Statistical generalized differentiation and asymptotic normality of estimator in a mixture of semiparametric models

YUICHI HIROSE[1]

[1] *Victoria University of Wellington, New Zealand.*
E-mail: Yuichi.Hirose@vuw.ac.nz

1. Introduction

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 $\theta_k\in\Theta\subset R^{m_k}$ and infinite dimensional parameter $\eta_k\in H_k$ where H_k is a subset of Banach space $\mathcal{B}_k$, and $\pi_1,\ldots,\pi_K$ are mixture probabilities. 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 $\pi=(\pi_1,\ldots,\pi_K)$. 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. This paper aim to establish large sample properties of the parameter θ using EM-algorithm and profile likelihood approach.

To discuss the EM-algorithm, we further introduce notations (we use notations from [Bishop (2006)]). Let $Z_i=(Z_{i1},\ldots,Z_{iK})$ be group indicator 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

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

Then the EM-algorithm utilize the identity

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

where $q(\mathbf{Z})$ is any distribution of $\mathbf{Z}$ ([McLachlan & Krishnan (2008)], Equation (3.3)).

In the E-step put

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

then it is well known that the gradient for the $\log p(\mathbf{X};\theta,\eta,\pi)$ coincides with the one for $\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z};\theta,\eta,\pi)$ at $(\theta^{old},\eta^{old},\pi^{old})$. In the M-step, maximize 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.

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 $\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z};\theta,\eta,\pi)$ ([McLachlan & Krishnan (2008)], Section 3.4.1).

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. In the following, we aim to establish asymptotic properties of the maximum likelihood estimator of θ using profile likelihood estimation with the EM-algorithm.

1.1. Estimation and asymptotic normality of the estimator

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

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

where

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

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

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\theta,\eta,\pi) \quad = \quad \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_k(X_i;\theta,\eta)[\log \pi_k + \log p_k(X_i;\theta_k,\eta_k)]. \quad (7)$$

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:

$$\hat{\pi}_k(\theta, \eta) = \frac{\sum_{i=1}^n \gamma_k(X_i; \theta, \eta)}{n}, \quad k = 1, \ldots, K. \tag{8}$$

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.

The efficient score function and information matrix in the mixture model: The score function for θ and score operator for η in the mixture model given in (1) are, respectively,

$$\dot{\ell}(x; \theta, \eta) = \frac{\partial}{\partial \theta} \log \left(\sum_{k=1}^K \pi_k p_k(x; \theta_k, \eta_k) \right) = \sum_{k=1}^K \gamma_k(x; \theta, \eta) \frac{\partial}{\partial \theta} \log p_k(x; \theta_k, \eta_k), \tag{9}$$

and

$$B(x; \theta, \eta) = d_\eta \log \left(\sum_{k=1}^K \pi_k p_k(x; \theta_k, \eta_k) \right) = \sum_{k=1}^K \gamma_k(x; \theta, \eta) d_\eta \log p_k(x; \theta_k, \eta_k) \tag{10}$$

where $\gamma_k(x; \theta, \eta)$ is given in (6) with X_i is replaced with x. The notation d_η is the Hadamard derivative operator with respect to the parameter η.

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 $B_0(x) = B(x; \theta_0, \eta_0)$. Then, it follows from the standard theory ([van der Vaart (1998)], page 374) that the efficient score function $\tilde{\ell}_0$ and the efficient information matrix $\tilde{I}_0$ in the semiparametric mixture model are given by

$$\tilde{\ell}_0(x) = (I - B_0(B_0^* B_0)^{-1} B_0^*) \dot{\ell}_0(x), \tag{11}$$

and

$$\tilde{I}_0 = E[\tilde{\ell}_0 \tilde{\ell}_0^T]. \tag{12}$$

Note: Equations (9) and (10) show that the score functions in the semiparametric mixture model (1) coincide with the ones for the expected complete data likelihood (7).

The score function for the profile likelihood: In the estimation of (θ, η) we use the profile likelihood approach: 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

$$\phi(x;\theta,F)=\frac{\partial}{\partial\theta}\log\left(\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})\right)=\sum_{k=1}^{K}\gamma_k(x;\theta,\hat{\eta}_{\theta,F})\frac{\partial}{\partial\theta}\log p_k(x;\theta_k,\hat{\eta}_{k,\theta,F}) \quad (13)$$

and

$$\psi(x;\theta,F)=d_F\log\left(\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})\right)=\sum_{k=1}^{K}\gamma_k(x;\theta,\hat{\eta}_{\theta,F})d_F\log p_k(x;\theta_k,\hat{\eta}_{k,\theta,F}), \quad (14)$$

We require that $\eta_0=\hat{\eta}_{\theta_0,F_0}$ and the condition (R2) below assumes $\phi(x;\theta_0,F_0)$ is the efficient score function $\tilde{\ell}_0(x)$ in the model where θ_0, η_0 and F_0 are the true values of the parameters θ, η and cdf F.

Assumptions: We list assumptions used for Theorem 1.1 and Theorem 1.2 given below.

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

$$\log p(x;\theta,F)=\log\left(\sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})\right) \quad (15)$$

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 $\psi(x;\theta,F)=d_F\log p(x;\theta,F)$ and they are given in (13) and (14).

(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.

(R5) If $\theta_t\to\theta_0$ and $F_t\to F_0$ as $t\to 0$, we have for $F\in\mathcal{C}_\rho$ and $\theta\in\Theta$,

$$\begin{aligned}&\phi(x;\theta_t,F)-\phi(x;\theta_0,F)=O(\theta_t-\theta_0)\quad\text{and}\quad\phi(x;\theta,F_t)-\phi(x;\theta,F_0)=O(F_t-F_0)\\&\psi(x;\theta_t,F)-\psi(x;\theta_0,F)=O(\theta_t-\theta_0)\quad\text{and}\quad\psi(x;\theta,F_t)-\psi(x;\theta,F_0)=O(F_t-F_0)\end{aligned} \quad (16)$$

(R6)

$$\left\|E[\phi(x;\theta_0,F_0)\phi^T(x;\theta_0,F_0)\psi(x;\theta_0,F_0)]\right\| < \infty \tag{17}$$

Main result: statistical generalized derivative and asymptotic linearity of the estimator. To calculate the second derivative of the score function $\phi(x;\theta,F,\gamma)$ given in (13), we use the idea similar to the derivative of generalized function. Let $\varphi \to (f,\varphi) = \int_{-\infty}^{\infty} f(x)\varphi(x)dx$ be a generalized 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.

Using condition (R1) and suppose the density for the profile likelihood $p(x;\theta,F)$ given in (15) is twice differentiable with respect to θ, then by differentiating 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 expressions for the efficient information matrix in terms of the score function $\phi(x;\theta_0,F_0)$:

$$\tilde{I}_0 = E[\phi\phi^T(X;\theta_0,F_0)] = -E\left[\frac{\partial}{\partial\theta^T}\phi(X;\theta_0,F_0)\right]. \tag{18}$$

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 $E[\phi\phi^T(X;\theta_0,F_0)]$. In the following theorem, we show that the definition is valid even when the derivative of the score function $\frac{\partial}{\partial\theta^T}\phi(x;\theta,F)$ does not exist.

Theorem 1.1 *Let $p(x;\theta,F) = \sum_{k=1}^{K}\pi_k p_k(x;\theta_k,\hat{\eta}_{k,\theta,F})$, $\phi(x;\theta,F) = \frac{\partial}{\partial\theta}\log p(x;\theta,F)$, and $\psi(x;\theta,F) = d_F \log p(x;\theta,F)$ as defined in (15), (13) and (14), respectively.*

Suppose (R1) and (R5), then, for $\theta_t \to \theta_0$ and $F_t \to F_0$ as $t \to 0$, we have that

$$\begin{aligned} &E\left[t^{-1}\{\phi(X;\theta_t,F_0) - \phi(X;\theta_0,F_0)\}\right] \\ &\quad = -E\left[\phi(X;\theta_0,F_0)\phi^T(X;\theta_0,F_0)\right]\{t^{-1}(\theta_t-\theta_0)\} + o\{1+t^{-1}(\theta_t-\theta_0)\}. \end{aligned} \tag{19}$$

Further suppose (R2) and (R6) then

$$E\left[t^{-1}\{\phi(X;\theta_t,F_t) - \phi(X;\theta_t,F_0)\}\right] = O\{t^{-1}(\theta_t-\theta_0)(F_t-F_0)\} + o\{1+t^{-1}(F_t-F_0)^2\}. \tag{20}$$

Note. *Note that even when the derivative $\frac{\partial}{\partial\theta}\phi(x;\theta,F)$ does not exist the equation (21) in the proof holds. Together with the derivative $\frac{\partial}{\partial\theta}p(x;\theta,F)$ exists imply that the derivative of the map $\theta \to E[\phi(x;\theta,F)]$ exists and it is given by (19). We may call the derivative the statistical generalized derivative. Similar comment for (20) holds.*

Proof.
First we prove (19). For each t, the equality

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

holds. It follows that

$$\begin{aligned} &\lim_{t\to 0}\int t^{-1}\{\phi(x;\theta_t,F_0)-\phi(x;\theta_0,F_0)\}p(x;\theta_t,F_0)dx \\ &\qquad = -\lim_{t\to 0}\int \phi(x;\theta_0,F_0)t^{-1}\{p(x;\theta_t,F_0)-p(x;\theta_0,F_0)\}dx. \end{aligned} \tag{21}$$

By the differentiability of $p(x;\theta,F)$ with respect to θ, the right hand side is equal to

$$\begin{aligned} &-\int \phi(x;\theta_0,F_0)\left[\frac{\partial}{\partial\theta^T}p(x;\theta_0,F_0)\left\{\lim_{t\to 0}t^{-1}(\theta_t-\theta_0)\right\}\right]dx \\ &= -\int \phi(x;\theta_0,F_0)\phi^T(x;\theta_0,F_0)p(x;\theta_0,F_0)dx\left\{\lim_{t\to 0}t^{-1}(\theta_t-\theta_0)\right\}. \end{aligned}$$

As long as we understood the limit $t\to 0$ is taken before the integral $\int\cdot dx$, the above can be written as

$$\begin{aligned} &\int t^{-1}\{\phi(x;\theta_t,F_0)-\phi(x;\theta_0,F_0)\}p(x;\theta_t,F_0)dx \\ &\qquad = -\int \phi(x;\theta_0,F_0)\phi^T(x;\theta_0,F_0)p(x;\theta_0,F_0)dx\left\{t^{-1}(\theta_t-\theta_0)\right\}+o(1). \end{aligned} \tag{22}$$

Using assumption (16), we get

$$\begin{aligned} &\left\|\int t^{-1}\{\phi(x;\theta_t,F_0)-\phi(x;\theta_0,F_0)\}p(x;\theta_t,F_0)dx - \int t^{-1}\{\phi(x;\theta_t,F_0)-\phi(x;\theta_0,F_0)\}p(x;\theta_0,F_0)dx\right\| \\ =\ &\left\|\int t^{-1}\{\phi(x;\theta_t,F_0)-\phi(x;\theta_0,F_0)\}\{p(x;\theta_t,F_0)-p(x;\theta_0,F_0)\}dx\right\| \\ \le\ &\left|\int O\{t^{-1}(\theta_t-\theta_0)\}\{p(x;\theta_t,F_0)-p(x;\theta_0,F_0)\}dx\right| = \int O\{t^{-1}(\theta_t-\theta_0)\}o(1)dx = o\{t^{-1}(\theta_t-\theta_0)\}. \end{aligned} \tag{23}$$

Altogether, we have (19):

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

Now we prove (20).
By Taylor's expansion,

$$p(x;\theta_t,F_0) = p(x;\theta_0,F_0) + \phi(x;\theta_0,F_0)p(x;\theta_0,F_0)(\theta_t-\theta_0) + o(\theta_t-\theta_0). \tag{24}$$

Further,

$$\begin{aligned} p(x;\theta_t,F_t) &= p(x;\theta_0,F_0) + \phi(x;\theta_0,F_0)p(x;\theta_0,F_0)(\theta_t-\theta_0) \\ &\quad +\psi(x;\theta_0,F_0)p(x;\theta_0,F_0)(F_t-F_0) + o\{(\theta_t-\theta_0)+(F_t-F_0)\}. \end{aligned} \tag{25}$$

Using assumption (16)

$$\begin{aligned} &\phi(x;\theta_t,F_0)\psi(x;\theta_t,F_0) \\ &= \{\phi(x;\theta_0,F_0) + O(\theta_t-\theta_0)\}\{\psi(x;\theta_0,F_0) + O(\theta_t-\theta_0)\} \\ &= \phi(x;\theta_0,F_0)\psi(x;\theta_0,F_0) + \{\phi(x;\theta_0,F_0) + \psi(x;\theta_0,F_0)\}O(\theta_t-\theta_0) + O(\theta_t-\theta_0)^2 \end{aligned} \tag{26}$$

Since $\psi(x;\theta_0,F_0)$ is in the nuisance tangent space and $\phi(x;\theta_0,F_0)$ is the efficient score function, we have

$$E[\phi(x;\theta_0,F_0)\psi(x;\theta_0,F_0)] = \int \phi(x;\theta_0,F_0)\psi(x;\theta_0,F_0)p(x;\theta_0,F_0)dx = 0. \tag{27}$$

As before, for each t, the following equation holds:

$$\begin{aligned} &\int t^{-1}\{\phi(x;\theta_t,F_t) - \phi(x;\theta_t,F_0)\}p(x;\theta_t,F_t)dx \\ &\quad = -\int \phi(x;\theta_t,F_0)t^{-1}\{p(x;\theta_t,F_t) - p(x;\theta_t,F_0)\}dx. \end{aligned} \tag{28}$$

By the differentiability of $p(x;\theta,F)$ with respect to F and (16), a similar proof as (22) can show that, as $t\to 0$, the equation (28) is equivalent to

$$\begin{aligned} &\int t^{-1}\{\phi(x;\theta_t,F_t) - \phi(x;\theta_t,F_0)\}p(x;\theta_t,F_t)dx \\ = \quad &-\int \phi(x;\theta_t,F_0)\psi(x;\theta_t,F_0)p(x;\theta_t,F_0)dx\{t^{-1}(F_t-F_0)\} + o(1). \end{aligned} \tag{29}$$

The left hand side of (29) is as $t\to 0$, using (16) and (25),

$$\begin{aligned} &\int t^{-1}\{\phi(x;\theta_t,F_t) - \phi(x;\theta_t,F_0)\}p(x;\theta_t,F_t) \\ &= \int t^{-1}\{\phi(x;\theta_t,F_t) - \phi(x;\theta_t,F_0)\}[p(x;\theta_0,F_0) + \phi(x;\theta_0,F_0)p(x;\theta_0,F_0)(\theta_t-\theta_0) \\ &\qquad +\psi(x;\theta_0,F_0)p(x;\theta_0,F_0)(F_t-F_0) + o\{(\theta_t-\theta_0)+(F_t-F_0)\}]dx \\ &= \int t^{-1}\{\phi(x;\theta_t,F_t) - \phi(x;\theta_t,F_0)\}p(x;\theta_0,F_0)dx + t^{-1}o\{(\theta_t-\theta_0)(F_t-F_0) + (F_t-F_0)^2\}. \end{aligned}$$

The integral in the right hand side of (29) is, using (24), (26) and (27),

$$\begin{aligned}
&\int \phi(x;\theta_t,F_0)\psi(x;\theta_t,F_0)p(x;\theta_t,F_0)dx \\
&= \int \{\phi(x;\theta_0,F_0)\psi(x;\theta_0,F_0) + [\phi(x;\theta_0,F_0)+\psi(x;\theta_0,F_0)]O(\theta_t-\theta_0) + O(\theta_t-\theta_0)\} \\
&\quad \times\{p(x;\theta_0,F_0) + \phi^T(x;\theta_0,F_0)p(x;\theta_0,F_0)(\theta_t-\theta_0) + o(\theta_t-\theta_0)\}dx \\
&= E[\phi(x;\theta_0,F_0)\phi^T(x;\theta_0,F_0)\psi(x;\theta_0,F_0)]O(\theta_t-\theta_0) + O(\theta_t-\theta_0).
\end{aligned}$$

Altogether, (29) is equivalent to

$$\begin{aligned}
&\int t^{-1}\{\phi(x;\theta_t,F_t)-\phi(x;\theta_t,F_0)\}p(x;\theta_0,F_0)dx + t^{-1}o\{(\theta_t-\theta_0)(F_t-F_0)+(F_t-F_0)^2\} \\
= \quad & -E[\phi(x;\theta_0,F_0)\phi^T(x;\theta_0,F_0)\psi(x;\theta_0,F_0)]t^{-1}O\{(\theta_t-\theta_0)(F_t-F_0)\} + t^{-1}O\{(\theta_t-\theta_0)(F_t-F_0)\} + o(1).
\end{aligned}$$

(20) follows from this with (17).

Using the result in Theorem 1.1, we can get the following result:

Theorem 1.2 *Suppose sets of assumptions* $(R1)-(R6)$*. Then a consistent solution* $\hat{\theta}_n$ *to the estimating equation*

$$\sum_{i=1}^{n}\phi(X_i;\hat{\theta}_n,F_n) = 0 \tag{30}$$

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

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

Proof

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

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi(X_i;\hat{\theta}_n,F_n)-\phi(X_i;\theta_0,F_0)\} = \sqrt{n}E\{\phi(X;\hat{\theta}_n,F_n)-\phi(X;\theta_0,F_0)\} + o_P(1). \tag{31}$$

Using (19) and (20) , the right hand side of (31) is

$$\begin{aligned}
&\sqrt{n}E\{\phi(X;\hat{\theta}_n,F_n)-\phi(X;\theta_0,F_0)\} \\
= \quad & \sqrt{n}E\{\phi(X;\hat{\theta}_n,F_n)-\phi(X;\hat{\theta}_n,F_0)\} + \sqrt{n}E\{\phi(X;\hat{\theta}_n,F_0)-\phi(X;\theta_0,F_0)\} \\
= \quad & \sqrt{n}(\hat{\theta}_n-\theta_0)O_p(F_n-F_0) - \tilde{I}_0\sqrt{n}(\hat{\theta}_n-\theta_0) + o_p\{1+\sqrt{n}(\hat{\theta}_n-\theta_0)+\sqrt{n}(F_n-F_0)^2\},
\end{aligned} \tag{32}$$

where $\tilde{I}_0 = E\{\phi(X;\theta_0,F_0)\phi^T(X;\theta_0,F_0)\}$. Since we assumed 4th-root-n-consistency, we have $\sqrt{n}(F_n - F_0)^2 = O_p(1)$ and $F_n - F_0 = o_p(1)$. Finally, (31) together with (32) and $\frac{1}{\sqrt{n}}\sum_{i=1}^n \phi(X_i;\hat{\theta}_n,F_n) = 0$ imply that

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

References

[Averbukh and Smolyanov (1968)] Averbukh, V.I. and Smolyanov, O.G. (1968) The various definitions of the derivative in linear topological spaces. *Russ. Math. Surv.* **23**, 67. doi:10.1070/RM1968v023n04ABEH003770

[Bishop (2006)] Bishop, C.M. (2006). *Pattern recognition and Machine learning*, Springer.

[Gill (1989)] Gill, R.D. (1989) Non-and semi-parametric maximum likelihood estimators and the von Mises method (part 1). *Scandinavian Journal of Statistics*, **16**, 97–128.

[Hirose (2011)] Hirose, Y. (2011). Efficiency of profile likelihood in semi-parametric models, *Ann. Inst. Statist. Math.* **63** 1247–1275.

[Hirose (2016)] Hirose, Y. (2016). On differentiability of implicitly defined function in semi-parametric profile likelihood estimation, *Bernoulli* **22** 589–614.

[Kolmogorov and Fomin (1975)] Kolmogorov, A.N. and Fomin, S.V. (1975) *Introductory Real Analysis*. Dover, New York.

[Kosorok (2008)] Kosorok, M.R. (2008) *Introduction to Empirical Processes and Semiparametric Inference* Springer, New York.

[McLachlan & Krishnan (2008)] McLachlan, G. & Krishnan, T. (2008). *The EM Algorithm and Extensions Second Edition*, Wiley, New York.

[Murphy and van der Vaart (2000)] Murphy, S.A. and van der Vaart, A.W. (2000) On profile likelihood (with discussion). *J. Amer. Statist. Assoc.* **95** 449–485.

[Shapiro (1990)] Shapiro, A. (1990) On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, **66**, 477–487.

[Sova (1966)] Sova, M. (1966) Condition of differentiability in linear topological spaces. *Czech. Math. J.* **14**, 485–508.

[van der Vaart (1998)] van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge university press, Cambridge.

[Vaart and Wellner (1996)] Van der Vaart, A. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.