Semiparametrics by way of parametrics and contiguity

Emil Aas Stoltenberg
joint work with Adam Lee
Department of Data Science, BI Norwegian Business School

Godt Hjort

December 5, 2023, Blindern

Complicated stuff

The Annals of Statistics
1990, Vol. 18, No. 3, 1259–1294

NONPARAMETRIC BAYES ESTIMATORS BASED ON BETA PROCESSES IN MODELS FOR LIFE HISTORY DATA

By Nils Lid Hjort

Norwegian Computing Centre and University of Oslo

Let A be any Lévy process. There exists a separable version with right-continuous paths [Breiman (1968), page 299], i.e., $\mathscr{P}(\mathscr{B}) = 1$, where $\mathscr{P}$ is the probability measure governing A. Let $\mathscr{E}$ be the expectation operator associated with $\mathscr{P}$ and let $t_1, t_2, \dots$ be the times at which A a.s. is discontinuous, say with jumps $S_j = A\{t_j\} = A(t_j) - A(t_j -)$. Then A admits a *Lévy representation*

$$(3.6)\quad \mathscr{E}\exp\{-\theta A(t)\} = \left[\prod_{j:\, t_j \le t} \mathscr{E}\exp(-\theta S_j)\right]\exp\left\{-\int_0^\infty (1 - \varepsilon^{-\theta s})\, dL_t(s)\right\}, \qquad t \ge 0, \theta \ge 0,$$

where $\{L_t;\, t \ge 0\}$ is a continuous Lévy measure. This means that L_t for each t is a measure on $(0, \infty)$, $L_t(D)$ is nondecreasing and continuous in t for each Borel set D in $(0, \infty)$, and $L_0(D) = 0$. It holds that $A(t)$ is finite a.s. whenever $\int_0^\infty s/(1+s)\, dL_t(s)$ is finite. In the Lévy formula (3.6), which follows from Ferguson [(1974) page 623], it is assumed that A contains no nonrandom part. The distribution of such a $\mathscr{P}$ is specified by $\{t_1, t_2, \dots\}$, the distributions of $S_1, S_2, \dots$ and $\{L_t;\, t \ge 0\}$.

…this is Hjort (1990).

Start 'simple', start finite-dimensional

NONPARAMETRIC BAYES ESTIMATORS 1261

search are briefly discussed in Section 7, along with some complementing remarks.

2. Nonparametric time-discrete survival analysis.

2.1. *A time-discrete model with censoring.* Let X be a variable taking values in $\mathscr{X} = \{0, b, 2b, \dots\}$ and let

$$f(jb) = \Pr\{X = jb\}, \qquad F(jb) = \Pr\{X \le jb\} = \sum_{i=0}^{jb} f(ib),$$

$$\alpha(jb) = \Pr\{X = jb | X \ge jb\} = f(jb)/F[jb, \infty), \tag{2.1}$$

$$A(jb) = \sum_{i=0}^{j} \alpha(ib),$$

for $j \ge 0$. α is the *hazard rate*, while A will be called the *cumulative hazard rate*. Note that F and f can be recovered from knowledge of A:

$$\begin{aligned} F(jb) &= 1 - \prod_{i=0}^{j} \{1 - \alpha(ib)\}, \\ f(jb) &= \left[\prod_{i=0}^{j-1} \{1 - \alpha(ib)\}\right] \alpha(jb), \qquad j \ge 0. \end{aligned} \tag{2.2}$$

...this is also Hjort (1990).

Semiparametric models

A semiparametric model is of the form

$$\{P_{\theta,\eta} \colon \theta \in \Theta, \eta \in H\},$$

where $\Theta \subset \mathbb{R}^p$ and H is a function space.

- Partial linear regression $Y = \eta(z) + x^{\mathrm{t}}\theta + \sigma\epsilon$;
- the Cox model $\alpha(t \,|\, x) = \eta(t)\exp(x^{\mathrm{t}}\theta)$;
- partially linear logistic regression

$$\mathrm{pr}(x, z) = 1/\{1 + \exp(-\eta(z) - x^{\mathrm{t}}\theta)\}.$$

- partly parametric Aalen models (McKeague and Sasieni, 1994)

$$\alpha(t \,|\, x, z) = z^{\mathrm{t}}\eta(t) + x^{\mathrm{t}}\theta.$$

  or its Hjort and Stoltenberg (2023) version, and so on.

Throughout this presentation, we seek inference for the parametric part θ, or in Nils jargon, θ is our focus parameter.

Again, start simple

Had Nils been presented with any of these models – before their theory had been worked out, that is – I conjecture that he would have said[1]

. . . did you try a parametric version?

. . . then take limits?, perhaps.

[1] In view of the Beta process paper, other papers, and personal communication.

Parametric partial linear regression

For example, instead of directly attacking

$$Y = \eta(z) + x^{t}\theta + \sigma\epsilon,$$

with θ as our focus parameter and an infinite dimensional nuisance η, one ought first to master (and perhaps even settle for?)

$$Y = \eta_\gamma(z) + \theta x + \sigma\epsilon, \quad \text{for } \gamma \in \mathbb{R}^m\text{, say.}$$

with θ the focus and a finite dimensional nuisance γ_m.

Also, if $\eta_{\gamma_{0,m}}$ is close enough to η_0, inference for θ in the parametric model shouldn't differ that much from inference for θ in the semiparametric one.

This idea leads to that of semiparametric sieves.

.

Semiparametric sieves

If we have data from P_{θ_0,η_0} where

$$P_{\theta_0,\eta_0} \text{ is in } \{P_{\theta,\eta}\colon \theta \in \Theta, \eta \in H\},$$

where $\Theta \subset \mathbb{R}$ and H is a function space, the idea is to instead consider a family of parametric models

$$\{P_{\theta,\eta}\colon \theta \in \Theta, \eta \in H_m\},$$

where H_m is a collection of parametric functions, indexed by the m parameters

$$\gamma_m = (\gamma_1^{(m)}, \ldots, \gamma_m^{(m)}) \in \mathbb{R}^m,$$

where, for any $\eta \in H$, there is a sequence η_{γ_m} such that

$$\eta_{\gamma_m} \to \eta, \quad \text{as } m \text{ tends to infinity.}$$

In other words, $\cup_{m\geq 1} H_m$ is dense in H.

We denote $\gamma_{0,m}$ the sequence such that $\eta_{\gamma_{0,m}} \to \eta_0$, i.e., the limit is the true value in the big model.

Not only parametric modelling

…but why stop at parametric *modelling*? Let's instead go further and pretend that the world is parametric, that is, work under the parametric measure(s) $P_{\theta_0,\gamma_{0,m}}$.

This idea we have from Mykland and Zhang (2009), who studied inference for $\int_0^t \sigma_s^2 \,\mathrm{d}s$ (and other estimands) in continuous time models of the type,

$$\mathrm{d}X_t = \sigma_t \,\mathrm{d}B_t, \quad t \in [0,1],\ X_0 = x_0.$$

by pretending that the data $X_{t_0}, \ldots, X_{t_n}$ were realisations of the discrete time (thus parametric) process

$$\Delta \breve{X}_{t_i} = \sigma_{t_{i-1}} \sqrt{\Delta t_i}\, \mathrm{N}(0,1), \quad \text{for } i = 1, \ldots, n,\ X_0 = x_0,$$

where $\Delta \breve{X}_{t_i} = \breve{X}_{t_i} - \breve{X}_{t_{i-1}}$ and $\Delta t_i = t_i - t_{i-1}$.

The key is contiguity.

Contiguity

Let Q_n and P_n be probability measures on $(\Omega_n, \mathcal{A}_n)$. The sequence Q_n is contiguous w.r.t. the sequence P_n if

$$P_n(A_n) \to 0 \text{ implies } Q_n(A_n) \to 0,$$

for every sequence events A_n. Write $Q_n \triangleleft P_n$.

Le Cam's third lemma: If X_n is a sequence of random variables, and $Q_n \triangleleft P_n$, and[2]

$$(X_n, \frac{\mathrm{d}Q_n}{\mathrm{d}P_n}) \overset{P_n}{\rightsquigarrow} (X, V),$$

then $\mu(B) = \mathrm{E}\, I_B(X)V$ is a probability measure, and $X_n \overset{Q_n}{\rightsquigarrow} \mu$.

[2] If Q_n is not absolutely continuous w.r.t. P_n, the expression $\mathrm{d}Q_n/\mathrm{d}P_n$ should be read as the ratio of $\mathrm{d}Q_n/\mathrm{d}\nu_n$ and $\mathrm{d}P_n/\mathrm{d}\nu_n$ where $\nu_n = (Q_n + P_n)/2$, for example.

Le Cam's third lemma

In particular, if $\widehat{\theta}_n$ is an estimator of $\theta_0 \in \mathbb{R}^p$, and $Q_n \lhd P_n$, and

$$(\sqrt{n}(\widehat{\theta}_n - \theta_0), \log \frac{\mathrm{d}Q_n}{\mathrm{d}P_n}) \overset{P_n}{\rightsquigarrow} N_{p+1}\Big(\begin{pmatrix} 0 \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & b \\ b^{\mathrm{t}} & \sigma^2 \end{pmatrix} \Big),$$

then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \overset{Q_n}{\rightsquigarrow} b + \mathrm{N}_p(0, \Sigma).$$

Parametric building blocks

Given a sample $X_1, \ldots, X_n$ from P_{θ_0,η_0} where

$$P_{\theta_0,\eta_0} \text{ is in } \{P_{\theta,\eta} \colon \theta \in \Theta, \eta \in H\}, \text{ and } H \text{ is infinite dimensional}$$

we pretend that the sample stems from $P_{\theta_0,\eta_{\gamma_{0,m}}}$, where

$$P_{\theta_0,\eta_{\gamma_{0,m}}} \text{ is in } \{P_{\theta,\eta_{\gamma_m}} : \theta \in \Theta, \eta_{\gamma_m} \in H_m\},$$

where H_m is a collection of parametric functions, indexed by m-dimensional parameter vector $\gamma_m = (\gamma_1^{(m)}, \ldots, \gamma_m^{(m)})$.

Let $f_{\theta,\eta_{\gamma_m}}$ be the density of $P_{\theta,\eta_{\gamma_m}}$. Being parametric we proceed as usual and differentiate

$$\dot{\ell}_{\theta_0,\gamma_{0,m}} := \frac{\partial}{\partial\theta} \log f_{\theta,\eta_{\gamma_{0,m}}}\Big|_{\theta=\theta_0}, \quad \& \quad \dot{v}_{\theta_0,\gamma_{0,m}} := \frac{\partial}{\partial\gamma_m} \log f_{\theta_0,\eta_{\gamma_m}}\Big|_{\gamma_m=\gamma_{0,m}},$$

and form the Fisher information matrix

$$J_m = \begin{pmatrix} J_{\theta_0\theta_0} & J_{\theta_0\gamma_{0,m}} \\ J_{\gamma_{0,m}\theta_0} & J_{\gamma_{0,m}\gamma_{0,m}} \end{pmatrix}.$$

The efficient score and information for θ. Fixed m

We can now form the efficient score and efficient information for estimating θ under the mth parametric model $P_{\theta_0,\eta_{\gamma_{0,m}}}$, they are

$$\tilde{\ell}_{\theta_0,\gamma_{0,m}} = \dot{\ell}_{\theta,\gamma_{0,m}} - (J^{-1}_{\gamma_{0,m}\gamma_{0,m}} J_{\gamma_{0,m}\theta_0})^{\mathrm{t}} \dot{v}_{\theta,\gamma_m},$$

and $\tilde{J}_m = J_{\theta_0\theta_0} - J_{\theta_0\gamma_{0,m}} J^{-1}_{\gamma_{0,m}\gamma_{0,m}} J_{\gamma_{0,m}\theta_0}$.

The estimator sequence (in n) $\widehat{\theta}_{m,n}$ is efficient under $P_{\theta_0,\eta_{\gamma_{0,m}}}$, or 'best regular', if and only if,[3]

$$\sqrt{n}(\widehat{\theta}_{m,n} - \theta_0) = \tilde{J}_m^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\gamma_{0,m}}(X_i) + o_{P_m^n}(1),$$

as n tends to infinity.

[3] See, e.g., van der Vaart (1998, p. 369).

A growing parametric profiled LAN theorem

Recall the i.i.d. observations $X_1, \ldots, X_n$, and write

$$P^n = P_{\theta_0,\eta_0} \times \cdots \times P_{\theta_0,\eta_0}, \quad \text{and} \quad P^n_m = P_{\theta_0,\eta_{\gamma_{0,m}}} \times \cdots \times P_{\theta_0,\eta_{\gamma_{0,m}}},$$

for the n-fold product measures. Form the sieved profile likelihood,

$$\mathrm{pl}_{m,n}(\theta) = \sup_{\eta_{\gamma_m} \in H_m} \sum_{i=1}^n \log f_{\theta,\eta_\gamma}(X_i) = \sup_{\gamma_m \in \mathbb{R}^m} \sum_{i=1}^n \log f_{\theta,\eta_{\gamma_m}}(X_i),$$

and a version of one of Nils' favourite processes,

$$A_{m,n}(h) = \mathrm{pl}_{m,n}(\theta_0 + h/\sqrt{n}) - \mathrm{pl}_{m,n}(\theta_0).$$

We prove a growing parametric profiled LAN theorem:[4] Assuming '(1), (2), (3)' (that I will not go into here), and that m_n is a subsequence such that $P^n \triangleleft P^n_{m_n}$,

$$A_{m_n,n}(h) = \frac{h^{\mathrm{t}}}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0,\gamma_{0,m_n}}(X_i) - \tfrac{1}{2} h^{\mathrm{t}} \tilde{J}_{m_n} h + o_{P^n}(1).$$

[4] This is a sieved version of a theorem due to due to Murphy and van der Vaart (2000).

What this theorem does

…it provides conditions (the ones I failed to mention) under which the profile score is only $o_{P^n_{m_n}}(1)$ away from the efficient score, that is

$$\frac{1}{\sqrt{n}} \frac{\mathrm{d}}{\mathrm{d}\theta} \mathrm{pl}_{m_n,n}(\theta)\big|_{\theta=\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\gamma_{0,m_n}} + o_{P^n_{m_n}}(1),$$

Due to the assumed contiguity of $P^n_{m_n}$ with respect to P^n, the $o_{P^n_{m_n}}(1)$ can be replaced by $o_{P^n}(1)$ (this is Le Cam's first lemma).

The nice thing about going parametric here, is that the model with $\tilde{\ell}_{\theta_0,\gamma_{0,m}}$ as its score[5] always takes the form

$$P_{\theta,\gamma_m(\theta)}, \quad \text{with} \quad \gamma_m(\theta) = \gamma_{0,m} + J^{-1}_{\gamma_{0,m}\gamma_{0,m}} J_{\gamma_{0,m}\theta_0}(\theta_0 - \theta),$$

so you don't have to be clever about finding it (which you do have to be in the semiparametric world).

[5] i.e., the least favourable submodel.

Semiparametric efficiency, $m_n \to \infty$

Let $\widehat{\theta}_{m,n}$ be the maximiser of $\mathrm{pl}_{m,n}(\theta)$, i.e., the maximum likelihood estimator under the mth parametric model.

We show that under the same assumptions invoked above and also assuming consistency of $\widehat{\theta}_{m_n,n}$ for θ_0 under $P^n_{m_n}$, …

…or, via a concavity argument à la Hjort and Pollard (1993),

$$\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0) = \tilde{J}_{m_n}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\gamma_{0,m_n}}(X_i) + o_{P^n_{m_n}}(1),$$

where $\tilde{\ell}_{\theta_0,\gamma_{0,m_n}}$ is a sequence of efficient scores in growing parametric models, and $\tilde{J}_{m_n} = \mathrm{E}_{\theta_0,\eta_{\gamma_{0,m_n}}} \tilde{\ell}_{\theta_0,\gamma_{0,m_n}} \tilde{\ell}^{\mathrm{t}}_{\theta_0,\gamma_{0,m_n}}$.

A theorem and a lemma

With the efficient score $\tilde{\ell}_{\theta_0,\gamma_{0,m}}$ and efficient information $\tilde{J}_m^{-1}$ we form the efficienct influence function for estimating θ under $P_{\theta_0,\eta_{\gamma_{0,m}}}$

$$\tilde{\psi}_m = \tilde{J}_m^{-1}\tilde{\ell}_{\theta_0,\gamma_{0,m}}.$$

Let $\tilde{\psi}$ be the efficient influence function for estimating θ under the semiparametric model P_{θ_0,η_0}.

Let $\theta \in \mathbb{R}$ for simplicity.

Theorem: If $\mathrm{E}\,(\tilde{\psi}_{m_n} - \tilde{\psi})^2 \to 0$, then $\widehat{\theta}_{m_n,n}$ is efficient for θ under P_{θ_0,η_0}.

Lemma: The sieve construction, i.e., $\cup_{m\geq 1} H_m$ being dense in H, ensures the convergence in the theorem, provided

$$\mathrm{E}\,(\dot{\ell}_{\theta_0,\gamma_{0,m}} - \dot{\ell}_{\theta_0,\eta_0})^2 \to 0.$$

... from which we conclude that

$$A_{m_n,n} = \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\gamma_{0,m_n}} - \tfrac{1}{2} h^2 \tilde{J}_{m_n} + o_{P^n_{m_n}}(1)$$
$$= \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0} - \tfrac{1}{2} h^2 \tilde{J} + o_{P^n_{m_n}}(1),$$

which, combined with

- consistency of $\widehat{\theta}_{m_n,n}$ for θ_0 under $P^n_{m_n}$;
- or, concavity of $\mathrm{pl}_{m,n}(\theta)$ and Hjort and Pollard (1993),

yields,

$$\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0) \overset{P^n}{\rightsquigarrow} \mathrm{N}(0, \tilde{J}^{-1}),$$

provided m_n is chosen so that $\mathrm{d}P_n/\mathrm{d}P^n_{m_n} \to 1$ in $P^n_{m_n}$-probability; where $\tilde{J}$ is the efficient information under the big semiparametric model P_{θ_0,η_0}.

... and $\tilde{J}$ is the limit of $\tilde{J}_m$.

An test case: The partial linear model

Let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ be independent replicates of (X, Y, Z), where the covariates X and Z take their values in $[0, 1]$; have a joint density; and $Z \sim F_Z$, with $F_Z' = f_Z$ a continuous density, bounded below.

The big semiparametric model is

$$P_{\theta_0,\eta_0}: \quad Y = \eta_0(Z) + \theta_0 X + \sigma\epsilon,$$

for $\epsilon \sim \mathrm{N}(0, 1)$, with (θ_0, η_0) denoting the true parameter value, and η_0 assumed continuously differentiable. Consider the smaller parametric approximations (the sieves)

$$P_{\theta_0,\eta_{\gamma_{0,m}}}: \quad Y = \eta_{\gamma_{0,m}}(Z) + \theta_0 X + \sigma\epsilon',$$

with $\epsilon' \sim \epsilon \sim \mathrm{N}(0, 1)$, and $\eta_{\gamma_{0,m}} = \sum_{j=1}^{m} \gamma_{0,m} I_{W_{m,j}}(z)$.

Parametric inference, fixed m

Let's first pretend that $(X_1, Y_1, Z_1), \ldots, (X_1, Y_1, Z_1)$ are i.i.d. from the parametric model, $P_m^n = P_{\theta_0, \gamma_{0,m}} \times \cdots \times P_{\theta_0, \gamma_{0,m}}$ for some fixed m.

Estimating θ_0 is then a least squares problem, and with $\widehat{\theta}_{m,n}$ the least squares estimator

$$\sqrt{n}(\widehat{\theta}_{m,n} - \theta_0) \overset{P_m^n}{\rightsquigarrow} \mathrm{N}(0, J_m^{-1}),$$

as $n \to \infty$, where J_m the sum ($\approx$ a Riemann–Stieltjes sum)

$$J_m = \frac{1}{\sigma^2} \sum_{j=1}^{m} \mathrm{Var}(X \mid Z \in W_{m,j}) \{F_Z(j\Delta_m) - F_Z((j-1)\Delta_m)\},$$

and F_Z is the distribution function of Z (covariate distributions are the same under all models).

From parametric likelihood theory we know that $\widehat{\theta}_{m,n}$ is efficient under P_m. End of parametric story.

Semiparametric inference, $m \to \infty$ with n

We get a semiparametric problem when we let the models grow, i.e., when m tends to infinity with the sample size n.

The profile likelihood takes the form

$$\mathrm{pl}_{m,n}(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m} \{(Y_i - \bar{Y}_{m,j}) - \theta(X_i - \bar{X}_{m,j})\}^2 I_{W_{m,j}}(Z_i).$$

where $\bar{X}_{m,j} = \sum_{i=1}^{n} X_i I_{W_{m,j}}(Z_i) / \sum_{i=1}^{n} X_i I_{W_{m,j}}(Z_i)$, and $\bar{Y}_{m,j}$ similarly defined. The profile score evaluated in θ_0 is then

$$\frac{1}{\sqrt{n}} \frac{\mathrm{d}}{\mathrm{d}\theta} \mathrm{pl}_{m,n}(\theta)\big|_{\theta=\theta_0} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{m} (X_i - \bar{X}_{m,j}) I_{W_{m,j}}(Z_i) \epsilon_i,$$

and provided $n\Delta_{m_n} \to \infty$ as $n \to \infty$ and $\Delta_{m_n} \to 0$,

$$\frac{1}{\sqrt{n}} \frac{\mathrm{d}}{\mathrm{d}\theta} \mathrm{pl}_{m_n,n}(\theta)\big|_{\theta=\theta_0} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{m_n} (X_i - \mu_{m_n,j}) I_{W_{m_n,j}}(Z_i) \epsilon_i + o_{P^n_{m_n}}(1),$$

where $\mu_{m,j} = \mathrm{E}\,(X \mid Z \in W_{m,j})$.

. . . which is close the the efficient score

Recall that the least favourable submodel alwyas takes the form takes the form $P_{\theta,\gamma_m(\theta)}$ with $\gamma_m(\theta) = \gamma_{0,m} + J^{-1}_{\gamma_{0,m}\gamma_{0,m}} J_{\gamma_{0,m}\theta_0}(\theta_0 - \theta)$.

The efficient score under the m parametric model is therefore

$$\begin{aligned}\tilde{\ell}_{\theta_0,\gamma_{0,m}} &= \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_{\theta,\gamma_m(\theta)}\big|_{\theta=\theta_0} \\ &= \sigma^{-1} \sum_{j=1}^{m} (X - \{J^{-1}_{\gamma_{0,m}\gamma_{0,m}} J_{\gamma_{0,m}\theta_0}\}_j) I_{W_{m,j}}(Z)\epsilon'.\end{aligned}$$

and doing the multiplication $J^{-1}_{\gamma_{0,m}\gamma_{0,m}} J_{\gamma_{0,m}\theta_0} = \mu_{m,j}$.

Can check directly check that $\mathrm{E}\,(\tilde{\psi}_{m_n} - \tilde{\psi})^2 \to 0$, because the efficient score for θ under the semiparametric model $P_{\theta,\eta}$ is

$$\tilde{\ell}_{\theta_0,\eta_0} = \sigma^{-1}(X - \mathrm{E}\,(X \mid Z))\epsilon',$$

and we see that

$$\mathrm{E}\,(\tilde{\ell}_{\theta,\gamma_{0,m}}(X,Y,Z) - \tilde{\ell}_{\theta_0,\eta_0}(X,Y,Z))^2 \to 0,$$

as $m \to \infty$.

Switching back to P_{θ_0,η_0}

From the above we get that

$$A_{m_n,n} = \frac{h}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathrm{E}\,(X \mid Z_i))\epsilon_i - \tfrac{1}{2} h^2 \tilde{J} + o_{P^n_{m_n}}(1),$$

where $\tilde{J} = \sigma^{-2}\mathrm{E}\,\mathrm{Var}(X \mid Z)$. Here, since $\mathrm{pl}_{m,n}(\theta)$ is indeed concave,

$$\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0) = \tilde{J}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathrm{E}\,(X \mid Z_i))\epsilon_i + o_{P^n_{m_n}}(1).$$

Using the assumption that η_0 is continuously differentiable,

$$\frac{\mathrm{d}P^n}{\mathrm{d}P^n_{m_n}} \overset{P^n_{m_n}}{\rightsquigarrow} 1, \quad \text{(so in probability)}$$

provided $\sqrt{n}\Delta_{m_n} \to 0$. Le Cam's third lemma then allows us to switch back to the semiparametric world, and

$$\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0) \overset{P^n}{\rightsquigarrow} \mathrm{N}(0, \tilde{J}^{-1}),$$

as $n \to \infty$. Conclude that $\widehat{\theta}_{m_n,n}$ is efficient for θ under the semiparametric model P_{θ_0,η_0}.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120.

Donald, S. G. and Newey, W. K. (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis*, 50:30–40.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10:401–414.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *The Annals of Statistics*, 18:1259–1294.

Hjort, N. L. and Pollard, D. B. (1993). Asymptotics for minimisers of convex processes. Technical report, Department of Mathematics, University of Oslo.

Hjort, N. L. and Stoltenberg, E. A. (2023). The partly parametric and partly nonparametric additive risk model. *Lifetime Data Analysis*, 29:372–402.

Karr, A. F. (1987). Maximum likelihood estimation in the multiplicative intensity model via sieves. *The Annals of Statistics*, 15:473–490.

Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics*, 25:1014–1035.

McKeague, I. W. (1986). Estimation for a semimartingale regression model using the method of sieves. *The Annals of Statistics*, 14:579–589.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81:501–514.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95:449–465.

Mykland, P. A. and Zhang, L. (2009). Inference for continuous semimartingales observed at high frequency. *Econometrica*, 77:1403–1445.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Same story for the Cox model (I think)

Survival data (T, δ, X) observed over $[0,1]$. The mth parametric model $P_{\theta,\gamma_{0,m}}$ is one in which the baseline hazard is locally constant, as above.

With standard notation and assumptions (Andersen and Gill, 1982), the profile score for the mth model, evaluted in the true parameter value, θ_0, is

$$\frac{1}{\sqrt{n}}\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{pl}_{m,n}(\theta)|_{\theta=\theta_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{m}\{X_i - \frac{\int_{W_{m,j}} S_n^{(1)}(s,\theta)\,\mathrm{d}s}{\int_{W_{m,j}} S_n^{(0)}(s,\theta)\,\mathrm{d}s}\}\int_{W_{m,j}} \mathrm{d}M_{i,t}^{(m)},$$

under $P_{\theta_0,\gamma_{0,m}}$, which is $o_{P^n_{m_n}}(1)$ away from the efficient score (found via the parametric least favourable submodel approach)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\ell}_{\theta_0,\gamma_{0,m}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{m}\{X_i - \frac{\int_{W_{m,j}} s_m^{(1)}(s)}{\int_{W_{m,j}} s_m^{(0)}(s)}\}\int_{W_{m,j}} \mathrm{d}M_{i,t}^{(m)},$$

where $s_m^{(k)}(t) = \mathrm{E}_{\theta_0,\gamma_{0,m}}\, Y(t)X^k\exp(\theta_0 X)$.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\ell}_{\theta_0,\gamma_{0,m_n}},$$

is $o_{P^n_{m_n}}(1)$ away from the discrete time martingale

$$Z_{m_n,n} := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{m_n}\{X_i - \frac{s^{(1)}_{m_n}((j-1)\Delta_{m_n})}{s^{(0)}_{m}((j-1)\Delta_{m_n})}\}\{M^{(m_n)}_{i,j\Delta_{m_n}} - M^{(m_n)}_{i,(j-1)\Delta_{m_n}}\},$$

whose variance process

$$\langle Z_{m_n,n}, Z_{m_n,n}\rangle = \int_0^1 (\frac{s^{(2)}_{m_n}(t)}{s^{(0)}_{m_n}(t)} - \frac{s^{(1)}_{m_n}(t)^2}{s^{(0)}_{m_n}(t)^2}) s^{(0)}_{m_n}(t)\eta_{\gamma_{0,m}}(t)\,\mathrm{d}t + o_{P^n_{m_n}}(\Delta_{m_n}).$$

as $n \to \infty$ and $\Delta_{m_n} \to 0$.

for $t \in (0,1]$,

$$\log \frac{\mathrm{d}P^n_{\theta_0,\eta_0}}{\mathrm{d}P^n_{\theta_0,\gamma_{0,m_n}}}\Big|_{\mathcal{F}_t} = \sum_{i=1}^n \{\xi_i^{(m_n)}(t) - \tfrac{1}{2}\langle \xi_i^{(m_n)}, \xi_i^{(m_n)}\rangle_t\} + o_{P^n_{m_n}}(1),$$

where

$$\xi_i^{(m_n)}(t) = -\frac{1}{\sqrt{n}} \int_0^t \frac{h_{m_n,n}(s)}{\eta_0(s)}\, \mathrm{d}M_i^{(m_n)}(s),$$

where $h_{m,n}(s) = \sqrt{n}(\eta_{\gamma_{0,m}}(s) - \eta_0(s))$, so with η_0 continuously differentiable, as above,

$$\frac{\mathrm{d}P^n_{\theta_0,\eta_0}}{\mathrm{d}P^n_{\theta_0,\gamma_{0,m_n}}} \overset{P^n_{\theta_0,\gamma_{0,m_n}}}{\rightsquigarrow} 1,$$

provided $\sqrt{n}\Delta_{m_n} \to 0$.

. . . to be continued