Asymptotic normality in a mixture of semi-parametric models using statistical generalised derivative

Yuichi Hirose & Ivy Liu

School of Mathematics and Statistics,
Victoria University of Wellington,
New Zealand

December 4, 2017

I dedicate this talk to
Alastair Scott (1939-2017)

Mixture of semiparametric models (1)

We consider a mixture of semiparametric models whose density is of the form

$$p(x;\theta,\eta,\pi)=\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\eta_k), \tag{1}$$

where for each $k=1,\ldots,K$,

$$p_k(x;\theta_k,\eta_k)$$

is a semiparametric model with

- finite dimensional parameter θ_k and
- infinite dimensinal parameter η_k, and
- $\pi_1,\ldots,\pi_K$ are mixture probbailities.

We assume that $\pi_k>0$ for each k and $\sum_{k=1}^{K}\pi_k=1$. We denote $\theta=(\theta_1,\ldots,\theta_K)$, $\eta=(\eta_1,\ldots,\eta_K)$ and $\pi=(\pi_1,\ldots,\pi_K)$.

Mixture of semiparametric models (2)

Once we observe iid data $X_1, \ldots, X_n$ from the mixture model, the joint probability function of the data $\mathbf{X} = (X_1, \ldots, X_n)$ is given by

$$p(\mathbf{X}; \theta, \eta, \pi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k p_k(X_i; \theta_k, \eta_k). \tag{2}$$

We consider θ is the parameters of interest, and η and π are nuisance parameters.

- We aim to establish large sample properties of the parameter θ using EM-algorithm and profile likelihood approach.

E-M Algorithm (1)

For the EM-algorithm, we further introduce notations.
Let

$$Z_i = (Z_{i1}, \ldots, Z_{iK})$$

be group indocator variable for the subject i:

- for each k, $Z_{ik} = 0$ or $= 1$ with $P(Z_{ik} = 1) = \pi_k$, and
- $\sum_{k=1}^{K} Z_{ik} = 1$.

Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$. The joint probability function of the complete data $(\mathbf{X}, \mathbf{Z})$ is

$$p(\mathbf{X}, \mathbf{Z}; \theta, \eta, \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k p_k(X_i; \theta_k, \eta_k)]^{Z_{ik}}. \tag{3}$$

E-M Algorithm (2)

Then the EM-algorithm utilize the identity

$$\begin{aligned} \log p(\mathbf{X}; \theta, \eta, \pi) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}; \theta, \eta, \pi) \\ &\quad - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}; \theta, \eta, \pi), \end{aligned} \tag{4}$$

where $q(\mathbf{Z})$ is any distribution of $\mathbf{Z}$.

E-M Algorithm (3)

- In the E-step put

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}; \theta^{old}, \eta^{old}, \pi^{old}).$$

- In the M-step, maximise the expectation of complete data log likelihood function

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}; \theta, \eta, \pi)$$

  to obtain $(\theta^{new}, \eta^{new}, \pi^{new})$.
- Then repeat E-step and M-step iteratively until we achieve the maximum.

E-M Algorithm (4)

- Under this procedure, the maximizer of the mixture log-likelihood function

$$\log p(\mathbf{X}; \theta, \eta, \pi)$$

with respect to θ, η and π is the same as the ones for the expectation of complete data log likelihood function

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}; \theta, \eta, \pi).$$

- The EM-algorithm gives us value of the maximum likelihood estimator $\hat{\theta}$ of the mixture model. However it does not give us the variance of the estimator.

E-M Algorithm (5)

From the complete data joint distribution (3), we can derive the conditional distribution $p(\mathbf{Z}|\mathbf{X};\theta,\eta,\pi)$:

$$\begin{aligned} p(\mathbf{Z}|\mathbf{X};\theta,\eta,\pi) &= \frac{p(\mathbf{X},\mathbf{Z};\theta,\eta,\pi)}{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z};\theta,\eta,\pi)} \\ &= \prod_{i=1}^{n}\prod_{k=1}^{K}\frac{[\pi_k p_k(X_i;\theta_k,\eta_k)]^{Z_{ik}}}{\sum_{j=1}^{K}\pi_j p_j(X_i;\theta_j,\eta_j)} \\ &= \prod_{i=1}^{n}\prod_{k=1}^{K}\gamma_k(X_i;\theta,\eta)^{Z_{ik}}. \end{aligned} \tag{5}$$

where

$$\gamma_k(X_i;\theta,\eta) = \frac{\pi_k p_k(X_i;\theta_k,\eta_k)}{\sum_{j=1}^{K}\pi_j p_j(X_i;\theta_j,\eta_j)}, \quad k = 1,\ldots,K. \tag{6}$$

E-M Algorithm (6)

Again from (3), the expected complete data log-likelihood under $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}; \theta, \eta, \pi)$ is

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\theta, \eta, \pi)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_k(X_i; \theta, \eta)[\log \pi_k + \log p_k(X_i; \theta_k, \eta_k)].$$

E-M Algorithm (7)

With the expected complete data log-likelihood (7), the method of Lagrange multiplier can be applied to get the MLE $\hat{\pi}_k$ of π_k:

$$\hat{\pi}_k(\theta, \eta) = \frac{\sum_{i=1}^{n} \gamma_k(X_i; \theta, \eta)}{n}, \quad k = 1, \ldots, K. \tag{7}$$

We require that, as $n \to \infty$,

$$\hat{\pi}_k(\theta_0, \eta_0) \xrightarrow{P} \pi_{0k}$$

where (θ_0, η_0) are the true value of (θ, η) and π_{0k}, $k = 1, \ldots, K$, are the true mixture probability.

Score functions

The score function for θ and score operator for η in the mixture model are, respectively,

$$\begin{aligned}\dot{\ell}(x;\theta,\eta) &= \frac{\partial}{\partial\theta}\log\left(\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\eta_k)\right)\\ &= \sum_{k=1}^{K}\gamma_k(x;\theta,\eta)\frac{\partial}{\partial\theta}\log p_k(x;\theta_k,\eta_k),\end{aligned}$$

and

$$\begin{aligned}B(x;\theta,\eta) &= d_\eta\log\left(\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\eta_k)\right)\\ &= \sum_{k=1}^{K}\gamma_k(x;\theta,\eta)d_\eta\log p_k(x;\theta_k,\eta_k).\end{aligned}$$

The notation d_η is the Hadamard derivative operator with respect to the parameter η.

Efficient score function and efficient information

Let θ_0, η_0 be the true values of θ, η and denote $\dot{\ell}_0(x) = \dot{\ell}(x; \theta_0, \eta_0)$ and $B_0(x) = B(x; \theta_0, \eta_0)$. The efficient score function $\tilde{\ell}_0$ and the efficient information matrix $\tilde{I}_0$ in the semiparametric mixture model are given by

$$\tilde{\ell}_0(x) = (I - B_0(B_0^* B_0)^{-1} B_0^*)\dot{\ell}_0(x), \tag{8}$$

and

$$\tilde{I}_0 = E[\tilde{\ell}_0 \tilde{\ell}_0^T]. \tag{9}$$

Note: The score functions in the semiparametric mixture model coincide with the ones for the expected complete data likelihood.

Score function for Profile likelihood

In the estimation of (θ, η) we use the profile likeihood aproach: we obtain a function $(\theta, F) \to \hat{\eta}_{\theta,F} = (\hat{\eta}_{1,\theta,F}, \ldots, \hat{\eta}_{K,\theta,F})$ whose values are in the space of the parameter $\eta = (\eta_1, \ldots, \eta_K)$.
Define the score functions for the profile likelihood in the model

$$\begin{aligned}
\phi(x; \theta, F) &= \frac{\partial}{\partial \theta} \log \left(\sum_{k=1}^{K} \pi_k p_k(x; \theta_k, \hat{\eta}_{k,\theta,F}) \right) \\
&= \sum_{k=1}^{K} \gamma_k(x; \theta, \hat{\eta}_{\theta,F}) \frac{\partial}{\partial \theta} \log p_k(x; \theta_k, \hat{\eta}_{k,\theta,F})
\end{aligned}$$

and

$$\begin{aligned}
\psi(x; \theta, F) &= d_F \log \left(\sum_{k=1}^{K} \pi_k p_k(x; \theta_k, \hat{\eta}_{k,\theta,F}) \right) \\
&= \sum_{k=1}^{K} \gamma_k(x; \theta, \hat{\eta}_{\theta,F}) d_F \log p_k(x; \theta_k, \hat{\eta}_{k,\theta,F}),
\end{aligned}$$

Example: Joint mixture model (1)

- Let Y_{ijm} be the ordered categorical response from 1 (poor) to L (excellent) on item (or question) j for subject i at the m^{th} protocol-specified time point.
- In total, there are J items in the questionnaire, collected at times $t_1, t_2, \ldots, t_M$.
- Given that subject i belongs to group r, a stereotype model can be written as

$$\log\left[\frac{P(Y_{ijm}=\ell \mid \theta_r)}{P(Y_{ijm}=1 \mid \theta_r)}\right] = a_\ell + \phi_\ell(b_j+\theta_r), \quad r=1,\ldots,R. \tag{10}$$

Example: Joint mixture model(2)

The ordinal response part of likelihood function for the ith subject is

$$P(\boldsymbol{Y}_i \mid \theta_r, \boldsymbol{\alpha}) = \prod_{m=1}^{M_i} \prod_{j=1}^{J} \prod_{\ell=1}^{L} \left(\frac{\exp(a_\ell + \phi_\ell(b_j + \theta_r))}{1 + \sum_{k=2}^{L} \exp(a_k + \phi_k(b_j + \theta_r))} \right)^{y_{ijm\ell}}$$

where $\boldsymbol{\alpha} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\phi})$.

Example: Joint mixture model(3)

We consider the Cox proportional hazards model for the survival part in the joint model. Let X be a time-independent covariate. The hazard function for the failure time T_i of the i^{th} subject is of the form

$$\lambda(t|X_i, \theta_r, \boldsymbol{\delta}) \quad = \quad \lambda_0(t) \exp(\theta_r \delta_0 + X_i \delta_1)$$

where $\lambda_0(t)$ is the baseline hazard function. The latent variable θ_r is linked with the ordinal response model and $\boldsymbol{\delta} = (\delta_0, \delta_1)$ are coefficients.

Example: Joint mixture model(4)

Assume that the hazard is zero between adjacent times so that the survival time is discrete. Let λ_i be the hazard at time t_i, where $t_1 < t_2 < \ldots < t_n$ are the ordered observed times.
The cumulative hazard function $\Lambda_0(t_i) = \sum_{p \le i} \lambda_p$ is a step function with jumps at the failure time t_i. Then the the survival part likelihood function of subject i is

$$\begin{aligned} P\big(T_i, d_i \,|\, \boldsymbol{\lambda}, \theta_r, \boldsymbol{\delta}\big) &= \big(\lambda_i \exp(\theta_r \delta_0 + X_i \delta_1)\big)^{d_i} \\ &\quad \times \exp\Big(-\sum_{p \le i} \lambda_p \exp(\theta_r \delta_0 + X_i \delta_1)\Big). \end{aligned}$$

The d_i is an indicator of censorship for individual i: if we observe failure time, then $d_i = 1$, otherwise $d_i = 0$.

Example: Joint mixture model(5)

Let π_r be the unknown probability $(r = 1, \ldots, R)$ that a subject lies in group r, and $\boldsymbol{\Theta}$ be all the unknown parameters of the joint model. The mixture model likelihood function is

$$L(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{D}) = \prod_{i=1}^{n} \left(\sum_{r=1}^{R} P(\boldsymbol{Y}_i \,|\, \theta_r, \boldsymbol{\alpha}) P(T_i, d_i \,|\, \boldsymbol{\lambda}, \theta_r, \boldsymbol{\delta}) \pi_r \right). \tag{11}$$

Example: Joint mixture model(6)

The expected complete data log likelihood under $q(\mathbf{Z}) = P(\mathbf{Z}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{d})$ is

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log L(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{d}, \boldsymbol{Z})$$

$$= \sum_{i=1}^{n}\sum_{r=1}^{R} \gamma(Z_{ir}) \log \pi_r$$

$$+ \sum_{i=1}^{n}\sum_{r=1}^{R} \gamma(Z_{ir}) \left\{ \log P\left(\boldsymbol{Y}_i \,|\, \theta_r, \boldsymbol{\alpha}\right) + \log P\left(T_i, d_i \,|\, \boldsymbol{\lambda}, \theta_r, \boldsymbol{\delta}\right) \right\}$$

Example: Joint mixture model (7)

Before starting the EM-step, we profile out the baseline hazard function $\lambda_0(t)$. The survival part of likelihood function can be separately maximized with respect to $\boldsymbol{\lambda}$.
We find the maximizer $\widehat{\lambda}_i$ by holding $(\boldsymbol{\theta}, \boldsymbol{\delta})$ fixed, and it is given by

$$\widehat{\lambda}_i(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{d_i}{\sum_{p \geq i} \sum_{r=1}^{R} \gamma(Z_{pr}) \exp(\theta_r \delta_0 + X_p \delta_1)}. \tag{12}$$

Denote $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{\delta}) = (\widehat{\lambda}_i(\boldsymbol{\theta}, \boldsymbol{\delta}), \ldots, \widehat{\lambda}_n(\boldsymbol{\theta}, \boldsymbol{\delta}))$.

Example: Joint mixture model (8)

The E-step: In the E-step, we use the current parameter estimates $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ to find the expected values of Z_{ir} of the complete data log likelihood:

$$\begin{aligned}
\gamma(Z_{ir}) &= E\big(Z_{ir} | \, \boldsymbol{Y}_i, T_i, d_i\big) \\
&= \frac{\pi_r \, P\big(\boldsymbol{Y}_i \, | \, \theta_r, \boldsymbol{\alpha}\big) P\big(T_i, d_i \, | \, \widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{\delta}), \theta_r, \boldsymbol{\delta}\big)}{\sum_{g=1}^{R} \pi_g \, P\big(\boldsymbol{Y}_i \, | \, \theta_g, \boldsymbol{\alpha}\big) P\big(T_i, d_i \, | \, \widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{\delta}), \theta_g, \boldsymbol{\delta}\big)}.
\end{aligned}$$

Example: Joint mixture model (9)

The M-step: In the M-step, we maximize equation (12) with respect to π_r and $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\delta})$. Due to the fact that there is no relationship between π_r and $\boldsymbol{\Theta}$, they can be estimated separately.

1. Calculate the estimates of π_r

$$\widehat{\pi}_r = \frac{\sum_{i=1}^{n} \gamma(Z_{ir})}{n}.$$

2. We maximize the second and third parts of equation (12) (with $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{\delta})$ in the place of $\boldsymbol{\lambda}$)

$$\sum_{i=1}^{n} \sum_{r=1}^{R} \gamma(Z_{ir}) \left\{ \log P\left(\boldsymbol{Y}_i \,|\, \theta_r, \boldsymbol{\alpha}\right) + \log P\left(T_i, d_i \,|\, \widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{\delta}), \theta_r, \boldsymbol{\delta}\right) \right\} \quad (13)$$

with respect to $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ to obtain $\widehat{\boldsymbol{\Theta}}$.

The estimated parameters from the M-step are returned into the E-step until convergence.

Theoretical challenge

" $\cdots$ the **theoretical challenge** regarding efficiency and the asymptotic distribution of the parametric estimators in the joint model frame work.
No distribution or asymptotic theory is available to date, and even the standard error (SE) are difficult to obtain. "
Hsieh, Tseng and Wang (2006)

Example continued.

An estimator of the cumulative hazard fucntion in the counting process notation is

$$\widehat{\Lambda}(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) \sum_{r=1}^R \gamma(Z_{ir}) \exp(\theta_r \delta_0 + X_i \delta_1)}$$

where $N_i(u) = 1_{\{T_i \leq u, d_i = 1\}}$ and $Y_i(u) = 1_{\{T_i \geq u\}}$.
Let us denote $E_{F_n} f = \int f dF_n$. Then the above $\widehat{\Lambda}(t)$ can be writte as

$$\widehat{\Lambda}(t; \Theta, F_n) = \int_0^t \frac{E_{F_n} dN(u)}{E_{F_n} Y(u) \sum_{r=1}^R \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)} \qquad (14)$$

where $N(u) = 1_{\{T \leq u, d = 1\}}$, $Y(u) = 1_{\{T \geq u\}}$ and similarly $\gamma(Z_r)$ is defined.

Differentiability of implicitly defined function (1)

Hirose (2016) considered the function $\hat{\eta}_{\theta,F}$ is given as the solution to the operator equation of the form

$$\eta = \Psi_{\theta,F}(\eta). \tag{15}$$

Differentiability of implicitly defined function (2)

Suppose the function $\Psi_{\theta,F}(\eta)$ is

(A1) two times continuously differentiable with respect to θ and two times Hadamard differentiable with respect to η and Hadamard differentiable with respect to F.

(A2) the true values (θ_0, η_0, F_0) satisfy $\eta_0 = \Psi_{\theta_0,F_0}(\eta_0)$.

(A3) the linear operator $d_\eta \Psi_{\theta_0,F_0}(\eta_0) : \mathcal{B} \to \mathcal{B}$ has the operator norm $\|d_\eta \Psi_{\theta_0,F_0}(\eta_0)\| < 1$.

Differentiability of implicitly defined function (3)

Then the solution $\eta_{\theta,F}$ to the equation

$$\eta = \Psi_{\theta,F}(\eta) \tag{16}$$

exists in an neighborhood of (θ_0, F_0) and it is two times continuously differentiable with respect to θ and Hadamard differentiable with respect to F in the neighborhood.

Differentiability of implicitly defined function (4)

The derivatives are given by

$$\dot{\eta}_{\theta,F} = [I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})]^{-1} \dot{\psi}_{\theta,F}(\eta_{\theta,F}), \tag{17}$$

$$\begin{aligned} \ddot{\eta}_{\theta,F} &= [I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})]^{-1} \Big[\ddot{\psi}_{\theta,F}(\eta_{\theta,F}) + d_\eta \dot{\psi}_{\theta,F}(\eta_{\theta,F}) \dot{\eta}_{\theta,F}^T \\ &\quad + d_\eta \dot{\psi}_{\theta,F}^T(\eta_{\theta,F}) \dot{\eta}_{\theta,F} + d_\eta^2 \Psi_{\theta,F}(\eta_{\theta,F}) \dot{\eta}_{\theta,F} \dot{\eta}_{\theta,F}^T \Big], \end{aligned} \tag{18}$$

and

$$d_F \eta_{\theta,F} = [I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})]^{-1} d_F \Psi_{\theta,F}(\eta_{\theta,F}). \tag{19}$$

Solution to the challenge

If I can work only with the first derivates (score function)

$$\begin{aligned}\phi(x;\theta,F) &= \frac{\partial}{\partial\theta}\log\left(\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})\right)\\ &= \sum_{k=1}^{K}\gamma_k(x;\theta,\hat{\eta}_{\theta,F})\frac{\partial}{\partial\theta}\log p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})\end{aligned}$$

and not dealing with the second derivatives.

$$\begin{aligned}&\frac{\partial^2}{\partial\theta\partial\theta^T}\log\left(\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})\right)\\ =\ &\frac{\partial}{\partial\theta^T}\left(\sum_{k=1}^{K}\gamma_k(x;\theta,\hat{\eta}_{\theta,F})\frac{\partial}{\partial\theta}\log p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})\right),\end{aligned}$$

the problem is so simple!

I introduce the **statistical generalised derivative**.

Derivative of generalised function

To calculate the second derivative of the score function $\phi(x;\theta,F)$, we use the idea similar to the derivative of generalised function. Let

$$\varphi \to (f,\varphi) = \int_{-\infty}^{\infty} f(x)\varphi(x)dx$$

be a generalised function, where φ vanishes outside of some interval. Then if f and φ are differentiable with derivative f' and φ', then by integration by parts,

$$(f',\varphi) = \int_{-\infty}^{\infty} f'(x)\varphi(x)dx = -\int_{-\infty}^{\infty} f(x)\varphi'(x)dx = -(f,\varphi').$$

We define the derivative (f',φ) of the generalized function $\varphi \to (f,\varphi)$ by $-(f,\varphi')$. This definition is valid even if f is not differentiable, provided φ is differentiable.

Statistical generalised derivative (Motivation)

Suppose the density for the profile likelihood $p(x;\theta,F)$ is twice differentiable with respect to θ, then by diferentiating the identity

$$\int \left\{ \frac{\partial}{\partial \theta} \log p(x;\theta,F) \right\} p(x;\theta,F) dx = 0,$$

with respect to θ at $(\theta,F) = (\theta_0,F_0)$, we get equivalent expressons for the efficient information matrix in terms of the score function $\phi(x;\theta_0,F_0)$:

$$\tilde{I}_0 = E[\phi\phi^T(X;\theta_0,F_0)] = -E\left[\frac{\partial}{\partial \theta^T}\phi(X;\theta_0,F_0)\right]. \qquad (20)$$

From this equation we are motivated to define the expected derivative of the score function $-E\left[\frac{\partial}{\partial \theta^T}\phi(X;\theta_0,F_0)\right]$ by $E[\phi\phi^T(X;\theta_0,F_0)]$.

Assumptions (1)

On the set of cdf functions $\mathcal{F}$, we use the sup-norm, i.e. for $F, F_0 \in \mathcal{F}$,

$$\|F - F_0\| = \sup_x |F(x) - F_0(x)|.$$

For $\rho > 0$, let

$$\mathcal{C}_\rho = \{F \in \mathcal{F} : \|F - F_0\| < \rho\}.$$

We assume that:

(R1) For each $(\theta, F) \in \Theta \times \mathcal{F}$, the log-profile-likelihood function for an observation x

$$\log p(x; \theta, F) = \log \left(\sum_{k=1}^{K} \pi_k p_k(x; \theta_k, \hat{\eta}_{k,\theta,F}) \right) \tag{21}$$

is continuously differentiable with respect to θ and Hadamard differentiable with respect to F for all x. Derivatives are respectively denoted by $\phi(x; \theta, F) = \frac{\partial}{\partial \theta} \log p(x; \theta, F)$ and $\psi(x; \theta, F) = d_F \log p(x; \theta, F)$.

Assumptions (2)

We assume that:

(R2) The 4th-root-n-consistency of F_n, $n^{1/4}(F_n - F_0) = O_P(1)$, and $\hat{\eta}_{\theta,F}$ satisfies $\hat{\eta}_{\theta_0,F_0} = \eta_0$ and the function

$$\tilde{\ell}_0(x) := \phi(x; \theta_0, F_0)$$

is the efficient score function.

(R3) The efficient information matrix $\tilde{I}_0 = E[\tilde{\ell}_0 \tilde{\ell}_0^T] = E[\phi\phi^T(X; \theta_0, F_0)]$ is invertible.

(R4) There exists a $\rho > 0$ and a neighborhood Θ of θ_0 such that the class of functions $\{\phi(x; \theta, F) : \ (\theta, F) \in \Theta \times \mathcal{C}_\rho\}$ is P_{θ_0,η_0}-Donsker with square integrable envelope function.

Main result (Statistical generalised derivative)

Let $p(x;\theta,F))$ be the density, $\phi(x;\theta,F) = \frac{\partial}{\partial\theta}\log p(x;\theta,F)$, and $\psi(x;\theta,F) = d_F \log p(x;\theta,F)$.
For $\theta_t \to \theta_0$ and $F_t \to F_0$ as $t \to 0$, we have that

$$
\begin{aligned}
&E\left[t^{-1}\{\phi(X;\theta_t,F_0) - \phi(X;\theta_0,F_0)\}\right] \\
&= -E\left[\phi(X;\theta_0,F_0)\phi^T(X;\theta_0,F_0)\right]\{t^{-1}(\theta_t-\theta_0)\} + o(1),
\end{aligned}
$$

and

$$
E\left[t^{-1}\{\phi(X;\theta_t,F_t) - \phi(X;\theta_t,F_0)\}\right] = o(1).
$$

Proof of Main result

For each t, the equality

$$
\begin{aligned}
0 &= t^{-1}\left\{\int \phi(x;\theta_t,F_0)p(x;\theta_t,F_0)dx - \int \phi(x;\theta_0,F_0)p(x;\theta_0,F_0)dx\right\} \\
&= \int t^{-1}\{\phi(x;\theta_t,F_0)-\phi(x;\theta_0,F_0)\}p(x;\theta_0,F_0)dx \\
&\quad + \int \phi(x;\theta_t,F_0)t^{-1}\{p(x;\theta_t,F_0)-p(x;\theta_0,F_0)\}dx
\end{aligned}
$$

holds. By rearranging this, we get

$$
\begin{aligned}
&\int t^{-1}\{\phi(x;\theta_t,F_0)-\phi(x;\theta_0,F_0)\}p(x;\theta_0,F_0)dx \\
&= -\int \phi(x;\theta_t,F_0)t^{-1}\{p(x;\theta_t,F_0)-p(x;\theta_0,F_0)\}dx
\end{aligned}
$$

Corollary (Asymptotic linearity of profile likelihood MLE $\hat{\theta}$)

Suppose sets of assumptions $(R1) - (R6)$. Then a consistent solution $\hat{\theta}_n$ to the estimating equation

$$\sum_{i=1}^{n} \phi(X_i; \hat{\theta}_n, F_n) = 0 \tag{22}$$

is an asymptotically linear estimator for θ_0 :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}_0^{-1} \tilde{\ell}_0(X_i) + o_P(1).$$

Hence we have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \tilde{I}_0^{-1}\right) \quad \text{as } n \to \infty.$$

Proof of Corollary

By Lemma 19.24 in van der Vaart (1998) together with the dominated convergence theorem and condition (R4) implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi(X_i;\hat{\theta}_n,F_n)-\phi(X_i;\theta_0,F_0)\}$$
$$=\sqrt{n}E\{\phi(X;\hat{\theta}_n,F_n)-\phi(X;\theta_0,F_0)\}+o_P(1).$$

Using the main result , the righthand side is

$$\begin{aligned}&\sqrt{n}E\{\phi(X;\hat{\theta}_n,F_n)-\phi(X;\theta_0,F_0)\}\\ =\;&\sqrt{n}E\{\phi(X;\hat{\theta}_n,F_n)-\phi(X;\hat{\theta}_n,F_0)\}\\ &+\sqrt{n}E\{\phi(X;\hat{\theta}_n,F_0)-\phi(X;\theta_0,F_0)\}\\ =\;&-\tilde{I}_0\sqrt{n}(\hat{\theta}_n-\theta_0)+o_P(1)\end{aligned}$$

where $\tilde{I}_0=E\{\phi(X;\theta_0,F_0)\phi^T(X;\theta_0,F_0)\}$. Finally, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(X_i;\hat{\theta}_n,F_n)=0$ imply that

$$\sqrt{n}(\hat{\theta}_n-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{I}_0^{-1}\phi(X_i;\theta_0,F_0)+o_P(1).$$

Example continued: Joint mixture model (10)
The score operator for Θ

The survival part of score function for Θ is

$$\begin{aligned}
\dot{\ell}_{\Theta,\Lambda} &= \sum_{r=1}^{R} \gamma(Z_r) \frac{\partial}{\partial \Theta} \log P_{r,\Theta,\Lambda}(T, d) \\
&= \sum_{r=1}^{R} \gamma(Z_r) \begin{pmatrix} \delta_0 \\ \theta_r \\ X \end{pmatrix} \{ d - \Lambda(T) \exp(\theta_r \delta_0 + X \delta_1) \}.
\end{aligned}$$

Example continued: Joint mixture model (11)
The score operator for Λ

Let $h : [0, \tau] \to R$ be a function on $[0, \tau]$. The path defined by

$$d\Lambda_s = (1 + sh)d\Lambda$$

is a submodel passing through Λ at $s = 0$.
The survival part of score operator for Λ is given by

$$\begin{aligned} B_{\Theta,\Lambda} h &= \left. \frac{d}{ds} \right|_{s=0} \sum_{r=1}^{R} \gamma(Z_r) \log P_{r,\Theta,\Lambda_s}(T, d) \\ &= \sum_{r=1}^{R} \gamma(Z_r) \left(dh(T) - \exp(\theta_r \delta_0 + X\delta_1) \int_0^T h(u) d\Lambda(u) \right). \end{aligned}$$

Example continued: Joint mixture model (12)
The efficient score function derived from definition

The efficient score function for the survival part of the model is given by

$$\begin{aligned}
&\tilde{\ell}_{\Theta,\Lambda} = \dot{\ell}_{\Theta,\Lambda} - B_{\Theta,\Lambda}(B^*_{\Theta,\Lambda}B_{\Theta,\Lambda})^{-1}B^*_{\Theta,\Lambda}\dot{\ell}_{\Theta,\Lambda} \\
= &\sum_{r=1}^{R}\gamma(Z_r)d\left[\begin{pmatrix}\delta_0\\ \theta_r\\ X\end{pmatrix} - \frac{M_1(T)}{M_0(T)}\right] \\
&-\sum_{r=1}^{R}\gamma(Z_r)\left(\exp(\theta_r\delta_0 + X\delta_1)\int_0^T\left[\begin{pmatrix}\delta_0\\ \theta_r\\ X\end{pmatrix} - \frac{M_1(u)}{M_0(u)}\right]d\Lambda(u)\right)
\end{aligned}$$

Example continued: Joint mixture model (13)
Efficient score function derived from profile likelihood

The score function for Θ in the survival part of the profile likelihood (at the true value of parameters Θ) is

$$\begin{aligned}
& \sum_{r=1}^{R} \gamma(Z_r) \frac{\partial}{\partial \Theta} \log P\left(T_i, d_i \mid \widehat{\Lambda}(\Theta, F_0), \theta_r, \delta\right) \\
= & \sum_{r=1}^{R} \gamma(Z_r) d \left[\begin{pmatrix} \delta_0 \\ \theta_r \\ X \end{pmatrix} - \frac{M_1(T)}{M_0(T)} \right] \\
& - \sum_{r=1}^{R} \gamma(Z_r) \left(\exp(\theta_r \delta_0 + X \delta_1) \int_0^T \left[\begin{pmatrix} \delta_0 \\ \theta_r \\ X \end{pmatrix} - \frac{M_1(u)}{M_0(u)} \right] d\Lambda(u) \right)
\end{aligned}$$

The expression above is the efficient score function in the model derived from definition.

Overview and Reference

- Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood. J. Amer. Statist. Assoc.
- Hirose, Y. (2011). Efficiency of profile likelihood in semi-parametric models, Ann. Inst. Statist. Math.
- Hirose, Y. (2016). On differentiability of implicitly defined function in semi-parametric profile likelihood estimation, Bernoulli.

Reference

- Bishop, C.M. (2006) Pattern recognition and machine learning, Springer.
- Preedalikit, K. Liu, I. Hirose, Y. Sibanda, N. and Fernández, D. (2015) Joint modeling of survival and longitudinal ordered data using a semiparametric approach, Aust. N. Z. J.
- Hsieh, F. Tseng, Y.K. and Wang, J.L. (2006) Joint modeling of survival and longitudinal data: likelihood approach revisited, Biometrics.