Asymptotic Properties of Extended Least Squares Estimators

Bradley M. Bell and Alan Schumitzky

University of Washington, Seattle

September 11, 1997

Abstract. We analyze the asymptotic properties of estimators based on optimizing an extended least squares objective function. This corresponds to maximum likelihood estimation when the measurements are normally distributed. These estimators are used in models where there are unknown parameters in both the mean and variance of measurements. Our approach is based on the analysis of optimization estimators. We prove consistency and asymptotic normality under the general conditions of independent, but not necessarily identically distributed, measurement data. Asymptotic covariance formulas are derived for the cases where the data are both normally and arbitrarily distributed.

Keywords: asymptotic normality, consistency, M-estimators, optimization estimators,u pharmacokinetic models

1 Introduction

In this paper we consider the asymptotic properties of estimators based on optimizing an extended least squares (ELS) objective function. Such estimators arise naturally in the method of the maximum likelihood and its variants. Our approach is based on the analysis of optimization estimators used by White (1994). Consistency and asymptotic normality are proved under general conditions of independent but not identically distributed measurement data. Formulas for the asymptotic covariance are derived for the cases where the data are both normally and arbitrarily distributed. The study is motivated by a class of nonlinear regression problems where there are unknown parameters in both the mean and variance models; see Beal and Sheiner (1988). Optimization estimators minimize an objective function that depends on random measurement data. Maximum likelihood (ML) estimators are a well known example in this class. The asymptotic properties of optimization estimators have been studied under a variety of conditions. Huber (1967) considered the independent and identically distributed measurement case and defined the class of M-estimators. Ljung and Caines, on the other hand, studied the dependent measurement case (Caines, 1988). Gallant (1987) considered general optimization estimators for both independent and dependent measurement data. Our approach uses the machinery developed by White (1994) for general optimization estimators. The ELS objective function, $L_N(\theta, y)$, is defined by

$$q_j(\theta, y_j) = (1/2)\,[y_j - S_j(\theta)]^T V_j(\theta)^{-1}[y_j - S_j(\theta)] + (1/2)\log\det V_j(\theta) \tag{1}$$

$$L_N(\theta, y) = \sum_{j=1}^{N} q_j(\theta, y_j)\ , \tag{2}$$

where $y = (y_1, y_2, \ldots)$ is a sequence of independent but not identically distributed measured column vectors, θ is a column vector belonging to a compact set Θ, and S_j and V_j are smooth functions defined on Θ. The value $S_j(\theta)$ has the same dimension as y_j, and the value $V_j(\theta)$ is a positive definite matrix with the same number of rows as y_j. An ELS estimator is any minimizer of the ELS objective with respect to $\theta \in \Theta$. The dimension of y_j may vary with j, which enables us to include the class of problems occurring with repeated measurement data; see Vonesh and Chinchilli (1997) or Davidian and Giltinan (1995). If there is an unknown parameter value $\theta_0 \in \Theta$ such that y_j is normally distributed with mean $S_j(\theta_0)$ and variance $V_j(\theta_0)$ for all j, then, up to an additive constant, $L_N(\theta, y)$ is the negative log likelihood of $(y_1, \ldots, y_N)$. In this case, under general hypotheses, it has been shown that the maximum likelihood estimate of θ_0 is consistent and asymptotically normal. See, for example, Hoadley (1971) or Philippou and Roussas (1975). In this paper, we do not make the normality assumption on y_j. We show, under general hypotheses, that the ELS estimator is still consistent and asymptotically normal. However, in this more general setting, the point θ_0 must satisfy a further identifiability condition. Our proof of consistency is similar to the consistency proof given by Bell, Burke, and Schumitzky (1996). Our proof of asymptotic normality follows the general template of White (1994). The special case where each y_j is a scalar, $\theta \equiv (x, u)$, and $V(\theta) \equiv V(S(x), u)$ is considered by Bell and Schumitzky (1997), where an extension of the Gauss-Newton method that minimizes $L_N(\theta, y)$ with respect to θ is presented. A final remark is of note. Hoadley (1971) and Philippou and Roussas (1975) established consistency and asymptotic normality for ML estimators in the general setting of independent but not identically distributed random measurement data. The methods of proof of these two works were general enough to actually include estimators based on objective functions other than the likelihood function, namely the optimization estimators. In a technical report, Beal (1984) defined the class of ELS problems and used the method of Hoadley (1971) to derive asymptotic properties of ELS estimators. Our approach is closer to the method of Philippou and Roussas (1975). When the objective function is suitably smooth it appears that this method is considerably simpler than that of Hoadley. A brief outline of the paper follows: In Sections 2 and 3 we define the basic notation and assumptions of the paper. In Section 4, we state the main theorems of consistency and asymptotic normality. We also state the formula for the asymptotic covariance. Proofs are given in Sections 5, 6, 7, and 8.

2 Notation

$L_N(\theta, y)$	extended least squares objective function (see Equation (2))
$q_j(\theta, y_j)$	jth term in the objective function (see Equation (1))
θ_0	true, but unknown, value for the parameter vector
$\widehat{\theta}_N$	the value of θ that minimizes $L_N(\theta, y)$, more precicely written $\widehat{\theta}_N(\omega)$
Θ	compact subset of finite dimensional Euclidean space
y_j	the jth measurement vector, more precisel written $y_j(\omega)$
$S_j(\theta)$	model for the mean of y_j
$V_j(\theta)$	model for the variance of y_j
$\|u\|$	square root of the sum of the squares of the elements of u
$E[g]$	expected value of $g(\omega)$ with respect to $\omega \in \Omega$
u^T	transpose of u
Ω	set of points in the probability space
ω	an element of the probability space
B	the sigma field of measurable sets of Ω
P	the probability measure on Ω
$\partial h(\theta)$	the derivative of h with respect to θ
$\partial^2 h(\theta)$	the second derivative of h with respect to θ
$\partial_k h(\theta)$	the derivative of h with respect to the $k-$th element of θ
$\\|f(\theta)\\|$	maximuim of $\|f(\theta)\|$ with respect to $\theta \in \Theta$
Σf_j	the sum from $j = 1$ to $j = N$ of f_j
$u_N \rightarrow u_0$	the sequence $\{u_N\}$ converges to u_0 as $N \rightarrow \infty$
sqrt(x)	square root of the value x

3 Assumptions

1. The elements of the sequence $\{y_j\}$ are independent random column vectors defined on the complete probability space (Ω, B, P) and there is a constant M such that for all j,

$$E[|y_j|^6] \leq M \ .$$

2. The column vector valued functions $\{S_j(\theta)\}$ and the positive definite matrix valued functions $\{V_j(\theta)\}$ are three times continuously differentiable on the compact space Θ such that there is a $\theta_0 \in \Theta$ with $E[y_j] = S_j(\theta_0)$ and $Var[y_j] = V_j(\theta_0)$. In addition there is a constant M such that for $i = 0, 1, 2$ and for all j

$$\left\|\partial^i S_j(\theta)\right\| \leq M, \left\|\partial^i V_j(\theta)\right\| \leq M, \text{ and } \left\|V_j(\theta)^{-1}\right\| \leq M \ .$$

3. There is a function $L(\theta)$ defined on Θ such that

$$\|L(\theta) - (1/N)\, E[L_N(\theta, y)]\| \to 0 \ .$$

In addition, θ_0 is in the interior of Θ, and it is the unique minimizer of $L(\theta)$ on Θ.

4. There is a matrix valued function $C(\theta)$ defined on Θ such that

$$\left\|C(\theta) - (1/N)\, E[\partial^2 L_N(\theta, y)]\right\| \to 0 \ .$$

In addition, $C(\theta_0)$ is positive definite.

5. There is a positive definite matrix D such that

$$(1/N)\, E[\partial L_N(\theta_0, y)^T \partial L_N(\theta_0, y)] \to D \ .$$

Given these assumptions it is shown by White (1994, Theorem 2.12) that there exists a measurable function $\widehat{\theta}_N(\omega)$ such that

$$L_N[\widehat{\theta}_N(\omega), y(\omega)] = \min_{\theta \in \Theta} L_N[\theta, y(\omega)] \ . \tag{3}$$

Under very general conditions, we prove that for almost all ω, $\widehat{\theta}_N(\omega)$ converges to θ_0 and that the sequence sqrt(N)($\widehat{\theta}_N(\omega) - \theta_0$) is asymptotically normal. The crux of our proofs is based on a uniform version of the strong law of large numbers. If all of the y_j are normally distributed, then $L_N(\theta,\ y)$ is the negative log likelihood function of the data (up to an additive constant). In this case, $\widehat{\theta}_N$ is the maximum likelihood estimate of θ_0 given the data $(y_1, \ldots, y_N)$. In this paper we do not assume normality, but we do point out special results for that case.

4 Consistency and Asymptotic Normality

In this section we state our main results. The proofs are given in Sections 5, 6, 7, and 8. The first theorem provides motivation for assuming that θ_0 is the unique minimizer of $L(\theta)$. The second theorem establishes that the estimates converge to the true parameter value. The third theorem establishes the asymptotic normality of the estimates. The last theorem provides a formula for calculating the covariance of the estimates.

Theorem 1. *If Assumption 2 is satisfied,*

$$E[L_N(\theta_0, y)] = \min_{\theta \in \Theta} E[L_N(\theta, y)] \ .$$

Theorem 2. *Suppose all the assumptions hold and* $\widehat{\theta}_N$ *is defined by Equation (3). It follows that for almost all* ω

$$\widehat{\theta}_N(\omega) \to \theta_0 \ .$$

Theorem 3. *Suppose all the assumptions hold and* $\widehat{\theta}_N$ *is defined by equation 3. It follows that the random column vector* $\mathrm{sqrt}(N)[\ \widehat{\theta}_N(\omega) - \ \theta_0]$ *converges in distribution to a normal random column vector with mean zero and covariance*

$$C(\theta_0)^{-1} D C(\theta_0)^{-1} \ .$$

In addition, if each y_j *is normally distributed,* $D = C(\theta_0)$*.*

Theorem 4. $E[\partial_m \partial_k L_N(\theta_0, y)]$ *is equal to*

$$\Sigma \partial_m S_j(\theta_0)^T V_j(\theta_0)^{-1} \partial_k S_j(\theta_0) + (1/2)\, \mathrm{trace}[\mathrm{V_j}(\theta_0)^{-1} \partial_\mathrm{m} \mathrm{V_j}(\theta_0) \mathrm{V_j}(\theta_0)^{-1} \partial_\mathrm{k} \mathrm{V_j}(\theta_0)] \ .$$

Remark 1. The result in Theorem 4 is known in the case where the elements of $\{y_j\}$ are scalar-valued measurements (Beal and Sheiner (1988), Section 2.6). It is less well known in the vector valued measurement case (Vonesh and Chinchilli (1997), Equation 9.2.24). We must approximate $C(\theta_0)$ by evaluating the expressions for $E[\partial_k \partial_m L_N(\theta_0, y)]$ with θ_0 replaced by $\widehat{\theta}_N$ because θ_0 is unknown. This is justified by that fact that $E[\partial_k \partial_m L_N(\theta, y)]$ is continuous and $\widehat{\theta}_N \to \theta_0$.

5 Proof of Theorem 1

Lemma 5. *Suppose u_0 is a column vector of length n and U_0 is an $n \times n$ symmetric positive definite matrix. For each column vector u and positive definite matrix U define*

$$H(u, U) = \mathrm{trace}[U^{-1}U_0] + \log\det(U) + (u - u_0)^T U^{-1}(u - u_0) \ .$$

The function $H(u, U)$ has a unique minimum at $(u, U) = (u_0, U_0)$.

Proof. Given a positive definite U, it follows that U^{-1} is positive definite and

$$\min_u H(u, U) = H(u_0, U) = \mathrm{trace}[U^{-1}U_0] + \log\det(U) \ .$$

Thus it suffices to show that U_0 minimizes $H(u_0, U)$ with respect to U. Let $F(U)$ be the logarithm of the determinant of U and let $\bullet$ denote the Frobenious inner product of matrices, i.e., the sum of the element-by-element product. It follows that $F(U)$ is concave, its derivative is U^{-1}, and

$$\begin{aligned}
F(U_0) &\leq F(U) + U^{-1} \bullet (U_0 - U) = F(U) + \mathrm{trace}[U^{-1}(U_0 - U)] \\
\log\det(U_0) &\leq \log\det(U) + \mathrm{trace}[U^{-1}(U_0 - U)] = \log\det(U) + \mathrm{trace}[U^{-1}U_0] - n \\
H(u_0, U_0) &\leq H(u_0, U) \ .
\end{aligned}$$

Lemma 6. *Suppose that u is a random column vector with mean u_0 and variance U_0, and w is a constant column vector. It follows that*

$$E[(u - w)(u - w)^T] = U_0 + (u_0 - w)(u_0 - w)^T \ .$$

Proof. $(u - w)(u - w)^T$ is equal to

$$(u - u_0)(u - u_0)^T + (u_0 - w)(u - u_0)^T + (u - u_0)(u_0 - w)^T + (u_0 - w)(u_0 - w)^T \ .$$

Taking the expected value of the expression above we obtain the conclusion of this lemma.

Lemma 7. *If Assumption 2 is satisfied, the argument θ_0 minimizes $E[q_j(\theta, y_j)]$ subject to $\theta \in \Theta$.*

Proof. Define $F_j(\theta) = E[2q_j(\theta, y_j)]$ which is equal to

$$
\begin{aligned}
& E\{[y_j - S_j(\theta)]^T V_j(\theta)^{-1}[y_j - S_j(\theta)] + \log\det V_j(\theta)\} \\
= \quad & \mathrm{traceE}\{\mathrm{V_j}(\theta)^{-1}[\mathrm{y_j} - \mathrm{S_j}(\theta)][\mathrm{y_j} - \mathrm{S_j}(\theta)]^{\mathrm{T}}\} + \log\det \mathrm{V_j}(\theta) \ .
\end{aligned}
$$

Applying Lemma 6 and the fact that trace(AB) is equal to trace(BA), we obtain

$$
\begin{aligned}
F_j(\theta) - \log\det V_j(\theta) &= \mathrm{trace}[\mathrm{V_j}(\theta)^{-1}\{\mathrm{V_j}(\theta_0)^{-1} + [\mathrm{S_j}(\theta_0) - \mathrm{S_j}(\theta)][\mathrm{S_j}(\theta_0) - \mathrm{S_j}(\theta)]^{\mathrm{T}}\}] \\
&= \mathrm{trace}[\mathrm{V_j}(\theta)^{-1}\mathrm{V_j}(\theta_0)^{-1}] + [\mathrm{S_j}(\theta_0) - \mathrm{S_j}(\theta)]^{\mathrm{T}}\mathrm{V_j}(\theta)^{-1}[\mathrm{S_j}(\theta_0) - \mathrm{S_j}(\theta)] \ .
\end{aligned}
$$

It now follows from Lemma 5 that θ_0 minimizes $F_j(\theta)$, which completes the proof of this lemma. It follows from Lemma 7 that θ_0 minimizes each of the terms in the summation $(1/2)\,\Sigma E[q_j(\theta, y_j)]$, which is equal to $E[L_N(\theta_0, y)]$. Thus θ_0 minimizes $E[L_N(\theta_0, y)]$ with respect to $\theta \in \Theta$. This completes the proof of Theorem 1.

Remark 2. Theorem 1 provides motivation for the statement that θ_0 minimizes $L(\theta)$ in Assumption 3. In addition, if the set of equations $S_j(\theta) = S_j(\theta_0)$ and $V_j(\theta) = V_j(\theta_0)$ for $j = 1, \ldots, N$ has the unique solution $\theta = \theta_0$, then θ_0 is the only minimizer of $E[L_N(\theta, y)]$.

6 Proof of Theorem 2

Definition. Given a sequence of matrix valued functions $\{f_j(\theta, \omega)\}$, each of which is defined on $\Theta \times \Omega$, let

$$
h_N(\theta, \omega) = (1/N)\Sigma\{f_j(\theta, \omega) - E[f_j(\theta, \omega)]\} \ .
$$

The sequence $\{f_j(\theta, \omega)\}$ satisfies the *pointwise strong law of large numbers* if for each $\theta \in \Theta$ and almost all $\omega \in \Omega$, $|h_N(\theta, \omega)| \to 0$. The sequence $\{f_j(\theta, \omega)\}$ satisfies the *uniform strong law of large numbers* if for almost all $\omega \in \Omega$, $\|h_N(\theta, \omega)\| \to 0$. Note that if a function does not depend on θ, the pointwise and uniform strong laws of large numbers are equivalent for the sequence. The following lemma is a special case of Andrews (1992, Theorem 3):

Lemma 8. *Suppose Θ is a compact subset of a real vector space, the sequence of Borel measurable vector valued functions $\{f_j(\theta, \omega)\}$ and scalar valued functions $\{B_j(\omega)\}$ satisfy the pointwise strong law of large numbers, and there is constant M such that for each θ_1, $\theta_2 \in \Theta$ and almost all $\omega \in \Omega$ and all j we have $E[B_j(\omega)] \leq M$, and*

$$
|f_j(\theta_1, \omega) - f_j(\theta_2, \omega)| \leq B_j(\omega)|\theta_1 - \theta_2| \ .
$$

It follows that the sequence of functions $\{f_j(\theta,\omega)\}$ satisfies the uniform strong law of large numbers.

Proof. Replacing the index t by the index j and the random variable $Z_t(\omega)$ by ω in the statement of Andrews (1992, Theorem 3 Part b) and noting that for all N

$$(1/N)\,\Sigma E[B_j(\omega)] \leq M \ ,$$

we obtain the conclusion of this lemma.

Lemma 9. *Suppose that $\{z_j(\omega)\}$ is a sequence of independent random variables, that the elements of $\{f_j(\theta,z_j)\}$ are Borel measurable column vector valued functions, that the elements of $\{B_j(z_j)\}$ are Borel measurable scalar valued functions, and that for each θ,*

$$E[|f_j(\theta,z_j)|^2] \leq M, \quad \text{and} \quad E[B_j(z_j)^2] \leq M \ .$$

In addition, suppose there is a constant M such that for all θ_1, $\theta_2 \in \Theta$ and almost all $\omega \in \Omega$

$$|f_j(\theta_1,z_j) - f_j(\theta_2,z_j)| \leq B_j(z_j)|\theta_1 - \theta_2| \ .$$

It follows that the sequence of functions $\{f_j(\theta,z_j)\}$ satisfies the uniform strong law of large numbers.

Proof. It follows from Chung (1968, Theorem 3.3.1) that the sequence $\{B_j(z_j)\}$ is independent because the sequence $\{z_j\}$ is independent and $\{B_j(z_j)\}$ are Borel measurable functions. This sequence of functions is uncorrelated and according to Chung (1968, Theorem 5.1.2), for almost all ω, $\{B_j(z_j)\}$ satisfies the pointwise strong law of large numbers. In a similar fashion, for a fixed θ and almost all ω, $\{f_j(\theta,z_j)\}$ satisfies the pointwise strong law of large numbers. By partitioning Ω into

$$\Omega = \{\omega \in \Omega : B_j[z_j(\omega)] \leq 1\} \cup \{\omega \in \Omega : B_j[z_j(\omega)] > 1\} \ ,$$

we conclude

$$E[B_j(z_j)] \leq 1 + E[B_j(z_j)^2] \leq 1 + M \ .$$

The conclusion of the lemma now follows from the previous lemma.

Lemma 10. *Suppose that $\{z_j\}$ is a sequence of independent column vector valued random variables, $\{g_j(\theta)\}$ is a sequence of column vector valued continuously differentiable function on the compact space Θ (such that the vectors z_j and $g_j(\theta)$ have the same length), and there is a constant M such that for all j and all $\theta \in \Theta$,*

$E[|z_j|^2] \le M$, $|g_j(\theta)| \le M$, and $|\partial g_j(\theta)| \le M$. If $f_j(\theta, z_j)$ is defined to be $(z_j)^T g_j(\theta)$, the sequence $\{f_j(\theta, z_j)\}$ satisfies the uniform strong law of large numbers.

Proof. We prove this lemma by verifying the conditions of the previous lemma. The first condition follows from

$$E[|f_j(\theta, z_j)|^2] = E[|(z_j)^T g_j(\theta)|^2] \le E[|z_j|^2]|g_j(\theta)|^2 \le M^3 \ .$$

For the second condition we define $B_j(z_j)$ to be $|z_j|M$ and obtain

$$E[B_j(z_j)^2] = E[|z_j|^2 M^2] \le M^3 \ .$$

The third condition is obtained as follows

$$|f_j(\theta_1, z_j) - f_j(\theta_2, z_j)| = |z_j^T(g_j(\theta_1) - g_j(\theta_2))| \le |z_j| \, \|\partial g_j(\theta)\| \, |\theta_1 - \theta_2| \le B_j(z_j)|\theta_1 - \theta_2| \ .$$

Note that we have used the fact that $|\partial g_j(\theta)|$ is greater than or equal to the operator norm of $\partial g_j(\theta)$. This completes the proof of this lemma.

Lemma 11. *If all the assumptions hold and $\{q_j(\theta, y_j)\}$ is defined by Equation (1), then $\{q_j(\theta, y_j)\}$, $\{\partial q_j(\theta, y_j)\partial q_j(\theta, y_j)^T\}$, and $\{\partial^2 q_j(\theta, y_j)\}$ satisfy the uniform strong law of large numbers.*

Proof. We show that the lemma is true for the sequence of functions $\{q_j(\theta, y_j)\}$. The other sequences have very similar proofs but involve more calculations.

$$q_j(\theta, y_j) = (1/2)\, y_j^T V_j(\theta)^{-1} y_j - y_j^T V_j(\theta)^{-1} S_j(\theta) + (1/2)\, S_j(\theta)^T V_j(\theta)^{-1} S_j(\theta) + \log\det V_j(\theta) \ .$$

Define the functions $f_j(\theta)$, $g_j(\theta, y_j)$, and $h_j(\theta, y_j)$ by

$$\begin{aligned} f_j(\theta) &= (1/2)\, S_j(\theta)^T V_j(\theta)^{-1} S_j(\theta) + \log\det V_j(\theta) \\ g_j(\theta, y_j) &= y_j^T V_j(\theta)^{-1} S_j(\theta) \\ h_j(\theta, y_j) &= (1/2)\, y_j^T V_j(\theta)^{-1} y_j \ . \end{aligned}$$

The assumptions ensure that the sequences $\{f_j(\theta)\}$ and $\{g_j(\theta, y_j)\}$ satisfy the conditions of the previous lemma (with $z_j \equiv 1$ and $z_j = y_j$ respectively). Thus it suffices to show that the sequence $\{h_j(\theta, y_j)\}$ also satisfies the conditions of the previous lemma to complete this proof. This follows from the following equalities:

$$h_j(\theta, y_j) = (1/2)\, y_j^T V_j(\theta)^{-1} y_j = (1/2)\, (y_j^T \otimes y_j^T)\mathrm{vec}[\mathrm{V_j}(\theta)^{-1}] \ ,$$

where $\otimes$ is the Kronker product and $\mathrm{vec}[\mathrm{V_j}(\theta)^{-1}]$ is the column vector consisting of the first row of $V_j(\theta)^{-1}$followed by its second row and so on. In addition

$$E(|y_j^T \otimes y_j^T|^2) \leq E(n|y_j|^4) \leq n[1 + E(|y_j|^6)] \leq n\,(M+1) \ ,$$

where n is the number of elements in the column vector y_j. Thus the sequence $\{h_j(\theta, y_j)\}$ also satisfies the conditions of the previous lemma.

Lemma 12. *If all the assumptions hold, for almost all $\omega \in \Omega$, the sequences $\{L_N(\theta, y)\}$ and $\{\partial^2 L_N(\theta, y)\}$ satisfy the uniform strong law of large numbers.*

Proof. This follows directly from the previous lemma and the definition of $L_N(\theta, y)$ in Equation (2). Assumption 3 and the Lemma 12 imply that

$$\|(1/N)\,\Sigma L_N(\theta, y) - (1/N)\,E[L_N(\theta, y)]\| \to 0, \text{ and } \|L(\theta) - (1/N)\,E[L_N(\theta, y)]\| \to 0$$

and that θ_0 is the unique minimizer of $L(\theta)$ on Θ. Uniform convergence implies epi-convergence (Wets (1980), Theorem 4) Thus, for almost all $\omega \in \Omega$, the minimizers of $L_N(\theta, y) \to \theta_0$. This completes the proof of Theorem 2.

7 Proof of Theorem 3

The following lemma is a special case of Chung's (1968) Theorem 7.1.2.

Lemma 13. *Suppose that $\{z_j(\omega)\}$ is a sequence of scalar valued mean zero independent random variables and define*

$$b_N = \mathrm{sqrt}(\Sigma \mathrm{E}[\mathrm{z_j^2}]), \quad \mathrm{S_N} = \mathrm{b_N^{-1}}\Sigma \mathrm{z_j}, \quad \text{and} \quad \Gamma_\mathrm{N} = \mathrm{b_N^{-3}}\Sigma \mathrm{E}[|\mathrm{z_j}|^3] \ .$$

If $\Gamma_N \to 0$, S_N converges in distribution to a normal random variable with mean zero and variance one.

Lemma 14. *Suppose that $\{z_j(\omega)\}$ is a sequence of scalar valued mean zero independent random variables and there is an $\alpha > 0$ and an M such that*

$$(1/N)\,\Sigma E[z_j^2] \to \alpha, \text{ and } E[|z_j|^3] \leq M \text{for all } j \ .$$

It follows that $\mathrm{sqrt}(1/N)\Sigma z_j$ *converges in distribution to a normal random variable with mean zero and variance* α.

Proof. Let b_N, S_N, and Γ_N be as in the previous lemma. For N sufficiently large, $(1/N)\Sigma E(z_j^2) \geq (\alpha/2)$ and

$$\Gamma_N = b_N^{-3}\Sigma E(|z_j|^3) \leq [\Sigma E(z_j^2)]^{-3/2}[\Sigma E([|z_j|^3)] \leq (\frac{2}{N\alpha})^{3/2} NM \ .$$

Thus $\Gamma_N \to 0$, the previous lemma applies, and S_N converges in distribution to a normal random variable with mean zero and variance one. From the assumptions of this lemma $b_N \mathrm{sqrt}(N) \to \mathrm{sqrt}(\alpha)$. Thus by the corollary below Theorem 4.4.6 in Chung (1968), the sequence $\{(b_N S_N/\mathrm{sqrt(N)}\}$ converges to a normal random variable with mean zero and variance α. Substituting the definition of b_N and S_N completes the proof of this lemma.

Lemma 15. *If all the assumptions are satisfied, the sequence* $\{\mathrm{sqrt(1/N)\partial L_N(\theta_0, y)^T}\}$ *converges in distribution to a normal random column vector with mean zero and covariance* D.

Proof. Using that fact that $\partial_k \log\det[V_j(\theta)]$ is equal to $\mathrm{trace[V_j(\theta)^{-1}\partial_k V_j(\theta)]}$ and $\partial_k V_j(\theta)^{-1}$ is equal to $V_j(\theta)^{-1}\partial_k V_j(\theta)V_j(\theta)^{-1}$, it follows from Equation (1) that

$$
\begin{aligned}
\partial_k q_j(\theta, y_j) \quad &= \quad -[y_j - S_j(\theta)]^T V_j(\theta)^{-1} \partial_k S_j(\theta) + (1/2)\,\mathrm{trace}[\mathrm{V_j}(\theta)^{-1} \partial_\mathrm{k} \mathrm{V_j}(\theta)] \\
&\quad - (1/2)\,[y_j - S_j(\theta)]^T V_j(\theta)^{-1} \partial_k V_j(\theta) V_j(\theta)^{-1} [y_j - S_j(\theta)] \ .
\end{aligned}
$$

Using the fact that a scalar is equal to its trace and that the trace of AB is equal to the trace of BA, we conclude that

$$
\begin{aligned}
\partial_k q_j(\theta, y_j) \quad &= \quad -[y_j - S_j(\theta)]^T V_j(\theta)^{-1} \partial_k S_j(\theta) \qquad (4) \\
&- \quad (1/2)\mathrm{trace}\left(V_j(\theta)^{-1} \partial_k V_j(\theta) V_j(\theta)^{-1} \{ [y_j - S_j(\theta)][y_j - S_j(\theta)]^T - V_j(\theta) \} \right) \ .
\end{aligned}
$$

Substituting θ_0 for θ and taking the expected value, we obtain $E[\partial_k q_j(\theta_0, y_j)] = 0$. We need to prove the central limit theorem for the column vector valued functions $\{\partial q_j(\theta_0, y_j)^T\}$. We do this by defining the inner product with a fixed deterministic direction h as $z_j = \partial q_j(\theta_0, y_j)\ h$. The elements of $\{z_j\}$ are the mean zero independent random variables. In addition,

$$
\begin{aligned}
(1/N)\,\Sigma E[z_j^2] \quad &= \quad (1/N)\,E[\Sigma z_j^2] = (1/N)\,E[(\Sigma z_j)(\Sigma z_j)] \\
&= \quad (1/N)\,E[(\Sigma \partial q_j(\theta_0, y_j) h)^T (\Sigma \partial q_j(\theta_0, y_j) h)] \\
&= \quad (1/N)\,h^T E[\partial L_N(\theta_0, y)^T \partial L_N(\theta_0, y)] h \ ,
\end{aligned}
$$

which by Assumption 5 converges to $h^T D h$. From the definition of z_j we have

$$
E[|z_j|^3] = E[|\partial q_j(\theta_0, y_j) h|^3] \ .
$$

From the equation above for $\partial_k q_j(\theta, y_j)$, the fact that $S_j(\theta)$, $V_j(\theta)^{-1}$, $\partial_k S_j(\theta)$, and $\partial_k V_j(\theta)$ are uniformly bounded (Assumption 2), and the fact that $E[|y_j|^6]$ is uniformly bounded (Assumption 1), there is a constant M such that $E[|z_j|^3] \leq M$ for all j. (Note that $\partial_k q_j(\theta, y_j)$ contains second order terms in y_j; hence $|\partial q_j(\theta_0, y_j) h|^3$ contains sixth order terms in y_j.) Therefore the previous lemma applies and $\mathrm{sqrt(1/N)\Sigma z_j}$ converges in distribution to a normal random variable with mean zero and variance $h^T D h$. The conclusion of this lemma now follows from the observation that Σz_j is equal to $\partial L_N(\theta_0, y) h$ and the fact that h was arbitrary. The following result is a special case of Theorem 6 in White (1994).

Lemma 16. *Suppose (Ω, B, P) is a complete probability, Θ is a compact subset of R^n, θ_0 is in the interior of Θ, and $Q_N(\theta, \omega)$ is twice continuously differentiable in θ for almost all ω. Define $\widehat{\theta}_N(\omega)$ to be a minimizer* *of $Q_N(\theta,\omega)$ with respect to θ, and suppose that for almost all ω, $\widehat{\theta}_N(\omega) \to \theta_0$ and there is a deterministic positive definite matrix B such that*

$$B^{-1/2}\mathrm{sqrt(N)}\partial \mathrm{Q_N}(\theta_0,\omega)^{\mathrm{T}}$$

converges to a normal random column vector with mean zero and variance equal to the identity. In addition there is a continuous matrix valued function $A(\theta)$,

$$\left\|\partial^2 Q_N(\theta,\omega) - A(\theta)\right\| \to 0$$

for almost all ω where $A(\theta_0)$ is positive definite. It follows that

$$\mathrm{sqrt(N)}[\widehat{\theta}_{\mathrm{N}}(\omega) - \theta_0]$$

converges in distribution to a normal random column vector with mean zero and variance equal to $A(\theta_0)^{-1}BA(\theta_0)^{-1}$.

We are now ready to complete the proof of Theorem 3 by applying the lemma above with the following identifications:

$$Q_N(\theta,\omega) = (1/N)\, L_N[\theta, y(\omega)], \quad A(\theta) = C(\theta), \quad \text{and } B = D\ .$$

Note that by the previous lemma

$$B^{-1/2}\mathrm{sqrt(N)}\partial \mathrm{Q_N}(\theta_0,\omega) = \mathrm{D}^{-1/2}\mathrm{sqrt(1/N)}\partial \mathrm{L_N}(\theta_0,\omega)$$

converges to a normal random column vector with mean zero and variance equal to the identity. In addition, by Lemma 12 and Assumption 4 for almost all ω,

$$(1/N)\left\|\partial^2 L_N(\theta,y) - E[\partial^2 L_N(\theta,y)]\right\| \to 0 \text{ and } \left\|(1/N)\, E[\partial^2 L_N(\theta,y)] - C(\theta)\right\| \to 0\ .$$

It follows that

$$\left\|(1/N)\, \partial^2 L_N(\theta,y) - C(\theta)\right\| \to 0; \text{ i.e., } \left\|\partial^2 Q_N(\theta,\omega) - A(\theta)\right\| \to 0\ .$$

By the previous lemma $\mathrm{sqrt(N)}[\widehat{\theta}_{\mathrm{N}}(\omega) - \theta_0]$ converges in distribution to a normal random column vector with mean zero and variance equal to

$$A(\theta_0)^{-1}BA(\theta_0)^{-1} = C(\theta_0)^{-1}DC(\theta_0)^{-1}\ ,$$

which is the first conclusion in Theorem 3. The function $L_N(\theta,z)$ is a constant plus the negative log-likelihood of $\{z_1,\cdots,z_N\}$ under the assumption that each z_j is normally distributed with mean $S_j(\theta)$ and variance

$V_j(\theta)$. Hence there is a fixed constant K independent of θ such that

$$\int \exp[-L_N(\theta, z)] dz_1 \cdots dz_N = K \ .$$

Taking the second partial derivative of both sides with respect to θ and passing the derivative under the integral sign, we obtain

$$\int [\partial^2 L_N(\theta, z) - \partial L_N(\theta, z)^T \partial L_N(\theta, z)] \exp[-L_N(\theta, z)] dz_1 \cdots dz_N = 0 \ .$$

The interchange of differentiation and integration is valid because $S_j(\theta)$, $V_j(\theta)$ and $V_j(\theta)^{-1}$ and their first and second derivatives are uniformly bounded by Assumption 2 (note that this implies that the minimum eigenvalue of $V_j(\theta)^{-1}$is bounded below and hence the exponential term dominates in the integral). Splitting the integral and substituting θ_0 for θ, we obtain

$$\int \partial^2 L_N(\theta_0, z) \exp[-L_N(\theta_0, z)] dz = \int \partial L_N(\theta_0, z)^T \partial L_N(\theta_0, z) \exp[-L_N(\theta_0, z)] dz \ ,$$

where dz denotes $dz_1, \ldots, dz_N$. If each y_j is normally distributed, the term on the left is $E[\partial^2 L_N(\theta_0, y)]$ and the term on the right is $E\left[\partial L_N(\theta_0, y)^T \partial L_N(\theta_0, y)\right]$. This completes the proof of the second part of Theorem 3.

8 Proof of Theorem 4

By Equation (5) it follows that

$$\partial_k q_j(\theta, y_j) = -r_j(\theta)^T V_j(\theta)^{-1} \partial_k S_j(\theta) - (1/2)\, \mathrm{trace}\{\mathrm{W_{jk}}(\theta)[\mathrm{r_j}(\theta)\mathrm{r_j}(\theta)^\mathrm{T} - \mathrm{V_j}(\theta)]\} \ ,$$

where

$$r_j(\theta) = [y_j - S_j(\theta)], \quad \text{and} \quad W_{jk}(\theta) = V_j(\theta)^{-1} \partial_k V_j(\theta) V_j(\theta)^{-1} \ .$$

It follows that

$$\begin{aligned}
\partial_m \partial_k q_j(\theta, y_j) &= +\partial_m S_j(\theta)^T V_j(\theta)^{-1} \partial_k S_j(\theta) - r_j(\theta)^T \partial_m [V_j(\theta)^{-1} \partial_k S_j(\theta)] \\
&- (1/2)\, \mathrm{trace}\{\partial_\mathrm{m} \mathrm{W_{jk}}(\theta)[\mathrm{r_j}(\theta)\mathrm{r_j}(\theta)^\mathrm{T} - \mathrm{V_j}(\theta)]\} \\
&+ (1/2)\, \mathrm{trace}\{\mathrm{W_{jk}}(\theta)[\partial_\mathrm{m} \mathrm{S_j}(\theta)\mathrm{r_j}(\theta)^\mathrm{T} + \mathrm{r_j}(\theta)\partial_\mathrm{m}\mathrm{S_j}(\theta)^\mathrm{T} + \partial_\mathrm{m}\mathrm{V_j}(\theta)]\}
\end{aligned}$$

because $\partial_m r_j(\theta) = -\partial_m S_j(\theta)$. It follows that the expected value $E[\partial_m \partial_k q_j(\theta_0, y_j)]$ is

$$
\begin{aligned}
& \partial_m S_j(\theta_0)^T V_j(\theta_0)^{-1} \partial_k S_j(\theta_0) + (1/2)\,\mathrm{trace}[\mathrm{W_{jk}}(\theta_0)\partial_\mathrm{m} \mathrm{V_j}(\theta_0)] \\
= \quad & \partial_m S_j(\theta_0)^T V_j(\theta_0)^{-1} \partial_k S_j(\theta_0) + (1/2)\,\mathrm{trace}[\mathrm{V_j}(\theta_0)^{-1} \partial_\mathrm{m} \mathrm{V_j}(\theta_0) \mathrm{V_j}(\theta_0)^{-1} \partial_\mathrm{k} \mathrm{V_j}(\theta_0)]
\end{aligned}
$$

because $E[r_j(\theta_0)]$ is equal to zero and $E[r_j(\theta_0) r_j(\theta_0)^T]$ is equal to $V_j(\theta_0)$. The conclusion of the theorem follows from the formula

$$
L_N(\theta_0, y) = \Sigma q_j(\theta, y_j) \ .
$$

References

Andrews, D. W. K. (1992). Generic uniform convergence. *Econometric Theory*, **8**, 241–257.

Beal, S. (1984). Asymptotic properties of optimization estimators for the independent not identically distributed case with application to extended least squares estimators. Tech. Report of the Division of Clinical Pharmacology, Univ. of California, San Francisco.

Beal, S. L. and Sheiner, L. B. (1988). Heteroscedastic nonlinear regression. *Technometrics* **30**, 327–338.

Bell, B. M., Burke, J., and Schumitzky, A. (1996). A relative weighting method for estimating parameters and variances in multiple data sets. *Computational Statistics and Data Analysis* **22**, 119–135.

Bell, B. M. and Schumitzky, A. (1997). An algorithm that simultaneously fits mean and variance parameters in nonlinear models. Submitted to *SIAM J. Opt.*

Caines, P.E. (1988). *Linear Stochastic Systems.* Wiley, New York.

Chung, K. L. (1968). *A Course in Probability Theory*, 2nd edition. Harcourt, Brace and World, New York.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data.* Chapman and Hall, New York.

Gallant, A. R. (1987). *Nonlinear Statistical Models.* Wiley, New York.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Stat.* **42**, 1977–1991.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statis. and Prob.* **I**, 221–233, University of California Press, Berkeley, CA.

Philippou, A. N. and Roussas, G. G. (1975). Asymptotic normality of the maximum likelihood estimate in the independent not identically distributed case. *Annals of the Institute of Statistical Mathematics, Tokyo* **27**, 45–55.

Vonesh, E. F. and Chinchilli, V. M. (1997). *Linear and Nonlinear Models for Analysis of Repeated Measurements.* Marcel Dekker, New York.

Wets, R. J. B. (1980). Convergence of convex functions, variational inequalities and convex optimization problems. In: *Variational Inequalities and Complementarity Problems* (Edited by R.W. Cottle, F. Giannessi, and J.-L. Lions), 375–403. Wiley, New York.

White, H. (1994). *Estimation, Inference and Specification Analysis.* Cambridge University Press.