(.04)	(.04)	(.05)	(.04)			
4th equation	.16 (1/4)	2.04 (2)	-.09 (0)	1.06 (1)		1.04	(1)
(st.error)	(.05)	(.04)	(.07)	(.05)			
5th equation	1.06 (1)	1.03 (1)	1.94 (2)	-.95 (-1)		.91	(1)
(st.error)	(.04)	(.04)	(.06)	(.04)			
6th equation	.91 (1)	-.65 (-½)	3.12 (3)	-.02 (0)		1.01	(1)
(st.error)	(.04)	(.05)	(.06)	(.05)			

Pseudo R^2 = .83

Remark: the program ran 1 hour and 20 minutes on a VAX8650, for 5 rounds with on average 40 iterations per round. (Within branckets to the right of the parameter its true values).

CONCLUSION.

In this paper we generalized the theory of ML-estimation on exact observations to deal with estimation on filtered observations in such a way that we got a feasible computer procedure. Obviously the method draws heavily on the newly developed simulation technology. From some simulated data sets it is clear that the method works on linear models with normal errors. However, the method is not restricted to that situation. Theoretically, the model is also applicable to non-linear models, non-normal errors and filtering partitions $\mathcal{A}_t$ that vary over observations t=1,...,T. The only requirement is that $\mathcal{A}_t$ consists of P-measurable subsets of R^{m+n}. Also we do not need the dichotomy into variables X and Y, but a similar method may be used for filtered observation of structural models, e.g.,of the type $Y' = \Gamma' \overset{''}{Y} + B' \overset{''}{X} + \varepsilon'$.

REFERENCES

Amemiya, T., (1981), "Qualitative Response Models: A Survey", Journal of Economic Literature 19, pp. 1483-1536.

Amemiya, T., (1985), Advanced Econometrics, Harvard University Press, Cambridge (Mass.), U.S.A.

Billingsley, P., (1979), Probability and Measure, Wiley and Sons, New York.

Cramèr, H., (1951), Mathematical Methods of Statistics, Princeton University Press, Princeton.

Domencich, T.A., and McFadden, D. (1975), Urban Travel Demand, North-Holland Publishing Cy., Amsterdam.

Fair, R.C., and Jaffee, D.M., (1972) "Methods of Estimation for Markets in Disequilibrium", Econometrica 49, pp. 497-514.

Gourièroux, C., Monfort, A., Renault, E., Trognon, A., (1984), "Rèsidus Géneralisés ou Interprétations Linéaires de l'Econométric Non-Linéaire", forthcoming in Annals of Econometrics.

Heckman, J.J., (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator For Such Models", Annals of Economic and Social Measurement 5, pp. 475-492.

Maddala, G.S., (1983), Limited-Dependent and Qualitative Variables in Econometrics, Cambridge University Press, England.

Rao, C.R., (1973) Linear Statistical Inference and its Applications, 2^{d} ed., Wiley and Sons, New York.

Van Dijk, H.K. (1984), Posterior analysis of Econometric Models using Monte Carlo Integration, Erasmus University, Rotterdam.