#1951

EFFICIENT SCORES IN SEMIPARAMETRIC MIXTURE MODELS

by

Leonard A. Stefanski[1]
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203

[1] This research was partially funded by the National Science Foundation, Grant DMS-8613681.

Abstract

A simple variational argument is employed to establish the efficient score function in semiparametric mixture models. This provides an alternate derivation to those offered by Begun *et al* (1983) and Lindsay (1983).

AMS 1980 subject classifications. Primary 62F12; secondary 62G99.

Key words and phrases: Efficient score, mixture model, semiparametric.

1. INTRODUCTION

This note employs a simple variational argument to establish the optimality of Lindsay's (1983) "conditional" score for the class of semiparametric mixture models described below. The result yields a lower bound for the covariance matrix of the limiting distribution of $n^{1/2}$-consistent estimators of the parametric component of the mixture model. Lindsay (1983) established optimality of the conditional score by studying "directional score statistics." Optimality of the conditional score can also be deduced from the work of Begun *et al* (1983) which involves the notion of "Hellinger-differentiable likelihoods." The ease with which the optimality result is established in Theorem 1 comes at the expense of restricting attention to the class of regular estimating equations delineated in (R1) - (R3) below and by assuming that the parameter space of mixing densities is complete as defined in Section 3. A special case of Theorem 1 can be found in Stefanski and Carroll (1987).

The model and efficient score function are presented in Sections 2 and 3 respectively. Section 4 contains examples, one of which shows that the conditional score need not be efficient if the family of mixing densities is not complete.

2. THE MODEL

Suppose that Θ is an open subset of $\mathbb{R}^p$ and that for each θ in Θ there exist functions $C_\theta\colon \mathbb{R}^q \to \mathbb{R}^k$, $T_\theta\colon \mathbb{R}^n \to \mathbb{R}^k$, $d_\theta\colon \mathbb{R}^q \to \mathbb{R}^1$ and $S_\theta\colon \mathbb{R}^n \to \mathbb{R}^1$ and an open subset, H_θ, of $\mathbb{R}^q$ such that for each η in H_θ

$$h(\bullet;\eta,\theta) = \exp\{C_\theta^T(\eta)T_\theta(\bullet) + d_\theta(\eta) + S_\theta(\bullet)\} \tag{2.1}$$

is a probability density with respect to a sigma-finite measure $m(\bullet)$ (not depending on θ or η) on $\mathbb{R}^n$.

For fixed θ, $\{h(\cdot;\eta,\theta): \eta \in H_\theta\}$ is a k-parameter exponential family with the natural sufficient statistic determined by $T_\theta(\cdot)$. Assume that this family is regular, $C_\theta(\cdot)$ is continuous and that the range of $C_\theta(\eta)$ for η in H_θ has a nonempty interior. This, in turn, implies that the family is complete. Let $\mathcal{G}$ be a collection of probability densities on $\mathbb{R}^q$ with respect to a sigma-finite measure $\nu(\cdot)$ containing a component which is absolutely continuous with respect to Lebesque measure. Now if g is a density in $\mathcal{G}$ for which supp(g) is a subset of H_θ, then

$$f(\cdot;\theta,g) = \int h(\cdot;\eta,\theta)g(\eta)d\nu(\eta) \qquad (2.2)$$

is a probability density on $\mathbb{R}^n$ with respect to the measure $m(\cdot)$. Equation (2.2) defines a semi-parametric model with typical "parameter" $\omega = (\theta,g)$. Let Ω be the parameter space for (2.2) specified as $\Omega = \{(\theta,g): \theta \in \Theta, g \in \mathcal{G}$ and $\text{supp}(g) \subset H_\theta\}$. Finally introduce the notation $\ell(\cdot,\theta,g) = \log f(\cdot;\theta,g)$ and $\dot{\ell}(\cdot,\theta,g) = (\partial/\partial\theta)\ell(\cdot,\theta,g)$ and let Z, Z_1, Z_2, ... denote i.i.d. random vectors with common density (2.2).

3. THE EFFICIENT SCORE FUNCTION

This section considers asymptotic efficiency of the class of M-estimators, i.e., those estimators satisfying equations of the form $\Sigma\, \psi(Z_i,\hat{\theta}) = 0$ where $\psi: \mathbb{R}^n \times \Theta \to \mathbb{R}^p$. Under an assumption concerning the richness of $\mathcal{G}$ an optimal estimating equation is identified using a simple variational argument.

Let Ψ be the class of regular estimating equations for Θ defined by the requirements that if ψ is in Ψ then for every ω in Ω:

(R1) $E_\omega\{\|\psi(Z,\Theta)\|^2\} < \infty$;

(R2) $E_\omega\{\psi(Z,\Theta)\} = 0$;

(R3) There exists a positive definite matrix $V_\psi = \{E_\omega(\psi\dot{\ell}^T)\}^{-1} E_\omega(\psi\psi^T)\{E_\omega(\dot{\ell}\psi^T)\}^{-1}$ and a sequence of estimators, $\{\hat{\Theta}\}$, satisfying $\Sigma\, \psi(Z_i,\hat{\Theta}) = 0$ such that $n^{1/2}(\hat{\Theta} - \Theta) \xrightarrow[\omega]{L} N(0,V_\psi)$.

It is now shown that the unbiasedness condition (R2) implies conditional unbiasedness with respect to $T_\Theta(Z)$, i.e., $E_\omega\{\psi(Z,\Theta)|T_\Theta(Z)\} = 0$, for every ψ in Ψ, provided the family of densities $\mathcal{G}$ is complete as defined below.

(D) A collection of functions, $\mathcal{S}$, is said to be complete with respect to a measure μ if a necessary condition for

$$\int r(t)s(t)d\mu(t) = 0 \text{ for all } s \text{ in } \mathcal{S}$$

is $r(\cdot) = 0$ μ-almost surely.

For a fixed Θ in Θ let $\mathcal{G}_\Theta = \{g \in \mathcal{G}: (\Theta,g) \in \Omega\}$ and assume:

(C) $\mathcal{G}_\Theta$ is complete with respect to ν for each Θ in Θ.

Assumption (C) plays a role similar to the convexity condition (C) of Bickel (1982), and to assumption (S) of Begun et. al. (1983). Note that if $\mathcal{G}_\Theta$ contains a complete parametric family of densities (in the familiar sense) then it is necessarily complete in the sense of (D).

The following Lemma establishes the conditional unbiasedness of scores in Ψ.

LEMMA 1. If $\psi \in \Psi$, then condition (C) implies $E_{\omega}\{\psi(Z,\Theta)|T_{\Theta}(Z)\} = 0$ for all ω in Ω.

PROOF. Fix Θ and let $T = T_{\Theta}(Z)$. For any ψ in Ψ we know that $E_{\Theta,g}\{\psi(Z,\Theta)\} = 0$ for all g in $\mathcal{B}_{\Theta}$. Conditioning first on $T_{\Theta}(Z)$ implies that $E_{\Theta,g}\{Q(T)\} = 0$ for all g in $\mathcal{B}_{\Theta}$ where $Q(T) = E_{\Theta,g}\{\psi(Z,\Theta)|T_{\Theta}(Z)\}$. But

$$E_{\Theta,g}\{Q(T)\} = \int Q\{T_{\Theta}(z)\}f(z;\Theta,g)dm(z)$$

$$= \int Q\{T_{\Theta}(z)\}\int h(z;\eta,\Theta)g(\eta)d\nu(\eta)dm(z)$$

$$= \int \left[\int Q\{T_{\Theta}(z)\}h(z;\eta,\Theta)dm(z)\right] g(\eta)d\nu(\eta)$$

where the interchange of integrations is justified by (R1) and Fubini's Theorem. Since $E_{\Theta,g}\{Q(T)\} = 0$ for all g in $\mathcal{B}_{\Theta}$, condition (C) implies that

$$\int Q\{T_{\Theta}(z)\}h(z;\eta,\Theta)dm(z) = 0 \qquad (3.1)$$

ν_{Θ}-almost surely where ν_{Θ} is the restriction of ν to H_{Θ}. Continuity of $C_{\Theta}(\bullet)$, the integrability condition (R1) and the exponential character of $h(\bullet;\eta,\Theta)$ imply that the left hand side of (3.1) is a continuous function of η. Thus since ν contains a component which is absolutely continuous with respect to Lebesgue measure, (3.1) holds for all η in H_{Θ}. Finally, completeness of the family $\{h(\bullet;\eta,\Theta): \eta \in H_{\Theta}\}$ implies that $Q\{T_{\Theta}(Z)\} = 0$ almost surely. ////

Lemma 1 shows that the only scores which are unbiased for all $\omega \in \Omega$ are those which are conditionally unbiased with respect to $T_\theta(Z)$. It also permits a simply proof that every score in Ψ is less efficient than $\psi^*(Z,\theta) = \dot{\ell}(Z,\theta,g) - E\{\dot{\ell}(Z,\theta,g)|T_\theta(Z)\}$ in that $V_\psi \geq V_{\psi^*} = \{E(\psi^*\psi^{*T})\}^{-1}$ in the sense of positive definiteness. In order for V_{ψ^*} to be a meaningful lower bound, (i.e., be finite) it is necessary that $E(\psi^*\psi^*)$ be positive definite. This means that the σ-field generated by $T_\theta(Z)$ must be strictly contained in the σ-field generated by Z and, in particular, that no linear combination of $\dot{\ell}$, $\lambda^T\dot{\ell}$, can be written as a function of $T_\theta(Z)$ almost surely. Under this assumption the following result is obtained.

THEOREM 1. Under the conditions stated above, $V_\psi \geq V_{\psi^*}$ for all ψ in Ψ.

PROOF. Let IC_ψ be the influence function for ψ, i.e., $IC_\psi = \{E(\psi\dot{\ell}^T)\}^{-1}\psi$; IC_{ψ^*} is the influence function for ψ^*. Again let $T = T_\theta(Z)$.

Pick any ψ in Ψ. Since $\psi\dot{\ell}^T = \psi\psi^{*T} + \psi E(\dot{\ell}^T|T)$, conditioning first on T and then appealing to Lemma 1 shows that $E(\psi\dot{\ell}^T) = E(\psi\psi^{*T})$ whenever ψ is conditionally unbiased. From this it follows that $V_{\psi^*}^{-1} = E(\psi^*\psi^{*T})$ and $E(IC_\psi IC_{\psi^*}^T) = E(IC_{\psi^*}IC_\psi^T) = V_{\psi^*}$. Using these identities shows that $E\{(IC_\psi - IC_{\psi^*})(IC_\psi - IC_{\psi^*})^T\} = E(IC_\psi IC_\psi^T) - E(IC_{\psi^*}IC_\psi^T) - E(IC_\psi IC_{\psi^*}^T) + E(IC_{\psi^*}IC_{\psi^*}^T) = V_\psi - V_{\psi^*}$. Thus V_ψ can be written as the sum of V_{ψ^*} and a nonnegative definite matrix which vanishes if and only if $IC_\psi = IC_{\psi^*}$ almost surely, establishing the desired result. ////

Provided differentiation and integration can be interchanged in (2.2) $\dot{\ell}(Z,\theta,g)$ is given by

$$\int \frac{\{(\partial/\partial\theta)\ \log\ h(Z;\eta,\theta)\}h(Z;\eta,\theta)g(\eta)d\nu(\eta)}{\int h(Z;\eta,\theta)g(\eta)d\nu(\eta)}$$

and upon taking expectations conditional on $T_\theta(Z)$ one finds that

$$\mathring{\ell}(Z,\theta,g) - E\{\mathring{\ell}(Z,\theta,g)|T_\theta(Z)\} =$$

$$\mathring{S}_\theta(Z) - E\{\mathring{S}_\theta(Z)|T_\theta(Z)\} +$$

$$\left[\mathring{f}_\theta(Z) - E\{\mathring{f}_\theta(Z)|T_\theta(Z)\}\right] \frac{\int C_\theta(\eta)\exp\{C_\theta^T(\eta)T_\theta(Z)+d_\theta(\eta)\}g(\eta)d\nu}{\int \exp\{C_\theta^T(\eta)T_\theta(Z)+d_\theta(\eta)\}g(\eta)d\nu} \qquad (3.2)$$

where "$\cdot$" denotes differentiation with respect to θ and for a function $R(\theta) = \{R_1(\theta),\ldots,R_s(\theta)\}^T$, $\dot{R}(\theta)$ denotes the pxs matrix with i,jth entry $\partial R_j(\theta)/\partial\theta_i$. The ratio of integrals in (3.2) is seen to be $E\{C_\theta(\eta)|T_\theta(Z)\}$.

4. EXAMPLES

Applications to measurement-error models are discussed in Stefanski and Carroll (1987). Example 4.1 has been previously considered by Lindsay (1983). Example 4.2 was motivated by some unpublished work of Brian Allen's at the University of Guelph citing difficulties with estimation in random coefficient models and by a recent problem discussed by Cox and Solomon (1988). Finally Section 4.3 illustrates the crucial role played by assumption (C).

4.1 Paired exponentials with proportional hazards.

Let $Z = (Y_1, Y_2)^T$ where Y_1 and Y_2 are independent exponentially distributed random variables with means $(\eta\theta)^{-1}$ and η^{-1} respectively. Then $h(Z;\eta,\theta)$ has the form (2.1) with $C_\theta(\eta) = \eta$, $T_\theta(Z) = -(\theta Y_1+Y_2)$, $d_\theta(\eta) = \log(\theta\eta^2)$ and $S_\theta(Z) = 0$. For this model $\mathring{f}_\theta(Z) = -Y_1$ and $E\{\mathring{f}_\theta(Z)|T_\theta(Z)\} = -(2\theta)^{-1}(\theta Y_1+Y_2)$ from which we get the efficient score

$$\psi^*(Z,\theta) = \left(\frac{Y_2}{2\theta} - \frac{Y_1}{2}\right) \frac{\int_0^\infty \eta^3 \exp(-\eta\theta Y_1 - \eta Y_2) g(\eta) d\eta}{\int_0^\infty \eta^2 \exp(-\eta\theta Y_1 - \eta Y_2) g(\eta) d\eta} \quad .$$

For this model it is assumed that ν is Lebesque measure on $(0,\infty)$ and $\mathcal{D}$ contains all probability densities on $(0,\infty)$. The asymptotic variance lower bound for this model is given by $[E\{\psi^*(Z,\theta)\}^2]^{-1}$ where

$$E\{\psi^*(Z,\theta)\}^2 = \int \frac{x^3\{Q'(x)\}^2}{12\theta^2\, Q(x)}\, dx \tag{4.1}$$

and

$$Q(x) = \int_0^\infty \eta^2 e^{-\eta x} g(\eta) d\eta \quad .$$

4.2 Random coefficient regression models.

Let $Z = (Y_1,\ldots,Y_n)^T$ and let X be an nxp matrix of rank p. Let $\Sigma(\rho)$ be an nxn matrix with i,jth entry $\rho^{|i-j|}$ $(|\rho| < 1)$. With $\eta = \beta$, a px1 vector of regression coefficients, and $\theta = (\sigma^2,\rho)^T$ suppose that the density of Z given η and θ has the form (2.1) where $C_\theta(\eta) = \eta/\sigma^2$, $T_\theta(Z) = X^T\Sigma^{-1}(\rho)Z$, $S_\theta(Z) = -(2\sigma^2)^{-1}\, Z^T\Sigma^{-1}(\rho)Z$ and $d_\theta(\eta) = -(2\sigma^2)^{-1}\, \eta^T X^T\Sigma^{-1}(\rho)X\eta$ $- (1/2)\log|\Sigma(\rho)| - (n/2)\log(2\pi\sigma^2)$. Under the assumption that β is a random vector the above delineates a random coefficient regression model with autoregressive errors. For this model

$$\dot{S}_\theta(Z) = \begin{pmatrix} \dfrac{Z^T\Sigma^{-1}Z}{2\sigma^4} \\[2ex] \dfrac{Z^T\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}Z}{2\sigma^2} \end{pmatrix} \tag{4.2}$$

and

$$\dot{T}_\theta(Z) = \begin{pmatrix} 0 \\ -Z^T\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}X \end{pmatrix} \qquad (4.3)$$

where $\Sigma^{-1} = \Sigma^{-1}(\rho)$ and $(\dot{\Sigma})_{i,j} = |i-j|\rho^{|i-j|-1}$.

Let $M = X^T\Sigma^{-1}X$, then some routine calculations show that

$$E\{Z|T_\theta(Z)\} = X M^{-1}T_\theta(Z) ;$$

$$E\{\frac{Z^T\Sigma^{-1}Z}{2\sigma^4} \,|\, T_\theta(Z)\} = \frac{n-p}{2\sigma^2} + \frac{T_\theta^T(Z)M^{-1}T_\theta(Z)}{2\sigma^4} ;$$

$$E\left\{\frac{Z^T\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}Z}{2\sigma^2} \,\middle|\, T_\theta(Z)\right\} = (1/2)\mathrm{tr}\left\{\Sigma^{-1}\dot{\Sigma}(I - \Sigma^{-1}XM^{-1}X^T)\right\}$$

$$+ (2\sigma^2)^{-1}\left\{T_\theta^T(Z)M^{-1}X^T\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}X\, M^{-1}T_\theta(Z)\right\} .$$

These conditional expectations determine the optimal score for θ using (4.2) (4.3) and (3.2). The resulting expression is very complicated. The feasibility of using this result in practice in full generality will be explored in a future paper. For now attention is restricted to two special cases of this model, one of long-standing interest (Neyman and Scott, 1948), the other of more recent interest (Cox and Solomon, 1988).

Suppose $\rho = 0$ and $X = (1,\ldots,1)^T$, thus Z_i consists of n normal measurements on the scalar β_i. Neyman and Scott (1948) discussed estimation of σ^2 in this setup. The optimal estimating equation reduces to

$$\psi^*(Z,\theta) = -\frac{Z^TZ}{2\sigma^4} + \frac{n-1}{2\sigma^2} + \frac{(Z^TX)^2}{2n\sigma^4}$$

which yields the familiar estimator

$$\hat{\sigma}^2 = \{N(n-1)\}^{-1}\Sigma_{i=1}^N\{Z_i^TZ_i - (Z_i^TX)^2/n\}$$

from the sample $\{Z_1, \ldots, Z_N\}$.

Now suppose $X = (1,1,1)^T$, so that Z_i forms a stationary first-order autoregressive process of correlation ρ and mean β_i of length 3, i=1, ..., N, see Cox and Solomon (1988). For this model

$$\dot{S}_\theta(Z) = \left\{ \begin{array}{c} \dfrac{Y_1^2 + (1-\rho^2)Y_2^2 + Y_3^2 - 2\rho\, Y_2(Y_1+Y_3)}{2\sigma^4(1-\rho^2)} \\ \\ \dfrac{-2\rho(Y_1^2 + 2Y_2^2 + Y_3^2) + 2(1+\rho^2)Y_2(Y_1+Y_3)}{2\sigma^2(1-\rho^2)^2} \end{array} \right\},$$

$$\dot{T}_\theta(Z) = \left\{ \begin{array}{c} 0 \\ \\ -\dfrac{Y_1 + 2Y_2+Y_3}{(1+\rho)^2} \end{array} \right\},$$

$$E\{\dot{S}_\theta(Z)|T_\theta(Z)\} = \left\{ \begin{array}{c} \dfrac{1}{\sigma^2} + \dfrac{(Y_1+(1-\rho)Y_2+Y_3)^2}{2\sigma^4(1+\rho)(3-\rho)} \\ \\ \dfrac{2(1-\rho)^2-4\rho}{(3-\rho)(1-\rho^2)} + \dfrac{2(Y_1+Y_2(1-\rho)+Y_3)^2}{\sigma^2(3-\rho)^2(1+\rho)^2} \end{array} \right\},$$

and

$$E\{\dot{T}_\theta(Z)|T_\theta(Z)\} = \left\{ \begin{array}{c} 0 \\ \\ -\dfrac{4(Y_1+(1-\rho)Y_2+Y_3)}{(1+\rho)^2(3-\rho)} \end{array} \right\}.$$

These four quantities are combined according to (3.2) to form the efficient estimating equations for $(\sigma^2,\rho)^T$. Since the resulting score depends on g through $E\{\beta|Y_1+(1-\rho)Y_2+Y_3\}$, fully efficient estimation in this model requires estimation of this regression function.

4.3 On the completeness of $\mathcal{G}$.

If $\mathcal{G}$ is not complete then the optimality result does not necessarily hold. Consider the paired-exponentials example (Sec. 4.1) and suppose now that $\mathcal{G} = \{g: \int \eta^{-1} g(\eta) d\eta = 1\}$. Then $E(Y_1) = E\{E(Y_1|\eta)\} = E(\theta\eta)^{-1} = \theta^{-1}E(\eta^{-1}) = \theta^{-1}$, independent of $g(\cdot)$, and $\text{var}(Y_1) = 2\theta^{-2}\{E(\eta^{-2}) - 1/2\}$ where $E(\eta^{-i}) = \int \eta^{-i} g(\eta) d\eta$. This implies that if $\tilde{\theta} = \bar{Y}_1^{-1}$, then $n^{1/2}(\tilde{\theta} - \theta) \xrightarrow[L]{\omega} N(0, 2\theta^2\{E(\eta^{-2}) - 1/2\})$. Now take a sequence, $\{g_k(\cdot)\}$, of densities in $\mathcal{G}$ such that $g_k(\cdot)$ has support (a_k, b_k) where both a_k and $1/b_k$ increase to one as k increases. Then $2\theta^2\{E(\eta^{-2}) - 1/2\} \to \theta^2$ while the inverse of (4.1) approaches $2\theta^2$. Thus there are choices of g in $\mathcal{G}$ under which $\tilde{\theta}$ beats the "optimal" estimator. Of course $\mathcal{G}$ is not complete in this case.

References

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-71.

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432-52.

Cox, D. R. and Solomon, P. J. (1988). On testing for serial correlation in large numbers of small samples. *Biometrika* **75**, 145-48.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

Lindsay, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11**, 486-97.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703-16.