INFERENCE FROM THE PRODUCT OF MARGINALS OF A DEPENDENT LIKELIHOOD

by

Bertrand S. Clarke
Purdue University

Brian W. Junker[1]
Carnegie Mellon University

Technical Report #91-10

Department of Statistics
Purdue University

February 1991

[1]Supported in part by grants ONR-N00014-90-J-1984, ONR-N00014-91-J-1208, and NIMH-MH15758.

◇

Abstract

In problems in which a psychological construct or latent trait is measured indirectly by a vector of responses or observations $\underline{x}^n = (x_1, \ldots, x_n)$ on each individual, a mixture model is often proposed in which the distribution of the observation vector, across individuals, is viewed as the mixture of some conditional distribution of the observation vector given the construct. The conditional distribution of observations given the trait θ_1 is often assumed to factor as $q^n(\underline{x}^n \mid \theta_1) = \prod_{i=1}^n q_i(x_i \mid \theta_1)$—though it may be conceded that conditional independence does not hold—because it is too difficult to work with the correct dependence model $\nu^n(\underline{x}^n \mid \theta_1)$. We show that taking q^n to be the product of marginals of ν^n is optimal, among conditional independence models, under criteria related to the Kullback-Leibler distance. Our main results give conditions on ν^n under which, as $n \to \infty$, the q^n-based MLE $\hat{\theta}_{q,1}$ is consistent under ν^n, and the q^n-based posterior $\omega_q(\theta_1|\underline{x}^n)$ is asymptotically normal, centered at $\hat{\theta}_{q,1}$ and scaled by the q^n-based empirical Fisher information. Sufficient conditions for these results involve laws of large numbers for ν^n which generalize "essential independence" criteria used in the modeling of standardized achievement tests.

Often, ν^n may be realized as the mixture over nuisance parameters $\underline{\theta}_2^d$ of some underlying higher-dimensional conditional independence model $p^n(\underline{x}^n|\theta_1, \underline{\theta}_2^d) = \prod_{i=1}^n p_i(x_i|\theta_1, \underline{\theta}_2^d)$. In this case the assumptions may be moved from ν^n to p^n, where laws of large numbers hold naturally. Again, we obtain consistency of $\hat{\theta}_{q,1}$, and show that when it is consistent it converges in distribution to a mixture of normals. However, $\omega_q(\theta_1|\underline{x}^n)$ remains asymptotically normal, with the same centering and scaling as before. This is significant, in that even when the full model p^n behaves well, asymptotic inference based on the product-of-marginals likelihood q^n yields different answers from asymptotic inference based on the product-of-marginals posterior ω_q. Moreover, model-fit considerations may be better served by asymptotic likelihood methods, which appear to be more sensitive to the true dependence structure of the data than asymptotic posterior methods.

We illustrate our basic results with some models from item response theory, and illustrate the extension to an underlying conditional independence structure with normal models in which either the location or the scale is a nuisance parameter.

Keywords: structural robustness, nuisance parameters, dependence, marginal likelihood, latent variable models, psychometrics, item response theory.

Contents

1 Introduction, Motivations, Examples

Suppose we wish to measure a construct such as social adjustment, job satisfaction, school math achievement, etc. To gain information about one of these traits, it is common to make a set of n observations (administer a test of n questions, ask an expert to assess the severity of n symptoms, etc.) on each individual. Often, the trait is then quantified as a latent (unobservable) random variable Θ_1. A numerical value is assigned to each observation used to measure Θ_1, giving rise to random variables $\underline{X}^n = (X_1, \ldots, X_n)$. Replications of $(\Theta_1, X_1, \ldots, X_n)$ across individuals are considered to be i.i.d.; and we will denote outcomes of random variables with the corresponding lower case letter.

The "ideal" model often proposed for data like this is a mixture of conditional independence models

$$m(\underline{x}^n) = \int r^n(\underline{x}^n \mid \theta_1)\, dF(\theta_1) \tag{1}$$

where F is the distribution of Θ_1, and $r^n(\underline{x}^n \mid \theta_1)$ factors as

$$r^n(\underline{x}^n \mid \theta_1) = \prod_{i=1}^{n} r_i(x_i \mid \theta_1)\,. \tag{2}$$

The main statistical task is inference about each individual's unobserved θ_1 from each individual's observed $\underline{x}^n$, based on the particular form of the right-hand side of (2).

Conditional independence models are an attractive and convenient data analysis tool, and are often assumed even though it may be agreed that (2) only approximately fits or reflects the mechanisms underlying the data. Suppose the correct formulation is

$$m(\underline{x}^n) = \int \nu^n(\underline{x}^n \mid \theta_1)\, dF(\theta_1), \tag{3}$$

where the conditional model for $\underline{X}^n$ given θ_1 is some dependent $\nu^n(\underline{x}^n \mid \theta_1)$ whose structure is not known in detail. How far could an analysis based on (1) and (2) go? We identify the product of one-dimensional marginals $q^n(\underline{x}^n \mid \theta_1) = \prod_{i=1}^n q_i(x_i \mid \theta_1)$, where

$$q_i(x_i \mid \theta_1) = \int \nu^n(\underline{x}^n \mid \theta_1)\, dx_1 \ldots dx_{i-1}\, dx_{i+1} \ldots dx_n, \tag{4}$$

as the optimal choice for $r^n(\underline{x}^n \mid \theta_1)$ under two different criteria; and we give conditions under which asymptotic inference (as $n \to \infty$) based on the product of marginals q^n may still be successful. In

particular we show the effects of an analysis based on q^n, when in fact the data came from ν^n, on asymptotic likelihood inference and asymptotic posterior inference. We also indicate some limitations on how much can be learned or predicted from this wrong-model analysis.

There are two ways in which ν^n may arise in applications. Let us first consider the case in which well-defined nuisance parameters prevent ν^n from factoring. In a test of school mathematics achievement, an examinee's performance on each question may also be affected by his verbal ability (since the questions and test instructions are written in English), by test anxiety, by special background knowledge not related to general math achievement, etc. Thus the unidimensional parameter Θ_1 should be regarded as the first coordinate in a vector valued parameter $\underline{\Theta}_1^d = (\Theta_1, \Theta_2, \ldots, \Theta_d)$, and the "ideal" density in (2) should be replaced with the conditionally independent density

$$p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right) = \prod_{i=1}^{n} p_i\left(x_i \mid \underline{\theta}_1^d\right) \tag{5}$$

and hence

$$\nu^n\left(\underline{x}^n \mid \theta_1\right) = \int p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right)\, dF(\underline{\theta}_2^d|\theta_1). \tag{6}$$

Typically the dimensionality d of $\underline{\theta}_1^d$ will be too high, or the details of the marginals for X_i in (5) will be too complicated, to simply estimate $\hat{\underline{\theta}}_1^d$ and then "throw away" $\hat{\underline{\theta}}_2^d$. Moreover, since $\underline{\theta}_2^d = (\theta_2, \ldots, \theta_d)$ represent nuisance factors not directly related to the trait of interest θ_1, the psychometric orthodoxy strongly favors the unidimensional model (2). Indeed, the practitioner will sometimes concede that the data is mildly multidimensional, in the sense of (5), but continue to use a fictional conditional model of the form (2) on the grounds that the more parsimonious unidimensional model (2) is simply not far wrong. Interest in this kind of analysis has been expressed by Ackerman (1987), Drasgow and Parsons (1983), Harrison (1986), Wang (1986, 1987), and Yen (1984).

Note that two kinds of nuisance parameters may be contemplated here. The first kind have a sustained influence on the distribution of $\underline{X}^n$ as $n \to \infty$: For instance, language ability will tend to help on every question in a math test. The second kind asymptotically attenuate in the sense that as n increases the distribution of $\underline{X}^n$ becomes less and less sensitive to their variations. An example in the educational testing context is specialized knowledge in a given subject area on a

general test. One expects that the specialized knowledge will be helpful on a cluster of questions but not on later or earlier questions (e.g. Rosenbaum, 1988; Stout, 1990; Wainer et al., 1990).

$\nu^n(\underline{x}^n \mid \theta_1)$ may also arise in settings for which (6) is not plausible. If the construct being measured is not sufficiently well-defined for (1) and (2) to hold with respect to a unidimensional Θ_1, it may not be clear that there are meaningful secondary traits $\underline{\Theta}_2^d$ for which (5) can be written. The correct model would seem to be (3) in which ν^n does not factor and does not arise by mixing out nuisance factors as in (6). Reiser (1989) reports results which suggest that major depressive disorder, as defined by the American Psychiatric Association's DSM-III criteria, may fall into this case. Thus it is also important to understand inference based on the model (1) and (2) when (3) holds but $\nu^n(\underline{x}^n \mid \theta_1)$ does not factor, and little further structure can be posited.

We shall restrict our attention to cases in which Θ_1 or $\underline{\Theta}_1^d$ have continuous distributions $dF(\theta_1) = \omega(\theta_1)d\theta_1$ and $dF(\underline{\theta}_1^d) = \omega(\underline{\theta}_1^d)d\underline{\theta}_1^d$. Our main interest is asymptotic inference, as $n \to \infty$, based upon the conditional independence model $q^n(\underline{x}^n \mid \theta_1) = \prod_{i=1}^n q_i(x_i \mid \theta_1)$ instead of the correct, conditionally dependent $\nu^n(\underline{x}^n \mid \theta_1)$. If a "full model" $p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right) = \prod_{i=1}^n p_i\left(x_i \mid \underline{\theta}_1^d\right)$ is assumed to exist, it follows that

$$\nu^n(\underline{x}^n \mid \theta_1) = \int p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right) \omega\left(\underline{\theta}_2^d \mid \theta_1\right) d\underline{\theta}_2^d \tag{7}$$

and, with the help of Fubini's theorem,

$$q_i(x_i \mid \theta_1) = \int p_i\left(x_i \mid \underline{\theta}_1^d\right) \omega\left(\underline{\theta}_2^d \mid \theta_1\right) d\underline{\theta}_2^d. \tag{8}$$

Note that, although we refer to θ_1 as unidimensional throughout this paper, the new results here all have obvious extensions to the case in which θ_1 is really of fixed dimension $d_1 \geq 1$.

In Section 2 we consider two criteria for choosing r^n in (2), both related to the Kullback-Leibler distance. The first is an estimation criterion: we show that $r^n = q^n$ results from minimizing what is essentially a Bayes risk over the collection of product densities. The second is a hypothesis testing criterion: we show that Stein's test based on q^n is near asymptotically minimax, under a uniformity assumption. It must be noted that in practice, the selection of r^n in (2) is itself often subject to uncertainty, in that the $r_i(\cdot|\theta_1)$ are typically selected from a parametric family $r_{\underline{\alpha}_i}(\cdot|\theta_1)$ whose parameters $\underline{\alpha}_1, \ldots, \underline{\alpha}_n$ are estimated from (some subset of) the data. This part of the problem

is important but we do not address it here. Instead, we provide an indication of what is "best possible" in that we work with the conditional independence likelihood q^n that is closest to the correct dependence model. There is some evidence that this "best possible" case is approximately achieved in certain applications (cf. Wang, 1987).

In analyzing inference based on q^n, it is more straightforward to first suppress consideration of the full model $p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right)$. Stout (1987, 1990) has developed a criterion for binary data $\underline{x}^n$ called *essential independence* which identifies θ_1 as the "dominant latent trait" in the sense that conditioning on θ_1 stabilizes linear combinations of the x_i's as $n \to \infty$. This criterion may be interpreted for more general $\underline{x}^n$ as imposing a law of large numbers (LLN) on ν^n:

$$\lim_{n\to\infty} \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} a_i(X_i)\middle|\theta_1\right) = 0 \tag{9}$$

for all bounded sequences of functions $\{a_i(\cdot) : i = 1,\ldots,\infty\}$. The condition (9) was applied by Junker (1991) in the educational measurement setting to the analysis of maximum likelihood estimators for θ_1 based on q^n when in fact $\nu^n\left(\underline{x}^n \mid \theta_1\right)$ is dependent. Equation (9) imposes conditions on the dependent likelihoods $\nu^n\left(\underline{x}^n \mid \theta_1\right)$ only; it does not make the further conditional independence assumptions (5) and (6). In this sense, (9) may be contemplated whether or not the full model p^n is assumed to exist. However if (6) does hold then (9) may be interpreted as requiring that influence of $\underline{\theta}_2^d$ on the distribution of $\underline{X}^n$ attenuates as $n \to \infty$.

In Section 3, making only assumptions on ν^n, we show that the q^n-based MLE $\hat{\theta}_{q,1}$ is ν^n-consistent for θ_1 (converges to θ_1 in ν^n-probability), under conditions including laws of large numbers related to (9). We also give an asymptotically normal approximation to the q^n-based posterior $\omega_q(\theta_1|\underline{x}^n)$ under these conditions, centered at $\hat{\theta}_{q,1}$ and scaled according to the q^n-based empirical Fisher information.

In Section 4 we establish ν^n-consistency of $\hat{\theta}_{q,1}$ without the LLN assumptions for ν^n, as long as regularity assumptions on p^n are made uniformly on compact sets of $\underline{\theta}_2^d$. We also show that if $\hat{\theta}_{q,1}$ is consistent, then it converges in distribution to a mixture of normals. On the other hand, $\omega_q(\theta_1|\underline{x}^n)$, the distribution of Θ_1 given $\underline{x}^n$, is still asymptotically normal, with the same q^n-based centering and scaling as before.

Sections 3 and 4 form two parts of a complete whole. Section 3 addresses the case in which

no underlying conditional independence model p^n is assumed to exist, as well as the case in which p^n exists and the nuisance parameters attenuate. Section 4 addresses the case in which it is reasonable to assume the existence of mathematically well defined nuisance parameters in p^n, without necessarily assuming they attenuate. The results of Sections 3 and 4 are illustrated in Section 5 with some models from item response theory, and with normal models in which either the location or the scale is a nuisance parameter. In Section 6 we consider some implications of our results.

Our results are interesting for several reasons. First, the weak law of large numbers appears to be a much stronger assumption than expected: assuming only a weak LLN for ν^n gives consistency for $\hat{\theta}_{q,1}$ and asymptotic normality for $\omega_q(\theta_1|\underline{x}^n)$. Second, when the assumptions are moved from ν^n to p^n, the posterior normality results depend crucially on using the empirical Fisher information to scale the distribution. Substituting the expected Fisher information for the empirical one leads to a result that is not useful. Third, even when the full model p^n behaves well, q^n-based likelihood inference is different from q^n-based posterior inference: the asymptotic distribution of $\hat{\theta}_{q,1}$ is a mixture of normals, whereas the posterior distribution $\omega_q(\theta_1|\underline{x}^n)$ continues to be normal with the "independence-based" location and scale parameters. This is true even though the scale for both theorems is determined by the q^n-based empirical Fisher information. Based on "i.i.d." intuition one expects Bayes and ML inference to be asymptotically equivalent, but this is a practical situation in which both analyses can be carried out on the same model and give different answers.

The present paper complements two existing literatures in the large sample theory of inference. One the one hand, Berk (1966) characterized the asymptotic carrier (support set) of the posterior distribution under a wrong-model analysis in which both the correct and incorrect models involve i.i.d. data. Yamada (1976) extended this characterization to situations in which both the correct and incorrect models may have more general dependence structures, but in this general setup the asymptotic carrier of the wrong-model posterior is difficult to actually calculate. It is interesting to note that Yamada's "sufficient condition for general cases" requires a condition like (9) to hold *uniformly* in the parameter of interest. We drop this uniformity requirement and focus on situations in which the correct model involves some form of dependence while the incorrect model assumes

independence. In addition, we obtain consistency and asymptotic distribution results for maximum likelihood and posterior distribution estimators.

On the other hand, the techniques used here are based on the clear description of Laplace's method by Walker (1969). A consequence of our results is that asymptotic posterior normality in the wrong model is rather insensitive to the true dependence structure of the data. This is consonant with Chen (1985) who shows that the success or failure of Laplace's method in establishing asymptotic posterior normality is an analytic property of the model and the particular data sequence observed, not a property of the true probability structure of the data. The work of Kass, Tierney and Kadane (1990) is also relevant here. In this context, we show that under (9) and related conditions, data sequences on which asymptotic q^n-based posterior normality can be observed are quite common under ν^n.

2 The Best Independence Model

2.1 An estimation interpretation

As discussed in Section 1, we would like to base inference for θ_1 on an objective function $r^n(\underline{x}^n \mid \theta_1) = \prod_{i=1}^n r_i(x_i \mid \theta_1)$, even though $\nu^n(\underline{x}^n \mid \theta_1)$ is the correct likelihood. When p^n is assumed to exist, we would like r^n to be as close as possible to p^n. When p^n cannot be assumed to exist, we would like r^n to be as close as possible to ν^n.

Recall the Kullback-Leibler distance $D(f \| g) = E[\log(f(\underline{X}^n)/g(\underline{X}^n))]$, where $\underline{X}^n$ has density $f(\cdot)$. See, for example, Section 4 of Bahadur (1971) for basic properties of $D(\cdot \| \cdot)$. Densities minimizing a Kullback-Leibler distance come up in various contexts, including minimax hypothesis tests, see Huber and Strasser (1973). Indeed, it follows from Stein's Lemma (Chernoff, 1954; Bahadur, 1971) that every Kullback-Leibler number is the exponent for the probability of type II error for some simple versus simple hypothesis test. When it is helpful to remember which parameters are fixed in the integration, the fixed parameters will appear as subscripts; for example

$$D(\nu^n \| r^n) \equiv D_{\theta_1}(\nu^n \| r^n) = \int \log \frac{\nu^n(\underline{x}^n \mid \theta_1)}{r^n(\underline{x}^n \mid \theta_1)} \nu^n(\underline{x}^n \mid \theta_1)\, d\underline{x}^n.$$

Proposition 2.1 *$D_{\theta_1}(\nu^n \| r^n)$ is minimized over r^n by taking $r^n \equiv q^n$.*

Proof. Following Aitchison (1975), we note that, by (4),

$$\begin{aligned} D_{\theta_1}\left(\nu^n \| r^n\right) &= D_{\theta_1}\left(\nu^n \| q^n\right) + \sum_{i=1}^{n} \int \log \frac{q_i\left(x_i \mid \theta_1\right)}{r_i\left(x_i \mid \theta_1\right)} q_i\left(x_i \mid \theta_1\right) dx_i \\ &= D_{\theta_1}\left(\nu^n \| q^n\right) + \sum_{i=1}^{n} D_{\theta_1}\left(q_i \| r_i\right), \end{aligned}$$

which is clearly minimized by taking $r_i \equiv q_i$ in each term of the summation at right. □

A measure of the discrepancy between $r^n\left(\cdot \mid \theta_1\right)$ and $p^n\left(\cdot \mid \underline{\theta}_1^d\right)$ at each value of θ_1 is

$$R_{\theta_1}\left(p^n, r^n\right) = E\left[D_{\left(\theta_1, \underline{\Theta}_2^d\right)}\left(p^n \| r^n\right) \middle| \theta_1\right], \tag{10}$$

which may be interpreted as the Bayes risk in estimating $p^n\left(\cdot \mid \underline{\theta}_1^d\right)$ by $r^n\left(\cdot \mid \theta_1\right)$.

Proposition 2.2 *The Bayes risk $R_{\theta_1}\left(p^n, r^n\right)$ is also minimized over r^n by taking $r^n \equiv q^n$.*

Remarks. Proposition 2.1 can be made to follow from Proposition 2.2 by expanding the integral defining $D_{\theta_1}\left(\nu^n \| r^n\right)$ and noting that

$$\begin{aligned} D_{\theta_1}\left(\nu^n \| r^n\right) &= E\left[\sum_{1}^{n} D\left(p_i \| r_i\right) \middle| \theta_1\right] - E\left[D\left(p^n \| \nu^n\right) | \theta_1\right] \\ &= R_{\theta_1}\left(p^n, r^n\right) - R_{\theta_1}\left(p^n, \nu^n\right). \end{aligned} \tag{11}$$

Proof. We note that

$$\begin{aligned} R_{\theta_1}\left(p^n, r^n\right) &= \int D_{\underline{\theta}_1^d}\left(p^n \| r^n\right) \omega\left(\underline{\theta}_2^d \mid \theta_1\right) d\underline{\theta}_2^d \\ &= \sum_{i=1}^{n} \int D_{\underline{\theta}_1^d}\left(p_i \| r_i\right) \omega\left(\underline{\theta}_2^d \mid \theta_1\right) d\underline{\theta}_2^d \\ &= \sum_{i=1}^{n} R_{\theta_1}\left(p_i, r_i\right). \end{aligned}$$

The summands may be decomposed, with the help of Fubini's theorem and (8), as $R_{\theta_1}\left(p_i, r_i\right) = R_{\theta_1}\left(p_i, q_i\right) + D\left(q_i \| r_i\right)$. Both terms are nonnegative, so the sum is clearly minimized by taking $r_i \equiv q_i$. □

Propositions 2.1 and 2.2 show that the best choice of $r^n\left(\underline{x}^n \mid \theta_1\right)$ is $q^n\left(\underline{x}^n \mid \theta_1\right) = \prod_{i=1}^{n} q_i\left(x_i \mid \theta_1\right)$, where $q_i\left(x_i \mid \theta_1\right)$ is the i^{th} marginal of $\nu^n\left(\underline{x}^n \mid \theta_1\right)$ specified in (4). However in practice r^n is used precisely because ν^n is not known; thus r^n must be somehow estimated from (a subset of) the data also. In the present paper we neglect this part of the problem and focus on what the "best case" analysis under q^n would be.

2.2 A Stein's Lemma interpretation

The basic data analyzed with (1) and (2) consists of i.i.d. vectors

$$(\Theta_{11}, \underline{X}_1^n), \ldots, (\Theta_{1m}, \underline{X}_m^n),$$

from m individuals, where the subvectors $\underline{X}_j^n$ are actually observed, one for each individual, and the Θ_{1j} are latent (unobserved) variables. Consider a statistical test which helps determine whether the fictional likelihood r^n is "close enough" to the true, dependent likelihood ν^n. A pair of hypotheses which leads to another interpretation of the Kullback-Leibler numbers we have calculated above is

$$H_0 : \omega(\theta_1)\nu^n(\underline{X}^n|\theta_1) \; \textit{versus} \; H_1 : \tau(\theta_1)r^n(\underline{X}^n|\theta_1)$$

where $r_{\theta_1}^n \equiv r^n(\underline{x}^n \mid \theta_1)$ is any fixed independence density, and τ is any marginal density for θ_1. Let φ be the indicator function for a rejection region for H_0, which we denote by A_φ^c, neglecting the possibility of point masses requiring randomization.

Stein's test for H_0 vs. H_1 has the acceptance region

$$A_{\text{Stein } \tau r_{\theta_1}^n, \epsilon} = \left\{ \left| \frac{1}{m} \sum_{j=1}^{m} \log \frac{\nu^n(\underline{X}_j^n|\theta_{1j})\omega(\theta_{1j})}{r^n(\underline{X}_j^n|\theta_{1j})\tau(\theta_j)} - D' \right| < \epsilon \right\},$$

for H_0, where $D' = D(\omega||\tau) + \int D(\nu_{\theta_1}^n||r_{\theta_1}^n)\omega(\theta_1)d\theta_1$. It is well known that Stein's test has type II error satisfying

$$[1 - o(\underline{1})]e^{-m(D'+\epsilon)} \leq P_{\tau r_{\theta_1}^n}\left(A_{\text{Stein } \tau r_{\theta_1}^n, \epsilon}\right) \leq e^{-m(D'-\epsilon)}.$$

As a result, if we choose ϵ so that Stein's test is level α, and we let $A_{\varphi\tau r_{\theta_1}^n}$ be the acceptance region for some other level α test φ then from the proof of the lower bound part of Proposition 3.C in Clarke and Barron (1990), we have that

$$P_{\tau r_{\theta_1}^n}\left(A_{\varphi\tau r_{\theta_1}^n}\right) \geq e^{m(D'+\epsilon)}[P_{\omega\nu_{\theta_1}^n}(A_{\varphi\tau r_{\theta_1}^n}) - P_{\omega\nu_{\theta_1}^n}\left(A_{\text{Stein } \tau r_{\theta_1}^n, \epsilon}\right)].$$

Since $P_{\omega\nu_{\theta_1}^n}(A_{\varphi\tau r_{\theta_1}^n})$ is greater than $1 - \alpha$, Stein's test is level α, and the probability of of type II error in Stein's test is bounded above by $e^{-m(D'-\epsilon)}$, we have that

$$P_{\tau r_{\theta_1}^n}(A_{\varphi\tau r_{\theta_1}^n}) \geq (1 - 2\alpha)e^{-2m\epsilon}P_{\tau r_{\theta_1}^n}\left(A_{\text{Stein } \tau r_{\theta_1}^n, \epsilon}\right). \tag{12}$$

Now replace the $1-2\alpha$ by $1-2\alpha-\eta$, where η is small enough that (12) remains nontrivial, and let Γ be a collection of densities of the form of H_1 above, containing $\omega q^n_{\theta_1}$. Consider the simple versus composite hypothesis test

$$H: \omega(\theta_1)\nu^n_{\theta_1} \text{ versus } K: \tau(\theta_1) r^n_{\theta_1} \in \Gamma.$$

Proposition 2.3 *Assume there is an $\eta > 0$ so that (12) holds uniformly over Γ then the Stein test based on $A_{\text{Stein } \omega q^n_{\theta_1}, \epsilon}$ is near asymptotically minimax in the sense that*

$$\lim_{\epsilon \to 0+} \lim_{m \to \infty} \frac{1}{m} \left(\log \min_{\varphi} \max_{\tau r^n_{\theta_1}} P_{\tau r^n_{\theta_1}}(A_\varphi) - \log P_{H_0}(A_{\text{Stein } \omega q^n_{\theta_1}, \epsilon}) \right) = 0. \tag{13}$$

Remarks. Thus, for some choices of Γ, the Stein test for H_0 *versus* H_1 with $\tau = \omega$ and $r^n_{\theta_1} = q^n_{\theta_1}$, is near asymptotically minimax. We note that the probability of type I error for the near asymptotically minimax test is

$$P_{H_0}(A^c) = P_{w\nu^n_{\theta_1}} \left(\left| \frac{1}{m} \sum_{j=1}^{m} \log \frac{\nu^n(\underline{X}^n_j | \theta_{1j})}{q^n(\underline{X}^n_j | \theta_{1j})} - D \right| > \epsilon \right),$$

where $D = \int D(\nu^n_{\theta_1} || q^n_{\theta_1}) \omega(\theta_1) d\theta_1$, which is a large deviation. Typically (e.g., Stroock, 1984) there will exist a rate function I_ϵ such that

$$\lim_{m \to \infty} \frac{1}{m} \log P_{H_0}(A^c_{\text{Stein } \omega q^n_{\theta_1}, \epsilon}) = I_\epsilon.$$

This means that we have, in principle, exact rates for the decrease of probabilities of errors for the minimax test.

Proof. A minimax test achieves $\max_\varphi \min_{\tau r^n_{\theta_1} \in \Gamma} P_{\tau r^n_{\theta_1}}(A^c_\varphi) = 1 - \min_\varphi \max_{\tau r^n_{\theta_1} \in \Gamma} P_{\tau r^n_{\theta_1}}(A_\varphi)$. Typically such tests exist (e.g. Lehmann, 1959, p. 341). For fixed $\epsilon > 0$,

$$\begin{aligned} \frac{1}{m} \log \min_{\varphi} \max_{\tau r^n_{\theta_1} \in \Gamma} P_{\tau r^n_{\theta_1}}(A_\varphi) &\leq \frac{1}{m} \log \max_{\tau r^n_{\theta_1} \in \Gamma} P_{\tau r^n_{\theta_1}}(A_{\text{Stein } \tau r^n_{\theta_1}, \epsilon}) \\ &\leq \frac{1}{m} \log \max_{\tau r^n_{\theta_1} \in \Gamma} e^{-m(D' - \epsilon)} \\ &= -D + \epsilon, \end{aligned}$$

since $\min_{\tau r^n_{\theta_j} \in \Gamma} D' = D$, by choice of $\tau = \omega$ and $r^n_{\theta_1} = q^n_{\theta_1}$. For a lower bound we note that by the uniformity of (12) over Γ that

$$\frac{1}{m} \log \min_{\varphi} \max_{\tau r^n_{\theta_j} \in \Gamma} P_{\tau r^n_{\theta_1}}(A_\varphi) \geq c \frac{1}{m} \log \max_{\tau r^n_{\theta_1} \in \Gamma} P_{\tau r^n_{\theta_1}} \left(A_{\text{Stein} \tau r^n_{\theta_1} \epsilon} \right) e^{-2m\epsilon}$$

where c is a positive constant. Using the uniformity of (12) over Γ, we have

$$\begin{aligned}\frac{1}{m}\log\min_{\varphi}\max_{\tau r^n_{\theta_j}\in\Gamma} P_{\tau r^n_{\theta_1}}(A_\varphi) &\geq \frac{1}{m}\log\max_{\tau r^n_{\theta_1}\in\Gamma} c(1-o(1))e^{-2m\epsilon}e^{-m(D'+\epsilon)} \\ &= -(D+\epsilon)+\frac{1}{m}\log c - 2\epsilon + \frac{1}{m}\log(1-O(1)).\end{aligned}$$

Thus we see that $\lim_{\epsilon\to 0+}\lim_{m\to\infty}\frac{1}{m}\log\min_\varphi\max_{\tau r^n_{\theta_j}\in\Gamma} P_{\tau r^n_{\theta_1}}(A_\varphi) = -D$. Since it is apparent that D is the exponent for the probability of type II error for the test based on $A_{\text{Stein }\omega q^n_{\theta_1},\epsilon}$ as $\epsilon\to 0^+$, the proposition is established. □

Since Stein's test is fairly sensitive, we have shown that the hardest independence model to test against is the product of marginals. Even though the minimax test is not ideal, it has some positive aspects. The rates of decrease for the probabilities of both type I and type II error are known, and they tend to zero exponentially. The minimaxity ensures that no test can have uniformly smaller probabilities of type I and type II error. Since we have used the minimality of $D(\nu^n_{\theta_1}||q^n_{\theta_1})$ pointwise in θ_1 this test is in accord with the estimation optimality proved in Section 2.1.

Finally, let us return to the methodological question raised in Section 1: Under what circumstances can a convenient independence model τr^n be substituted for a correct but intractable dependence model $\omega\nu^n$ for the purposes of making inferences about the latent variable Θ_1? The test H_0 *versus* H_1 asks "May we reject the correct model in favor of the convenient one?" (to which the desired answer is "yes"). In this sense, the test is our way of asking permission from the data to use the convenient independence model. However, the reverse question might be thought more appropriate: the test

$$H_0'\text{: } \tau r^n \textit{ versus } H_1'\text{: } \omega\nu^n$$

asks "Must we reject independence in favor of the mixture model?" (to which the desired answer is "no"). The problem with such an approach is its intractability: efforts to extend the simple versus simple case to cases in which even one of the hypotheses is composite run into problems with existence of the information projection, or with the restriction to independence models.

3 Direct analysis of q^n under ν^n

Our results are simplest to present, and most easily interpreted, when consideration of the dependence on $\underline{\theta}_2^d$ is suppressed. In this section, only the dependent measure $\nu^n(\cdot \mid \theta_1)$, its one-dimensional marginals $q_i(x_i \mid \theta_1)$, and the product measure $q^n(\underline{x}^n \mid \theta_1) = \prod_{i=1}^n q_i(x_i \mid \theta_1)$ are used. The law governing $\underline{X}^n$ is at all times $\nu^n(\underline{x}^n \mid \theta_1)$; but the likelihood we will analyze is $q^n(\underline{x}^n \mid \theta_1)$.

3.1 Consistency of the wrong-model MLE $\hat{\theta}_{q,1}$

Suppose that an M-estimator $\hat{\theta}_{q,1}(\underline{X}^n)$ is formally constructed as the MLE from $q^n(\underline{X}^n \mid \theta_1)$. Junker (1991) considers this estimator for discrete-valued $\underline{X}^n$ in a setting appropriate to educational measurement, and uses Cramér-style arguments to establish the weak consistency of $\hat{\theta}_{q,1}(\underline{X}^n)$ under the law ν^n, assuming Stout's (1990) essential independence condition (9). We present a Wald-style argument showing that consistency of the "wrong-model" $\hat{\theta}_{q,1}(\underline{X}^n)$ is quite widely true, under a generalization of the EI condition.

We will first develop general consistency conditions for the wrong-model MLE $\hat{\theta}_{q,1}$ that will also be useful in Section 4, and then we will show how the result applies under (9). Define

$$L_n(\theta) = \log q^n(\underline{X}^n \mid \theta) = \sum_{i=1}^n \log q_i(X_i|\theta). \tag{14}$$

and

$$D_n(\theta_1, \theta) \equiv \frac{1}{n}[L_n(\theta_1) - L_n(\theta)] = \frac{1}{n}\sum_{i=1}^n \log \frac{q_i(X_i|\theta_1)}{q_i(X_i|\theta)}. \tag{15}$$

Again abbreviating $q_\theta^n(\cdot) \equiv q^n(\cdot \mid \theta)$, we may use (4) to show that, under ν^n, $E[D_n(\theta_1, \theta)|\,\theta_1] = \frac{1}{n} D\left(q_{\theta_1}^n \middle\| q_\theta^n\right)$. $L_n(\theta)$ would be the log-likelihood under independence, but we are *not* assuming independence here: i.e., $q_\theta^n(\cdot)$ may not be the true likelihood function. Finally, for each $t \in \Omega_{\Theta_1}$, define $B_\delta(t) \equiv \{\theta \in \Omega_{\Theta_1} : |\theta - t| < \delta\}$. In the present context, Wald's key assumptions can be stated as follows.

Assumption C1. For each θ_1 and $t \neq \theta_1$, there exists $c(t) > 0$, such that

$$\lim_{n\to\infty} P[D_n(\theta_1, t) > c(t)|\,\theta_1] \;=\; 1.$$

Assumption C2. For all $t \neq \theta_1$ and all $\xi > 0$, there exists $\delta > 0$ such that

$$\lim_{n\to\infty} P\left[\inf_{\theta\in B_\delta(t)} D_n(t,\theta) \geq -\xi \middle| \theta_1\right] = 1.$$

Assumption C3. There exist $c_\Delta > 0$, such that for all $\delta > 0$ and Δ sufficiently large (depending on δ) that

$$\liminf_{n\to\infty} P\left[\inf_{|\theta|>\Delta} D_n(\theta_1,\theta) > c_\Delta \middle| \theta_1\right] \geq 1-\delta.$$

Under these assumptions we obtain the following proposition which gives Wald-style consistency, in that the usual asymptotic convexity condition holds: $L_n(\theta_1)$ dominates $L_n(\theta)$ as $n \to \infty$, for all θ "away from" θ_1. The domination will be used below to establish asymptotic posterior normality by Laplace's method. The proof of Proposition 3.1, which is straightforward, is deferred to the end of this section.

Proposition 3.1 *Under Assumptions C1 through C3, for all $\epsilon > 0$ and all $\delta > 0$, there exists $\gamma = \gamma(\epsilon,\delta) > 0$ such that*

$$\liminf_{n\to\infty} P\left[\inf_{\theta\notin B_\epsilon(\theta_1)} \frac{1}{n}[L_n(\theta_1) - L_n(\theta)] \geq \gamma \middle| \theta_1\right] \geq 1-\delta \tag{16}$$

and hence the formal MLE $\hat{\theta}_{q,1}(\underline{x}^n) \overset{\nu^n}{\to} \theta_1$ as $n \to \infty$ (where "$\overset{\nu^n}{\to}$" denotes convergence in ν^n-probability).

Assumptions C1–C3 are what is needed to make the proof work. In addition, it is useful to identify more readily interpretable sufficient conditions for C1 and C2. Ideally we would like Proposition 3.1 under the following.

Assumption EI. Under $\nu^n(\cdot \mid \theta_1)$, as $n \to \infty$,

$$\frac{1}{n}\sum_1^n \{a_i(X_i) - E[a_i(X_i)|\,\theta_1]\} \overset{\nu^n}{\to} 0,$$

for all sequences of uniformly bounded functions $\{a_i(\cdot)\}$.

This is a generalization of Stout's essential independence condition (9). Like Stout's original condition it is a weak law of large numbers for bounded transformations of the random variables X_i $(i = 1, 2, \ldots)$. However, for Wald-style calculations, we require a LLN that holds for sums of log-contrast functions $D_n(\theta_1, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{q_i(X_i|\theta_1)}{q_i(X_i|\theta)}$, whose summands need not be bounded. We make such an assumption in Lemma 3.1. Alternatively, one might adapt the Cramér proof given in Junker (1991) to produce a conclusion like (16). Then the additional LLN for log-contrast functions would not be needed.

Lemma 3.1 *Suppose*

(a) For each $t \neq \theta_1$ there exists $\beta(t) > 0$ such that

$$\liminf_{n\to\infty} \frac{1}{n} D\left(q^n_{\theta_1} \,\|\, q^n_t\right) \geq \beta(t).$$

(b) As $n \to \infty$,

$$D_n(\theta_1, t) - \frac{1}{n} D\left(q^n_{\theta_1} \,\|\, q^n_t\right) \xrightarrow{\nu^n} 0.$$

Then Assumption C1 holds.

Remarks. (a) can be seen to be a kind of minimum information or identifiability condition. In Section 5.1 we will see that for typical binary response data, Assumption EI implies (b).

Proof. By (a), there exists $\beta = \beta(t) > 0$ such that for all large n, $\frac{1}{n} D\left(q^n_{\theta_1} \,\|\, q^n_t\right) > \beta$. Therefore,

$$\begin{aligned}
\lim_{n\to\infty} P[D_n(\theta_1, t) > \beta/2 | \theta_1] &= \lim_{n\to\infty} P\{D_n(\theta_1, t) - \frac{1}{n} D\left(q^n_{\theta_1} \,\|\, q^n_t\right) > \beta/2 - \frac{1}{n} D\left(q^n_{\theta_1} \,\|\, q^n_t\right) | \theta_1\} \\
&\geq \lim_{n\to\infty} P\{D_n(\theta_1, t) - \frac{1}{n} D\left(q^n_{\theta_1} \,\|\, q^n_t\right) > \beta/2 - \beta = -\beta/2 | \theta_1\} \\
&\geq \lim_{n\to\infty} P\{\left| D_n(\theta_1, t) - \frac{1}{n} D\left(q^n_{\theta_1} \,\|\, q^n_t\right) \right| < \beta/2 | \theta_1\} \\
&= 1,
\end{aligned}$$

by (b). Now take $c(t) \equiv \beta/2$ to obtain Assumption C1. □

Lemma 3.2 *Suppose that, for all $t \neq \theta_1$ there exists $\delta_t > 0$ such that*

(a) $\forall\, \xi > 0\ \exists\, \delta \in (0, \delta_t)$, *such that*

$$\liminf_{n\to\infty} \inf_{\theta\in B_\delta(t)} E[D_n(t,\theta)|\,\theta_1] > -\xi;$$

(b) $\forall\, \xi > 0\ \exists\, \delta \in (0, \delta_t)$ *such that*

$$\lim_{n\to\infty} P\left[\sup_{\theta\in B_\delta(t)} |D_n(t,\theta) - E[D_n(t,\theta)|\,\theta_1]| < \xi \middle| \theta_1\right] = 1.$$

Then Assumption C2 holds.

Remarks. Under mild continuity and regularity conditions it follows that for all n, and all t, $\lim_{\theta\to t} E[D_n(t,\theta)|\,\theta_1] = \lim_{\theta\to t} \frac{1}{n}[D\left(q^n_{\theta_1} \middle\| q^n_\theta\right) - D\left(q^n_{\theta_1} \middle\| q^n_t\right)] = 0$. Hence (a) is a locally uniform one-sided version of this continuity condition on the map $\theta \mapsto q^n_\theta$. On the other hand, it follows from (b) that $D_n(t,\theta) - E[D_n(t,\theta)|\,\theta_1] \xrightarrow{\nu^n} 0$ pointwise in θ. Hence (b) is a locally uniform version of this WLLN.

Proof. By (a), we may choose $\delta \in (0, \delta_t)$, such that $\inf_{\theta\in B_\delta(t)} E[D_n(t,\theta)|\,\theta_1] \geq -\xi/2$ for all large n. By (b) we may make $\delta > 0$ enough smaller that $P\left[\sup_{\theta\in B_\delta(t)} |D_n(t,\theta) - E[D_n(t,\theta)|\,\theta_1]| < \xi/2 \middle| \theta_1\right]$ $\to 1$, and hence

$$\begin{aligned}
&P\left[\inf_{\theta\in B_\delta(t)} D_n(t,\theta) > -\xi \middle| \theta_1\right] \\
&\geq\ P\left[\inf_{\theta\in B_\delta(t)} E[D_n(t,\theta)|\,\theta_1] - \sup_{\theta\in B_\delta(t)} |D_n(t,\theta) - E[D_n(t,\theta)|\,\theta_1]| > -\xi \middle| \theta_1\right] \\
&\geq\ P\left[\sup_{\theta\in B_\delta(t)} |D_n(t,\theta) - E[D_n(t,\theta)|\,\theta_1]| < -\xi/2 + \xi \middle| \theta_1\right] \\
&\to\ 1.
\end{aligned}$$

Thus Assumption C2 holds. □

Proof of Proposition 3.1. Let $\Omega_{\Theta_1} = S_\Delta \cup C \cup B_\epsilon(\theta_1)$ where $S_\Delta = \{\theta :\ |\theta| > \Delta\}$, $C = \Omega_{\Theta_1} \setminus [S_\Delta \cup B_\epsilon(\theta_1)]$, and ϵ is fixed in (16). Given $\delta > 0$ in (16), fix Δ so large that

$$\liminf_{n\to\infty} P\left[\inf_{\theta\in S_\Delta} D_n(\theta_1,\theta) > c_\Delta \middle| \theta_1\right] \geq 1 - \delta \tag{17}$$

for some $c_\Delta > 0$, by Assumption C3. For $t \neq \theta_1$, take $\gamma(t) = c(t)/2$ from Assumption C1 and take $\xi = \gamma(t)/2$. Then for δ as in Assumption C2,

$$\begin{aligned}
\lim_{n\to\infty} P\left[\inf_{\theta\in B_\delta(t)} D_n(\theta_1,\theta) > \gamma(t)\middle|\theta_1\right] &= \lim_{n\to\infty} P\left[D_n(\theta_1,t) + \inf_{\theta\in B_\delta(t)} D_n(t,\theta) > \gamma(t)\middle|\theta_1\right] \\
&\geq \lim_{n\to\infty} P[D_n(\theta_1,t) + (-2)\cdot\gamma(t)/2 > \gamma(t)|\,\theta_1] \\
&= \lim_{n\to\infty} P[D_n(\theta_1,t) > 2\cdot\gamma(t) = c(t)|\,\theta_1] \\
&= 1. \qquad (18)
\end{aligned}$$

For fixed Δ, C is a compact set and so can be covered by finitely many balls $S_1 = B_{\delta_1}(t_1)$, ..., $S_m = B_{\delta_m}(t_m)$, such that (18) holds for each: $\lim_{n\to\infty} P\left[\inf_{\theta\in S_j} D_n(\theta_1,\theta) > \gamma_j\middle|\theta_1\right] = 1$, $j = 1,\ldots,m$. Then, letting $\gamma = \min\{\gamma_1,\ldots,\gamma_m,c_\Delta\}$, we have

$$\begin{aligned}
\liminf_{n\to\infty} P\left[\inf_{\theta\notin B_\epsilon(\theta_1)} D_n(\theta_1,\theta) \geq \gamma\middle|\theta_1\right] &\geq \liminf_{n\to\infty} P\left[\bigcap_{j=1,\ldots,m,\Delta}\{\inf_{\theta\in S_j} D_n(\theta_1,\theta)\geq\gamma\}\middle|\theta_1\right] \\
&\geq 1-\delta,
\end{aligned}$$

using (17) for S_Δ and (18) for each S_j, $j = 1,\ldots,m$. This is (16). □

3.2 The asymptotic distribution of $\hat{\theta}_{q,1}$

To make the estimator $\hat{\theta}_{q,1}(\underline{x}^n)$ a useful inferential tool we need to know something about the asymptotic behavior of the law

$$\mathcal{L}\left\{\sqrt{n}\frac{\hat{\theta}_{q,1}-\theta_1}{\sigma_n(\underline{x}^n)}\right\},$$

for some appropriate scale term $\sigma_n(\theta_1)$. Pursuing the usual Taylor expansion of the log-likelihood, we see that as usual

$$\sqrt{n}\,(\theta_1 - \hat{\theta}_{q,1}) = \frac{\sqrt{n}\,\overline{L}'_n(\theta_1)}{\overline{J}_n(\tilde{\theta}_1)}, \qquad (19)$$

where $\overline{L}'_n(\theta_1) = \frac{\partial}{\partial\theta_1}\frac{1}{n}L_n(\theta_1)$, and $\overline{J}_n(\tilde{\theta}_1) \equiv -\frac{1}{n}\frac{\partial^2}{\partial\theta_1{}^2}\log q^n(\underline{x}^n|\tilde{\theta}_1)$ for some $\tilde{\theta}_1 \in \{\theta : |\theta-\theta_1| < |\hat{\theta}_{q,1}-\theta_1|\}$. Assumptions such as those in Section 3.3 (see especially Assumption PN1 and Assumption PN3 below) guarantee that $\overline{J}_n(\tilde{\theta}_1)$ will behave well; so the main burden is the behavior of the sum

$$\sqrt{n}\,\overline{L}'_n(\theta_1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_1}\log q_i(X_i|\theta_1),$$

under ν^n. General conditions for asymptotic normality for dependent sums have been established by Dvoretzky (1972); particular cases that seem useful include mixing CLT's (Iosifescu and Theodorescu, 1969) and methods for associated random variables (Cox and Grimmett, 1984; Newman and Wright, 1982). Applications of these ideas to item response models are considered by Junker (1988, 1991).

In Section 4 we will consider another approach, in which the asymptotic behavior of $\hat{\theta}_{q,1}$ is first identified under p^n and then "marginalized" to produce a result under ν^n.

3.3 Posterior asymptotics

We now turn to the possibility of basing inference for θ_1 on the formal posterior distribution

$$\omega_q(\theta_1 \mid \underline{x}^n) = \frac{q^n(\underline{x}^n \mid \theta_1)\omega(\theta_1)}{\int_{-\infty}^{\infty} q^n(\underline{x}^n \mid \theta)\omega(\theta)\,d\theta}, \tag{20}$$

where $\omega(\theta_1)$ is the prior density on θ_1. Of course, the true posterior distribution is

$$\omega_\nu(\theta_1 \mid \underline{x}^n) = \frac{\nu^n(\underline{x}^n \mid \theta_1)\omega(\theta_1)}{\int_{-\infty}^{\infty} \nu^n(\underline{x}^n \mid \theta)\omega(\theta)\,d\theta}.$$

The point once again is to see whether a "wrong model analysis" based on q^n can work when ν^n is the correct conditional law.

Let us abbreviate $\hat{\theta}_n \equiv \hat{\theta}_{q,1}(\underline{X}^n)$ in what follows. The main result, Theorem 3.1, is that $\omega_q((\theta_1 - \hat{\theta}_n)/\sigma_n|\underline{x}^n)$ is asymptotically normal, in the sense of Walker (1969). The principal assumptions used are Assumption EI, local uniform continuity of $\frac{\partial^2}{\partial\theta_1^2} q_i(x_i \mid \theta_1)$, and the truth of Proposition 3.1. The standard error is the usual "independence" standard error, $\sigma_n = \{-L_n''(\hat{\theta}_n)\}^{-1/2}$ where L_n is defined as in (14), and there are no restrictions on the rate of convergence of $\hat{\theta}_n$ to θ_1.

Finally, although it is not emphasized in the remainder of the section, one does not have to use the right prior when calculating ω_q. The crucial assumptions are that q^n be constructed as the product of marginals of ν^n, that LLN's hold for ν^n, and that whatever prior is used in constructing ω_q be positive and continuous near the θ_1 that generated the data. More formally, we make the following regularity assumptions.

Assumption PN1. Let $I_i(\theta_1) = E\left[(\partial \log q_i(X_i|\theta_1)/\partial\theta_1)^2 \middle| \theta_1\right]$ and $\overline{I}_n(\theta) = \frac{1}{n}\sum_1^n I_i(\theta)$. We assume there exist $0 < \epsilon_{\theta_1} \leq M_{\theta_1} < \infty$ such that $\epsilon_{\theta_1} \leq \overline{I}_n(\theta_1) \leq M_{\theta_1}$, for all large n.

Assumption PN2. $\int \frac{\partial^2}{\partial \theta_1^2} q_i(x|\theta_1)dx = 0$;

Remarks. Hence the expected Fisher information can be found in a Taylor expansion for $L_n(\theta_1)$ via $I_i(\theta_1) = -E\left[\partial^2 \log q_i(X_i|\theta_1)/\partial {\theta_1}^2 \middle| \theta_1\right]$, $\forall\ i$. Although it is not needed for the proof, it may be natural to also assume $\int \frac{\partial}{\partial \theta_1} q_i(x|\theta_1)dx = 0$, so that $L_n'(\theta_1) = 0$ is an unbiased estimating equation for θ_1 (i.e., the expected score function $E[\frac{1}{n}L_n'(\theta_1)|\theta_1] = 0$).

Assumption PN3. There exists $\epsilon = \epsilon(\theta_1) > 0$, such that $M_{\epsilon,i}(x, \theta_1) = \sup_{\theta \in B_\epsilon(\theta_1)} \left| \frac{\partial^2}{\partial \theta^2} \log q_i(x|\theta) - \frac{\partial^2}{\partial \theta_1^2} \log q_i(x|\theta_1) \right|$ is bounded uniformly in x and i; and for $\overline{M}_n(\epsilon, \theta_1) = \frac{1}{n}\sum_1^n M_{\epsilon,i}(X_i, \theta_1)$,

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} E\left[\overline{M}_n(\epsilon, \theta_1) \middle| \theta_1\right] = 0.$$

Assumption PN4. The prior density $\omega(\theta)$ is positive and continuous throughout a small neighborhood of θ_1.

Before proving Theorem 3.1, we require a preliminary proposition which allows us to approximate $\overline{I}_n(\theta_1)$ with $-\frac{1}{n}L_n''(\hat{\theta}_1)$ in the usual way.

Proposition 3.2 *Suppose $\hat{\theta}_{q,1}(\underline{x}^n) \xrightarrow{\nu^n} \theta_1$, Assumptions PN1 through PN3 and Assumption EI hold.*

(a) Let $\theta_n^* = \hat{\theta}_n + r(\theta_1 - \hat{\theta}_n)$, where $r \in [0,1]$, and let $B_\epsilon(\theta_1)$ be as in Assumption PN3. Then for all $\xi > 0$ there exists ϵ sufficiently small that

$$\lim_{n \to \infty} P\left[\sup_{\{r:\theta_n^* \in B_\epsilon(\theta_1)\}} \left| \frac{1}{n} L_n''(\theta_n^*) + \overline{I}_n(\theta_1) \right| < \xi \middle| \theta_1 \right] = 1.$$

(b) In particular, $\frac{1}{n}L_n''(\hat{\theta}_n) + \overline{I}_n(\theta_1) \xrightarrow{\nu^n} 0$ as $n \to \infty$.

Proof. By Assumption PN2, $\overline{I}_n(\theta_1) = -\frac{1}{n}E[L_n''(\theta_1)|\theta_1]$; hence it suffices to show each of the following, for all $\xi > 0$:

$$P\left[\frac{1}{n}|L_n''(\theta_1) - E\left[L_n''(\theta_1)|\theta_1\right]| < \xi \middle| \theta_1\right] \to 1; \tag{21}$$

$$P\left[\sup_{\theta_n^* \in B_\epsilon(\theta_1)} \frac{1}{n}|L_n''(\theta_n^*) - L_n''(\theta_1)| < \xi \middle| \theta_1\right] \to 1. \tag{22}$$

The limit (21) follows from Assumptions EI and PN3. For (22), let $\epsilon > 0$ be small enough that Assumption PN3 holds, with $E\left[\overline{M}_n(\epsilon,\theta_1)\middle|\theta_1\right] < \xi/2$, for all large n. By assumption, both $\hat{\theta}_n \overset{\nu^n}{\to} \theta_1$ and $\theta_n^* \overset{\nu^n}{\to} \theta_1$; hence

$$\begin{aligned}
&\lim_{n\to\infty} P\left[\frac{1}{n}|L_n''(\theta_n^*) - L_n''(\theta_1)| < \xi\middle|\theta_1\right] \\
&= \lim_{n\to\infty} P\left[\theta_n^* \in B_\epsilon(\theta_1),\ \frac{1}{n}|L_n''(\theta_n^*) - L_n''(\theta_1)| < \xi\middle|\theta_1\right] \\
&\geq \lim_{n\to\infty} P\left[\overline{M}_n(\epsilon,\theta_1) < \xi\middle|\theta_1\right] \qquad (23) \\
&= \lim_{n\to\infty} P\{\overline{M}_n(\epsilon,\theta_1) - E\left[\overline{M}_n(\epsilon,\theta_1)\middle|\theta_1\right] < \xi - E\left[\overline{M}_n(\epsilon,\theta_1)\middle|\theta_1\right] \mid \theta_1\} \\
&\geq \lim_{n\to\infty} P\{|\overline{M}_n(\epsilon,\theta_1) - E\left[\overline{M}_n(\epsilon,\theta_1)\middle|\theta_1\right]| < \xi/2 \mid \theta_1\} \\
&= 1,
\end{aligned}$$

by Assumptions EI and PN3. Note that the bound in (23) is uniform on $B_\epsilon(\theta_1)$, giving the uniformity in (22). □

Theorem 3.1 *Assume the conclusion of Proposition 3.1 and Assumption EI. Under the additional assumptions PN1 through PN4,*

$$\sigma_n = \{-L_n''(\hat{\theta}_n)\}^{-1/2} \geq 0$$

(by Proposition 3.2 and Assumption PN1, $\sqrt{n}\sigma_n$ exists and is bounded away from 0 and ∞ with probability tending to 1 as $n \to \infty$). Then, for all $a < b$,

$$\int_{\hat{\theta}_n + a\sigma_n}^{\hat{\theta}_n + b\sigma_n} \omega_q(\theta \mid \underline{X}^n)\, d\theta \overset{\nu^n}{\to} \Phi(b) - \Phi(a) \qquad (24)$$

as $n \to \infty$, where $\Phi(\cdot)$ is the the standard normal c.d.f.

Remarks. The only place that the LLN's (Assumption EI and (b) of Lemma 3.1) are needed is in the proofs of Propositions 3.1 and 3.2. Thus we *could* replace reliance on Assumption EI with reliance on the conclusion of Proposition 3.2.

Let us break up the integral in (24) as follows:

$$\int_{\hat{\theta}_n + a\sigma_n}^{\hat{\theta}_n + b\sigma_n} \omega_q(\theta|\underline{X}^n)d\theta \quad = \quad \frac{\int_{\hat{\theta}_n + a\sigma_n}^{\hat{\theta}_n + b\sigma_n} q^n(\underline{X}^n|\theta)\omega(\theta)d\theta}{\int_{-\infty}^{\infty} q^n(\underline{X}^n|\theta)\omega(\theta)d\theta}$$

$$= \frac{\int_{\hat{\theta}_n+a\sigma_n}^{\hat{\theta}_n+b\sigma_n} q^n(\underline{X}^n|\theta)\omega(\theta)d\theta}{\left[\int_{B_\epsilon(\theta_1)} + \int_{B_\epsilon(\theta_1)^c}\right] q^n(\underline{X}^n|\theta)\omega(\theta)d\theta}$$
$$\equiv \frac{I_3}{I_1 + I_2},$$

with ϵ to be determined below. We will examine these three integrals in the order in which they are numbered. Although the proof is standard, we present the main points in Claims 3.1 through 3.3, to show that the probability structure is really not at issue, once Propositions 3.1 and 3.2 are established.

Claim 3.1 *For all $\xi > 0$, there exists ϵ small enough that*

$$\lim_{n\to\infty} P\left[|I_1/\{\sigma_n q^n(\underline{X}^n|\hat{\theta}_n)\} - (2\pi)^{1/2}\omega(\theta_1)| < \xi \middle| \theta_1\right] = 1.$$

Proof. Using a two-term Taylor expansion of $L_n(\theta)$ about $\hat{\theta}_n$,

$$\begin{aligned} I_1/q^n(\underline{X}^n|\hat{\theta}_n) &= \int_{B_\epsilon(\theta_1)} \frac{q^n(\underline{X}^n|\theta)}{q^n(\underline{X}^n|\hat{\theta}_n)}\omega(\theta)d\theta \\ &= \int_{B_\epsilon(\theta_1)} \exp\left\{-\frac{(\theta-\hat{\theta}_n)^2}{2\sigma_n^2}(-L_n''(\theta_n^*)\sigma_n^2)\right\}\omega(\theta_1)\frac{\omega(\theta)}{\omega(\theta_1)}d\theta, \end{aligned}$$

where $\theta_n^* = \hat{\theta}_n + r(\theta_1 - \hat{\theta}_n) \overset{\nu^n}{\to} \theta_1$ with $\hat{\theta}_n$ (by Proposition 3.1). For ξ_1 and ξ_2 to be determined momentarily, fix $\epsilon > 0$ so small that, by Proposition 3.2,

$$\lim_{n\to\infty} P\left[\sup_{\theta_n^*\in B_\epsilon(\theta_1)}\left|\frac{L_n''(\theta_n^*)}{L_n''(\hat{\theta}_n)} - 1\right| < \xi_1 \middle| \theta_1\right] = 1; \tag{25}$$

and by Assumption PN4,

$$1 - \xi_2 \le \inf_{\theta\in B_\epsilon(\theta_1)}\frac{\omega(\theta)}{\omega(\theta_1)} \le \sup_{\theta\in B_\epsilon(\theta_1)}\frac{\omega(\theta)}{\omega(\theta_1)} \le 1 + \xi_2. \tag{26}$$

Equation (26) follows directly from Assumption PN4. To see (25), rewrite

$$\left|\frac{L_n''(\theta_n^*)}{L_n''(\hat{\theta}_n)} - 1\right| = \left|\frac{(\frac{1}{n}L_n''(\theta_n^*) + \overline{I}_n(\theta_1)) - (\overline{I}_n(\theta_1) + \frac{1}{n}L_n''(\hat{\theta}_n))}{\frac{1}{n}L_n''(\hat{\theta}_n)}\right|$$

and apply Proposition 3.2, together with the observation that, by Assumption PN1, $|L_n''(\hat{\theta}_n)|$ is bounded away from 0 with probability tending to 1 as $n \to \infty$.

Hence, using the fact that $P\left[\hat{\theta}_n, \theta_n^* \in B_\epsilon(\theta_1) \middle| \theta_1\right] \to 1$ as $n \to \infty$, and recalling the definition of σ_n, we obtain

$$\begin{aligned} P\Bigg[&\omega(\theta_1)(1-\xi_2)\int_{B_\epsilon(\theta_1)} \exp\left\{-\frac{(\theta-\hat{\theta}_n)^2}{2\sigma_n^2}(1+\xi_1)\right\} d\theta \\ &\leq I_1/q^n(\underline{X}^n|\hat{\theta}_n) \leq \omega(\theta_1)(1+\xi_2)\int_{B_\epsilon(\theta_1)} \exp\left\{-\frac{(\theta-\hat{\theta}_n)^2}{2\sigma_n^2}(1-\xi_1)\right\} d\theta \middle| \theta_1\Bigg] \to 1. \end{aligned} \tag{27}$$

The outer expressions in this inequality may be readily identified as

$$\begin{aligned} &\sigma_n \omega(\theta_1)(1 \mp \xi_2)[2\pi/(1 \pm \xi_1)]^{1/2} \\ &\quad \times \left[\Phi\{(1 \pm \xi_1)^{1/2}\sigma_n^{-1}[\theta_1 + \epsilon - \hat{\theta}_n]\} - \Phi\{(1 \pm \xi_1)^{1/2}\sigma_n^{-1}[\theta_1 - \epsilon - \hat{\theta}_n]\}\right] \end{aligned} \tag{28}$$

The factor involving Φ tends to 1 in probability, since $\hat{\theta}_n \xrightarrow{\nu^n} \theta_1$ and $\sigma_n^{-1} \xrightarrow{\nu^n} \infty$. For each fixed ξ, an appropriate choice of ξ_1 and ξ_2 finishes the proof. □

Claim 3.2 *For each fixed $\epsilon > 0$,*

$$I_2/\{\sigma_n q^n(\underline{X}^n|\hat{\theta}_n)\} \xrightarrow{\nu^n} 0.$$

Proof.

$$\begin{aligned} I_2/q^n(\underline{X}^n|\hat{\theta}_n) &= \int_{B_\epsilon(\theta_1)^c} \frac{q^n(\underline{X}^n|\theta)}{q^n(\underline{X}^n|\hat{\theta}_n)}\omega(\theta)d\theta \\ &= \sigma_n \exp\{L_n(\theta_1) - L_n(\hat{\theta}_n)\}\int_{B_\epsilon(\theta_1)^c} \sigma_n^{-1}\exp\{L_n(\theta) - L_n(\theta_1)\}\omega(\theta)d\theta \\ &\leq \sigma_n \cdot 1 \cdot \int_{B_\epsilon(\theta_1)^c} O_p(\sqrt{n})O_p(\exp[-n\gamma])\omega(\theta)d\theta\ , \end{aligned}$$

where the bound 1 follows from the fact that $L_n(\theta) \leq L_n(\hat{\theta}_n)$, $\forall\ \theta$ (by the definition of the MLE); and the O_p bounds, which are uniform in $|\theta - \theta_1| \geq \epsilon$, $\forall\ \epsilon$, follow from Proposition 3.2, Assumption PN1, and Proposition 3.1. □

Claim 3.3 *For all $\xi > 0$,*

$$\lim_{n\to\infty} P\left[|I_3/\{\sigma_n q^n(\underline{X}^n|\hat{\theta}_n)\} - (2\pi)^{1/2}\omega(\theta_1)[\Phi(b) - \Phi(a)]| < \xi \middle| \theta_1\right] = 1.$$

Proof. Let $N_n = (\hat{\theta}_n + a\sigma_n, \hat{\theta}_n + b\sigma_n)$, and fix ϵ for ξ_1 and ξ_2 as in Claim 3.1. Since $P[N_n \subset B_\epsilon(\theta_1) | \theta_1] \to 1$, the argument for I_3 proceeds just as for I_1, but with N_n replacing $B_\epsilon(\theta_1)$ throughout. In particular, we again have

$$\begin{aligned} I_3/q^n(\underline{X}^n|\hat{\theta}_n) &= \int_{N_n} \frac{q^n(\underline{X}^n|\theta)}{q^n(\underline{X}^n|\hat{\theta}_n)}\omega(\theta)d\theta \\ &= \int_{N_n} \exp\left\{-\frac{(\theta-\hat{\theta}_n)^2}{2\sigma_n^2}(-L_n''(\theta_n^*)\sigma_n^2)\right\}\omega(\theta_1)\frac{\omega(\theta)}{\omega(\theta_1)}d\theta, \end{aligned}$$

and N_n replaces $B_\epsilon(\theta_1)$ in (25), (26) and (27). We can apply the same continuity arguments as in Claim 3.1 to discover that $I_3/q^n(\underline{X}^n|\hat{\theta}_n)$ is bounded above and below, with probability tending to 1 as $n \to \infty$, by integrals of the form

$$\begin{aligned} &\int_{N_n} \exp\left\{-\frac{(\theta-\hat{\theta}_n)^2}{2\sigma_n^2}(1 \pm \xi_1)\right\}\omega(\theta_1)(1 \mp \xi_2)d\theta \\ &\quad \approx \sigma_n\omega(\theta_1)(1 \mp \xi_2)[2\pi/(1 \pm \xi_1)]^{1/2}\left[\Phi\{(1 \pm \xi_1)^{1/2}b\} - \Phi\{(1 \pm \xi_1)^{1/2}a\}\right]. \end{aligned}$$

The proof is now completed as for Claim 3.1. □

Corollary 3.1 *In addition to the hypotheses of the Theorem 3.1, suppose*

$$\int_{-\infty}^{\infty} |t| q^n(\underline{X}^n \,|\, t)\, \omega(t)\, dt < \infty \tag{29}$$

with ν^n-probability tending to 1 as $n \to \infty$. Then $E_q[\Theta_1 | \underline{X}^n] \xrightarrow{\nu^n} \theta_1$ under ν^n; in other words,

$$\int_{-\infty}^{\infty} \theta\, \omega_q(\theta \,|\, \underline{X}^n)\, d\theta \xrightarrow{\nu^n} \theta_1. \tag{30}$$

Proof. The proof proceeds as for Theorem 3.1, except that Claim 3.3 becomes

$$\lim_{n\to\infty} P\left[\left|I_3'/\{\sigma_n q^n(\underline{X}^n|\hat{\theta}_n)\} - (2\pi)^{1/2}\omega(\theta_1)[\int_a^b t\phi(t)\,dt + \hat{\theta}_n(\Phi(b) - \Phi(a))]\right| < \xi \middle| \theta_1\right] = 1,$$

where $\phi(t) = \Phi'(t)$ is the standard normal density and

$$I_3' = \int_{\hat{\theta}_n + a\sigma_n}^{\hat{\theta}_n + b\sigma_n} \theta q^n(\underline{X}^n|\theta)\omega(\theta)d\theta.$$

The proof of this assertion proceeds exactly as for Claim 3.3. Now let $a, -b \to \infty$. □

Remarks. The fact that asymptotic normality results should depend little on the true dependence structure of the data was made clear by Chen (1985). Kass, Tierney and Kadane (1990) have identified a class of models in which the Chen/Walker-style argument works well, called the "Laplace regular" models. We note that (i) the continuity and boundedness conditions of Laplace regularity correspond to our Assumption PN3; (ii) the positivity of the Hessian for Laplace regularity corresponds to our Assumption PN1 and Proposition 3.2, and (iii) their asymptotic convexity condition is our Proposition 3.1.

4 Analysis of q^n under ν^n using the full model p^n

In many settings it is natural to assume that there exists a "full model" p^n from which ν^n and q^n can be derived as in (7) and (8). In this section we show that $\hat{\theta}_{q,1}(\underline{x}^n)$ is consistent and converges in distribution to a mixture of normals under ν^n, and asymptotic normality of the posterior ω_q under ν^n, replacing LLN assumptions on ν^n with assumptions on the full model $p^n\left(\cdot \mid \underline{\theta}_1^d\right)$.

If we simply apply the results of Section 3 to the full model case, we see that, under finite moment conditions, the LLN assumptions on ν^n imply that the effect of the nuisance parameters $\underline{\theta}_2^d$ in the full model $p^n\left(\cdot \mid \underline{\theta}_1^d\right)$ disappears asymptotically. Consider Assumption EI in this context, which asserts that the left side of the identity

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} a_i(X_i)\middle|\theta_1\right) = \mathrm{Var}\left(E\left[\frac{1}{n}\sum_{i=1}^{n} a_i(X_i)\middle|\underline{\theta}_1^d\right]\middle|\theta_1\right) + E\left[\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} a_i(X_i)\middle|\underline{\theta}_1^d\right)\middle|\theta_1\right] \quad (31)$$

tends to zero as $n \to \infty$, for bounded $a_i(\cdot)$. Also, the second term on the right will tend to zero by the weak law of large numbers for p^n. Hence the remaining term in (31) tends to zero, from which we may conclude, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left\{E\left[a_i(X_i)|\,\underline{\theta}_1^d\right] - E\left[a_i(X_i)|\,\theta_1\right]\right\}\right| < \epsilon\middle|\theta_1\right] \to 1.$$

As a result, for $\omega\left(\underline{\theta}_2^d \mid \theta_1\right)$-almost every $\underline{\theta}_2^d$, the first moment is asymptotically free of $\underline{\theta}_2^d$. If the $a_i(X_i)$'s have uniformly bounded $(k+1)^{\text{st}}$ moments and EI holds then by a uniform integrability argument higher moments are also asymptotically free of $\underline{\theta}_2^d$.

Our goal is to obtain estimators for θ_1 which do not involve explicitly estimating or accounting for $\underline{\theta}_2^d$. The representation $\nu^n(\underline{x}^n \mid \theta_1) = \int \prod_{i=1}^n p_i\left(x_i \mid \underline{\theta}_1^d\right) \omega\left(\underline{\theta}_2^d \mid \theta_1\right) d\underline{\theta}_2^d$ allows us to replace the LLN assumptions for ν^n with LLN's which hold naturally for the independent likelihood p^n. For consistency of $\hat{\theta}_{q,1}$ it is enough to require that the dependence on $\underline{\theta}_2^d$ is uniformly small on compact sets of $\underline{\theta}_2^d$; see Example 5.3 for an example of this. The asymptotic normality of the q^n-based posterior distribution may be handled in a similar fashion. As for the asymptotic distribution of $\hat{\theta}_{q,1}$, both first and second moments of $a_i(X_i)$ must be asymptotically free of $\underline{\theta}_2^d$ to obtain a conventional asymptotic normality result; otherwise one obtains various mixtures of normals which are less easy to work with in general.

The asymptotic distribution theory under ν^n necessarily involves mixing over $\underline{\theta}_2^d$ and this means that there is no well-defined Fisher information resulting from p^n—unless it too is free of $\underline{\theta}_2^d$. Consequently, the asymptotic normality of the q^n-based posterior requires that the empirical Fisher information be used for scaling. The result breaks down (c.f. Proposition 4.3 below, as well as the remarks following Theorem 4.1) if one tries to use the expected Fisher information under q^n, $\overline{I}_n(\theta_1) = E\left[-\frac{1}{n}L_n''(\theta_1)\middle|\theta_1\right]$; and, although a result can be obtained if one uses the p^n-based form $\overline{I}_n(\theta_1;\underline{\theta}_2^d) = E\left[-\frac{1}{n}L_n''(\theta_1)\middle|\underline{\theta}_1^d\right]$, the result is of little interest since it involves scaling by quantities functionally dependent on $\underline{\theta}_2^d$.

Recasting the problem in terms of a larger conditional independence model $p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right) = \prod_{i=1}^n p_i\left(x_i \mid \underline{\theta}_1^d\right)$ allows us to make two interpretations of the experimenter's pragmatic approach to inference using q^n, in terms of the behavior of the "full model" p^n:

1. We can regard $q_i(x_i \mid \theta_1)$ as marginals $\int p_i\left(x_i \mid \underline{\theta}_1^d\right) \omega(\underline{\theta}_2^d|\theta_1)\, d\underline{\theta}_2^d$ of some more complicated model. In this case the conventional procedure is to use an estimator based on q^n *and to assess its performance in* q^n *also.* Our results allow its performance to be assessed in the proper measure ν^n, and show that the MLE based on q^n is asymptotically sensitive to distortions due to $\underline{\theta}_2^d$, while the q^n-based posterior is not.

2. Alternatively we can regard a convenient model $q_i(x_i \mid \theta_1) \equiv p_{0,i}(x_i \mid \theta_1)$ as embedded in a larger model $p_i\left(x_i \mid \underline{\theta}_1^d\right)$, where fixing $\underline{\theta}_2^d$ at some "null value" produces $p_{0,i}$; see for example Cox and Wermuth (1990). Now if we allow $\underline{\theta}_2^d$ to vary, our techniques give estimators—

the q^n-based MLE and posterior—which do not depend functionally on $\underline{\theta}_2^d$, although their moments typically do. This contrasts with the usual MLE for θ_1 in p^n—the first coordinate $\hat{\theta}_1$ of $\hat{\underline{\theta}}_1^d$—which typically will depend functionally on $\underline{\theta}_2^d$.

Both interpretations admit the possibility of using larger models that are typically not fully specified since estimators for the parameters of interest can still be found and their performance assessed for sensitivity to the nuisance parameters. In the usual case where the distribution of the nuisance parameters is unknown, it may be sensible to simply choose a convenient prior for them (e.g., Berger and Bernardo, 1989). However a fully noninformative prior would typically not be expected to satisfy our assumptions below.

4.1 Consistency of $\hat{\theta}_{q,1}$

Our first result is an extension of Proposition 3.1 in the context of p^n. With the notation exactly as in Section 3.1 we assume:

Assumption C1′. $\forall\, t \neq \theta_1$, $\exists\, c(t) \equiv c(t;\underline{\theta}_1^d) > 0$ such that

$$\lim_{n\to\infty} P\left[D_n(\theta_1, t) > c(t) | \underline{\theta}_1^d\right] = 1.$$

Assumption C2′. $\forall\, t \neq \theta_1$, $\forall\, \xi > 0$, $\exists\, \delta > 0$ such that

$$\lim_{n\to\infty} P\left[\inf_{\theta\in B_\delta(t)} D_n(t,\theta) > -\xi \middle| \underline{\theta}_1^d\right] = 1.$$

Assumption C3′. $\forall\, \Delta$ large, $\exists\, c_\Delta \equiv c_\Delta(\underline{\theta}_1^d) > 0$ such that for all $\delta > 0$,

$$\liminf_{n\to\infty} P\left[\inf_{|\theta|>\Delta} D_n(\theta_1,\theta) > c_\Delta \middle| \underline{\theta}_1^d\right] \geq 1-\delta.$$

Proposition 4.1 *Under Assumption C1′ through Assumption C3′, for all $\epsilon > 0$ and all $\delta > 0$ there exists $\gamma > 0$ such that*

$$\liminf_{n\to\infty} P\left[\inf_{\theta\notin B_\epsilon(\theta_1)} D_n(\theta_1,\theta) \geq \gamma \middle| \underline{\theta}_1^d\right] \geq 1-\delta. \tag{32}$$

The proof is identical to that of Proposition 3.1, except that all probabilities and expectations are conditional on $\underline{\theta}_1^d$, not θ_1. Note that laws of large numbers hold under a wide variety of independence models p^n; arguments about the plausibility of Assumption C1′ and C2′ reduce to verifying the appropriate moment conditions (cf. e.g. Theorem 5.2.3 of Chung, 1974). Note that the $c(t)$ in Assumption C1′ and the c_Δ in Assumption C3′ depend on $\underline{\theta}_2^d$. This suggests what kind of uniformity argument to make, to obtain consistency in ν^n from consistency in p^n. Note that $\mathcal{K}$ appears in the *hypotheses* but not in the *conclusion* of Corollary 4.1, and assume that $\omega\left(\underline{\theta}_2^d \mid \theta_1\right)$ is σ-finite. An example illustrating this result is presented in Example 5.3.

Corollary 4.1 *Suppose that for any compact set* $\mathcal{K} \subset \operatorname{supp} \omega(\underline{\theta}_2^d|\theta_1)$,

$$\inf_{\underline{\theta}_2^d \in \mathcal{K}} c(t) > 0, \tag{33}$$

$$\inf_{\underline{\theta}_2^d \in \mathcal{K}} c_\Delta > 0. \tag{34}$$

Then $\forall\ \epsilon$ *and* $\forall\ \delta$, $\exists\ \gamma$ *such that*

$$\liminf_{n\to\infty} P\left[\inf_{\theta \notin B_\epsilon(\theta_1)} D_n(\theta_1, \theta) \geq \gamma \middle| \theta_1\right] \geq 1 - \delta.$$

Proof. Let $\delta' > 0$ so small, and $\mathcal{K}$ so large, that $\omega(\mathcal{K}|\theta_1)(1-\delta') \geq 1-\delta$. The proof of Proposition 4.1 goes through as before, with $c(t)$ and c_Δ set equal to their infima over $\mathcal{K}$, guaranteed positive by (33) and (34). The γ that one obtains for δ' and ϵ thereby is now uniform over $\mathcal{K}$. Then by a Fatou's Lemma argument,

$$\begin{aligned}
\liminf_{n\to\infty} P\left[\inf_{\theta \notin B_\epsilon(\theta_1)} D_n(\theta_1, \theta) \geq \gamma \middle| \theta_1\right] &= \liminf_{n\to\infty} E\left[P\left[\inf_{\theta \notin B_\epsilon(\theta_1)} D_n(\theta_1, \theta) \geq \gamma \middle| \underline{\theta}_1^d\right] \middle| \theta_1\right] \\
&\geq E\left[1_{\{\theta_1\}\times\mathcal{K}} \liminf_{n\to\infty} P\left[\inf_{\theta \notin B_\epsilon(\theta_1)} D_n(\theta_1, \theta) \geq \gamma \middle| \underline{\theta}_1^d\right] \middle| \theta_1\right] \\
&\geq \omega(\mathcal{K}|\theta_1)(1-\delta') \\
&\geq 1 - \delta,
\end{aligned}$$

which completes the proof. □

4.2 Asymptotic distribution of $\hat{\theta}_{q,1}$

For the discussion of asymptotic distribution properties of the $\hat{\theta}_{q,1}(\underline{X}^n)$, it is convenient to consider again the Taylor expression (19). For brevity, we denote $\overline{U} = \overline{L}'_n(\theta_1)$ below.

Assumption AN1′. Let $\gamma_n(\underline{\theta}_1^d) = \sqrt{n}\, E\left[\overline{U} \middle| \underline{\theta}_1^d\right]$, and assume there exist functions $\sigma_n^2(\underline{\theta}_1^d) > 0$ such that

$$\frac{\sqrt{n}\,\overline{U} - \gamma_n(\underline{\theta}_1^d)}{\sigma_n(\underline{\theta}_1^d)} \sim AN(0,1)$$

under $p^n\left(\cdot \,|\, \underline{\theta}_1^d\right)$.

Assumption AN2′. There exists $\epsilon = \epsilon(\underline{\theta}_1^d) > 0$ such that $M_{\epsilon,i}(x,\theta_1) = \sup_{\theta \in B_\epsilon(\theta_1)} \left|\frac{\partial^2}{\partial \theta^2} \log q_i(x_i|\theta) - \frac{\partial^2}{\partial \theta_1^2} \log q_i(x_i|\theta_1)\right|$ is dominated by some p^n-integrable function, uniformly in i, and for $\overline{M}_n(\epsilon,\theta_1) = \frac{1}{n}\sum_{i=1}^n M_{\epsilon,i}(X_i,\theta_1)$,

$$\lim_{\epsilon \to 0} \limsup_{n\to\infty} E\left[\overline{M}_n(\epsilon,\theta_1) \middle| \underline{\theta}_1^d\right] = 0.$$

Assumption AN3′. $\hat{\theta}_{q,1}(\underline{X}^n) \xrightarrow{\nu^n} \theta_1$ under $\nu^n(\cdot \,|\, \theta_1)$.

Remarks. Note that, since p^n is a product measure, Assumption AN1′ is a fairly mild assumption requiring only, say, the Lindeberg-Feller conditions on the summands of $\overline{U}$.

Proposition 4.2 *Let z_α be the standard normal cutoff, $\alpha = \Phi(z_\alpha)$ for $Z \sim N(0,1)$, and assume Assumption AN1′ through Assumption AN3′. Then*

1. *For all t,*

$$\lim_{n\to\infty} P\left[\sqrt{n}\,(\theta_1 - \hat{\theta}_{q,1}(\underline{X}^n)) \le t \middle| \theta_1\right] - E\left[\Phi\left(\frac{t\overline{J}_n(\theta_1) - \gamma_n(\theta_1, \underline{\Theta}_2^d)}{\sigma_n(\theta_1, \underline{\Theta}_2^d)}\right) \middle| \theta_1\right] = 0.$$

2. *For all $\alpha \in (0,1)$, and any "centering" and "scale" terms $b(\theta_1)$ and $c(\theta_1)$,*

$$\lim_{n\to\infty} P\left[\frac{\sqrt{n}\,(\overline{U} - b(\theta_1))}{c(\theta_1)} \le z_\alpha \middle| \theta_1\right] - E\left[\Phi\left(\frac{z_\alpha c(\theta_1) - (\gamma_n(\theta_1, \underline{\Theta}_2^d) - \sqrt{n}\, b(\theta_1))}{\sigma_n(\theta_1, \underline{\Theta}_2^d)}\right) \middle| \theta_1\right] = 0$$

Proof. Recall from (19) that $\sqrt{n}(\theta_1 - \hat{\theta}_{q,1}) = \sqrt{n}\,\overline{U}/\overline{J}_n(\tilde{\theta}_1)$, where $\tilde{\theta}_1 \in \overline{B_{|\theta_1 - \hat{\theta}_{q,1}|}(\theta_1)}$. By Assumption AN3′, we may assume without loss that for some small δ, $\overline{B_{|\theta_1 - \hat{\theta}_{q,1}|}(\theta_1)} \subset B_\delta(\theta_1)$, since

this is true with probability approaching 1 as $n \to \infty$. Now for part 1, we calculate

$$\begin{aligned} & P\left[\sqrt{n}(\theta_1 - \hat{\theta}_{q,1}) \le t \middle| \theta_1\right] \\ &= E\left[P\left[\frac{\sqrt{n}\,\overline{U}}{\overline{J}_n(\tilde{\theta}_1)} \le t \middle| \underline{\theta}_1^d\right] \middle| \theta_1\right] \\ &= E\left[P\left[\frac{\sqrt{n}(\overline{U} - E\left[\overline{U}\middle|\underline{\theta}_1^d\right])}{\sigma_n(\underline{\theta}_1^d)} \le \frac{t\overline{J}_n(\tilde{\theta}_1) - \gamma_n(\underline{\theta}_1^d)}{\sigma_n(\underline{\theta}_1^d)} \middle| \underline{\theta}_1^d\right] \middle| \theta_1\right] \end{aligned}$$

and part 1 follows from the continuity in Assumption AN2′. For part 2,

$$\begin{aligned} & P\left[\frac{\sqrt{n}(\overline{U} - b(\theta_1))}{c(\theta_1)} \le z_\alpha \middle| \theta_1\right] \\ &= E\left[P\left[\frac{\sqrt{n}\,\overline{U} - \gamma_n(\underline{\theta}_1^d)}{\sigma_n(\underline{\theta}_1^d)} \le \frac{z_\alpha c(\theta_1) - (\gamma_n(\underline{\theta}_1^d) - \sqrt{n}\,b(\theta_1))}{\sigma_n(\underline{\theta}_1^d)} \middle| \underline{\theta}_1^d\right] \middle| \theta_1\right] \end{aligned}$$

by Assumption AN3′ and the same Taylor expansion as before. □

Remarks. Part 1 here shows the distortion of the usual confidence intervals based on $\hat{\theta}_{q,1}$. Part 2 shows what happens if we try to force centering and scaling terms which depend only on θ_1. If we insist on having a "standard" asymptotic normality result, we are faced with investigating the stability and fixed points of integral operators, as $n \to \infty$:

$$\begin{aligned} \alpha &= \lim_{n\to\infty} E\left[\Phi\left(\frac{z_\alpha \overline{J}_n(\theta_1) - \gamma_n(\theta_1, \underline{\Theta}_2^d)}{\sigma_n(\theta_1, \underline{\Theta}_2^d)}\right) \middle| \theta_1\right] \\ \alpha &= \lim_{n\to\infty} E\left[\Phi\left(\frac{z_\alpha c(\theta_1) - (\gamma_n(\theta_1, \underline{\Theta}_2^d) - \sqrt{n}\,b(\theta_1))}{\sigma_n(\theta_1, \underline{\Theta}_2^d)}\right) \middle| \theta_1\right] \end{aligned}$$

It is suggestive to consider the easy case in which we may interchange limit and expectation. For example, in part 2, we would require

$$z_\alpha = \lim_{n\to\infty} \frac{z_\alpha c(\theta_1) - (\gamma_n(\underline{\theta}_1^d) - \sqrt{n}\,b(\theta_1))}{\sigma_n(\underline{\theta}_1^d)}$$

for all z_α; clearly this requires $c(\theta_1)/\sigma_n(\underline{\theta}_1^d) \to 1$ and $\sqrt{n}(E[\overline{U}|\underline{\theta}_1^d] - b(\theta_1))/\sigma_n(\underline{\theta}_1^d) \to 0$.

4.3 Posterior asymptotics

Although Proposition 4.2 implies likelihood-based inference is complicated, the situation for posterior inference is more straightforward. As in Section 3.3, the principal ingredients are consistency

of $\hat{\theta}_n \equiv \hat{\theta}_{q,1}(\underline{x}^n)$ for θ_1, and the approximation of the asymptotic information function with an appropriate "scoring" function.

Assumption PN1′. For each $\underline{\theta}_1^d$, there exists $0 < \epsilon \leq M < \infty$ such that

$$\epsilon \leq \liminf_{n\to\infty} \overline{I}_n(\theta_1; \underline{\theta}_2^d) \leq \limsup_{n\to\infty} \overline{I}_n(\theta_1; \underline{\theta}_2^d) \leq M,$$

where $\overline{I}_n(\theta_1; \underline{\theta}_2^d) = -E\left[\frac{1}{n} L_n''(\theta_1) \middle| \underline{\theta}_1^d\right]$.

Assumption PN2′. The weak law of large numbers holds for $p^n\left(\cdot \mid \underline{\theta}_1^d\right)$. In particular, we assume that

$$\frac{1}{n} L_n''(\theta_1) + \overline{I}_n(\theta_1; \underline{\theta}_2^d) \xrightarrow{p^n} 0.$$

Remarks. Since p^n is a product measure, satisfying Assumption PN2′ really just amounts to verifying appropriate moment conditions. Also, note that the "information" here is only computed one way; there is no analogue to Assumption PN2.

Assumption PN3′. There exists $\epsilon = \epsilon(\underline{\theta}_1^d) > 0$ such that $M_{\epsilon,i}(x, \theta_1) = \sup_{\theta \in B_\epsilon(\theta_1)} \left| \frac{\partial^2}{\partial \theta^2} \log q_i(x_i|\theta) - \frac{\partial^2}{\partial \theta_1^2} \log q_i(x_i|\theta_1) \right|$ is dominated by some p^n-integrable function, uniformly in i, and for $\overline{M}_n(\epsilon, \theta_1) = \frac{1}{n} \sum_{i=1}^n M_{\epsilon,i}(X_i, \theta_1)$,

$$\lim_{\epsilon \to 0} \limsup_{n\to\infty} E\left[\overline{M}_n(\epsilon, \theta_1) \middle| \underline{\theta}_1^d\right] = 0.$$

Remarks. Note that Assumption PN3′ is the same continuity condition as Assumption AN2′.

Assumption PN4′. The prior density $\omega(\theta)$ is positive and continuous throughout a small neighborhood of θ_1.

Our approximation result is now

Proposition 4.3 *Suppose $\hat{\theta}_n \equiv \hat{\theta}_{q,1} \xrightarrow{\nu^n} \theta_1$, and assumptions PN1′ through PN3′ hold.*

(a) Let $\theta_n^* = \hat{\theta}_n + r(\theta_1 - \hat{\theta}_n)$, $r \in [0,1]$. Then for all $\xi > 0$, there exists $\epsilon > 0$ such that

$$\lim_{n\to\infty} P\left[\sup_{r:\theta_n^* \in B_\epsilon(\theta_1)} \left|\frac{1}{n} L_n''(\theta_n^*) + \overline{I}_n(\theta_1; \underline{\Theta}_2^d)\right| < \xi \middle| \underline{\theta}_1^d\right] = 1;$$

(b) In particular, $\frac{1}{n} L_n''(\hat{\theta}_n) + \overline{I}_n(\theta_1; \underline{\theta}_2^d) \xrightarrow{p^n} 0$.

Proof. The proof of part (a) proceeds as for Lemma 3.2, replacing conditioning on θ_1 with conditioning on $\underline{\theta}_1^d$. Note that part (b) would follow immediately as long as $\hat{\theta}_n \xrightarrow{p^n} \theta_1$. But this must be true, for almost all $\underline{\theta}_2^d$: Suppose on some measurable $\mathcal{K} \subset \operatorname{supp} \omega(\underline{\theta}_2^d|\theta_1)$, that $|L_n''(\hat{\theta}_n)+\overline{I}_n(\theta_1;\underline{\theta}_2^d)|$ $> \xi'$. Then

$$E\left[1_{\mathcal{K}} P\left[|L_n''(\hat{\theta}_n)+\overline{I}_n(\theta_1;\underline{\theta}_2^d)|>\xi' \middle| \underline{\theta}_1^d\right] \middle| \theta_1\right] \leq P\left[|L_n''(\hat{\theta}_n)+\overline{I}_n(\theta_1;\underline{\theta}_2^d)|>\xi' \middle| \theta_1\right] \to 0$$

and hence $\omega(\mathcal{K}|\theta_1)=0$. □

Theorem 4.1 *Assume the conclusion of Proposition 4.1, and suppose Assumption PN1′ through Assumption PN4′ hold. Let*

$$\sigma_n = \{-L_n''(\hat{\theta}_n)\}^{-1/2} \geq 0$$

(by Proposition 4.3 and Assumption PN1′, $\sqrt{n}\sigma_n$ exists and is bounded away from 0 and ∞ with probability tending to 1 as $n \to \infty$). Then, for all $a < b$,

$$\int_{\hat{\theta}_n+a\sigma_n}^{\hat{\theta}_n+b\sigma_n} \omega_q(\theta \mid \underline{X}^n)\, d\theta \xrightarrow{\nu^n} \Phi(b)-\Phi(a) \tag{35}$$

under $\nu^n(\cdot \mid \theta_1)$, as $n \to \infty$.

Remarks. Proposition 4.3 allows us to replace the definition of σ_n here with $\sigma_n' \equiv \overline{I}_n(\theta_1;\underline{\theta}_2^d)^{-1/2}$, but the result would be of little practical use since the scaling would depend in an unwieldy fashion upon $\underline{\theta}_2^d$. On the other hand, this proof of Proposition 4.3 based only on LLN's for p^n fails to go through, if we replace $\overline{I}_n(\theta_1;\underline{\theta}_2^d)$ with the q^n-based expected Fisher information $\overline{I}_n(\theta_1) = E[\overline{I}_n(\theta_1;\underline{\Theta}_2^d) \mid \theta_1]$, unless $\overline{I}_n(\theta_1;\underline{\theta}_2^d)$ is asymptotically free of $\underline{\theta}_2^d$. Thus we see that scaling with $\sigma_n = \{-L_n''(\hat{\theta}_n)\}^{-1/2}$ is essentially required to obtain a useful result.

Proof. The main idea is to modify the Claims 3.1 through 3.3 to assert convergence under p^n, obtaining

$$\int_{\hat{\theta}_n+a\sigma_n}^{\hat{\theta}_n+b\sigma_n} \omega_q(\theta \mid \underline{X}^n)\, d\theta \xrightarrow{p^n} \Phi(b)-\Phi(a) \tag{36}$$

pointwise in $\underline{\theta}_1^d$, and then integrate over $\underline{\theta}_2^d$ to obtain (35). We will restate the three claims and indicate what modifications are necessary in the proofs.

Claim 4.1 *For all $\xi > 0$, there exists ϵ small enough that*

$$\lim_{n\to\infty} P\left[|I_1/\{\sigma_n q^n(\underline{X}^n|\hat{\theta}_n)\} - (2\pi)^{1/2}\omega(\theta_1)| < \xi \middle| \underline{\theta}_1^d\right] = 1.$$

The proof of Claim 3.1 can be used without change, except that all probability statements are with respect to $p^n\left(\cdot \mid \underline{\theta}_1^d\right)$ and not $\nu^n(\cdot \mid \theta_1)$, and we write

$$\frac{L_n''(\theta)}{L_n''(\hat{\theta}_n)} - 1 = \frac{[\frac{1}{n}L_n''(\theta) + \overline{I}_n(\theta_1; \underline{\Theta}_2^d)] - [\frac{1}{n}I_n(\theta_1; \underline{\Theta}_2^d) + \frac{1}{n}L_n''(\hat{\theta}_n)]}{\frac{1}{n}L_n''(\hat{\theta}_n)}.$$

Claim 4.2 *For each fixed $\epsilon > 0$,*

$$I_2/\{\sigma_n q^n(\underline{X}^n|\hat{\theta}_n)\} \xrightarrow{p^n} 0.$$

The proof is formally identical to that of Claim 3.2, but (32), Proposition 4.3 and Assumption PN1′ are used to justify the steps, and probability is assessed in p^n rather than ν^n.

Claim 4.3 *For all $\xi > 0$,*

$$\lim_{n\to\infty} P\left[|I_3/\{\sigma_n q^n(\underline{X}^n|\hat{\theta}_n)\} - (2\pi)^{1/2}\omega(\theta_1)[\Phi(b) - \Phi(a)]| < \xi \middle| \underline{\theta}_1^d\right] = 1.$$

The argument proceeds as for Claim 3.3, except that (32) is used to show that $P[N_n \subset B_\epsilon(\theta_1)|\underline{\theta}_1^d]$ and all probabilities are assessed in p^n.

This completes the proof of (36); dominated convergence now yields (35). □

5 Examples

5.1 Item response theory; inference under ν^n alone

Item response theory, IRT, treats models for subjects' responses to individual items (questions) on a standardized multiple-choice questionnaire, in terms of unobserved or latent factors. Suppose each observable variable x_i has k_i values $\xi_{i1}, \ldots, \xi_{ik_i}$ (the subject makes one of k_i responses for each item), with $k_i \leq k_0$ for some fixed $k_0 < \infty$. In practice the model often used to analyze the data is the product of marginals $q^n(\underline{x}^n \mid \theta_1) = \prod_{i=1}^n q_i(x_i \mid \theta_1)$, where

$$q_i(x_i \mid \theta_1) = \prod_{j=1}^{k_i} P_{ij}(\theta_1)^{Y_{ij}},$$

and $Y_{ij} = 1_{\{X_i=\xi_{ij}\}}$.

In many settings, the curves $P_{ij}(\theta_1)$ are considered well-enough estimated that they are taken to be known, and the practitioner is primarily interested in estimating θ_1 for each examinee. This is the case in large-scale educational testing, for example. Wang (1986, 1987) argues that when the full model $p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right)$ is assumed, and $d = 2$, the popular logistic response curve fitting program LOGIST produces stable estimates for $q_i(x_i|\vartheta_1)$, where ϑ_1 is essentially the first component of an appropriate rotation $\underline{\vartheta}_1^d$ of $\underline{\theta}_2^d$, when $\underline{x}^n$ consists of binary responses. Thus the error made in not modeling for or estimating $\underline{\theta}_2^d$ is often argued to be negligible.

Stout's notions of *essential independence* and *essential unidimensionality* (Stout, 1987, 1990; Junker, 1988, 1991) provide conditions under which not modeling for nuisance factors seems reasonable. Traditional analysis of educational tests is based on averages of item response scores $\overline{A}_n = \frac{1}{n}\sum_1^n A_i$, where $A_i = \sum_{j=1}^{k_i} a_{ij}Y_{ij}$, subject to the constraints that

$$\begin{aligned} &\text{(a)} \quad \exists M < \infty : \ -M \leq a_{i1} \leq a_{i2} \leq \ldots \leq a_{ik_i} \leq M, \ \forall\ i; \text{ and} \\ &\text{(b)} \quad \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_{ik_i} - a_{i1} > 0. \end{aligned} \tag{37}$$

Assumption EI applies directly to such scores, and directly generalizes Stout's definition of *strong essential independence* for binary data (Definition 3.5, Stout, 1990). Note also that, because of the bounded nature of X_i, Assumption EI implies (b) of Lemma 3.1, as long as the $P_{ij}(\theta_1)$ are bounded away from 0 and 1. Since estimation of θ_1 is the goal, some sort of minimum information—or discrimination, as it is called in educational testing models—condition is needed. Let $\overline{A}_n(\theta_1) = E\left[\overline{A}_n \middle| \theta_1\right]$; a minimum-information criterion that is appealing in the educational testing context is that, for every set of item scores satisfying (37) and every θ_1, there is an interval $B = B_\delta(\theta_1)$ and an $\epsilon > 0$ such that

$$\liminf_{n\to\infty} \frac{\overline{A}_n(t) - \overline{A}_n(\theta_1)}{t - \theta_1} \geq \epsilon, \ \forall t \in B, \ t \neq \theta_1. \tag{38}$$

This generalizes Stout's "local asymptotic discrimination," LAD, condition for binary items (Stout, 1990, Definition 3.8). Under mild smoothness assumptions, the conditions and results of Section 3—in which only q^n and ν^n play a role—hold under EI and LAD. Example 5.2 below provides a concrete example.

Proposition 5.1 *Suppose that* EI *and* LAD *hold, and that the response curves* P_{ij} *satisfy*

$$\text{For each } t,\ 0 < \inf_{i,j} P_{ij}(t) \leq \sup_{i,j} P_{ij}(t) < 1; \tag{39}$$

$$P_{ij}(t) \text{ is continuous at each } t, \text{ uniformly in } i \text{ and } j \tag{40}$$

and suppose Assumption C3 holds. Then the "wrong model MLE" $\hat{\theta}_{q,1}(\underline{x}^n)$ *is* ν^n*-consistent for* θ_1*, as* $n \to \infty$*.*

Proof. We will verify the conditions of Lemma 3.1 and Lemma 3.2. It follows from an inequality of Csiszar (1975), $D(f\|g) \geq \frac{1}{4}[\int |f(t) - g(t)|\, dt]^2$, that

$$\begin{aligned} \frac{1}{n} D(q^n_{\theta_1} \| q^n_{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} D(q_{i,\theta_1} \| q_{i,\theta}) \\ &\geq \frac{1}{4} \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{k_i} |P_{ij}(\theta_1) - P_{ij}(\theta)| \right]^2, \end{aligned} \tag{41}$$

which is bounded away from zero under (38) (consider $\{a_{ij}\}$ for which $a_{ik_i} = 1$, and $a_{ij} \equiv 0$ for all $j < k_i$). This is (a) of Lemma 3.1. As noted above, (b) of Lemma 3.1 follows from Assumption EI and (39).

The continuity condition (a) of Lemma 3.2 follows from (40). (b) of Lemma 3.2 requires that

$$\lim_{n\to\infty} P\left[\sup_{\theta \in B_\delta(t)} |D_n(t,\theta) - E[D_n(t,\theta)|\theta_1]| < \epsilon \middle| \theta_1 \right] = 1$$

for every ϵ and appropriate δ. The expression in absolute values may be written as

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_i} [Y_{ij} - P_{ij}(\theta_1)] \log \frac{P_{ij}(t)}{P_{ij}(\theta)}$$

which will tend to zero uniformly in $\theta \in B_\delta(t)$ by Assumption EI, (39) and (40). □

Example 5.1 Assumption C3 may often be verified directly. Consider the case of binary response data, in which $k_i \equiv 2$, $\xi_{i1} \equiv 0$ and $\xi_{i2} \equiv 1$. A commonly used model for the response curves is

$$P_{i2}(\theta_1) = c_i + (1 - c_i) \frac{1}{1 + \exp\{-a_i(\theta_1 - b_i)\}},$$

and $P_{i1}(\theta_1) = 1 - P_{i2}(\theta_1)$. Then $D_n(\theta_1, \theta) = \frac{1}{n}\sum_1^n t_i(\theta_1) - t_i(\theta)$, where

$$t_i(\theta) = X_i \log \frac{c_i + e^{a_i(\theta - b_i)}}{1 - c_i} - \log\left[1 + \frac{c_i + e^{a_i(\theta - b_i)}}{1 - c_i}\right].$$

Hence

$$\begin{aligned} \lim_{\theta \to \infty} -t_i(\theta) &= \begin{cases} 0, \text{ if } X_i = 1, \\ \infty, \text{ if } X_i = 0; \end{cases} \\ \lim_{\theta \to -\infty} -t_i(\theta) &= -\log c_i^{X_i}(1 - c_i)^{1 - X_i}, \end{aligned}$$

and we see that Assumption C3 holds as long as $P[X_i = 1 \,\forall i|\theta_1] = P[X_i = 0 \,\forall i|\theta_1] = 0$; this in turn follows from Assumption EI and (39), which has a natural interpretation in terms of the a_i's b_i's and c_i's. □

Proposition 5.2 *Suppose, in addition to the assumptions of Proposition 5.1, that*

$$\frac{\partial^2}{\partial \theta^2} \log P_{ij}(\theta) \text{ is bounded pointwise in } \theta, \text{ uniformly in } i \text{ and } j. \tag{42}$$

Then, in the sense of (24),

$$\mathcal{L}_q\left\{ \sqrt{n}\frac{\Theta_1 - \hat{\theta}_{q,1}(\underline{x}^n)}{\sigma_n} \middle| \underline{x}^n \right\} \xrightarrow{\nu^n} N(0,1).$$

Proof. It is enough to show that Assumptions PN1 through PN4 are satisfied; only Assumption PN1 and Assumption PN3 are problematic. Proposition 4.1 of Junker (1991) shows that Assumption PN1 holds under (38) and differentiability conditions (the argument is similar to the one bounding (41) away from zero). The uniform continuity condition of Assumption PN3 focuses on a locally uniform bound for

$$\left| \sum_{j=1}^{k_i} Y_{ij} \left[\frac{\partial^2}{\partial \theta^2} \log P_{ij}(\theta) - \frac{\partial^2}{\partial {\theta_1}^2} \log P_{ij}(\theta_1) \right] \right|; \tag{43}$$

which follows from (42), due to the boundedness of the Y_{ij}'s. □

In the following example the asymptotic MLE and posterior distributions are different; this is an explicit case in which interval estimates for θ_1 based on the q^n-likelihood are wider than intervals based on the q^n-posterior.

Example 5.2 Consider binary responses $X_1, X_2, X_3, \ldots$, having the same response curve $P_{i2}(\theta) = \theta$ (so the latent scale is the interval $(0,1)$ and $P[X_i = 1 \mid \theta] \equiv \theta$). Suppose that the items are arranged in successive groups of g_o items as $X_1, X_2, \ldots, X_{g_o}$; $X_{g_o+1}, X_{g_o+2}, \ldots, X_{2g_o}$; etc., such that different groups of g_o items are independent of one another, given θ, and items within a single group are positively correlated, given θ, and with

$$\mathrm{Corr}(X_i, X_j|\theta) = \begin{cases} c \text{ if } X_i \text{ and } X_j \text{ are in the same group,} \\ 0 \text{ if not,} \end{cases}$$

for some fixed $c \in (0,1]$. This ν^n is a naive model for a paragraph comprehension test in which several paragraphs are presented and g_o questions are asked for each paragraph. Here, θ represents a trait common to all the items, which we might wish to think of as reading comprehension; and the nonzero correlations are induced by nuisance traits, for example, specific knowledge about the subject matter of the paragraph at hand. This example is also considered by Stout (1990) and Junker (1991). Current interest in (more realistic) block-dependent structures like this is evidenced, for example, by Wainer, Lewis, Kaplan and Braswell (1990).

One may easily verify Assumption EI, (38), and the subsidiary continuity conditions used above to verify Lemmas 3.1 and 3.2. Assumption C3 is not an issue, since the parameter space $(0,1)$ has compact closure. Also, because of the block-dependent structure, it is trivial to obtain an asymptotic normality result for $\hat{\theta}_{q,1}(\underline{X}^n) \equiv \overline{X}_n$; we see that

$$\sqrt{n}\,(\hat{\theta}_{q,1}(\underline{X}^n) - \theta_1) \sim AN(0, \sigma^2)$$

where $\sigma^2 = \theta_1(1-\theta_1)[1 + c(g_0 - 1)]$ is somewhat inflated over the anticipated asymptotic variance $\theta_1(1-\theta_1)$ under q^n. Thus the q^n-based MLE is consistent and asymptotically normal, but has a somewhat larger asymptotic variance than would normally be expected.

Turning to the posterior distribution of θ_1, the continuity condition (43) is easily verified, for $\theta_1 \in (0,1)$. Hence, in the sense of (24),

$$\mathcal{L}_q \left\{ \sqrt{n} \frac{\Theta_1 - \hat{\theta}_{q,1}(\underline{x}^n)}{\sqrt{\hat{\theta}_{q,1}(\underline{x}^n)(1 - \hat{\theta}_{q,1}(\underline{x}^n))}} \middle| \underline{x}^n \right\} \xrightarrow{\nu^n} N(0,1),$$

where the standard error $\sqrt{\hat{\theta}_{q,1}(\underline{x}^n)(1-\hat{\theta}_{q,1}(\underline{x}^n))}$ is calculated directly from $(-L_n''(\hat{\theta}_{q,1}(\underline{x}^n)))^{-1/2}$ in the statement of Theorem 3.1, using $\hat{\theta}_{q,1}(\underline{x}^n) = \overline{x}_n$. □

5.2 Inference when p^n is also present

In this section we illustrate the results of Section 4. As will be seen, explicit calculation of integrals over $\underline{\theta}_2^d$, and other needed quantities, becomes complicated fairly quickly. Example 5.3 treats a multivariate normal p^n in which the location is a nuisance parameter, and Example 5.4 treats a multivariate normal p^n in which the scale is a nuisance parameter. In Example 5.3 we identify a rate at which the nuisance parameter must attenuate if asymptotic normality is to hold for $\hat{\theta}_{q,1}$. In Example 5.4, p^n continues to be highly dependent on θ_2 as $n \to \infty$; this shows that the requirement that dependence on $\underline{\theta}_2^d$ attenuate as $n \to \infty$ is not necessary when embedding in a full conditional independence model p^n.

Example 5.3 Suppose $\mathcal{L}(X_i|\theta_1,\theta_2) = N(\frac{\theta_2}{\alpha_i},\theta_1)$, independent of one another, and $\mathcal{L}(\Theta_2|\theta_1) = N(0,\theta_1)$. It is easy to verify that $\mathcal{L}(X_i|\theta_1) = N(0,\frac{\alpha_i^2+1}{\alpha_i^2})$, that

$$\frac{1}{n}\log q^n(\underline{x}^n|t) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\theta_1 - \frac{1}{2n}\sum_{i=1}^n \log\frac{\alpha_i^2+1}{\alpha_i^2} - \frac{1}{2n\theta_1}\sum_{i=1}^n \frac{\alpha_i^2}{1+\alpha_i^2}X_i^2,$$

and that consequently $\hat{\theta}_{q,1}(\underline{x}^n) = \frac{1}{n}\sum_1^n \frac{\alpha_i^2}{1+\alpha_i^2}X_i^2$. Let $\overline{\alpha}_n = \frac{1}{n}\sum_1^n \frac{\alpha_i^2}{1+\alpha_i^2}$ and $\overline{\beta}_n = 1-\overline{\alpha}_n$; then $E\left[\hat{\theta}_{q,1}(\underline{X}^n)\middle|\theta_1,\theta_2\right] = \overline{\alpha}_n\theta_1 + \overline{\beta}_n{\theta_2}^2$, a convex combination of the parameter of interest and the nuisance parameter. If we assume

$$\lim_{n\to\infty}|\alpha_i| = \infty \tag{44}$$

then, since $\overline{\alpha}_n = 1-\overline{\beta}_n \to 1$, it is easy to verify that $\hat{\theta}_{q,1} \xrightarrow{p^n} \theta_1$ and hence $\hat{\theta}_{q,1} \xrightarrow{\nu^n} \theta_1$. It is also easy to verify that

$$D_n(\theta_1,t) = \frac{1}{n}\log\frac{q^n(\underline{x}^n|\theta_1)}{q^n(\underline{x}^n|t)} = \frac{1}{2}\log\frac{t}{\theta_1} + \frac{1}{2}(\frac{1}{t}-\frac{1}{\theta_1})\hat{\theta}_{q,1} \approx \frac{1}{2}\log\frac{t}{\theta_1} + \frac{1}{2}(\frac{\theta_1}{t}-1)$$

as $n \to \infty$. Analysis of the function $f(u) = \log u + (\frac{1}{u}-1)$ shows that the assumptions C1′, C2′, C3′ and the uniformity assumptions (33) and (34) are satisfied, assuming (44). Hence the somewhat stronger consistency results Proposition 4.1 and Corollary 4.1 hold also.

Turning to the asymptotic distribution of $\hat{\theta}_{q,1}$, we may calculate in Assumption AN1′ that

$$\sigma_n(\theta_1,\theta_2) = [\mathrm{Var}\,(\overline{U}|\theta_1,\theta_2)]^{1/2} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\alpha_i^2}{1+\alpha_i^2}\right)\left(\frac{1}{2{\theta_1}^2}+\frac{4\theta_1{\theta_2}^2}{4{\theta_1}^4\alpha_i^2}\right)\right]^{1/2},$$

which tends to $c(\theta_1) = 1/(\sqrt{2}\theta_1)$ as $n \to \infty$, under (44). To obtain a recognizable asymptotic normality result for $\hat{\theta}_{q,1}$, we also need to identify a function $b(\theta_1)$ such that $\sqrt{n}(E[\overline{U}|\theta_1,\theta_2] - b(\theta_1))/\sigma_n(\theta_1,\theta_2) \to 0$ (see remarks following the proof of Proposition 4.2). We calculate $\gamma_n(\theta_1,\theta_2) = \sqrt{n}E[\overline{U}|\theta_1,\theta_2] = \frac{-\sqrt{n}}{2\theta_1}(1 - \frac{\overline{\alpha}_n}{\theta_1} + \frac{\overline{\beta}_n{\theta_2}^2}{\theta_1})$; such a function $b(\theta_1)$ can be identified in this case only by supplementing (44) with the stronger rate condition that $\sqrt{n}\,\overline{\beta}_n \to 0$.

Finally we may turn to the posterior normality results of Section 4. Note that

$$\begin{aligned}\overline{I}(\theta_1;\theta_2) = E\left[-\frac{1}{n}L_n''(\theta_1)\middle|\theta_1,\theta_2\right] &= \frac{1}{2{\theta_1}^2} - \frac{1}{{\theta_1}^3}E\left[\hat{\theta}_{q,1}\middle|\theta_1,\theta_2\right] \\ &= \frac{1}{{\theta_1}^2}(\overline{\alpha}_n - \frac{1}{2}) - \frac{\overline{\beta}_n{\theta_2}^2}{{\theta_1}^3} \\ &\to \frac{1}{2{\theta_1}^2},\end{aligned}$$

as $n \to \infty$, and

$$\frac{1}{n}L_n'(\theta_1) + \overline{I}_n(\theta_1;\theta_2) = \frac{1}{{\theta_1}^3}[\hat{\theta}_{q,1} - (\overline{\alpha}_n\theta_1 + \overline{\beta}_n{\theta_2}^2)] \xrightarrow{p^n} 0,$$

as $n \to \infty$. From these it is easy to verify Assumptions PN1′ and PN2′. For Assumption PN3′, we note that

$$E\left[\overline{M}_n(\delta,\theta_1)\middle|\theta_1,\theta_2\right] = E\left[\sup_{t\in B_\delta(\theta_1)}\left|\frac{t-2}{2t^3}\right|\hat{\theta}_{q,1}(\underline{X}^n)\middle|\theta_1,\theta_2\right] = (\overline{\alpha}_n\theta_1 + \overline{\beta}_n{\theta_2}^2)\sup_{t\in B_\delta(\theta_1)}\left|\frac{t-2}{2t^3}\right|,$$

from which Assumption PN3′ follows. In this case the scale of the asymptotic posterior distribution is the same as for the asymptotic distribution of $\hat{\theta}_{q,1}(\underline{X}^n)$. □

Example 5.4 Let $\mathcal{L}(X_i|\theta_1,\theta_2)$ be i.i.d. $N(\theta_1,\theta_2^{-2})$, so that the common marginal density under p^n is

$$p(x|\theta_1,\theta_2) = \frac{\theta_2 e^{-\theta_2^2(x-\theta_1)^2/2}}{\sqrt{2\pi}}$$

for $\theta_1 \in \mathbb{R}$, $\theta_2 \in \mathbb{R}^+$. Assume that the prior density for (θ_1,θ_2) factors, and that the marginal for θ_2 is $\omega(\theta_2) = (2/\sqrt{2\pi})\exp\{-\theta_2^2/2\}$, for $\theta_2 > 0$. Then $q(x|\theta_1) = 1/\pi[1+(x-\theta_1)^2]$, a location Cauchy; so that $\hat{\theta}_{q,1} = \arg\min_t \frac{1}{n}\sum_{i=1}^{n} \log 1 + (x_i - t)^2$, uniquely.

Instead of verifying Assumption C1′ through Assumption C3′ we will establish the conclusion of Proposition 4.1 directly, for fixed (θ_1, θ_2). First we show that there is a positive constant Δ so large that we may ignore parameter values outside of $[-2\Delta,\ 2\Delta]$, in the sense that it is very unlikely that $|\hat{\theta}_{q,1}|$ exceeds 2Δ. We also show that $L_n(\theta_1)$ dominates $L_n(t)$ for $|t| > 2\Delta$. This will allow us to restrict attention to the compact set $[-2\Delta,\ 2\Delta]$, which will be handled afterwards.

Let $\gamma > 0$ and consider the event

$$\mathcal{E}_{n,\gamma}(\Delta) = \{\min_{t \in [-2\Delta,\ 2\Delta]^c} \frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i - t)^2 \geq \gamma + \frac{1}{n}\sum_{i=1}^{n} \log 1 + X_i^2\}$$

We show that for Δ large enough

$$\lim_{n\to\infty} P(\mathcal{E}_{n\gamma}(\Delta)|\theta_1, \theta_2) = 1. \tag{45}$$

Note that for $\gamma > 0$ (45) is analogous to Assumption C3′: the density outside $[-2\Delta,\ 2\Delta]$ is beaten by the density for $t = 0$, by a factor of $e^{-n\gamma}$.

Consider the random set of indices $\mathcal{I} = \{i : F^{-1}_{\theta_1,\theta_2}(0.02) \leq X_i \leq F^{-1}_{\theta_1,\theta_2}(0.98)\}$, where $F_{\theta_1,\theta_2}(x) = \int_{-\infty}^{x} p(t|\theta_1, \theta_2)dt$, and the event $\mathcal{B} = \{\frac{1}{n}\,\mathrm{card}\,(\mathcal{I}) \geq 0.95\}$. Then

$$\begin{aligned}
\mathcal{E}_{n\gamma}(\Delta) &\supset \{\min_{t \in [-2\Delta,\ 2\Delta]^c} \frac{1}{n}\sum_{i\in\mathcal{I}} \log 1 + (X_i - t)^2 \geq \gamma + \frac{1}{n}\sum_{i=1}^{n} \log 1 + X_i^2\} \\
&\supset \{\min_{t \in [-2\Delta,\ 2\Delta]^c} \frac{1}{n}\sum_{i\in\mathcal{I}} \log 1 + (X_i - t)^2 \geq \gamma + \frac{1}{n}\sum_{i=1}^{n} \log 1 + X_i^2\} \cap \mathcal{B}
\end{aligned} \tag{46}$$

For $\Delta > \max\{|F^{-1}_{\theta_1,\theta_2}(0.02)|, |F^{-1}_{\theta_1,\theta_2}(0.98)|\}$ we have, on the last event, $\log 1 + (X_i - t)^2 \geq \log 1 + \Delta^2 \geq 2\log\Delta$. Now, because of the intersection and restriction we have that for $|t| > 2\Delta$, $\frac{1}{n}\sum_{i\in\mathcal{I}} \log 1 + (X_i - t)^2 \geq n\frac{(.95)2\ \log\Delta}{n} \geq \log\Delta$. Hence

$$\mathcal{E}_{n\gamma}(\Delta) \supset \{\log\Delta \geq \gamma + \frac{1}{n}\sum_{i=1}^{n} \log 1 + X_i^2\} \cap \mathcal{B} \tag{47}$$

For given $\gamma > 0$ we can choose Δ so large that $\log\Delta > \gamma + \delta + E[\log 1 + X^2|\theta_1, \theta_2]$ for some $\delta > 0$. Now $P[\mathcal{E}_{n\gamma}(\Delta)|\theta_1, \theta_2] \geq P(\textit{Event on right in } (47)|\theta_1, \theta_2) \to 1$ so (45) is proved since both conditions on the right in (47) are met with probability 1 as $n \to \infty$ (by the LLN and definition of percentiles.)

Next we obtain an asymptotic convexity condition analogous to Proposition 4.1, using the restriction to $[-2\Delta, 2\Delta]$. An easy calculus argument shows that

$$E[\log 1 + (X-\theta_1)^2|\theta_1,\theta_2] < \min_{t\in\Delta_\epsilon} E[\log 1 + (X-t)^2|\theta_1,\theta_2] \tag{48}$$

where Δ_ϵ is the union $[-2\Delta, \theta_1-\epsilon]\cup[\theta_1+\epsilon, 2\Delta]$ (take first and second derivatives with respect to θ_1 in the argument of the expectation on the left). Then there exists $\gamma > 0$ such that

$$\lim_{n\to\infty} P\{\min_{t\in\Delta_\epsilon} \frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i-t)^2 \geq (1+\gamma)\frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i-\theta_1)^2|\theta_1,\theta_2\} = 1. \tag{49}$$

Indeed,

$$\begin{aligned}
&P(\textit{event in } (49)|\theta_1,\theta_2)\\
&\quad\geq P\{\min_{t\in\Delta_\epsilon}[E(\log 1 + (X-t)^2|\theta_1,\theta_2) \geq (1+\epsilon)\frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i-\theta_1)^2\\
&\qquad -|\frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i-t)^2 - E(\log 1 + (X-t)^2|\theta_1,\theta_2)|]\\
&\qquad \geq (1+\epsilon)\frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i-\theta_1)^2 \qquad (50)\\
&\qquad \text{and } \sup_{t\in\Delta_\epsilon} |\frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i-t)^2 - E(\log 1 + (X-t)^2|\theta_1,\theta_2)| < \epsilon|\theta_1,\theta_2\}\\
&\quad\geq P\{(-\epsilon + E[\log 1 + (X-t_{\min})|\theta_1,\theta_2])/1+\gamma \geq \frac{1}{n}\sum_{i=1}^{n} \log 1 + (X_i-\theta_1)^2\\
&\qquad \text{and ULLN's}|\theta_1,\theta_2\} \qquad (51)
\end{aligned}$$

where "ULLN's" (uniform LLN) refers to the second event in (50) and $t_{\min}$ is the element of Δ_ϵ which achieves the minimum on the right in (48). By the strictness of the inequality in (48) we see that for given ϵ there are positive numbers η and γ so that

$$\frac{-\eta + E[\log 1 + (X-t_{\min})^2|\theta_1,\theta_2]}{1+\gamma} > E[\log 1 + (X-\theta_1)^2|\theta_1,\theta_2]$$

Now both events in (51) have probability tending to 1 so (49) is proved. Explicitly using a uniform LLN seems a more expedient application of the intuition behind Wald's proof in examples of wrong model analysis; the same technique works for the double exponential for instance and in both cases is easier than establishing Assumptions C1′ and C2′ (which we believe are in fact true). The

conditions for asymptotic mixture normality of $\hat{\theta}_{q,1}(\underline{x}^n)$ and asymptotic normality of $\omega_q(\theta_1|\underline{x}^n)$ given in Section 4 are now easy to check, since X_i is i.i.d. under ν^n. However, the quantities are difficult to work with analytically so we have been unable to verify conclusively, as in Example 5.2, that the asymptotic distributions are different. □

6 Discussion

When assumptions cannot safely be made about the dependence structure of a model a natural approach is to regard the data as coming from a distribution $\nu^n(\underline{x}^n \mid \theta_1)$ conditioned only on the parameter of interest, but otherwise unspecified. Because ν^n may be difficult to work with, practitioners are sometimes lead to the product of marginals $q^n(\underline{x}^n \mid \theta_1) = \prod_1^n q_i(x_i \mid \theta_1)$, where $q_i(x_i \mid \theta_1)$ is the marginal for X_i from ν^n. Although this natural choice is also optimal in the senses indicated in Section 2, it is not clear that inference based on the product of marginals can be relied upon. This methodological concern has been raised explicitly in applied psychological measurement in particular; and in fact has implications in much of applied statistical practice, where the assumption of conditional independence is often an incompletely-justified convenience. What is the asymptotic behavior of q^n-based estimators under the correct law ν^n?

We have identified two broad categories of problems in which asymptotic inference based on the product of marginals can, at least in part, proceed: Section 3 treats the case in which laws of large numbers (LLN's) are imposed upon ν^n; these LLN's, which generalize conditions used in item response theory, arise as a way to stabilize the asymptotic behavior of familiar sums derived from the log-likelihood $\overline{L}_n(\theta_1) = \frac{1}{n}\sum_{i=1}^n \log q_i(x_i \mid \theta_1)$. Section 4 treats the case in which ν^n can be embedded, as a mixture over nuisance parameters $\underline{\theta}_2^d$, in a larger conditional independence model $p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right) = \prod_{i=1}^n p_i\left(x_i \mid \underline{\theta}_1^d\right)$. In this case, LLN's need not be imposed upon ν^n, since the LLN's which hold naturally under p^n are enough. In both categories of problems we have obtained consistency of the q^n-based MLE $\hat{\theta}_{q,1}$ and consistency of the q^n-based posterior distribution $\omega_q(\theta_1|\underline{x}^n)$ under $\nu^n(\underline{x}^n \mid \theta_1)$, in the sense that $\hat{\theta}_{q,1} \overset{\nu^n}{\to} \theta_1$ and $\omega_q(\theta_1|\underline{x}^n)$ concentrates at $\hat{\theta}_{q,1}$ with ν^n-probability tending to one as $n \to \infty$.

The asymptotic distribution of the q^n-based MLE cannot be determined without further as-

sumptions on ν^n or, if it is assumed to exist, p^n. In addition to indicating a few situations in which a conventional asymptotic normality result is possible, we have considered in some detail the situation in which ν^n is obtained by mixing nuisance parameters $\underline{\theta}_2^d$ out of the full model $p^n\left(\underline{x}^n \mid \underline{\theta}_1^d\right)$. In this second category of problems, we have identified the asymptotic distribution of $\hat{\theta}_{q,1}$ as a mixture of normals, which is determined by the way in which p^n depends upon $\underline{\theta}_2^d$.

In contrast, the q^n-based posterior ω_q is asymptotically normal in both categories of problems, centered at $\hat{\theta}_{q,1}$ and scaled by σ_n, where σ_n^{-2} is the the q^n-based empirical Fisher information (other possible scaling terms do not lead to so simple a result for q^n); thus the asymptotic normality of ω_q has little to do with the true dependence structure of the data. This is consonant with recent interpretations of Laplace's method (especially Chen, 1985; and Kass, Tierney and Kadane, 1990), which show that asymptotic posterior normality is really an analytic property of the model (right or wrong) along a particular data sequence. The stochastic behavior of the data only enters into the asymptotic distribution of the centered and scaled ω_q-measure of an interval $[a,b] \in \Omega_{\Theta_1}$, viewed as a functional of the stochastic process $(X_1, X_2, \ldots)$ which "estimates" the corresponding $N(0,1)$-measure of the same interval $[a,b]$.

This disparity between MLE-based and posterior-based asymptotic inference can be illustrated in a practical setting. With no structure on ν^n except for the LLN, Example 5.2 gives a situation in educational testing for which asymptotic $\hat{\theta}_{q,1}$-based confidence intervals can be calculated and they are significantly wider than the highest posterior density (HPD) intervals based on asymptotic normality of ω_q. In this case there is one model in which both likelihood and Bayes analysis can be conducted in full detail; and the analyses clearly give different answers, even asymptotically.

When ν^n is represented as the mixture over nuisance parameters of p^n, imposing an LLN on ν^n has the effect of forcing the dependence of $p^n\left(\cdot \mid \underline{\theta}_1^d\right)$ on the nuisance parameters $\underline{\theta}_2^d$ to attenuate as $n \to \infty$. However this attenuation seems to be neither necessary nor sufficient for useful asymptotic results. In Example 5.3 a sufficiently rapid rate of attenuation is required for asymptotic normality of $\hat{\theta}_{q,1}$, and in Example 5.4 consistency and asymptotic normality obtain, in both the Bayes and likelihood senses, even though there is no attenuation of dependence on the nuisance parameter.

It is widely believed that the two paradigms, likelihood-based inference and posterior-based

inference, are philosophically different but "asymptotically the same", except in bizarre situations. The setting of this paper, in which the analyst replaces the correct dependent likelihood with a convenient independent likelihood, is one in which the asymptotics come out differently for "typical" cases. How can we make sense of this?

On the one hand, our q^n-based MLE is really an M-estimator with a particular choice of objective function, namely the product of the one-dimensional data marginals of ν^n, which we have denoted q^n. We have given consistency and asymptotic distribution results for this M-estimator under suitable assumptions on ν^n or, if it is assumed to exist, p^n. We may interpret the asymptotic distribution of the M-estimator as a measure of estimation error under ν^n without difficulty; in particular we need not assume that the data actually came from q^n to arrive at this interpretation.

On the other hand, our approximation to the q^n-based posterior shows that it concentrates at the q^n-based M-estimator—cf. equations (24) or (35)—but its "asymptotic rate of concentration" is harder to interpret: q^n-based asymptotic posterior standard errors say how much the q^n-based posterior is concentrated around the M-estimator, but not how much the q^n-based posterior is concentrated around the θ_1 which "generated" $\underline{x}^n$. If the data actually came from q^n then Bayes' rule would allow one to interpret the q^n-based posterior, and hence its asymptotically normal approximation, in the usual sense of updating belief about where θ_1 was after looking at the data. If the data didn't come from q^n, then one cannot appeal to Bayes' rule for this interpretation, and the q^n-based posterior is interesting only because it corresponds to what is done in practice. Perhaps the only justifiable interpretation of ω_q is a counterfactual: "If the data had come from q^n this is where one would think θ_1 was."

Finally, the greater sensitivity of $\hat{\theta}_{q,1}(\underline{x}^n)$ to dependence in the data suggests that the standard error of $\hat{\theta}_{q,1}(\underline{x}^n)$ may be a good starting place for diagnostic checks of conditional independence: for example, inflated standard errors would indicate positive dependence in ν^n that might make it worthwhile to model ν^n directly; see Junker (1991) in this regard. Although both MLE and Bayes paradigms lead to consistent estimators when the product of marginals q^n is substituted for the correct likelihood ν^n, correct calculation and interpretation of the variability of the estimators depends on a more careful analysis of the stochastic behavior of the data-generating mechanism.

References

Ackerman, T. A. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 20, 1987.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

Bahadur, R. R. (1971). *Some limit theorems in statistics.* Philadelphia: Society for Industrial and Applied Mathematics.

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, **84**, 200–207.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, **37**, 51–58.

Chen, C.-F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society, Series B*, **47**, 540-546.

Chung, K. L. (1974). *A Course in Porobability Theory.* Second Edition. New York: Academic Press.

Clarke, B. S. and Barron, A. R. (1989a). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, **36**, xxx–xxx.

Chernoff, H. (1956). Large sample theory: parametric case. *Annals of Mathematical Statistics*, **27**, 1–22.

Cox, D. R. and Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika*, **77**, 747–762.

Cox, J. T. and Grimmett, G. (1984). Central limit theorem for associated random variables and the percolation model. *Annals of Probability*, **12**, 514–528.

Csiszar, I. (1975). Information type measures of difference of probability distributions and direct observations. *Studia Scientiarum Mathematicum Hungarica*, **2**, 299–318.

Drasgow, F. and Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, **7**, 189–199.

Dvoretzky, A. (1972). Asymptotic normality for sums of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Probability and Statistics*, Volume II, 513–535.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, **11**, 91–115.

Huber, P. J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson Lemma for capacities. *Annals of Statistics*, **1**, 251–263.

Iosifescu, M. and Theodorescu, M. (1969). *Random processes and learning.* New York: Springer-Verlag.

Junker, B. W. (1988). *Statistical aspects of a new latent trait model*, unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, **56**, xxx–xxx.

Kass, R. E., Tierney, L. and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In Geisser, S., Hodges, J. S., Press, S. J, and Zellner, A., ed's. (1990). *Bayesian and likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard* (pp. 473–488). New York: North-Holland.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* John Wiley and Sons. New York.

Newman, C. M. and Wright, A. L. (1982). Associated random variables and martingale inequalities. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **59**, 361–371.

Reiser, M. (1990). An application of the item-response model to psychiatric epidemiology. *Sociological Methods and Research*, **18**, 66–103.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, **53**, 349-359.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, **52**, 79–98.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, **55**, 293–325.

Stroock, D. (1984). *An introduction to the theory of large deviations.* Springer-Verlag. New York.

Wainer, H., Lewis, C., Kaplan, B., and Braswell, J. (1990). *An adaptive algebra test: a testlet-based, hierarchically-structured test with validity-based scoring.* Program Statistics Research Technical Report 90–92, Educational Testing Service, Princeton, New Jersey.

Walker, A. M. (1969). On the asymptotic behavior of posterior distributions. *Journal of the Royal Statistical Society, Series B*, **31**, 80–88.

Wang, M.-M. (1986). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at Office of Naval Research Model-Based Measurement Contractors' Meeting, Knoxville, TN, April 28, 1986.

Wang, M.-M. (1987). *Estimation of ability parameters from response data that are precalibrated with a unidimensional model.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 22, 1987.

Yamada, K. (1976). Asymptotic behavior of distributions for random processes under incorrect models. *Journal of Mathematical Analysis and Applications*, **56**, 294–308.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, **8**, 125-145.