Rate of Convergence of k-Step Newton Estimators to Efficient Likelihood Estimators

Steve Verrill[1], Richard A. Johnson[2]

November 26, 2006

[1]Mathematical Statistician, USDA Forest Products Laboratory, 1 Gifford Pinchot Drive, Madison, Wisconsin 53726

[2]Professor of Statistics, University of Wisconsin, Madison, Wisconsin

Abstract

We make use of Cramér conditions together with the well-known local quadratic convergence of Newton's method to establish the asymptotic closeness of k-step Newton estimators to efficient likelihood estimators. In Verrill and Johnson (2006), we use this result to establish that estimators based on Newton steps from $\sqrt{n}$-consistent estimators may be used in place of efficient solutions of the likelihood equations in likelihood ratio, Wald, and Rao tests. Taking a quadratic mean differentiability approach rather than our Cramér condition approach, Lehmann and Romano (2005) have outlined proofs of similar results. However, their Newton step estimator results actually rely on unstated assumptions about Cramér conditions. Here we make our Cramér condition assumptions and their use explicit.

Keywords: one-step Newton estimators, Newton's method, $\sqrt{n}$-consistent estimators, Cramér conditions, quadratic mean differentiability, likelihood ratio test, Wald test, Rao test, asymptotics

1 Introduction

Lehmann (1983) (Theorems 3.1 and 4.2 of Chapter 6) demonstrates that efficient likelihood estimators, $\hat{\boldsymbol{\theta}}_n$, in

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \stackrel{D}{\to} \mathrm{N}(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1})$$

can be replaced by Newton one-step estimators, $\boldsymbol{\theta}_{n,\mathrm{Newt}}$, that are generated from $\sqrt{n}$-consistent[1] estimators, $\boldsymbol{\theta}_{n,c}$, via

$$\boldsymbol{\theta}_{n,\mathrm{Newt}} \equiv - \left[\frac{\partial^2 \ln L}{\partial\theta_l \partial\theta_m}\right]^{-1}_{s\times s} |_{\boldsymbol{\theta}_{n,c}} \begin{pmatrix} \partial \ln L/\partial\theta_1 \\ \vdots \\ \partial \ln L/\partial\theta_s \end{pmatrix} |_{\boldsymbol{\theta}_{n,c}} + \boldsymbol{\theta}_{n,c}$$

where $\boldsymbol{\theta} = (\theta_l, \ldots, \theta_s)^T$ and L is the likelihood.

In this paper we establish that under Lehmann's versions of the Cramér conditions, we have

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n,\mathrm{Newt}} = \mathbf{O}_p(n^{-1}) \tag{1}$$

An immediate corollary of our approach is that if $\boldsymbol{\theta}_{n,k,\mathrm{Newt}}$ denotes the result from the kth Newton step from a $\sqrt{n}$-consistent initial estimate then

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n,k,\mathrm{Newt}} = \mathbf{O}_p(n^{-(2^{k-1})})$$

In Verrill and Johnson (2006) we have used result (1) to establish that estimators based on Newton steps from $\sqrt{n}$-consistent estimators may be used in place of efficient solutions of the likelihood equations in likelihood ratio, Wald, and Rao tests. Taking a quadratic mean differentiability approach rather than our Cramér condition approach, Lehmann and Romano (2005) outline proofs of results closely related to those of Verrill and Johnson (2006). However, their Newton step estimator results actually rely on unstated assumptions about Cramér conditions. Here we make our Cramér condition assumptions and their use explicit.

Under regularity conditions, Janssen *et al* (1985) demonstrate in the one-dimensional case that

$$\hat{\theta}_n - T_n^{(1)} = O_p(n^{-1})$$

where $T_n^{(1)} \approx \theta_{n,\mathrm{Newt}}$, and

$$\hat{\theta}_n - T_n^{(2)} = O_p(n^{-3/2})$$

where

$$T_n^{(2)} \approx - \left(\frac{\partial^2 \ln L}{\partial\theta^2}|_{\theta_{n,\mathrm{Newt}}}\right)^{-1} \frac{\partial \ln L}{\partial\theta}|_{\theta_{n,\mathrm{Newt}}} + \theta_{n,\mathrm{Newt}}$$

That is, $T_n^{(2)}$ is approximately equal to the result of a second Newton step.

Janssen *et al*'s conditions are related to Lehmann's (1983) version of the Cramér conditions. However, neither set is strictly weaker than the other.

[1] $\hat{a}$ is a $\sqrt{n}$-consistent estimator of a if $\sqrt{n}(\hat{a} - a) = O_p(1)$

2 Lehmann's (1983) versions of the Cramér conditions

Let the parameter space be denoted by $\Theta \subset R^s$. Let $\boldsymbol{\theta}_0 \in \Theta$ denote the true parameter value.

(A0) The distributions $P(\boldsymbol{\theta})$ of the observations are distinct. That is, distinct $\boldsymbol{\theta}$'s cannot correspond to the same distribution.

(A1) The distributions $P(\boldsymbol{\theta})$ have common support.

(A2) The observations are $\mathbf{X} = (X_1 \ldots X_n)^T$ where the X_i are iid with probability density $f(x;\boldsymbol{\theta})$. (The X_i may be vector valued.)

(A) There exists an open subset T of Θ that contains the true parameter value $\boldsymbol{\theta}_0$ such that for almost all x, the density $f(x;\boldsymbol{\theta})$ has continuous third derivatives, $\partial^3 f(x;\boldsymbol{\theta})/\partial\theta_l\partial\theta_m\partial\theta_p$ for all $\boldsymbol{\theta} \in T$.

(B) For all $\boldsymbol{\theta}$ in T, the first and second logarithmic derivatives of f satisfy the equations

$$E_{\boldsymbol{\theta}}(\partial \ln f(X;\boldsymbol{\theta})/\partial\theta_l) = 0$$

for $l = 1, \ldots, s$ and

$$I_{lm}(\boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}}(\partial \ln f(X;\boldsymbol{\theta})/\partial\theta_l \times \partial \ln f(X;\boldsymbol{\theta})/\partial\theta_m) = E_{\boldsymbol{\theta}}(-\partial^2 \ln f(X;\boldsymbol{\theta})/\partial\theta_l\partial\theta_m)$$

for $l, m = 1, \ldots, s$. The $I_{lm}(\boldsymbol{\theta})$ are finite.

(C) $I(\boldsymbol{\theta}) \equiv [I_{lm}]_{s\times s}$ is positive definite for all $\boldsymbol{\theta}$ in T.

(D) For all l, m, p, $\partial^3 \ln f(x;\boldsymbol{\theta})/\partial\theta_l\partial\theta_m\partial\theta_p$ is a continuous function of $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in T$. Further, there exist integrable functions $M_{lmp}(x)$ such that

$$\left|\partial^3 \ln f(x;\boldsymbol{\theta})/\partial\theta_l\partial\theta_m\partial\theta_p\right| \leq M_{lmp}(x)$$

for all $\boldsymbol{\theta} \in T$, and

$$m_{lmp} \equiv E_{\boldsymbol{\theta}_0}(M_{lmp}(X)) < \infty$$

for all l, m, p.

Given conditions **(A0)** through **(D)**, Lehmann establishes (Theorem 4.1 of his Chapter 6) that there exists a consistent solution of the likelihood equations, $\hat{\boldsymbol{\theta}}_n$, that satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \stackrel{D}{\to} N(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1}) \tag{2}$$

3 The Theorem

Assume that conditions **(A0)** through **(D)** hold. Let $\boldsymbol{\theta}_{n,c}$ be a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$. That is, assume that

$$\sqrt{n}(\boldsymbol{\theta}_{n,c} - \boldsymbol{\theta}_0) = \mathbf{O}_p(1) \tag{3}$$

Then with probability approaching one as $n \to \infty$, the Newton estimator,

$$\boldsymbol{\theta}_{n,\mathrm{Newt}} \equiv -\left[\frac{\partial^2 \ln L}{\partial\theta_l \partial\theta_m}\right]^{-1}_{s\times s} |_{\boldsymbol{\theta}_{n,c}} \begin{pmatrix} \partial \ln L/\partial\theta_1 \\ \vdots \\ \partial \ln L/\partial\theta_s \end{pmatrix} |_{\boldsymbol{\theta}_{n,c}} + \boldsymbol{\theta}_{n,c} \tag{4}$$

is well-defined (that is the partials exist and the matrix is invertible), and

$$\boldsymbol{\theta}_{n,\mathrm{Newt}} - \hat{\boldsymbol{\theta}}_n = \mathbf{O}_p(n^{-1}) \tag{5}$$

where $\hat{\boldsymbol{\theta}}_n$ is a consistent solution of the likelihood equations guaranteed by Lehmann's theorem.
Proof

We will be making use of the fact that the Newton method yields quadratic convergence. In particular, we will verify the conditions of Theorem 5.2.1 in Dennis and Schnabel (1983).

By assumption **(D)** we can define

$$J_n(\boldsymbol{\theta}) \equiv -\left(\left[\frac{\partial^2 \ln L}{\partial\theta_l \partial\theta_m}\right]_{s\times s} |_{\boldsymbol{\theta}}\right)/n$$

We have

$$J_n(\hat{\boldsymbol{\theta}}_n) - J_n(\boldsymbol{\theta}_0) = -\left[\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f(X_i;\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m}|_{\hat{\boldsymbol{\theta}}_n} - \frac{\partial^2 \ln f(X_i;\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m}|_{\boldsymbol{\theta}_0}\right)/n\right]_{s\times s}$$

and, making use of assumption **(D)**, by Taylor's theorem

$$\frac{\partial^2 \ln f(X_i;\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m}|_{\hat{\boldsymbol{\theta}}_n} - \frac{\partial^2 \ln f(X_i;\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m}|_{\boldsymbol{\theta}_0} = \left(\frac{\partial^3 \ln f(X_i;\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m \partial\theta_1}, \dots, \frac{\partial^3 \ln f(X_i;\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m \partial\theta_s}\right)|_{\boldsymbol{\theta}^*_{lm,n}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

where $\boldsymbol{\theta}^*_{lm,n}$ lies on the line segment between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$.

Thus, by assumption **(D)**, for $\hat{\boldsymbol{\theta}}_n \in T$ (an open neighborhood of $\boldsymbol{\theta}_0$), the absolute value of the lmth element of $J_n(\hat{\boldsymbol{\theta}}_n) - J_n(\boldsymbol{\theta}_0)$ is bounded by

$$\sum_{i=1}^{n}\sum_{p=1}^{s} M_{lmp}(X_i)\left|\hat{\theta}_{pn} - \theta_{p0}\right|/n \tag{6}$$

Since (by assumptions **(A2)** and **(D)** and the strong law of large numbers)

$$\sum_{i=1}^{n} M_{lmp}(X_i)/n_j \overset{a.s.}{\to} m_{lmp} < \infty$$

for $l, m, p \in \{1, \dots, s\}$, results 2 and 6 imply that

$$||J_n(\hat{\boldsymbol{\theta}}_n) - J_n(\boldsymbol{\theta}_0)||_F \overset{p}{\to} 0 \tag{7}$$

where $||M||_F$ denotes the Frobenius norm of the matrix M.

Now by assumptions **(A2)** and **(B)** and the strong law of large numbers, we know that

$$J_n(\boldsymbol{\theta}_0) \overset{a.s.}{\to} I(\boldsymbol{\theta}_0) \tag{8}$$

Results 7 and 8 imply that

$$||J_n(\hat{\boldsymbol{\theta}}_n) - I(\boldsymbol{\theta}_0)||_F \xrightarrow{p} 0 \tag{9}$$

By assumption **(C)**, $I(\boldsymbol{\theta}_0)$ is positive definite. Since the inverse and norm of a matrix are continuous functions of the elements of the matrix, this implies that given any $\delta > 0$, we can find an $N_{\delta,1}$ such that $n > N_{\delta,1}$ implies that

$$\text{Prob}\left(||J_n(\hat{\boldsymbol{\theta}}_n)^{-1}||_F < 2||I(\boldsymbol{\theta}_0)^{-1}||_F\right) > 1 - \delta \tag{10}$$

Since (see, for example, Theorem 3.1.3 of Dennis and Schnabel (1983))

$$||M_{s\times s}||_F/\sqrt{s} \leq ||M_{s\times s}||_2 \leq ||M_{s\times s}||_F$$

where $||M||_2$ denotes the l_2 induced matrix norm of M (see, for example, pages 43 and 44 of Dennis and Schnabel), result 10 implies that for $n > N_{\delta,1}$

$$\text{Prob}\left(||J_n(\hat{\boldsymbol{\theta}}_n)^{-1}||_F < 2\sqrt{s}||I(\boldsymbol{\theta}_0)^{-1}||_2\right) > 1 - \delta$$

or

$$\text{Prob}\left(||J_n(\hat{\boldsymbol{\theta}}_n)^{-1}||_F < \beta\right) > 1 - \delta \tag{11}$$

where $\beta \equiv 2\sqrt{s}/\lambda$ and λ is the smallest eigenvalue of $I(\boldsymbol{\theta}_0)$.

Let $r > 0$ be such that $D(\boldsymbol{\theta}_0; 2r) \subset T$, the open neighborhood of $\boldsymbol{\theta}_0$ in assumptions **(A)** through **(D)**. (Here, $D(\boldsymbol{\theta}_0; 2r)$ denotes the open ball of radius $2r$ centered at $\boldsymbol{\theta}_0$.) Since (result 2) $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, given any $\delta > 0$, we can find an $N_{\delta,2}$ such that $n > N_{\delta,2}$ implies that $\text{Prob}(\hat{\boldsymbol{\theta}}_n \in D(\boldsymbol{\theta}_0; r)) > 1 - \delta$.

Now, provided that $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in D(\boldsymbol{\theta}_0; 2r)$,

$$||J_n(\boldsymbol{\theta}_1) - J_n(\boldsymbol{\theta}_2)||_F = ||[a_{lm}]_{s\times s}||_F$$

where

$$a_{lm} \equiv \sum_{i=1}^{n} \left(\frac{\partial^2 \ln f(X_i; \boldsymbol{\theta})}{\partial\theta_l \partial\theta_m}|_{\boldsymbol{\theta}_1} - \frac{\partial^2 \ln f(X_i; \boldsymbol{\theta})}{\partial\theta_l \partial\theta_m}|_{\boldsymbol{\theta}_2}\right) / n$$

$$= \sum_{i=1}^{n} \left(\frac{\partial^3 \ln f(X_i; \boldsymbol{\theta})}{\partial\theta_l \partial\theta_m \partial\theta_1}, \ldots, \frac{\partial^3 \ln f(X_i; \boldsymbol{\theta})}{\partial\theta_l \partial\theta_m \partial\theta_s}\right) |_{\boldsymbol{\theta}^*_{lm,n}} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)/n$$

where $\boldsymbol{\theta}^*_{lm,n}$ lies on the line segment between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

Thus, by assumption **(D)**, if $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ are within r of $\hat{\boldsymbol{\theta}}_n$, then for $n > N_{\delta,2}$, with probability greater than $1 - \delta$, we have

$$||J_n(\boldsymbol{\theta}_1) - J_n(\boldsymbol{\theta}_2)||_F^2 = \sum_{l=1}^{s}\sum_{m=1}^{s} a_{lm}^2 \leq \left(\sum_{l=1}^{s}\sum_{m=1}^{s} |a_{lm}|\right)^2 \tag{12}$$

$$\leq \left(\sum_{l=1}^{s}\sum_{m=1}^{s}\sum_{i=1}^{n}\sum_{p=1}^{s} (M_{lmp}(X_i)/n) \, |\theta_{p1} - \theta_{p2}|\right)^2$$

Since (by assumption **(D)**)

$$\sum_{i=1}^{n} M_{lmp}(X_i)/n \stackrel{a.s.}{\rightarrow} m_{lmp} < \infty$$

for $l, m, p \in \{1, \ldots, s\}$, result 12 implies that given any $\delta > 0$, we can find an $N_{\delta,3}$ such that $n > N_{\delta,3}$ implies that, if $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ are within r of $\hat{\boldsymbol{\theta}}_n$, then with probability greater than $1 - \delta$,

$$||J_n(\boldsymbol{\theta}_1) - J_n(\boldsymbol{\theta}_2)||_F^2 \leq \left(\sum_{l=1}^{s} \sum_{m=1}^{s} \sum_{p=1}^{s} (m_{lmp} + 1)|\theta_{p1} - \theta_{p2}| \right)^2$$

$$\leq \gamma^2 ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2$$

where

$$\gamma \equiv \left(\sum_{l=1}^{s} \sum_{m=1}^{s} \sum_{p=1}^{s} (m_{lmp} + 1) \right) < \infty$$

That is, for $n > N_{\delta,3}$, with probability greater than $1 - \delta$,

$$J_n \in \mathrm{Lip}_{\gamma}(D(\hat{\boldsymbol{\theta}}_n, r)) \tag{13}$$

Results 11 and 13 permit us to invoke Dennis and Schnabel's (1983) Theorem 5.2.1 to conclude that given any $\delta > 0$, we can find an $N_{\delta,4}$ such that $n > N_{\delta,4}$ implies that with probability greater than $1 - \delta$

$$\boldsymbol{\theta}_{n,\mathrm{Newt}} \equiv - \left[\frac{\partial^2 \ln L}{\partial\theta_l \partial\theta_m} \right]^{-1} |_{\boldsymbol{\theta}_{n,c}} \begin{pmatrix} \partial \ln L/\partial\theta_1 \\ \vdots \\ \partial \ln L/\partial\theta_s \end{pmatrix} |_{\boldsymbol{\theta}_{n,c}} + \boldsymbol{\theta}_{n,c}$$

is well-defined (that is the partials exist and the matrix is invertible), and

$$||\boldsymbol{\theta}_{n,\mathrm{Newt}} - \hat{\boldsymbol{\theta}}_n|| \leq \beta \times \gamma \times ||\boldsymbol{\theta}_{n,c} - \hat{\boldsymbol{\theta}}_n||^2 \tag{14}$$

provided that

$$||\boldsymbol{\theta}_{n,c} - \hat{\boldsymbol{\theta}}_n|| < \epsilon \equiv \min(r, \frac{1}{2\beta\gamma})$$

But by result 2 and the fact that $\boldsymbol{\theta}_{n,c}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$, we have $\sqrt{n}(\boldsymbol{\theta}_{n,c} - \hat{\boldsymbol{\theta}}_n) = \mathbf{O}_p(1)$ so given any $\delta > 0$ we can find a K_δ and an $N_{\delta,5}$ such that $n > N_{\delta,5}$ implies

$$\mathrm{Prob}(\sqrt{n}||\boldsymbol{\theta}_{n,c} - \hat{\boldsymbol{\theta}}_n|| \leq K_\delta) \geq 1 - \delta \tag{15}$$

If we require that $N_{\delta,5} > K_\delta^2/\epsilon^2$, then $n > N_{\delta,5}$ also implies

$$\mathrm{Prob}(||\boldsymbol{\theta}_{n,c} - \hat{\boldsymbol{\theta}}_n|| < \epsilon) \geq 1 - \delta \tag{16}$$

Results 14, 15, and 16 imply that given any $\delta > 0$, we can find an N such that $n > N$ implies that with probability greater than $1 - \delta$,

$$||\boldsymbol{\theta}_{n,\text{Newt}} - \hat{\boldsymbol{\theta}}_n|| \leq \beta \times \gamma \times K_\delta^2/n \tag{17}$$

which completes the proof of the theorem. ■

Corollary

Assume that conditions **(A0)** through **(D)** hold. Let $\boldsymbol{\theta}_{n,c}$ be a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$. That is, assume that

$$\sqrt{n}(\boldsymbol{\theta}_{n,c} - \boldsymbol{\theta}_0) = \mathbf{O}_p(1) \tag{18}$$

Define $\boldsymbol{\theta}_{n,0,\text{Newt}} \equiv \boldsymbol{\theta}_{n,c}$. Then with probability approaching one as $n \to \infty$, the kth Newton estimator,

$$\boldsymbol{\theta}_{n,k,\text{Newt}} \equiv - \left[\frac{\partial^2 \ln L}{\partial\theta_l \partial\theta_m}\right]^{-1}_{s\times s} |_{\boldsymbol{\theta}_{n,k-1,\text{Newt}}} \begin{pmatrix} \partial \ln L/\partial\theta_1 \\ \vdots \\ \partial \ln L/\partial\theta_s \end{pmatrix} |_{\boldsymbol{\theta}_{n,k-1,\text{Newt}}} + \boldsymbol{\theta}_{n,k-1,\text{Newt}} \tag{19}$$

is well-defined (that is the partials exist and the matrix is invertible), and

$$\boldsymbol{\theta}_{n,k,\text{Newt}} - \hat{\boldsymbol{\theta}}_n = \mathbf{O}_p(n^{-(2^{k-1})}) \tag{20}$$

where $\hat{\boldsymbol{\theta}}_n$ is a consistent solution of the likelihood equations guaranteed by Lehmann's theorem.

Proof

The proof is essentially the same as that of the main theorem. We simply replace $\boldsymbol{\theta}_{n,c}$ with $\boldsymbol{\theta}_{n,k-1,\text{Newt}}$ and $\boldsymbol{\theta}_{n,\text{Newt}}$ with $\boldsymbol{\theta}_{n,k,\text{Newt}}$. Result 17 then becomes

$$||\boldsymbol{\theta}_{n,k,\text{Newt}} - \hat{\boldsymbol{\theta}}_n|| \leq \beta \times \gamma \times K_\delta^2/(n^{2^{k-2}} \times n^{2^{k-2}}) = \beta \times \gamma \times K_\delta^2/n^{2^{k-1}}$$

■

REFERENCES

Dennis, J.E. and Schnabel, R.B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, New Jersey.

Janssen, P., Jureckova, J., and Veraverbeke, N. (1985), "Rate of Convergence of One- and Two-Step M-Estimators with Applications to Maximum Likelihood and Pitman Estimators," *The Annals of Statistics*, **13**, 1222–1229.

Lehmann, E.L. (1983), *Theory of Point Estimation*, John Wiley, New York.

Lehmann, E.L. and Romano, J.P. (2005), *Testing Statistical Hypotheses, Third Edition*, Springer, New York.

Verrill, S. and Johnson, R.A. (2007), "Confidence Bounds and Hypothesis Tests for Normal Distribution Coefficients of Variation," USDA Forest Products Laboratory Research Paper FPL-RP-638.