Partial likelihood and estimating equations

Priscilla E. Greenwood
University of British Columbia

Wolfgang Wefelmeyer
University of Siegen

Abstract

Consider a regression model for discrete-time stochastic processes, with a (partially specified) model for the conditional distribution of the response given the covariate and the past observations. Suppose we also have some knowledge about how the parameter of interest affects the conditional distribution of the covariate given the past. We assume that these two model assumptions give rise to two martingale estimating functions, and determine an optimal combination. We indicate for the case of jump processes how our result carries over to continuous time. The resulting estimators are efficient.

1 Introduction

Suppose we know something about how the parameter of interest in a regression model appears both in the conditional distribution of the response given the covariate, and in the distribution of the covariate. How can we exploit this knowledge? Let us illustrate our approach in the case of independent and identically distributed observations (X_i, Y_i), with X_i the covariate and Y_i the response. In a regression model one usually specifies the conditional distribution of Y given X, either fully, by a parametric model, or partially. An example of a partial specification is a model for the conditional mean of Y given X, say $\mathrm{E}(Y|X) = \vartheta X$. More generally, we specify a function $\overline{g}_\vartheta(X, Y)$ such that $\mathrm{E}(\overline{g}_\vartheta(X, Y)|X) = 0$. In the example, $\overline{g}_\vartheta(X, Y) = Y - \vartheta X$. We assume a similar partial specification of the distribution of the covariate X, say $\mathrm{E}g_{*\vartheta}(X) = 0$. The two functions $\overline{g}_\vartheta$ and $g_{*\vartheta}$ give rise to two estimating equations

$$\sum_{i=1}^n \overline{g}_\vartheta(X_i, Y_i) = 0, \quad \sum_{i=1}^n g_{*\vartheta}(X_i) = 0.$$

By the usual Taylor expansion argument, their solutions $\vartheta = \hat{\vartheta}_n$ are asymptotically normal with variances I^{-1} and I_*^{-1}, respectively, where

$$I = \frac{(\mathrm{E}\overline{g}'_\vartheta)^2}{\mathrm{E}\overline{g}^2_\vartheta}, \quad I_* = \frac{(\mathrm{E}g'_{*\vartheta})^2}{\mathrm{E}g^2_{*\vartheta}}.$$

[1] Work supported by NSERC, Canada.

Since $\overline{g}_\vartheta(X_i, Y_i)$ and $g_{*\vartheta}(X_i)$ are uncorrelated, the combined estimating equation

$$\sum (w\overline{g}_\vartheta(X_i, Y_i) + w_* g_{*\vartheta}(X_i)) = 0$$

leads to an estimator with asymptotic variance

$$\frac{w^2 \mathrm{E}\overline{g}_\vartheta^2 + w_*^2 \mathrm{E} g_{*\vartheta}^2}{(w\mathrm{E}\overline{g}'_\vartheta + w_* \mathrm{E} g'_{*\vartheta})^2}.$$

Applying the Schwarz inequality to the denominator, one sees that this variance is minimized for

$$w^{\mathrm{opt}} = \frac{\mathrm{E}\overline{g}'_\vartheta}{\mathrm{E}\overline{g}_\vartheta^2}, \quad w_*^{\mathrm{opt}} = \frac{\mathrm{E} g'_{*\vartheta}}{\mathrm{E} g_{*\vartheta}^2}.$$

The minimal asymptotic variance is $(I + I_*)^{-1}$. The weights w^{opt} and w_*^{opt} depend on ϑ and, in general, also on other features of the distribution of (X, Y). In the estimating function, they must be replaced by estimators $\hat{w}_n^{\mathrm{opt}}$ and $\hat{w}_{*n}^{\mathrm{opt}}$, say by using empirical estimators for the distributions involved. This does not change the asymptotic variance $(I + I_*)^{-1}$.

Can we do better than using the combined estimating equation? Note that we can multiply $\overline{g}_\vartheta(X, Y)$ by a function $\overline{w}(X)$ of X and still have conditional expectation zero,

$$\mathrm{E}(\overline{w}(X)\overline{g}_\vartheta(X, Y)|X) = 0.$$

This leads to new estimating equations

$$\sum (\overline{w}(X_i)\overline{g}_\vartheta(X_i, Y_i) + w_* g_{*\vartheta}(X_i)) = 0 \tag{1.1}$$

with asymptotic variance $(I_{\overline{w}} + I_*)^{-1}$, where

$$I_{\overline{w}} = \frac{(\mathrm{E}\overline{w}\,\overline{g}'_\vartheta)^2}{\mathrm{E}\overline{w}^2\overline{g}_\vartheta^2} = \frac{(\mathrm{E}(\overline{w}(X)\,\mathrm{E}(\overline{g}'_\vartheta(X, Y)|X)))^2}{\mathrm{E}(\overline{w}(X)^2 \mathrm{E}(\overline{g}_\vartheta(X, Y)^2|X))}.$$

Applying again the Schwarz inequality, one sees that $I_{\overline{w}}$ is maximized by

$$\overline{w}^{\mathrm{opt}}(X) = \frac{\mathrm{E}(\overline{g}'_\vartheta(X, Y)|X)}{\mathrm{E}(\overline{g}_\vartheta(X, Y)^2|X)}.$$

The weight again depends on ϑ and, in general, also on other features of the distribution of (X, Y) and must be estimated, say by using appropriate nonparametric estimators for the conditional expectations involved. With $\hat{\overline{w}}_n^{\mathrm{opt}}(X)$ denoting such an estimator, we arrive at the estimating equation

$$\sum \left(\hat{\overline{w}}_n^{\mathrm{opt}}(X_i)\overline{g}_\vartheta(X_i, Y_i) + \hat{w}_{*n}^{\mathrm{opt}} g_{*\vartheta}(X_i)\right) = 0. \tag{1.2}$$

The asymptotic variance of the estimator corresponding to this equation is $(\overline{I} + I_*)^{-1}$ with

$$\overline{I} = \mathrm{E}\frac{(\mathrm{E}(\overline{g}'_\vartheta(X, Y)|X))^2}{\mathrm{E}(\overline{g}_\vartheta(X, Y)^2|X)}.$$

By the Schwarz inequality (2.10) below, $\overline{I}$ is strictly larger than I unless both conditional expectations do not depend on X.

For the example given above, $\overline{g}_\vartheta(X,Y) = Y - \vartheta X$, we have $\mathrm{E}(\overline{g}'_\vartheta(X,Y)|X) = -X$ and $\mathrm{E}(\overline{g}_\vartheta(X,Y)^2|X) = \mathrm{E}((Y-\vartheta X)^2|X)$, the conditional variance of Y given X.

The estimating equation (1.2) is not only optimal among estimating equations (1.1) but even efficient among *all* (regular) estimators as long as we do not impose additional restrictions on the distribution of (X,Y) which involve ϑ. Let us give a sketch of the argument, referring to Bickel, Klaassen, Ritov and Wellner (1993) for an account of the concepts involved.

The model is described by all distributions $p(dx,dy) = p_*(dx)\overline{p}(x,dy)$ of (X,Y) such that

$$\int \overline{p}(x,dy)\overline{g}_\vartheta(x,y) = \mathrm{E}(\overline{g}_\vartheta(x,Y)|x) \;=\; 0 \quad \text{for all } x, \tag{1.3}$$

$$\int p_*(dx)g_{*\vartheta}(x) = \mathrm{E}g_{*\vartheta} \;=\; 0 \tag{1.4}$$

if ϑ is true. Introduce a local model by perturbing $\overline{p}$ and p_* as $\overline{p}(x,dy)(1+n^{-1/2}u\overline{h}(x,y))$ and $p_*(dx)(1+n^{-1/2}uh_*(x))$ such that the two conditions (1.3) and (1.4) hold (approximately) with ϑ replaced by $\vartheta + n^{-1/2}u$:

$$\begin{aligned}\int \overline{p}(x,dy)\left(1+n^{-1/2}u\overline{h}(x,y)\right)\overline{g}_{\vartheta+n^{-1/2}u}(x,y) &= 0 \quad \text{for all } x,\\ \int p_*(dx)\left(1+n^{-1/2}uh_*(x)\right)g_{*,\vartheta+n^{-1/2}u}(x) &= 0.\end{aligned}$$

Then $\overline{h}$ and h_* must fulfill

$$\int \overline{p}(x,dy)\left(\overline{h}(x,y)\overline{g}_\vartheta(x,y) + \overline{g}'_\vartheta(x,y)\right) \;=\; 0 \quad \text{for all } x, \tag{1.5}$$

$$\int p_*(dx)\left(h_*(x)g_{*\vartheta}(x) + g'_{*\vartheta}(x)\right) \;=\; 0. \tag{1.6}$$

The perturbed p is approximately

$$p(dx,dy)\left(1+n^{-1/2}u\left(\overline{h}(x,y)+h_*(x)\right)\right).$$

This means that the *tangent space* in the sense of Bickel et al. (1993, p. 50, Definition 2) consists of the functions $u(\overline{h}+h_*)$ with $\overline{h}$ and h_* fulfilling (1.5) and (1.6). We view the parameter ϑ as a function of p and determine its *canonical gradient* in the sense of Bickel et al. (1993, p. 58). This is a function $\dot{\nu}$ in the tangent space such that

$$n^{1/2}(\vartheta + n^{-1/2}u - \vartheta) = u = u\mathrm{E}\dot{\nu}(\overline{h}+h_*) \quad \text{for all } \overline{h}, h_* \text{ fulfilling } (1.5), (1.6).$$

According to Bickel et al. (1993, p. 63, Theorem 2B, and p. 65, Theorem 1A), an estimator $\hat{\vartheta}_n$ is regular and efficient if and only if

$$n^{1/2}(\hat{\vartheta}_n - \vartheta) = n^{-1/2}\sum_{i=1}^n \dot{\nu}(X_i,Y_i) + o_P(1). \tag{1.7}$$

In particular, a lower bound for the asymptotic variance of regular estimators of ϑ is $\mathrm{E}\dot{\nu}^2$.

Since the tangent space is generated by the affine space of functions $\overline{h}+h_*$ with $\overline{h}$ and h_* fulfilling (1.5) and (1.6), we can write the canonical gradient as $\dot{\nu} = (\mathrm{E}(\overline{s}+s_*)^2)^{-1}(\overline{s}+s_*)$, where $\overline{s}+s_*$ is the *optimal score function*, minimizing $\mathrm{E}(\overline{h}+h_*)^2$ over all $\overline{h}$ and h_* fulfilling (1.5) and (1.6). In particular, the lower bound for the asymptotic variance of regular estimators can be written $1/\mathrm{E}(\overline{s}+s_*)^2$. The function $\overline{s}+s_*$ is characterized by

$$\mathrm{E}(\overline{s}+s_*)(\overline{h}+h_*) = \mathrm{E}(\overline{s}+s_*)^2 \quad \text{for all } \overline{h}, h_* \text{ fulfilling } (1.5), (1.6).$$

Since $\overline{h}$ and h_* are orthogonal, this is equivalent to

$$\mathrm{E}\overline{s}\overline{h} = \mathrm{E}\overline{s}^2 \quad \text{for all } \overline{h} \text{ fulfilling } (1.5), \quad \mathrm{E}s_*h_* = \mathrm{E}s_*^2 \quad \text{for all } h_* \text{ fulfilling } (1.6).$$

One easily checks that the solution is

$$\overline{s}(x,y) = -\overline{w}^{\mathrm{opt}}(x)\overline{g}_\vartheta(x,y), \quad s_*(x) = -w_*^{\mathrm{opt}}g_{*\vartheta}(x).$$

By the usual Taylor series argument, the solution $\vartheta = \hat{\vartheta}_n$ of the optimal estimating equation (1.2) is seen to fulfill

$$n^{1/2}(\hat{\vartheta}_n - \vartheta) = \left(\mathrm{E}(\overline{s}+s_*)^2\right)^{-1} n^{-1/2}\sum_{i=1}^n (\overline{s}(X_i,Y_i)+s_*(X_i)) + o_P(1).$$

By the characterization (1.7), this means that this estimator is efficient.

In Sections 2 and 3 we show how the calculation of the optimal estimating function carries over to ergodic discrete-time stochastic processes and jump processes, respectively. Efficiency also carries over, but we will not give the details. All results extend immediately to vectors ϑ and vector-valued functions $\overline{g}_\vartheta$ and $g_{*\vartheta}$. We do not give precise regularity conditions for our results.

2 Discrete-time stochastic processes

Suppose we observe a stochastic process (X_i, Y_i) at times $i = 1, \ldots, n$. The law of the process is determined by the conditional distributions $p_i(dx, dy)$ of (X_i, Y_i) given the past observations. Here and in the following, we suppress the dependence of p_i and similar objects on the past, $(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1})$. As in the i.i.d. case considered in Section 1, we describe a regression model by (partial) specifications of (1) the conditional distribution of the response given the present value of the covariate and now also the past observations, and of (2) the conditional distribution of the covariate, now also given the past. We factor p_i into marginal and conditional,

$$p_i(dx, dy) = p_{*i}(dx)\overline{p}_i(x, dy), \tag{2.1}$$

and specify two functions $\overline{g}_{i\vartheta}(x,y)$ and $g_{*i\vartheta}(x)$, possibly depending on the past, such that

$$\begin{aligned}\overline{\mathrm{E}}_i^x \overline{g}_{i\vartheta} = \int \overline{p}_i(x,dy)\overline{g}_{i\vartheta}(x,y) &= 0 \quad \text{for all } x,\\ \mathrm{E}_{*i} g_{*i\vartheta} = \int p_{*i}(dx) g_{*i\vartheta}(x) &= 0.\end{aligned}$$

They give rise to estimating equations

$$\sum_{i=1}^n \overline{g}_{i\vartheta}(X_i,Y_i) = 0, \quad \sum_{i=1}^n g_{*i\vartheta}(X_i) = 0.$$

How can we combine them in an optimal way? Our result holds for geometrically ergodic processes and under appropriate smoothness and moment conditions which can be seen from the sketch of the proof.

Result 1. *From estimating equations of the form*

$$\sum_{i=1}^n \left(\overline{w}_i(X_i)\overline{g}_{i\vartheta}(X_i,Y_i) + w_{*i} g_{*i\vartheta}(X_i)\right) = 0, \tag{2.2}$$

an estimator with minimal asymptotic variance is obtained using weights which are consistent estimators of

$$\overline{w}_i(X_i) = \overline{\mathrm{E}}_i^{X_i} \overline{g}'_{i\vartheta} / \overline{\mathrm{E}}_i^{X_i} \overline{g}^2_{i\vartheta}, \tag{2.3}$$

$$w_{*i} = \mathrm{E}_{*i} g'_{*i\vartheta} / \mathrm{E}_{*i} g^2_{*i\vartheta}. \tag{2.4}$$

The estimator is asymptotically normal. Its asymptotic variance is the limit of

$$n\left(\sum_{i=1}^n \mathrm{E}_{*i} \frac{\left(\overline{\mathrm{E}}_i^{X_i} \overline{g}'_{i\vartheta}\right)^2}{\overline{\mathrm{E}}_i^{X_i} \overline{g}^2_{i\vartheta}} + \sum_{i=1}^n \frac{\left(\mathrm{E}_{*i}^{X_i} g'_{*i\vartheta}\right)^2}{\mathrm{E}_{*i}^{X_i} g^2_{*i\vartheta}}\right)^{-1}. \tag{2.5}$$

Sketch of proof. To simplify the notation, we introduce $w = (\overline{w}, w_*)$ and $g_\vartheta = (\overline{g}_\vartheta, g_{*\vartheta})$, and write the estimating equation (2.2) as

$$\sum w_i g_{i\vartheta} = \sum \overline{w}_i \overline{g}_{i\vartheta} + \sum w_{*i} g_{*i\vartheta} = 0.$$

Let $\hat{\vartheta}$ be a solution. Under appropriate differentiability conditions, a Taylor expansion gives

$$0 = \sum w_i g_{i\hat{\vartheta}} \approx \sum w_i g_{i\vartheta} + (\hat{\vartheta} - \vartheta) \sum w_i g'_{i\vartheta}. \tag{2.6}$$

Then

$$n^{1/2}(\hat{\vartheta} - \vartheta) \approx -n^{-1/2} \sum w_i g_{i\vartheta} \Big/ \frac{1}{n} \sum w_i g'_{i\vartheta}. \tag{2.7}$$

Conditionally on the past, the martingale increments $\overline{w}_i(X_i)\overline{g}_{i\vartheta}(X_i, Y_i)$ and $w_{*i}g_{*i\vartheta}(X_i)$ are orthogonal:

$$\mathrm{E}_i\overline{w}_i\overline{g}_{i\vartheta}w_{*i}g_{*i\vartheta} = w_{*i}\int p_{*i}(dx)g_{*i\vartheta}(x)\overline{w}_i(x)\int \overline{p}_i(x,dy)\overline{g}_{i\vartheta}(x,y) = 0. \tag{2.8}$$

Introduce an inner product

$$(v, w) = \sum \mathrm{E}_{*i}\overline{v}_i\overline{w}_i + \sum v_{*i}w_{*i}$$

with corresponding norm $\|w\|^2 = (w, w)$. Interpret products vw of vectors componentwise,

$$vw = (\overline{v}_1\overline{w}_1, \dots, \overline{v}_n\overline{w}_n, v_{*1}w_{*1}, \dots, v_{*n}w_{*n}).$$

Consider first the numerator in (2.7). With (2.8), the predictable quadratic variation of $\sum w_i g_{i\vartheta}$ is $(w^2, v) = \|wv^{1/2}\|^2$, where

$$\overline{v}_i(x) = \overline{\mathrm{E}}_i^x\overline{g}_{i\vartheta}^2, \quad v_{*i} = \mathrm{E}_{*i}g_{*i\vartheta}^2.$$

Consider now the denominator in (2.7). The compensator of $\sum w_i g'_{i\vartheta}$ is (w, m), where

$$\overline{m}_i(x) = \overline{\mathrm{E}}_i^x\overline{g}'_{i\vartheta}, \quad m_{*i} = \mathrm{E}_{*i}g'_{*i\vartheta}.$$

Since $\sum w_i g'_{i\vartheta} - (w, m)$ is a martingale, $\frac{1}{n}\left(\sum w_i g'_{i\vartheta} - (w, m)\right)$ is asymptotically negligible, and we may replace $\frac{1}{n}\sum w_i g_{i\vartheta}$ by $\frac{1}{n}(w, m)$. If the process is ergodic, $\frac{1}{n}(w, m)$ is asymptotically constant. Hence the predictable quadratic variation of $n^{1/2}(\hat{\vartheta} - \vartheta)$ is approximately $n\|wv^{1/2}\|^2/(w, m)^2$. By the Schwarz inequality,

$$|(w, m)| = |(wv^{1/2}, mv^{-1/2})| \le \|wv^{1/2}\|\,\|mv^{-1/2}\|. \tag{2.9}$$

In other words,

$$\frac{\|wv^{1/2}\|^2}{(w, m)^2} \ge \frac{1}{\|mv^{-1/2}\|^2} = \frac{\|mv^{-1}\cdot v^{1/2}\|^2}{(mv^{-1}, m)^2}.$$

Hence $\|wv^{1/2}\|^2/(w, m)^2$ is minimized by $w = mv^{-1}$, and the minimum is

$$\|mv^{-1/2}\|^{-2} = \left(\sum \mathrm{E}_{*i}\frac{\overline{m}_i^2}{\overline{v}_i} + \sum \frac{m_{*i}^2}{v_{*i}}\right)^{-1}.$$

By an appropriate martingale central limit theorem, $n^{1/2}(\hat{\vartheta} - \vartheta)$ is asymptotically normal with variance equal to the limit of $n\|mv^{-1/2}\|^{-2}$, and the assertion follows. □

Efficiency of the estimator based on the optimal estimating equation can be proved by an approach similar to that outlined in Section 1 for the i.i.d. case.

If we use *predictors* to estimate $\overline{w}_i(x)$ and w_{*i}, i.e. estimators involving only the past observations $(X_1, Y_1), \dots, (X_{i-1}, Y_{i-1})$, then the optimal estimating function is a martingale.

We may allow weights $\overline{w}_i$, w_{*i} to depend on ϑ. Then the derivative of $\sum w_i g_{i\vartheta}$ in the expansion (2.6) has a second term $\sum w_i' g_{i\vartheta}$. It is asymptotically negligible since the $g_{i\vartheta}$ are martingale increments.

Remark 1. Usually one takes *predictable* weights to combine two martingale estimating functions; e.g. Heyde (1987). For the estimating functions $\sum \overline{g}_{i\vartheta}(X_i, Y_i)$ and $\sum g_{*i\vartheta}(X_i)$ this would mean using weights $\overline{w}_i$ rather than $\overline{w}_i(X_i)$. Then the best weights would be

$$\overline{w}_i = \mathrm{E}_i \overline{g}_{i\vartheta}' / \mathrm{E}_i \overline{g}_{i\vartheta}^2$$

and w_{*i} as above, and the minimal asymptotic variance would be the limit of an expression of the form (2.5) with $\mathrm{E}_{*i}\left((\overline{\mathrm{E}}_i^{X_i} \overline{g}_{i\vartheta}')^2 / \overline{\mathrm{E}}_i^{X_i} \overline{g}_{i\vartheta}^2\right)$ replaced by the simpler $(\mathrm{E}_i \overline{g}_{i\vartheta}')^2 / \mathrm{E}_i \overline{g}_{i\vartheta}^2$. The resulting variance is, in general, larger than our minimum variance (2.5) because

$$\frac{(\mathrm{E} f)^2}{\mathrm{E} g^2} \le \mathrm{E} \frac{(\mathrm{E}^X f)^2}{\mathrm{E}^X g^2}. \tag{2.10}$$

This inequality follows from the Schwarz inequality:

$$\begin{aligned} (\mathrm{E} f)^2 &= (\mathrm{E}\mathrm{E}^X f)^2 = \left(\mathrm{E} \frac{\mathrm{E}^X f}{(\mathrm{E}^X g^2)^{1/2}} (\mathrm{E}^X g^2)^{1/2} \right)^2 \\ &\le \mathrm{E} \frac{(\mathrm{E}^X f)^2}{\mathrm{E}^X g^2} \cdot \mathrm{E}\mathrm{E}^X g^2 = \mathrm{E} \frac{(\mathrm{E}^X f)^2}{\mathrm{E}^X g^2} \cdot \mathrm{E} g^2. \end{aligned}$$

Remark 2. The weight $\overline{w}_i$ depends only on $\overline{p}_i$ and $\overline{g}_{i\vartheta}$, and w_{*i} depends only on p_{*i} and $g_{*i\vartheta}$. This is due to orthogonality (2.8). Indeed, the weights (2.3) are optimal for estimating functions of the form $\sum \overline{w}_i(X_i) \overline{g}_{i\vartheta}(X_i, Y_i)$, and the weights (2.4) are optimal for estimating functions of the form $\sum w_{*i} g_{*i\vartheta}(X_i)$.

Remark 3. Suppose we have a parametric model $\overline{p}_{i\vartheta}(X_i, dy)$ for the conditional distribution of the response Y_i given the present covariate X_i and the past observations. Differentiating under the integral, we obtain

$$0 = \left(\overline{\mathrm{E}}_{i\vartheta}^x \overline{g}_{i\vartheta}\right)' = \overline{\mathrm{E}}_{i\vartheta}^x \overline{g}_{i\vartheta}' + \overline{\mathrm{E}}_{i\vartheta}^x \overline{\ell}_{i\vartheta}' \overline{g}_{i\vartheta},$$

where

$$\overline{\ell}_{i\vartheta}'(x, y) = \partial_{\tau=\vartheta} \frac{d\overline{p}_{i\tau}(x, \cdot)}{d\overline{p}_{i\vartheta}(x, \cdot)}(y).$$

Hence, by the Schwarz inequality, $\left(\overline{\mathrm{E}}_{i\vartheta}^x \overline{g}_{i\vartheta}'\right)^2 \Big/ \overline{\mathrm{E}}_{i\vartheta}^x \overline{g}_{i\vartheta}^2$ is maximal for $\overline{g}_{i\vartheta} = \overline{\ell}_{i\vartheta}'$, and the optimal weight (2.3) for $\overline{g}_{i\vartheta} = \overline{\ell}_{i\vartheta}'$ is $\overline{w}_i(X_i) = -1$.

In particular, the estimating function $\sum \overline{\ell}'_{i\vartheta}(X_i, Y_i)$ is optimal among estimating functions $\sum \overline{w}_i(X_i)\overline{g}_{i\vartheta}(X_i, Y_i)$. The optimal estimating function is the *partial score function*, i.e., the derivative of the *partial likelihood ratio* of Cox (1975),

$$\prod \frac{d\overline{p}_{i\tau}(X_i, \cdot)}{d\overline{p}_{i\vartheta}(X_i, \cdot)}(Y_i).$$

Hence the optimal estimating function gives the *maximum partial likelihood estimator.* If the observations (X_i, Y_i) are independent, the partial likelihood ratio is the *conditional* likelihood ratio for $Y_1, \ldots, Y_n$ given the covariates $X_1, \ldots, X_n$.

Similarly, if there is a parametric model $p_{*i\vartheta}$ for the distribution of the covariate X_i given the past, the optimal $g_{*i\vartheta}$ is

$$g_{*i\vartheta}(x) = \partial_{\tau=\vartheta} \frac{dp_{*i\tau}}{dp_{*i\vartheta}}(x) =: \ell'_{*i\vartheta}(x),$$

and the optimal weight (2.4) is -1.

Moreover, if there is a *fully* specified parametric model

$$p_{i\vartheta}(dx, dy) = p_{*i\vartheta}(dx)\overline{p}_{i\vartheta}(x, dy),$$

then the likelihood ratio can be written

$$\prod \frac{dp_{*i\tau}}{dp_{*i\vartheta}}(X_i) \prod \frac{d\overline{p}_{i\tau}(X_i, \cdot)}{d\overline{p}_{i\vartheta}(X_i, \cdot)}(Y_i), \tag{2.11}$$

and the optimal estimating function is the *score function* $\sum \ell'_{i\vartheta}(X_i, Y_i)$ with

$$\ell'_{i\vartheta}(x, y) := \partial_{\tau=\vartheta} \frac{dp_{i\tau}}{dp_{i\vartheta}}(x, y) = \overline{\ell}'_{i\vartheta}(x, y) + \ell'_{*i\vartheta}(x).$$

Hence the optimal estimator is the maximum likelihood estimator, and its asymptotic variance is the limit of

$$n \left(\sum \mathrm{E}_i \overline{\ell}'^2_{i\vartheta} + \sum \mathrm{E}_{*i} \ell'^2_{*i\vartheta} \right)^{-1}.$$

Remark 4. For discrete-time processes, it is common to model the conditional distribution of the response given the past *and* the present value of the covariate. In the continuous-time setting of Section 3 one usually models the conditional distribution of the response given only the past. This is just a convention: We may consider X_{i-1} rather than X_i the 'present' covariate of the response.

3 Jump processes

Suppose we observe a jump process $(X, Y) = (X_s, Y_s)_{s \geq 0}$ on a finite time interval $[0, t]$. The corresponding multivariate point process is given by the jump measure

$$\mu(ds, dx, dy) = \sum_{s:(\Delta X_s, \Delta Y_s) \neq 0} \varepsilon_{(s, \Delta X_s, \Delta Y_s)}(ds, dx, dy).$$

The law of the process is determined by the compensator of the jump measure. Assume, for simplicity, that the compensator has the form $K_s(dx, dy)ds$, so that there are no time points at which the process has a positive probability of jumping. We can write $K_s(dx, dy) = a_s p_s(dx, dy)$ with p_s a *probability* measure, the jump size distribution at time s given the past, and a_s the jump intensity. For the theory of continuous-time processes and limit theorems we refer to Jacod and Shiryaev (1987).

The multivariate point process corresponding to the response process Y is

$$\mu^Y(ds, dy) = \sum_{s:\Delta Y_s \neq 0} \varepsilon_{(s, \Delta Y_s)}(ds, dy).$$

A regression model is given by a (partial) specification of the compensator of μ^Y, say $\overline{K}_s(dy)ds$. As noted in Remark 4, this is not exactly analogous to the discrete-time case. We specify a predictable function $\overline{g}_{s\vartheta}(y)$ such that

$$\overline{K}_s \overline{g}_{s\vartheta} = \int \overline{K}_s(dy) \overline{g}_{s\vartheta}(y) = 0$$

and obtain a martingale estimating equation

$$\sum_{s \leq t:\Delta Y_s \neq 0} \overline{g}_{s\vartheta}(\Delta Y_s) = 0.$$

We want to assume a similar partial specification of the distribution of the covariate process X. It will be based on a factorization of $\overline{K}_s(dx, dy)$ analogous to the factorization (2.1) of the distribution $p_i(dx, dy)$. We must take into account the possibility that X jumps while Y does not. Following Arjas and Haara (1984) and Greenwood and Wefelmeyer (1996), we write

$$K_s(dx, dy) = K_{-0s}(dx, dy) + K_{*s}(dx)\varepsilon_0(dy),$$

where $K_{-0s}(dx, dy)$ does not charge the subspace described by $y = 0$. Then K_{*s} governs those jumps of X that do not occur simultaneously with jumps of Y. As in (2.1), but with the roles of X and Y interchanged, we factor

$$K_{-0s}(dx, dy) = \overline{K}_s(dy) K_{-*s}(dx, y). \tag{3.1}$$

We note that $K_{-*s}(dx, y)$ is a *probability* measure, the conditional distribution of the jump size of the covariate given a jump of size y of the response, and given the past.

Additional specifications of the model may now be given by predictable functions $g_{*s\vartheta}(x)$ and $g_{-*s\vartheta}(x,y)$ such that

$$\begin{aligned} K_{*s}g_{*s\vartheta} = \int K_{*s}(dx)g_{*s\vartheta}(x) &= 0, \\ K^y_{-*s}g_{-*s\vartheta} = \int K_{-*s}(dx,y)g_{-*s\vartheta}(x,y) &= 0 \quad \text{for all } y. \end{aligned}$$

They give rise to additional martingale estimating equations

$$\begin{aligned} \sum_{s\le t:\Delta X_s\neq 0,\Delta Y_s=0} g_{*s\vartheta}(\Delta X_s) &= 0, \\ \sum_{s\le t:\Delta Y_s\neq 0} g_{-*s\vartheta}(\Delta X_s,\Delta Y_s) &= 0. \end{aligned}$$

How can we combine the three estimating functions in an optimal way? Again, our result holds for geometrically ergodic processes under appropriate smoothness and moment conditions which can be seen from the sketch of the proof.

Result 2. *From estimating equations of the form*

$$\begin{aligned} &\sum_{s\le t:\Delta Y_s\neq 0} \overline{w}_s\overline{g}_{s\vartheta}(\Delta Y_s) \\ &+ \sum_{s\le t:\Delta X_s\neq 0,\Delta Y_s=0} w_{*s}g_{*s\vartheta}(\Delta X_s) \\ &+ \sum_{s\le t:\Delta Y_s\neq 0} w_{-*s}(\Delta Y_s)g_{-*s\vartheta}(\Delta X_s,\Delta Y_s) = 0, \end{aligned} \tag{3.2}$$

an estimator with minimal asymptotic variance is obtained using weights which are consistent estimators of

$$\overline{w}_s = \overline{K}_s\overline{g}'_{s\vartheta}/\overline{K}_s\overline{g}^2_{s\vartheta}, \tag{3.3}$$

$$w_{*s} = K_{*s}g'_{*s\vartheta}/K_{*s}g^2_{*s\vartheta}, \tag{3.4}$$

$$w_{-*s}(y) = K^y_{-*s}g'_{-*s\vartheta}/K^y_{-*s}g^2_{-*s\vartheta}. \tag{3.5}$$

The estimator is asymptotically normal. Its asymptotic variance is the limit of

$$t\left(\int_0^t \frac{(\overline{K}_s\overline{g}'_{s\vartheta})^2}{\overline{K}_s\overline{g}^2_{s\vartheta}}ds + \int_0^t \frac{(K_{*s}g'_{*s\vartheta})^2}{K_{*s}g^2_{*s\vartheta}}ds + \int_0^t \overline{K}_s(dy)\frac{(K^y_{-*s}g'_{-*s\vartheta})^2}{K^y_{-*s}g^2_{-*s\vartheta}}ds\right)^{-1}. \tag{3.6}$$

Sketch of proof. To simplify the notation, we introduce $w = (\overline{w}, w_*, w_{-*})$ and $g_\vartheta = (\overline{g}_\vartheta, g_{*\vartheta}, g_{-*\vartheta})$, and write the estimating equation (3.2) as

$$\sum w_s g_{s\vartheta} = \sum \overline{w}_s\overline{g}_{s\vartheta} + \sum w_{*s}g_{*s\vartheta} + \sum w_{-*s}g_{-*s\vartheta} = 0.$$

Let $\hat{\vartheta}$ be a solution. Under appropriate differentiability conditions, a Taylor expansion gives

$$0 = \sum w_s g_{s\hat{\vartheta}} \approx \sum w_s g_{s\vartheta} + (\hat{\vartheta} - \vartheta) \sum w_s g'_{s\vartheta}. \tag{3.7}$$

Then

$$t^{1/2}(\hat{\vartheta} - \vartheta) \approx -t^{-1/2} \sum w_s g_{s\vartheta} \Big/ \frac{1}{t} \sum w_s g'_{s\vartheta}. \tag{3.8}$$

The martingale $\sum w_{*s} g_{*s\vartheta}$ is orthogonal to the martingales $\sum \overline{w}_s \overline{g}_{s\vartheta}$ and $\sum w_{-*s} g_{-*s\vartheta}$ because it lives on time points s with $\Delta Y_s = 0$ while the two other martingales do not jump at these time points. Because $K_{-0s}(dx, dy)$ does not charge $y = 0$, we may and will assume that $g_{-*s\vartheta}(x, 0) = 0$. Then

$$\begin{aligned} K_s \overline{w}_s \overline{g}_{s\vartheta} w_{-*s} g_{-*s\vartheta} &= K_{-0s} \overline{w}_s \overline{g}_{s\vartheta} w_{-*s} g_{-*s\vartheta} \\ &= \overline{w}_s \int \overline{K}_s(dy) \overline{g}_{s\vartheta}(y) w_{-*s}(y) \int K_{-*s}(dx, y) g_{-*s\vartheta}(x, y) = 0. \end{aligned} \tag{3.9}$$

Hence $\sum \overline{w}_s \overline{g}_{s\vartheta}$ and $\sum w_{-*s} g_{-*s\vartheta}$ are also orthogonal. Introduce an inner product

$$(v, w) = \int_0^t \overline{v}_s \overline{w}_s + \int_0^t v_{*s} w_{*s} + \int_0^t \int \overline{K}_s(dy) v_{-*s}(y) w_{-*s}(y) ds$$

with corresponding norm $\|w\|^2 = (w, w)$. Consider first the numerator in (3.8). With (3.9) and orthogonality of $\sum w_{*s} g_{*s\vartheta}$ and $\sum \overline{w}_s \overline{g}_{s\vartheta}$, the predictable quadratic variation of $\sum w_s g_{s\vartheta}$ is $(w^2, v) = \|wv^{1/2}\|^2$, where

$$\overline{v}_s = \overline{K}_s \overline{g}^2_{s\vartheta}, \quad v_{*s} = K_{*s} g^2_{*s\vartheta}, \quad v_{-*s}(y) = K^y_{-*s} g^2_{-*s\vartheta}.$$

Consider now the denominator in (3.8). The compensator of $\sum w_s g_{s\vartheta}$ is (w, m), where

$$\overline{m}_s = \overline{K}_s \overline{g}'_{s\vartheta}, \quad m_{*s} = K_{*s} g'_{*s\vartheta}, \quad m_{-*s}(y) = K^y_{-*s} g'_{-*s\vartheta}.$$

Since $\sum w_s g'_{s\vartheta} - (w, m)$ is a martingale, $\frac{1}{t}\left(\sum w_s g'_{s\vartheta} - (w, m)\right)$ is asymptotically negligible, and we may replace $\frac{1}{t} \sum w_s g'_{s\vartheta}$ by $\frac{1}{t}(w, m)$. If the process is ergodic, $\frac{1}{t}(w, m)$ is asymptotically constant. Hence the predictable quadratic variation of $t^{1/2}(\hat{\vartheta} - \vartheta)$ is approximately $t\|wv^{1/2}\|^2/(w, m)^2$. By the inequality (2.9), this is minimized by $w = mv^{-1}$, and the minimum is

$$t\|mv^{-1/2}\|^{-2} = t \left(\int_0^t \frac{\overline{m}_s^2}{\overline{v}_s} ds + \int_0^t \frac{m_{*s}^2}{v_{*s}} ds + \int_0^t \int \overline{K}_s(dy) \frac{m_{-*s}(y)^2}{v_{-*s}(y)} ds \right)^{-1}.$$

By an appropriate martingale central limit theorem, $t^{1/2}(\hat{\vartheta} - \vartheta)$ is asymptotically normal with variance equal to the limit of $t\|mv^{-1/2}\|^{-2}$, and the assertion follows. □

As in Remark 2, the weight (3.3) is optimal for estimating equations

$$\sum_{s\le t:\Delta Y_s\neq 0} \overline{w}_s \overline{g}_{s\vartheta}(\Delta Y_s) = 0,$$

and the weights (3.4) and (3.5) have analogous optimality properties on their own.

Remark 5. Suppose we have a parametric model $\overline{K}_{s\vartheta}(dy)$ for the compensator of the jump measure μ^Y of the response. Write

$$\overline{V}_{s\vartheta\tau} = \frac{d\overline{K}_{s\tau}}{d\overline{K}_{s\vartheta}}, \quad \overline{\ell}'_{s\vartheta} = \partial_{\tau=\vartheta}\overline{V}_{s\vartheta\tau}.$$

When the intensity $\overline{K}_{s\vartheta}(\mathbf{R})$ of the response depends on ϑ, then $\overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta}$ will not be zero in general. This differs from the discrete-time case. Differentiating under the integral, we obtain

$$0 = (\overline{K}_{s\vartheta}\overline{g}_{s\vartheta})' = \overline{K}_{s\vartheta}\overline{g}'_{s\vartheta} + \overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta}\overline{g}_{s\vartheta} = \overline{K}_{s\vartheta}\overline{g}'_{s\vartheta} + \overline{K}_{s\vartheta}(\overline{\ell}'_{s\vartheta} - \overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta})\overline{g}_{s\vartheta}.$$

Using the Schwarz inequality, we see that

$$\frac{(\overline{K}_{s\vartheta}\overline{g}'_{s\vartheta})^2}{\overline{K}_{s\vartheta}\overline{g}^2_{s\vartheta}} = \frac{\left(\overline{K}_{s\vartheta}(\overline{\ell}'_{s\vartheta} - \overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta})\overline{g}_{s\vartheta}\right)^2}{\overline{K}_{s\vartheta}\overline{g}^2_{s\vartheta}}$$

is maximal for $\overline{g}_{s\vartheta} = \overline{\ell}'_{s\vartheta} - \overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta}$, and then the optimal weight is $\overline{w}_s = -1$.

In particular,

$$\sum_{s\le t:\Delta Y_s\neq 0} \left(\overline{\ell}'_{s\vartheta}(\Delta Y_s) - \overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta}\right) \tag{3.10}$$

is optimal among estimating functions of the form

$$\sum_{s\le t:\Delta Y_s\neq 0} \overline{w}_s \overline{g}_{s\vartheta}(\Delta Y_s),$$

and the asymptotic variance of the optimal estimator is the limit of

$$t\left(\int_0^t \overline{K}_{s\vartheta}(\overline{\ell}'_{s\vartheta} - \overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta})^2 ds\right)^{-1}.$$

As in Remark 3, the optimal estimating function turns out to be the partial score function in the following sense. A *partial likelihood ratio* for jump processes was introduced by Arjas and Haara (1984) as

$$\prod_{s\le t:\Delta Y_s\neq 0} \overline{V}_{s\vartheta\tau}(\Delta Y_s)\exp\left(-\int_0^t \overline{K}_{s\vartheta}(\overline{V}_{s\vartheta\tau} - 1)ds\right).$$

See also Andersen, Borgan, Gill and Keiding (1993) and, for general semimartingales, Jacod (1987) and (1990). The *partial score function* is the derivative of the partial likelihod ratio at $\tau = \vartheta$. Using $\overline{V}_{s\vartheta\vartheta} = 1$ we see that the derivative equals (3.10).

Remark 6. Suppose we have a parametric model $K_{*s\vartheta}(dx)ds$ for the compensator of the jump measure of those jumps of the covariate X that do not occur simultaneously with jumps of the response Y. Write

$$V_{*s\vartheta\tau} = \frac{dK_{*s\tau}}{dK_{*s\vartheta}}, \quad \ell'_{*s\vartheta} = \partial_{\tau=\vartheta} V_{*s\vartheta\tau}.$$

As in Remark 5, the best $g_{*s\vartheta}$ is $\ell'_{*s\vartheta} - K_{*s\vartheta}\ell'_{*s\vartheta}$, and then the optimal weight (3.4) is $w_{*s} = -1$.

In particular,

$$\sum_{s \le t : \Delta X_s \neq 0, \Delta Y_s = 0} (\ell'_{*s\vartheta}(\Delta X_s) - K_{*s\vartheta}\ell'_{*s\vartheta}) \tag{3.11}$$

is optimal among estimating functions of the form

$$\sum_{s \le t : \Delta X_s \neq 0, \Delta Y_s = 0} w_{*s} g_{*s\vartheta}(\Delta X_s).$$

Remark 7. Suppose we have a parametric model for the compensator $K_{-*s\vartheta}(dx, y)$ of the conditional jump size distribution at time s of X given a jump $Y = y$ and the past. Write

$$V_{-*s\vartheta\tau}(x, y) = \frac{dK_{-*s\tau}(\cdot, y)}{dK_{-*s\vartheta}(\cdot, y)}, \quad \ell'_{-*s\vartheta} = \partial_{\tau=\vartheta} V_{-*s\vartheta\tau}.$$

Since $K_{-*s\vartheta}(dx, y)$ is a *probability* measure, we have $K^y_{-*s\vartheta}\ell'_{-*s\vartheta} = 0$. As in Remark 5, the best $g_{-*s\vartheta}$ is $\ell'_{-*s\vartheta}$, and then the optimal weight (3.5) is $w_{-*s} = -1$.

In particular,

$$\sum_{s \le t : \Delta Y_s \neq 0} \ell'_{-*s\vartheta}(\Delta X_s, \Delta Y_s) \tag{3.12}$$

is optimal among estimating functions of the form

$$\sum_{s \le t : \Delta Y_s \neq 0} w_{-*s} g_{-*s\vartheta}(\Delta X_s, \Delta Y_s).$$

Remark 8. Suppose we have a fully parametric model $\overline{K}_{s\vartheta}$, $K_{*s\vartheta}$, $K_{-*s\vartheta}$. According to Result 2 and Remarks 5 to 7, the best estimating equation is

$$\begin{aligned} &\sum_{s \le t : \Delta Y_s \neq 0} \left(\overline{\ell}'_{s\vartheta}(\Delta Y_s) - \overline{K}_{s\vartheta}\overline{\ell}'_{s\vartheta}\right) \\ &+ \sum_{s \le t : \Delta X_s \neq 0, \Delta Y_s = 0} (\ell'_{*s\vartheta}(\Delta X_s) - K_{*s\vartheta}\ell'_{*s\vartheta}) \\ &+ \sum_{s \le t : \Delta Y_s \neq 0} \ell'_{-*s\vartheta}(\Delta X_s, \Delta Y_s) = 0. \end{aligned}$$

To show that this gives the maximum likelihood estimator, we recall a representation of Greenwood and Wefelmeyer (1996) of the likelihood ratio,

$$\prod_{s\le t:\Delta Y_s\neq 0} \overline{V}_{s\vartheta\tau}(\Delta Y_s)\exp\left(-\int_0^t \overline{K}_{s\vartheta}(\overline{V}_{s\vartheta\tau}-1)ds\right) \tag{3.13}$$
$$\prod_{s\le t:\Delta X_s\neq 0,\Delta Y_s=0} V_{*s\vartheta\tau}(\Delta X_s)\exp\left(-\int_0^t K_{*s\vartheta}(V_{*s\vartheta\tau}-1)ds\right)$$
$$\prod_{s\le t:\Delta Y_s\neq 0} V_{-*s\vartheta\tau}(\Delta X_s,\Delta Y_s).$$

For a heuristic derivation in terms of product integrals, see Andersen et al. (1993, p. 107). We have already noted in Remark 5 that the derivative of the first factor, the partial likelihood ratio, equals the partial score function (3.10). The derivative of the second factor is obtained similarly. Finally, $\partial_{\tau=\vartheta}V_{-*s\vartheta\tau}=\ell'_{-*s\vartheta}$ by definition.

The representation (3.13) of the likelihood can be used in the partially specified model of Result 2 to prove that the optimal estimating function obtained there is efficient as long as no additional restrictions involving ϑ are imposed on the model. The arguments are similar to those outlined in Section 1 for the i.i.d. case. In Greenwood and Wefelmeyer (1996) a representation analogous to (3.13) is given for general semimartingales and can be used to generalize the results obtained here to partially specified semimartingale regression models.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Arjas, E. and Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates. *Scand. J. Statist.* 11, 193–209.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Cox, D. R. (1975). Partial likelihood. *Biometrika* 62, 269–276.

Greenwood, P. E. and Wefelmeyer, W. (1996). Cox's factoring of regression model likelihoods for continuous time processes. To appear in: *Bernoulli.*

Heyde, C.C. (1987). On combining quasi-likelihood estimating functions. *Stochastic Process. Appl.* **25** 281–287.

Jacod, J. (1987). Partial likelihood process and asymptotic normality. *Stochastic Process. Appl.* 26, 47–71.

Jacod, J. (1990). Sur le processus de vraisemblance partielle. *Ann. Inst. Henri Poincaré* 26, 299–329.

Jacod, J. and Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes.* Grundlehren der mathematischen Wissenschaften 288, Springer-Verlag, Berlin.