Chapter 5
Adaptive Estimation

Jon A. Wellner

5.1 Introduction to Four Papers on Semiparametric and Nonparametric Estimation

5.1.1 Introduction: Setting the Stage

I discuss four papers of Peter Bickel and coauthors: Bickel (1982), Bickel and Klaassen (1986), Bickel and Ritov (1987), and Ritov and Bickel (1990).

The four papers by Peter Bickel (and co-authors Chris Klaassen and Ya'acov Ritov) to be discussed here all deal with various aspects of estimation in semiparametric and nonparametric models. All four papers were published in the period 1982–1990, a time when semiparametric theory was in rapid development. Thus it might be useful to briefly review some of the key developments in statistical theory prior to 1982, the year in which Peter Bickel's Wald lectures (given in 1980) appeared, in order to give some relevant background information. Because I was personally involved in some of these developments in the early 1980s, my account will necessarily be rather subjective and incomplete. I apologize in advance for oversights and a possibly incomplete version of the history.

A key spur for the development of theory for semiparametric models was the clear recognition by Neyman and Scott (1948) that maximum likelihood estimators are often inconsistent in the presence of an unbounded (with sample size) number of nuisance parameters. The simplest of these examples is as follows: suppose that

$$(X_i, Y_i) \sim N_2((\mu_i, \mu_i), \sigma^2), \quad i = 1, \ldots, n \tag{5.1}$$

J.A. Wellner (✉)
Department of Statistics, University of Washington, Seattle, WA, USA
e-mail: jaw@stat.washington.edu

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_5,

are independent where $\mu_i \in \mathbb{R}$ for $i = 1, \ldots, n$ and $\sigma^2 > 0$. Then the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}_n^2 = (4n)^{-1} \sum_{i=1}^{n} (X_i - Y_i)^2 \to_p \frac{\sigma^2}{2}.$$

This is an example of what has come to be known as a "functional model". The corresponding "structural model" (or mixture or latent variable model) is: (X_i, Y_i) are i.i.d. with density $p_{\sigma,G}$ where

$$p_{\sigma,G}(x,y) = \int \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) dG(\mu)$$

where ϕ is the standard normal density, $\sigma > 0$, and G is a (mixing) distribution on $\mathbb{R}$. Equivalently,

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Z \\ Z \end{pmatrix} + \sigma \begin{pmatrix} \delta \\ \varepsilon \end{pmatrix}$$

where $Z \sim G$ is independent of $(\delta, \varepsilon) \sim N_2(0, I)$, and only (X, Y) is observed. Here the nuisance parameters $\{\mu_i, i = 1, \ldots, n\}$ of the functional model (5.1) have been replaced by the (nuisance) mixing distribution G. Kiefer and Wolfowitz (1956) studied general semiparametric models of this "structural" or mixture type, $\{p_{\theta,G} : \theta \in \Theta \subset \mathbb{R}^d,\ G \text{ a probability distribution}\}$, and established consistency of maximum likelihood estimators $(\hat{\theta}_n, \hat{G}_n)$ of (θ, G). (Further investigation of the properties of maximum likelihood estimators in structural models (or semiparametric mixture models) was pursued by Aad van der Vaart in the mid 1990s; I will return to this later.)

Nearly at the same time as the work by Kiefer and Wolfowitz (1956) and Stein (1956) studied efficient testing and estimation in problems with many nuisance parameters (or even nuisance functions) of a somewhat different type. In particular Stein considered the one-sample symmetric location model

$$\mathscr{P}_1 = \{p_{\theta,f}(x) = f(x-\theta) : \ \theta \in \mathbb{R},\ f \text{ symmetric about } 0,\ I_f < \infty\}$$

and the two-sample (paired) shift model

$$\mathscr{P}_2 = \{p_{\mu,\nu,f}(x,y) = f(x-\mu)f(y-\nu) : \ \mu, \nu \in \mathbb{R},\ I_f < \infty\};$$

here $I_f \equiv \int (f'/f)^2 f dx$. Stein (1956) studied testing and estimation in models $\mathscr{P}_1$ and $\mathscr{P}_2$, and established necessary conditions for "adaptive estimation": for example, conditions under which the information bounds for estimation of θ in the model $\mathscr{P}_1$ are the same as for the information bounds for estimation of θ in the sub-model in which f is known. Roughly speaking, these are both cases in which the efficient score and influence functions are orthogonal to the "nuisance tangent space" in $L_2^0(P)$; i.e. orthogonal to all possible score functions for regular parametric submodels for the infinite-dimensional part of the model. Models of this type, and in particular the symmetric location model $\mathscr{P}_1$, remained as a focus of research during the period 1956–1982.

Over the period 1956–1982, considerable effort was devoted to finding sufficient conditions for the construction of "adaptive estimators" and "adaptive tests" in the context of the model $\mathscr{P}_1$: Hájek (1962) gave conditions for the construction of adaptive tests in the model $\mathscr{P}_1$, while van Eeden (1970) gave a construction for the sub-model of $\mathscr{P}_1$ consisting of log-concave densities (for which the score function for location is monotone non-decreasing), Beran (1974) constructed efficient estimators based on ranks, while Stone (1975) gave a construction of efficient estimators based on an "estimated" one-step approach.

This, modulo a key paper by Efron (1977) on asymptotic efficiency of Cox's partial likelihood estimators, was roughly the state of affairs of semiparametric theory in 1980–1982. Of course this is an oversimplification: much progress had been underway from a more nonparametric perspective from several quarters: the group around Lucien Le Cam in Berkeley, including P. W. Millar and R. Beran, the Russian school including I. Ibragimov and R. Has'minskii in (now) St. Petersburg and Y. A. Koshevnik and B. Levit in Moscow, and J. Pfanzagl in Cologne. Over the decade from 1982 to 1993 these two directions would merge and be understood as a whole piece of cloth, but that was not yet the case in 1980–1982, the period when Peter Bickel gave his Wald Lectures (and prepared them for publication).

5.1.2 Paper 1

The first of these four papers, *On Adaptive Estimation*, represents the culmination and summary of the first period of research on the phenomena of adaptive estimation uncovered by Stein (1956): it gives a masterful exposition of the state of "adaptive estimation" in the early 1980s, and new constructions of efficient estimators in several models satisfying Stein's necessary conditions for "adaptive estimation" in the sense of Stein (1956). Bickel (1982) begins in Sect. 5.1.2 with an explanation of "adaptive estimation", with focus on the "i.i.d. case", and introduces four key examples to be treated: (1) the one-sample symmetric location model $\mathscr{P}_1$ introduced above; (2) linear regression with symmetric errors; (3) linear regression with a constant and arbitrary errors, a model closely related to the two-sample shift model $\mathscr{P}_2$ introduced above; and (4) location and variance-covariance parameters of elliptic distributions. The paper then moves to an explanation of Stein's necessary condition and presentation of a (new) set of sufficient conditions for adaptive estimation involving $L_2(P_{\theta_m,G})-$consistent estimation of the efficient influence function ("Condition H"). Bickel shows that the sufficient conditions are satisfied in the Examples (1)–(4), and hence that adaptive estimators exist in each of these problems. It was also conjectured that Condition H is necessary for adaptation. Necessary and sufficient conditions only slightly stronger than "Condition H" were established by Schick (1986) and Klaassen (1987); also see Bickel et al. (1993, 1998), Sect. 7.8.

According to the ISI Web of Science, as of 20 June 2011, this paper has received 228 citations, and thus is the most cited of the four papers reviewed

here. It inspired the search for necessary and sufficient conditions for adaptive estimation (including the papers by Schick (1986) and Klaassen (1987) mentioned above). It also implicitly raised the issue of understanding efficient estimation in semiparametric models more generally. This was the focus of my joint work with Janet Begun, W. J. (Jack) Hall, and Wei-Min Huang at the University of Rochester during the period 1979–1983, resulting in Begun et al. (1983), which I will refer to in the rest of this discussion as BHHW.

5.1.3 Paper 2

Neyman and Scott (1948) had focused on inconsistency of maximum likelihood estimators in functional models, and Kiefer and Wolfowitz (1956) showed that inconsistency of likelihood-based procedures was not a difficulty for the corresponding structural (or mixture) models. Bickel and Klaassen (1986) initiated the exploration of efficiency issues in connection with functional models, with a primary focus on functional models connected with the symmetric location model $\mathscr{P}_1$. In particular, this paper examined the functional model with $X_i \sim N(\theta, \sigma_i^2)$ independent with $\sigma_i^2 \in \mathbb{R}^+$, $\theta \in \mathbb{R}$, for $1 \le i \le n$. The corresponding structural model is the normal scale mixture model with shift parameter θ, and hence is a subset of $\mathscr{P}_1$. In fact, it is a very rich subset with nuisance parameter tangent spaces (for "typical" points in the model) agreeing with that of the model $\mathscr{P}_1$. The main result of the paper is a theorem giving precise conditions under which a modified version of the estimator of Stone (1975) is asymptotically efficient, again in a precise sense defined in the paper.

This paper inspired further work on efficiency issues in functional models: see e.g. Pfanzagl (1993) and Strasser (1996). According to the ISI Web of Science (20 June 2011), it has been cited 15 times. These types of models remain popular (in September 2011, MathSciNet gives 414 hits for "functional model" and 480 hits for "structural model"), but many problems remain.

Between 1982 and publication of this paper in 1986, the paper Begun et al. (1983) appeared. In June 1983 Peter Bickel and myself had given a series of lectures at Johns Hopkins University on semiparametric theory as it stood at that time, and had started writing a book on the subject together with Klaassen and Ritov, Bickel et al. (1993, 1998), which was optimistically announced in the references for this paper as "BKRW (1987)".

5.1.4 Paper 3

This paper, Bickel and Ritov (1987), treats efficiency of estimation in the structural (or mixture model) version of the errors-in-variables model dating back at least to Neyman and Scott (1948) and Reiersol (1950), and perhaps earlier. As noted by the

authors: "Estimates of β in the general Gaussian error model, with Σ_0 diagonal, have been proposed by a variety of authors including Neyman and Scott (1948) and Rubin (1956). In the arbitrary independent error model, Wolfowitz in a series of papers ending in 1957, Kiefer, Wolfowitz, and Spiegelman (1979) by a variety of methods gave estimates, which are consistent and in Spiegelman's case $n^{1/2}-$consistent and asymptotically. Little seems to be known about the efficiency of these procedures other than that in the restricted Gaussian model ...". This model is among the first semiparametric mixture models involving a nontrivial projection in the calculation of the efficient score function to receive a thorough analysis and constructions of asymptotically efficient estimators. The authors gave an explicit construction of estimators achieving the information bound in a very detailed analysis requiring 17 pages of careful argument.

The type of construction used by the authors involves kernel smoothing estimators of the nonparametric part of the model, and hence brings in choices of smoothing kernels and smoothing parameters (ε_n, c_n and ν_n in the authors' notation, with $nc_n^2\nu_n^6 \to \infty$). This same approach was used by van der Vaart (1988) to construct efficient estimators in a whole class of structural models of this same type; van der Vaart's construction involved the choice of seven different smoothing parameters. On the other hand, Pfanzagl (1990a) pages 47 and 48 (see also Pfanzagl 1990b) pointed out that the resulting estimators are rather artificial in some sense, and advocated in favor of maximum likelihood or other procedures requiring no (or at least fewer) smoothing parameter choices. This approach was pursued in van der Vaart (1996). Forty years after Kiefer and Wolfowitz established consistency of maximum likelihood procedures, Van der Vaart proved, efficiency of maximum likelihood in several particular structural models (under moment conditions which are sufficient but very likely not necessary), including the errors-in-variables model treated in the paper under review. The proofs in van der Vaart (1996) proceed via careful use of empirical process theory. Furthermore, Murphy and van der Vaart (1996) succeeded in extending the maximum likelihood estimators to confidence sets via profile likelihood considerations.

This paper has 35 citations in the ISI Web of Science as of 20 June 2011, but it inspired considerable further work on efficiency bounds and especially on alternative methods for construction of efficient estimators.

5.1.5 Paper 4

In the period 1988–1991 several key questions on the "boundary" between nonparametric and semiparametric estimation came under close examination by van der Vaart, Bickel and Ritov, and Donoho and Liu. The lower bound theory under development for publication in BKRW (1993) relied upon Hellinger differentiability of real-valued functionals. (The lower bound theory based on pathwise Hellinger differentiability was put in a very nice form by van der Vaart (1991).)

But the possibility of a gap between the conditions for differentiability and sufficient conditions to attain the bounds became a nagging question. In Ritov and Bickel (1990), Peter and Ya'acov analyzed the situation in complete detail for the real-valued functional $\nu(P) = \int p^2(x)dx$ defined for the collection $\mathscr{P}$ of distributions P on $[0,1]$ with a density p with respect to Lebesgue measure. This functional turns out to be Hellinger differentiable at all such densities p with an information lower bound given by

$$I_\nu^{-1} = 4Var(p(X)) = 4\int (p(x) - \nu(P))^2 p(x)dx.$$

However, Theorem 1 of Ritov and Bickel (1990) shows that there exist distributions $P \in \mathscr{P}$ such every sequence of estimators of $\nu(p)$ converges to $\nu(p)$ more slowly than $n^{-\alpha}$ for every $\alpha > 0$. It had earlier been shown by Ibragimov and Hasminskii (1979) that the $\sqrt{n}-$ convergence rate could be achieved for densities satisfying a Hölder condition of order at least $1/2$, and in a companion paper to the one under discussion Bickel and Ritov (1988), Peter and Ya'acov showed that this continued to hold for densities p satisfying a Hölder condition of at least $1/4$.

These results have been extended to obtain rates of convergence in the "non-regular" or nonparametric domain: see Birgé and Massart (1993, 1995) and Laurent and Massart (2000). More recently the techniques of analysis have been extended still further Tchetgen et al. (2008) and Robins et al. (2009). As of 20 June 2011, this paper has been cited 45 times (ISI Web of Science).

5.1.6 Summary and Further Problems

The four papers reviewed here represent only a small fraction of Peter Bickel's work on the theory of semiparametric models, but they illustrate his superb judgement in the choice of problems suited to push both the theory of semiparametric models in general terms and having relevance for applications. They also showcase his wonderful ability to see his way through the technicalities of problems to solutions of theoretical importance and which point the way forward to further understanding. Paper 1 was clearly important in development of general theory for the adaptive case beyond the location and shift models $\mathscr{P}_1$ and $\mathscr{P}_2$. Paper 2 initiated efficiency theory for estimation in functional models quite generally. Paper 3 played an important role in illustrating how semiparametric theory could be applied to the structural (or mixing) form of the classical errors in variables model, hence yielding one of the first substantial models to be discussed in detail in the "non-adaptive case" in which calculation of the efficient score and efficient influence function requires a non-trivial projection.

As noted by Kosorok (2009) semiparametric models continue to be of great interest because of their "... genuine scientific utility ... combined with the breadth and depth of the many theoretical questions that remain to be answered".

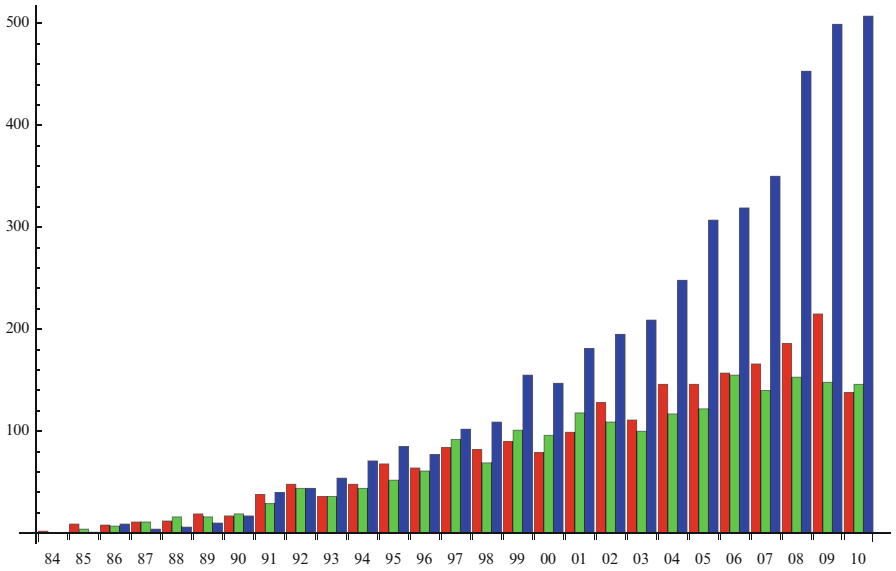

Fig. 5.1 Numbers of papers with "semiparametric" in title, keywords, or abstract, by year, 1984–2010. *Red* = MathSciNet; *Green* = Current Index of Statistics (CIS); *Blue* = ISI Web of Science

Figure 5.1 gives an update of Fig. 2.1 of Wellner et al. (2006). The trend is clearly increasing!

Acknowledgements Supported in part by NSF Grant DMS-0804587, and by NI-AID grant 2R01 AI291968-04.

References

Begun JM, Hall WJ, Huang W-M, Wellner JA (1983) Information and asymptotic efficiency in parametric–nonparametric models. Ann Stat 11(2):432–452

Beran R (1974) Asymptotically efficient adaptive rank estimates in location models. Ann Stat 2:63–74

Bickel PJ (1982) On adaptive estimation. Ann Stat 10(3):647–671

Bickel PJ, Klaassen CAJ (1986) Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. Adv Appl Math 7(1):55–69

Bickel PJ, Ritov Y (1987) Efficient estimation in the errors in variables model. Ann Stat 15(2):513–540

Bickel PJ, Ritov Y (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhyā Ser A 50(3):381–393

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, Baltimore

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1998) Efficient and adaptive estimation for semiparametric models. Springer, New York. Reprint of the 1993 original
Birgé L, Massart P (1993) Rates of convergence for minimum contrast estimators. Probab Theory Relat Fields 97(1–2):113–150
Birgé L, Massart P (1995) Estimation of integral functionals of a density. Ann Stat 23(1):11–29
Efron B (1977) The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc 72(359):557–565
Hájek J (1962) Asymptotically most powerful rank-order tests. Ann Math Stat 33:1124–1147
Ibragimov IA, Khasminskii RZ (1981) Statistical estimation: asymptotic theory. Springer Verlag, New York (Russian ed. 1979)
Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Stat 27:887–906
Klaassen CAJ (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat 15(4):1548–1562
Kosorok MR (2009) What's so special about semiparametric methods? Sankhyā 71(2, Ser A):331–353
Laurent B, Massart P (2000) Adaptive estimation of a quadratic functional by model selection. Ann Stat 28(5):1302–1338
Murphy SA, van der Vaart AW (1996) Likelihood inference in the errors-in-variables model. J Multivar Anal 59(1):81–108
Neyman J, Scott EL (1948) Consistent estimates based on partially consistent observations. Econ 16:1–32
Pfanzagl J (1990a) Estimation in semiparametric models. Lecture notes in statistics, vol 63. Springer, New York. Some recent developments
Pfanzagl J (1990b) Large deviation probabilities for certain nonparametric maximum likelihood estimators. Ann Stat 18(4):1868–1877
Pfanzagl J (1993) Incidental versus random nuisance parameters. Ann Stat 21(4):1663–1691
Reiersol O (1950) Identifiability of a linear relation between variables which are subject to error. Econometrica 18:375–389
Ritov Y, Bickel PJ (1990) Achieving information bounds in non and semiparametric models. Ann Stat 18(2):925–938
Robins J, Tchetgen Tchetgen E, Li L, van der Vaart A (2009) Semiparametric minimax rates. Electron J Stat 3:1305–1321
Rubin H (1956) Uniform convergence of random functions with applications to statistics. Ann Math Statist 27:200–203
Schick A (1986) On asymptotically efficient estimation in semiparametric models. Ann Stat 14(3):1139–1151
Stein C (1956) Efficient nonparametric testing and estimation. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability, 1954–1955, vol I. University of California Press, Berkeley/Los Angeles, pp 187–195
Stone CJ (1975) Adaptive maximum likelihood estimators of a location parameter. Ann Stat 3:267–284
Strasser H (1996) Asymptotic efficiency of estimates for models with incidental nuisance parameters. Ann Stat 24(2):879–901
Tchetgen E, Li L, Robins J, van der Vaart A (2008) Minimax estimation of the integral of a power of a density. Stat Probab Lett 78(18):3307–3311
van der Vaart AW (1988) Estimating a real parameter in a class of semiparametric models. Ann Stat 16(4):1450–1474
van der Vaart A (1991) On differentiable functionals. Ann Stat 19(1):178–204
van der Vaart A (1996) Efficient maximum likelihood estimation in semiparametric mixture models. Ann Stat 24(2):862–878
van Eeden C (1970) Efficiency-robust estimation of location. Ann Math Stat 41:172–181
Wellner JA, Klaassen CAJ, Ritov Y (2006) Semiparametric models: a review of progress since BKRW (1993). In: Frontiers in statistics. Imperial College Press, London, pp 25–44

The Annals of Statistics
1982, Vol 10, No. 3, 647-671

THE 1980 WALD MEMORIAL LECTURES

ON ADAPTIVE ESTIMATION

BY P. J. BICKEL[1]

University of California, Berkeley

We simplify a general heuristic necessary condition of Stein's for adaptive estimation of a Euclidean parameter in the presence of an infinite dimensional shape nuisance parameter and other Euclidean nuisance parameters. We derive sufficient conditions and apply them in the construction of adaptive estimates for the parameters of linear models and multivariate elliptic distributions. We conclude with a review of issues in adaptive estimation.

1. Introduction. In 1956, C. Stein published a paper in the Third Berkeley Symposium which deserves to be as well known as its celebrated companion piece on the inadmissibility of the normal mean. In this work Stein dealt with the problem of estimating and testing hypotheses about a Euclidean parameter θ or, more generally, a function $q(\theta)$ in the presence of an infinite dimensional "nuisance" shape parameter G. The question he asked (framed in estimation terms) was, "When can one estimate θ as well asymptotically not knowing G as knowing G?" He gave a simple necessary condition, which he checked in several important examples and, in one of these—testing that the center of symmetry has a specified value—he indicated a procedure that should work.

In recent years there has been considerable interest in an important situation where Stein's condition is satisfied, estimating the center of symmetry of an unknown symmetric distribution. Completely definitive results for this problem were obtained by Beran (1974) and Stone (1975). In this paper we return to Stein's original general formulation in the i.i.d. case. Motivated by his necessary condition for existence of adaptive estimates we obtain a simple sufficient condition for adaptation and apply it to a variety of important examples.

The paper is organized as follows. In Section 2 we define what we mean by adaptive estimation of θ; more precisely, we review some known results in the area and introduce the examples with which we will deal. In Section 3 we recall Stein's necessary condition for adaptation, and introduce a condition which we prove is sufficient. In Section 4 we check that our sufficient condition is satisfied in our examples. Section 5 contains a discussion of the connections between our work and recent research of Lindsay (1978, 1980), Hammerstrom (1978), Levitt (1974) and others, as well as a discussion of open questions. Finally, in Section 6, we gather technical parts of the proofs of our results.

2. What is adaptation? For simplicity we restrict ourselves throughout to the i.i.d. case. This is quite unnecessary for the heuristics of the paper. However, at least some of our proofs employ the assumed independence of the observations quite heavily.

Let $X_1, \cdots, X_n$ be i.i.d. k dimensional vectors with common distribution F. Let us recall the basic facts about the asymptotic theory of estimation when F ranges over a parametric model as put into their most elegant form by Le Cam.

Suppose that F is of the form F_θ where $\theta \in \Theta$, an open subset of R^p, and the F_θ have densities which we denote by $f(\cdot, \theta)$ with respect to a sigma-finite measure μ on R^k. Write

Received March 1981; revised November 1981.

[1] Research partially supported by ONR Grant No. N00014-80-C-0163 and the Adolph C. and Mary Sprague Miller Foundation for Basic Research in Science.

AMS 1970 *subject classification.* 62F20, 62G20.

Key words and phrases. Adaptation, efficient estimation, linear models, elliptic distributions.

E_θ, P_θ, $\mathscr{L}_\theta$ respectively for expectations, probabilities, and laws when θ holds. Let $\ell(x, \theta) = \log f(x, \theta)$, and define the following regularity conditions.

CONDITIONS R. *For all* $\theta \in \Theta$,

(i) $\ell(\cdot, \theta)$ *is differentiable in (the components of)* θ a.e. P_θ *and* $\dot{\ell} = (\partial\ell/\partial\theta_1, \cdots, \partial\ell/\partial\theta_p)$.

(ii) *The Fisher information matrix* $I(\theta)$ *exists,* $I(\theta) = E_\theta\{\dot{\ell}^T\dot{\ell}\,(X_1, \theta)\} < \infty$;

(iii) *Square root likelihood is differentiable in quadratic means, i.e. as* $t \to 0$,

$$E_\theta\left[\left\{\frac{f(X_1, \theta + t)}{f(X_1, \theta)}\right\}^{1/2} - 1 - \frac{t}{2}\dot{\ell}^T(X_1, \theta)\right]^2 = o(|t|^2),$$

and

$$P_{\theta+t}\{f(X_1, \theta) = 0\} = o(|t|^2),$$

where $|\cdot|$ *denotes the Euclidean norm (cf.* b_1 *and* b_2 *on page* 10 *of Le Cam,* 1969).

(iv) *There exist* $n^{1/2}$*-consistent estimates of* θ, *i.e.* $\{\tilde{\theta}_n(X_1, \cdots, X_n)\}$ *such that* $n^{1/2}(\tilde{\theta}_n - \theta) = O_{P_\theta}(1)$.

Under these conditions the following theorem holds (Le Cam, 1969; Fabian and Hannan, 1980). Call θ a *regular* point if $I(\theta)$ is nonsingular and if $I(\cdot)$ is continuous at θ.

THEOREM 2.1. *Under Conditions* R *there exist estimates* $\{\hat{\theta}_n\}$ *such that*

(a) *For all regular,* θ, $\mathscr{L}_{\theta_n}\{n^{1/2}(\hat{\theta}_n - \theta_n)\} \to \mathscr{N}(0, I^{-1}(\theta))$ *whenever* $n^{1/2}|\theta_n - \theta| \le M$ *for all* n, $M < \infty$.

(b) *The estimates* $\{\hat{\theta}_n\}$ *are asymptotically locally sufficient in the sense of Le Cam* (1969) *and locally asymptotically minimax in the sense of Hájek* (1972) *as modified by Fabian and Hannan* (1980).

Statement (a) says that $\{\hat{\theta}_n\}$ are efficient in the usual sense. Hájek (1972) also establishes, for $k = 1$, that any estimates satisfying (a) also are efficient in the sense of Rao. That is, if we define $\Delta_n(\cdot)$ by

$$\hat{\theta}_n = \theta + n^{-1}\textstyle\sum_{i=1}^n \dot{\ell}(X_i, \theta)I^{-1}(\theta) + \Delta_n(\theta), \tag{2.1}$$

then

$$n^{1/2}\Delta_n(\theta) \to_{P_\theta} 0, \tag{2.2}$$

for θ_n as in the theorem. In Theorem 6.1 (Section 6.4) we extend this result to general k.

REMARK 1. The construction of $\hat{\theta}_n$ used by Le Cam will prove useful to us later. Let $R_n^k = \{n^{-1/2}(i_1, \cdots, i_k), i_1, \cdots, i_k$ are arbitrary integers$\}$, and let

$$\bar{\theta}_n = \text{the point in } R_n^k \text{ closest to } \tilde{\theta}_n. \tag{2.3}$$

If $\dot{\ell}^*(\mathrm{x}, \theta)$ has the property that

$$n^{-1/2}\textstyle\sum_{i=1}^n \{\dot{\ell}^*(X_i, \theta_n) - \dot{\ell}(X_i, \theta)\} + n^{1/2}(\theta_n - \theta)I(\theta) = o_{P_\theta}(1)$$

whenever $n^{1/2}|\theta_n - \theta| \le M$, then Theorem 4 of Le Cam (1969) shows that

$$\hat{\theta}_n = \bar{\theta}_n + n^{-1}\textstyle\sum_{j=1}^n \dot{\ell}^*(X_j, \bar{\theta}_n)I^-(\bar{\theta}_n) \tag{2.4}$$

is efficient in the sense of Theorem 2.1; where I^- is a generalized inverse of I. Of course, this construction is not unique and has unpleasant aspects such as the "discretization" of $\tilde{\theta}_n$ and its non-iterative character. However, the construction works in great generality, i.e., under the mild and natural Conditions R(i)–R(iv).

We shall actually want to take $\dot{\ell}^* = \dot{\ell}$. To do so we need an inconsequential strengthening of R(iii) which is valid in all our examples. We call UR(iii) the assumption that for all θ

$\in \Theta$, the differentiability condition of R(iii) holds uniformly in some neighbourhood of θ. We show in Theorem 6.2 (Section 6.4) that R(i), R(ii) and UR(iii) enable us to take $\dot{\ell}^* = \dot{\ell}$ in (2.4).

REMARK 2. Condition R(iv), although clearly necessary, appears hard to verify. In fact, Le Cam shows that if we assume identifiability of θ and nonsingularity of $I(\theta)$ for all $\theta \in \Theta$, R(i)-R(iii) imply R(iv). We have chosen to leave R(iv) in its present form for reasons which will be apparent later.

In a preprint which we saw after our lectures were prepared, Fabian and Hannan (1980) give a very careful treatment of estimation in locally asymptotically normal families. They present, among other results, the "right" version of Hájek's local asymptotic minimaxity, as well as a rigorous discussion of Stein's (1956) necessary conditions for adaptation. Their notion of adaptation agrees with ours (in their more general framework).

The models for which we will discuss adaptation may be described as follows: The common d.f. F of the X_i ranges over a set which can be parametrized by a Euclidean parameter θ of interest, and a shape nuisance parameter G, i.e.,

$$\mathcal{F} = \{F_{(\theta, G)} : \theta \in \Theta, G \in \mathcal{G}\} \tag{2.5}$$

where Θ is an open subset of R^p, $\mathcal{G}$ is a set of distributions on some space, and the map $(\theta, G) \to F_{(\theta, G)}$ is known.

For each $G \in \mathcal{G}$, define

$$\mathcal{F}_G = \{F_{(\theta, G)} : \theta \in \Theta\}. \tag{2.6}$$

The models $\mathcal{F}_G$ are parametric models. Suppose that $\mathcal{F}_G$ satisfies R(i), R(ii) and UR(iii) for each $G \in \mathcal{G}$. Define $f(\cdot, \theta, G)$, $\ell(\cdot, \theta, G)$, $I(\theta, G)$ respectively as density, log likelihood, and information in $\mathcal{F}_G$. Call (θ, G) *regular* if θ is regular in $\mathcal{F}_G$. Finally, in view of the Le Cam theorem, we can state the following definition.

DEFINITION. *A sequence of estimates $\{\hat{\theta}_n\}$ is adaptive if and only if, for every regular (θ, G),*

$$\mathcal{L}_{\theta_n}\{n^{1/2}(\hat{\theta}_n - \theta_n)\} \to \mathcal{N}(0, I^{-1}(\theta, G)) \tag{2.7}$$

whenever $n^{1/2}|\theta_n - \theta|$ stays bounded. Thus adaptive estimates, if they exist, are efficient for every $\mathcal{F}_G$ even though knowledge of the true G may not be used in the construction of the estimates.

Adaptive estimates of θ have been constructed in the first of our examples.

EXAMPLE 1. *Estimation of the center of symmetry.* Let $k = p = 1$. Take $\Theta = R$, $\mathcal{G}$ = {All distributions symmetric about 0}, $F_{(\theta, G)}(x) = G(x - \theta)$.

The problem of adaptive estimation of θ in this model began to be studied by van Eeden (1970) and Takeuchi (1971), although the corresponding testing problem was earlier considered by Stein (1956) and solved by Hájek (1962). The definitive theorem was obtained by Beran (1974) and Stone (1975).

Let

$$I(G) = \int \{g'(x)\}^2/g(x)\, dx \tag{2.8}$$

whenever g, the density of G, is absolutely continuous, and let $I(G) = \infty$ otherwise.

THEOREM 2.2. *There exist translation and scale equivariant estimates, $\{\hat{\theta}_n\}$ such that*

$$\mathcal{L}_{(0, G)}(n^{1/2}\hat{\theta}_n) \to \mathcal{N}(0, I^{-1}(G)) \tag{2.9}$$

for all $G \in \mathcal{G}$ *with* $I(G) < \infty$.

Hájek (1962) has shown that for this model (θ, G) is regular if $I(\theta, G) = I(G) < \infty$. The converse is also true. Thus $\{\hat{\theta}_n\}$ are adaptive according to our general definition. In fact, Stone (1975) shows that the estimates he constructs satisfy (2.9) with $I^{-1}(G) = 0$ whenever $I(G) = \infty$. □

We will construct adaptive estimates of θ in the following generalization of Example 1.

EXAMPLE 2. *Estimation of regression with symmetric errors.* We describe the model structurally in terms of a variable $X \sim F_{(\theta, G)}$. Here $k = p + 1$ and $\Theta = R^p$. Let

$$X = (C, Y) \tag{2.10}$$

where C is a p dimensional random vector and Y a scalar. Further,

$$Y = C\theta^T + \varepsilon \tag{2.11}$$

where $\varepsilon \sim G$, and ε and C are independent. We again take

$$\mathcal{G} = \{\text{All distributions } G \text{ on } R \text{ symmetric about } 0\}.$$

Finally, we suppose

$$E(C^TC) \text{ is nonsingular.} \tag{2.12}$$

This is just a stochastic version of the usual multiple regression model,

$$X_i = C_i\theta^T + \varepsilon_i, \qquad i = 1, \cdots, n,$$

where $C_1, \cdots, C_n$ are p dimensional vectors of constants such that $C^T = (C_1^T, \cdots, C_n^T)$ and C^TC is nonsingular.

We deliberately do not specify that the distribution of C is known. The adaptive estimates we construct depend only on the data and work for any distribution of C satisfying (2.12). □

In many interesting situations a parameter θ for which efficient estimates exist in every model $\mathcal{F}_G$ cannot be consistently estimated in $\mathcal{F}$ because the parameter becomes unidentifiable. This is true in the next two examples. However, in both, natural functions $q(\theta)$ can be so estimated. In fact, adaptive estimation of these functions is possible. The definition of adaptive estimation of q is straightforward:

DEFINITION. *Suppose* $q: \Theta \to R^d$, $d \leq p$, *has a total differential* $\dot{q}(\theta)$, *a* $d \times p$ *matrix. A sequence of estimates* $\{\hat{q}_n\}$ *of* q *is adaptive if and only if, for every regular* (θ, G),

$$\mathcal{L}_{\theta_n}\{n^{1/2}(q_n - q(\theta_n))\} \to \mathcal{N}(0, \dot{q}(\theta)I^{-1}(\theta, G)\dot{q}(\theta)^T) \tag{2.13}$$

whenever $n^{1/2} \mid \theta_n - \theta \mid$ *stays bounded.*

EXAMPLE 3. *Regression with a constant and arbitrary errors.* In Example 2, let $C = (C^\circ, 1)$, C° a $p - 1$ dimensional vector. Define X, Y, ε as before and suppose ε and C are independent. However, let $\mathcal{G} = \{\text{all distributions on } R\}$, and replace (2.12) by

$$E(C^\circ - EC^\circ)^T(C^\circ - EC^\circ) \quad \text{nonsingular.} \tag{2.14}$$

Evidently θ is not identifiable in $\mathcal{F}$ since a change in the constant θ_p could equally well be a change in G. However, $q(\theta) = (\theta_1, \cdots, \theta_{p-1})$ can be adaptively estimated, as we shall see.

A special case of this model, where $p = 2$ and

$$C^\circ = \begin{cases} 1 & \text{with probability } \lambda \\ 0 & \text{with probability } 1 - \lambda, \end{cases}$$

can be thought of as a two-sample model with random sample sizes, i.e., we observe N observations with distribution $G(x - \theta_1 - \theta_2)$ and $n - N$ observations with distribution $G(x - \theta_2)$, where N has a binomial (n, λ) distribution.

Adaptation in the two-sample model with fixed sample sizes (and unknown scale) was studied by Stein (1956), Weiss and Wolfowitz (1970), and Wolfowitz (1974). A definitive result was obtained by Beran (1974). Weiss and Wolfowitz (1971) considered the fixed sample size multiple regression model and obtained partial results. □

EXAMPLE 4. *Parameters of elliptic distributions.* The following multivariate generalization of the symmetric one-sample location and scale model has been considered by Huber (1977) and others. Let

$$X = \mu + \varepsilon V^{-1/2}$$

where μ is an unknown $1 \times k$ vector, V is a positive definite $k \times k$ symmetric matrix, and $V^{-1/2}$ is the unique positive definite symmetric square root of V^{-1}. We suppose $\varepsilon \sim G$, where

$$\mathcal{G} = \{G : G \text{ absolutely continuous, spherically symmetric on } R^k\}.$$

Take $\theta = (\mu, [V])$ where for any symmetric $k \times k$ matrix $M = \| m_{ij} \|$, we define $[M]$ to be the lexicographically written row vector of the lower $k(k+1)/2$ entries of M. Thus, $p = k(k+3)/2$ and

$$\Theta = \{(\mu, [V]) : V \text{ symmetric positive definite}\}$$

is an open subset of R^p.

Here θ is efficiently estimable at regular points of $\mathcal{F}_G$ but is not identifiable in $\mathcal{F}$. A common scale change in all coordinates is ascribable to either V or G, yet $(\mu, V/\text{tr } V)$ can be estimated consistently, in fact, adaptively, as we shall see.

3. Stein's considerations and a sufficient condition for adaptation. We begin by recalling Stein's necessary condition for adaptation. Define a parametric subfamily of $\mathcal{G}$ as a set $\{\mathcal{G}_\eta\}$, $\eta \in T$, where T is an open set in R^t and the map $\eta \to G_\eta$ is smooth. The parametric submodel of $\mathcal{F}$ corresponding to the parametric subfamily $\{G_\eta\}$ is naturally defined by $\{F_{(\theta, G_\eta)} : \theta \in \Theta, \eta \in T\}$. Here is Stein's necessary condition.

CONDITION S. *For every parametric submodel obeying* R(i)-R(iv) *with* $G_{\eta_0} = G_0$

$$\int \left\{ \frac{\partial}{\partial \theta_i} \ell(x, \theta, G_\eta) \frac{\partial}{\partial \eta_j} \ell(x, \theta, G_\eta) \right\}_{\theta = \theta_0, \eta = \eta_0} f(x, \theta_0, G_0) \mu(dx) = 0 \tag{3.1}$$

$$i = 1, \cdots, p, \quad j = 1, \cdots, t.$$

Stein (1956) shows that if an adaptive estimate of θ exists and (θ_0, G_0) is regular, then Condition S must hold. The argument is simple. Let

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

where I_{11} is $p \times p$ and I_{22} is $t \times t$, be the $(p+t) \times (p+t)$-dimensional Fisher information matrix of the parametric submodel $F_{(\theta, G_\eta)}$ evaluated at (θ_0, η_0), and write

$$I^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}.$$

Now, by definition, if $\{\hat{\theta}_n\}$ is adaptive, then $I_{11}^{-1} = I^{-1}(\theta_0, G_0)$ is the asymptotic variance covariance matrix of $n^{1/2}(\hat{\theta}_n - \theta_n)$ whenever $n^{1/2} | \theta_n - \theta |$ stays bounded. But, by Hájek's (1972) theorem, I^{11} is the smallest variance covariance matrix achievable in this way. Thus $I_{11}^{-1} = I^{11}$ which is equivalent to $I_{12} = 0$, which is Condition S.

Condition S suffers from two defects: (i) it can be awkward to verify, (ii) it is unclear how to proceed from it to the construction of adaptive procedures. We now proceed to derive a simpler condition which is at least heuristically necessary and which in turn leads to a verifiable sufficient condition.

All the examples we have studied exhibit the following simple convexity structure:

CONDITION C. *$\mathcal{G}$ is convex and $G_0, G_1 \in \mathcal{G}$ implies that for $0 \leq \alpha \leq 1$*

$$F_{(\theta,\alpha G_0+(1-\alpha)G_1)} = \alpha F_{(\theta,G_0)} + (1-\alpha)F_{(\theta,G_1)}.$$

This structure suggests that we examine Condition S for the following $\{G_\eta\}$. Fix G_0 and G_1, take $T = (0, 1)$, and let

$$G_\eta = \eta G_0 + (1-\eta)G_1.$$

Then Condition S becomes for $\eta > 0$, $i = 1, \cdots, p$,

$$\int \frac{\partial}{\partial\theta_i}\ell(x, \theta, G_\eta)\{f(x, \theta, G_1) - f(x, \theta, G_0)\}\mu(dx) = 0.$$

Letting $\eta \to 0$ formally we get for "all" $G_0, G_1 \in \mathcal{G}$ that the following holds.

CONDITION S*.

$$\int \dot{\ell}(x, \theta, G_0)f(x, \theta, G_1)\mu(dx) = 0.$$

It may be shown formally that if Condition S* holds, so does Condition S (Bickel, 1979). Condition S* has a simple heuristic interpretation. If G_0 is a fixed shape in $\mathcal{G}$ let θ_n^* be the M-estimate corresponding to G_0, i.e., solving

$$\textstyle\sum_{i=1}^n \dot{\ell}(x_i, \theta_n^*, G_0) = 0.$$

We know that, under regularity conditions (Huber, 1967), if Condition S* holds, then $n^{1/2}(\theta_n^* - \theta)$ is asymptotically normal under $F_{(\theta,G)}$ with mean 0 and variance covariance matrix $A^{-1}B(A^T)^{-1}$, where

$$\begin{aligned} A &= \left\| -\int \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(x, \theta, G_0)f(x, \theta, G)\mu(dx)\right\|, \\ B &= \int \dot{\ell}^T(x, \theta, G_0)\dot{\ell}(x, \theta, G_0)f(x, \theta, G)\mu(dx). \end{aligned} \tag{3.2}$$

A heuristic summary of this is as follows. Firstly, M-estimates corresponding to a fixed shape G_0 should be $n^{-1/2}$ consistent for θ under every shape G_1. Secondly, suppose we can estimate the true G by data-dependent $\{G_n\}$ so that the score functions $\ell(\cdot, \cdot, G_n)$ converge to $\ell(\cdot, \cdot, G)$ and so that the matrices A_n, B_n obtained by replacing G_0 by G_n in (3.2) converge to $I(\theta, G)$. It then seems plausible that the sequence of M-estimates corresponding to G_n is adaptive.

Motivated by these considerations we now formulate two conditions, GR(iv) and H.

CONDITION GR(iv). *There exist estimates $\{\tilde{\theta}_n\}$ such that $n^{1/2}(\tilde{\theta}_n - \theta) = O_{P_{(\theta,G)}}(1)$ at all regular points (θ, G).*

Let

$$\begin{aligned} \mathcal{H} = \{h: &h \text{ maps } R^k \times \Theta \text{ to } R^k \text{ and} \\ &\int h(x, \theta)F_{(\theta,G)}(dx) = 0 \text{ for all } \theta \in \Theta, G \in \mathcal{G}\}. \end{aligned} \tag{3.3}$$

In view of Condition S*, $\mathscr{H}$ includes the space of possible score functions. For convenience we introduce

$$\tilde{\ell}(x, \theta, G) = \dot{\ell}(x, \theta, G)I^{-}(\theta, G), \tag{3.4}$$

where I^{-} is any generalized inverse. (In fact we only need $\tilde{\ell}$ for θ such that $I(\theta, G)$ is nonsingular.) Note that $\tilde{\ell}$ can be substituted for $\dot{\ell}$ in Condition S*. Here is our main condition:

CONDITION H. *Appropriate consistent estimation of score functions is possible. That is, there exists a sequence of maps* $\hat{\hat{\ell}}_m : (R^k)^m \to \mathscr{H}$, $m = 1, 2, \cdots$, *taking* $(x_1, \cdots, x_m)$ *into* $\hat{\hat{\ell}}(\cdot, \cdot; x_1, \cdots, x_m)$ *such that for all regular* (θ, G) *and any* $|\theta_m - \theta| = O(m^{-1/2})$,

$$\int |\hat{\hat{\ell}}_m(x, \theta_m; X_1, \cdots, X_m) - \tilde{\ell}(x, \theta_m, G)|^2 F_{(\theta_m, G)}(dx) \to 0 \tag{3.5}$$

in $P_{(\theta,G)}$ *probability.*

Note that GR(iv) is evidently a necessary condition for adaptive estimation and is the natural generalization of R(iv). Under Condition S*, M-estimates corresponding to a fixed shape are natural candidates for $\tilde{\theta}_n$. In view of Stein's necessary Condition S*, we conjecture that Condition H is necessary for adaptation. W. R. van Zwet pointed out a suggestive inequality bolstering this conjecture (Klaassen, 1980, Theorem 3.2.1). In any case these conditions are sufficient.

THEOREM 3.1. *If Conditions* GR(iv) *and* H *hold, then adaptive estimates exist.*

NOTE. The construction is closely related to that given for adaptive rank tests in the linear model by Hájek (1962). A related construction for Example 1 has been given by Bretagnolle (private communication). See also Hasminskii and Ibragimov (1978).

PROOF. Define $\tilde{\theta}_n$ as in (2.3). Let $\{m(n)\}$ be a sequence of subsample sizes with $m(n) = o(n)$. Write m for $m(n)$ and let $\bar{n} = n - m$.
Define

$$\hat{\theta}_n = \bar{\theta}_n + \bar{n}^{-1} \sum_{i=m+1}^{n} \hat{\hat{\ell}}(X_i, \bar{\theta}_n; X_1, \cdots, X_m). \tag{3.6}$$

We claim $\{\hat{\theta}_n\}$ is adaptive. By Theorem 6.2,

$$\bar{\theta}_n + \bar{n}^{-1} \sum_{i=m+1}^{n} \tilde{\ell}(X_i, \bar{\theta}_n, G)$$

is efficient for every regular (θ, G). Write P_θ for $P_{(\theta,G)}$. Then to prove the theorem it is enough to show

$$\bar{n}^{-1/2} \sum_{i=m+1}^{n} \{\hat{\hat{\ell}}_m(X_i, \bar{\theta}_n; X_1, \cdots, X_m) - \tilde{\ell}(X_i, \bar{\theta}_n, G)\} = o_{P_\theta}(1). \tag{3.7}$$

Now we use a trick of Le Cam's and note that we need only establish (3.7) with $\tilde{\theta}_n$ replaced by $\theta_n = \theta + t_n n^{-1/2}$, where t_n is an arbitrary convergent deterministic sequence. This follows since $\bar{\theta}_n$ is $\sqrt{n}$ − consistent and the intersection of its range with any sphere of radius $Mn^{-1/2}$ about θ is finite with cardinality bounded independent of n. Having made the replacement, we prove (3.7). Note that R(i) − R(iii) imply that the $\bar{n}$ dimensional product measures of $X_{m+1}, \cdots, X_n$ under P_θ and under P_{θ_n} are contiguous. Therefore, it suffices to prove (3.7) in P_{θ_n} probability. Condition on $X_1, \cdots, X_m$ for this probability. Since $\hat{\hat{\ell}}(\cdot, \cdot; X_1, \cdots, X_m) \in \mathscr{H}$,

$$\int \hat{\hat{\ell}}_m(x, \theta_n; X_1, \cdots, X_m) f(x, \theta_n, G) \mu(dx) = 0 \tag{3.8}$$

and by R(i) – R(iii),

$$\int \tilde{\ell}(x, \theta_n, G) f(x, \theta_n, G) \mu(dx) = 0. \tag{3.9}$$

Therefore

$$E_{\theta_n}[|\bar{n}^{-1/2} \textstyle\sum_{i=m+1}^{n} \{\hat{\tilde{\ell}}_m(X_i, \theta_n; X_1, \cdots, X_m) - \tilde{\ell}(X_i, \theta_n, G)\}|^2 | X_1, \cdots, X_m] \tag{3.10}$$
$$= \int |\hat{\tilde{\ell}}_m(x, \theta_n; X_1, \cdots, X_m) - \tilde{\ell}(x, \theta_n, G)|^2 f(x, \theta_n, G) \mu(dx) \to 0$$

in P_θ probability by Condition H and hence, by contiguity again, in P_{θ_n} probability. Claim (3.7) is proved, and the theorem follows. □

NOTES. It is possible to replace Condition H by the following condition H′ which permits separate estimation of $\dot{\ell}$ and I^{-1}.

CONDITION H'. (a) *There exist maps* $\hat{\dot{\ell}}_m (R^k)^m \to \mathscr{H}$ *such that for all regular* (θ, G), $|\theta_m - \theta| = O(m^{-1/2})$

$$\int |\hat{\dot{\ell}}_m(x, \theta_m; X_1, \cdots, X_m) - \dot{\ell}(x, \theta_m, G)|^2 f(x, \theta_m, G) \mu(dx) = o_{P_\theta}(1). \tag{3.11}$$

(b) *There exist estimates* $\hat{I}_m(X_1, \cdots, X_m)$ *of* $I(\theta, G)$ *consistent for all regular* (θ, G).

It is easy to show that if GR(iv) and H′ both hold, and if we define

$$\theta_n^* = \bar{\theta}_n + \bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \hat{\dot{\ell}}(X_i, \bar{\theta}_n; X_1, \cdots, X_m) \hat{I}_n^- \tag{3.12}$$

then

$$\theta_n^* = \bar{\theta}_n + \bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \hat{\dot{\ell}}(X_i, \bar{\theta}_n; X_1, \cdots, X_m) I^{-1}(\theta, G) + o_{P_\theta}(n^{-1/2}) \tag{3.13}$$

and θ_n^* is adaptive.

A natural choice of $\hat{I}_n$ is provided by

$$\hat{I}_n = \bar{n}^{-1} \textstyle\sum_{i=m+1}^{n} \hat{\dot{\ell}}^T \hat{\dot{\ell}}(X_i, \theta_n; X_1, \cdots, X_m) \tag{3.14}$$

We show in Section 6.2 that this choice of $\hat{I}_n$ is consistent for regular (θ, G) provided that GR(iv) and (3.11) hold, and if

$$m^{-1} \textstyle\sum_{i=1}^{m} \dot{\ell}^T \dot{\ell}(X_i, \theta_m, G) \to I(\theta, G) \tag{3.15}$$

in P_θ probability for all regular (θ, G).

These are the results we will apply to Example 2 and which are applicable to other situations where all of θ is estimable. To deal with Examples 3 and 4 we need an extension of our theory. First we study the analogue of Condition S* when we only ask that $q(\theta)$, rather than all of θ, be estimated adaptively. Stein considers this question in a slightly different formulation. He writes $\theta = (q, t)$ with $q = q(\theta)$ and t, the rest of θ, is a nuisance parameter, and he introduces the model $\{F_{(\theta, G\eta)}\}$. He notes that adaptive estimation of q is possible only if the upper left-hand corner of the inverse of the information matrix for (q, t) with $\eta = \eta_0$ fixed is the same as the upper left-hand corner of the inverse of the information matrix for (q, t, η) evaluated at η_0. We do not pursue further his matrix formulation of this condition, but only note that in the presence of convexity Condition C, Stein's condition is heuristically equivalent to the d equations

CONDITION S* (*generalized*).

$$\int \dot{\ell}(x, \theta, G_0) I^{-1}(\theta, G_0) \dot{q}^T(\theta) f(x, \theta, G_1) \mu(dx) = 0$$

for every shape $G_0, G_1 \in \mathcal{G}$. For $q(\theta) = \theta$, $\dot{q}$ is the identity and our more general formulation of S* agrees with our old one.

New difficulties are introduced by the possible lack of identifiability of θ. Of course we need to have q identifiable. That is, if

$$F_{(\theta_0, G_0)} = F_{(\theta_1, G_1)} = F \tag{3.16}$$

then

$$q(\theta_0) = q(\theta_1).$$

But adaptation requires more. If F can be embedded in both $\mathcal{F}_{G_0}$ and $\mathcal{F}_{G_1}$ as in (3.16), then the information bound for estimation of q must be the same in both parametric families. That is, (3.16) implies

$$\dot{q}(\theta_0) I^-(\theta_0, G_0) \dot{q}^T(\theta_0) = \dot{q}(\theta_1) I^-(\theta_1, G_1) \dot{q}^T(\theta_1). \tag{3.17}$$

This condition is satisfied in all our examples because if $\mathcal{F}_{G_0}$ and $\mathcal{F}_{G_1}$ have a member in common then they are the same, or, rather, one is a smooth relabelling of the other. For instance, in Example 3, (3.16) holds if and only if G_1 is obtained from G_0 by a translation. We shall use this structural feature in a stronger way to reduce $\mathcal{G}$ and make θ identifiable. Here is a formal statement of our structural assumptions. They are obviously satisfied in Examples 3 and 4.

ASSUMPTION A1. *Either* $\mathcal{F}_{G_0} = \mathcal{F}_{G_1}$ *or* $\mathcal{F}_{G_0} \cap \mathcal{F}_{G_1} = \emptyset$, *for all* $G_0, G_1 \in \mathcal{G}$.

ASSUMPTION A2. *There exists* $T \subset R^{p-d}$ *and a smoothly invertible map from* Θ *to* $Q \times T$ *where* $Q = q(\Theta)$ *which carries* θ *into* $(q(\theta), t(\theta))$. *That is, we can identify* q *with a piece of* θ.

ASSUMPTION A3. *If we replace* θ *by* (q, t) *and* $\mathcal{F}_{G_0} = \mathcal{F}_{G_1}$ *there exists a unique smoothly invertible mapping* $\tau(q, \cdot)$ *of* T *into itself defined by* $F_{(q,t,G_0)} = F_{(q,\tau,G_1)}$.

Assumption A1 implies that there exists an "identifying subset" $\mathcal{G}_0 \subset \mathcal{G}$ such that (i) $\mathcal{F} = \{F_{(\theta,G)} : G \in \mathcal{G}_0, \theta \in \Theta\}$, and (ii) θ is identifiable when G is restricted to $\mathcal{G}_0$ provided that it is identifiable in each $\mathcal{F}_G$. We can select $\mathcal{G}_0$ as a set of representatives of the equivalence classes generated by the relation $G_1 \equiv G_2 \Leftrightarrow \mathcal{F}_{G_1} = \mathcal{F}_{G_2}$. For instance, in Example 3 we can take $\mathcal{G}_0 = \{G : \mu(G) = 0\}$ where μ is a location parameter. As we noted, Assumptions A2 and A3 imply that if (a) $\mathcal{F}_G = \mathcal{F}_{G_0}$, $G_0 \in \mathcal{G}_0$, and (b) $F_{(\theta,G)} = F_{(\theta_0,G_0)}$, then $q(\theta) = q(\theta_0)$ and (3.16) holds. That is, it does not matter in which parametric model $\mathcal{F}_G$ we embed a distribution F. The value of q and the ease with which q can be estimated remain the same. Since we can talk about estimation of θ for $(\theta, G) \in \Theta \times \mathcal{G}_0$ it is natural to propose the following extensions of the conditions for $\sqrt{n}$-consistency and appropriate consistent estimation of score functions.

GENERALIZED CONDITION GR(iv). *There exists* $\mathcal{G}_0$ *satisfying* (i) *and* (ii) *above and estimates* $\{\tilde{\theta}_n\}$ *such that*

$$n^{1/2}(\tilde{\theta}_n - \theta) = O_{P_{(\theta,G)}}(1)$$

for all (θ, G), $G \in \mathcal{G}_0$.

We now redefine $\tilde{\ell}$, $\mathcal{H}$ *for given* q. Our definitions agree with the old ones when q is the identity. Let

$$\mathcal{H} = \left\{ h : h \text{ maps } R^k \times \Theta \text{ into } R^d \text{ so that } \int h(x, \theta) f(x, \theta, G) \mu(dx) = 0 \quad \text{for all} \quad (\theta, G) \right\}.$$

$$\tilde{\ell}(x, \theta, G) = \ell(x, \theta, G) I^-(\theta, G) \dot{q}^T(\theta). \tag{3.18}$$

Condition H is now generalized as was condition GR(iv), merely by substituting $\mathscr{G}_0$ for $\mathscr{G}$. The easy extension of Theorem 3.1 is as follows.

THEOREM 3.2. *If Assumptions* A1–A3 *and the generalized conditions* GR(iv) *and* H *hold, then adaptive estimates* $\{\hat{q}_n\}$ *of* $q(\theta)$ *exist.*

The proof is the same as for Theorem 3.1 when we propose as estimate

$$\hat{q}_n = q(\bar{\theta}_n) + \bar{n}^{-1} \sum_{i=m+1}^{n} \hat{\dot{\ell}}_m(X_i, \bar{\theta}_n; X_1, \cdots, X_m). \tag{3.19}$$

4. Adaptation in Examples 1–4. For the examples we leave verification of the trivial structural Assumptions A1 through A3 to the reader. In each example we shall proceed through the following steps:

Step A. Formally verify Stein's orthogonality Condition S* and in the process construct what we can think of as the "space of possible score functions" $\mathscr{H}$ or a suitable subset $\mathscr{H}_0$.

Step B. Find a suitable identifying subset $\mathscr{B}_0$ and construct $\sqrt{n}$-consistent estimates $\{\tilde{\theta}_n\}$ so as to satisfy GR(iv).

Step C. Construct score function estimates $\hat{\dot{\ell}}$ satisfying (3.5) and taking values in $\mathscr{H}_0$ i.e. satisfy Condition H for the appropriate consistent estimation of score functions, or satisfy its modification H′ providing for separate estimation of $\dot{\ell}$ and I.

Since Example 1 is a special case of Example 2 and has already been dealt with satisfactorily, we begin with Example 2. For convenience from now on we write P for P_θ.

EXAMPLE 2. *Step A.* If the distribution of C has density r with respect to some ν, and if G has density g, then $X = (C, Y)$ has density (with respect to the product measure)

$$f(c, y, \theta, G) = r(c)g(y - c\theta^T), \tag{4.1}$$

and

$$\dot{\ell}(c, y, \theta, G) = c\frac{g'}{g}(y - c\theta^T). \tag{4.2}$$

Then

$$E_{(\theta,G_0)}\dot{\ell}(C, Y, \theta, G) = E_{(\theta,G_0)}\left\{C\frac{g'(\varepsilon)}{g(\varepsilon)}\right\} = E(C)E_{G_0}\left\{\frac{g'(\varepsilon)}{g(\varepsilon)}\right\} = 0,$$

since g'/g is antisymmetric and G_0 is symmetric about 0. Thus, Condition S* is satisfied and by our argument, $\mathscr{H} \supset \mathscr{H}_0$ where $h \in \mathscr{H}_0$ if and only if

$$h(c, y, \theta) = c\psi(y - c\theta^T) \tag{4.3}$$

for ψ bounded and antisymmetric, i.e.

$$\psi(y) = -\psi(-y). \tag{4.4}$$

So we will use score function estimates of the form (4.3).

Step B. Let $\psi: R \to R$ be such that ψ is twice continuously differentiable, with ψ and its derivatives bounded. Suppose, moreover, that $\psi' > 0$ and that ψ is antisymmetric. Let $\{\tilde{\theta}_n\}$ be the M-estimates corresponding to ψ, i.e., the unique solutions of

$$\sum_{i=1}^{n} C_{ij}\psi(Y_i - C_i\tilde{\theta}_n^T) = 0, \quad j = 1, \cdots, p, \tag{4.5}$$

where $X_i = (C_i, Y_i)$, $C_i = (C_{i1}, \cdots, C_{ip})$. Then by Huber's theorem (Huber, 1973), $\{\tilde{\theta}_n\}$ are $\sqrt{n}$-consistent. (This is just the construction suggested in the previous section.)

Step C. By modifying the arguments of Hájek (1972) it is easy to see that (θ, G) is regular if g is absolutely continuous with derivative g' and if $I(G)$, the Fisher information

for location given in Section 2, is finite. The converse is also true (proof available from author).

By (4.2) we calculate

$$\tilde{\ell}(c, y, \theta, G) = c\frac{g'}{g}(y - c\theta^T)\{E(C^TC)I(G)\}^{-1}, \tag{4.6}$$

where the last term is just $I^{-1}(\theta, G)$. To apply Condition H or H′ we need to estimate g'/g and $I(G)$. This is achieved by the following lemma whose proof is given in Section 6.1.

LEMMA 4.1. *Let $\varepsilon_1, \varepsilon_2, \cdots$ be* i.i.d. *random variables. There exists a sequence of function estimates $q_m: R \times R^m \to R$, $m = 1, 2, \cdots$, such that q_m is bounded for each m and such that as $m \to \infty$*

$$\int \left\{q_m(y; \varepsilon_1, \cdots, \varepsilon_m) - \frac{g'(y)}{g(y)}\right\}^2 g(y)\, dy \to 0 \tag{4.7}$$

in probability whenever the common d.f. *of the ε_i is G with density g and $I(G) < \infty$.*

We proceed to show how to estimate $\dot{\ell}$ and $I(G)$ separately and verify Condition H′. Let

$$\hat{\varepsilon}_i = Y_i - C_i\bar{\theta}_m^T(X_1, \cdots, X_m), \quad i = 1, \cdots, m, \tag{4.8}$$

be the residuals with respect to the "discretized" estimate based on the first m observations. Define

$$\psi_m(y; X_1, \cdots, X_m) = \tfrac{1}{2}\{q_m(y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - q_m(-y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m)\} \tag{4.9}$$

and

$$\hat{\dot{\ell}}_m(c, y, \theta; X_1, \cdots, X_m) = c\psi_m(y - c\theta^T; X_1, \cdots, X_m). \tag{4.10}$$

Clearly $\hat{\dot{\ell}}(\cdot; X_1, \cdots, X_m) \in \mathscr{H}_0$ and

$$\begin{aligned}
&\int |\hat{\dot{\ell}}_m(c, y, \theta_m; X_1, \cdots, X_m) - \dot{\ell}(c, y, \theta_m, G)|^2 f(c, y, \theta_m, G)\, dy\, \nu(dc) \\
&= \int c\left|\psi_m(y - c\theta_m^T; X_1, \cdots, X_m) - \frac{g'}{g}(y - c\theta_m^T)\right|^2 c^T g(y - c\theta_m^T)\, dy\, \nu(dc) \qquad (4.11) \\
&\leq \left[\left\{\int \left|q_m(y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - \frac{g'}{g}(y)\right|^2 g(y)\, dy\right\} ECC^T\right].
\end{aligned}$$

Now let $\theta_m = \theta + t_m$, where t_m and $c_1, \cdots, c_m$ are p-dimensional vectors such that $|t_m| = O(m^{-1/2})$ and $\sum_{i=1}^m c_i t_m^T t_m c_i^T$ is bounded independent of m. Then the sequence of m-dimensional product measures induced by $\varepsilon_1, \cdots, \varepsilon_m$ and $\varepsilon_1 - c_1 t_m^T, \cdots, \varepsilon_m - c_m t_m^T$ are contiguous if $I(G) < \infty$ (Hájek and Sidák, 1967, page 211). Since ECC^T is finite, if $|t_m| = O(m^{-1/2})$, $\sum_{i=1}^m c_i t_m^T t_m c_i^T = O_{P_\theta}(1)$. Thus, by Lemma 4.1,

$$\int \left|q_m(y; \varepsilon_1 - C_1 t_m^T, \cdots, \varepsilon_m - C_m t_m^T) - \frac{g'}{g}(y)\right|^2 g(y)\, dy \to_{P_\theta} 0. \tag{4.12}$$

But, as usual, by the structure of $\bar{\theta}_m$ and its $m^{1/2}$-consistency, this result is enough to establish

$$\int \left\{q_m(y; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) - \frac{g'}{g}(y)\right\}^2 g(y)\, dy \to_{P_\theta} 0. \tag{4.13}$$

Substituting in (4.11), we see that $\hat{\dot{\ell}}_m$ is a consistent estimate of $\dot{\ell}$ in the sense of part (a) of Condition H′, in (3.11).

There are various ways to construct $\hat{I}_n$. For instance, we can verify (3.15) in this case as follows:

$$m^{-1}\sum_{i=1}^{m}\dot{\ell}^T\dot{\ell}(X_i, \theta_m, G) = m^{-1}\sum_{i=1}^{m} C_i^T C_i\left(\frac{g'}{g}\right)^2 (Y_i - C_i\theta_m^T)$$

(4.14)

$$\to_{P_{\theta_m}} E(C^TC)I(G) = I(\theta, G)$$

by the weak law of large numbers. By contiguity we can replace θ_m by θ in P_{θ_m}. This yields as the consistent estimate of (3.14),

$$\hat{I}_n^{(1)} = \bar{n}^{-1}\sum_{i=m+1}^{n} C_i^T C_i\psi_m^2(Y_i - C_i\bar{\theta}_n^T; \hat{\epsilon}_1, \cdots, \hat{\epsilon}_m). \tag{4.15}$$

A more familiar alternative, which may similarly be shown to work, is

$$\hat{I}_n^{(2)} = (n^{-1}\sum_{i=1}^{n} C_i^T C_i)\bar{n}^{-1}\sum_{i=m+1}^{n}\psi_m^2(Y_i - C_i\bar{\theta}_m^T; \hat{\epsilon}_1, \cdots, \hat{\epsilon}_m). \tag{4.16}$$

We have proved the following result.

THEOREM 4.1. *Let $\tilde{\theta}_n$ be defined as in* (4.5), *ψ_m as in* (4.9). *Let*

$$\hat{\theta}_n = \bar{\theta}_n + \bar{n}^{-1}\sum_{i=m+1}^{n} C_i\psi_m(Y_i - C_i\bar{\theta}_m^T; \hat{\epsilon}_1, \cdots, \hat{\epsilon}_m) \tag{4.17}$$

where $\hat{I}_n$ is given by (4.15) *or* (4.16). *Then $\{\hat{\theta}_n\}$ is adaptive in Example* 2.

EXAMPLE 3.

Step A. If $c = (c^\circ, 1)$, $q(\theta) = (\theta_1, \cdots, \theta_{p-1})$ and $\tilde{\ell}$ is defined by (3.18), we get

$$\tilde{\ell}(c, y, \theta, G) = (c^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}\frac{g'}{g}(y - c\theta^T)I^{-1}(G). \tag{4.18}$$

Thus, formally

$$E_{(\theta,G_0)}\tilde{\ell}(X, \theta, G) = E(C^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}E\frac{g'}{g}(\varepsilon)I^{-1}(G) = 0$$

and Condition S* is satisfied. In view of (4.18) it is natural to choose

$$\mathcal{H}_0 = \{h: h(c, y, \theta) = (c^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}\psi(y - c\theta^T), \psi \text{ bounded}\}. \tag{4.19}$$

Step B. Let ψ be as in Step B of Example 2 and define

$$\mathcal{G}_0 = \left\{G: \int \psi(y)G(dy) = 0\right\}. \tag{4.20}$$

Evidently $\mathcal{G}_0$ is an identifying subset and, by Huber's theorem, $\{\tilde{\theta}_n\}$ corresponding to ψ are $\sqrt{n}$-consistent when G is restricted to $\mathcal{G}_0$.

Step C. A possible definition of $\hat{\tilde{\ell}}$ is just

$$\hat{\tilde{\ell}}_m(c, y, \theta; X_1, \cdots, X_m) = (c^\circ - EC^\circ)(\text{Var } C^\circ)^{-1}q_m(y - c\theta^T; \hat{\epsilon}_1, \cdots, \hat{\epsilon}_m)\hat{I}^{-1}, \tag{4.21}$$

where

$$\hat{I} = \bar{n}^{-1}\sum_{i=m+1}^{n} q_m^2(Y_i - C_i\bar{\theta}_n^T; \hat{\epsilon}_1, \cdots, \hat{\epsilon}_m), \tag{4.22}$$

q_m is given in Lemma 4.1 and the $\hat{\epsilon}_i$ are defined by (4.8). That $\hat{\tilde{\ell}}$ works is evident by the same argument as we gave for Theorem 4.1, since regular (θ, G) again correspond to $I(G) < \infty$. This is not satisfactory, however, because the resultant estimates depend on the first and second moments of the unknown distribution of C°. We claim that estimating these

does just as well. Here is one way of proceeding. Define

$$\bar{C}_n^\circ = n^{-1}\sum_{i=1}^n C_i^\circ \tag{4.23}$$

$$\hat{\mathrm{V}}\mathrm{ar}\, C^\circ = n^{-1}\sum_{i=1}^n (C_i^\circ - \bar{C}_n^\circ)^T(C_i^\circ - \bar{C}_n^\circ).$$

Let

$$\hat{q}_n = \bar{\theta}_n^{(p-1)} + \bar{n}^{-1}\sum_{i=m+1}^n (C_i^\circ - \bar{C}_n^\circ)(\hat{\mathrm{V}}\mathrm{ar}\, C_n^\circ)^- q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m)\hat{I}^{-1} \tag{4.24}$$

where $\bar{\theta}_n^{(p-1)}$ is the vector of the initial $p-1$ elements of $\bar{\theta}_n$.

THEOREM 4.2. *The estimates $\hat{q}_n$ defined by* (4.24) *adaptively estimate* $(\theta_1, \cdots, \theta_{p-1})$ *in Example* 3.

PROOF. We know that

$$\bar{n}^{-1}\sum_{i=m+1}^n (C_i^\circ - EC^\circ)q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m)(\mathrm{Var}\, C^\circ)^- = o_P(n^{-1/2}) \tag{4.25}$$

and

$$\hat{\mathrm{V}}\mathrm{ar}\, C^\circ = \mathrm{Var}\, C^\circ + o_P(1). \tag{4.26}$$

Therefore, replacing Var C° by $\hat{\mathrm{V}}\mathrm{ar}\, C^\circ$ in (4.21) will still lead to adaptive estimates. Thus to establish that the estimates given by (4.24) are adaptive it suffices to prove that

$$\bar{n}^{-1}\sum_{i=m+1}^n (\bar{C}_n^\circ - EC^\circ)(\hat{\mathrm{V}}\mathrm{ar}\, C^\circ)^- q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) = o_P(n^{-1/2}) \tag{4.27}$$

or, since $\bar{C}_n^\circ - EC^\circ = O_P(n^{-1/2})$, that

$$\bar{n}^{-1}\sum_{i=m+1}^n q_m(Y_i - C_i\bar{\theta}_n^T; \hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_m) = o_P(1). \tag{4.28}$$

To prove (4.28) we show that we can replace q_m by g'/g and $Y_i - c_i\bar{\theta}_n^T$ by ε_i and then apply the law of large numbers. Details are given in Section 6.2. □

EXAMPLE 4. *Step A.* In this case if $\theta = (\mu, [V])$, then

$$f(x, \theta, G) = \{\det(V)\}^{1/2}\gamma(\{(x-\mu)V(x-\mu)^T\}^{1/2}) \tag{4.29}$$

where det denotes determinant, and γ maps R^+ into itself. Of course, $\gamma(|x|)$ is the density of G. We want to estimate

$$q([\mu, [V]) = (\mu, q_0([V])) \tag{4.30}$$

where q_0 is any homogeneous function of $[V]$. A "most general" choice is $q_0([V]) = [V]/\mathrm{tr}(V)$. We can write, for (θ, G_0) regular,

$$\dot{\ell}(x, \mu, G_0)I^{-1}(\theta, G_0) = (\psi^\circ(x, \mu, V), [\chi^\circ(x, \mu, V)])$$

where ψ° is $1 \times k$, χ° is $k \times k$ symmetric, and $[\chi]$ denotes the $k(k+1)/2$ dimensional vector of the lower half of χ. It is shown in Section 6.3 that

$$\psi^\circ(x, \mu, V) = \psi^\circ((x-\mu)V^{1/2}, 0, J)V^{-1/2} \tag{4.31}$$

$$\chi^\circ(x, \mu, V) = V^{1/2}\chi^\circ((x-\mu)V^{1/2}, 0, J)V^{1/2}, \tag{4.32}$$

where J is the $k \times k$ identity matrix. We further show in Section 6.3 that, if $|\cdot|$ is the Euclidean norm and $\gamma_0(|x|)$ is the density of G_0, then

$$\psi^\circ(x, 0, J) = -\frac{x}{|x|}\frac{\gamma_0'}{\gamma_0}(|x|)kI_1^{-1}(G_0) \tag{4.33}$$

and

$$\chi_{ij}^{\circ}(x, 0, J) = \begin{cases} I_2^{-1}(G_0)k(k+2)\dfrac{x_i x_j}{|x|}\dfrac{\gamma_0'}{\gamma_0}(|x|), & i \neq j, \\ 2\left\{I_2(G_0)\dfrac{3}{k(k+2)} - 1\right\}^{-1}\left\{\dfrac{x_i^2}{|x|}\dfrac{\gamma_0'}{\gamma_0}(|x|) + 1\right\}, & i = j, \end{cases} \tag{4.34}$$

where

$$I_1(G) = c_k \int_0^{\infty} r^{k-1}\frac{[\gamma']^2}{\gamma}(r)\, dr \tag{4.35}$$

$$I_2(G) = c_k \int_0^{\infty} r^{k+1}\frac{[\gamma']^2}{\gamma}(r)\, dr \tag{4.36}$$

and c_k is the surface area of the unit sphere in R^k. Then by (4.31) and (4.32),

$$E_{(\theta,G)}\{\psi^{\circ}(X, \mu, V), [\chi^{\circ}(X, \mu, V)]\}\dot{q}^T(\theta) \tag{4.37}$$

$$= E_{(0,[J],G)}\{\psi^{\circ}(X, 0, J)V^{-1/2}, [V^{1/2}\chi^{\circ}(X, 0, J)V^{1/2}]\}\dot{q}^T(\theta).$$

Moreover, if $i \neq j$, X_{ij}° changes sign if all the coordinates of x other than x_i are left unchanged while $x_i \to -x_i$. Since if $\theta = (0, [J])$, all the X_i are identically distributed and the distributions of $(X_1, \cdots, X_k)$ and $(\pm X_1, \cdots, \pm X_k)$ are the same, we conclude that

$$E_{(0,[J],G)}\psi^{\circ}(X, 0, J) = 0 \tag{4.38}$$

$$E_{(0,[J],G)}\chi^{\circ}(X, 0, J) = cJ, \tag{4.39}$$

where c depends on G and G_0. Therefore

$$E_{(0,[J],G)}[V^{1/2}\chi^{\circ}(X, 0, J)V^{1/2}] = c[V]. \tag{4.40}$$

Substituting (4.38) and (4.40) back into (4.37) we find that all components of (4.37) vanish either by (4.38) or by Euler's equation $\sum_{k \geq \ell} v_{k\ell}\, \partial q_0/\partial v_{k\ell} = 0$.

The orthogonality Condition S* follows and our argument makes it clear that $\mathscr{H}$ defined in (3.3), contains the set $\mathscr{H}_0$ of $h(x, \theta)$ defined by

$$h(x, \theta) = (\psi((x-\mu)V^{1/2})V^{-1/2}, [V^{1/2}\chi((x-\mu)V^{1/2})V^{1/2}])\dot{q}^T(\theta), \tag{4.41}$$

where ψ is $1 \times k$ and χ is symmetric $k \times k$ with forms

$$\psi(x) = \omega(|x|)\frac{x}{|x|}a_1 \tag{4.42}$$

$$\chi_{ij}(x) = \begin{cases} \omega(|x|)\dfrac{x_i x_j}{|x|}a_2, & i \neq j, \\ \left\{\omega(|x|)\dfrac{x_i^2}{|x|} + 1\right\}a_3, & i = j, \end{cases} \tag{4.43}$$

where ω is bounded and a_1, a_2, a_3 are constant. Clearly $\mathscr{H}$ is much bigger than $\mathscr{H}_0$, but $\mathscr{H}_0$ is the space of natural estimates of $\tilde{\ell}$.

Step B. Thanks to Maronna (1976) we can find an identifying subset $\mathscr{G}_0$ and corresponding $\sqrt{n}$ − consistent $\tilde{\theta}_n$ as follows. Let u_1 and u_2 be functions on R^+. Define the M-estimate $(\tilde{\mu}_n, \tilde{V}_n)$ corresponding to u_1 and u_2 to be any solution of

$$n^{-1}\sum_{i=1}^n u_1(\{(X_i - \tilde{\mu}_n)\tilde{V}_n(X_i - \tilde{\mu}_n)^T\}^{1/2}) = 0 \tag{4.44}$$

$$n^{-1}\sum_{i=1}^n u_2(\{(X_i - \tilde{\mu}_n)\tilde{V}_n(X_i - \tilde{\mu}_n)^T\})(X_i - \tilde{\mu}_n)^T(X_i - \tilde{\mu}_n) = [\tilde{V}_n]^{-1}$$

if one exists, and arbitrarily otherwise.

It is easy to see that the maximum likelihood estimates for a particular G are of this type. Let u_1, u_2 satisfy conditions (A) – (D) on page 53 of Maronna (1976). In addition, if $\psi_i(s) = su_i(s)$, $i = 1, 2$, suppose that $s\psi_j'(s)$ are bounded, $j = 1, 2$, and $\psi_1' > 0$. By Theorem 5.6 of Maronna, under these conditions $n^{1/2}(\tilde{\mu}_n - \tilde{\mu}, \tilde{V}_n - \tilde{V}) = O_P(1)$ for all $F \in \mathscr{F}$ where $\tilde{\mu}(\mu, V, G)$, $\tilde{V}(\mu, V, G)$ satisfy *uniquely*

$$\int u_1(\{(x-\tilde{\mu})\tilde{V}(x-\tilde{\mu})^T\}^{1/2})(x-\tilde{\mu})f(x, \theta, G)\,dx = 0 \tag{4.45}$$

$$\int u_2(\{(x-\tilde{\mu})\tilde{V}^T(x-\tilde{\mu})^T\})(x-\tilde{\mu})^T(x-\tilde{\mu})f(x, \theta, G)\,dx = [\tilde{V}]^{-1}. \tag{4.46}$$

It is clear by the unicity of $\tilde{\mu}$, $\tilde{V}$ that

$$\tilde{\mu}(\mu, V, G) = \mu, \tag{4.47}$$

$$\tilde{V}(\mu, V, G) = c(G)V, \tag{4.48}$$

where $c(G)$ is that measure of scale which is the unique solution of the equation

$$E\{u_2(c\,\varepsilon\,\varepsilon^T)\} = \frac{1}{c};$$

existence is guaranteed by the monotonicity of u_2. Clearly we can take as an identifying subset

$$\mathscr{G}_0 = \{G : c(G) = 1\} \tag{4.49}$$

and $\tilde{\theta}_n = (\tilde{\mu}_n, \tilde{V}_n)$ defined by (4.44).

Step C. It may be shown that regularity of (θ, G) is equivalent to absolute continuity of γ on $(0, \infty)$ and finiteness of $I_1(G)$ and $I_2(G)$. (Proof available from author.) We will show how to construct adaptive estimates of $q_0(V)$ in a simple fashion and then discuss the simultaneous adaptive estimation of μ.

Note that if X has density given by (4.29), then $\log|(X-\mu)V^{1/2}|$ has density j given by

$$j(z) = c_k e^{kz}\gamma(e^z). \tag{4.50}$$

Thus

$$\frac{\gamma'}{\gamma}(y) = y^{-1}\left\{\frac{j'}{j}(\log y) - k\right\}, \quad y > 0, \tag{4.51}$$

and this leads to the following construction of an estimate of γ'/γ.

Let $\bar{\mu}_m$ be obtained by discretizing $\tilde{\mu}_m$ as usual while $[\bar{V}_m]$ is the closest member of the $m^{-1/2}$ lattice to $\tilde{V}_m$ which itself corresponds to a positive definite matrix. Let

$$z_{im} = \log|(X_i - \bar{\mu}_m)\bar{V}_m^{1/2}|, \quad i = 1, \cdots, m,$$

and define

$$\omega_m(y; X_1, \cdots, X_m) = y^{-1}\{q_m'(\log y; z_{1m}, \cdots, z_{mm}) - k\}. \tag{4.52}$$

We claim that

$$\int |x|^2 \left|\omega_m(|x|; X_1, \cdots, X_m) - \frac{\gamma'}{\gamma}(|x|)\right|^2 \gamma(|x|)\,dx \to 0 \tag{4.53}$$

in P_θ probability if (θ, G) is regular. The proof follows the usual lines. By construction of $\bar{\mu}_m$, $\bar{V}_m$ it is possible to treat them as deterministic sequences such that $|\bar{\mu}_m - \mu|$ and $|\bar{V}_m - V| = O(m^{-1/2})$. Since (θ, G) is regular the m-dimensional product measures induced by $\varepsilon_1, \cdots, \varepsilon_m$ and $(X_1 - \bar{\mu}_m)\bar{V}_m^{1/2}, \cdots, (X_m - \bar{\mu}_m)\bar{V}_m^{1/2}$ are contiguous. If we also use (4.51) we can conclude that (4.53) is equivalent to

$$\int \left|q_m(\log|x|; \log|\varepsilon_1|, \cdots, \log|\varepsilon_m|) - \frac{j'}{j}(\log|x|)\right|^2 \gamma(|x|)\,dx \to 0 \tag{4.54}$$

in probability whenever $\varepsilon_1, \cdots, \varepsilon_m$, are i.i.d. with common distribution G such that $I_1(G)$ and $I_2(G)$ are finite. But the integral in (4.54) equals

$$\int_{-\infty}^{\infty} \left| q_m(z; \log|\varepsilon_1|, \cdots, \log|\varepsilon_m|) - \frac{j'}{j}(z) \right|^2 g(z)\, dz. \tag{4.55}$$

Moreover,

$$\int_{-\infty}^{\infty} \frac{(j')^2}{j}(z)\, dz = \int_{-\infty}^{\infty} \left\{ e^z \frac{\gamma'}{\gamma}(e^z) + k \right\}^2 g(z)\, dz = I_2(G) - k^2 \tag{4.56}$$

using integration by parts. Thus the integral in (4.55) tends to 0 whenever $I_2(G) < \infty$ by Lemma 4.1 and (4.54) and hence (4.53) holds. Now that we have an estimate $\omega_m(\cdot; X_1, \cdots, X_m)$ of γ'/γ we can estimate $I_2(G)$ by, for instance, splitting our preliminary sample of m, taking $m = 2\ell$ and letting

$$\hat{I}_2 = \ell^{-1} \sum_{i=\ell+1}^{m} q_m^2(z_{im}; z_{1m}, \cdots, z_{\ell m}) + k^2. \tag{4.57}$$

Evidently $\hat{I}_2$ depends only on $X_1, \cdots, X_m$. Moreover, we can argue as for (4.28) that, whenever (θ, G) is regular,

$$\hat{I}_2 \to I_2(G) \text{ in probability.} \tag{4.58}$$

Now define $\hat{\chi}_0(\cdot, O, J)$ by substituting $\hat{I}_2$ for $I_2(G_0)$ and $\omega_m(\cdot; X_1, \cdots, X_m)$ for γ_0'/γ_0 in (4.34) and let

$$\hat{\tilde{\ell}}_m(x, \theta; X_1, \cdots, X_m) = [V_m^{1/2}\, \hat{\chi}_0((x-\mu)V^{1/2}, O, J) V_m^{1/2}] \dot{q}_0^T([V]). \tag{4.59}$$

This is the natural estimate of $\tilde{\ell}$ corresponding to $q_0([V])$. Now after some algebra, if $\theta_m = (\mu_m, [V_m])$,

$$\begin{aligned} \int |\hat{\tilde{\ell}}_m(x, \theta_m; X_1, \cdots, X_m) - \dot{\ell}(x, \theta_m, G) I^{-1}(\theta_m, G)(0, \dot{q}_0([V]))^T|^2 f(x, \theta_m, G)\, dx \\ = O_P\Bigg(\int |(x-\mu_m)V_m^{1/2}|^2 \Bigg| \omega_m((x-\mu_m)V_m^{1/2}; X_1, \cdots, X_m) \\ - \frac{\gamma'}{\gamma}(|(x-\mu_m)V_m^{1/2}|) \Bigg|^2 f(x, \theta_m, G)\, dx \Bigg) + O_P(\hat{I}_2 - I_2). \end{aligned} \tag{4.60}$$

But the right-hand side of (4.60) is $o_p(1)$ by (4.53) and (4.58). From (4.60) and the structure of $\hat{\tilde{\ell}}$ we see that $\hat{\tilde{\ell}}$ falls in $\mathscr{H}_0$ given by (4.41) and is appropriately consistent. We have proved the following result.

THEOREM 4.3. *In Example* 4, *if we define*

$$\hat{q}_{on} = q_0([\bar{V}_n]) + \bar{n}^{-1} \sum_{i=m+1}^{n} \hat{\tilde{\ell}}_m(X_i, \bar{\theta}_n; X_1, \cdots, X_m), \tag{4.61}$$

then $\{\hat{q}_{on}\}$ *is an adaptive estimate of* $q_0([V])$.

To estimate μ simultaneously and adaptively using the estimate of γ'/γ we need to show that

$$\int \left| \omega_m(|x|; X_1, \cdots, X_m) - \frac{\gamma'}{\gamma}(|x|) \right|^2 \gamma(|x|)\, dx \to 0 \tag{4.62}$$

in probability, or equivalently that

$$\int_{-\infty}^{\infty} e^{-2z} \left| q_m(z; \log|\varepsilon_1|, \cdots, \log|\varepsilon_m|) - \frac{j'}{j}(z) \right|^2 g(z)\, dz \tag{4.63}$$

in probability. Unfortunately, to show (4.63) we need

$$\int_{-\infty}^{\infty} e^{-2z} \frac{(j')^2}{j}(z)\, dz = c_k \int_0^{\infty} y^{k-1}\left\{\frac{\gamma'}{\gamma}(y) + ky^{-1}\right\}^2 \gamma(y)\, dy < \infty \tag{4.64}$$

and this happens if $I_1(G) < \infty$ and

$$\int_0^{\infty} y^{k-3}\gamma(y)\, dy < \infty, \tag{4.65}$$

a superfluous condition.

To get rid of (4.65) we need to estimate γ'/γ differently by smoothing the multivariate empirical distribution of $(X_i - \bar{\mu}_m)\bar{V}_m^{1/2}$ and constructing an estimate of γ'/γ out of the first partial derivatives of the smoothed empirical distribution. This can be done but we omit the tedious and rather technical definition of the estimate and the necessary argument.

5. Questions raised by this work and other issues in adaptive estimation.

5.1 *When is adaptation not possible?* We have seen heuristically the necessity of the $\sqrt{n}$-consistency condition GR(iv) and the orthogonality Condition S when there are no nuisance parameters. In parametric models $\sqrt{n}$-consistency is available under mild smoothness and identifiability conditions while orthogonality is special. Orthogonality seems special in these nonparametric nuisance parameter models as well. We illustrate with a famous example of Neyman and Scott. The failure of adaptation in this case was already noted by Wolfowitz (1953).

EXAMPLE 5. *Estimation in Model II.* Suppose $X_i = (X_{i1}, X_{i2})$, $i = 1, \cdots, n$, such that

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad j = 1, 2, \tag{5.1}$$

where the ε_{ij} are independent identically distributed $\mathcal{N}(0, \theta)$, and the μ_i are independent and identically distributed with common distribution G. Let $\Theta = R^+$, $\mathcal{G} = \{\text{all distributions on } R\}$. It is easy to see that all (θ, G) are regular, and there is a natural $\sqrt{n}$-consistent estimate, the best unbiased estimate when the μ_i are treated as constants,

$$\bar{\theta}_n = \frac{1}{2n}\sum_{i=1}^n (X_{i1} - X_{i2})^2. \tag{5.2}$$

Thus Condition GR(iv) holds. But Condition H does not. For instance, take G_0 to be point mass at 0. Then

$$\dot{\ell}(x_1, x_2, \theta, G_0) = \frac{1}{\theta}\left\{\frac{(x_1^2 + x_2^2)}{2\theta} - 1\right\} \tag{5.3}$$

and

$$E_{(\theta,G)}\dot{\ell}(X, \theta, G_0) = \frac{1}{\theta^2}\int \mu^2 dG(\mu) > 0 \tag{5.4}$$

unless $G = G_0$. Thus adaptation in the sense we have discussed is not possible. Note that the natural estimate $\bar{\theta}_n$ has asymptotic variance $2\theta^2/n$ in this case while $I^{-1}(\theta, G_0) = \theta^2/n$. Lindsay (1978, 1980) and Hammerstrom (1978) have independently studied situations such as this one (which are the rule rather than the exception) where adaptation is not possible. They have obtained what may be viewed as a minimax optimality property of $\bar{\theta}_n$ in Example 5 and analogous results in other problems of this type. We are investigating the natural extension of adaptation in this context.

5.2 *Better estimates.* The estimates we construct in Examples 2–4 have some serious

failings: (i) the estimate of $\dot{\ell}$ is based on a small subsample rather than all the data; (ii) the estimates do not have natural invariance properties possessed by reasonable estimates in these problems, primarily because of the discretization of $\bar{\theta}_n$; and (iii) the behavior of the estimates when $I(\theta, G)$ is singular is not analyzed.

We believe that analogues of Stone's procedures in the location problem (which meet all these criticisms) can be constructed using the special structures of our examples. We have not pursued this since our interest lies primarily in illustrating the applicability of the general Condition H.

5.3 *Extensions to other asymptotic structures.* The theory we have developed extends naturally to cases where the observations are independent but not identically distributed, e.g., the usual linear model context. It can be applied, we believe, to the linear model and, as Stein's calculations and Wolfowitz (1974) indicate, to multiple regression models where both the location and the scale of the dependent variable are functions (possibly nonlinear) of the independent variables. Other extensions to non-independent situations, such as that treated in part in Beran (1976), should also be possible.

5.4 *Efficient estimation of functionals.* Levitt (followed by Ibragimov and Khazminski and others), in a series of papers starting with Levitt (1974), has studied how best to estimate functions $\theta(F)$ in nonparametric models, basing this work in part on Stein (1956). In some sense our problem can be viewed as the estimation of the solution $\theta(F)$ of $\int \dot{\ell}\ (x, \theta, G)\ dF_{(\theta,G)}(x) = 0$ which is meaningful (though possibly nonexistent) for $F \in \mathcal{F}$. Beyond this formal connection there seems to be no real link between our studies.

5.5 *Uniformity of adaptation.* Beran (1978) notes in the location problem (Example 1) that adaptive estimates converge to their limiting distributions uniformly on (shrinking n-dependent) "contiguous" neighborhoods of each G. This property can, we believe, be suitably re-expressed to apply generally. However, the weakness of this property is pointed out by Klaassen (1980) who shows (in Example 1, his Theorems 3.2.1 and 3.3.2) that for reasonable fixed neighborhoods the convergence is far from uniform. Thus from a practical point of view adaptive estimates may not work nearly as well for moderate samples as we might expect.

5.6 *Practical questions.* The difficulty of nonparametric estimation of score functions suggests that a more practical goal is partial adaptation, the construction of estimates which are (i) always $\sqrt{n}$-consistent, and (ii) efficient over a large parametric subfamily of $\mathcal{F}$. Our results indicate that when the orthogonality Condition S* and $\sqrt{n}$-consistency Condition GR(iv) hold, this goal should be achievable by using a one-step Newton approximation to the maximum likelihood estimate for the parametric subfamily by starting with an estimate which is $\sqrt{n}$-consistent for all of $\mathcal{F}$. Partial adaptation in Example 2 is discussed in Hogg (1980). This highlights an important practical and theoretical question in problems of this type, how to construct $\sqrt{n}$-consistent estimates. When there are no nuisance parameters present and adaptation is possible, maximum likelihood estimates for fixed shapes are natural candidates. In general, this question deserves further study. The constructions of Birgé (1980) may prove useful.

6. Theoretical Details.

6.1 *Proof of Lemma* 4.1. We use Stone's (1975) approach. Let ϕ_σ be the $\mathcal{N}(0, \sigma^2)$ density, g be any density, and define the convolution of g and ϕ_σ

$$g_\sigma = g * \phi_\sigma \tag{6.1}$$

and the convolution of the empirical d.f. and ϕ_σ

$$\hat{g}_\sigma(y) = m^{-1} \sum_{i=1}^{m} \phi_\sigma(y - \varepsilon_i). \tag{6.2}$$

We suppress dependence on $\varepsilon_1, \cdots, \varepsilon_m$ in what follows.

For given σ_m, c_m, d_m, $e_m > 0$ define

$$q_m(y) = \begin{cases} \dfrac{\hat{g}'_{\sigma_m}}{\hat{g}}(y) & \text{if } \hat{g}_{\sigma_m}(y) \geq d_m, \quad |y| \leq e_m \text{ and } |\hat{g}'_{\sigma_m}(y)| \leq c_m \hat{g}_{\sigma_m}(y), \\ 0 \quad \text{otherwise.} \end{cases} \tag{6.3}$$

We claim that if $c_m \to \infty$, $e_m \to \infty$, $\sigma_m \to 0$ and $d_m \to 0$ in such a way that

$$\sigma_m c_m \to 0, \tag{6.4}$$

$$e_m \sigma_m^{-3} = o(m), \tag{6.5}$$

then q_m satisfies the conclusions of Lemma 4.1. The argument proceeds by

LEMMA 6.1. *If the conditions of Lemma* 4.1 *hold and* q_m *satisfies* (6.3)–(6.5), *then*

$$\int_{[R>0]} \left\{ q_m(y) - \frac{g'_{\sigma_m}}{g_{\sigma_m}}(y) \right\}^2 g_{\sigma_m}(y)\, dy \to_P 0. \tag{6.6}$$

PROOF. We use the elementary estimates noted in Stone. For κ_i universal constants and all y,

$$\operatorname{Var} \hat{g}_\sigma^{(i)}(y) \leq \kappa_i \sigma^{-(2i+1)} m^{-1} g_\sigma(y), \qquad i = 0, 1, \cdots. \tag{6.7}$$

Denote the conditions in (6.3) by A, B, C and the left-hand side of (6.6) by $I_1 + I_2$, where

$$I_1 = \int_{ABC} \left\{ \frac{\hat{g}'_{\sigma_m}}{\hat{g}_{\sigma_m}}(y) - \frac{g'_{\sigma_m}}{g_{\sigma_m}}(y) \right\}^2 g_{\sigma_m}(y)\, dy \tag{6.8}$$

$$I_2 = \int_{[ABC]^c} \frac{[g'_{\sigma_m}]^2}{g_{\sigma_m}}(y)\, dy. \tag{6.9}$$

Bound $E(I_1)$ by

$$2\left[\int_{ABC} g_{\sigma_m}^{-1}(y) E\{\hat{g}'_{\sigma_m}(y) - g'_{\sigma_m}(y)\}^2\, dy + \int_{ABC} c_m^2 g_{\sigma_m}^{-1}(y) E\{\hat{g}_{\sigma_m}(y) - g_{\sigma_m}(y)\}^2\, dy \right] = o(1) \tag{6.10}$$

by (6.7), (6.4) and (6.5). Bound

$$\begin{aligned} E(I_2) \leq \int \frac{[g'_{\sigma_m}]^2}{g_{\sigma_m}}(y) [P\{|\hat{g}'_{\sigma_m}(y)| > c_m \hat{g}_{\sigma_m}(y)\} \\ + P\{\hat{g}_{\sigma_m}(y) < d_m, g(y) > 0\} + I(|y| > e_m)]\, dy. \end{aligned} \tag{6.11}$$

We claim that

$$\hat{g}_{\sigma_m}(y) \to g(y) \quad \text{in probability for all } y \quad \text{if} \quad m\sigma_m \to \infty, \tag{6.12}$$

$$\hat{g}'_{\sigma_m}(y) \to g'(y) \quad \text{in probability a.e. } y \quad \text{if} \quad m\sigma_m^3 \to \infty, \tag{6.13}$$

$$\int \frac{g'^2_{\sigma_m}}{g_{\sigma_m}}(y)\, dy \leq \int \frac{g'^2}{g}(y)\, dy \quad \text{for all } m. \tag{6.14}$$

Evidently (6.12) and (6.13) imply that if $c_m \to \infty$ and $d_m \to 0$, then the two probabilities in (6.11) tend to 0 a.e. y, while (6.12)–(6.14) imply uniform integrability of $g'^2_{\sigma_m}/g_{\sigma_m}(y)$ and

hence that

$$EI_2 \to 0. \tag{6.15}$$

Together (6.10) and (6.15) will establish Lemma 6.1. It remains to prove (6.12)-(6.14). Now by (6.7), for all y,

$$\hat{g}_{\sigma_m}(y) - g_{\sigma_m}(y) \to 0 \quad \text{in probability if} \quad m\sigma_m \to \infty, \tag{6.16}$$

$$\hat{g}'_{\sigma_m}(y) - g'_{\sigma_m}(y) \to 0 \quad \text{in probability if} \quad m\sigma_m^3 \to \infty. \tag{6.17}$$

Continuity of g and (6.16) imply (6.12). To prove (6.13) write (using the absolute continuity of g),

$$\begin{aligned} \int_{-\infty}^{\infty} |g'_{\sigma_m}(y) - g'(y)|\, dy &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} (g'(y - \sigma_m x) - g'(y))\phi(x)\, dx \right| dy \\ &\le \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g'(y - \sigma_m x) - g'(y)|\, dy \phi(x)\, dx. \end{aligned} \tag{6.18}$$

Note that $I(G) < \infty$ implies $\int_{-\infty}^{\infty} |g'(y)|\, dy < \infty$. Thus we can apply the L_1 continuity theorem and the dominated convergence theorem to conclude that the right-hand side of (6.18) tends to 0 as $\sigma_m \to 0$ and (6.13) follows from (6.17) and (6.18). Finally, (6.14) is a well known inequality (see Hájek and Šidák, 1967, page 17). The lemma is proved. □

Next we need

LEMMA 6.2. *If* $\sigma \to 0$,

$$\int_{[g>0]} \left\{ \frac{g'_\sigma}{\sqrt{g_\sigma}}(y) - \frac{g'}{\sqrt{g}}(y) \right\}^2 dy \to 0. \tag{6.19}$$

PROOF. Apply (6.12)-(6.14).

LEMMA 6.3. *If* $\sigma_m c_m \to 0$,

$$\int_{[g>0]} q_m^2(y)(\sqrt{g_{\sigma_m}(y)} - \sqrt{g(y)})^2\, dy \to_P 0. \tag{6.20}$$

PROOF. Write, using Cauchy's form of Taylor's theorem,

$$\begin{aligned} \sqrt{g_\sigma(y)} - \sqrt{g(y)} &= \sigma \int_0^1 \left\{ \frac{\partial}{\partial \sigma} g_{\sigma\lambda}(y)/2g_{\sigma\lambda}^{1/2}(y) \right\} d\lambda \\ &= -\frac{\sigma}{2} \int_0^1 g_{\sigma\lambda}^{-1/2}(y) \int_{-\infty}^{\infty} zg'(y - \lambda\sigma z)\phi(z)\, dz\, d\lambda. \end{aligned} \tag{6.21}$$

Thus we can bound the square in the integrand of (6.20) by

$$\begin{aligned} &\frac{\sigma_m^2}{4} \int_0^1 g_{\lambda\sigma_m}^{-1}(y) \left\{ \int_{-\infty}^{\infty} zg'(y - \lambda\sigma_m z)\phi(z)\, dz \right\}^2 d\lambda \\ &\le \frac{\sigma_m^2}{4} \int_0^1 \int_{-\infty}^{\infty} \frac{\{zg'(y - \lambda\sigma_m z)\}^2}{g(y - \lambda\sigma_m z)} \phi(z)\, dz\, d\lambda \end{aligned} \tag{6.22}$$

by convexity of $(u, v) \to u^2/v$. Substitute (6.22) in (6.20) and use $|q_m| \le c_m$ to bound (6.20)

by

$$\frac{c_m^2\sigma_m^2}{4}\int_0^1\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{g'^2}{g}(v)\,z^2\phi(z)\,dz\,dv\,d\lambda.$$

Since the integrals stay bounded independent of m, the result follows. □

Lemma 4.1 follows from Lemmas 6.1 and 6.3 since

$$\begin{aligned}
&\int\left\{q_m(y)-\frac{g'}{g}(y)\right\}^2 g(y)\,dy\\
&\le 3\Bigg[\int_{[g>0]}\left\{q_m(y)-q_m\left(\frac{g_{\sigma_m}}{g}\right)^{1/2}(y)\right\}^2 g(y)\,dy\\
&\quad+\int_{[g>0]}\left\{q_m\left(\frac{g_{\sigma_m}}{g}\right)^{1/2}(y)-\left(\frac{g'_{\sigma_m}}{g_{\sigma_m}}\right)\left(\frac{g_{\sigma_m}}{g}\right)^{1/2}(y)\right\}^2 g(y)\,dy\\
&\quad+\int_{[g>0]}\left\{\left(\frac{g'_{\sigma_m}}{g_{\sigma_m}}\right)\left(\frac{g_{\sigma_m}}{g}\right)^{1/2}(y)-\frac{g'}{g}(y)\right\}^2 g(y)\,dy\Bigg],
\end{aligned}\tag{6.23}$$

and the first term tends to 0 by Lemma 6.3, the second by Lemma 6.1, and the last by Lemma 6.2. □

6.2 *Consistency Proofs.*

(i) *Consistency of $\hat{I}_n$ in* (3.14). As usual, we can take $\bar{\theta}_n$ to be deterministic, and in view of (3.15) we need only check that

$$\Delta_n=\bar{n}^{-1}\textstyle\sum_{i=m+1}^n\{\dot{\ell}^T\dot{\ell}(X_i,\theta_n;X_1,\cdots,X_m)-\dot{\ell}^T\dot{\ell}(X_i,\theta_n,G)\}\to_{P_{\theta_n}}0\tag{6.24}$$

whenever $|\theta_n-\theta|=O(n^{-1/2})$. But by (3.11),

$$\begin{aligned}
&E_{\theta_n}\{|\Delta_n|\,|X_1,\cdots,X_m\}\\
&\le E\{|\hat{\dot{\ell}}^T\hat{\dot{\ell}}(X_{m+1},\theta_n;X_1;X_1,\cdots,X_m)-\dot{\ell}^T\dot{\ell}(X_{m+1},\theta_n,G)|\,X_1,\cdots,X_m\}\\
&=o_{P_{\theta_n}}(1)
\end{aligned}\tag{6.25}$$

and the result follows.

(ii) *Consistency in Theorem* 4.2. Again we can treat $\bar{\theta}_n$ as deterministic. Define measures $\{Q_n\}$ on $(R^{p+1})^n$ with densities

$$\textstyle\prod_{i=1}^m r(c_i)g(y_i-c_i\theta)^T\prod_{i=m+1}^n r(c_i)g(y_i-c_i(\theta-\bar{\theta}_n)^T).$$

We can argue as in the proof of (4.12) that the measures $\{Q_n\}$ are contiguous to the product measures specifying the distribution of the observations when θ is true. It follows that (4.28) is equivalent to

$$\bar{n}^{-1}\textstyle\sum_{i=m+1}^n q_m(\varepsilon_i;\hat{\varepsilon}_1,\cdots,\hat{\varepsilon}_m)=o_P(1).\tag{6.26}$$

By the usual calculation, conditioning on the first m observations,

$$\begin{aligned}
&E\left(\left[\bar{n}^{-1}\textstyle\sum_{i=m+1}^n\left\{q_m(\varepsilon_i;\hat{\varepsilon}_1,\cdots,\hat{\varepsilon}_m)-\frac{g'}{g}(\varepsilon_i)\right\}\right]^2\Bigg|\,\hat{\varepsilon}_1,\cdots,\hat{\varepsilon}_m\right)\\
&\quad=\int\left\{q_m(y;\hat{\varepsilon}_1,\cdots,\hat{\varepsilon}_m)-\frac{g'}{g}(y)\right\}^2 g(y)\,dy=o_P(1)
\end{aligned}$$

by (4.13) and we can substitute g'/g for q_m in (6.26). With this final substitution, (4.28) follows from the WLLN. □

6.3 *Identities of Example* 4.

Verification of (4.31) *and* (4.32). Write $\dot{\ell} = (\dot{\ell}_1, \dot{\ell}_2)$ where

$$\dot{\ell}_1 = \left(\frac{\partial \ell}{\partial \mu_1}, \cdots, \frac{\partial \ell}{\partial \mu_k}\right), \qquad \dot{\ell}_2 = \left\{\frac{\partial \ell}{\partial v_{ij}} : i \geq j\right\}.$$

Evidently

$$\begin{aligned} \dot{\ell}_1(x, \theta, G_0) &= -\,|(x-\mu)V^{1/2}|^{-1}\frac{\gamma_0'}{\gamma_0}(|(x-\mu)V^{1/2}|)(x-\mu)V \\ &= \dot{\ell}_1((x-\mu)V^{1/2}, 0, [J], G_0)V^{1/2}, \end{aligned} \tag{6.27}$$

$$\dot{\ell}_2(x, \theta, G_0) = \left\{\left(\frac{(x_i-\mu_i)(x_j-\mu_j)}{|(x-\mu)V^{1/2}|}\frac{\gamma_0'}{\gamma_0}(|(x-\mu)V^{1/2}|) - v^{ij}\right)\left(1-\frac{\delta_{ij}}{2}\right)\right\}, \tag{6.28}$$

where $V^{-1} = \| v^{ij} \|$ and $x = (x_1, \cdots, x_k)$.

Define a linear operator L_B on $R^{k(k+1)/2}$, corresponding to a $k \times k$ matrix $B = \| b_{ij} \|$, by the $\frac{k(k+1)}{2} \times \frac{k(k+1)}{2}$ matrix

$$L_B = \left\| (b_{ir}b_{sj} + b_{jr}b_{is})\left(1-\frac{\delta_{ij}}{2}\right) \right\|, \qquad r \geq s,\ i \geq j,$$

where (r, s) indexes rows and (i, j) columns. It is easy to verify that

$$\dot{\ell}_2(x, \theta, G_0) = \dot{\ell}_2((x-\mu)V^{1/2}, 0, [J], G_0)L_{V^{1/2}}. \tag{6.29}$$

By (6.27) and (6.29) we have

$$I(\theta, G_0) = \begin{pmatrix} V^{1/2} & 0 \\ 0 & L_{V^{-1/2}} \end{pmatrix}^T I(0, [J], G_0)\begin{pmatrix} V^{1/2} & 0 \\ 0 & L_{V^{-1/2}} \end{pmatrix} \tag{6.30}$$

and, finally,

$$\dot{\ell}(x, \theta, G_0)I^{-1}(\theta, G_0) = \dot{\ell}((x-\mu)V^{1/2}, 0, [J], G_0)I^{-1}(0, [J], G_0) \times \begin{pmatrix} V^{1/2} & 0 \\ 0 & L^T_{V^{-1/2}} \end{pmatrix}^{-1}$$

Since $V^{1/2}$ is symmetric, (4.31) follows. To get (4.32) it is enough to verify that

$$L_B^{-1} = L_{B^{-1}} \quad \text{for any} \quad B, \tag{6.31}$$

and that if x is a triangular array

$$xL_B^T = [BQ(x)B^T], \tag{6.32}$$

where $Q(x)$ is the symmetric matrix whose ij-th entry is x_{ij} if $i \geq j$, or x_{ji} if $i < j$. The verifications of (6.31) and (6.32) are straightforward exercises in matrix multiplication.

Verification of (4.33) *and* (4.34). In this case $V^{1/2} = J$. For convenience suppress $(0, [J], G_0)$ in the arguments of functions for this discussion. We have

$$\dot{\ell}_1(x) = -\frac{x}{|x|}\frac{\gamma_0'}{\gamma_0}(|x|), \tag{6.33}$$

$$E\dot{\ell}_1^T\dot{\ell}_1(X) = E\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\frac{X^TX}{|X|^2} = \frac{1}{k}\left\{E\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\}J \tag{6.34}$$

by symmetry. Next, note that

$$\dot{\ell}_2(x) = \left\{\left(\frac{x_ix_j}{|x|}\frac{\gamma_0'}{\gamma_0}(|x|) - \delta_{ij}\right)(1-\delta_{ij}/2)\right\}_{i \geq j} \tag{6.35}$$

and by symmetry

$$E\dot{\ell}_2^T\dot{\ell}_1(X) = 0 \tag{6.36}$$

$$E\dot{\ell}_2^T\dot{\ell}_2(X) = \|a_{rs,ij}\|_{r\geq s,i\geq j}, \tag{6.37}$$

where $X = (X_1, \cdots, X_k)$

$$\begin{aligned} a_{rs,ij} &= 0, \quad \text{unless } r = i,\ s = j, \\ a_{rs,rs} &= E\left\{\frac{X_1^2X_2^2}{|X|^2}\left(\frac{\gamma_0'}{\gamma_0}\right)(|X|)\right\}, \quad r \neq s, \\ a_{rr,rr} &= E\left\{\frac{X_1^2}{|X|}\frac{\gamma_0'}{\gamma_0}(|X|) + 1\right\}^2. \end{aligned} \tag{6.38}$$

$$E\left\{\frac{X_1^2X_2^2}{|X|^2}\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\} = E\left\{\frac{X_1^2X_2^2}{|X|^4}\right\}E\left\{|X|^2\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\} \tag{6.39}$$

by spherical symmetry of G_0. The second term in (6.39) is just $I_2(G_0)$, while the first term is independent of G_0 and may be shown to equal $k^{-1}(k+2)^{-1}$ by taking G_0 to be the spherical normal distribution. Thus

$$a_{rs,rs} = k^{-1}(k+2)^{-1}I_2(G_0), \quad r \neq s. \tag{6.40}$$

A similar computation gives

$$a_{rr,rr} = \frac{1}{4}E\left\{\frac{X_1^4}{|X|^2}\left(\frac{\gamma_0'}{\gamma_0}\right)^2(|X|)\right\} - 1 = \frac{1}{4}3k^{-1}(k+2)^{-1}I_2(G_0) - 1. \tag{6.41}$$

We see from (6.37) that $I(0, [J], G_0)$ is a diagonal matrix with entries given by (6.40) and (6.41). Upon inverting it and substituting (6.40) and (6.41) in $\dot{\ell}(x, 0, [J], G_0)$, we obtain (4.33) and (4.34).

6.4 *Two Theorems on efficient estimates.*

THEOREM 6.1. *Under R suppose* $\{\hat{\theta}_n\}$ *are such that, for a given* $\theta, \mathscr{L}_{\theta_n}\{n^{1/2}(\hat{\theta}_n - \theta_n)\} \to \mathscr{N}(0, I^{-1}(\theta))$ *whenever* $n^{1/2}|\theta_n - \theta| \leq M$ *for all* n, $M < \infty$. *Then,*

$$n^{1/2}(\hat{\theta}_n - \theta) = n^{-1/2}\textstyle\sum_{i=1}^n \dot{\ell}(X_i, \theta)I^{-1}(\theta) + o_{P_\theta}(1). \tag{6.42}$$

NOTE. This claim is in fact valid in great generality if the local asymptotic normality (LAN) condition of Hájek (1972) holds with $\Delta_n(\theta)$ replacing $n^{-1/2}\sum_{i=1}^n \dot{\ell}(X_i, \theta)$. Moreover it is clear that everything is local so that the condition and conclusion need only hold at a point θ on which $\hat{\theta}_n$ can depend.

PROOF. Since the sequence of joint laws $\mathscr{L}_n$ of $n^{1/2}(\hat{\theta}_n - \theta)$ and $n^{-1/2}\sum_{i=1}^n \dot{\ell}(X_i, \theta)I^{-1}(\theta)$ is tight under P_θ it is enough to show that if $\mathscr{L}_{m_n}$ is any subsequence weakly convergent to $\mathscr{L}^*$ (say) then $\mathscr{L}^*$ must concentrate on the diagonal. by a contiguity and analyticity argument, see Roussas (1972, pages 136–141), we can show that the joint characteristic function $\phi^*(u, v)$ of $\mathscr{L}^*$ satisfies the equation

$$\phi^*(u, v) = \phi^*(u, 0)\exp\{-uI^{-1}(\theta)v^T\}\exp\{-\tfrac{1}{2}vI^{-1}(\theta)v^T\}$$

(Substitute $\Gamma = I(\theta)$, $h = vI^{-1}(\theta)$ in (3.11) of Roussas.) But, by hypothesis,

$$\phi^*(u, 0) = \exp\{-\tfrac{1}{2}uI^{-1}(\theta)u^T\}$$

so that

$$\phi^*(u, v) = \exp\{-\tfrac{1}{2}(u+v)I^{-1}(\theta)(u+v)^T\},$$

and the theorem follows. □

THEOREM 6.2. *If* R(i), R(ii) *and* UR(iii) *hold and if* $\bar{\theta}_n$ *is* $\sqrt{n}$*-consistent and discretized as in* (2.3) *and*

$$\hat{\theta}_n = \bar{\theta}_n + n^{-1} \sum_{j=1}^n \dot{\ell}(X_j, \bar{\theta}_n) I^-(\bar{\theta}_n),$$

then $\hat{\theta}_n$ *is efficient in the usual sense.*

PROOF. In view of the arguments leading to Theorem 4 of Le Cam (1968), it is enough to show that for θ regular and any sequence θ_n such that $n^{1/2} | \theta_n - \theta | \leq M$ for all n

$$n^{-1/2} \sum_{i=1}^n \{\dot{\ell}(X_i, \theta_n) - \dot{\ell}(X_i, \theta)\} + n^{1/2}(\theta_n - \theta) I(\theta) = o_{P_\theta}(1). \tag{6.43}$$

We claim that (6.43) is implied by the fact that

$$\sum_{i=1}^n \{\ell(X_i, \theta_n + hn^{-1/2}) - \ell(X_i, \theta_n)\} = hn^{-1/2} \sum_{i=1}^n \dot{\ell}(X_i, \theta_n) - \tfrac{1}{2} h I(\theta_n) h^T + o_{P_\theta}(1) \tag{6.44}$$

for all h. To see this, note that from the usual LAN condition

$$\sum_{i=1}^n \{\ell(X_i, \theta_n + hn^{-1/2}) - \ell(X_i, \theta)\} = n^{1/2}(\theta_n - \theta) + hn^{-1/2} \sum_{i=1}^n \ell(X_i, \theta) - \tfrac{1}{2}\{n^{1/2}(\theta_n - \theta) + h\} I(\theta) \{n^{1/2}(\theta_n - \theta) + h\}^T + o_{P_\theta}(1); \tag{6.45}$$

$$\sum_{i=1}^n \{\ell(X_i, \theta_n) - \ell(X_i, \theta)\} = n^{1/2}(\theta_n - \theta) n^{-1/2} \sum_{i=1}^n \dot{\ell}(X_i, \theta) - \frac{n}{2} \{(\theta_n - \theta) I(\theta) (\theta_n - \theta)^T\} + o_{P_\theta}(1). \tag{6.46}$$

Subtracting (6.46) from (6.45) and matching the coefficient of h in (6.44) yields (6.43).

Finally, (6.44) is just the usual statement of LAN with θ replaced by θ_n. It is argued in exactly the same way as the usual equivalence,—see pages 54–63 of Roussas (1972) for example,—but, of course, we use the uniformity in UR(iii). The theorem follows. □

Acknowledgement V. Fabian and J. Hannan corrected my mistaken impression that R(i) – R(iii) were sufficient to establish the efficiency of $\bar{\theta}_n + n^{-1/2} \sum \dot{\ell}(X_i, \bar{\theta}_n)$. I am grateful to them for prompting me to prove Theorems 6.1 and 6.2 as well as other valuable comments. I am also grateful to Chris A. J. Klaassen for a careful reading of the paper resulting in several substantial corrections and to J. Pfanzagl for the Ibragimov-Hasminskii reference.

REFERENCES

BERAN, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.* **2** 63–74.

BERAN, R. (1975). Adaptive estimates for autoregressive processes. *Ann. Inst. Statist. Math.* **28** 77–89.

BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313.

BICKEL, P. J. (1981). Lectures on robustness and adaptation, to appear in 1979 *St. Flour Conference Lecture Notes in Mathematics* 876 Springer, Berlin.

BIRGÉ, L. (1980). Approximation dans les espaces métrique et théorie de l'estimation. Unpublished thesis, University of Paris.

DIONNE, L. (1981). Efficient nonparametric estimators of parameters in the general linear hypothesis. *Ann. Statist.* **9** 457–460.

FABIAN V. and HANNAN J. (1980). On estimation and adaptive estimation for locally asymptotically normal families. *Z. Wahrsch. verw. Gebiete.* To appear.

HÁJEK, J. (1962). Asymptotically most powerful rank order tests. *Ann. Math. Statist.* **33** 1124–1147.

HÁJEK, J. and SIDÁK, Z. (1967). *Theory of Rank Tests*, Academic, New York and Academia, Prague.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **1** 175–194. University of California Press, Berkeley.

HAMMERSTROM, T. (1978). Ph.D. Thesis, University of California, Berkeley.

HASMINSKII, P. Z. and IBRAGIMOV I. A. (1978). On the nonparametric estimation of functionals. *Proc. Second Prague Symp. Asympt. Statistics and Probability*, J. Jureçkova Ed., Prague.

HOGG, R. (1980). On adaptive robust inference. Tech. Report, Univ. of Iowa.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings Fifth Berkeley Symp. Math. Statist. Prob.* **1** 221–233 University of California Press, Berkeley.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Math. Statist.* **1** 799–821.

HUBER, P. J. (1977). Robust covariances *Statistical Decision Theory and Related Topics*, II. S. S. Gupta and D. S. Moore, eds. 165–192 Academic, New York.

KLAASSEN, C. (1980). Statistical performance of location estimators. Thesis, University of Leiden, Netherlands.

LE CAM, L. (1969). *Théorie Asymptotique de la Décision Statistique*, Les Presses de l'Univérsite de Montreal.

LEVITT, B. (1974). On optimality of some statistical estimates, *Proc. Prague Symp. Asympt. Statist.* 215–238 J. Jureckova Ed. Prague.

LINDSAY, B. (1978). Information in the presence of nuisance parameters. Thesis, University of Washington, Seattle.

LINDSAY, B. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood. *Phil. Trans. Roy. Soc. London* **296** 639–665.

MARONNA, R. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

ROUSSAS, G. (1972). *Contiguity of Probability Measures: Applications in Statistics.* Cambridge University Press.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 187–196. University of California Press.

STONE, C. (1975). Adaptive maximum likelihood estimation of a location parameter. *Ann. Statist.* **3** 267–284.

TAKEUCHI, K. (1971). A uniformly asymptotically efficient estimator of a location parameter, *J. Amer. Statist. Assoc.* **66** 292–301.

VAN EEDEN, C. (1970). Efficiency-robust estimation of location. *Ann. Math. Statist.* **41** 172–181.

WEISS, L. and WOLFOWITZ, J. (1970). Asymptotically efficient nonparametric estimators of location and scale parameters. *Z. Wahrsch. verw. Gebiete* **16** 134–150.

WEISS, L. and WOLFOWITZ, J. (1971). Asymptotically efficient estimation of nonparametric regression coefficients, *Statistical Decision Theory and Related Topics*, S. Gupta and J. Yackel, 29–40 eds. Academic, New York.

WOLFOWITZ, J. (1953). The method of maximum likelihood and the Wald theory of decision functions. *Indag. Mathemat.* **56** 114–119.

WOLFOWITZ, J. (1974). Asymptotically efficient nonparametric estimators of location and scale parameters. II. *Z. Wahrsch. verw. Gebiete* **30** 117–128.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

ADVANCES IN APPLIED MATHEMATICS 7, 55–69 (1986)

Empirical Bayes Estimation in Functional and Structural Models, and Uniformly Adaptive Estimation of Location

P. J. BICKEL*

Department of Statistics, University of California, Berkeley, California 94720

AND

C. A. J. KLAASSEN*,†

Department of Mathematics, University of Leiden, Postbus 9512, 2300 RA Leiden, Netherlands

DEDICATED TO HERBERT ROBBINS ON THE OCCASION OF HIS 70TH BIRTHDAY

We discuss estimation of parameters in functional and structural models in relation to Robbins' empirical Bayes and compound decision theories. We construct an efficient estimate of ν in the normal functional model, X_i independent $\mathcal{N}(\nu, \theta_i)$ where $\varepsilon \le \theta_i^2 \le 1/\varepsilon, \varepsilon > 0, 1 \le i \le n$.

1. INTRODUCTION

In 1956, Robbins [15] (see also Good [4]) initiated the systematic study of nonparametric empirical Bayes procedures. Robbins [16] is a good entry to the large literature. The focus of his work and that of its many successors has been the model:

I: We observe random variables or vectors $X_1, \ldots, X_n$ i.i.d. F where F ranges over all (or most) mixtures of a parametric family $\{F_\theta : \theta \in \Theta\}$ with

*Research partially supported by ONR Contract N00014-80-C-0163.

†Research carried out in part with the support of the Mathematical Sciences Research Institute (Berkeley).

0196-8858/86 $7.50

$\Theta \subset R^p$. That is,

$$F = \int F_\theta dG(\theta)$$

for some probability G on Θ, belonging to a set $\mathcal{G}$. Equivalently, we observe X_i, $1 \le i \le n$ where (θ_i, X_i) are i.i.d. with $\theta_i \sim G$ and given θ_i, $X_i \sim F_{\theta_i}$. Work in the area has focused on questions such as simultaneous estimation of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ with squared error loss, $L(\boldsymbol{\theta}, \mathbf{d}) = n^{-1}\sum_{i=1}^n (\theta_i - d_i)^2$, $\mathbf{d} = (d_1, \ldots, d_n)^T$, and the possibility of constructing decision rules

$$\delta^*(\mathbf{X}) = \left(h^*(X_1; \mathbf{X}), \ldots, h^*(X_n; \mathbf{X})\right)^T, \qquad \mathbf{X} = (X_1, \ldots, X_n)^T, \quad (1.1)$$

which to first order approximate the Bayes rule,

$$\delta(\mathbf{X}, G) = \left(h(X_1, G), \ldots, h(X_n, G)\right)^T$$

where

$$h(X, G) = E(\theta | X), \qquad (\theta, X) \sim (\theta_1, X_1).$$

Robbins came to the empirical Bayes formulation from his 1951 consideration of the compound decision problem [14].

II: Observe X_i independent with $X_i \sim F_{\theta_i}$, $\theta_i \in K$ compact $\subset \Theta$, for $1 \le i \le n$. A typical problem now is to simultaneously estimate $\theta_1, \ldots, \theta_n$ as well as possible, asymptotically, i.e., to find $\delta_n^*(X_1, \ldots, X_n) = (h_{1n}^*(\mathbf{X}), \ldots, h_{nn}^*(\mathbf{X}))^T$ such that

$$\liminf_n n^{-1} \sum_{i=1}^n \left\{ E_{\theta_i}\left(h_{in}(\mathbf{X}) - \theta_i\right)^2 - E_{\theta_i}\left(h_{in}^*(\mathbf{X}) - \theta_i\right)^2 \right\} \le 0 \quad (1.2)$$

for any competing sequence $\delta_n(\mathbf{X}) = (h_{1n}(\mathbf{X}), \ldots, h_{nn}(\mathbf{X}))^T$. The solution, heuristically, is to use δ^* given by (1.1) since the risks in (1.2) should be close to model I risks when $G = G_n$ is the empirical distribution of $\theta_1, \ldots, \theta_n$.

A key element in the transition from I to II evidently lies in establishing that the approximation of $\delta(\mathbf{X}, G)$ by $\delta^*(\mathbf{X})$ is in a suitable sense, uniform in G.

An analogous set of questions was investigated by Neyman and Scott [12], Kiefer and Wolfowitz [8], and notably, recently Lindsay [10] and others. Their focus is on estimating a parameter ν common to the X_i in the

presence of random (structural models) or fixed (functional models) nuisance parameters $\theta_1, \ldots, \theta_n$. The corresponding models are:

I′: (Structural) $X_1, \ldots, X_n$ i.i.d. F where

$$F = \int F_{(\nu, \theta)} dG(\theta) \tag{1.3}$$

$G \in \mathcal{G}$, $\nu \in H$ open $\subset R^m$.

II′: (Functional) X_i independent with $X_i \sim F_{(\nu, \theta_i)}$, $\theta_i \in K \subset \Theta$, K compact, $1 \leq i \leq n$.

Again $\{F_{(\nu, \theta)} : \nu \in H, \theta \in \Theta\}$ is a postulated parametric model.

In various examples discussed by these authors it is clear that ν can be estimated at rate $n^{-1/2}$. For instance, if $F_{(\nu, \theta)}$ is the $\mathcal{N}(\nu, \theta^2)$ distribution, $\bar{X}$ is a $n^{1/2}$ consistent estimate of ν in model I′ if $\int \theta^2 dG(\theta) < \infty$ and in II′ if the empirical second moment of θ, $n^{-1}\sum_{i=1}^n \theta_i^2$ is bounded. What are optimal procedures in this context? For simplicity take $m = 1$.

Let $F_{(\nu, G)}$ denote the distribution (1.3), $P_{(\nu, G)}$ the associated probability measure, etc. Call a (sequence of) estimate(s) *regular* (I′) if

$$\mathcal{L}_{(\nu_n, G_n)}\left(n^{1/2}(T_n - \nu_n)\right) \to \mathcal{N}\left(0, \sigma_T^2(\nu_0, G_0)\right) \tag{1.4}$$

whenever $\nu_n \to \nu_0$ and $G_n \to G_0$ (weakly) for all $\nu_0 \in H, G_0 \in \mathcal{G}$. Call T_n^* *efficient* (I′) if $\{T_n^*\}$ is regular (I′) and

$$\sigma_T^{2*}(\nu_0, G_0) \leq \sigma_T^2(\nu_0, G_0) \tag{1.5}$$

for all regular $\{T_n\}$, (ν_0, G_0).

In model I′ let $\mathcal{G}$ be the set of all probability distributions on K. Call an estimate *regular* (II′) if

(i) $T_n(x_1, \ldots, x_n)$ is symmetric in $(x_1, \ldots, x_n)$

(ii) $\mathcal{L}_{(\nu_n, \theta_1, \ldots, \theta_n)}(n^{1/2}(T_n - \nu_n)) \to \mathcal{N}(0, \sigma_T^2(\nu_0, G_0))$

whenever $\nu_n \to \nu_0$ and G_n, the empirical distribution, $n^{-1}\sum_{i=1}^n I(\theta_i \leq \cdot)$, of $\{\theta_1, \ldots, \theta_n\}$, tends (weakly) to $G_0 \in \mathcal{G}$. An estimate T_n^* is *efficient* (II′) if it is regular (II′) and satisfies (1.5) for regular (II′) competitors T_n.

In problem I′ sufficiency of the order statistics permits us to restrict to symmetric estimates. In problem II′ invariance of the problem under permutations of the θ_i leads less forcefully to the same conclusion. The passage from efficiency (I′) to efficiency (II′) is as in Robbins' problems a question of uniformity.

Evidently,

PROPOSITION 1.1. *Suppose $\mathcal{G}$ for both models is the set of all distributions on K.*

(i) *if T_n is regular* (II′) *it is regular* (I′)

(ii) *If T_n^* is efficient* (I′) *and regular* (II′) *then T_n^* is efficient* (II′).

An extension of the theory of information (Cramér–Rao) bounds to models with infinite dimensional nuisance parameters such as I′ has been developed by Koshevnik and Levit [9], Pfanzagl [13], and Begun *et al.* [1] on the basis of a fundamental paper of Stein [17]. Under regularity conditions, efficient (I′) estimates are regular (I′) estimates achieving these information bounds. Methods for constructing such estimates in a general context are discussed in [13, 2, 3] among others. We do not study the general situation further but show in an important special case how to construct estimates which are not only efficient (I′) but also regular (II′) and hence efficient (II′).

The example we consider and extend somewhat is the normal location problem with variances possibly changing from observation to observation.

$$F_{(\nu,\theta)} = \mathcal{N}(\nu, \theta^2) \tag{1.6}$$

with $\Theta = R^+$. Take $K = [\varepsilon, 1/\varepsilon]$ for fixed $\varepsilon > 0$ and $\mathcal{G}$, all distributions on K. Then $F_{(\nu,G)}$ is still a symmetric location family in ν. If G is known, efficient estimates are asymptotically $\mathcal{N}(\nu, I^{-1}(H)/n)$ where $H = \int F_{(0,\theta)} dG(\theta)$,

$$I(H) = \int \frac{[h']^2}{h}(t)\, dt$$

$$h(t) = \int_0^\infty \theta^{-1}\varphi(t\theta^{-1})\, dG(\theta).$$

The general information bound theory indicates that it should be possible to adapt perfectly in this case, i.e., do as well not knowing G as knowing it. In fact, Stone [18] constructs an estimate $\hat{\nu}_n$ which is location and scale equivariant and such that,

$$\mathcal{L}_F\big(n^{1/2}(\hat{\nu}_n - \nu)\big) \to \mathcal{N}\big(0, I^{-1}(H)\big) \tag{1.7}$$

whenever $X_1, \ldots, X_n$ are i.i.d. F and

$$F(\cdot) = F(\cdot + \nu) \text{ is symmetric about } 0. \tag{1.8}$$

Here, we define generally for H on $[-\infty, \infty] = \bar{R}$, $H(R) > 0$

$$I(H) = \int_R \frac{[h']^2}{h}(t)\,dt \qquad \text{if } H \text{ has an absolutely continuous density } h \text{ on } R,$$
$$= \infty \qquad \text{otherwise.} \tag{1.9}$$

For convenience, in the sequel, distribution functions are defined by capital letters and their densities, by convention, are the corresponding lower case letters. In Section 2 of this paper we construct a modified and simplified translation but not scale equivariant version of Stone's estimate, ν_n^*, which satisfies (1.7) and is also regular (II') for the model (1.6). In fact, we show for the symmetric location model,

THEOREM 1.1. *$\mathcal{L}_{H_n}(n^{1/2}\nu_n^*)$ converges to $\mathcal{N}(0, I^{-1}(H_0))$ whenever*

(a) $H_n \xrightarrow{w} H_0$, $H_0(R) = 1$

(b) $I(H_n) \to I(H_0) < \infty$.

Then, in Theorem 2.1, we show that uniformity of convergence persists in a generalization of model II'.

Theorem 1.1 is the best that one can hope for in adaptive estimation of location since $\mathcal{L}_{H_m}(n^{1/2}\hat{\nu}_n) \to \mathcal{N}(0, I^{-1}(H_m))$ as $n \to \infty$, uniformly in m, $H_m \xrightarrow{w} H_0$ as $m \to \infty$, and $\sup_m I(H_m) < \infty$ imply that $I(H_m) \to I(H_0)$.

This estimate is also asymptotically minimax in Huber's [6] sense and can be used for the construction of an adaptive confidence interval, $\nu_n^* \pm z(nI_n^*)^{-1/2}$ where, $\inf_{\mathcal{H}} P_H[\nu_n^* - z(nI_n^*)^{-1/2} \le \nu \le \nu_n^* + z(nI_n^*)^{-1/2}] \to 2\Phi(z) - 1$ for any family $\mathcal{H}$ of distributions symmetric about 0 which does not have point mass at $\pm\infty$ as a weak limit point. The details of these results and other robustness properties of $\{\nu_n^*\}$ will appear in Bickel *et al.* [3].

2. THE RESULTS

Suppose the common distribution of $X_1, \ldots, X_n$ i.i.d. is H as in (1.8), with $H \in \mathcal{H}$. Suppose $\mathcal{H}$ does not have point mass at $\pm\infty$ as a weak limit point. Then there exist uniformly $n^{1/2}$ consistent translation equivariant estimates $\tilde{\nu}_n$ of ν, such that,

$$\mathcal{L}_H\big(n^{1/2}(\tilde{\nu}_n - \nu)\big) \to \mathcal{N}\big(0, \sigma^2(H)\big) \tag{2.1}$$

uniformly on $\mathscr{H}$ and $\sup_{\mathscr{H}} \sigma^2(H) < \infty$. For instance let

$$k(t) = \frac{e^{-t}}{(1 + e^{-t})^2} \tag{2.2}$$

be the logistic density. If $\tilde{\nu}_n$ is the unique solution of

$$\sum_{i=1}^{n} \frac{k'}{k}(X_i - \nu) = 0$$

then it is easy to see that $\tilde{\nu}_n$ satisfies (2.1).

To define ν_n^* we proceed as in Stone [18], but use the logistic rather than the normal kernel for smoothing. Let

$$k_\sigma(x) = \frac{1}{\sigma} k\left(\frac{x}{\sigma}\right).$$

If $\hat{F}_n$ is the empirical d.f. of $X_1, \ldots, X_n$, define

$$\hat{h}_\sigma(x) = \int k_\sigma(x - z) d\hat{F}_n(z) = \frac{1}{n} \sum_{i=1}^{n} k_\sigma(x - X_i).$$

Next let

$$\hat{q}_\sigma(x) = \frac{\hat{h}'_\sigma}{\hat{h}_\sigma}(x),$$

$$\bar{q}_\sigma(x, \nu) = \frac{1}{2}\left[\hat{q}_\sigma(x + \nu) - \hat{q}_\sigma(-x + \nu)\right].$$

Let ψ be symmetric and continuous at 0 with support $[-1, 1]$, $0 \le \psi \le 1$ and $\psi(0) = 1$. Let

$$\psi_n(x) = \psi(c_n x)$$

and $\sigma_n \downarrow 0, c_n \downarrow 0$ at a rate to be determined later. Write $\hat{q}_n, \bar{q}_n, \hat{h}_n$ for $\hat{q}_{\sigma_n}, \bar{q}_{\sigma_n}, \hat{h}_{\sigma_n}$. Then we define

$$\nu_n^*(\nu) = \nu - \hat{I}_n^{-1}(\nu) \int \bar{q}_n(x, \nu) \psi_n(x) \hat{h}_n(x + \nu)\, dx$$

where

$$\hat{I}_n(\nu) = \int \bar{q}_n^2(x, \nu) \psi_n(x) \hat{h}_n(x + \nu)\, dx.$$

Finally our estimate is

$$\nu_n^* = \nu_n^*(\tilde{\nu}_n).$$

Since we have selected $\tilde{\nu}_n$ to be translation equivariant, the second term of ν_n^* is translation invariant and ν_n^* itself is translation equivariant, and therefore we may and do assume that the true value of $\nu = 0$, i.e., that H_n is the common distribution of the X_i. We then define the density and score function of the convolution $\tilde{H}_n$ of H_n with the logistic distribution with mean 0 and variance σ_n^2

$$\tilde{h}_n(x) = \int k_{\sigma_n}(x - z)h_n(z)\,dz$$

$$\tilde{q}_n(x) = \frac{\tilde{h}_n'}{\tilde{h}_n}(x).$$

Then,

$$\tilde{h}_n(x) = E\hat{h}_n(x) \tag{2.3}$$

and $\hat{I}_n(\nu_n^*)$ estimates the quantity

$$I_n(\tilde{H}_n) = \int \tilde{q}_n^2(x)\psi_n(x)\tilde{h}_n(x)\,dx.$$

We prove Theorem 1.1 by a series of lemmas. The proof is somewhat simpler than our original thanks to an idea of J. Ritov. Uniformly for $H_n \in \mathscr{H}$,

LEMMA 2.1. *Write* $\bar{q}_n(x)$ *for* $\bar{q}_n(x, 0)$ *etc. Then,*

$$\int [\bar{q}_n(x, \nu)\psi_n(x)\hat{h}_n(x + \nu) - \bar{q}_n(x)\psi_n(x)\hat{h}_n(x)]\,dx$$

$$-\nu\int \bar{q}_n(x)\psi_n(x)\hat{h}_n'(x)\,dx$$

$$= 0_p\left(\nu\int(\hat{q}_n'(x) - \hat{q}_n'(-x))\psi_n(x)\hat{h}_n(x)\,dx\right) + 0_p(\sigma_n^{-3}\nu^2) \tag{2.4}$$

$$\hat{I}_n(\nu) = \hat{I}_n + 0_p(\sigma_n^{-3}\nu). \tag{2.5}$$

LEMMA 2.2.

$$\int(\hat{q}_n'(x) - \hat{q}_n'(-x))\psi_n(x)\hat{h}_n(x)\,dx = 0_p(\sigma_n^{-3}c_n^{-1}n^{-1}) \tag{2.6}$$

$$\int \bar{q}_n(x)\psi_n(x)\hat{h}_n'(x)\,dx = \hat{I}_n + 0_p(\sigma_n^{-3}c_n^{-1}n^{-1}). \tag{2.7}$$

LEMMA 2.3.

$$\int \bar{q}_n(x)\psi_n(x)\hat{h}_n(x)\,dx = \int \tilde{q}_n(x)\psi_n(x)\hat{h}_n(x)\,dx + 0_p\left(n^{-1}c_n^{-1}\sigma_n^{-2}\right) \tag{2.8}$$

$$\hat{I}_n = I_n(\tilde{H}_n) + 0_p\left(\sigma_n^{-5/2}c_n^{-1/2}n^{-1/2}\right) \tag{2.9}$$

LEMMA 2.4. *If* $c_n = \sigma_n$, $n\sigma_n^6 \to \infty$, *and* $\sup_n I(H_n) < \infty$, *then*

$$n^{1/2}\nu_n^* = -n^{-1/2}I_n^{-1}(\tilde{H}_n)\sum_{i=1}^{n}\int \tilde{q}_n(x)\psi_n(x)k_{\sigma_n}(x - X_i)\,dx + o_p(1). \tag{2.10}$$

LEMMA 2.5. *If* $H_n \xrightarrow{w} H_0$ *and* $\sup_n I(H_n) < \infty$,

$$\liminf_n I_n(H_n) \geq I(H_0) \tag{2.11}$$

and

$$\liminf_n I(H_n) \geq \liminf_n I_n(\tilde{H}_n) \geq I(H_0). \tag{2.12}$$

If also $I(H_n) \to I(H_0) < \infty$, *then*

$$\int \left(h_n^{-1/2}h_n' - h_0^{-1/2}h_0'\right)^2(x)\,dx \to 0. \tag{2.13}$$

LEMMA 2.6. *If* $H_n \xrightarrow{w} H_0$, $H_0(R) = 1$, *and* $I(H_n) \to I(H_0) < \infty$, *the family of product measures* $Q_{n,\theta}$ *with density* $\{\pi_{i=1}^n h_n(x_i - \theta/n^{1/2})\}$ *satisfies Le Cam's L.A.N. condition and*

$$\log\frac{dQ_{n,\theta}}{dQ_{n,0}} = \theta n^{-1/2}\sum_{i=1}^{n}\frac{h_n'}{h_n}(X_i) - \frac{1}{2}\theta^2 I(H_n) + o_p(1) \tag{2.14}$$

and

$$\mathcal{L}_{H_n}\left(n^{-1/2}\sum_{i=1}^{n}\frac{h_n'}{h_n}(X_i)\right) \to \mathcal{N}(0, I(H_0)).$$

Proof of Lemma 2.1. Taylor expand, about $\nu = 0$, to find that (2.4) equals

$$\frac{\nu}{2}\int\left(\hat{q}_n'(x) - \hat{q}_n'(-x)\right)\psi_n(x)\hat{h}_n(x)\,dx$$

$$+\nu^2\int\int_0^1(1-\lambda)\left[\frac{\partial^2}{\partial\mu^2}\left(\bar{q}_n(x,\mu)\hat{h}_n(x+\mu)\right)\Big|_{\mu=\lambda\nu}\right]\psi_n(x)\,d\lambda\,dx.$$

Note that if $\|\cdot\|$ is the sup norm,

$$\left\|\frac{\hat{h}_n^{(r)}}{\hat{h}_n}\right\| = 0_p(\sigma_n^{-r}), \qquad \left\|\frac{\tilde{h}_n^{(r)}}{\tilde{h}_n}\right\| = 0_p(\sigma_n^{-r}),$$

since there exist finite constants C_r with

$$\left|\int k^{(r)}(x)\,dG(x)\right| \le C_r \int k(x)\,dG(x) \qquad \text{for all } r, G. \tag{2.15}$$

Hence,

$$\begin{aligned}\left\|\frac{\partial^r \overline{q}_n}{\partial \nu^r}(\cdot, \nu)\right\| &= 0_p\left(\sum_{s=1}^{r+1}\left\|\frac{\hat{h}_n^{(s)}}{\hat{h}_n}\right\|^{r+1/s}\right)\\ &= 0_p\left(\sigma_n^{-(r+1)}\right)\end{aligned}$$

and (2.4) follows. A similar argument yields (2.5).

Proof of Lemma 2.2. Write, using symmetry,

$$\begin{aligned}&\int (\hat{q}_n'(x) - \hat{q}_n'(-x))\psi_n \hat{h}_n(x)\,dx\\ &\quad = \int (\hat{q}_n'(x) - \hat{q}_n'(-x))\psi_n(\hat{h}_n - \tilde{h}_n)(x)\,dx\\ &\quad = 0_p\left(\left[\int (\hat{q}_n' - \tilde{q}_n')^2 \psi_n \tilde{h}_n(x)\,dx\right]^{1/2}\right.\\ &\qquad \left.\times\left(\int \frac{(\hat{h}_n - \tilde{h}_n)^2}{\tilde{h}_n}\psi_n(x)\,dx\right)^{1/2}\right).\end{aligned} \tag{2.16}$$

By (2.15),

$$E\left(\hat{h}_n^{(r)} - \tilde{h}_n^{(r)}\right)^2(x) \le \tfrac{1}{4}C_r^2 n^{-1}\sigma_n^{-(2r+1)}\tilde{h}_n(x) \tag{2.17}$$

and consequently

$$\int \left(\hat{h}_n^{(r)} - \tilde{h}_n^{(r)}\right)^2 \tilde{h}_n^{-1}\psi_n(x)\,dx = 0_p\left(n^{-1}\sigma_n^{-(2r+1)}c_n^{-1}\right). \tag{2.18}$$

Next write

$$(\hat{q}_n' - \tilde{q}_n')^2 \le 2\left\{\left(\frac{\hat{h}_n''}{\hat{h}_n} - \frac{\tilde{h}_n''}{\tilde{h}_n}\right)^2 + (\hat{q}_n^2 - \tilde{q}_n^2)^2\right\} \tag{2.19}$$

$$\frac{\hat{h}_n''}{\hat{h}_n} - \frac{\tilde{h}_n''}{\tilde{h}_n} = \frac{\hat{h}_n''}{\hat{h}_n}\left(\frac{\tilde{h}_n - \hat{h}_n}{\tilde{h}_n}\right) + \tilde{h}_n^{-1}(\hat{h}_n'' - \tilde{h}_n'') \tag{2.20}$$

$$\begin{aligned} |\hat{q}_n^2 - \tilde{q}_n^2| &= |\hat{q}_n + \tilde{q}_n||\hat{q}_n - \tilde{q}_n| \\ &\le 2\sigma_n^{-1}\left|\frac{\hat{h}_n'}{\hat{h}_n}\tilde{h}_n^{-1}(\tilde{h}_n - \hat{h}_n) + \tilde{h}_n^{-1}(\hat{h}_n' - \tilde{h}_n')\right|. \end{aligned} \tag{2.21}$$

Using (2.19)–(2.21) and (2.15) we get

$$\begin{aligned} &\int (\hat{q}_n' - \tilde{q}_n')^2 \psi_n \tilde{h}_n(x)\,dx \\ &\quad = 0_p\Big(\sigma_n^{-4}\int(\tilde{h}_n - \hat{h}_n)^2\tilde{h}_n^{-1}\psi_n(x)\,dx + \int(\tilde{h}_n'' - \hat{h}_n'')^2\tilde{h}_n^{-1}\psi_n(x)\,dx \\ &\qquad\qquad + \sigma_n^{-2}\int(\hat{h}_n' - \tilde{h}_n')^2\tilde{h}_n^{-1}\psi_n(x)\,dx\Big) \\ &\quad = 0_p\big(\sigma_n^{-5}c_n^{-1}n^{-1}\big) \end{aligned}$$

by (2.18). From this, (2.16), and (2.18), we obtain (2.6). Similarly,

$$\begin{aligned} \int \bar{q}_n\psi_n\hat{h}_n'(x)\,dx - \hat{I}_n &= \frac{1}{4}\int(\hat{q}_n^2(x) - \hat{q}_n^2(-x))\psi_n\hat{h}_n(x)\,dx \\ &= \frac{1}{4}\int(\hat{q}_n^2(x) - \hat{q}_n^2(-x))\psi_n(\hat{h}_n - \tilde{h}_n)(x)\,dx \\ &= 0_p\left(\int|\hat{q}_n^2 - \tilde{q}_n^2|\psi_n|\hat{h}_n - \tilde{h}_n|(x)\,dx\right) \\ &= 0_p\left(\left(\int|\hat{q}_n^2 - \tilde{q}_n^2|^2\psi_n\tilde{h}_n(x)\,dx\right)^{1/2}\right. \\ &\quad \left.\times\left(\int(\hat{h}_n - \tilde{h}_n)^2\tilde{h}_n^{-1}\psi_n(x)\,dx\right)^{1/2}\right) \\ &= 0_p\big(\sigma_n^{-3}c_n^{-1}n^{-1}\big). \end{aligned} \tag{2.22}$$

Proof of Lemma 2.3. For (2.8) write

$$\int(\bar{q}_n - \tilde{q}_n)\psi_n\hat{h}_n(x)\,dx = \int(\bar{q}_n - \tilde{q}_n)\psi_n(\hat{h}_n - \tilde{h}_n)(x)\,dx$$

and proceed as for (2.16).

For (2.9) write

$$\begin{aligned}\hat{I}_n - \int \bar{q}_n^2 \psi_n \tilde{h}_n(x)\,dx &= \int \bar{q}_n^2 \psi_n (\hat{h}_n - \tilde{h}_n)(x)\,dx + \int (\bar{q}_n^2 - \tilde{q}_n^2)\psi_n \tilde{h}_n(x)\,dx \\ &= 0_p\left(\sigma_n^{-2}\left(\int (\hat{h}_n - \tilde{h}_n)^2 \tilde{h}_n^{-1}\psi_n(x)\,dx\right)^{1/2}\right. \\ &\qquad \left. + \sigma_n^{-1}\left(\int (\bar{q}_n - \tilde{q}_n)^2 \psi_n \tilde{h}_n(x)\,dx\right)^{1/2}\right) \\ &= 0_p\left(\sigma_n^{-5/2} c_n^{-1/2} n^{-1/2}\right)\end{aligned}$$

as in (2.21)–(2.22). □

Lemma 2.4 follows from Lemmas 2.1–2.3 and $\liminf_n I_n(\tilde{H}_n) > 0$, a consequence of Lemma 2.5 and our assumption on $\mathscr{H}$.

Proof of Lemma 2.5. For the proof of (2.11), without loss of generality suppose $\psi^{1/2}$ is continuously differentiable since for any ψ_1 satisfying our conditions and $\varepsilon > 0$, there exists a ψ_2 satisfying them such that $\psi_2^{1/2}$ is continuously differentiable and

$$(1 - \varepsilon)\psi_2(x) \le \psi_1(x) \qquad \text{for all } x.$$

If $H_n \xrightarrow{w} H_0$, $H_0(R) > 0$, and $\sup_n I(H_n) \le M < \infty$ by Cauchy–Schwarz,

$$\begin{aligned}|h_n^{1/2}(x) - h_n^{1/2}(y)| &= \frac{1}{2}\left|\int_x^y h_n^{-1/2} h_n'(t)\,dt\right| \\ &\le \frac{1}{2} I^{1/2}(H_n)|x - y|^{1/2} \\ &\le \frac{M^{1/2}}{2}|x - y|^{1/2}. \qquad (2.23)\end{aligned}$$

Since $\int h_n(x)\,dx = 1$ for all n, (2.23) implies $\{h_n(x_0)\}$ bounded for any x_0. By Ascoli's theorem, (2.23) then implies $\{h_n^{1/2}\}$ and hence $\{h_n\}$ compact in the sup norm on $[-a, a]$ for all $a < \infty$. Since $H_n \xrightarrow{w} H_0$, a subsequence argument yields

$$h_n^{1/2}(x) \to h_0^{1/2}(x) \qquad (2.24)$$

uniformly on $[-a, a]$. Next define an operator T_n on $L_2(R)$ by

$$T_n(v) = \frac{1}{2}\int_R h_n' h_n^{-1/2}\psi_n^{1/2} v(x)\,dx = \int_R \psi_n^{1/2} v(x)\,dh_n^{1/2}(x)$$

and

$$T(v) = \frac{1}{2}\int_R h_0' h_0^{-1/2} v(x)\, dx.$$

If v is continuously differentiable with compact support,

$$T_n(v) = -\int_R h_n^{1/2}\left(\left[\psi_n^{1/2}\right]' v + v'\psi_n^{1/2}\right)(x)\, dx \to -\int_R h_0^{1/2} v'(x)\, dx = T(v)$$

by (2.24) since the integrand is bounded and vanishes off a compact. Moreover

$$\begin{aligned} 4\|T_n\|^2 &= \int_R \psi_n \frac{[h_n']^2}{h_n}(x)\, dx \\ &= I_n(H_n) \le I(H_n) \le M. \end{aligned} \tag{2.25}$$

By the Banach Steinhaus theorem,

$$T_n(v) \to T(v) \tag{2.26}$$

for all v and

$$\liminf_n \|T_n\|^2 \ge \|T\|^2 = \frac{1}{4} I(H_0) \tag{2.27}$$

and (2.11) follows.

Since $\tilde{H}_n \xrightarrow{w} H_0$ and $I_n(\tilde{H}_n) \le I(\tilde{H}_n) \le I(H_n)$, (2.12) follows from (2.11).

Now take $\psi_n = 1$. The argument leading to (2.25)–(2.27) is valid. Therefore if $I(H_n) \to I(H_0)$, by (2.25) and (2.27),

$$\|T_n\| \to \|T\|. \tag{2.28}$$

But (2.26) and (2.28) imply

$$\|T_n - T\| \to 0$$

which is equivalent to (2.13).

Proof of Lemma 2.6. By Theorem 3.1, p. 124 of [7], we need only check that

$$\sup\left\{\int\left(h_n' h_n^{-1/2}\left(x - \frac{\theta}{n^{1/2}}\right) - h_n' h_n^{-1/2}(x)\right)^2 dx : |\theta| \le M\right\} \to 0, \qquad \forall M < \infty,$$

and

$$\int\left[\frac{h_n'}{h_n}\right]^2 I\left[\left|\frac{h_n'}{h_n}\right| \ge \varepsilon n^{1/2}\right] h_n(x)\, dx \to 0, \qquad \forall \varepsilon > 0.$$

The first claim follows from (2.13) and the L_2 continuity theorem, the

second from (2.13) and (2.24).

Proof of Theorem 1.1. By Lemmas 2.4 and 2.5, $\mathscr{L}_{H_n}(n^{1/2}v_n^*)$ and $\mathscr{L}_{H_n}(n^{-1/2}I^{-1}(H_0)\Sigma_{i=1}^n - \int \tilde{q}_n\psi_n(x)k_{\sigma_n}(x - X_i)\,dx)$ are asymptotically equal. Moreover,

$$\left|\int \tilde{q}_n\psi_n(x)k_{\sigma_n}(x - X_i)\,dx\right| \le \sigma_n^{-1} = o(n^{1/2})a.s. \tag{2.29}$$

and, by (2.12),

$$\begin{aligned}\limsup_n &\int\left(\int \tilde{q}_n\psi_n(x)k_{\sigma_n}(x - z)\,dx\right)^2 h_n(z)\,dz \\ &\le \limsup_n I_n(\tilde{H}_n) = I(H_0),\end{aligned}$$

if $H_n \xrightarrow{w} H_0$ and $I(H_n) \to I(H_0)$. By Lindeberg's theorem, the sequence $\mathscr{L}_{H_n}(n^{1/2}v_n^*)$ is then tight and all its limit points are $\mathscr{N}(0, \sigma^2)$ with $\sigma^2 \le I^{-1}(H_0)$. If $H_0(R) = 1$ and $I(H_0) < \infty$, by Lemma 2.6, and Cor. 11.1, p. 161 of [7], $\sigma^2 \ge I^{-1}(H_0)$ and the theorem follows. As a by-product we obtain

$$\int\left(\int \tilde{q}_n\psi_n(x)k_{\sigma_n}(x - z)\,dx\right)^2 h_n(z)\,dz \to I(H_0). \tag{2.30}$$

□

THEOREM 2.1. *Suppose $X_{1n}, \ldots, X_{nn}$ are independent, X_{in} has density $h_{in}(\cdot - v) = \theta_{in}^{-1}f(\theta_{in}^{-1}(\cdot - v))$, $i = 1, \ldots, n$, f symmetric about 0 and $I(F) < \infty$. By H_n we denote the distribution function of $h_n = n^{-1}\Sigma_{i=1}^n h_{in}$. If H_n and H_0 satisfy the conditions of Theorem* 1.1, *then*

$$\mathscr{L}_{(0,\theta_{1n},\ldots,\theta_{nn})}(n^{1/2}v_n^*) \to \mathscr{N}(0, I^{-1}(H_0)). \tag{2.31}$$

Proof of Theorem 2.1. The proofs of Lemmas 2.1–2.4 are essentially unchanged for this new model, as we can see by noting that the key inequalities (2.15) and (2.17) continue to hold. Moreover,

$$\begin{aligned}&\operatorname{var}_{(0,\theta_{1n},\ldots,\theta_n)}\left(n^{-1/2}\sum_{i=1}^n \int \tilde{q}_n\psi_n(x)k_{\sigma_n}(x - X_i)\,dx\right) \\ &\quad = \int\left(\int \tilde{q}_n\psi_n(x)k_{\sigma_n}(x - z)\,dx\right)^2 h_n(z)\,dz \to I(H_0),\end{aligned}$$

by (2.30). Consequently, (2.29) and Lindeberg's theorem yield (2.31).

NOTES. (1) If $f = \phi$, Theorem 2.1 shows that ν_n^* is regular (II′) and hence by Theorem 1.1 and Proposition 1.1 efficient (II′). We conjecture that it is in fact efficient within the class of all asymptotically normal translation equivariant estimates which are symmetric and even depend on G_n. That is, ν_n^* does as well as if we knew the θ_{in} up to a permutation.

(2) The companion problem, $X_i = (X_{i1}, X_{i2})$, X_{i1}, X_{i2} independent $\mathcal{N}(\theta_i, \nu)$ is much easier. Lindsay [10] and Hammerstrom [4] showed that the UMVU estimate $(2n)^{-1}\Sigma_{i=1}^n (X_{i1} - X_{i2})^2$ is efficient.

ACKNOWLEDGMENTS

We thank I. Johnstone, J. Ritov, and K. Takeuchi for helpful discussions.

REFERENCES

1. J. M. BEGUN, W. J. HALL, W. M. HUANG, AND J. A. WELLNER, Information and asymptotic efficiency in parametric–nonparametric models, *Ann. Statist.* **11** (1983), 432–452.
2. P. J. BICKEL, On adaptive estimation, *Ann. Statist.* **10** (1982), 647–671.
3. P. J. BICKEL, C. A. J. KLAASSEN, J. RITOV, AND J. A. WELLNER, "Efficient and Adaptive Statistical Inference," to be published by Johns Hopkins University Press, 1987.
4. I. J. GOOD, The population frequencies of species and the estimation of population parameters, *Biometrika* **40** (1953), 237–264.
5. T. HAMMERSTROM, "On Asymptotic Optimality Properties of Tests and Estimates in the Presence of Increasing Numbers of Nuisance Parameters," Ph.D. dissertation, University of California, Berkeley, 1978.
6. P. J. HUBER, Robust estimation of a location parameter, *Ann. Math. Statist.* **35** (1964), 73–101.
7. I. A. IBRAGIMOV AND R. Z. HASMINSKII, "Statistical Estimation: Asymptotic Theory," Springer-Verlag, New York, 1981.
8. J. KIEFER AND J. WOLFOWITZ, Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *Ann. Math. Statist.* **27** (1956), 887–906.
9. YU. A. KOSHEVNIK AND B. YA. LEVIT, On a non-parametric analogue of the information matrix, *Theory Probab. Appl.* **21** (1976), 738–753.
10. B. G. LINDSAY, Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators, *J. Philos. Trans. R. Soc. London Ser A* **296** (1980), 639–665.
11. B. G. LINDSAY, Efficiency of the conditional score in a mixture setting, *Ann. Statist.* **11** (1983), 486–497.
12. J. NEYMAN AND E. SCOTT, Consistent estimates based on partially consistent observations, *Econometrica* **16** (1948), 1–32.
13. J. PFANZAGL, "Contributions to a General Asymptotic Statistical Theory," Lecture Notes in Statistics, Springer Pub., New York, 1982.
14. H. ROBBINS, Asymptotically subminimax solutions of compound statistical decision problems, *in* "Proc. Second Berkeley Symp. Math. Statist. Probab.," pp. 131–148, Univ. of California Press, Berkeley, 1951.

15. H. ROBBINS, An empirical Bayes approach to statistics, *in* "Proc. Third Berkeley Symp. Math. Statist. Probab.," Vol. 1, pp. 157–163, Univ. of California Press, Berkeley, 1956.
16. H. ROBBINS, Some thoughts on empirical Bayes estimation, *Ann. Statist.* **11** (1983), 713–724.
17. C. STEIN, Efficient nonparametric testing and estimation, *in* "Proc. Third Berkeley Symp. Math. Statist. Prob.," Vol. 1, pp. 187–195, Univ. of California Press, Berkeley, 1956.
18. C. STONE, Adaptive maximum likelihood estimators of a location parameter, *Ann. Statist.* **3** (1975), 267–284.

The Annals of Statistics
1987, Vol. 15, No. 2, 513–540

EFFICIENT ESTIMATION IN THE ERRORS IN VARIABLES MODEL[1]

BY P. J. BICKEL AND Y. RITOV

University of California, Berkeley and The Hebrew University of Jerusalem

We consider efficient estimation of the slope in the errors in variables model with normal error when either the ratio of error variances is known and the distribution of the independent is arbitrary and unknown or the distribution of the independent variable is not Gaussian or degenerate. We calculate information bounds and exhibit estimates achieving these bounds using an initial minimum distance estimate and suitable estimates of the efficient score function.

1. Introduction. Errors in variables models have been the subject of an enormous amount of literature. A fairly recent reference with a good bibliography is Anderson (1984).

In its simplest form the model assumes n independent observations $\mathbf{X}_i = (X_i, Y_i)$, which are written as

$$\begin{aligned} X_i &= X_i' + \varepsilon_{i1}, \\ Y_i &= \alpha + \beta X_i' + \varepsilon_{i2}. \end{aligned} \tag{1.1}$$

The X_i' are viewed either as

(i) unknown constants;
(ii) independent identically distributed random variables.

Model (i) is called functional and (ii) structural by Kendall and Stuart (1979), Chapter 29.

The $(\varepsilon_{i1}, \varepsilon_{i2})$ are considered random vectors, which are identically distributed with mean 0, as well as independent of the X_i' in model (ii). In this paper we will deal exclusively with large sample theory in the structural model, although we believe our results generalize to the functional model. Our aim in this paper is the construction of efficient estimates of β under various assumptions in various special cases of (1.1). We also suggest how our results may be extended to instrumental variable models through the special case of repeated observations at the same X_i'.

Write $\mathbf{X}$, X', ε_1, ε_2 for "generic" observations. If we do not make any assumptions on the distributions of X and $(\varepsilon_1, \varepsilon_2)$, then β is clearly unidentifiable. In fact, β is unidentifiable even if we assume ε_1, ε_2 to be independent Gaussian variables with unknown variances and suppose X' is also Gaussian.

Received October 1984; revised September 1986.
[1]This research was supported in part by ONR Contract N00014-80-C-0163.
AMS 1980 *subject classifications.* Primary 62G20; secondary 62P20.
Key words and phrases. Reiersøl models, semiparametric, minimum distance, structural.

However, β has been shown to be identifiable under various sets of assumptions. These fall into two broad classes:

(A) *Gaussian errors.* $(\varepsilon_1, \varepsilon_2)$ have a bivariate Gaussian distribution with variance–covariance matrix Σ. The usual way to make β identifiable in the literature is to assume $\varepsilon_1, \varepsilon_2$ independent and either

$$\text{Var}(\varepsilon_1) = c_0 \text{Var}(\varepsilon_2) \tag{1.2}$$

or

$$\text{Var}(\varepsilon_1) = c_0, \tag{1.3}$$

with c_0 assumed known. Both (1.2) and (1.3) are plausible under special circumstances [see Kendall and Stuart (1979), Chapter 29, for a discussion]. We shall explore a generalization of (1.2),

$$\Sigma = \sigma^2 \Sigma_0, \tag{1.4}$$

where Σ_0 is known. Model (1.3) can be analyzed in the same way. We shall call (1.4) the *restricted Gaussian error* model. This model and its generalizations to more complicated situations have been extensively studied; see Anderson (1984), for example. A second model in which the identifiability of β was established by Reiersøl (1950) puts no restriction on Σ but requires X' to be non-Gaussian (where constants are viewed as Gaussian). We shall call this the *general Gaussian error* model.

(B) *General independent errors.* Assume $\varepsilon_1, \varepsilon_2$ independent. If (1.2) holds, β is identifiable. This *restricted independent error* has also been extensively studied. If (1.2) is not present but either X' is non-Gaussian or $\varepsilon_1, \varepsilon_2$ have no Gaussian component, then, again according to Reiersøl (1950), β is identifiable. This *arbitrary independent error* model is probably most satisfactory but our results do not bear on it.

We review briefly some results on these models.

The restricted Gaussian model can be reduced to case (1.2) with $c_0 = 1$. The maximum likelihood estimate for β in this case is $\hat{\beta}_P$, which minimizes the sum of squared perpendicular distances of observed points from the fitted line

$$\sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta X_i)^2}{1 + \beta^2}. \tag{1.5}$$

This estimate is well known to be $n^{1/2}$-consistent and asymptotically normal not only under the restricted Gaussian model but also under the restricted independent error model, see Gleser (1981) who considers multivariate generalizations. In the presence of fourth moments, it is not hard to show that $n^{1/2}$-consistency and asymptotic normality persist under the restricted independent error model when Σ_0 is the identity. Estimates of β in the general Gaussian error model, with Σ_0 diagonal, have been proposed by a variety of authors including Neyman and Scott (1948) and Rubin (1956). In the arbitrary independent error model, Wolfowitz in a series of papers ending in 1957, Kiefer and Wolfowitz (1956) and Spiegelman (1979) by a variety of methods gave estimates, which are consistent and in Spiegelman's case $n^{1/2}$-consistent and asymptotically normal.

Little seems to be known about the efficiency of these procedures other than that in the restricted Gaussian model the estimate $\hat{\beta}_P$ is efficient if X' is Gaussian by the classical results for M.L.E.'s in parametric models. Our main aims in this paper are:

In the general Gaussian error model:

(i) To give the structure that efficient estimates in the sense of Stein (1956), Koshevnik and Levit (1976) and Pfanzagl (1982) must have (Theorem 2.1).

(ii) To exhibit a reasonable efficient estimate (Theorem 2.2). In addition, we extend Theorem 2.1 to the simplest instrumental variable model, m repeated measurements with Gaussian errors,

$$X_{ij} = X_i' + \varepsilon_{ij1},$$
$$Y_{ij} = \alpha + \beta X_i + \varepsilon_{ij2}, \qquad j = 1, \ldots, m,\ i = 1, \ldots, r,\ n = mr,$$

and

$$\mathbf{X}_i = \{(X_{ij}, Y_{ij}),\ j = 1, \ldots, m\},$$

where $m \geq 2$.

The ε_{ij2} are independent and identically distributed Gaussian and independent of ε_{ij1} which are also Gaussian. We refer this as the multiple *Gaussian measurements model.* Note that in this model if $m \geq 2$, the assumption of non-Gaussianity of the distribution of X' is unnecessary.

We speak of efficient estimation in the sense of Stein (1956) as developed by Koshevnik and Levit (1976), Pfanzagl (1982), Begun, Hall, Huang and Wellner (1983) and in a forthcoming monograph by Klaassen, Wellner and ourselves. Let $\mathbf{P}$ be the set of possible joint distributions of $\mathbf{X}$. We call $\mathbf{P}_0$ a parametric submodel of $\mathbf{P}$ if $\mathbf{P}_0 \subset \mathbf{P}$ and $\mathbf{P}_0$ can be represented as $\{P_{(\beta,\eta)};\ \beta \in \mathbf{R},\ \eta \in E \text{ open } \subset \mathbf{R}^k\}$. A parametric submodel is regular if at every (β_0, η_0) the mapping $(\beta, \eta) \to P_{(\beta,\eta)}$ is continuously Hellinger differentiable. Suppose that P belongs to $\mathbf{P}_0$—a regular parametric submodel of $\mathbf{P}$. Then the notion of information bound and efficient estimation of β are well defined [e.g., Ibragimov and Has'minskii (1981), pages 158–169]. Let $n^{-1}I^{-1}(P; \beta, \mathbf{P}_0)$ denote the asymptotic variance of an efficient estimate of β when P ranges over $\mathbf{P}_0$. Clearly, if we only assume that $P \in \mathbf{P}$ we can estimate no better than if we assumed that $P \in \mathbf{P}_0$. Accordingly, let $I(P; \beta, \mathbf{P}) = \inf\{I(P; \beta, \mathbf{P}_0)$: $\mathbf{P}_0$ a regular parametric submodel, $P \in \mathbf{P}_0\}$, be the information bound for estimating β under $\mathbf{P}$.

Loosely speaking, $\hat{\beta}_n$ is regular and efficient in $\mathbf{P}$ if

$$\mathbf{L}_P\big(\sqrt{n}\,(\hat{\beta}_n - \beta(P))\big) \to \mathbf{N}(0, I^{-1}(P; \beta, \mathbf{P})),$$

in some sense uniformly in $P \in \mathbf{P}$. Here $\mathbf{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The weakest kind of uniformity acceptable is that

$$\mathbf{L}_{P_n}\big(\sqrt{n}\,(\hat{\beta}_N - \beta(P_n))\big) \to \mathbf{N}(0, I^{-1}(P; \beta, \mathbf{P})), \tag{1.6}$$

for sequences $P_n \in \mathbf{P}_0$, a regular parametric submodel as above, with $P_n = P_{(\beta_n, \eta_n)}$, $|\beta_n - \beta_0| = O(n^{-1/2}) = |\eta_n - \eta_0|$ for some β_0, η_0, $P = P_{(\beta_0, \eta_0)}$.

If $I^{-1}(P; \beta, \mathbf{P})$ is assumed at some $\mathbf{P}_0$, we obtain from the Hájek–Le Cam convolution theorem, Ibragimov and Has'minskii (1981), that $\hat{\beta}_n$ is asymptotically linear

$$\hat{\beta}_n = \beta(P) + n^{-1} \sum_{i=1}^{n} \tilde{l}(\mathbf{X}_i, P; \beta, \mathbf{P}) + o_p(n^{-1/2}),$$

where $\tilde{l}$ is defined as the efficient influence function, which has the properties

$$E_P \tilde{l}(\mathbf{X}_i, P; \beta, \mathbf{P}) = 0,$$

$$E_P \tilde{l}^2(\mathbf{X}_i, P, \beta, \mathbf{P}) = I^{-1}(P; \beta, \mathbf{P}).$$

Finding $\tilde{l}$ is equivalent to finding a suitable least favorable $\mathbf{P}_0$ (at each P). We discuss the theory which guides us in this search in Section 3.

Note that an estimate is efficient if

(a) it converges in law uniformly [as in (1.6)] on $\mathbf{P}$ and

(b) it is efficient in some parametric submodel $\mathbf{P}_0$ at each P. By the Hájek–Le Cam theorem (b) holds iff the efficient influence function is the influence function of the (local) maximum likelihood estimate of β in $\mathbf{P}_0$.

In Section 2 (Theorem 2.1), we exhibit $\tilde{l}$ and $\mathbf{P}_0$ for the general Gaussian error model and the restricted Gaussian model and discuss the main features of $I(P; \beta, \mathbf{P})$. In Theorem 2.2 we exhibit, for each of the two models, an estimate $\hat{\beta}$, converging in law uniformly [as in (1.6)] on $\mathbf{P}$, which has $\tilde{l}$ as influence function. By (a) and (b), $\hat{\beta}$ is necessarily efficient. The proof of Theorem 2.1 is deferred to Section 3, and the proof of Theorem 2.2 to Section 4.

2. The main results. Without loss of generality let $(\varepsilon_{i1}, \varepsilon_{i2}) \sim \mathbf{N}(0, \Sigma)$ where $\Sigma = [\sigma_{ij}]_{2\times 2}$ is nonsingular. Let $\theta = (\alpha, \beta, \Sigma)$ and

$$U(\theta) = U(\mathbf{X}, \theta) = \frac{Y - \alpha - \beta X}{\bar{\sigma}(\theta)}, \tag{2.1}$$

$$T(\theta) = T(\mathbf{X}, \theta) = \bar{\sigma}^{-2}(\theta)\left[(\sigma_{22} - \beta\sigma_{12})X + (\beta\sigma_{11} - \sigma_{12})(Y - \alpha)\right], \tag{2.2}$$

where $\bar{\sigma}^2(\theta)$ is the variance of $Y - \alpha - \beta X$ if θ is true,

$$\bar{\sigma}^2(\theta) = \beta^2\sigma_{11} - 2\beta\sigma_{12} + \sigma_{22}. \tag{2.3}$$

Then given θ, $T(\theta)$ is a complete and sufficient statistic for X' treated as a parameter, i.e., for the model $\{\mathbf{L}_\theta(\mathbf{X}|X' = \eta)\colon \eta \in R\}$. This follows since given $X' = \eta$, (X, Y) have an $\mathbf{N}(\eta, \alpha + \beta\eta, \Sigma)$ distribution. Moreover, $U(\theta)$ is ancillary in this problem. It is necessarily independent of $T(\theta)$ in the original model and is distributed $\mathbf{N}(0, 1)$. $T(\theta)$ is also the unbiased predictor of X', i.e., given $X' = \eta$, $T(\theta)$ has a $\mathbf{N}(\eta, \tilde{\sigma}^2(\theta))$ distribution, where

$$\tilde{\sigma}^2(\theta) = \bar{\sigma}^{-2}(\theta)\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right).$$

We can write the joint density of $\mathbf{X}$ under (θ, G), where G is the distribution of X',

$$p(\mathbf{x}, \theta, G) = \int K(\mathbf{x}, z, \theta) G(dz), \tag{2.4}$$

where

$$K(\mathbf{x}, z, \theta) = \left[2\pi\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right)^{1/2}\right]^{-1} \times \exp\left\{-\left[2\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right)\right]^{-1} \times \left[\sigma_{22}(x - z)^2 - 2\sigma_{12}(x - z) \times (y - \alpha - \beta z) + \sigma_{11}(y - \alpha - \beta z)^2\right]\right\}$$

$$= \left[2\pi\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right)^{1/2}\right]^{-1} \exp\left\{-\frac{1}{2}U^2(\mathbf{x}, \theta)\right\} \times \exp\left\{-\frac{\tilde{\sigma}^{-2}(\theta)}{2}\left(T(\mathbf{x}, \theta) - z\right)^2\right\},$$

is the conditional density of $\mathbf{X}$ given $X' = z$.

Fix $\theta = \theta_0$, $G = G_0$. Drop the argument θ in $U(\theta)$, $T(\theta)$, $\bar{\sigma}^2(\theta)$, and $\tilde{\sigma}^2(\theta)$. Let

$$\omega(t) = \omega(t, \theta, G) = \tilde{\sigma}^{-1} \int \phi\left(\tilde{\sigma}^{-1}(t - z)\right) G(dz) \tag{2.5}$$

be the density of T and let

$$I_0 = \int \frac{[\omega']^2}{\omega}(t)\, dt$$

be the Fisher information for location of ω. Let $\eta = (\mu, \tau)$, $\mu \in R$, $\tau > 0$, and

$$G(\cdot, \eta) = G_0\left(\frac{\cdot - \mu}{\tau}\right).$$

Define

$$\mathbf{P}_0 = \left\{P_{(\theta, G(\cdot, \eta))}\right\}. \tag{2.6}$$

That is, in $\mathbf{P}_0$ we assume G known up to location and scale. $\mathbf{P}_0$ is not the same in the general Gaussian error model and the restricted Gaussian error model since Σ varies freely in the former!

THEOREM 2.1. *Assume* $\int \eta^2 G(d\eta) < \infty$. *Then* $\mathbf{P}_0$ *is the least favorable regular parametric submodel and the information bounds and the efficient influence functions for estimating* β *at* $\theta = \theta_0$, $G = G_0$, *are as follows*:

Restricted Gaussian error model. *Define the random variable*

$$l_a^* = \bar{\sigma}^{-1} U\left(T - E(T) + \tilde{\sigma}^2 \frac{\omega'}{\omega}(T)\right). \tag{2.7}$$

This is the efficient score function defined by Begun, Hall, Huang and Wellner (1983). *The information bound of* (1.5), *which we write as* I_a, *is given by*

$$I_a = E_0(l_a^*)^2 = \bar{\sigma}^{-2}\left(\mathrm{Var}(T) + \tilde{\sigma}^4 I_0 - 2\tilde{\sigma}^2\right) = \bar{\sigma}^{-2}\left(\mathrm{Var}(X') + \tilde{\sigma}^2\left(\tilde{\sigma}^2 I_0 - 1\right)\right) \tag{2.8}$$

and the efficient influence function is given by

$$\tilde{l}_a = l_a^*/I_a. \tag{2.9}$$

General Gaussian error model. *Define*

$$l_b^* = \bar{\sigma}^{-1}U\Big(T - E(T) + I_0^{-1}\frac{\omega'}{\omega}(T)\Big). \tag{2.10}$$

The information bound is given by

$$\begin{aligned} I_b = E(l_b^*)^2 &= \bar{\sigma}^{-2}\big(\mathrm{Var}(T) - I_0^{-1}\big) \\ &= \bar{\sigma}^{-2}\big(\mathrm{Var}(X') + \tilde{\sigma}^2 - I_0^{-1}\big) \end{aligned} \tag{2.11}$$

and the efficient influence function by

$$\tilde{l}_b = l_b^*/I_b. \tag{2.12}$$

NOTES.

Restricted Gaussian error model.

(1) If $\sigma_{11} = 0$, then $\tilde{\sigma} = 0$ and we are in the case where $T = X = X'$ is observed without error. In this case,

$$I_a = \mathrm{Var}(X')/\mathrm{Var}(Y - \alpha - \beta X)$$

is the reciprocal of the asymptotic variance of $n^{1/2}$ times the ordinary least-squares estimate as it should be.

(2) If X' is normal, $\mathrm{Var}(T) = I_0^{-1}$ and (2.7) becomes

$$\begin{aligned} \bar{\sigma}^{-2}\big(\mathrm{Var}(X') + \tilde{\sigma}^2\big(\tilde{\sigma}^2 - \mathrm{Var}(T)\big)I_0\big) &= \bar{\sigma}^{-2}(\mathrm{Var}\,X')\big(1 - \tilde{\sigma}^2 I_0\big) \\ &= \bar{\sigma}^{-2}\mathrm{Var}^2(X')/\mathrm{Var}_0(T), \end{aligned}$$

which we shall call I_c.

This is just the asymptotic variance of $\hat{\beta}_P$ if $\Sigma_0 =$ identity [see, e.g., Gleser (1981)], whatever be G. So we conclude that we can do as well not knowing G as knowing it is Gaussian. This is a special instance of the claim that $\mathbf{P}_0$ given by (2.6) is least favorable.

(3) We can study the asymptotic efficiency I_c/I_a of $\hat{\beta}_P$ if G_0 is *not* normal. We show in Section 5 that, $I_c/I_a \geq (1 + \sigma^2/(\beta^2 + 1)(\mathrm{Var}(X') + \sigma^2))^{-1}$. In particular, if the signal-to-noise ratio in X, $\mathrm{Var}(X')/\sigma^2$, is large $\hat{\beta}_P$ is close to efficiency.

(4) The score function l_a^* can be written as

$$l_a^* = \bar{\sigma}^{-1}U(E(X'|T) - E(X')).$$

The least-squares estimate if X' were known is based on the score function

$$\bar{\sigma}^{-1}U(X' - E(X')).$$

Thus the efficient estimate replaces the unobservable X' by its best "estimate" $E(X'|T)$.

(5) Suppose that with $\Sigma = \sigma^2\Sigma_0$ we have m repeated observations at each X_i'. Then by sufficiency l_a^*, evaluated at the mean of each set of observations with Σ_0 replaced by Σ_0/m, is the efficient score function.

General Gaussian error model.

(1) Normality of X', under which β is unidentifiable, corresponds to G = point mass at 0. Appropriately, $I_b \to 0$ as G tends to point mass since then T approaches normality and $\tilde{\sigma}^2 \sim I_0^{-1}$.

(2) Necessarily, $I_a \geq I_b$. The inequality is always strict since

$$\tilde{\sigma}^2(I_a - I_b) = I_0^{-1}\left(\tilde{\sigma}^4 I_0^2 - 2\tilde{\sigma}^2 I_0 + 1\right)$$
$$= I_0^{-1}\left(\tilde{\sigma}^2 I_0 - 1\right)^2 > 0,$$

since I_0, the Fisher information for $X' + \varepsilon_1$, is always smaller than the Fisher information for ε_1 which is just $\tilde{\sigma}^{-2}$.

Multiple Gaussian measurements model. The efficient influence function can be calculated as for the general Gaussian error model, but is much more complicated.

Let $\mathbf{X} = (X_j, Y_j)$, $j = 1, \ldots, m$, where $X_j = X' + \varepsilon_{j1}$, $Y_j = \alpha + \beta X' + \varepsilon_{j2}$ is a generic observation. We assume the ε_{ji} are independent Gaussian with mean 0 and $\mathrm{Var}(\varepsilon_{j1}) = \sigma_{11}$, $\mathrm{Var}(\varepsilon_{j2}) = \sigma_{22}$. Let

$$U = \left(\overline{Y} - \beta\overline{X} - \alpha\right)/\sigma_0,$$
$$T = \left(\sigma_{22}\overline{X} + \beta\sigma_{11}(\overline{Y} - \alpha)\right)/\left(\sigma_{22} + \beta^2\sigma_{11}\right), \tag{2.13}$$

where $\overline{Y} = m^{-1}\Sigma_{j=1}^m Y_j$, $\overline{X} = m^{-1}\Sigma_{j=1}^m X_j$. Let

$$\sigma_0^2 = \left(\sigma_{22} + \beta^2\sigma_{11}\right)/m,$$
$$\tilde{\sigma}^2 = \sigma_{11}\sigma_{22}/m^2\sigma_0^2, \tag{2.14}$$
$$I_0 = \int\left(\frac{w'}{w}\right)^2 w(t)\,dt, \quad \text{where } w \text{ is the density of } T \text{ given by (2.13).}$$

The efficient score function is then

$$l^* = \frac{UT}{\sigma_0\tilde{\sigma}^2} + a_2\frac{U}{\sigma_0\tilde{\sigma}^2}\frac{\omega'}{\omega}(T) + a_3(U^2 - 1) + a_4 S_1 + a_5 S_2, \tag{2.15}$$

where

$$S_1 = \sum_{j=1}^m \frac{\left(Y_j - \overline{Y}\right)^2}{\sigma_{22}} - (m-1), \qquad S_2 = \sum_{j=1}^m \frac{\left(X_j - \overline{X}\right)^2}{\sigma_{11}} - (m-1)$$

and the a's are functions of m, σ^2, σ_0^2 and I_0. For $m = 1$ the form of l^* agrees with l_b^* as it should. As $m \to \infty$,

$$a_2 \sim \tilde{\sigma}^2,$$

which corresponds to l_a^*. This is as expected since m large corresponds to σ_{11}, σ_{22} essentially known. The information I_d for this problem is I_b plus a complicated positive term vanishing for $m = 1$.

We now construct efficient estimates. The idea is to proceed as in the classical estimation of the location problem:

(a) Find a good estimate $\tilde{\beta}_n$ of β.

(b) (i) Consider $\tilde{l}$ as $\tilde{l}(\mathbf{x}, \beta, \eta, G)$ where $\theta = (\beta, \eta)$, G are now viewed as dummy variables and the argument $\mathbf{x}$ replaces $\mathbf{X}$. For example,

$$\tilde{l}_a(\mathbf{x}, \theta, G) = \bar{\sigma}^{-1}(\theta)U(\mathbf{x}, \theta)\Big(T(\mathbf{x}, \theta) - \int T(\mathbf{x}, \theta)P_{(\theta, G)}(d\mathbf{x}) + \tilde{\sigma}^2(\theta)\frac{\omega'}{\omega}(T(\mathbf{x}, \theta), \theta)\Big)\Big/I_a(\theta, G),$$

where T is given by (2.2) and $\omega(\cdot, \theta)$ is the marginal density of $T(\mathbf{X}, \theta)$, under $P_{(\theta, G)}$. Construct a suitable estimate $\hat{\tilde{l}}(\mathbf{x}, \beta; \mathbf{X}_1, \ldots, \mathbf{X}_n)$ of $\tilde{l}(\mathbf{x}, \beta, \eta, G)$.

(ii) Form

$$\hat{\beta}_n = \tilde{\beta}_n + n^{-1}\sum_{i=1}^{n}\hat{\tilde{l}}(\mathbf{X}_i, \tilde{\beta}_n; \mathbf{X}_1, \ldots, \mathbf{X}_n)$$

as the efficient estimate.

PRELIMINARY ESTIMATE. We motivate our $\tilde{\beta}_n$ as follows. If we calculate under P_0 and $\beta = \beta_0$, $\mathrm{Var}(Y) \geq \mathrm{Var}(\beta X)$, then

$$\mathbf{L}(Y) = \mathbf{L}(\beta X + \sigma Z + \mu), \tag{2.16}$$

for $Z \sim \mathbf{N}(0, 1)$ independent of X and

$$\mu = E(Y) - \beta E(X),$$
$$\sigma^2 = \mathrm{Var}(Y) - \beta^2\mathrm{Var}(X).$$

If $\mathrm{Var}(Y) < \mathrm{Var}(\beta X)$, then

$$\mathbf{L}(X) = \mathbf{L}\Big(\frac{Y}{\beta} + \sigma Z + \mu\Big), \tag{2.17}$$

for $Z \sim \mathbf{N}(0, 1)$ independent of Y, some σ, μ. For $|\beta| \neq |\beta_0|$ neither identity (2.16) nor (2.17) can hold; see Proposition 5.1. Our initial estimate is essentially a minimizing value for the distance between the natural estimates of the laws in (2.16) or (2.17). We believe our estimate may be improved by considering the joint distribution of (X, Y) and not only the marginals. For that note that if (2.16) holds, then

$$\mathbf{L}(\beta X + \sigma Z + \mu, Y) = \mathbf{L}(Y, \beta X + \sigma Z + \mu).$$

Another possible estimate is given by Spiegelman (1979) who does not assume Gaussianity of the errors but does assume $\varepsilon_1, \varepsilon_2$ independent. Different estimates $\tilde{\beta}_a, \tilde{\beta}_b$ are appropriate for the restricted Gaussian error model and the general Gaussian error model. Essentially, $\tilde{\beta}_b$ works whenever $\tilde{\beta}_a$ does except when G is

Gaussian. We give $\tilde{\beta}_b$ formally and sketch the difference for $\tilde{\beta}_a$. Without loss of generality, we assume $E(\varepsilon_1) = E(\varepsilon_2) = 0$.

Let $\hat{F}_1$ be the empirical distribution function of X_i, $i = 1, \ldots, n$, and $F_1(\cdot)$ be the distribution function of X. Let $\hat{F}_2(\cdot)$ and $F_2(\cdot)$ be the empirical distribution function of Y_i and the distribution function of Y, respectively. Let

$$\hat{\mu}(\beta) = \bar{Y} - \beta\bar{X}, \qquad \hat{\sigma}^2(\beta) = |\hat{\sigma}_y^2 - \beta^2\hat{\sigma}_x^2|,$$

$$\hat{\sigma}_y^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad \hat{\sigma}_x^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \lambda = \hat{\sigma}_y/\hat{\sigma}_x. \tag{2.18}$$

Define, for $\hat{\sigma}_x^2 > 0$, $\hat{\sigma}_y^2 > 0$,

$$\Delta_n(\beta) = \sqrt{n}\int\left|\hat{F}_2(y) - \int\Phi\left(\frac{y - \beta x - \hat{\mu}(\beta)}{\hat{\sigma}(\beta)}\right)d\hat{F}_1(x)\right|^2\phi(y)\,dy,$$

$$\text{if } \hat{\sigma}_y^2 > \beta^2\hat{\sigma}_x^2$$

$$= \sqrt{n}\int\left|\hat{F}_1(x) - \int\Phi\left(\frac{\beta x - y + \hat{\mu}(\beta)}{(\operatorname{sgn}\beta)\hat{\sigma}(\beta)}\right)d\hat{F}_2(y)\right|^2\lambda\phi(\lambda y)\,dy, \tag{2.19}$$

$$\text{if } \hat{\sigma}_y^2 < \beta^2\hat{\sigma}_x^2.$$

Note that $\Delta_n(\beta)$ can be defined by continuity at $\sigma(\beta) = 0$ since $P[|\beta| + \hat{\sigma}^2(\beta) > 0, \forall\beta] = 1$. For given $a > 0$, let $\Delta_n(\beta, a)$ be the corresponding quantity with Y_i replaced by $Y_i + aX_i$, $i = 1, \ldots, n$. Let $\beta_n^*(a)$ minimize $\Delta_n(\beta, a)$. $\beta_0 = 0$ poses difficulties but we can always shift away from this value. Accordingly, let

$$\beta_n^* = \beta_n^*(0), \qquad \text{if } |\beta_n^*(0)| \geq \delta_0$$

$$= \beta_n^*(2\delta_0) - 2\delta_0, \quad \text{if } |\beta_n^*(0)| < \delta_0.$$

Finally, we need to distinguish between $\pm\beta_n^*$. For that let $\hat{W}_n^+$ be the empirical distribution function of $\hat{\sigma}^{-1}(Y_i - \mu(\beta_n^*) - \beta_n^*X_i)$, where

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \mu(\beta_n^*) - \beta_n^*X_i)^2$$

and $\hat{W}_n^-$ the corresponding quantity for $-\beta_n^*$. Let

$$\tilde{\beta} = \beta_n^*, \qquad \text{if } \int|\hat{W}_n^+(y) - \Phi(y)|^2\phi(y)\,dy \leq \int|W_n^-(y) - \Phi(y)|^2\phi(y)\,dy$$

$$= -\beta_n^*, \quad \text{otherwise.}$$

For the restricted Gaussian error model, $\Sigma_0 =$ identity we proceed as above but change the definition of $\hat{\sigma}^2(\beta)$ to, using the new information,

$$\hat{\sigma}_a^2(\beta) = \frac{|1 - \beta^2|}{1 + \beta^2}n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}(\beta) - \beta X_i)^2$$

and switch the definition of $\Delta_n(\beta)$ as $\beta^2 \leq 1$ or > 1.

EFFICIENT ESTIMATES. Note that

$$\beta\sigma_{11} - \sigma_{12} = \beta \operatorname{Var}(X) - \operatorname{cov}(X, Y), \tag{2.20}$$

$$\sigma_{22} - \beta\sigma_{12} = \operatorname{Var}(Y) - \beta \operatorname{cov}(X, Y), \tag{2.21}$$

$$\alpha = E(Y) - \beta E(X). \tag{2.22}$$

We can reparametrize the general Gaussian error model using $(\beta, \alpha, \gamma_1, \gamma_2, \sigma_{11}, G)$, where $\gamma_1, \gamma_2, \alpha$ are the expressions in (2.20)–(2.22), respectively. Abusing notation, let $\theta = (\beta, \alpha, \gamma_1, \gamma_2)$ so that

$$U_i(\theta) = (Y_i - \alpha - \beta X_i)/(\beta\gamma_1 + \gamma_2)^{1/2},$$

$$T_i(\theta) = (\gamma_2 X_i + \gamma_1(Y_i - \alpha))/(\beta\gamma_1 + \gamma_2).$$

Define $\tilde{\theta}_n = (\tilde{\beta}_n, \tilde{\alpha}_n, \tilde{\gamma}_{1n}, \tilde{\gamma}_{2n})$ by substituting sample moments and $\tilde{\beta}_n$ in the definitions (2.20)–(2.22) for $\beta, \alpha, \gamma_1, \gamma_2$. Let

$$\lambda(t) = e^{-t}(1 + e^{-t})^2,$$

$$\lambda_\nu(t) = \frac{1}{\nu}\lambda\left(\frac{t}{\nu}\right).$$

For sequences $c_n, \nu_n \downarrow 0$, to be characterized later, let $\lambda_n = \lambda_{\nu_n}$ and estimate ω_0 by the kernel estimator,

$$\hat{\omega}_n(t, \theta) = \frac{1}{n}\sum_{i=1}^{n} \lambda_\nu(t - T_i(\theta)) + c_n.$$

Define the efficient estimate for the general Gaussian error model by

$$\tilde{\beta}_{nb} = \tilde{\beta}_n + n^{-1}\hat{I}_b^{-1}\sum_{i=1}^{n} \frac{\tilde{U}_i}{\sigma(\tilde{\theta}_n)}\left(\tilde{T}_i - \tilde{T}_\cdot + \hat{I}_0^{-1}\frac{\hat{\omega}_n'}{\hat{\omega}_n}(\tilde{T}_i, \tilde{\theta}_n)\right), \tag{2.23}$$

where $\tilde{U}_i, \tilde{T}_i$ are used for $U_i(\tilde{\theta}_n), T_i(\tilde{\theta}_n)$, and $\tilde{T}_\cdot = n^{-1}\Sigma_{i=1}^n \tilde{T}_i$,

$$\hat{I}_0 = n^{-1}\sum_{i=1}^{n}\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}\right)^2(T_i(\tilde{\theta}_n), \tilde{\theta}_n), \tag{2.24}$$

$$\hat{I}_b = (\tilde{\beta}_n\tilde{\gamma}_{12} + \tilde{\gamma}_{2n})^{-1}n^{-1}\sum_{i=1}^{n}\left(\tilde{T}_i - \tilde{T}_\cdot + \hat{I}_0^{-1}\frac{\omega_n'}{\omega_n}(\tilde{T}_i, \tilde{\theta}_n)\right)^2. \tag{2.25}$$

Similarly, we define the efficient estimate $\hat{\beta}_{na}$ for the restricted Gaussian error model by

$$\hat{\beta}_{na} = \tilde{\beta}_{na} + n^{-1}\hat{I}_a^{-1}\sum_{i=1}^{n}\frac{\tilde{U}_i}{\hat{\sigma}_n}\left(\tilde{T}_{ia} - \tilde{T}_{\cdot a} + (1 + \tilde{\beta}_n^2)^{-1}\hat{\sigma}_n^2\frac{\hat{\omega}_n'}{\hat{\omega}_n}(\tilde{T}_{ia}, \tilde{\theta}_n)\right),$$

where

$$\hat{\sigma}_n^2 = \left(1 + \tilde{\beta}_n^2\right)^{-1} n^{-1} \sum_{i=1}^{n} \left(Y_i - \mu\left(\tilde{\beta}_n\right) - \tilde{\beta}_n X_i\right)^2,$$

$$\tilde{T}_{ia} = \left(\tilde{\beta}_n(Y_i - \tilde{\alpha}_n) + X_i\right)\left(1 + \tilde{\beta}_n^2\right)^{-1},$$

$$\hat{I}_a = \hat{\sigma}_n^{-2} n^{-1} \sum_{i=1}^{n} \left(\tilde{T}_{ia} - \tilde{T}_{\cdot a} + \left(1 + \tilde{\beta}_n^2\right)^{-1} \hat{\sigma}_n^2 \frac{\omega_n'}{\omega_n}\left(\tilde{T}_{ia}, \tilde{\theta}_n\right)\right)^2,$$

in accordance with (2.7) and (2.8).

Let $\{c_n\}, \{\nu_n\}$ be such that

$$c_n \to 0, \qquad \nu_n \to 0, \qquad n c_n^2 \nu_n^6 \to \infty.$$

THEOREM 2.2. (i) *Suppose G_0 is non-Gaussian, $\int x^2\, dG_0(x) < \infty$ and $\mathbf{P}_0 = \{P_{(\theta, G_0)}\colon \theta \in \Theta\}$ is regular. Then, if $P_0 = P_{(\theta_0, G_0)}$ satisfies the general Gaussian error model,*

$$\mathbf{L}_{P_0}\left(n^{1/2}\left(\hat{\beta}_{bn} - \beta(P_0)\right)\right) \to \mathbf{N}\left(0, I_b^{-1}(P_0)\right), \tag{2.26}$$

for all $P_0 \in \mathbf{P}_0$.

(ii) *If also $n\nu_n^{-6} \log n \to 0$, the convergence in* (2.26) *continues to hold if P_0 is replaced by $P_n = P_{(\theta_n, G_n)}$, where*

$$\theta_n = (\beta_n, \alpha_n, \gamma_{1n}, \gamma_{2n}, \sigma_{11n}) \to \theta = (\beta, \alpha, \gamma_1, \gamma_2, \sigma_{11})$$

and $G_n \to G$ weakly and $\int z^2 G_n(dz) \to \int z^2 G(dz) < \infty$.

(iii) *Write* (2.21)–(2.23) *as $\hat{\beta}_n = \hat{\beta}_n(\tilde{\beta}_n)$ and let $\hat{\beta}_{0n} = \tilde{\beta}_n$, $\hat{\beta}_{in} = \hat{\beta}_n(\hat{\beta}_{i-1,n})$, $i = 1, 2, 3, \ldots$. Then, for $i \geq 1$, all $\hat{\beta}_{in}$ are efficient and $|\hat{\beta}_{in} - \hat{\beta}_{i-1,n}| = o_p(n^{-1/2})$ for all $i \geq 2$.*

(iv) *If $\hat{\beta}_n$ is replaced by $\hat{\beta}_{an}$ and the restricted Gaussian error model is considered then claims* (i)–(iii) *continue to hold with I_b replaced by I_a.*

NOTES.

(1) Let $K \subset \mathbf{P}$ be compact in the total variation norm topology. Part (ii) of the theorem shows that the convergence in (2.24) is uniform over K if $P \to I_b(P)$ is continuous on K. These are the largest sets over which we may expect uniform convergence.

(2) Part (iii) of the theorem may be interpreted in terms of running the iteration $\hat{\beta}_{in}$ to convergence. Suppose the stopping rule is of the form: Stop as soon as $|\hat{\beta}_{in} - \beta_{i-1,n}| \leq \varepsilon_n$, where $\varepsilon_n \downarrow 0$, $n^{1/2}\varepsilon_n > c > 0$. This is reasonable since the random fluctuations in the estimate are of order $n^{-1/2}$. Then, by part (iii), with probability tending to 1 the iteration stops with $\hat{\beta}_{2n}$.

Under more stringent conditions on ν_n, c_n we conjecture that tedious calculations will show that, in fact, $\lim_i \hat{\beta}_{in}$ exists with probability tending to 1 and is efficient.

3. Information bounds and proof of Theorem 2.1. Let $\mathbf{P}_0$ be a regular parametric submodel of a model $\mathbf{P}$ written in the form $\{P_{(\beta,\gamma)}\colon \beta \in R,\ \gamma \in E \subset R^k\}$. Let $l(X,\beta,\gamma)$ denote the log likelihood of an observation from $P_{(\beta,\gamma)}$ and let $\dot{l}_0(X) = \partial l/\partial\beta|_{(\beta_0,\gamma_0)}$, $\dot{l}_j(X) = \partial l/\partial\gamma_j|_{(\beta_0,\gamma_0)}$, $1 \le j \le k$, where $\gamma = (\gamma_1,\ldots,\gamma_k)$. Begun, Hall, Huang and Wellner (1983) [see also Efron (1977) and Neyman (1957)] show (in slightly different terms) that, if $P_0 = P_{(\beta_0,\gamma_0)}$

$$I(P_0;\beta,\mathbf{P}_0) = \min\left\{E\left(\dot{l}_0(X) - \sum_{j=1}^{k} c_j\dot{l}_j(X)\right)^2 : (c_1,\ldots,c_k) \in R^k\right\}$$

$$= E\{[l^*]^2(X)\},$$

where

$$l^* = \dot{l}_0 - \sum_{j=1}^{k} c_j^*\dot{l}_j, \tag{3.1}$$

and the c_j^* are uniquely determined by the orthogonality condition

$$El^*\dot{l}_j(X) = 0, \qquad j = 1,\ldots,k. \tag{3.2}$$

Moreover, the efficient influence function for $\mathbf{P}_0$ is given by

$$\tilde{l}(X, P_0|\beta,\mathbf{P}_0) = l^*(X)/I(P_0;\beta,\mathbf{P}_0). \tag{3.3}$$

Therefore, to calculate $\tilde{l}$ for $\mathbf{P}_0$ we need only calculate the projection $\Sigma_{j=1}^{k} c_j^*\dot{l}_j(X)$, in $L_2(P_0)$, of $\dot{l}_0$ into $[\dot{l}_j\colon 1 \le j \le k]$, the linear span of $\dot{l}_1,\ldots,\dot{l}_k$. Let $\Pi(h|L)$ denote the projection of $h \in L_2(P_0)$ into a closed linear space $L \subset L_2(P_0)$.

To prove Theorem 2.1 we go through the following steps for the restricted Gaussian error model and an analogous series for the general Gaussian error model.

(i) Identify $(\gamma_1,\gamma_2) = (\alpha,\sigma^2)$, where σ^2 is given by (1.4) and let $\eta = (\eta_1,\ldots,\eta_{k-2})$ index G, i.e.,

$$\mathbf{P}_0 = \left\{P_{(\theta,G_\eta)}\colon \eta \in E,\ \theta = (\alpha,\beta,\sigma^2),\ \alpha,\beta \in R\right\}.$$

Calculate formally $\dot{l}_j$, $0 \le j \le k$, at $P_0 = P_{(\theta_0,G_{\eta_0})}$, where $j = 0 \leftrightarrow \beta$, $j = 1,2 \leftrightarrow \alpha,\sigma^2$, $j \ge 3 \leftrightarrow \eta$.

We project $\dot{l}_0$ into $[\dot{l}_j\colon j \ge 1]$ in two steps. First, calculate, for $0 \le j \le 2$, $\Pi(\dot{l}_j|V)$, where

$$V = \left[\dot{l}_j\colon j \ge 3\right],$$

$$l^* = \dot{l}_0 - \Pi\left(\dot{l}_0|V\right) - \Pi\left(\dot{l}_0 - \Pi\left(\dot{l}_0|V\right)|W\right), \tag{3.4}$$

where

$$W = \left[\dot{l}_j - \Pi\left(\dot{l}_j|V\right)\colon 1 \le j \le 2\right].$$

Claim (3.4) is well known and can be verified by checking (3.2). We establish that:

(ii) For any regular parametric submodel $\mathbf{P}_0$

$$\left[\dot{l}_j\colon j \geq 3\right] \subset \{a(T)\colon a(T) \in L_2(P_0),\ Ea(t) = 0\}$$

and then prove:

(iii) If $\mathbf{P}_0$ is given by (2.6), then $\mathbf{P}_0$ is regular and

$$\left[\dot{l}_j\colon j \geq 3\right] \supset \left[E\left(\dot{l}_0(X)|T\right)\right]. \tag{3.5}$$

The existence of a model $\mathbf{P}_0$ having property (3.5), but not the specific choice (2.6), follows from Theorem 14.3.12 of Pfanzagl (1982). Note that

$$E(h(X) - E(h(X)|T))a(T) = 0, \quad \text{for all } a(T),\ h \in L_2(P_0). \tag{3.6}$$

Now (ii) and (iii) imply that, for $\mathbf{P}_0$ given by (2.6),

$$\Pi\left(\dot{l}_i|V\right) = E\left(\dot{l}_i(X)|T\right), \qquad 0 \leq i \leq 2,$$

and hence by (3.4) if l_0^* is the l^* of $\mathbf{P}_0$ given by (2.6),

$$l_0^*(X) = \dot{l}_0(X) - E\left(\dot{l}_0(X)|T\right) - \sum_{j=1}^{2} d_j\left(\dot{l}_j(X)\right) - E\left(\dot{l}_j(X)|T\right), \tag{3.7}$$

with $\{d_j\colon 1 \leq j \leq 2\}$ determined by (3.2) for $j = 1, 2$. Take $\mathbf{P}_0$ to be any regular parametric submodel. By (ii) and (3.6)

$$El_0^*(X)\dot{l}_j(X) = 0, \qquad j \geq 3.$$

By (3.2)

$$El_0^*(X)\dot{l}_j(X) = 0, \qquad j = 1, 2.$$

Therefore,

$$\begin{aligned} &E(l^*(X))^2 - E(l_0^*(X))^2 \\ &\quad = E(l^*(X) - l_0^*(X))^2 + 2E(l_0^*(X)(l^* - l_0^*)(X)) \\ &\quad = E(l^*(X) - l_0^*(X))^2 \geq 0, \end{aligned} \tag{3.8}$$

since $l^* - l_0^* \in [\dot{l}_j\colon j \geq 1]$. We conclude that $\mathbf{P}_0$ given by (2.6) is least favorable.

PROOF OF THEOREM 2.1. For mnemonic convenience we write $\dot{l}_0 = l_\beta$ and $\dot{l}_j = l_\alpha$, l_{σ^2}, $l_{\sigma_{11}}$, etc., as appropriate.

Restricted Gaussian error model. (i) Differentiating (2.4) we get, for $\theta = \theta_0$, $G = G_0$,

$$\begin{aligned} l_\beta(\mathbf{X}) &= p^{-1}(\mathbf{X}, \theta, G)\int\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right)^{-1}\left(\sigma_{11}(Y - \alpha - \beta z) - \sigma_{12}(X - z)\right) \\ &\qquad \times zK(\mathbf{X}, z, \theta)G(dz) \\ &= \tilde{\sigma}^{-1}(\theta)\int\left(\frac{U}{\sigma(\theta)} + \frac{\beta\sigma_{11} - \sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}(T - z)\right) \\ &\qquad \times z\phi\left(\tilde{\sigma}^{-1}(\theta)(T - z)\right)G(dz)/\omega(T), \end{aligned} \tag{3.9}$$

since

$$X = T - \bar{\sigma}^{-1}(\theta)(\beta\sigma_{11} - \sigma_{12})U,$$
$$Y - \alpha = \beta T + \bar{\sigma}^{-1}(\theta)(\sigma_{22} - \beta\sigma_{12})U.$$

Similarly,

$$l_\alpha = [\omega(T)\tilde{\sigma}(\theta)]^{-1} \times \int\left(\frac{U}{\sigma(\hat{\theta})} + \frac{\beta\sigma_{11} - \sigma_{22}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}(T - z)\right)\phi\big(\tilde{\sigma}^{-1}(\theta)(T - z)\big)G(dz), \tag{3.10}$$

$$l_{\sigma^2} = \frac{1}{2\sigma^2}\Big((U^2 - 1) + \tilde{\sigma}^{-1}(\theta) \times \int\big(\tilde{\sigma}^{-2}(\theta)(T - z)^2 - 1\big)\phi\big(\tilde{\sigma}^{-1}(\theta)(T - z)\big)G(dz)/\omega(T)\Big). \tag{3.11}$$

(ii) Suppose $\mathbf{P}_0 = \{P_{(\theta, G_\eta)}\}$ is a regular submodel with $G_\eta \ll G_0 = G$. If $g_\eta = dG_\eta/dG$, $g_0 = 1$, and, formally,

$$\dot{l}_{j+2}(X) = \int\exp\left\{-\frac{\tilde{\sigma}^{-2}}{2}(T - z)^2\right\}\frac{\partial g_\eta}{\partial \eta_j}(z)G(dz)/\omega(T), \tag{3.12}$$

a function of T only. If $\dot{l}_{j+2}$ exists only in the Hellinger sense it is easy to check that $\dot{l}_{j+2}$ is an L_2 limit of functions of T and hence T measurable.

(iii) If $\mathbf{P}_0$ is given by (2.6),

$$\frac{\partial l}{\partial \mu}(X, \theta, G_\eta)\Big|_{\mu=0,\,\tau=1} = \omega^{-1}(T)\frac{\partial}{\partial \mu}\int\exp\left\{-\frac{\tilde{\sigma}^{-2}}{2}\left(T - \frac{(z - \mu)}{\tau}\right)^2\right\}G(dz) = \omega^{-1}(T)\int(T - z)\exp\left\{-\frac{\tilde{\sigma}^{-2}}{2}(T - z)^2\right\}G(dz), \tag{3.13}$$

$$\frac{\partial l}{\partial \tau}(X, \theta, G_\eta)\Big|_{\mu=0,\,\tau=1} = \omega^{-1}(T)\int z(T - z)\exp\big\{-\tilde{\sigma}^{-2}(T - z)^2\big\}G(dz). \tag{3.14}$$

The independence of U and T and $EU = 0$ yield from (3.9)

$$E(l_\beta|T) = \tilde{\sigma}^{-1}\frac{\beta\sigma_{11} - \sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\int z(T - z)\phi\big(\tilde{\sigma}^{-1}(T - z)\big)G(dz)/\omega_0(T),$$

which is proportional to $\partial l/\partial\tau$ as required. Therefore,

$$l_\beta - E(l_\beta|T) = \bar{\sigma}^{-1}U\left[\int\phi\big(\tilde{\sigma}^{-1}(T - z)\big)G(dz)\right]^{-1}\int z\phi\big(\tilde{\sigma}^{-1}(T - z)\big)G(dz).$$

From (2.5)

$$l_\beta - E(l_\beta|T) = \bar{\sigma}^{-1}U\left(T + \tilde{\sigma}^2\frac{\omega'}{\omega}(T)\right). \tag{3.15}$$

Similarly,

$$l_\alpha - E(L_\alpha|T) = \tilde{\sigma}^{-1}U,$$

$$l_{\sigma^2} - E(l_{\sigma^2}|T) = \frac{1}{2\tilde{\sigma}^2}(U^2 - 1).$$

Now, from (3.10) and (3.13)

$$l_\alpha - \Pi(l_\alpha|V) = l_\alpha - E(l_\alpha|T) \tag{3.16}$$

and necessarily by (ii)

$$l_{\sigma^2} - \Pi(l_{\sigma^2}|V) = l_{\sigma^2} - E(l_{\sigma^2}|T) + b(T). \tag{3.17}$$

Therefore, (3.17) is orthogonal to both (3.16) and (3.15) so that $d_2 = 0$. On the other hand, it is easy to see that $d_1 = E(T)$. From (3.7), (3.15) and (3.16) we obtain Theorem 2.1 for a restricted Gaussian model.

General Gaussian error model. We find after some computation

$$\begin{aligned} l_{\sigma_{11}} &= \alpha_{11}(U^2 - 1) + \beta_{11}U\frac{\omega'}{\omega}(T) + \gamma_{11}b(T), \\ l_{\sigma_{22}} &= \alpha_{22}(U^2 - 1) + \beta_{22}U\frac{\omega'}{\omega}(T) + \gamma_{22}b(T), \\ l_{\sigma_{12}} &= \alpha_{12}(U^2 - 1) + \beta_{12}U\frac{\omega'}{\omega}(T) + \gamma_{12}b(T), \end{aligned} \tag{3.18}$$

where

$$b(T) = \tilde{\sigma}^{-1}\int z^2\phi(\tilde{\sigma}^{-1}(T - z))G(dz)/\omega(T),$$

and the matrix $\begin{pmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{22} & \beta_{22} \\ \alpha_{12} & \beta_{12} \end{pmatrix}$ has dimension 2. Let $V = [l_\mu(\mathbf{x}), l_\tau(\mathbf{x})]$.

From (3.18) the linear span of $l_\alpha - E(l_\alpha|T)$, $l_{\sigma_{ij}} - \Pi(l_{\sigma_{ij}}|V)$, $i, j = 1, 2$, is

$$\left[U, U^2 - 1, U\frac{\omega'}{\omega}(T), c(T)\right], \tag{3.19}$$

where $c(T) = \Pi(b(T)|V)$. We find the projection of $l_\beta - E(l_\beta|T)$ on (3.19) by using the independence of U and T, $EU = 0$, $EU^2 = 1$. We obtain

$$\begin{aligned} &\tilde{\sigma}^{-1}(\Pi(UT|[U])) + \Pi\left(UT\middle|\left[U\frac{\omega'}{\omega}(T)\right]\right) + \tilde{\sigma}^2U\frac{\omega'}{\omega}(T) \\ &\quad = \tilde{\sigma}^{-1}UE(T) + \left(\tilde{\sigma}^2 - \frac{1}{I_0}\right)U\frac{\omega'}{\omega}(T), \end{aligned}$$

since $E(T(\omega'/\omega)(T) = -1$. We conclude that under the submodel (2.6), with Σ varying freely, l_b^* is the efficient score function. But clearly, $El_b^*(\mathbf{X})a(T) = 0$ for all $a(T) \in L_2(P_0)$ and, in view of (ii), the argument leading to (3.8) applies to l_b^* also and (2.6) is least favorable. □

4. Proof of Theorem 2.2 and miscellaneous results. We begin by studying β_n^*.

PROPOSITION 4.1. *If either*

$$\mathbf{L}_{P_0}(Y) = \mathbf{L}_{P_0}(\beta X) * \mathbf{N} \tag{4.1}$$

or

$$\mathbf{L}_{P_0}(X) = \mathbf{L}_{P_0}(Y/\beta) * \mathbf{N} \tag{4.2}$$

(*where* $\mathbf{N}$ *is a Gaussian law and* $*$ *denotes convolution*), *then* $|\beta| = |\beta_0|$ *or* G_0 *is Gaussian. If* $\beta = \beta_0$ *one of these relations holds.*

PROOF. Let ψ be the characteristic function of X'. The case $\beta_0 = 0$ is simple. Assume $\beta_0 \neq 0$. Without loss of generality, take $E_0(X) = E_0(Y) = 0$ and $\beta_0 = 1$. Suppose $|\beta| \neq 1$ and without loss of generality, take $|\beta| > 1$. Then (4.1) becomes

$$\psi(t) = \psi(\beta t)e^{at^2}, \tag{4.3}$$

for some a. Iterating (4.3) we get for all k, t

$$\psi(\beta^k t) = \exp\left(-at^2\frac{(\beta^{2k}-1)}{(\beta^2-1)}\right)\psi(t).$$

Putting $u = \beta^k t$ and letting $k \to \infty$,

$$\psi(u) = \exp\left(-au^2(\beta^2-1)^{-1}(1+o(1))\right)(1+o(1))$$

and we get G_0 Gaussian. The same argument works for (4.2). □

PROPOSITION 4.2. *Suppose that* $\mathbf{P}$ *consists of all probabilities satisfying the general Gaussian error model with* $\int x^2\,dG(x) < \infty$. *Then for every* $P_0 \in \mathbf{P}$

$$\lim_{M\to\infty}\lim_{n\to\infty} P_0\left[\sqrt{n}\,\tilde{\beta}_n - \beta(P_0)\right| \geq M\Big] = 0.$$

PROOF. Let

$$Z_n(y,\beta) = \sqrt{n}\left\{\left(\hat{F}_2(y) - F_2(y)\right) - \int\Phi\left(\frac{y-\beta x-\mu(\beta)}{\sigma(\beta)}\right)d\left(\hat{F}_1(x) - F_1(x)\right)\right\}$$

$$= \sqrt{n}\left\{\left(\hat{F}_2(y) - F_2(y)\right) + \operatorname{sgn}\beta\int\left(\hat{F}_1((y-\mu(\beta)-z\sigma(\beta))/\beta)\right.\right. \tag{4.4}$$

$$\left.\left. - F_1((y-\mu(\beta)-z\sigma(\beta))/\beta)\right)\phi(z)\,dz\right\},$$

where F_1, F_2 are the marginal distribution functions of X and Y under P_0 and $\mu(\beta)$, $\sigma(\beta)$ are obtained by substituting population for sample moments in (2.18) and (2.19). By strong approximation, e.g., Csörgő (1981), we can construct $Z(\cdot,\cdot)$, a mean 0 Gaussian process in $C([-\infty,\infty]\times[-\infty,\infty])$ such that

$$\sup_{y,\beta}|Z_n(y,\beta) - Z(y,\beta)| = o_p(1). \tag{4.5}$$

Let $\hat{Z}_n(\cdot,\cdot)$ be defined by replacing $\mu(\beta), \sigma(\beta)$ by $\hat{\mu}(\beta), \hat{\sigma}(\beta)$ in (4.4). For $\sigma(\beta) \geq \varepsilon$, the family of functions $x \to \Phi((y - \beta x - \mu(\beta))/\sigma(\beta))$ is uniformly bounded and equicontinuous. Moreover,

$$\sup\{\sigma^{-1}(\beta)\beta\phi((y - \beta x - \mu(\beta)/\sigma(\beta)) - \hat{\sigma}(\beta)\beta\phi(y - \beta x - \mu(\hat{\beta}))/\sigma(\hat{\beta})): \sigma(\beta) \geq \varepsilon\} \to_P 0.$$

From (4.4) we then conclude that

$$\sup_y \{|\hat{Z}_n(y,\beta) - Z_n(y,\beta)|: \sigma(\beta) \geq \varepsilon\} \to_P 0.$$

Now there exist $\varepsilon, \delta > 0$ such that $\inf\{\sigma(\beta): |\beta| \leq \delta\} \geq \varepsilon$ and so

$$\sup_y \{|\hat{Z}_n(y,\beta) - Z_n(y,\beta)|: |\beta| \leq \delta\} \to_P 0.$$

On the other hand, from (4.5)

$$\sup_y \{|\hat{Z}_n(y,\beta) - Z_n(y,\beta)|: \delta \leq |\beta|\} \to_P 0$$

and so

$$\sup\{|\hat{Z}_n(y,\beta) - Z(y,\beta)|\} \to_P 0. \tag{4.6}$$

Similarly,

$$\sup\{|\hat{Z}_n^*(x,\beta) - Z^*(x,\beta)|\} \to_P 0, \tag{4.7}$$

where

$$\hat{Z}_n^*(x,\beta) = \sqrt{n}\left(\hat{F}_1(x) - F_1(x) - \int \Phi\left(\frac{\beta x - y + \hat{\mu}(\beta)}{\hat{\sigma}(\beta)}\right) d(\hat{F}_2(y) - F_2(y))\right)$$

and Z^* is an appropriately defined Gaussian process. A weak consequence of (4.6) and (4.7) is that for all $\varepsilon > 0$,

$$\inf\{\Delta_n(\beta): \varepsilon \leq |\beta^2 - \beta_0^2|\} \to_P \infty$$

and

$$\Delta_n(\beta_0) = O_p(1).$$

Therefore, by Proposition 4.1

$$\min\{|\beta_n^*(0) - \beta_0|, |\beta_n^*(0) + \beta_0|\} \to_P 0.$$

Since $Y - \mu(\beta) - \beta X$ is normal if and only if $\beta = \beta_0$, we conclude that $\tilde{\beta}_n$ is consistent.

We need to distinguish several cases for $n^{1/2}$-consistency:

(a) $|\beta_0| \geq \frac{3}{2}\delta_0$, $\sigma^2(\beta_0) > 0$;
(b) $|\beta_0| \geq \frac{3}{2}\delta_0$, $\sigma^2(\beta_0) = 0$;
(c) $\frac{1}{2}\delta_0 \leq |\beta_0| < \frac{3}{2}\delta_0$;
(d) $|\beta_0| \leq \frac{1}{2}\delta_0$.

(a) Suppose also that $\mathrm{Var}(Y) > \beta_0^2 \mathrm{Var}(X)$. Then, by (4.4) and (4.5)

$$\Delta_n(\beta) = \int \left| \sqrt{n} \left(F_2(y) - \int \Phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x) \right) + Z(y, \beta) \right|^2 \phi(y)\, dy + Q_n(\beta),$$

where

$$\sup\{ |Q_n(\beta)| : |\beta - \beta_0| \le \varepsilon_n \} = o_p(1).$$

Now, under these conditions,

$$\frac{\partial}{\partial \beta} \int \Phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x)$$

$$= -\sigma^{-1}(\beta) \int \phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) (x - E(X) - \beta \sigma^{-2}(\beta) \times (y - \beta x - \mu(\beta)) \mathrm{Var}\, X)\, dF_1(x),$$

$$\frac{\partial}{\partial \beta} \int \Phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x) \Big|_{\beta_0} \tag{4.8}$$

$$= -\beta_0^{-1}((y - EY) f_2(y) + \mathrm{Var}\, Y f_2'(y)),$$

which cannot vanish identically as a function of y unless Y is normal (i.e., $\beta_0 = 0$ or G_0 is normal). Moreover, the derivative in (4.8) is bounded as a function of y and continuous in β. We can conclude that $\tilde{\beta}_n$ is $n^{1/2}$-consistent in this case. This follows since $\Delta_n(\beta_0) = O_p(1)$ and

$$\Delta_n(\tilde{\beta}_n) \ge \int (Z(y, \beta_0) + n^{1/2}(\beta_n - \beta_0) c(y))^2 \phi(y)\, dy + o_p(1),$$

where $c(y)$ is the derivative in (4.6). Unboundedness of $n^{1/2}(\tilde{\beta}_n - \beta_0)$ leads to a contradiction since $c(y)$ does not vanish identically.

Case (a) with $\mathrm{Var}(Y) < \beta_0^2 \mathrm{Var}(X)$ is dealt with similarly using Z^*.

(b) If $\sigma(\beta_0) = 0$, calculate (taking $\beta_0 > 0$)

$$\lim_{\beta \to \beta_0} (\beta - \beta_0)^{-1} \left[\int \phi\left(\frac{y - \beta x - \mu(\beta)}{\sigma(\beta)} \right) dF_1(x) - F_1\left(\frac{y - \mu(\beta_0)}{\beta_0} \right) \right]$$

$$= \lim_{\beta \to \beta_0} (\beta - \beta_0)^{-1} \int (F_1((y - \mu(\beta) - z\sigma(\beta))/\beta) - F_1((y - \mu(\beta_0))/\beta_0))\, d\Phi(z)$$

$$= -\frac{y - EY}{\beta_0^2} f_1((y - \mu(\beta_0))/\beta_0) - \mathrm{Var}\, X f_1'((y - \mu(\beta_0))/\beta_0).$$

Again this expression cannot vanish identically in y unless F_1 and hence G_0 is normal. Boundedness in y and continuity in β again hold. (i) and case (b) follow.

(c) In this range since $\tilde{\beta}_n$ is consistent, we are driven to minimizing either $\Delta_n(\beta, 0)$ or $\Delta_n(\beta, 2\delta)$. In the first case, we are minimizing at $|\beta_0| > \delta/2$ and get $n^{1/2}$-consistency. In the second case, after reparametrization, we again minimize at $\delta/2 \le \beta_0 \le 7\delta/2$ and again get $n^{1/2}$-consistency.

(d) In this range since $\tilde{\beta}_n$ is consistent, we minimize $\Delta_n(\beta, 2\delta)$ with probability tending to 1. But after reparametrizing this corresponds to minimizing at $\beta_0 \ge 3\delta/2$ and we again get $n^{1/2}$-consistency. □

NOTES.

(1) For cases (ii) and (iii) of Theorem 2.2 we need to check that convergence in our arguments holds uniformly for sequences with $\|P_n - P_0\| \to 0$, $\int x^2\, dG_n(x) \to \int x^2\, dG_0(x)$, where $\|\cdot\|$ is total variation. A careful examination of the argument shows that for consistency, we need only check that

$$\overline{Y} \to_{P_n} E_0(Y), \qquad \hat{\sigma}_x^2 \to_{P_n} \mathrm{Var}_{P_0}(X),$$
$$\overline{X} \to_{P_n} E_0(X), \qquad \hat{\sigma}_y^2 \to_{P_n} \mathrm{Var}_{P_0}(Y).$$

For $n^{1/2}$-consistency, the derivatives in (4.8) and (4.9) are now evaluated at $\beta_{0n} \leftrightarrow P_n$ and depend on the marginals of T, $F_{1n} \leftrightarrow P_n$ with $\|F_{1n} - F_{10}\| \to 0$ and $F_{10} \leftrightarrow P_0$ non-Gaussian. The derivatives still converge to that for F_{10} uniformly for β bounded and are bounded uniformly in y, since $\sup_n \int |x|\, dF_{1n} < \infty$. The argument can now be made at the limit F_{10} as before.

(2) Under the restricted Gaussian error model the same argument yields that $\tilde{\beta}_{na}$ is $n^{1/2}$-consistent.

We now proceed to study the correction term which gives efficiency.

PROPOSITION 4.3. *Whatever be* G_0

$$\left|\frac{\omega_0'}{\omega_0}(t)\right| \le \tilde{\sigma}^{-2}(\theta_0)\left(|t| + \int |\eta| G_0(d\eta)\right). \tag{4.9}$$

PROOF. By a standard Laplace transform theorem, writing $\tilde{\sigma}$ for $\tilde{\sigma}(\theta_0)$,

$$\frac{\omega_0'}{\omega_0}(t) = \tilde{\sigma}^{-2}\, \frac{\int (\eta - t)\phi\big(\tilde{\sigma}^{-1}(t-\eta)\big) G_0(d\eta)}{\int \phi\big(\tilde{\sigma}^{-1}(t-\eta)\big) G_0(d\eta)},$$

$$\left|\int (\eta - t)\phi\big((\tilde{\sigma}^{-1}(t-\eta))\big) G_0(d\eta)\right| \le \int |\eta - t| \phi\big((\tilde{\sigma}^{-1}(t-\eta))\big) G_0(d\eta)$$
$$\le \int |\eta - t| G_0(d\eta) \int \phi\big((\tilde{\sigma}^{-1}(t-\eta))\big) G_0(d\eta),$$

by an inequality of Chebyshev [Hardy, Littlewood and Pólya (1952), page 43] since $\phi(t)$ is decreasing for $t \ge 0$. □

PROPOSITION 4.4. *Suppose* $H_n \to H$ *weakly and* $\int x^2\, dH_n(x) \to \int x^2\, dH(x)$. *Then*

$$I(H_n * \Phi) \to I(H * \Phi),$$

where I denotes Fisher information for location.

PROOF. By dominated convergence for all t

$$H_n * \phi(t) \to H * \phi(t),$$
$$[H_n * \phi]'(t) \to [H * \phi]'(t).$$

By Proposition 4.3

$$\frac{|[H_n * \phi]'(t)|^2}{[H_n * \phi]} \le V(t, H_n), \tag{4.10}$$

where

$$V(t, H) = 4[H * \phi](t)\left(t^2 + \int \eta^2 H(d\eta)\right).$$

But

$$V(t, H_n) \to V(t, H) \quad \text{for all } t$$

and

$$\int V(t, H_n)\, dt = 8\int \eta^2 H_n(d\eta) + 4 \to 8\int \eta^2 H(d\eta) + 4 = \int V(t, H)\, dt.$$

The sequence in (4.10) is uniformly integrable and the result follows. □

PROPOSITION 4.5. *Let*

$$\omega_{0n}(t) = \int \omega_0(t - \sigma_n s)\lambda(s)\, ds + c_n. \tag{4.11}$$

Then if we write T_i *for* $T_i(\theta_0)$,

$$E\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_1) - \frac{\omega_{0n}'}{\omega_{0n}}(T_1)\right)^2 \to 0, \tag{4.12}$$

$$E\left(\frac{\omega_{0n}'}{\omega_{0n}}(T_1) - \frac{\omega_0'}{\omega_0}(T_1)\right)^2 \to 0. \tag{4.13}$$

PROOF. We repeatedly use the inequalities

$$|\omega_{0n}^{(i)}| \le \sigma_n^{-i}\omega_{0n}, \qquad \omega_{0n} \le \sigma_{0n}^{-1}.$$

Write

$$\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_1)\frac{\omega_{0n}'}{\omega_{0n}}(T_1) = \frac{n^{-1}\Sigma_{j=1}^n\left[\lambda_n'(T_1 - T_j) - \omega_{0n}'(T_1)\right]}{\hat{\omega}_n}$$
$$- \frac{\omega_{0n}'(T_1)}{\omega_{0n}\hat{\omega}_n}\left(\frac{1}{n}\sum_{j=1}^n \lambda_n(T_1 - T_j) - \omega_{0n}(T_1)\right).$$

The first term has L_2 norm bounded by

$$c_n n^{-1/2} E^{1/2}\left([\lambda_n']^2(T_1 - T_2)\right) = O\left(c_n^{-1}\sigma_n^{-2}n^{-1/2}\right).$$

The second term is similarly norm bounded by

$$O\left(c_n^{-1}\sigma_n^{-2}n^{-1/2}\right)$$

and (4.12) follows.

For (4.13) note that, for all t, by dominated convergence,

$$\frac{\omega'_{0n}}{\omega_{0n}}(t) \to \frac{\omega'_0}{\omega_0}(t). \tag{4.14}$$

Without loss of generality, take $\tilde{\sigma}(\theta_0) = 1$. Then

$$\omega_{0n}(t) = \int \phi(t-\eta)\, d(G_0 * \lambda_n)(\eta) + c_n,$$

$$\omega'_{0n}(t) = \int \omega'_0(t - \sigma_n s)\lambda(s)\, ds.$$

By Proposition 4.3 we get

$$\frac{[\omega'_{0n}]^2}{\omega_{0n}^2}(t) \le 2\left(t^2 + \int \eta^2\, dG_{s0} * \lambda_n(\eta)\right).$$

But

$$\int t^2\omega_0(t)\, dt < \infty,$$

so that by dominated convergence and (4.14)

$$\int \left(\frac{\omega'_{0n}}{\omega_{0n}}\right)^2 (t)\omega_0(t)\, dt \to \int \frac{[\omega'_0]^2}{\omega_0}(t)\, dt.$$

L_2 convergence of ω'_{0n}/ω_{0n} to ω'_0/ω_0 follows. □

PROPOSITION 4.6. *For sequences* $\{P_n\}, \{c_n\}, \{\nu_n\}$ *as in Theorem* 2.2(ii), *and all* M *finite,*

$$\sup\left\{\left| n^{-1/2} \sum_{i=1}^{n} U_i(\theta)\left(\frac{\hat{\omega}'_n}{\hat{\omega}_n}(T_i(\theta), \theta) - \frac{\omega'_{0n}}{\omega_{0n}}(T_i(\theta))\right)\right| : n^{1/2}|\theta - \theta_{0n}| \le M\right\} \to_{P_n} 0, \tag{4.15}$$

where $\theta_{0n} \leftrightarrow P_{0n}$,

$$\sup\left\{n^{1/2}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{U_i(\theta)\left(T_i(\theta) - E_{P_n}(T_i(\theta)) + I_{0n}^{-1}\frac{\omega'_{0n}}{\omega_{0n}}(T_i(\theta))\right)\right.\right.\right.$$

$$\left. - U_i(\theta_{0n})\left(T_i(\theta_{0n}) - E_{P_n}(T_i(\theta_{0n})) - I_{0n}^{-1}\frac{\omega'_{0n}}{\omega_{0n}}(T_i(\theta_{0n}))\right)\right\} \tag{4.16}$$

$$\left. + I_{bn}(\theta - \beta_{0n})\right| : n^{1/2}|\theta - \theta_{0n}| \le M\Bigg\} \to_{P_n} 0.$$

This proposition reduces the proof of case (ii) to establishing that if $U_i \triangleq U_i(\theta_{0n})$, $T_i \triangleq T_i(\theta_{0n})$

$$\text{(4.17)}\quad \mathbf{L}_{P_0}\left(n^{-1/2}\sum_{i=1}^{n}\left\{U_i\left(T_i - E_{P_0}(T_i)\right) + I_{0n}^{-1}\frac{\omega_{0n}'}{\omega_{0n}}(T_i)\right\}\right) \to \mathbf{N}\left(0, I_b^{-1}(P_0)\right)$$

and

$$\text{(4.18)}\quad n^{-1}\sum_{i=1}^{n}\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)^2(T_i) \to_{P_n} I_0(P_0),$$

$$\text{(4.19)}\quad n^{-1}\sum_{i=1}^{n}U_i^2\left(T_i + I_0^{-1}(P_0)\frac{\omega_{0n}'}{\omega_{0n}}(T_i)\right)^2 \to_{P_n} I_b(P_0).$$

All three claims follow since

$$\mathbf{L}_{P_n}(U_1, T_1) \to \mathbf{L}_{P_0}(U_1, T_1),$$
$$\frac{\omega_{0n}'}{\omega_{0n}}(t) \to \frac{\omega_0'}{\omega_0}(t), \quad \text{for all } t,$$

and $E_{P_n}(U_1^2)$, $E_{P_n}(T_1^2)$, $\int([\omega_{0n}']^2/\omega_{0n})(t)\,dt$ all converge to the appropriate limits under P_0. The last claim is a consequence of Proposition 4.4.

PROOF OF PROPOSITION 4.6. Denote the (random) functions in absolute values in (4.15) by

$$Q_n(\Delta), \quad \text{where } \Delta = (\theta - \theta_{0n})n^{1/2}.$$

Now

$$\text{(4.20)}\quad Q_n(0) \to_{P_n} 0$$

by Proposition 4.5.

Write

$$Q_{1n}(\Delta) = n^{-1}\sum_{i=1}^{n}T_i\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_i(\theta), \theta) - \frac{\omega_{0n}'}{\omega_{0n}}(T_i(\theta))\right),$$

$$Q_{2n}(\Delta) = n^{-1/2}\sum_{i=1}^{n}U_{in}\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_i(\theta), \theta) - \frac{\omega_{0n}'}{\omega_{0n}}(T_i(\theta))\right).$$

It is easy to see that for (4.15) we need only check that

$$\text{(4.21)}\quad \sup\left\{|Q_{in}(\Delta)|: |\Delta| \le M\right\} \to_{P_n} 0, \qquad i = 1, 2.$$

Throughout this calculation we write $\lambda_n = \lambda_{\nu_n}$ and repeatedly use

$$\hat{\omega}_n \ge c_n, \qquad \omega_{0n} \ge c_n, \qquad |\lambda_n^{(i)}| \le \nu_n^{-i}\lambda_n.$$

We begin with $i = 1$. Let

$$V_{1n}(\Delta) = \frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_1(\theta), \theta) - \frac{\omega_{0n}'}{\omega_{0n}}(T_1(\theta), \theta).$$

By Cauchy-Schwarz and uniform integrability of T_1^2 (as P_n varies), it is enough

to check that

$$E\sup_{\Delta}\big(V_{1n}(\Delta)\big)^2 = O\big(n^{-1}\nu_n^{-4}\big(c_n^{-2} + \log n\big)\big). \tag{4.22}$$

Note first that

$$|V_{1n}(0)| \le c_n^{-1}|\hat{\omega}_n'(T_1,\theta_{0n}) - \omega_{0n}'(T_1)| + c_n^{-1}\nu_n^{-1}|\hat{\omega}_n(T_1,\theta_{0n}) - \omega_{0n}(T_1)|. \tag{4.23}$$

Let $\hat{F}_n$ be the empirical distribution function of $T_1,\dots,T_n$ and F its expectation. Then

$$\begin{aligned}|\hat{\omega}_n'(t,\theta_{0n}) - \omega_{0n}'(t)| &= O\big(n^{-1}\sigma_n^{-2}\big) + n^{-1}\sum_{i=2}^{n}\big[\lambda_n'(t-T_i) - E\lambda_n'(t-T_i)\big]\\ &= O\big(n^{-1}\sigma_n^{-2}\big) + O(1)\int\big(\hat{F}_n(s) - F(s)\big)\lambda_n''(t-s)\,ds\\ &\le O\big(n^{-1}\sigma_n^{-2}\big) + O(1)\sup_s\big|\hat{F}_n(s) - F(s)\big|\int|\lambda''(s)|\,ds,\end{aligned}$$

where the 0 terms are nonstochastic and independent of t. A similar bound holds for the second term in (4.23) and hence

$$EV_{1n}^2(0) = O\big(n^{-1}\nu_n^{-4}c_n^{-2}\big). \tag{4.24}$$

Next we write

$$T_i(\theta) = T_i + \frac{a}{\sqrt{n}}U_i + \frac{b}{\sqrt{n}}T_i,$$

so that a, b are well defined functions of Δ and note that

$$\begin{aligned}\frac{\partial}{\partial a}V_{1n}(\Delta) = n^{-1/2}\Bigg\{&\frac{\Sigma(U_1-U_j)\lambda_n''\big(T_1(\theta)-T_j(\theta)\big)}{nc_n + \Sigma\lambda_n\big(T_1(\theta)-T_j(\theta)\big)} - \frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_1(\theta),\theta)\\ &\times\frac{\Sigma(U_1-U_j)\lambda_n'\big(T_1(\theta)-T_j(\theta)\big)}{nc_n + \Sigma\lambda_n\big(T_1(\theta)-T_j(\theta)\big)} - U_1\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_1(\theta))\Bigg\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&E\sup\left\{\left|\frac{\partial}{\partial a}V_{1n}(\Delta)\right| : (\Delta)\le M\right\}^2\\ &\le C(M)n^{-1}\nu_n^{-4}E\Big(\max_j\big(U_1-U_j\big)^2 + U_1^2\Big) = O\left(\frac{\log n}{n}\sigma_n^{-4}\right).\end{aligned} \tag{4.25}$$

Similarly, we can bound

$$\begin{aligned}\left|\frac{\partial}{\partial b}V_{1n}(\Delta)\right| \le n^{-1/2}\Bigg|&\frac{\Sigma(T_1-T_j)\lambda_n''\big(T_1(\theta)-T_j(\theta)\big)}{\hat{\omega}_n(T_1(\theta),\theta)}\\ &-\frac{\hat{\omega}_n'}{\hat{\omega}_n^2}(T_1(\theta),\theta)\Sigma|T_1-T_j|\lambda_n'\big(T_1(\theta)-T_j(\theta)\big)\\ &-T_1\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_1(\theta))\Bigg|.\end{aligned} \tag{4.26}$$

Representing $T_i = kT_i(\theta) + (c/\sqrt{n})U_i$, $k \to 1$, we can bound (4.26) by

$$An^{-1/2}\left\{\nu_n^{-2}\frac{\Sigma|T_1(\theta) - T_j(\theta)|\lambda_n(T_1(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_1(\theta) - T_j(\theta))} + n^{-1/2}\left|\frac{\partial V_{1n}}{\partial a}(\Delta)\right| + \nu_n^{-2}(|U_1| + |T_1|)\right\}.$$

Representing $T_i = kT_i(\theta) + (c/\sqrt{n})U_i$, $k \to 1$, we can bound (4.26) by

$$An^{-1/2}\left\{\nu_n^{-2}\frac{\Sigma|T_1(\theta) - T_j(\theta)|\lambda_n(T_1(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_1(\theta) - T_j(\theta))} + n^{-1/2}\left|\frac{\partial V_{1n}}{\partial a}(\Delta)\right| + \nu_n^{-2}(|U_1| + |T_1|)\right\},$$

for a constant A depending on M only. Since $\lambda_\nu(|t|)$ is decreasing, the first term in curly brackets is bounded using the Chebyshev inequality by

$$n^{-1}\sum|T_1(\theta) - T_j(\theta)|. \tag{4.27}$$

Since (4.27) is bounded by

$$B\{n^{-1}\sum(|T_j| + |T_1| + n^{-1/2}(|U_j| + |U_1|))\},$$

for B depending on M only, we obtain

$$E\sup\left\{\left|\frac{\partial V_{1n}}{\partial b}(\Delta)\right|^2 : |\Delta| \le M\right\} = O(n^{-1}\nu_n^{-4}\log n). \tag{4.28}$$

Combining (4.24), (4.25) and (4.28), we get (4.21) for $i = 1$.

The proof of (4.21) for $i = 2$ is similar, but more complicated using the almost independence of $U_i(\theta)$, $T_i(\theta)$.

First, since $\hat{\omega}_n(\cdot, \theta_0)$ does not depend on the U_i,

$$\begin{aligned} EQ_{2n}^2(0) &= EU_1^2E(V_{1n}^2(0)) \\ &= O(n^{-2}\nu_n^{-4}c_n^{-2}). \end{aligned} \tag{4.29}$$

Next,

$$\begin{aligned} \frac{\partial Q_{2n}}{\partial a}(\Delta) &= \frac{1}{n}\sum_j U_j^2\left(\left(\frac{\hat{\omega}_n'}{\hat{\omega}_n}\right)'(T_j(\theta), \theta_0) - \left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_j(\theta))\right) \\ &\quad + n^{-1}\sum_i U_i\frac{\Sigma_j U_j\lambda_n''(T_i(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_i(\theta) - T_j(\theta))} \\ &\quad - n^{-1}\sum_i U_i\frac{\hat{\omega}_n'}{\hat{\omega}_n}(T_i(\theta), \theta)\frac{\Sigma_j U_j\lambda_n'(T_i(\theta) - T_j(\theta))}{nc_n + \Sigma\lambda_n(T_i(\theta) - T_j(\theta))} \\ &= R_{1n}(\Delta) + R_{2n}(\Delta) + R_{3n}(\Delta), \quad \text{say.} \end{aligned}$$

By arguing as for (4.22)

$$\sup\{ER_{1n}^2(\Delta): |\Delta| \le M\} = O\big(n^{-1}\nu_n^{-6}(c_n^{-2} + \log n)\big).$$

The additional ν_n^{-2} comes from the third derivatives in λ_n we have to deal with.

To deal with R_{2n} and R_{3n}, note that we can define $c(\theta)$ such that the Gaussian random variable

$$\tilde{U}_i(\theta) = U_i + \frac{c(\theta)}{\sqrt{n}}(T_i - X_i') \tag{4.30}$$

is independent of $T_i(\theta)$. This follows since $T_i(\theta)$ is a linear combination of X_i' and the Gaussian variables U_i and $T_i - X_i'$, both of which are independent of X_i'. Using (4.30)

$$\begin{aligned} ER_{2n}^2(\Delta) &\le 4E\left(n^{-2}\sum_{i,j}\tilde{U}_i\tilde{U}_j(\theta)\frac{\lambda_n''(T_i(\theta) - T_j(\theta))}{\hat{\omega}_n(T_i(\theta),\theta)}\right)^2 \\ &\quad + 4E\left(n^{-2}\sum_{i,j}\big(U_iU_j - \tilde{U}_i\tilde{U}_j(\theta)\big)\frac{\lambda_n''(T_i(\theta) - T_j(\theta))}{\hat{\omega}_n(T_i(\theta),\theta)}\right)^2 \\ &= O\big(n^{-1}\nu_n^{-4}\big) + O\big(n^{-1}\log n\nu_n^{-4}\big), \end{aligned}$$

since

$$E\tilde{U}_i^2(\theta) = O(1),$$

$$E\max\big(\tilde{U}_i\tilde{U}_j(\theta) - U_iU_j\big)^2 = O\big(n^{-1}\log n\big).$$

We can bound $ER_{3n}^2(\Delta)$ similarly to get

$$\sup\left\{E\left(\frac{\partial}{\partial a}Q_{2n}(\Delta)\right)^2 : |\Delta| \le M\right\} = O\big(n^{-1}\nu_n^{-6}(c_n^{-2} + \log n)\big). \tag{4.31}$$

Finally, we need to study $(\partial/\partial b)Q_{2n}(\Delta)$. It is possible to pass from the bound on $E((\partial/\partial a)Q_{2n}(\Delta))^2$ to the bound on $E((\partial/\partial b)Q_{2n}(\Delta))^2$ as was done in the passing from the bound on $(\partial/\partial a)V_{1n}(\Delta)$ to the bound on $(\partial/\partial b)V_{1n}(\Delta)$. We conclude

$$\sup\left\{E\left(\frac{\partial}{\partial b}Q_{2n}(\Delta)\right)^2 : |\Delta| \le M\right\} = O\big(n^{-1}\sigma_n^{-6}(c_n^{-2} + \log n)\big). \tag{4.32}$$

If we combine (4.31) and (4.32) with (4.29), we get by the standard Billingsley–Chentsov fluctuation inequalities [Billingsley (1968)],

$$\sup\{|V_{2n}(\Delta)|: |\Delta| \le M\} = O_{P_n}\big(n^{-1}\sigma_n^{-6}(c_n^{-2} + \log n)\big).$$

The proof of (4.15) is complete.

We now prove (4.16). Let

$$W_n(\Delta) = n^{-1/2}\bar{\sigma}(\theta)\sum_{i=1}^n U_i(\theta)\left(T_i(\theta) - E_{P_n}(T_i(\theta)) + I_{0n}^{-1}\frac{\omega_{0n}'}{\omega_{0n}}(T_i(\theta))\right),$$

where $\theta = \theta_0 + \Delta n^{-1/2}$, $\Delta = (\Delta_1, \ldots, \Delta_4)$, $\Delta_1 = \beta$, etc. Claim (4.16) is equivalent to

$$\sup\left\{\left|W_n(\Delta) - W_n(0) - \sum_{j=1}^{4} \frac{\partial W_n(0)}{\partial \Delta_j}\Delta_j\right| : |\Delta| \le M\right\} \to_{P_n} 0 \tag{4.33}$$

and

$$\left|\frac{\partial W_n(0)}{\partial \Delta_j} - I_{bn}\bar{\sigma}(\theta_0)\delta_{1j}\right| \to_{P_n} 0, \qquad j = 1, \ldots, 4. \tag{4.34}$$

Now,

$$\frac{\partial W_n(0)}{\partial \Delta_1} = n^{-1}\sum_{i=1}^{n}\left[X_i\left(T_i + I_{0n}^{-1}\frac{\omega_{0n}'}{\omega_{0n}}(T_i)\right) + U_i(\gamma_1 U_i + \gamma_2(T_i - ET_i)) \times\left(1 + I_{0n}^{-1}\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)'(T_i)\right)\right],$$

for suitable γ_1, γ_2, the laws of the summands converge to $\mathbf{L}_0(A)$, where

$$A = X\left(T + I_0^{-1}\frac{\omega_0'}{\omega_0}(T)\right) + U(\gamma_1 U + \gamma_2(T - E_0 T))\left(1 + I_0^{-1}\left(\frac{\omega_0'}{\omega_0}\right)'(T)\right),$$

and the summands are uniformly integrable (P_n) by Proposition 4.4. Therefore,

$$\frac{\partial W_n}{\partial \Delta_1}(0) \to_{P_n} E_0(A) = I_b\bar{\sigma}(\theta_0),$$

after some computation. A similar argument establishes (4.34) for $j > 1$. For (4.33) we check that for $1 \le j \le k \le 4$,

$$\sup\left\{\left|\frac{\partial^2 W_n}{\partial \Delta_j\, \partial \Delta_k}(\Delta)\right| : |\Delta| \le M\right\} \to 0. \tag{4.35}$$

We give the argument for a typical term, $\Delta_3 \leftrightarrow \nu_1$,

$$\frac{\partial^2 W_n}{\partial \Delta_3^2} = n^{-3/2}\sum_{i=1}^{n}\sigma(\theta)U_i(\theta)I_{0n}^{-1}X_{in}^2\left(\frac{\omega_{0n}'}{\omega_{0n}}\right)''(T_i(\theta)). \tag{4.36}$$

Since $|\omega_{0n}^{(i)}/\omega_{0n}| \le \sigma_n^{-i}$, we bound (4.35) uniformly in $|\Delta| \le M$ by

$$n^{-1/2}\sigma_n^{-2}O(1)\left\{n^{-1}\sum_{i-1}^{n}|U_i|(T_i^2 + U_i^2) + n^{-3/2}\sum_{i=1}^{n}|T_i|^3\right\}. \tag{4.37}$$

Since T_i^2 are uniformly integrable under P_n,

$$n^{-1/2}\max_i|T_i| \to_{P_n} 0. \tag{4.38}$$

Claim (4.35) for $j = k = 3$ follows from (4.37) and (4.38). The other terms are dealt with similarly and the result follows.

Proposition 4.6 establishes claim (ii) of the theorem. For part (iii) note that Proposition 4.6 shows that if β_n^* is $n^{1/2}$-consistent so is $\hat{\beta}_n(\beta_n^*)$ and, in fact,

$$\hat{\beta}_n(\beta_n^*) = \beta_{0n} + n^{-1}\sum_{i=1}^{n} \tilde{I}_b(X_i, P_n) + o_{P_n}(n^{-1/2}).$$

Therefore, taking β_n^* successively as $\hat{\beta}_{0n}, \hat{\beta}_{1n}, \ldots$, we get

$$\hat{\beta}_{in} - \hat{\beta}_{1n} = o_{P_n}(n^{-1/2})$$

and claim (iii) follows. Claim (iv) is established in exactly the same way as claims (i)–(iii). □

PROPOSITION 4.7. *The efficiency of* $\hat{\beta}_P$ *under model* (*Identity*, Φ), I_c/I_a, *satisfies*

$$I_c/I_a \geq \left(1 + \sigma^2/(\beta^2 + 1)(\mathrm{Var}(X') + \sigma^2)\right)^{-1}.$$

PROOF.

$$I_a/I_c = [\mathrm{Var}(X')]^{-2}\mathrm{Var}(T)\left[\mathrm{Var}(T) - 2\tilde{\sigma}^2 + \tilde{\sigma}^4 I_0\right]$$

$$= 1 + \tilde{\sigma}^4(I_0\mathrm{Var}(T) - 1)/(\mathrm{Var}(X'))^2,$$

since $\mathrm{Var}(T) = \mathrm{Var}(X') + \tilde{\sigma}^2$. Since T is, in general, an inefficient estimate of η in the location model $T = \eta + \varepsilon$ we must have $\tilde{\sigma}^2 \geq I_0^{-1}$ so that

$$I_a/I_c - 1 \leq \tilde{\sigma}^4(\mathrm{Var}(T)/\tilde{\sigma}^2 - 1)/(\mathrm{Var}(X'))^2$$

$$= \tilde{\sigma}^2/\mathrm{Var}(X') = \sigma^2/(\beta^2 + 1)\mathrm{Var}(X')$$

and the result follows. □

Acknowledgment. We thank Cliff Spiegelman for helpful discussions.

REFERENCES

ANDERSON, T. W. (1984). Estimating linear statistical relationships. *Ann. Statist.* **12** 1–45.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

CSÖRGŐ, M. (1981). *Strong Approximations in Probability and Statistics*. Academic, New York.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565.

GLESER, L. (1981). Estimation in a multivariate "errors in variables" regression model: Large sample results. *Ann. Statist.* **9** 24–44.

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1952). *Inequalities*, 2nd ed. Cambridge Univ. Press.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

KENDALL, M. G. and STUART, A. (1979). *The Advanced Theory of Statistics* **2**, 4th ed. Hafner, New York.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

KOSHEVNIK, YU. A. and LEVIT, B. YA. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.* **21** 738–753.

NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics* (U. Grenander, ed.) 213–234. Wiley, New York.

NEYMAN, J. and SCOTT, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory. Lectures Notes in Statist.* **13**. Springer, New York.

REIERSØL, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica* **18** 375–389.

RUBIN, H. (1956). Uniform convergence of random functions with applications to statistics. *Ann. Math. Statist.* **27** 200–203.

SPIEGELMAN, C. (1979). On estimating the slope of a straight line when both variables are subject to error. *Ann. Statist.* **7** 201–206.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–195. Univ. of California Press.

COURANT INSTITUTE OF MATHEMATICAL SCIENCES
NEW YORK UNIVERSITY
251 MERCER STREET
NEW YORK, NEW YORK 10012

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM 91905
ISRAEL

The Annals of Statistics
1990, Vol. 18, No. 2, 925–938

ACHIEVING INFORMATION BOUNDS IN NON AND SEMIPARAMETRIC MODELS[1]

BY Y. RITOV AND P. J. BICKEL

The Hebrew University of Jerusalem and University of California, Berkeley

We consider in this paper two widely studied examples of nonparametric and semiparametric models in which the standard information bounds are totally misleading. In fact, no estimators converge at the $n^{-\alpha}$ rate for any $\alpha > 0$, although the information is strictly positive "promising" that $n^{-1/2}$ is achievable. The examples are the estimation of $\int p^2$ and the slope in the model of Engle et al. A class of models in which the parameter of interest can be estimated efficiently is discussed.

1. Introduction. Consider the standard simple random sampling model on a sample space $\mathbf{X}$: $X_1, \ldots, X_n$ i.i.d. according to $P \in \mathbf{P}$, a set of probability measures on $\mathbf{X}$ dominated by μ. Let p denote the density of P and $\theta: \mathbf{P} \to R$ be a parameter. Suppose $\mathbf{P}$ is a regular parametric model, that is,

1. $\mathbf{P} = \{P_{(\theta,\eta)}: \theta \in R, \eta \in R^m\}$, where if $s(\theta,\eta) = [dP_{(\theta,\eta)}/d\mu]^{1/2}$, the map $(\theta,\eta) \to s(\theta,\eta)$ is continuously Fréchet differentiable from R^{m+1} to $L_2(\mu)$, with derivative $\dot{s}(\theta,\eta)$ an $m+1$ vector of elements of $L_2(\mu)$.
2. The Fisher information matrix, $I(\theta,\eta) = 4[\int \dot{s}_i(\theta,\eta)\dot{s}_j(\theta,\eta)\,d\mu]_{(m+1)\times(m+1)}$ (where the $\dot{s}_i$ are the components of $\dot{s}$), is nonsingular.

Then it is known [see, for example, Hájek (1972)] that if θ is identifiable it can be estimated at rate $1/\sqrt{n}$. In fact, there exist $\hat{\theta}_n$ of "maximum likelihood" type which have the property that, if I^{11} is the first element of I^{-1}, then

$$\mathbf{L}_\theta \mathbf{X}\big(n^{1/2}(\hat{\theta} - \theta)\big) \to \mathbf{N}\big(0, I^{11}(\theta,\eta)\big)$$

uniformly on compact subsets of R^{m+1} and I^{11} is the smallest asymptotic variance achievable by uniformly converging estimates.

Levit (1978), Pfanzagl (1982) and Begun, Hall, Huang and Wellner (1983) have used an idea of Stein (1956) to extend those lower bounds to $\mathbf{P}$ nonparametric or semiparametric, provided that θ is pathwise Hellinger differentiable on $\mathbf{P}$.

In this paper we investigate the question: Under the conditions of the above authors, are the bounds necessarily sharp if we drop the restriction that $\mathbf{P}$ is a regular parametric model?

Received October 1987; revised July 1989.
[1]Research supported by ONR grant N00014-80-C-0163.
AMS 1980 *subject classifications.* 62G20, 62G05.
Key words and phrases. Rate of convergence, nonparametric estimations, functionals of a density.

We begin, in Section 2, by showing in the context of two widely studied examples, estimation of $\int p^2$, and of the regression coefficient in the model of Engle, Granger, Rice and Weiss (1986) that the answer is, in general, no. In fact, the rate $n^{-1/2}$ is not even achievable pointwise. Although the arguments are specific, they can evidently be generalized to show similar results for much broader classes of parameters. A general view of these phenomena is given in Donoho and Liu (1988).

In Section 3 we show that the information bounds are valid for a general class of semiparametric models. The class includes the regular parametric models and is rich enough to contain models having essentially any tangent space structure.

2. The bounds are not sharp. The first example we consider is

$$\mathbf{P} \equiv \{P \text{ on } [0,1]\colon P \text{ absolutely continuous with density } p \le M\},$$

where M is a finite constant and,

$$\theta(p) = \int p^2(x)\,dx.$$

Since the functional $\theta(p)$ is differentiable along every Hellinger path in $\mathbf{P}$, the regularity conditions required for validity of the information bound are satisfied. This functional appears in the asymptotic variance of the Hodges–Lehmann estimator. Similar functions (the integral of the square of the derivative of the density) appear in the theory of optimal density estimation.

It is well known [Pfanzagl (1982) and Donoho and Liu (1988)] that the information bound in this case is

$$4 \operatorname{Var} p(X) = 4\int (p(x) - \theta(p))^2 p(x)\,dx. \tag{2.1}$$

Hasminskii and Ibragimov (1979), following work of Schweder (1975), exhibit an estimate $\hat{\theta}_n$ such that $\sqrt{n}\,(\hat{\theta}_n - \theta(p))/2[\operatorname{Var} p(X)]^{1/2}$ converges in law to $\mathbf{N}(0,1)$ uniformly on $\{P$ with densities p such that $\|p\|_\infty + \|p'\|_\infty \le L\}$. Yet we can establish the following.

THEOREM 1. *For any $\varepsilon > 0$, there exists a subset $\mathbf{P}_0 \subset \mathbf{P}$ (compact in the topology induced by the variational norm and having diameter less than ε) such that for every sequence of estimators $\hat{\theta}_n$ and every $\alpha > 0$, there exists $P \in \mathbf{P}_0$ such that*

$$\liminf_n P\left[\left|\hat{\theta}_n - \theta\right| \ge n^{-\alpha}\right] > 0. \tag{2.2}$$

A consequence of this result is that the rate of convergence on $\mathbf{P}_0$, as defined, for example, by Stone (1980), is slower than $n^{-\alpha}$ for any $\alpha > 0$. In fact, no sequence of estimators which is $n^{-\alpha}$ consistent at each point of $\mathbf{P}_0$ exists. So the information bound is totally misleading for $\mathbf{P}$.

To see what goes wrong, we consider the behaviour of a plausible type of estimator. It is proved in Pfanzagl (1982)—see also Bickel, Klaassen, Ritov and Wellner (to which we refer in the sequel as BKRW)—that if $\hat{\theta}_{\text{eff}}$ is efficient, then

$$\hat{\theta}_{\text{eff}} = \theta(p) + 2n^{-1} \sum_{i=1}^{n} (p(X_i) - \theta(p)) + o_p(n^{-1/2}).$$

The naive approach to estimating θ efficiently is to try $\tilde{\theta} = \theta(\hat{p}_n) + 2n^{-1}\sum_{i=1}^{n}[\hat{p}_n(X_i) - \theta(\hat{p}_n)]$ for $\hat{p}_n$ an estimator of the density. For simplicity, suppose $\hat{p}_n(\cdot)$ is based on an auxiliary sample. If $\tilde{\theta} = \hat{\theta}_{\text{eff}} + o_p(n^{-1/2})$, we would expect

$$E(\tilde{\theta}|\hat{p}_n) = \int p^2(x)\,dx + O_p(n^{-1/2}).$$

But,

$$E(\tilde{\theta}|\hat{p}_n) - \int p^2(x)\,dx = 2\int \hat{p}_n(x)p(x)\,dx - \int \hat{p}_n^2(x)\,dx - \int p^2(x)\,dx$$

$$= -\int (\hat{p}_n(x) - p(x))^2\,dx.$$

According to Bretagnolle and Huber (1979), to have this last term be of order $n^{-1/2}$ uniformly for $p \in \mathbf{P}$ we need a Hölder condition of order at least $\frac{1}{2}$ on p in $\mathbf{P}$, viz. $|p(x) - p(y)| \leq c|x - y|^{1/2}$. A positive result when p is so restricted has been obtained by Ibragimov and Haminskii (1979). This argument cannot be translated into a proof since we have considered only estimates of a particular type in the discussion of the rate at which p can be estimated. In fact, a cleverer construction [see Bickel and Ritov (1988)] shows that a Hölder condition of order $\frac{1}{4}$ suffices. However, we hope the point is clear. The calculations leading to the information bound are local. They are irrelevant to actual performance if you can't even get to within $o_p(n^{-1/4})$ of $\theta(p)$.

We begin with a simpler construction which establishes the following.

THEOREM 2. *For any sequence of estimates $\hat{\theta}_n$ there exists a compact $\mathbf{P}_0$ for which the uniform rate of convergence is slower than a_n, for any sequence $a_n \to 0$, viz.*

$$\liminf_n \sup_{\mathbf{P}_0} P\left[\left|\hat{\theta}_n - \theta\right| \geq a_n\right] > 0. \tag{2.3}$$

Note that (2.3) *implies the existence of $\varepsilon > 0$ such that*

$$\liminf_n \sup_{p_0} P\left[\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right] > 0.$$

The main idea of the proof is a "Bayesian" construction. We exhibit a sequence of prior distributions π_n, assigning mass $\frac{1}{2}$ each to finite subsets H_{0n} of $\{P\colon \theta(P) = 1 + \frac{4}{3}a_n\}$ and H_{1n} of $\{P\colon \theta(P) = 1 + \frac{16}{3}a_n\}$, whose size $k(n)\uparrow\infty$ such that the posterior probabilities of H_{1n}, H_{0n} given $X_1, \ldots, X_n$ are, with

probability tending to 1, still equal to $\frac{1}{2}$. More explicitly, the members p_{jln}, $l = 1, \ldots, k(n)$, of H_{jn}, $j = 0, 1$, are equally likely a priori and are chosen so that, with probability tending to 1,

$$k^{-1}(n) \sum_{l=1}^{k(n)} \prod_{i=1}^{n} p_{0ln}(X_i) = k^{-1}(n) \sum_{l=1}^{k(n)} \prod_{i=1}^{n} p_{1ln}(X_i) = \prod_{i=1}^{n} p(X_i),$$

where p is the uniform distribution on $(0, 1)$ (though this is inessential). Define $\mathbf{P}_0$ to be this countable collection of P_{jlm}'s together with their limit, the uniform distribution. An immediate consequence from which (2.3) follows is that,

$$\inf_{\hat{\theta}_n} \int P\left[\left|\hat{\theta}_n - \theta\right| \geq a_n\right] \pi_n(dP) \to \tfrac{1}{2},$$

and this establishes the theorem. This construction differs from similar constructions appearing in the density estimation literature where the corresponding H_{0n}, H_{1n} are simple (consist of one point).

PROOF OF THEOREM 2. Here is the sequence of priors, the union of whose carriers is a set having the uniform distribution on $(0, 1)$ as its limit. We prescribe π_n through some auxiliary variables.

(1) Let

$$\alpha_n = \begin{cases} c_n, & \text{with probability } \frac{1}{2}, \\ 2c_n, & \text{with probability } \frac{1}{2}; \end{cases}$$

the sequence $c_n \downarrow 0$ is to be chosen later.

(2) Let $\Delta_0, \ldots, \Delta_m$, $m = n^3$, be independent identically distributed random variables independent of α_n and equal to ± 1 with probability $\frac{1}{2}$.

π_n is the distribution of the random density p given by

$$p\left((i + y)(m + 1)^{-1}\right) = 1 + \Delta_i \alpha_n h(y), \qquad i = 0, \ldots, m, 0 \leq y \leq 1,$$

where (say)

$$h(t) = \begin{cases} t, & 0 \leq t < \frac{1}{2}, \\ -(1 - t), & \frac{1}{2} \leq t \leq 1. \end{cases}$$

The support of each π_n is finite and $\int |p_\cdot - 1| \leq 2c_n$ with π_n probability 1, so the union of the supports of π_n is a sequence tending to the uniform distribution. Now, if P corresponds to the random p,

$$\theta(P) = \int p^2(x)\, dx = (m + 1)^{-1} \sum_{i=0}^{m} \int_0^1 (1 + \Delta_i \alpha_n h(y))^2\, dy = 1 + \frac{\alpha_n^2}{12}.$$

This construction, since $m = n^3$, has the property that the π_n probability that at most one of the observed $X_1, \ldots, X_n$ will fall into any of the intervals $[i/(m + 1), (i + 1)/(m + 1))$ is $1 - O(n^{-1})$. But one observation in a cell

gives no new information on whether $\alpha_n = c_n$ or $2c_n$ and so the posterior probability,

$$\pi_n\left\{\theta = 1 + \frac{c_n^2}{12}\middle| X_1, \ldots, X_n\right\} = \pi_n\left\{\theta = 1 + \frac{c_n^2}{3}\middle| X_1, \ldots, X_n\right\} = \frac{1}{2} + o_{\pi_n}(1). \tag{2.4}$$

Let $c_n = 3a_n^{1/2}$. Then (2.4) implies that

$$\inf_{\hat{\theta}} P\left(|\hat{\theta}_n - \theta| > a_n | X_1, X_2, \ldots, X_n\right) \to_{\pi_n} \tfrac{1}{2},$$

or, for any $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$,

$$\int P\left[|\hat{\theta}_n - \theta| \geq a_n\right] \pi_n(dP) \to \tfrac{1}{2}.$$

Then

$$\liminf_n \sup_{\mathbf{P}_0} P\left[|\hat{\theta}_n - \theta| > a_n\right] \geq \liminf_n \int P\left[|\hat{\theta}_n - \theta| > a_n\right] \pi_n(dP) = \tfrac{1}{2}$$

and (2.3) follows. To check (2.4), note that if at most one X_i falls in each interval, the posterior distribution of $(\alpha_n, \Delta_0, \ldots, \Delta_m)$ is

$$\begin{aligned} &\pi_n(\alpha, \Delta_0, \ldots, \Delta_m | X_1, \ldots, X_n) \\ &\quad = 2^{-(m+2)} \prod_{i=0}^{m} \left\{\frac{1 + \Delta_i}{2} f_\alpha^+(Y_i) + \frac{1 - \Delta_i}{2} f_\alpha^-(Y_i)\right\}^{\delta_i} c(X_1, \ldots, X_n) \\ &\quad = \prod_{i=0}^{m} \{1 + \Delta_i \alpha h(Y_i)\}^{\delta_i}, \end{aligned} \tag{2.5}$$

where

$$f_\alpha^{\pm}(y) = 1 \pm \alpha h(y),$$

$$\delta_i = \begin{cases} 1, & \text{if there exists } X_{j_i} \in \left[\dfrac{i}{m+1}, \dfrac{i+1}{m+1}\right), \\ 0, & \text{otherwise,} \end{cases}$$

and Y_i is the fractional part of $(m + 1)X_{j_i}$. By symmetry, from (2.5),

$$\pi_n(\alpha_n = c_n | X_1, \ldots, X_n) = \tfrac{1}{2}$$

and (2.4) follows. □

Theorem 1 again uses a Bayesian construction. For the conclusion we cannot reduce our problem from estimation to testing but have to construct a prior distribution with infinite support whose Bayes risk for the loss function $l_n(\theta, \hat{\theta}) = 1(|\hat{\theta} - \theta| \geq a_n)$ is bounded away from 0.

PROOF OF THEOREM 1. We exhibit a $\mathbf{P}_0$ contained in the ε ball around $\mathbf{U}(0,1)$ and π_0 concentrating on $\mathbf{P}_0$ such that for all $\alpha > 0$,

$$\liminf_n \inf_{\hat{\theta}_n} \int P\left[\left|\hat{\theta}_n - \theta\right| \geq n^{-\alpha}\right] \pi_0(dP) \geq \tfrac{1}{4}. \tag{2.6}$$

Then (2.2) follows. Otherwise, we could exhibit $\alpha > 0$, $\hat{\theta}_n$ such that for all P,

$$P\left[\left|\hat{\theta}_n - \theta\right| \geq n^{-\alpha}\right] \to 0,$$

which by dominated convergence would imply

$$\int P\left[\left|\hat{\theta}_n - \theta\right| \geq n^{-\alpha}\right] \pi_0(dP) \to 0,$$

contradicting (2.6). Here is π_0. Let $\alpha_k, \Delta_k(0), \ldots, \Delta_k(2^k - 1)$, $k = 1, 2, \ldots$ be independent, $\alpha_k = 0$ or 1 with probability $\frac{1}{2}$, each $\Delta_k(i) = \pm 1$ with probability $\frac{1}{2}$ each. Define the random functions

$$h_k(x) = \begin{cases} \Delta_k(i), & i2^{-k} \leq x < (i + \frac{1}{2})2^{-k}, \\ -\Delta_k(i), & (i + \frac{1}{2})2^{-k} \leq x < (i + 1)2^{-k}. \end{cases} \tag{2.7}$$

Finally, the random density p is given by

$$p(x) = 1 + \sum_{k=1}^{\infty} c_k \alpha_k h_k(x),$$

where the c_k are positive $\sum_{k=1}^{\infty} c_k < \varepsilon/2$. Note that since $\int h_i(x)\,dx = 0$, $\int h_i h_j(x)\,dx = \delta_{ij}$,

$$\begin{aligned} \theta(P) &= 1 + \sum_{i=1}^{\infty} \alpha_i^2 c_i^2 \\ &= 1 + \sum_{i=1}^{m-1} \alpha_i^2 c_i^2 + \sum_{i=m}^{\infty} \alpha_i^2 c_i^2. \end{aligned}$$

Let $\beta = (\alpha_1, \ldots, \alpha_{k-1})$ and $\pi_{0\beta}$ be the conditional distribution of all the α's and Δ's given β. For any bounded loss function $L(\theta, a)$,

$$\inf_{\delta} E_{\pi_0} L(\theta, \delta) = \inf_{\delta} \int E_{\pi_{0\beta}} L(\theta, \delta) \nu(d\beta) \geq \int \inf_{\delta} E_{\pi_{0\beta}} L(\theta, \delta) \nu(d\beta), \tag{2.8}$$

where δ ranges over all estimates of θ based on $X_1, \ldots, X_n$ and ν is the marginal distribution of β. Therefore, there exists a value β_0 of β such that the Bayes risk of π_0 is no smaller than the Bayes risk of $\pi_{0\beta_0} \equiv \pi_{00}$. Under π_{00}, if $m = [3 \log_2 n]$ any interval of the form $[i2^{-m}, (i + 1)2^{-m})$ contains at most one of $X_1, \ldots, X_n$ with probability $\geq 1 - (2n)^{-1}$. Arguing as before, under π_{00}, except on a set of probability $O(n^{-1})$ the conditional distribution of $\Delta \equiv \{\Delta_k(i)\colon 1 \leq i \leq 2^k, k \geq m\}$ given $X_1, \ldots, X_n$ is the same as the marginal distribution. We claim that the same is true of the conditional distribution of $\alpha = \{\alpha_k, \ldots, k \geq m\}$. Write the joint density of $(\alpha, \Delta, X_1, \ldots, X_n)$ with respect to the measure μ, where, under μ, the α_k's and $\Delta_k(i)$ have the distribution

specified earlier and $X_1, \ldots, X_n$ are independent of α, Δ and are uniform $(0, 1)$ as

$$\prod_{i=1}^{n}\left(1 + \sum_{k=1}^{m-1} c_k \alpha_{k0} h_k(X_i) + \sum_{k=m}^{\infty} c_k \alpha_k h_k(X_i)\right).$$

The posterior density, if at most one X_i is in each interval $[i/2^k, (i+1)/2^k)$, $k \geq m$, is proportional to

$$\prod_{i=1}^{n}\left(A_i(X_i) + \sum_{k=m}^{\infty} c_k \alpha_k \varepsilon_k(X_i)\Delta_{ki}\right),$$

where $A_i(x) = 1 + \sum_{k=1}^{m-1} c_k \alpha_{k0} h_k(x)$, $\Delta_{ki} = \Delta_k(j)$ iff j is such that $X_i \in [j2^{-k}, (j+1)2^{-k})$ and

$$\varepsilon_k(X_i) = \begin{cases} +1, & \text{if } X_i \in [j2^{-k}, (j + \tfrac{1}{2})2^{-k}), \\ -1, & \text{if } X_i \in [(j + \tfrac{1}{2})2^{-k}, (j+1)2^{-k}). \end{cases}$$

Then the posterior probability that $(\alpha_{m+1}, \ldots, \alpha_{m+t}) = (\alpha^0_{m+1}, \ldots, \alpha^0_{m+t})$ given $X_1 = x_1, \ldots, X_n = x_n$ is proportional to

$$E_\mu\left\{\prod_{i=1}^{n}\left(A_i(X_i) + \sum_{k=m+1}^{m+t} c_k \alpha^0_k \varepsilon_k(X_i)\Delta_{ki} + \sum_{k=m+t+1}^{\infty} c_k \alpha_k \varepsilon_k(X_i)\Delta_{ki}\right) \times 1(\alpha_{m+1} = \alpha^0_{m+1}, \ldots, \alpha_{m+t} = \alpha^0_{m+t})\right\}.$$

But the α_k and the Δ_{ki} are independent under μ. Multiplying out the product and using the symmetry of the Δ_{ki}, we obtain that the posterior probability is proportional to $\prod_{i=1}^{n} A_i(X_i)$ and our claim follows. To complete the argument note that, under π_{00} if $B_m = \sum_{k=m}^{\infty} c_k^2(\alpha_k^2 - \frac{1}{2})$,

$$P\left[B_m \geq \tfrac{1}{2}c_m^2\right] \geq P\left[\alpha_m = 1, \sum_{k=m+1}^{\infty} c_k^2(\alpha_k^2 - \tfrac{1}{2}) \geq 0\right] \geq \tfrac{1}{4}$$

by the symmetry and independence of α_m, and $\alpha_k^2 - \frac{1}{2}$, $k = m+1, \ldots$. A similar argument shows

$$P\left[B_m \leq -\tfrac{1}{2}c_m^2\right] \geq \tfrac{1}{4}.$$

Hence, if at most one X_i falls in each interval,

$$\begin{aligned} \inf_a P\left[|\theta - a| \geq \tfrac{1}{2}c_m^2 | X_1, \ldots, X_n\right] &\geq \min\left\{P\left[B_m \geq \tfrac{1}{2}c_m^2 | X_1, \ldots, X_n\right], P\left[B_m \leq -\tfrac{1}{2}c_m^2 | X_1, \ldots, X_n\right]\right\} \\ &\geq \tfrac{1}{4} + O_p(n^{-1}), \end{aligned}$$

since, except on a set of probability $O(n^{-1})$, the marginal and conditional distributions of B_m agree. So the Bayes risk of π_{00} for the loss function

$L_m(\theta, a) = 1[|\theta - a| \geq \frac{1}{2}c_m^2]$ is $\geq \frac{1}{4} + O(n^{-1})$. If $c_m = 9\varepsilon^2[\log n]^{-1-\varepsilon}$, say, then (2.6) follows from (2.8). □

In the model of Engle, Granger, Rice and Weiss (1986) we observe $X_i = (W_i, Z_i, Y_i)$, $i = 1, \ldots, n$, where

$$Y = \beta W + t(Z) + \varepsilon \tag{2.9}$$

and $\varepsilon \sim \mathbf{N}(0, \sigma^2)$. The joint distribution of (W, Z) and t are unknown. In recent work, Chen (1988) and Cuzick (1987) have exhibited, under various smoothness restrictions on t, estimates $\hat{\beta}$ which are asymptotically $\mathbf{N}(0, I^{-1}/n)$, where

$$I = \sigma^{-2}E(W - E(W|Z))^2 > 0 \tag{2.10}$$

unless W is a function of Z. Local calculations yield this as the information bound whenever $W \in L_2$. Let

$\mathbf{P} = \{\text{All distributions } (W, Z, Y) \text{ given by (2.9) such that } I > 0 \text{ and well defined}\}$.

THEOREM 3. (1) *Even if* $\sigma = 0$ [*or, equivalently, I given by* (2.10) *equals* ∞], *there exists a subset* $\mathbf{P}_0$ *of* $\mathbf{P}$ *such that for all estimates* $\hat{\beta}_n$,

$$\sup_{\mathbf{P}_0} P\left[|\hat{\beta}_n - \beta| \geq \varepsilon\right] > 0 \quad \textit{for any } \varepsilon > 0. \tag{2.11}$$

(2) *For* $\sigma > 0$ *there exists a compact subset* $\mathbf{P}_0$ *of* $\mathbf{P}$ *such that for all estimates* $\hat{\beta}_n$ *and all* $\gamma > 0$,

$$\liminf_n \sup_{\mathbf{P}_0} \left[|\hat{\beta} - \beta| \geq n^{-\gamma}\right] > 0.$$

We argue as for Theorem 2.

PROOF OF THEOREM 3. (1) We give the simpler construction for $\sigma = 0$ and $\mathbf{P}_0$ noncompact and sketch if for $\sigma > 0$ and $\mathbf{P}_0$ compact. Here is the prior π_n. Take $W = \pm 1$ with probability $\frac{1}{2}$ and $0 \leq Z \leq 1$.

Let $\alpha, \Delta_0, \ldots, \Delta_m$, $m = n^3$, be i.i.d. and equal to ± 1 with probability $\frac{1}{2}$. If $\alpha = -1$, then $\beta = 0$, $Z \sim \mathbf{U}(0, 1)$ independent of W and $t(z) \equiv 0$. If $\alpha = 1$, then $\beta = c$ and the conditional density of $Z|W$ and $t(\cdot)$ are given by

$$\begin{aligned} &p(z|w) = 1 - \Delta_i w, \quad t(z) = c\Delta_i, \quad \frac{i}{m+1} \leq z < \frac{i + 1/2}{m+1}, \\ &p(z|w) = 1 + \Delta_i w, \quad t(z) = -c\Delta_i, \quad \frac{i + 1/2}{m+1} \leq z < \frac{i+1}{m+1}. \end{aligned} \tag{2.12}$$

Again with probability $1 - O(n^{-1})$, the posterior of $\Delta_1, \ldots, \Delta_m$ is the same as the prior distribution. Note also by construction that $\beta W + t(Z) \equiv 0$. So, with

probability $1 - O(n^{-1})$,

$$P[\alpha = 1|W_i, Z_i, Y_i, i = 1, \ldots, n] = P[\alpha = 1|W_i, Z_i, i = 1, \ldots, n]$$

is proportional to,

$$E\left\{\prod_{i=1}^{n}(1 - \Delta_i W_i)^{\delta_i}(1 + \Delta_i W_i)^{1-\delta_i}\right\}, \tag{2.13}$$

where $W_1, Z_1, \ldots, W_n, Z_n$ are fixed. If Z_i falls in $[j_i/(m + 1), (j_i + 1)/(m + 1))$, we define $\delta_i = 1$ if Z_i is in the first half of that interval and 0 if it is in the second. The expectation in (2.13) is again 1 and we conclude that the posterior distribution of α is the same as its prior and hence that the Bayes risk of π_n is bounded away from 0. (2.11) follows.

(2) If $\sigma = 1$ (say), proceed as follows. Let $\alpha, \Delta_1, \ldots, \Delta_m$ be as above. Suppose $P[W = 0] = P[W = 1] = \frac{1}{2}$ and that the conditional distribution of Z given $W = 0$ is $\mathbf{U}(0, 1)$. Under π_n if $\alpha = -1$, $\beta = 0$ and Z given $W = 1$ is also $\mathbf{U}(0, 1)$. Let

$$t_n(z) = \begin{cases} a_n\Delta_i, & i/(m + 1) \le z < \left(i + \frac{1}{2}\right)/(m + 1), \\ -a_n\Delta_i, & \left(i + \frac{1}{2}\right)/(m + 1) \le z < i/(m + 1). \end{cases}$$

If $\alpha = 1$, $\beta = c_n$ and

$$p(z|W = 1)$$

$$= \begin{cases} 1 - b_n\Delta_i, & i/(m + 1) \le z < \left(i + \frac{1}{2}\right)/(m + 1), \\ 1 + b_n\Delta_i, & \left(i + \frac{1}{2}\right)/(m + 1) \le z < (i + 1)/(m + 1). \end{cases} \tag{2.14}$$

With probability $1 - O(n^{-1})$, there is at most one Z_i in each interval $[i(m + 1)^{-1}, (i + 1)(m + 1)^{-1})$. Conditional on that event, being given (W_i, Z_i, Y_i) is the same as being given (W_i, V_i, Y_i), where V_i is the fractional part of $(m + 1)Z_i$. Further, the posterior distribution of β is the same as the conditional distribution of β given $\{(V_i, Y_i)\colon W_i = 1\}$. Given $W_i = 1$, V_i is $\mathbf{U}(0, 1)$ by (2.14) since the conditional distribution of Δ_{ji} given $W_i = 1$, where $Z_i \in (j_i/(m + 1), (j_i + 1)/(m + 1))$, is the same as its prior.

Finally, the conditional density of Y_i given $W_i = 1$, V_i, $\alpha = 1$, is

$$\tfrac{1}{2}(1 - b_n)\phi(y - c_n - a_n) + \tfrac{1}{2}(1 + b_n)\phi(y - c_n + a_n)$$

$$= \phi(y) + y\phi(y)(c_n - a_n b_n) + O\left(c_n^2 + a_n^2\right).$$

If $a_n = c_n^{1-\delta}$, $b_n = c_n^{\delta}$, $\delta > 0$, the density of Y_i given $W_i = 1$, V_i, $\alpha = 1$ is $\phi(y)(1 + c_n^{2-2\delta}h(y) + O(c_n^3 + a_n^3))$, where $\int\phi(y)h(y)\,dy = 0$. One can show the joint distribution of $\{(V_i, Y_i)\colon W_i = 1\}$ under $\alpha = 1$ is contiguous to that under $\alpha = 0$ provided $c_n^{2-2\delta} = O(n^{-1/2})$. Hence, by taking $c_n = n^{-1/4+\varepsilon}$, $\varepsilon > 0$, arbitrary, we can deduce that β cannot be estimated at a rate better than $n^{-1/4+\varepsilon}$.

□

3. Validity of the bounds for a class of models. We consider semiparametric models with the following structure:

$$\mathbf{P} = \bigcup_{m=1}^{\infty} \mathbf{P}_m, \qquad \mathbf{P}_m \subset \mathbf{P}_{m+1}, \forall\, m, \tag{3.1}$$

and $\mathbf{P}_m$ regular parametric. That is, we can write

$$\mathbf{P}_m = \{P_{(\theta,\eta^m)}: \theta \in \Theta, \eta^m = (\eta_1, \ldots, \eta_{d-1}), \text{ with } d = d(m) \text{ and } \eta_j \in E_j, j = 1, \ldots, d-1, E_j, \Theta \text{ open subsets of } R\}.$$

1. $\mathbf{P} \ll \mu$.
2. The maps $(\theta, \eta^m) \to P_{(\theta,\eta^m)}$ are 1–1 for all m. Further, if $P \in \mathbf{P}_m = \mathbf{P}_m \cap \mathbf{P}_{m'}$, $m' > m$, then the first $d(m)$ coordinates of $\eta^{m'}$ agree with η^m.
3. The maps $(\theta, \eta^m) \to s(\theta, \eta^m) \equiv (dP_{(\theta,\eta^m)}/d\mu)^{1/2} \in L_2(\mu)$ are continuously Fréchet differentiable with derivative $\dot{s}(\theta, \eta^m) = (\dot{s}_1, \ldots, \dot{s}_d)(\theta, \eta^m)$, $\dot{s}_j \in L_2(\mu)$, $j = 1, \ldots, d$.
4. The information matrix,

$$I(\theta, \eta^m) \equiv 4\left[\int \dot{s}_i \dot{s}_j(\theta, \eta^m)\, d\mu\right]_{d\times d} = \left[E_{(\theta,\eta^m)} \dot{l}_i \dot{l}_j(\theta, \eta^m)\right]_{d\times d},$$

is nonsingular for all (θ, η^m), where $\dot{l}(\theta, \eta^m) = 2(\dot{s}/s)(\theta, \eta^m)$ is the derivative of the log likelihood.

In words, every member of $\mathbf{P}$ belongs to a nice parametric model whose dimension d can, however, be arbitrarily large. A moment's thought will show that most if not all semiparametric models proposed in the literature can be thought of as the closures (for weak convergence) of such $\mathbf{P}$. For example, the symmetric location model $\{P$: P is absolutely continuous on R, symmetric about some $\theta \in R\}$ is the closure of $\mathbf{P}$ as in (3.1), where $P_{(\theta,\eta^m)}$, for example, has

$$\log p_{(\theta,\eta^m)}(x) = h(x - \theta, \eta^m),$$

where

$$h''(x, \eta^m) = \sum_{k=1}^{d-1} \eta_k 1(|x| < b_{km}),$$

where $d = 2^m + 1$, $b_{km} = mk2^{-m}$, $k = 1, \ldots, d-1$. That is, we assume that the log density of $X - \theta$ is a symmetric quadratic spline with knots at $\pm b_{km}$, which is constant for $|x| > m$. Such models have been considered by Faraway (1987) and Stone (1986) among others. It is well known [see Le Cam (1956) and Bickel (1982)] that there exist estimates $\hat{\theta}_{mn}, \eta_{mn}$ which are efficient on $\mathbf{P}_m$. In particular,

$$\hat{\theta}_{mn} - \theta_0 = n^{-1} \sum_{i=1}^{n} \tilde{l}_{0m}(X_i) + o_{P_0}(n^{-1/2}), \tag{3.2}$$

where

$$\tilde{l}_{0m} = \frac{s^{-1}}{2} \frac{s_1^*}{\|s_1^*\|^2}$$

and

$$s_1^* = \dot{s}_1 - \Pi(\dot{s}_1 | [\dot{s}_2, \ldots, \dot{s}_d]),$$

$\Pi(h|L)$ denotes the projection of $h \in L_2(\mu)$ on the closed linear subspace L in the $L_2(\mu)$ norm, $\|\cdot\|$, and $[\dot{s}_2, \ldots, \dot{s}_d]$ is the linear span of $\{\dot{s}_2, \ldots, \dot{s}_d\}$. $\hat{\eta}_{mn} - \eta_0$ has a similar expansion but we can only note that

$$\hat{\eta}_{mn} - \eta_0 = O_{P_0}(n^{-1/2}). \tag{3.3}$$

These relations hold for each m fixed, all $P_0 \in \mathbf{P}_m$, as $n \to \infty$. Frequently, we achieve (3.2) and (3.3) using the maximum likelihood estimates of θ, η^m under $\mathbf{P}_m$. For any $P \in \mathbf{P}$, let $\eta = (\eta_1, \ldots, \eta_{d(P)})$, and $d(P)$ is the smallest m such that $P \in \mathbf{P}_m$. For the model $\mathbf{P}$, the information bound in estimating θ at $P_0 = P_{(\theta_0, \eta_0)}$ is given by

$$\|I^{-1}(P_0; \theta) = \tfrac{1}{4}\|\dot{s}_1 - \Pi\|(\dot{s}_1 | \dot{\zeta}_2(\theta_0, \eta_0))\|^{-2},$$

where

$$\dot{\zeta}_2(\theta_0, \eta_0) = \text{closure of the linear span of } \{\dot{s}_2(\theta_0, \eta_0), \ldots, \dot{s}_l(\theta_0, \eta_0), \ldots\}.$$

Here, for $m \geq m(P_0)$, we consider P_0 as a member of $\mathbf{P}_m$, i.e., corresponding to (θ_0, η_0^m) such that $P_0 = P_{(\theta_0, \eta_0^m)}$.

Suppose $I(P_0; \theta) > 0$ for all $P_0 \in \mathbf{P}$. Let

$$\tilde{l}(\theta_0, \eta_0) = 2s^{-1}(\theta_0, \eta_0)(\dot{s}_1(\theta_0, \eta_0) - \Pi(\dot{s}_1(\theta_0, \eta_0) | \dot{\zeta}_2(\theta_0, \eta_0))/I(P_0; \theta)) \tag{3.4}$$

be the efficient influence function for estimating θ in $\mathbf{P}$ at P_0; $\tilde{l}$ depends on (θ_0, η_0).

THEOREM 4. *Suppose that if* $P_{(\theta_k, \eta_k^m)} \in \mathbf{P}_m$, $\theta_k \to \theta_0$, $\eta_k^m \to \eta_0^m$, *then*

$$\Pi(v | \dot{\zeta}_2(\theta_k, \eta_k^m)) \to \Pi(v | \dot{\zeta}_2(\theta_0, \eta_0)) \tag{3.5}$$

for all $v \in L_2(\mu)$ *and*

$$\limsup_k \|\tilde{l}(\theta_k, \eta_k^m)\|_\infty < \infty, \tag{3.6}$$

where $\|\cdot\|_\infty$ *is the* sup *norm.*

Then there exists $\hat{\theta}_n$ *such that,*

$$\hat{\theta}_n = \theta_0 + n^{-1} \sum_{i=1}^n \tilde{l}_0(X_i) + o_{P_0}(n^{-1/2}),$$

where $\tilde{l} = \tilde{l}(\theta_0, \eta_0)$.

Moreover, the $\hat{\theta}_n$ are at least locally regular. That is, for all $P_0 \in \mathbf{P}$, $\{P_\tau: |\tau| < 1\}$ is a regular parametric submodel of $\mathbf{P}$, $\tau_n = O(n^{-1/2})$, we have $\mathbf{L}_{\tau_n}(n^{1/2}(\hat{\theta}_n - \theta(P_{\tau_n})))$ tending to a limit law independent of $\{P_\eta\}$.

The construction is essentially to pick the lowest dimensional submodel $\mathbf{P}_{\hat{m}_n}$ which is close enough to the empirical distribution, then treat $\hat{m}_n$ as fixed, compute the efficient estimate $\hat{\eta}_{\hat{m}_n n}$ of $\eta_{\hat{m}_n}$ in that model and then "solve the equation;"

$$\sum_{i=1}^{n} \tilde{l}(\theta, \hat{\eta}_{m_n n}) = 0. \tag{3.7}$$

The resulting estimate is well behaved if $P \in \mathbf{P}$. However, if $P \in \overline{\mathbf{P}} - \mathbf{P}$, we necessarily have $\hat{m}_n \to \infty$ and no guarantee that the solution of (3.7) is even consistent, much less efficient. In fact, the examples of the previous section make it clear that there is no hope for such a general consistency theorem. The question remains whether one can formulate reasonable conditions on the structure of $\tilde{l}$ and the behaviour of the distance in suitable metrics $\mathbf{P}_m$ and members of $\overline{\mathbf{P}} - \mathbf{P}$ as a function of m which yield the validity of the information bounds for members of $\mathbf{P}$. An attempt in this direction is the work of Severini and Wong (1987). However, we do not pursue this, in part, because we believe that the checking of any such conditions in models of interest will be at least as difficult as the construction of efficient estimates by one of a number of heuristic methods which have been developed—see BKRW, Chapter 7 for a discussion.

PROOF. Let d_K be the Kolmogorov distance between distributions. Let $\hat{\theta}_{mn}, \hat{\eta}_{mn}$ be as in (3.2) and (3.3) and let

$$\hat{P}_m \text{ be the corresponding member of } \mathbf{P}_m.$$

Let $\hat{m}_n$ be the first m such that $d_K(\hat{P}_m, P_n) \le \varepsilon_n$, where $\varepsilon_n \to 0$, $n^{1/2}\varepsilon_n \to \infty$, P_n is the empirical distribution. Evidently, if $m_0 = m(P_{(\theta_0, \eta_0)})$,

$$P_0[\hat{m}_n = m_0] \to 1.$$

Moreover, $\hat{P}_{\hat{m}_n} \leftrightarrow (\hat{\theta}_{\hat{m}_n n}, \hat{\eta}_{m_n n}) = (\theta_0, \eta_0) + O_{p_0}(n^{-1/2})$. Therefore, by (3.5),

$$\int \left(\tilde{l}(\theta_n, \hat{\eta}_{m_n n}) - \tilde{l}(\theta_n, \eta_n)\right)^2 s^2(\theta_n, \eta_n)\, d\mu = o_{p_0}(1), \tag{3.8}$$

for all sequences $P_{(\theta_n, \eta_n)} \in \mathbf{P}_{m_0}$ with $|\theta_n - \theta_0| = O(n^{-1/2})$, $|\eta_n - \eta_0| = O(n^{-1/2})$.

Moreover, using (3.6), we see that,

$$\begin{aligned}
&\int \tilde{l}(\theta_n, \hat{\eta}_{\hat{m}_n n}) s^2(\theta_n, \eta_n)\, d\mu \\
&\quad = 2\int \tilde{l}(\theta_n \hat{\eta}_{m_0 n})\big(s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n})\big) s\big(\hat{\theta}_n, \hat{\eta}_{m_0 n}\big)\, d\mu \\
&\qquad + O_{p_0}\big(\|s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n})\|^2\big) \\
&\quad = 2\int \tilde{l}(\theta_n, \hat{\eta}_{m_0 n})\big(\dot{s}_2(\theta_n, \hat{\eta}_{m_0 n}), \ldots, \dot{s}_{m_0}(\theta_m, \hat{\eta}_{m_0 n})\big) \\
&\qquad \times (\eta_n - \hat{\eta}_{m_0 n})' s\big(\hat{\theta}_n, \hat{\eta}_{m_0 n}\big)\, d\mu \\
&\qquad + o_{p_0}\big(|\eta_n - \hat{\eta}_{m_0 n}|\big) + O_{p_0}\big(\|s(\theta_n, \eta_n) - s(\theta_n, \hat{\eta}_{m_0 n})\|^2\big).
\end{aligned} \tag{3.9}$$

The first term on the right in (3.9) is 0 by (3.4). The last two terms are

$o_{P_0}(n^{-1/2})$ by (3.2) and (3.3), so

$$\int \tilde{l}(\theta_n, \hat{\eta}_{\hat{m}_{nn}}) s^2(\theta_n, \eta_n)\, du = o_{p_0}(n^{-1/2}). \tag{3.10}$$

Together, (3.8) and (3.10) yield the existence of $\hat{\theta}_n$—see Klassen (1987), for example. □

Thus the $\hat{\theta}_n$ are at least locally regular and $n^{1/2}(\hat{\theta}_n - \theta_0)$ is asymptotically normal $(0, I^{-1}(P_0; \theta))$, i.e., achieves the information bound.

NOTE. (1) Conditions (3.5) and (3.6) are trivially satisfied by the symmetric location example. Condition (3.6) can be interpreted as a robustness condition for efficient estimates in $\mathbf{P}_m$. That is, on the model $\mathbf{P}_m$, efficient influence functions are bounded and bounded uniformly in small Hellinger neighbourhoods of any P.

(2) It is easy to check that if in the model of Engle, Granger, Rice and Weiss we, for instance, let $\mathbf{P}_m$ be such that $t(Z)$ and $\log P(W = 1|Z)$ are representable as splines with $d(m)$ knots, condition (3.5) is satisfied. Although condition (3.6) fails for ε Gaussian, $\tilde{l}$ is of the form ε times functions which are uniformly $\|\cdot\|_\infty$ bounded and (3.7) continues to hold.

(3) A further peculiarity of these models is that, if we only consider the asymptotic behaviour of $\hat{\theta}_n$ at fixed (θ, η), it is asymptotically inadmissible. However, when we consider its behaviour over "contiguous" neighbourhoods in $\mathbf{P}$, it is uniquely asymptotically minimax. More precisely, let $\{P_t, |t| < 1\}$ be a regular parametric submodel of $\mathbf{P}$ passing through $P_0 = P_{(\theta_0, \eta_0)}$. Corresponding to this model is its score function at (θ_0, η_0) given by (say) $s_0^{-1}v$, where $v \in \dot{\zeta}_2(\theta_0, \eta_0)$. Consider $\hat{\hat{\theta}} \equiv \hat{\theta}_{\hat{m}_{nn}}$. By Le Cam's third lemma, if $\theta_n \equiv \theta_n(t) = \theta(P_{tn^{-1/2}})$, $\eta_n \equiv \eta_n(t) = \eta(P_{tn^{-1/2}})$, then

$$L_{(\theta_n, \eta_n)}\sqrt{n}\left(\hat{\hat{\theta}} - \theta_n\right) \to \mathbf{N}\left(2t\int v s_1^*\, d\mu,\ \tfrac{1}{4}\|s_1^*\|^{-2}\right). \tag{3.11}$$

On the other hand, by the same argument,

$$L_{(\theta_n, \eta_n)}\sqrt{n}\left(\hat{\theta} - \theta_n\right) \to \mathbf{N}\left(0, I^{-1}(P_0; \theta)\right).$$

Now,

$$I(P_0; \theta) = \tfrac{1}{4}\left\|\dot{s}_1 - \Pi\left(\dot{s}_1 | \dot{\zeta}_2(\theta_0, \eta_0)\right)\right\|^2$$

$$\leq \frac{\|s_1^*\|^2}{4}.$$

So, at (θ_0, η_0), i.e., $t = 0$, both $\sqrt{n}(\hat{\theta} - \theta_0)$ and $\sqrt{n}(\hat{\hat{\theta}} - \theta_0)$ are asymptotically normal with mean 0 and the asymptotic variance of $\sqrt{n}\,\hat{\hat{\theta}}$ is smaller than that of $\hat{\theta}$. However, evidently, on each parametric submodel, for any bounded bowl-shaped loss function l,

$$\liminf_M \liminf_n \sup\left\{E_{(\theta_n(t), \eta_n(t))} l\left(n^{1/2}\left(\hat{\hat{\theta}} - \theta_n\right)\right) : |t| \leq Mn^{-1/2}\right\} = \sup_d l(d),$$

higher than the comparable asymptotic minimax risk for $\hat{\theta}$.

This is a superefficiency phenomenon. The estimator $\hat{\hat{\theta}}$ is, in view of (3.11), not locally regular, i.e., the limit of $\mathbf{L}_{(\theta_n, \eta_n)}(\sqrt{n}(\hat{\hat{\theta}} - \theta_n))$ is not independent of t.

REFERENCES

BEGUN, J. M., HALL. W. J., HUANG, W. M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BICKEL, P. J. and RITOV, J. (1988). Estimating integrated squared derivatives. *Sankhyā*. To appear.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1989). Efficient and Adaptive Inference in Semiparametric Models. Forthcoming monograph, Johns Hopkins University Press, Baltimore.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: Risque minimax. *Z. Warsch. Verw. Gebiete* **47** 119–137.

CHEN, H. (1988). Convergence rates for the parametric component in a partially linear model. *Ann. Statist.* **16** 136–146.

CUZICK, J. (1987). Semiparametric additive regression. Technical Report, Imperial Cancer Research Laboratories, London.

DONOHO, D. L. and LIU, R. C. (1988). Geometrizing rates of convergence. Technical Report, Dept. Statist., Univ. California, Berkeley.

ENGLE, R. F., GRANGER, C. W. J., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

FARAWAY, J. J. (1987). Smoothing in adaptive estimation. Ph.D. dissertation, Univ. California, Berkeley.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math Statist. Prob.* **1** 175–194. Univ. California Press.

HASMINSKII, R. and IBRAGIMOV, I. A. (1979). On the nonparametric estimation of functionals. In *Proc. 2nd Prague Symp. Asymptotic Statist.* (P. Mandl and M. Huskova, eds.) 41–51. North-Holland, Amsterdam.

KLAASSEN, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15** 1548–1562.

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math Statist. Prob.* **1** 129–156. Univ. California Press.

LEVIT, B. Y. (1978). Infinite dimensional information bounds. *Theor. Probab. Appl.* **20** 723–740.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

SCHWEDER, T. (1975). Window estimation of the asymptotic variance of rank estimation of location. *Scand. J. Statist.* **2** 113–126.

SEVERINI, T. A. and WONG, W.-H. (1987). Profile likelihood and semiparametric models. Technical Report, Univ. Chicago.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 187–195. Univ. California Press.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1986). A nonparametric framework for statistical modeling. Technical Report, Dept. Statist., Univ. California, Berkeley.

DEPARTMENT OF STATISTICS
THE HEBREW UNIVERSITY OF JERUSALEM
JERUSALEM 91905
ISRAEL

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720