Chapter 5

Square Array Asymptotics

5.1 Asymptotic efficiency

This chapter considers efficiency issues associated with problems with many nuisance parameters. The context will be a Neyman-Scott type problem, in which there is a parameter of interest ψ, and observations are observed from p independent strata, each strata having an associated nuisance parameter λ_i, $i = 1, 2, \ldots, p$. The within stratum observations are *i.i.d.* with density $f(\psi, \lambda_i)$, and the sample size from each strata is n_p. The type of asymptotics considered is termed "rectangular array" asymptotics, because both p and n_p are allowed to go to infinity simultaneously. The situation is illustrated in Figure 5.1, where the arrows point in the direction of growth of the table.

Parameters

(λ_1, ψ)	X_{11}	X_{12}	X_{13}	$\ldots$	X_{1n_p}
(λ_2, ψ)	X_{21}	X_{22}	X_{23}	$\ldots$	X_{2n_p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
(λ_p, ψ)	X_{p1}	X_{p2}	X_{p3}	$\ldots$	X_{pn_p}

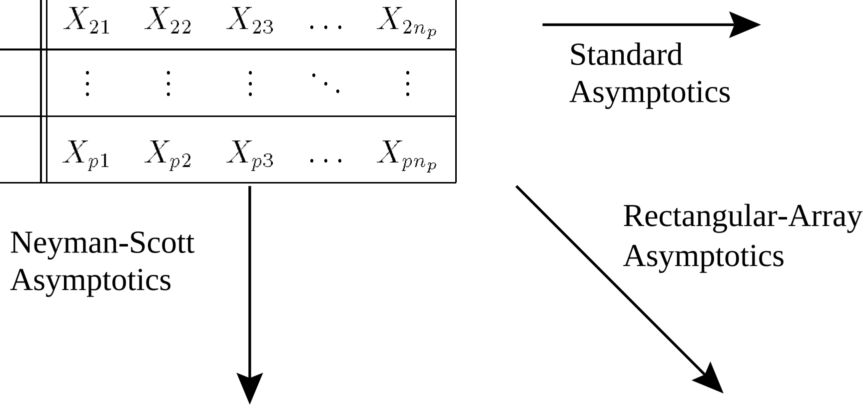

Figure 5.1: Graphical representation of 3 possible asymptotics in Neyman-Scott type problems.

Before starting to analyze the asymptotics it is useful to review the definition of *asymptotic efficiency* in the *i.i.d.* case, (this follows Lehmann (1983, ch.6)). Interest is in estimation of ψ. Suppose that $\delta_n = \delta_n(\mathbf{x}_n)$ is asymptotically normal,

$$\sqrt{n}\,[\delta_n - \psi] \xrightarrow{\ell} N[0, \nu(\psi, \lambda)], \tag{5.1}$$

then a sequence $\{\delta_n\}$ satisfying (5.1) is said to be asymptotically efficient if

$$\nu(\psi, \lambda) = [I(\psi, \lambda)]^{-1}_{\psi\psi},$$

where $I(\psi, \lambda)$ is the information matrix for a single observation. That is, the inverse of the asymptotic variance is equal to the marginal Fisher's Information in ψ.

In the rectangular array situation an estimator will be said to be *asymptotically efficient* if

$$\sqrt{pn_p}\,[\delta_n - \psi] \xrightarrow{\ell} N[0, \nu(\psi, \lambda)] \tag{5.2}$$

and the inverse of the variance in the limiting distribution is the limiting average of the marginal Fisher's information over the strata.

Recall that the per observation Fishers information matrix for the i^{th} stratum is

$$M_1^i = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix},$$

so that the marginal Fisher's information is equal to $i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi}$. This is simply the per observation information in the U_1 estimating equation, and will be denoted by I_1^i (the superscript i here, is needed because this information is associated with the i^{th} stratum). So an equivalent way of expressing the asymptotic variance of an efficient estimator is to state that

$$\nu(\psi, \lambda)^{-1} = \overline{I}_1, \tag{5.3}$$

where $\overline{I}_1$ is the limiting average of the I_1^i, which is assumed to exist.

Focusing now on the classic Neyman-Scott problem, Example 2, consider the consistency and efficiency of the *m.l.e.* in the three different possible asymptotic embeddings pictured in Figure 5.1. For the standard asymptotics the *m.l.e.* of ψ is both consistent and efficient. In the Neyman-Scott asymptotic setting the *m.l.e.* is inconsistent, so it makes no sense to consider efficiency, and in the Square Array asymptotic embedding, that is when $n_p = p$, the following argument demonstrates that $\hat{\psi}$ is consistent, *but* fails to be efficient. These results are displayed in Table 5.1.

Table 5.1: Properties of $\hat{\psi}$ in the three asymptotic embeddings.

Asymptotics	Consistency	Efficiency
Standard	✓	✓
Square Array	✓	✗
Neyman-Scott	✗	✗

The *m.l.e.* of ψ in the Neyman-Scott problem is the average of the within strata estimates, that is, denoting the within strata estimate by $\hat{\psi}_i$, then

$$\hat{\psi} = \tfrac{1}{p}\textstyle\sum_{i=1}^{p} \hat{\psi}_i.$$

Since $\hat{\psi}_i = \frac{1}{n_p}\sum_j (x_{ij} - \overline{x}_i)^2$, then $\hat{\psi}_i = \frac{\psi}{n_p}\sum_j \frac{(x_{ij}-\overline{x}_i)^2}{\psi} = \frac{\psi}{n_p} Y_i$, where Y_i is a χ^2

random variable on $n_p - 1$ degrees of freedom. Consequently

$$\hat{\psi} = \frac{\psi}{pn_p} \sum_{i=1}^{p} Y_i,$$

where the Y_i are independent χ^2 random variables, each on $n_p - 1$ degrees of freedom. It follows that $\hat{\psi}$ can be expressed as

$$\hat{\psi} = \frac{(n_p - 1)\psi}{n_p} \frac{1}{(n_p - 1)p} \sum_{k=1}^{(n_p-1)p} Z_k, \tag{5.4}$$

where the Z_k are *i.i.d.* χ^2 random variables on 1 degree of freedom. Applying the Weak Law of Large Numbers to $\frac{1}{(n_p-1)p} \sum_{k=1}^{(n_p-1)p} Z_k$ shows that it goes in probability to 1, so that

$$\hat{\psi} \xrightarrow{p} \psi,$$

as required. This establishes the consistency of $\hat{\psi}$. For the efficiency result, consider the limiting distribution of

$$\sqrt{pn_p}[\hat{\psi} - \psi].$$

The per observation Fisher's information matrix for the i^{th} stratum is

$$I_i = \begin{pmatrix} \frac{1}{2\psi^2} & 0 \\ 0 & \frac{1}{\psi} \end{pmatrix} \quad \text{and} \quad I_i^{-1} = \begin{pmatrix} 2\psi^2 & 0 \\ 0 & \psi \end{pmatrix},$$

so that $\nu(\psi, \lambda) = (\overline{I}_1)^{-1} = 2\psi^2$. Writing $\hat{\psi}$ as in (5.4), allows $\sqrt{pn_p}[\hat{\psi} - \psi]$ to be expressed as

$$\psi\sqrt{pn_p}\left[\frac{1}{pn_p} \sum_{k=1}^{(n_p-1)p} Z_k - 1\right], \tag{5.5}$$

where again the Z_k are independent χ^2 variates on one degree of freedom, so that their expectation and variance are 1 and 2 respectively. Rearranging (5.5) gives

$$\psi\sqrt{pn_p}\left[\frac{1}{pn_p} \sum_{k=1}^{(n_p-1)p} (Z_k - 1) - \frac{1}{n_p}\right],$$

but since $n_p = p$ this can be written as

$$\psi\left[\frac{1}{\sqrt{pn_p}}\sum_{i=1}^{(n_p-1)p}(Z_k-1)-1\right],$$

which has limiting distribution $N(-\psi, 2\psi^2)$. Consequently the *m.l.e.* fails to be efficient because of the bias term, $-\psi$, in the limiting distribution, even though it has the correct variance term. Heuristically, what is happening is that the within stratum estimates of ψ are biased by a factor of order $O(n_p^{-1})$, and since there are $p = n_p$ strata, this bias accumulates as the within strata *m.l.e.*'s are averaged over the strata.

The key result in this chapter, is that if the root of the U_2 estimating equation is used, rather than the *m.l.e.*, then bias will have been reduced to a high enough order to ensure full efficiency, in the sense that the limiting distribution will be mean zero, with the inverse of the asymptotic variance equal to (5.3).

5.2 Projective scores in the independent strata model

In this section the projective scores, and extended information matrices will be examined in the independent strata model. The case of univariate ψ and λ^i will be described, solely for ease of exposition, rather than technical necessity.

In the independent strata model the joint density of the observations is

$$f(\mathbf{x}, \psi, \lambda^1, \lambda^2\cdots\lambda^p) = \prod_{i=1}^{p}\prod_{j=1}^{n_p} f(x_{ij}, \psi, \lambda^i),$$

with log-likelihood

$$l(\psi, \lambda^1, \lambda^2\cdots\lambda^p) = \sum_i\sum_j \log(f(x_{ij}, \psi, \lambda^i)),$$

so that the score for the parameter of interest is

$$U_0 = \sum_i U_0^i, \tag{5.6}$$

where U_0^i is the score from the i^{th} stratum. The score for the nuisance parameter λ^i is

$$V_i = \sum_j \log(f(x_{ij}, \psi, \lambda^i)). \tag{5.7}$$

Notice that the nuisance parameter scores are mutually independent because the strata themselves are independent by assumption.

The fact that each nuisance parameter appears in only one stratum, and that the strata are independent has important consequences for the elements of the Bhattacharyya basis and their covariance matrix. To establish these results it is useful to think of the Bhattacharyya basis elements as members of one of two classes. The first class, $\mathcal{C}_1$ consists of elements that are from *within* strata, so that they would have the form $V_{i_1,i_2 \cdots i_t}$ with $i_1 = i_2 \cdots = i_t$, and the second class $\mathcal{C}_2$ are the *between* strata elements, that is those of the form $V_{i_1 i_2 \cdots i_t}$ with at least one $i_j \neq i_k$.

The next lemma gives a simple decomposition of the elements of $\mathcal{C}_2$ in terms of the elements of $\mathcal{C}_1$:

Lemma 5.1 *Let $V_{i_1 i_2 \cdots i_t} \in \mathcal{C}_2$, and without loss of generality assume that the subscripts have been permuted so that they fall into groups with a single index, for example $V_{1321213} = V_{1112233}$, and write this in general as $V_{v_1, v_2, \cdots, v_s}$, then*

$$V_{v_1, v_2, \cdots, v_s} = V_{v_1} V_{v_2} \cdots V_{v_s}.$$

For the given example the lemma states that $V_{1112233} = V_{111} V_{22} V_{33}$, in other words the between strata basis element has been expressed as a product of independent within strata elements.

Proof

The result is easily established using induction. The critical fact is that because each nuisance parameter appears in only a single strata, an element of the form V_{ij}, which by Theorem 3.1, can be written as

$$V_{ij} = \frac{\partial}{\partial \lambda^j} V_i + V_i V_j,$$

has the derivative, $\frac{\partial}{\partial \lambda^j} V_i$ equal to zero because V_i is a function of λ^i alone. Consequently $V_{ij} = V_i V_j$ and the general result follows by induction. ∎

Lemma 5.1 leads to the following observations:

- The covariance of a between strata element, and a within strata element is 0.
- The covariance of a between strata element and U_0 is 0.
- The covariance of two within strata elements, from different strata, is 0.

The results are consequences of the facts that the Bhattacharyya basis functions are mean zero, and that the strata are independent. The relevance of the results lies in the fact that all the between element members of the Bhattacharyya basis can be discarded from the information and projective score calculations, because they are orthogonal to both U_0, *and* orthogonal to the within strata elements.

If the Bhattacharyya basis of order t, $\{\mathbf{V}\}$, is partitioned according to the strata, that is $\{\mathbf{V}\} = \{\mathbf{V}^1, \mathbf{V}^2, \cdots, \mathbf{V}^p\}$, (here, all the between element members of the basis have been discarded) where the superscript i indicates that $\mathbf{V}^i$ is the Bhattacharyya basis based on the i^{th} stratum, then it is clear from the previous remarks that the relevant part of the extended information matrix takes on the form

$$M_t = \begin{pmatrix} E[U_0^2] & E[U_0\mathbf{V}^{1T}] & E[U_0\mathbf{V}^{2T}] & \cdots & [U_0\mathbf{V}^{pT}] \\ E[\mathbf{V}^1 U_0] & M_t^1 & 0 & \cdots & 0 \\ E[\mathbf{V}^2 U_0] & 0 & M_t^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ E[\mathbf{V}^p U_0] & 0 & 0 & \cdots & M_t^p \end{pmatrix}, \tag{5.8}$$

where M_t^i is the extended information matrix of order t based on the i^{th} stratum.

The expected value of $U_0 U_0$ is by definition

$$E[U_0 U_0] = E[(\sum_i U_0^i)^2] = \sum_i E[(U_0^i)^2], \tag{5.9}$$

since the strata are independent. Putting (5.9), together with (5.8) results in

Lemma 5.2 *In the Neyman-Scott sampling scheme described in section 5.1,*

$$\begin{aligned} U_t &= \sum_i U_t^i, \ \textit{and} \\ I_{U_t} &= E[U_t U_t^T] = \sum_i E[U_t^i U_t^{iT}]. \end{aligned}$$

Simply put, the lemma states that both the projective scores and the information are additive over the strata.

5.3 Efficiency in the square array

In the following developments there are many heavy assumptions, some concern the existence of higher order moments of the estimating equations, others put severe conditions on the nuisance parameter sequence $\{\lambda_i\}$. It would undoubtably be possible to weaken and change the assumptions, but that is not the purpose of this section. The aim is to demonstrate the importance of bias reduction in the estimating equations, with respect to obtaining efficiency in estimation in the square array asymptotic embedding. Since the asymptotics being considered involves allowing the number of nuisance parameters to increase to infinity, some form of structure or regularity needs to be placed on them. Three commonly used assumptions are (Andersen, 1970)

1. Bounded set: Assume that there exists a bounded closed interval in $\mathbb{R}$, say I, such that $\lambda_i \in I$ for all i.

2. Lindenberg Feller type conditions: these allow the sequence $\{\lambda_i\}$ to contain arbitrarily large values of λ_i, but that the sequence is dominated by moderate λ_i.

3. Random effects assumption: The λ_i are independent identically distributed random variables with common distribution $H(\lambda)$.

Here, the first of these will be assumed, though with some appropriate modifications the proofs would go through for the other assumptions. It will be necessary in the proofs to assert that many of the properties associated with projective scores and their expectations that hold for an individual stratum, also hold *uniformly* over the strata. The fact that the nuisance parameters are elements of a bounded set makes this uniformity assumption reasonable.

5.4 Fully efficient estimates in Neyman-Scott type problems

An argument for establishing the conditions, under which the solution of the 2^{nd} projective score will be fully efficient is given.

To do this a slight modification to the U_2^i projective score is needed in order to achieve the efficiency results. The modification is motivated from the ideas of Bickel (1982) and Schick (1986). Assume that the sample size is $2n_p$, and that the sample is partitioned into two independent groups of size n_p. Then the modified U_2^i is obtained by forming the U_2^i based on the first half of the data, and evaluating it at $\hat{\lambda}_\psi^i$ estimated from the second half of the data, and then finding U_2^i based on the second half of the data, and evaluating it at $\hat{\lambda}_\psi^i$ from the first half. These two estimating equations are then added to form the modified U_2^i; that is

$$U_{2M}^i(\hat{\lambda}_\psi^i, \mathbf{x}_i) = U_2^i(\hat{\lambda}_\psi^i(\mathbf{x}_{i2}), \mathbf{x}_{i1}) + U_2^i(\hat{\lambda}_\psi^i(\mathbf{x}_{i1}), \mathbf{x}_{i2}) \tag{5.10}$$

where $\mathbf{x}_i$ denotes the entire data, $\mathbf{x}_{i1}$ the first half of the data and $\mathbf{x}_{i2}$ the second half of the data. The reason why this device is needed is to ensure independence between $U_2^i(\lambda^i(\mathbf{x}_{i2}))$ and $\hat{\lambda}_\psi^i(\mathbf{x}_{i1})$, and similarly between $U_2^i(\lambda^i(\mathbf{x}_{i1}))$ and $\hat{\lambda}_\psi^i(\mathbf{x}_{i2})$, the need for this will become apparent in the proof of Theorem 5.1. It seems clear that (5.10) could be improved, in the sense that the data could be split over all permutations of the observations, so that the resulting estimating equation is permutation invariant. However, this will not be investigated here, but makes an interesting topic for future research.

The main theorem is

Theorem 5.1 *Denote the solution to the modified 2^{nd} projective score equation by $\tilde{\psi}$. Then under suitable regularity conditions, to be subsequently listed, $\tilde{\psi}$ is fully efficient in the square array asymptotic framework. That is, with $n_p = p$,*

$$\sqrt{pn_p}\left(\tilde{\psi} - \psi\right) \xrightarrow{\ell} N\left(0, (\overline{I}_1)^{-1}\right).$$

Proof

To establish the result note that $\tilde{\psi}$ solves

$$\sum_{i=1}^{p} U_2^i(\hat{\lambda}_\psi^i(\mathbf{x}_{i2}), \mathbf{x}_{i1}) + U_2^i(\hat{\lambda}_\psi^i(\mathbf{x}_{i1}), \mathbf{x}_{i2}) = 0.$$

That is,

$$\sum_{i=1}^{p} U_2^i(\hat{\lambda}_\psi^i(\mathbf{x}_{i2}), \mathbf{x}_{i1}) + U_2^i(\hat{\lambda}_\psi^i(\mathbf{x}_{i1}), \mathbf{x}_{i2})\Big|_{\psi=\tilde{\psi}} = 0. \tag{5.11}$$

A Taylor series expansion of (5.11) about $\tilde{\psi} = \psi$ yields

$$\begin{aligned} 0 = \sum_i U_{2M}^i(\tilde{\psi}, \hat{\lambda}_\psi^i) &= \sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i) + (\tilde{\psi} - \psi)\frac{\partial}{\partial\psi}\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i) \\ &+ (\tilde{\psi} - \psi)^2 \frac{\partial^2}{\partial\psi^2}\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i)\Big|_{\psi=\psi^*} \end{aligned} \tag{5.12}$$

for $|\psi^* - \psi| < |\tilde{\psi} - \psi|$. Rearranging (5.12) gives

$$\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i) = (\tilde{\psi}-\psi)\left[-\frac{\partial}{\partial\psi}\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i) + (\tilde{\psi} - \psi)\,\frac{\partial^2}{\partial\psi^2}\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i)\Big|_{\psi=\psi^*}\right]. \tag{5.13}$$

Recall that the per observation information in the U_1^i projective score is $I_1^i(\psi, \lambda_i)$, and the average of these informations over the p strata is denoted by $\overline{I}_1$, that is $\overline{I}_1 = \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} I_1^i(\psi, \lambda_i)$.

Dividing both sides of (5.13) through by $\sqrt{pn_p}\,\overline{I}_1$ gives

$$\frac{\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i)}{\sqrt{pn_p}\,\overline{I}_1} = \sqrt{pn_p}(\tilde{\psi}-\psi)\left(\frac{-\frac{\partial}{\partial\psi}\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i) + (\tilde{\psi} - \psi)\,\frac{\partial^2}{\partial\psi^2}\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i)\Big|_{\psi=\psi^*}}{pn_p\overline{I}_1}\right) \tag{5.14}$$

If it can be shown that the LHS of (5.14) goes in law to a Normally distributed random variable, with mean zero and variance $(\overline{I}_1)^{-1}$, and that the multiplicative factor on the RHS is $1 + o_p(1)$, that is,

$$\frac{-\frac{\partial}{\partial\psi}\left(\sum U_{2M}^i(\psi, \hat{\lambda}_\psi^i)\right) + (\tilde{\psi} - \psi)\,\frac{\partial^2}{\partial\psi^2}\left(\sum U_{2m}^i(\psi, \hat{\lambda}_\psi^i)\right)\Big|_{\psi=\psi^*}}{pn_p\overline{I}_1} = 1 + o_p(1)$$

then it follows that,

$$\sqrt{pn_p}\left(\tilde{\psi} - \psi\right) \xrightarrow{\ell} N(0, (\overline{I}_1)^{-1}),$$

so that $\tilde{\psi}$ is asymptotically fully efficient as required. The remainder of this chapter serves to verify these two statements. ∎

The first result seeks to establish that asymptotically the substitution version of the U_{2M} score, has the same limiting distribution as the U_{2M} score evaluated at λ_0, so that there is no asymptotic penalty for estimating λ.

Some notation needs to be established: first define Y_{in_p} as

$$Y_{in_p} = U_2^i(\hat{\lambda}_\psi(\mathbf{x}_{i2}), \mathbf{x}_{i1}) + U_2^i(\hat{\lambda}_\psi(\mathbf{x}_{i1}), \mathbf{x}_{i2}) - U_2^i(\lambda, \mathbf{x}_{i1}) - U_2^i(\lambda, \mathbf{x}_{i2})$$

and denote the mean of Y_{in_p} by μ_{in_p}. One can show in a similar fashion to Theorem 4.5 that Y_{in_p} has a mean zero limiting distribution as $n_p \to \infty$, denoted by Y_i, so that its stochastic order of magnitude is fixed and μ_{in_p} can be expected to become small as $n_p \to \infty$. This motivates the following reasonable assumptions:

- (A1) Uniformly over the strata, $E[Y_{in_p}^4] \to E[Y_i^4]$, as $n_p \to \infty$.
- (A2) There exist constants $B > 0$ and $C < \infty$, and an integer N_0 such that

$$\mathrm{Var}[Y_{in_p}] \geq B \quad \text{for all } i, \quad \text{and } n_p \geq N_0 \tag{5.15}$$

$$E[(Y_{in_p} - \mu_{in_p})^4] \leq C \quad \text{for all } i, \quad \text{and } n_p \geq N_0 \tag{5.16}$$

- (A3) The "average" bias in Y_{in_p} is negligible, in the sense

$$\lim_{p\to\infty} \frac{\sum_{i=1}^p \mu_{in_p}}{\sqrt{pn_p}} = o(1).$$

(This assumption will be discussed further after the results are presented.)

Then under conditions A1 - A3 we have

Proposition 5.1

$$\frac{\sum_i U_{2M}^i(\hat{\lambda}_\psi)}{\sqrt{pn_p}} - \frac{\sum_i U_{2M}^i(\lambda_0)}{\sqrt{pn_p}} \xrightarrow{p} 0. \tag{5.17}$$

Proof

Consider the standardized sum

$$\frac{1}{\sqrt{p}}\sum_{i=1}^{p}(Y_{in_p}-\mu_{in_p}),$$

and make the following identifications:

$$X_{pi}=Y_{in_p}-\mu_{in_p} \qquad \sigma_{pi}^2=E[X_{pi}^2] \qquad s_p^2=\sum_{i-1}^{p}\sigma_{pi}^2.$$

Then the Lindenberg condition

$$\lim_{p\to\infty}\sum_{i=1}^{p}\frac{1}{s_p^2}\int_{|X_{pi}|\geq\epsilon s_p}X_{pi}^2dP=0 \qquad \text{for } \epsilon>0$$

is satisfied. This follows because the moment assumptions (5.15) and (5.16) are sufficient to ensure that Lyapounov's condition, see Billingsley (1986, (27.16)), with $\delta=2$ holds. It is well known that Lyapounov's condition implies the Lindenberg condition, so that by the Lindenberg Central Limit Theorem (Billingsley, 1986, Theorem 27.2),

$$\frac{\sum_{i=1}^{p}X_{pi}}{s_p}\xrightarrow{\ell}N(0,1). \tag{5.18}$$

Denoting $\lim_{p\to\infty}\sqrt{p}/s_p$ as c, $0<c<\infty$, then

$$\frac{\sum_{i=1}^{p}X_{pi}}{\sqrt{p}}\xrightarrow{\ell}N(0,c^{-2}). \tag{5.19}$$

On substituting X_{pi} by $Y_{in_p}-\mu_{in_p}$, (5.19) becomes

$$\frac{\sum_{i=1}^{p}Y_{in_p}}{\sqrt{p}}-\frac{\sum_{i=1}^{p}\mu_{in_p}}{\sqrt{p}}\xrightarrow{\ell}N(0,\sigma^2),$$

and on division by $\sqrt{n_p}$ we obtain

$$\frac{\sum_{i=1}^{p}U_2^i(\hat{\lambda}_\psi(\mathbf{x}_{i2}),\mathbf{x}_{i1})+U_2^i(\hat{\lambda}_\psi(\mathbf{x}_{i1}),\mathbf{x}_{i2})-U_2^i(\lambda_0^i,\mathbf{x}_{i1})-U_2^i(\lambda_0^i,\mathbf{x}_{i2})}{\sqrt{pn_p}}-\frac{\sum_{i=1}^{p}\mu_{in_p}}{\sqrt{pn_p}}\xrightarrow{p}0.$$

Since the term

$$\frac{\sum_{i=1}^{p} \mu_{in_p}}{\sqrt{pn_p}},$$

is from assumption (A3), $o(1)$,

$$\frac{\sum_{i=1}^{p} U_{2M}^{i}(\hat{\lambda}_{\psi}^{i}) - U_{2M}^{i}(\lambda)_{0}^{i}}{\sqrt{pn_p}} \xrightarrow{p} 0$$

as required. ■

Examining the bias term $\frac{\sum_{i=1}^{p} \mu_{in_p}}{\sqrt{pn_p}} = \frac{\sqrt{p}}{\sqrt{n_p}} \overline{B}_p$, where $\overline{B}_p = \frac{1}{p} \sum_i \mu_{in_p}$, indicates a tradeoff between the the order of $\overline{B}_p$ and the rate that n_p is allowed to go to ∞. That is, the greater the degree of bias removal, the slower the rate that n_p can be allowed to go to infinity. This suggests an entire spectrum of rectangular array asymptotics, in which increased bias reduction is rewarded with a decrease in the need for information within the strata.

Assumption (A3) is of course critical to Proposition 5.1. The following argument establishes some conditions under which the order of $\frac{\sqrt{p}}{\sqrt{n_p}} \overline{B}_p$ is $\sqrt{p} O(n_p^{-1})$, so that when $p = n_p$, the bias is of order $O(p^{-\frac{1}{2}})$, that is $o(1)$. The argument requires the following assumptions:

- (B1) Within each stratum, there is an appropriate λ-parameterization such that the sequence $\{\zeta_{n_p}^{i}\} = \sqrt{n_p}(\hat{\lambda}_{\psi}^{i} - \lambda_{0}^{i})$ is uniformly integrable, so that $E[\zeta_{n_p}^{i}] = O(1)$ and $E[(\hat{\lambda}^{i} - \lambda_{0}^{i})^{k}] = O(n_p^{-\frac{k}{2}})$. Furthermore, this holds uniformly over the strata.

- (B2) The derivatives of the within strata projective scores behave like *i.i.d.* sums, in particular $E[U_{2M,k}^{i}(\lambda_{0}^{i})] = O(n_p)$.

- (B3) The remainder term in the Taylor series expansion of $U_{2M}^{i}(\hat{\lambda}^{i})$ has a negligible expectation in comparison with the higher order terms.

Proposition 5.2 *In the square array asymptotic embedding, under conditions B1-B3, the bias term* $\frac{\sqrt{p}}{\sqrt{n_p}}\overline{B}_p$ *is* $O(p^{-\frac{1}{2}})$.

Proof

Consider the i^{th} stratum, and a Taylor series expansion of $U^i_{2M}(\hat{\lambda}^i) - U^i_{2M}(\lambda^i_0)$ about λ^i_0 (k as a subscript here is denoting the k^{th} derivative with respect to λ^i):

$$\begin{aligned} & U^i_{2M}(\hat{\lambda}^i) - U^i_{2M}(\lambda^i_0) = \\ & \quad \sum_{k=1}^{3} \frac{1}{k!}(\hat{\lambda}^i_\psi(\mathbf{x}_{i2}) - \lambda^i_0)^k U^i_{2M,k}(\lambda^i_0, \mathbf{x}_{i1}) + \frac{1}{4!}(\hat{\lambda}^i_\psi(\mathbf{x}_{i2}) - \lambda^i_0)^4 U^i_{2M,4}(\lambda^*) \\ & + \sum_{k=1}^{3} \frac{1}{k!}(\hat{\lambda}^i_\psi(\mathbf{x}_{i1}) - \lambda^i_0)^k U^i_{2M,k}(\lambda^i_0, \mathbf{x}_{i2}) + \frac{1}{4!}(\hat{\lambda}^i_\psi(\mathbf{x}_{i1}) - \lambda^i_0)^4 U^i_{2M,4}(\lambda^*) \end{aligned} \tag{5.20}$$

where $|\lambda^* - \lambda^i_0| < |\hat{\lambda}_\psi(\mathbf{x}_{i2})^i - \lambda^i_0|$ and $|\lambda^* - \lambda^i_0| < |\hat{\lambda}_\psi(\mathbf{x}_{i1}) - \lambda^i_0|$.

Taking the expectation of (5.20) and noting that by Corollary 3.1 to Theorem 3.2

$$E_{\lambda_0}[U^i_{2M,k}] = 0 \text{ for } k = 0, 1, 2,$$

implies that

$$\begin{aligned} E_{\lambda_0}[U^i_{2M}(\hat{\lambda}^i_0) - U^i_t(\lambda^i_0)] &= E\left[\frac{1}{3!}(\hat{\lambda}^i_\psi(\mathbf{x}_{i2}) - \lambda^i_0)^3 U^i_{2M,3}(\lambda^i_0, \mathbf{x}_{i1})\right] + \\ & \quad E\left[\frac{1}{3!}(\hat{\lambda}^i_\psi(\mathbf{x}_{i1}) - \lambda^i_0)^3 U^i_{2M,3}(\lambda^i_0, \mathbf{x}_{i2})\right] + E[Rem]. \end{aligned}$$

Now the reason for using the modified form of U_2 is revealed: since the estimator $\hat{\lambda}^i_\psi(\mathbf{x}_{i2})$ and the score function $U^i_{2M,3}(\lambda^i_0, \mathbf{x}_{i1})$ are *independent* the expectation of their product is the product of their expectations, so that

$$E_{\lambda_0}[U^i_{2M}(\lambda^i) - U^i_{2M}(\lambda^i_0)] = O(n_p^{-\frac{3}{2}})O(n) = O(n_p^{-\frac{1}{2}})$$

and $\frac{\sqrt{p}}{\sqrt{n_p}}\overline{B}_p = \sqrt{p}O(n_p^{-1})$. As the situation under consideration is a *square array*, with $n_p = p$, the bias is $O(p^{-\frac{1}{2}})$ as required since the remainder term has by assumption a negligible expectation. ■

The next result shows that the per observation information in the modified U_2 estimating equation is asymptotically equivalent to the per observation information in the U_1 estimating equation. The relevance of the result is that it will ensure that the root of U_{2M} has the correct limiting variance.

Proposition 5.3

$$\frac{1}{n_p} I_{U_{2M}^i(\lambda_0^i, \mathbf{x}_i)} = \frac{1}{n_p} I_{U_1^i(\lambda_0^i, \mathbf{x}_i)} + O(n_p^{-1})$$

Proof

First note that by definition

$$I_{U_{2M}^i}(\lambda_0^i) = E[\nabla_\psi(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))](E[(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))^2])^{-1} \times E[\nabla_\psi(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))] \quad (5.21)$$

But, because U_0^i is an *i.i.d.* sum,

$$\begin{aligned} E[\nabla_\psi(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))] &= E[(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))U_0^i(\mathbf{x})] \\ &= E[(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))(U_0^i(\mathbf{x}_{i1}) + U_0^i(\mathbf{x}_{i2}))] \end{aligned}$$

Consequently, using the independence of $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$

$$E[\nabla_\psi(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))] = E[U_2^i(\lambda_0^i, \mathbf{x}_{i1})^2 + U_2^i(\lambda_0^i, \mathbf{x}_{i2})^2].$$

By a similar argument one can show that

$$E[(U_2^i(\lambda_0^i, \mathbf{x}_{i1}) + U_2^i(\lambda_0^i, \mathbf{x}_{i2}))^2] = E[U_2^i(\lambda_0^i, \mathbf{x}_{i1})^2 + U_2^i(\lambda_0^i, \mathbf{x}_{i2})^2].$$

It therefore follows that

$$I_{U_{2M}^i}(\lambda_0^i) = E[U_2^i(\lambda_0^i, \mathbf{x}_{i1})^2 + U_2^i(\lambda_0^i, \mathbf{x}_{i2})^2].$$

Also, since U_1^i itself is an *i.i.d.* sum, the U_1^i information, $I_{U_1^i}$ can be expressed as,

$$I_{U_1^i} = E[(U_1(\lambda_0^i, \mathbf{x}_{i1}))^2] + E[(U_1(\lambda_0^i, \mathbf{x}_{i2}))^2]$$

Consequently

$$I_{U_1^i(\lambda_0^i)} - I_{U_{2M}^i(\lambda_0^i)} = E[U_2^i(\lambda_0^i, \mathbf{x}_{i1})^2] - E[U_1^i(\lambda_0^i, \mathbf{x}_{i1})^2] + E[U_2^i(\lambda_0^i, \mathbf{x}_{i2})^2] - E[U_1^i(\lambda_0^i, \mathbf{x}_{i2})^2]$$

Appealing to Theorem 3.5, $E[U_2^i(\lambda_0^i, \mathbf{x}_{i1})^2] - E[U_1^i(\lambda_0^i, \mathbf{x}_{i1})^2] = O(1)$ which gives that

$$\frac{1}{n_p}(I_{U_1^i(\lambda_0^i)} - I_{U_{2M}^i}(\lambda_0^i)) = O(n_p^{-1})$$

as required. ■

In order to establish that the LHS of (5.14) does indeed have the appropriate limiting Normal distribution, one final result is needed. The necessary assumption is that the random variables $\tilde{U}_{2M}^i = \frac{1}{\sqrt{n_p}} U_{2M}^i$ satisfy the Lindenberg condition. That is defining $\text{Var}(\tilde{U}_{2M}^i) = \sigma_{pi}^2$, and $s_p^2 = \sum_i \sigma_{pi}^2$

$$\lim_{p\to\infty} \sum_{i=1}^{p} \frac{1}{s_p^2} \int_{|\tilde{U}_{2M}^i| \geq \epsilon s_p} \tilde{U}_{2M}^{i\,2} dP = 0 \qquad \text{for } \epsilon > 0$$

Proposition 5.4 *Under the above assumptions*

$$\frac{1}{\sqrt{pn_p}} \sum_i U_{2M}^i(\lambda_0) \xrightarrow{\ell} N(0, (\overline{I}_1)^{-1})$$

Proof

Clearly $\tilde{U}_{2M}^i$ is a mean zero random variable, with variance

$$\sigma_{pi}^2 = E[(\tilde{U}_{2M}^i)^2] = \frac{1}{n_p} E[(U_{2M}^i)^2] = \frac{1}{n_p} I_{U_1} + O(n_p^{-1}) = I_1^i + O(n_p^{-1}),$$

where I_1^i is the per observation I_1 information in the i^{th} stratum. The last equality is established using Proposition 5.3. Therefore $s_p^2 = \sum_i \sigma_{pi}^2 = \sum_i I_1^i + O(1)$.

Since by assumption, the Lindenberg condition holds,

$$\frac{\sum_i \tilde{U}_{2M}^i}{s_p} \xrightarrow{\ell} N(0,1).$$

Denoting the limit of $\sqrt{p}/s_p$ as c,

$$\frac{\sum_i \tilde{U}_{2M}^i}{\sqrt{p}} \xrightarrow{\ell} N(0,c^{-2}).$$

Now $c^{-2} = lim_{p\to\infty} s_p^2/p = \overline{I}_1$, and therefore

$$\frac{\sum_i \tilde{U}_{2M}^i}{\sqrt{p}\overline{I}_1} \xrightarrow{\ell} N(0,\overline{I}_1{}^{-1}),$$

which is the required result. ∎

Propositions 5.1 and 5.4 together imply that

$$\frac{\sum U_{2M}^i(\psi,\hat{\lambda}_\psi^i)}{\sqrt{pn_p}\,\overline{I}_1} \xrightarrow{\ell} N(0,(\overline{I}_1)^{-1}),$$

which was the claim for the LHS of (5.14).

It still remains to deal with the RHS of (5.14): recall that the claim was that

$$\frac{-\frac{\partial}{\partial\psi}\left(\sum U_{2M}^i(\psi,\hat{\lambda}_\psi^i)\right) + (\tilde{\psi}-\psi)\,\frac{\partial^2}{\partial\psi^2}\left(\sum U_{2m}^i(\psi,\hat{\lambda}_\psi^i)\right)\Big|_{\psi=\psi^*}}{pn_p\overline{I}_1} = 1 + o_p(1).$$

To take care of the remainder term it is enough to assume that

$$\frac{\frac{\partial^2}{\partial\psi^2}\left(\sum U_{2M}^i(\psi,\hat{\lambda}_\psi^i)\right)\Big|_{\psi=\psi^*}}{pn_p},$$

is bounded in the limit, and that $\tilde{\psi}-\psi_0$ is $o_p(1)$.

It remains to be shown that

$$\frac{-\frac{\partial}{\partial\psi}\left(\sum U_{2M}(\psi,\hat{\lambda}_\psi^i)\right)}{p\,n_p\,\overline{I}}$$

really does go to 1 in probability.

Differentiating for the numerator yields

$$\frac{\partial}{\partial \psi}\left(\sum U^i_{2M}(\psi, \hat{\lambda}^i_\psi)\right) = \sum U^i_{2M,\psi}(\psi, \hat{\lambda}^i_\psi) + \sum \frac{d\hat{\lambda}^i_\psi}{d\psi} U^i_{2M,\lambda_i}(\psi, \hat{\lambda}^i_\psi)$$

and the result will follow if

$$\frac{\sum U^i_{2M,\psi}(\psi, \hat{\lambda}^i_\psi)}{pn_p} = -\overline{I}_1 + o_p(1), \tag{5.22}$$

and if

$$\frac{\sum \frac{d\hat{\lambda}^i_\psi}{d\psi} U^i_{2M,\lambda_i}(\psi, \hat{\lambda}^i_\psi)}{pn_p} = o_p(1). \tag{5.23}$$

To obtain (5.22), consider an expansion about λ^i_0, with an appropriate bounding condition on the remainder term. Then

$$\frac{\sum U^i_{2M,\psi}(\psi, \lambda^i_0)}{pn_p} \xrightarrow{p} -\overline{I}_1$$

by a Weak Law of Large Numbers for triangular arrays, (see Billingsley (1986, Theorem 6.2)). For (5.23) the fact that

$$\frac{d\hat{\lambda}^i_\psi}{d\psi} = -\frac{\frac{\partial}{\partial \psi} V_i}{\frac{\partial}{\partial \lambda^i} V_i}$$

means that this term goes in probability to a constant, since $\frac{d\hat{\lambda}^i_\psi}{d\psi}$ is a ratio of sums of *i.i.d.* random variables. Again expanding (5.23) about λ^i_0 and bounding the remainder term, shows that

$$\frac{\sum \frac{d\hat{\lambda}^i_\psi}{d\psi} U^i_{2M,\lambda_i}(\psi, \hat{\lambda}^i_\psi)}{pn_p} = \frac{\sum \frac{d\hat{\lambda}^i_\psi}{d\psi} U^i_{2M,\lambda_i}(\psi, \lambda^i_0)}{pn_p} + o_p(1). \tag{5.24}$$

But, $\frac{d\hat{\lambda}^i_\psi}{d\psi} \xrightarrow{p} c_i$, and within each stratum $\frac{U^i_{2M,\lambda_i}(\psi,\lambda^i_0)}{n_p} \xrightarrow{p} 0$, so that (5.23) can be written as

$$\frac{1}{p}\sum_i x_{pi},$$

where $x_{pi} = \frac{d\hat{\lambda}_\psi^i}{d\psi} \frac{U_{2M,\lambda_i}^i(\psi,\lambda_0^i)}{n_p}$ and $x_{pi} \xrightarrow{p} 0$ as $n_p \to \infty$. We now make a final assumption that the $x_{pi} \xrightarrow{p} 0$ uniformly over the strata, so that

$$\frac{1}{p} \sum_i x_{pi} \xrightarrow{p} 0,$$

which is the required result.

5.5 The exponential family case

The purpose of this section is to verify that the assumptions made in the proof of Theorem 5.1 are satisfied for a canonical parameter of interest in an exponential family setting. First consider the canonically parameterized exponential family model for the i^{th} stratum:

$$f(x_i, \psi, \lambda) = \exp(\psi y_i + \lambda s_i - k(\psi, \lambda)).$$

Since the projective scores are invariant under reparameterization of the nuisance parameters, one can equally well consider the parameterization in which the nuisance parameter is a mean value parameter. That is, we consider λ to be a function of ψ and η, where $\eta = \frac{\partial}{\partial \lambda} k(\psi, \lambda)$.

Then based on a sample of size n_p, the $m.l.e.$ of η is $\frac{1}{n_p} \sum s_j$, and $E[\hat{\eta}_{n_p}] = \eta$. Consequently the Central Limit theorem applies directly to the sequence

$$\left\{ \sqrt{n_p} \left(\hat{\eta}_{n_p} - \eta \right) \right\}$$

so that it has a limiting Normal distribution. Furthermore,

$$\mathrm{Var}[\sqrt{n_p} \left(\hat{\eta}_{n_p} - \eta \right)] = \mathrm{Var}[s_i] < \infty,$$

since all moments of exponential family random variables are finite. Consequently the condition

$$sup_{n_p} E\{|\sqrt{n_p}\left(\hat{\eta}_{n_p} - \eta\right)|^{1+\epsilon}\} < \infty$$

is satisfied for $\epsilon = 1$, which is sufficient for the sequence $\{\sqrt{n_p}\left(\hat{\eta}_{n_p} - \eta\right)\}$ to be uniformly integrable. This was the result required to establish that

$$E[\left(\hat{\eta}_{n_p} - \eta\right)] = O(n_p^{-\frac{1}{2}}), \tag{5.25}$$

and that $E[\left(\hat{\eta}_{n_p} - \eta\right)^k] = O(n_p^{-\frac{k}{2}})$. In fact, it is clear that $E[\left(\hat{\eta}_{n_p} - \eta\right)]$ is exactly equal to 0, but (5.25) is needed to get the magnitude of the higher order moments. See McCullagh (1987, p.209) for results of this nature in a more general setting.