LIKELIHOOD INFERENCE IN THE ERRORS-IN-VARIABLES MODEL

By S.A. Murphy[1] And A.W. Van Der Vaart[2]

Pennsylvania State University and Free University Amsterdam

October 1995

1 Research partially supported by NSF grant DMS-9307255.

2 Research partially carried while on leave at Université de Paris-sud.

Running head: Errors-in-variables.

Corresponding author:
A.W. van der Vaart
Department of Mathematics
Free University
De Boelelaan 1081a
1081 HV Amsterdam
Netherlands

Abstract

We consider estimation and confidence regions for the parameters α and β based on the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ in the errors-in-variables model $X_i = Z_i + e_i$ and $Y_i = \alpha + \beta Z_i + f_i$ for normal errors e_i and f_i of which the covariance matrix is known up to a constant. We study the asymptotic performance of the estimators defined as the maximum likelihood estimator under the assumption that $Z_1, \ldots, Z_n$ is a random sample from a completely unknown distribution. These estimators are shown to be asymptotically efficient in the semi-parametric sense if this assumption is valid. These estimators are shown to be asymptotically normal even in the case that $Z_1, Z_2, \ldots$ are arbitrary constants satisfying a moment condition. Similarly we study the confidence regions obtained from the likelihood ratio statistic for the mixture model and show that these are asymptotically consistent both in the mixture case and in the case that $Z_1, Z_2, \ldots$ are arbitrary constants.

MSC 1991 subjectclassifications: 62G15, 62G20, 62F12, 62F25

Keywords and phrases: Errors-in-variables, Maximum Likelihood, Likelihood Ratio Test, Semi-parametric model, Mixture model, Donsker class, Asymptotic efficiency, Efficient score equation

1. Introduction and Main Result

Suppose we observe independent random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfying the model

$$X_i = Z_i + e_i$$
$$Y_i = \alpha + \beta Z_i + f_i.$$

Here $Z_1, Z_2, \ldots$ are unobservable, independent random variables with unknown distributions $\eta_1, \eta_2, \ldots$ independent from the unobservable, independent zero-mean bivariate normal variables (e_i, f_i) having covariance matrix $\sigma^2\Sigma_0$ for a known nonsingular matrix Σ_0 and unknown parameter σ^2. By a linear transformation it can be ensured that Σ_0 is the identity matrix. For simplicity we assume this throughout the paper.

This formulation of this errors-in-variables model covers two versions. In the *functional model* the sequence $Z_1, Z_2, \ldots$ are unknown constants $z_1, z_2, \ldots$ referred to as *incidental parameters* ; this corresponds to the submodel obtained by assuming the distributions η_j to be degenerate. In the *structural model* the sequence $Z_1, Z_2, \ldots$ is assumed to be a random sample from a fixed unknown distribution η; then the observations (X_i, Y_i) are a sample from the *mixture density*

$$p_{\theta,\gamma}(x, y) = \int \frac{1}{\sigma}\phi\Big(\frac{x-z}{\sigma}\Big)\,\frac{1}{\sigma}\phi\Big(\frac{y-\alpha-\beta z}{\sigma}\Big)\,d\eta(z).$$

Here $\theta = (\alpha, \beta)$, $\gamma = (\sigma, \eta)$ and ϕ is the standard normal density. Since in practice it is hard to decide which of these models is more relevant, it is useful to treat the two models at the same time. The terminology 'incidental' and 'structural' is due to Neyman and Scott (1948).

In this paper we are interested in obtaining point estimators and confidence regions for the regression parameter $\theta = (\alpha, \beta)$, considering the remaining parameters $(\sigma, \eta_1, \eta_2, \ldots)$ as nuisance parameters. As point estimators we propose the first coordinate $\hat{\theta}$ of the pair (θ, γ) that maximizes the function

$$(\theta, \gamma) \mapsto \prod_{i=1}^{n} p_{\theta,\gamma}(X_i, Y_i) \tag{1.1}$$

over all pairs (θ, γ) in the parameter set for the mixture version of the model, which we take to be $\mathbb{R}^2 \times [m, M] \times \mathcal{H}$ for known constants $0 < m < M < \infty$ and $\mathcal{H}$ being the set of all probability distributions on $\mathbb{R}$. In the structural version of the model this estimator is the maximum likelihood estimator, but in the functional version it is not. We shall show that the estimator is well-behaved in both versions of the model, provided the average $n^{-1}\sum_{i=1}^{n}\eta_j$ does not diverge to infinity. In fact we prove the following theorem.

THEOREM 1.1. *Assume that the sequence of distributions* $\bar\eta_n = n^{-1}\sum_{i=1}^n \eta_j$ *converges weakly to a distribution* η_0 *and satisfies* $\int |z|^{7+\delta}\, d\bar\eta_n(z) = O(1)$ *for some* $\delta > 0$. *Then the sequence* $\sqrt{n}(\hat\theta_n - \theta_0)$ *converges under* $(\theta_0, \sigma_0, \eta_1, \eta_2, \ldots)$ *in distribution to a normal distribution with mean zero and covariance matrix* $\tilde I^{-1}_{\theta_0,\gamma_0}$ *for* $\tilde I_{\theta,\gamma}$ *given by (3.1)-(3.2).*

To obtain confidence regions for θ we propose to invert the likelihood ratio test for the mixture model. Thus we define

$$L_n(\theta_0) = 2\log \frac{\sup_{\theta,\gamma} \prod_{i=1}^n p_{\theta,\gamma}(X_i, Y_i)}{\sup_{\gamma} \prod_{i=1}^n p_{\theta_0,\gamma}(X_i, Y_i)}$$

and take as confidence region the set of parameters θ such that $L_n(\theta)$ does not exceed the upper α-quantile of the chi-squared distribution with two degrees of freedom.

We can similarly obtain confidence sets for the slope parameter β alone. Define the statistics

$$K_n(\beta_0) = 2\log \frac{\sup_{\theta,\gamma} \prod_{i=1}^n p_{\theta,\gamma}(X_i, Y_i)}{\sup_{\alpha,\gamma} \prod_{i=1}^n p_{\alpha,\beta_0,\gamma}(X_i, Y_i)}.$$

As a confidence set for β take the set of all β_0 such that $K_n(\beta_0)$ does not exceed the upper α-quantile of the chi-squared distribution with one degree of freedom.

The following theorem implies that the asymptotic confidence level of these sets is $1-\alpha$, under both versions of the model.

THEOREM 1.2. *Assume that the sequence of distributions* $\bar\eta_n = n^{-1}\sum_{i=1}^n \eta_j$ *satisfies* $\int |z|^{7+\delta}\, d\bar\eta_n(z) = O(1)$ *for some* $\delta > 0$. *Then the sequences of statistics* $K_n(\beta_0)$ *and* $L_n(\theta_0)$ *converge under* $(\theta_0, \sigma_0, \eta_1, \eta_2, \ldots)$ *in distribution to chi-squared distributions with one and two degrees of freedom, respectively.*

It is reasonable to expect that some stability condition on the sequence $\eta_1, \eta_2, \ldots$ is necessary to obtain results of this type. The condition that the $7+\delta$-th absolute moments of the averages remain bounded is fairly weak, but probably more restrictive than necessary. The assumption in Theorem 1.1 that the sequence $\bar\eta_n$ converges to a limit is not necessary, but made for convenience. Inspection of our proofs shows that, given bounded $7+\delta$-th moments, the theorem is valid along every subsequence for which the averages converge.

Estimation and setting confidence regions appears to be particularly difficult in the incidental version of the model, since one has to deal with an increasing number of nuisance parameters. In an important sense our estimator of θ is preferable over the usual estimator (the maximum likelihood estimator in the incidental version of the model). In Section 6 we show that its asymptotic variance is strictly smaller than the asymptotic variance of the usual estimator, unless the empirical distribution of $z_1, z_2, \ldots$ approaches a normal distribution, in which case the procedures are

equivalent. The gain in efficiency depends on the limit of this empirical distribution and is shown to range between 0 and 20 % for reasonable 'designs'. In the mixture version of the model our estimator is asymptotically efficient in the semiparametric sense (cf. Begun, Huang, Hall and Wellner (1983) or Bickel, Ritov, Klaassen and Wellner (1993).) Estimators with a limiting behaviour as our estimator are generally also thought to be asymptotically efficient in an appropriate sense in the incidental version of the model. However, an appropriate definition of asymptotic efficiency in incidental models is not easy. See e.g. the discussions in Van der Vaart (1988, Section 5.4.2), and Pfanzagl (1993). Also see Gleser and Hwang (1987), whose results show that (uniform) finite sample confidence intervals of finite expected length are possible in the incidental model only by restricting the range of the parameter. Pfanzagl (1993), who gives counterexamples to show the difficulties, concludes with the advice to "scholars interested in applications" (page 1675) "to use estimator sequences which are asymptotically efficient among all S-regular estimator sequences [i.e. efficient in the mixture model], but make sure that these estimator sequences are asymptotically linear with a remainder term converging stochastically [in the incidental model] to zero [..]". Theorem 1.1 and its proof shows that the latter is true for the maximum likelihood estimator for the mixture model.

The situation as regards testing hypotheses about θ and setting confidence intervals is similar. The likelihood ratio tests proposed in this paper have a Pitman efficiency strictly bigger than one relative to the usual procedures unless the empirical distribution of $z_1, z_2, \ldots$ approaches a normal distribution, in which case the relative efficiency is one. The same improvement in asymptotic efficiency could be gained by using a Wald type test based on our estimator $\hat{\theta}$, but it is a general phenomenon that likelihood ratio based tests and confidence sets have better finite sample properties, presumably because they do not impose a-priori symmetry.

Models with incidental parameters were considered by Neyman and Scott (1948), who drew attention to the fact that the maximum likelihood estimators for the structural parameter (θ, σ) of the model, obtained by maximizing the likelihood over all parameters $(\theta, \sigma, z_1, \ldots, z_n)$, can be asymptotically inconsistent. In the present model the estimator for θ obtained in this manner is consistent, but the estimator for σ^2 converges to $\sigma^2/2$. The resulting estimator for θ appears to be the accepted procedure in the literature. See e.g. Kendall and Stuart (1979), Chapter 29 or Fuller (1987), Chapter 1, and also Section 7 of this paper. Kiefer and Wolfowitz (1956) showed that usually (in particular in our model) the maximum likelihood estimator $(\hat{\theta}, \hat{\sigma}, \hat{\gamma})$ in a structural (or mixture) model is consistent for the product of Euclidean and weak topology. They give an open-ended discussion of the practical relevance of the two types of models. In Section 2 we extend their consistency result to our general version of the model: it is shown that the distance between the maximizer $(\hat{\theta}, \hat{\sigma}, \hat{\eta})$ of (1.1)

and $(\theta, \sigma, \bar{\eta}_n)$ converges to zero. Thus in the case of incidental parameters our estimator $\hat{\eta}$ can be viewed as an estimator of the empirical distribution $n^{-1}\sum_{i=1}^n \delta_{z_i}$ of the incidental parameters. We prove asymptotic consistency under the weak condition that the sequence $n^{-1}\sum_{i=1}^n |z_i|^{2+\delta}$ remains bounded for some $\delta > 0$.

There is a large literature on the errors-in-variables model and its variations. Good starting points are Chapter 29 of Kendall and Stuart (1979) or Chapter 1 of Fuller (1987). Gleser (1981) gives a detailed derivation of the asymptotic properties of the standard estimators (in multivariate version of the model.) Anderson (1984) gives a long list of references and connections with other problems. We review the most relevant results in Section 6 of this paper, where we compare our procedures with the standard procedures. These standard procedures are inefficient from an asymptotic point of view. Efficient estimators for θ in the mixture model were first constructed by Bickel and Ritov (1987). They constructed a one-step estimator with the efficient score function estimated by using a kernel estimator. An extension of their result to models with incidental parameters is contained in Van der Vaart (1988a, 1988b). Since the maximum likelihood estimator does not require appropriate tuning of smoothing parameters, it seems preferable over these one-step estimators. In the case of a mixture model Van der Vaart (1995) proved the asymptotic normality of the maximum likelihood estimators $\hat{\theta}$ under a strong moment condition. Theorem 1.1 improves the moment condition, but, more importantly, extends his result to the incidental version of the model. Theorem 1.2 appears to have no precursors and the likelihood ratio procedure appears new, in particular for incidental models.

A different version of the errors-in-variables model is obtained by assuming that the covariance matrix Σ of the errors is completely unknown, but the mixing distribution (or the limit of the sequence $n^{-1}\sum \eta_j$) is not Gaussian. Then the parameters are still identifiable (Reiersøl (1950)), but our results have no bearing on the asymptotic behaviour of their maximum likelihood estimators. The technical reason is that the estimators for θ and σ are no longer orthogonal, so that it is necessary to consider (θ, σ) jointly. However our approach can, at present, not handle σ, because the efficient score equation ((3.4) in Section 3) for σ appears to fail and we have not been able to show that it is sufficiently small.

A negative aspect of the estimator and confidence intervals considered in this paper is a stronger dependence on the Gaussian error structure. While the standard procedure for estimation can be motivated by a least squares criterion and therefore yields asymptotically normal estimators under just moment conditions, our procedures use the fact that the variables $X_i + \beta(Y_i - \alpha)$ are sufficient for η_i, which is true for Gaussian errors, but not in general. It may be remarked that in the literature the Gaussian assumption is often made and rarely contested. Furthermore, Gaussianity is essential for the standard (exact) procedure to set confidence intervals (cf. Sec-

tion 6). In practice one will have to weigh the gain in efficiency (which depends on η_0) against one's belief in the normality of the errors. See Spiegelman (1979) for a further discussion of the non-Gaussian case.

Another disadvantage of our proposal is the computational complexity, in particular to compute the maximum likelihood estimator for the mixing distribution. However this problem has been investigated by a number of authors. Lindsay (1983b) has shown that for every fixed (θ, σ) the likelihood is maximized with respect to η by a discrete distribution $\hat{\eta}_n(\theta, \sigma)$ having at most n support points. Several algorithms to compute these suppport points and the corresponding weights are reviewed in Lesperance and Kalbfleisch (1992). Since computing the maximum likelihood estimator for the mixing distribution in our problem is equivalent to computing the maximum likelihood estimator in a normal deconvolution problem, the convex minorant algorithm considered by Groeneboom (1991) and Jongbloed (1995) can be used as well. The maximum likelihood estimator $(\hat{\theta}, \hat{\sigma})$ can be calculated by maximizing the profile likelihood $(\theta, \sigma) \mapsto \mathrm{lik}\big(\theta, \sigma, \hat{\eta}_n(\theta, \sigma)\big)$ or, preferably, by building an updating procedure for initial estimators for (θ, σ) into the iteration steps for computing the mixing distribution.

The paper is organized as follows. In Section 2 we derive the consistency of our estimator. Section 3 contains a discussion of least favorable submodels and an outline of the proofs of Theorems 1.1 and 1.2. The proofs of Theorems 1.1 and 1.2 are given in Sections 4 and 5. In Section 6 we conpute the relative efficiencies of our procedures and the standard procedures. In particular we derive the asymptotic power of the likelihood ratio test on which our confidence sets are based. Section 7 is an appendix and contains a number of technical lemmas.

2. Consistency

Kiefer and Wolfowitz (1956, Section 4) show that the maximum likelihood estimator $(\hat{\theta}, \hat{\gamma})$ in the mixture version of the errors-in-variables with free covariance matrix Σ is consistent for the product of the Euclidean topology and the weak topology (provided η_0 is not normal). Their proof can also be applied to the mixture model with Σ known up to a constant. At first one might expect that the resulting estimator, which is defined as the maximum likelihood estimator for the mixture model, will behave erratically in the functional version of the model. This is not true: in the functional version of the model the estimator $\hat{\eta}$ may be considered an estimator for the empirical measure $n^{-1}\sum_{i=1}^{n}\delta_{z_i}$ of the incidental parameters.

THEOREM 2.1. *Assume that the sequence of distributions $\bar{\eta}_n = n^{-1}\sum_{i=1}^{n}\eta_j$ satisfies $\int |z|^{2+\delta}\, d\bar{\eta}_n(z) = O(1)$ for some $\delta > 0$. Then $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, $\hat{\sigma}_n \xrightarrow{\mathrm{P}} \sigma_0$ and $d(\hat{\eta}_n, \bar{\eta}_n) \xrightarrow{\mathrm{P}} 0$ under $(\theta_0, \sigma_0, \eta_1, \eta_2, \ldots)$ for d a distance that generates the weak topology.*

Proof. It is clear from the form of the likelihood that $(\hat{\theta}_n, \hat{\sigma}_n, \hat{\eta}_n)$ is also the point of maximum if the parameter space for η is enlarged to all subprobability measures on $\mathbb{R}$. The latter set is compact and metrizable for the vague topology and the vague topology restricted to the set of probability measures is identical to the weak topology. Thus it suffices to show that $d(\hat{\eta}_n, \bar{\eta}_n) \xrightarrow{\mathrm{P}} 0$ in this setting.

We adapt the proofs of Wald (1949) and Kiefer and Wolfowitz (1956), sketching only the main steps. Assume without loss of generality that the sequence $\bar{\eta}_n$ converges weakly to a limit η_0; otherwise argue along subsequences. Compactify the parameter set for θ to $\bar{\mathbb{R}}^2$, defining $p_{\theta,\gamma}$ to be identically zero if $\theta \notin \mathbb{R}^2$. The parameter (θ_0, γ_0) is identifiable in the mixture model. This implies that

$$\int \log \frac{p_{\theta,\gamma}}{p_{\theta_0,\gamma_0}}\, p_{\theta_0,\gamma_0}\, d\lambda^2 < 0, \qquad \text{every } (\theta,\gamma) \neq (\theta_0,\gamma_0).$$

The densities $p_{\theta,\gamma}(x,y)$ are uniformly bounded in θ, γ and (x,y). Furthermore by Jensen's inequality $|\log p_{\theta_0,\gamma_0}|$ is bounded up to a constant by x^2+y^2+1 and $p_{\theta_0,\sigma_0,\bar{\eta}_n}$ converges pointwise and in mean to $p_{\theta_0,\sigma_0,\eta_0}$. Apply Fatou's lemma to see that for every $m_n \to \infty$, $M_n \downarrow -\infty$ and neighbourhoods U_m decreasing to an arbitrary pair (θ,γ) we have

$$\limsup_{n\to\infty} \int \Bigl(\sup_{(\theta',\gamma')\in U_{m_n}} \log \frac{p_{\theta',\gamma'}}{p_{\theta_0,\gamma_0}} \vee M_n \Bigr)\, p_{\theta_0,\sigma_0,\bar{\eta}_n}\, d\lambda^2 < 0.$$

Thus for every (θ,γ) there exists an open neighbourhood and a constants M (both depending on (θ,γ)) such that

$$\limsup_{n\to\infty} \int \Bigl(\sup_{(\theta',\gamma')\in U} \log \frac{p_{\theta',\gamma'}}{p_{\theta_0,\gamma_0}} \vee M \Bigr)\, p_{\theta_0,\sigma_0,\bar{\eta}_n}\, d\lambda^2 < 0. \tag{2.1}$$

Given an open neighbourhood V of (θ_0, γ_0), its complement, which is compact, can be covered with finitely many neighbourhoods $U_1, \ldots, U_k$ attached to some (θ_i, γ_i) in this manner. If $(\hat\theta_n, \hat\eta_n)$ is not in V, then it is in one of these neighbourhoods. It suffices to show that for every neighbourhood U the probability that it contains $(\hat\theta_n, \hat\eta_n)$ tends to zero as $n \to \infty$. By the definition of $(\hat\theta_n, \hat\eta_n)$ this probability is bounded by

$$\mathrm{P}_{\theta_0,\sigma_0,\eta_1,\eta_2,\ldots}\Big(\mathbb{P}_n \log \frac{\sup_{(\theta,\gamma)\in U} p_{\theta,\gamma}}{p_{\theta_0,\gamma_0}} \vee M \geq 0\Big)$$

Here the variables $A_{ni} = \log \sup_{(\theta,\gamma)\in U} p_{\theta,\gamma}/p_{\theta_0,\gamma_0}(X_i, Y_i) \vee M$ are bounded below by M and bounded above by a multiple of $X_i^2 + Y_i^2 + 1$. Under the conditions $n^{-1}\sum_{i=1}^n \mathrm{E}|A_{ni}| =)(1)$ and $n^{-1}\sum_{i=1}^n \mathrm{E}|A_{ni}|\{|A_{ni}| > \varepsilon n\} \to 0$ for every $\varepsilon > 0$, which are implied by the moment condition on $\bar\eta_n$, the averages $\bar A_n$ satisfy the weak law of large numbers: $\bar A_n - \mathrm{E}\bar A_n \to 0$ in probability. Since $\mathrm{E}\bar A_n$ is asymptotically negative by (2.1) it follows that the probability in the preceding display converges to zero. ∎

For the proof of Theorem 1.2 we shall also need the consistency of the corresponding estimators of the nuisance parameters under the null hypotheses. Let $\hat\gamma_{00}$ be defined analogously to $\hat\gamma$, but with the parameter θ fixed at the value θ_0. Thus $\hat\gamma_{00}$ maximizes the function

$$\gamma \mapsto \prod_{i=1}^n p_{\theta_0,\gamma}(X_i, Y_i), \tag{2.2}$$

over $[m, M] \times \mathcal{H}$ for $\mathcal{H}$ the probability distributions on $\mathbb{R}$. Similarly let $(\hat\alpha_0, \hat\gamma_0)$ maximize the function

$$(\alpha, \gamma) \mapsto \prod_{i=1}^n p_{\alpha,\beta_0,\gamma}(X_i, Y_i), \tag{2.3}$$

The proof of the following theorem is similar to the proof of the preceding theorem and omitted.

THEOREM 2.2. *Assume that the sequence of distributions $\bar\eta_n = n^{-1}\sum_{i=1}^n \eta_j$ satisfies $\int |z|^{2+\delta}\, d\bar\eta_n(z) = O(1)$ for some $\delta > 0$. Then $\hat\sigma_{n,00} \xrightarrow{\mathrm{P}} \sigma_0$ and $d(\hat\eta_{n,00}, \bar\eta_n) \xrightarrow{\mathrm{P}} 0$ under $(\theta_0, \sigma_0, \eta_1, \eta_2, \ldots)$ for d a distance that generates the weak topology. Similarly $\hat\sigma_{n,0} \xrightarrow{\mathrm{P}} \sigma_0$ and $d(\hat\eta_{n,0}, \bar\eta_n) \xrightarrow{\mathrm{P}} 0$.*

A final result on consistency that is useful in the proof of Theorem 1.2 concerns the consistency of the mean of our estimator for $\bar\eta_n$. This is also of independent interest.

THEOREM 2.3. *Assume that the sequence of distributions $\bar\eta_n = n^{-1}\sum_{i=1}^n \eta_j$ satisfies $\int |z|^{2+\delta}\, d\bar\eta_n(z) = O(1)$ for some $\delta > 0$. Then the differences between $\int z\, d\hat\eta_{n,00}(z)$, $\int z\, d\hat\eta_{n,0}(z)$, $\int z\, d\bar\eta_n(z)$ and $\int z\, d\eta(z)$ converge to zero in probability under the model $(\theta_0, \sigma_0, \eta_1, \eta_2, \ldots)$.*

Proof. We give the proof for $\hat{\eta}$, the other cases being similar. Inspection of the likelihood shows that our estimator $\hat{\eta}$ maximizes

$$\eta \mapsto \prod_{i=1}^{n} \int \phi\big((T_i - z)/\hat{\sigma}(1+\hat{\beta}^2)^{-1/2}\big)\, d\eta(z), \tag{2.4}$$

for $T_i = \big(X_i + \hat{\beta}(Y_i - \hat{\alpha})\big)/(1+\hat{\beta}^2)$. Define submodels

$$d\hat{\eta}_t = \big(1 + (z - \hat{\eta}z)\big)\, d\hat{\eta},$$
$$\hat{\eta}_t(B) = \hat{\eta}(B - t).$$

For fixed $\hat{\eta}$ these are well defined for t sufficiently close to zero. (Remember that $\hat{\eta}$ is discrete.) Inserting these submodels in the likelihood and differentiating with respect to t at $t = 0$ we obtain the equations

$$\mathbb{P}_n \frac{\int (z - \hat{\eta}z)\, \phi\big((T_i - z)/\hat{\sigma}(1+\hat{\beta}^2)^{-1/2}\big)\, d\eta(z)}{\int \phi\big((T_i - z)/\hat{\sigma}(1+\hat{\beta}^2)^{-1/2}\big)\, d\eta(z)} = 0,$$
$$\mathbb{P}_n \frac{\int (T_i - z)\, \phi\big((T_i - z)/\hat{\sigma}(1+\hat{\beta}^2)^{-1/2}\big)\, d\eta(z)}{\int \phi\big((T_i - z)/\hat{\sigma}(1+\hat{\beta}^2)^{-1/2}\big)\, d\eta(z)} = 0.$$

Here $\mathbb{P}_n$ is the empirical measure of $T_1, \ldots, T_n$. It follows that $\hat{\eta}z = n^{-1}\sum_{i=1}^{n} T_i$. The result follows by inserting the defining equations for X_i and Y_i and applying the law of large numbers. ■

3. Least Favorable Submodels

The proofs of both Theorem 1.1 and Theorem 1.2 are based on differentiating the log (mixture) likelihood along a least favorable submodel. Given a distribution η on $\mathbb{R}$ and pairs $\theta = (\alpha, \beta)$ and $t = (a, b)$ define, with $\kappa_b = b(1 + b^2)^{-1}$,

$$\eta_\theta(t, \eta)(B) = \eta\Big(B\big(1 + (b - \beta)\kappa_b\big)\Big)^{-1} + (\alpha - a)\kappa_b\Big).$$

For $|b - \beta|\kappa_b < 1$ this defines a probability distribution on $\mathbb{R}$. Thus we obtain a submodel $\theta \mapsto \eta_\theta(t, \eta)$ that passes through η at $\theta = t$. The gradient (vector of partial derivatives) of the function the function $\log p_{\theta,\sigma,\eta_\theta(t,\eta)}(x, y)$ with respect to θ can be found by inserting the path η_θ in the mixture density, a change of variables, and straightforward calculations. Evaluating the gradient at $\theta = t$ we obtain

$$\begin{aligned}\tilde{\ell}_{t,\gamma}(x, y) &:= \frac{\partial}{\partial\theta} \log p_{\theta,\sigma,\eta_\theta(t,\eta)}(x, y)_{|\theta=t} \\ &= \frac{-bx + y - a}{\sigma^2(1 + b^2)} \frac{\displaystyle\int \begin{pmatrix} 1 \\ z \end{pmatrix} \phi\Big(\frac{x - z}{\sigma}\Big)\phi\Big(\frac{y - a - bz}{\sigma}\Big) \frac{1}{\sigma^2}\, d\eta(z)}{\displaystyle\int \phi\Big(\frac{x - z}{\sigma}\Big)\phi\Big(\frac{y - a - bz}{\sigma}\Big) \frac{1}{\sigma^2}\, d\eta(z)}. \end{aligned} \tag{3.1}$$

This is well-known to be the efficient score function for θ (the score function minus its projection on the linear span of the nuisance scores) for the mixture version of the model at (θ, γ). See e.g. Bickel, Klaassen, Ritov and Wellner (1993), page 135–139 or Van der Vaart (1995), Section 5. In this sense the submodel $\theta \mapsto \eta_\theta(t, \eta)$ is least favorable at (t, γ) for estimating θ. Note that σ does not play a role in this submodel: the parameters θ and σ are orthogonal in the sense that efficient estimators for θ and σ are asymptotically independent. The asymptotic covariance matrix of the best estimators of θ in the mixture model (which include the maximum likelihood estimators by Theorem 1.1) is the inverse of the efficient information matrix

$$\tilde{I}_{\theta,\gamma} = \mathrm{E}_{\theta,\gamma}\tilde{\ell}_{\theta,\gamma}(X, Y)\tilde{\ell}_{\theta,\gamma}(X, Y)'. \tag{3.2}$$

It can also be checked that (3.1) gives the 'conditional score function' (Lindsay (1983a) defined as (with $\dot{\ell}_{\theta,\gamma}$ the partial derivative with respect to θ of the log mixture density)

$$\tilde{\ell}_{\theta,\gamma}(X, Y) = \dot{\ell}_{\theta,\gamma}(X, Y) - \mathrm{E}_\theta\big(\dot{\ell}_{\theta,\gamma}(X, Y)|\, X + (Y - \alpha)\beta\big).$$

Note that $X + (Y - \alpha)\beta$ is a sufficient statistic for the nuisance parameters $\eta_1, \eta_2, \ldots$, which, however, depends on the parameter of interest. An important property, which may be checked using the fact that $-\beta X + Y - \alpha$ is independent from $X + \beta(Y - \alpha)$, is that

$$\mathrm{E}_{\theta,\gamma}\tilde{\ell}_{\theta,\gamma'}(X, Y) = 0, \qquad \text{every } \theta, \gamma, \gamma'. \tag{3.3}$$

Thus the efficient score function yields an estimation equation for θ that is unbiased in the nuisance parameter: using the methods of this paper the equation

$\sum \tilde{\ell}_{\theta,\gamma}(X_i, Y_i) = 0$ can be shown to give an asymptotically normal estimator for θ, for any choice of γ, even random choices. Choosing a random sequence γ_n that converges to γ_0 we obtain an efficient estimator for θ. As we shall now show Theorem 1.1 corresponds to choosing the maximum likelihood estimator for γ.

Denote the empirical measure of the observations by $\mathbb{P}_n$ and write taking expectations in the operator notion; thus $\mathbb{P}_n f(X,Y) = n^{-1}\sum_{i=1}^n f(X_i, Y_i)$. Since the estimator $\hat{\theta}$ maximizes the function

$$\theta \mapsto \prod_{i=1}^n p_{\theta,\hat{\sigma},\eta_\theta(\hat{\theta},\hat{\eta})}(x_i, y_i),$$

(note that $\eta_{\hat{\theta}}(\hat{\theta}, \hat{\eta}) = \hat{\eta}$), we can conclude that

$$\mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{\gamma}}(X,Y) = 0. \tag{3.4}$$

This 'efficient score equation' combined with the unbiasedness (3.3) is the basis of our proof of asymptotic normality of $\hat{\theta}$. This is carried out by linearizing the efficient score equation in $\hat{\theta} - \theta$, keeping $\hat{\gamma}$ fixed. The terms in the linearization, which are sums of functions dependent on $\hat{\gamma}$ evaluated at the observations, are controlled by using a uniform central limit theorem for empirical processes.

Let $\hat{\gamma}_{00}$ be the maximum likelihood estimator of γ in the mixture model under the hypothesis that $\theta = \theta_0$, i.e. the maximizer of (2.2) over $[m, M]$ times the probability distributions on $\mathbb{R}$. Then the likelihood ratio statistic for testing $H_0{:}\theta = \theta_0$ can be 'sandwiched' in the following manner:

$$2n\mathbb{P}_n \log \frac{p_{\hat{\theta},\hat{\sigma}_{00},\eta_{\hat{\theta}}(\theta_0,\hat{\eta}_{00})}}{p_{\theta_0,\hat{\sigma}_{00},\hat{\eta}_{00}}} \le L_n(\theta_0) \le 2n\mathbb{P}_n \log \frac{p_{\hat{\theta},\hat{\sigma},\hat{\eta}}}{p_{\theta_0,\hat{\sigma},\eta_{\theta_0}(\hat{\theta},\hat{\eta})}}. \tag{3.5}$$

The proof of the second assertion in Theorem 1.2 is based on two-term Taylor expansions in $\hat{\theta} - \theta_0$ of the left and right side, again keeping the other estimators fixed. Since in the left side we can write $\hat{\eta}_{00} = \eta_{\theta_0}(\theta_0, \hat{\eta}_{00})$, and in the right side $\hat{\eta} = \eta_{\hat{\theta}}(\hat{\theta}, \hat{\eta})$, these are ordinary Taylor expansions along (2-dimensional) least favorable models. Both sides are shown to converge to a chi-squared distribution. For a nontechnical motivation of this method of proof we refer to Murphy and Van der Vaart (1995).

The proof of the first assertion of Theorem 1.2 is based on a 'sandwich' approach as well. In this case we use a least favorable submodel for β only, which should include a perturbation in both the α and η space. If the efficient score for α and β jointly is written in the form $\tilde{\ell}_{\theta,\gamma} = (\tilde{\ell}_{\theta,\gamma|\alpha}, \tilde{\ell}_{\theta,\gamma|\beta})$, then the efficient score function for β in the presence of (α, γ) can be found as

$$\tilde{\ell}_{\theta,\gamma|\beta} - \frac{(\tilde{I}_{\theta,\gamma})_{1,2}}{(\tilde{I}_{\theta,\gamma})_{1,1}} \tilde{\ell}_{\theta,\gamma|\alpha} = \tilde{\ell}_{\theta,\gamma|\beta} - \int z\, d\eta(z)\, \tilde{\ell}_{\theta,\gamma|\alpha}. \tag{3.6}$$

For $\alpha = a$ this is the score function at $\beta = b$ of the submodel indexed by the parameters $\beta \mapsto (\theta,\gamma)_\beta(t,\gamma) := \big(\alpha_\beta(t), \beta, \sigma, \eta_{\alpha_\beta(t),\beta}(t,\eta)\big)$ with $\alpha_\beta(t) = a + (b-\beta)\int z\, d\eta$

(where $t = (a, b)$ and $\eta_\theta(t, \eta)$ are as before). Thus this submodel is least favourable and motivates the sandwich

$$2n\mathbb{P}_n \log \frac{p_{(\theta,\gamma)_{\hat\beta}(\hat\alpha_0,\beta_0,\hat\sigma_0,\hat\eta_0)}}{p_{\hat\alpha_0,\beta_0,\hat\sigma_0,\hat\eta_0}} \le K_n(\beta_0) \le 2n\mathbb{P}_n \log \frac{p_{\hat\theta,\hat\sigma,\hat\eta}}{p_{(\theta,\gamma)_{\beta_0}(\hat\theta,\hat\gamma)}}. \tag{3.7}$$

We next proceed by a two-term Taylor expansion in the one-dimensional parameter $\hat\beta - \beta_0$, noting that $(\hat\alpha_0, \beta_0, \hat\sigma_0, \hat\eta_0) = (\theta, \gamma)_{\beta_0}(\hat\alpha_0, \beta_0, \hat\sigma_0, \hat\eta_0)$ and $(\hat\theta, \hat\sigma, \hat\eta) = (\theta, \gamma)_{\hat\beta}(\hat\theta, \hat\gamma)$.

The technical details of the program outlined in the preceding paragraphs are not trivial because of the presence of estimators for the nuisance parameters γ and γ_0. In both proofs the expansions contain random terms of the form $\mathbb{P}_n\ell(\cdot|\,\tilde\theta, \tilde\gamma)$ for deterministic functions $\ell(x, y|\,\theta, \gamma)$ and estimators $(\tilde\theta, \tilde\gamma)$ depending on all the data. The following propositions are used to control these expressions.

The propositions are stated for independent random elements $X_1, \ldots, X_n$ in an arbitrary measurable space $(\mathcal{X}, \mathcal{A})$ and arbitrary collections $\mathcal{F}$ of measurable functions $f: \mathcal{X} \mapsto \mathbb{R}$. The function F is a measurable envelope function of the class $\mathcal{F}$: $|f| \le F$ for every $f \in \mathcal{F}$. The $L_r(P)$-bracketing number $N_{[]}\big(\varepsilon, \mathcal{F}, L_r(P)\big)$ is defined as the minimal number of pairs of functions $[l, u]$ such that $P(u - l)^r \le \varepsilon^r$ and every $f \in \mathcal{F}$ is contained in some bracket: $l \le f \le u$ for some pair $[l, u]$.

PROPOSITION 3.1. *Let $X_1, \ldots, X_n$ be independent random elements with distributions $P_1, \ldots, P_n$. For $\bar P_n = n^{-1}\sum_{i=1}^n P_i$ suppose*

$$\sup_n N_{[]}\big(\varepsilon, \mathcal{F}, L_1(\bar P_n)\big) < \infty, \qquad \text{every } \varepsilon > 0$$

$$\bar P_n F = O(1), \qquad \bar P_n F\{F \ge \varepsilon n\} \to 0, \qquad \text{every } \varepsilon > 0.$$

Then the sequence $\sup_{f\in\mathcal{F}}\big|n^{-1}\sum_{i=1}^n (f(X_i) - P_i f)\big|$ converges in outer probability to zero.

Proof. By the moment assumptions on the envelope function F the sequence $(\mathbb{P}_n - \bar P_n)f_n$ converges to zero in probability for every sequence of measurable functions f_n with $|f_n| \le F$. If $l_n \le f \le u_n$, then $(\mathbb{P}_n - \bar P_n)f \le (\mathbb{P}_n - \bar P_n)u_n + \bar P_n(u_n - l_n)$. For every fixed n choose a minimal number of brackets $[l_{n,i}, u_{n,i}]$ of size ε in $L_1(\bar P_n)$ that cover $\mathcal{F}$. By assumptions the number of brackets is uniformly bounded in n. Thus

$$\sup_f (\mathbb{P}_n - \bar P_n)f \le \sup_i \big|(\mathbb{P}_n - \bar P_n)u_{n,i}\big| + \varepsilon,$$

where the number of terms in the supremum on the right is uniformly bounded in n. The bracketing functions u_n can be chosen to satisfy $|u_n| \le F$. Conclude that the right side of the display converges in probability to ε. This being true for every

$\varepsilon > 0$, and a similar argument applied with the lower bracketing functions, yields the proposition. ∎

PROPOSITION 3.2. *Let $X_1, \ldots, X_n$ be independent random elements with distributions $P_1, \ldots, P_n$. For $\bar{P}_n = n^{-1}\sum_{i=1}^n P_i$ and an arbitrary probability measure P_0 suppose*

$$\int_0^{\delta_n} \sqrt{\log N_{[]}\bigl(\varepsilon, \mathcal{F}, L_2(\bar{P}_n)\bigr)}\, d\varepsilon \to 0, \qquad \text{every } \delta_n \downarrow 0.$$

$$\bar{P}_n F^2 = O(1), \qquad \bar{P}_n F^2\{F \geq \varepsilon\sqrt{n}\} \to 0, \qquad \text{every } \varepsilon > 0.$$

$$\sup_{f,g\in\mathcal{F}} \bigl|(\bar{P}_n - P_0)(f-g)^2\bigr| \to 0.$$

Then the sequence $\{n^{-1/2}\sum_{i=1}^n (f(X_i) - P_i f)\colon f \in \mathcal{F}\}$ converges in distribution in the space $\ell^\infty(\mathcal{F})$ to a tight Brownian P_0-bridge.

Proof. This follows along the lines of Andersen, Giné, Ossiander and Zinn (1987), or alternatively and more directly from Theorem 2.11.9 of Van der Vaart and Wellner (1995). Note that $\mathcal{F}$ is totally bounded in $L_2(P_0)$ for every $\varepsilon > 0$ by the first and third condition: by the first the sequence $N_{[]}\bigl(\varepsilon, \mathcal{F}, L_2(\bar{P}_n)\bigr)$ is bounded in n for every $\varepsilon > 0$; by the third its limsup is a bound for the covering numbers of $\mathcal{F}$ under P_0. ∎

4. Proof of Theorem 1.1

Let E_0 denote expectation under the true parameters $(\theta_0, \sigma_0, \eta_1, \eta_2, \ldots)$ and let $P_{0,\eta}$ denote expectation under the mixture distribution with parameters $(\theta_0, \sigma_0, \eta)$. Define an $\mathbb{R}^2$-valued stochastic process indexed by the parameters (θ, γ) by

$$G_n(\theta,\gamma) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Bigl(\tilde{\ell}_{\theta,\gamma}(X_i, Y_i) - \mathrm{E}_0\tilde{\ell}_{\theta,\gamma}(X_i, Y_i)\Bigr). \tag{4.1}$$

By Lemma 7.3 there exists a neighbourhood U of $(\theta_0, \sigma_0, \eta_0)$ for the product of the Euclidean and weak topology, such that the set $\mathcal{F}$ of all functions $\tilde{\ell}_{\theta,\sigma,\eta|\alpha}$, $\tilde{\ell}_{\theta,\sigma,\eta|\beta}$ with (θ, σ, η) ranging over this neighbourhood satisfies, for every $V \geq 1/\alpha$ and $0 < \alpha \leq 1$, and $\delta > 0$

$$\log N_{[]}\bigl(\varepsilon, \mathcal{F}, L_2(P_{0,\bar{\eta}_n})\bigr) \leq C\Bigl(\frac{1}{\varepsilon}\Bigr)^V \Bigl(P_{0,\bar{\eta}_n}\bigl(1 + |x| + |y|\bigr)^{5+\delta+2\alpha+2/V}\Bigr)^{V/2}.$$

For V close to 2, δ close to zero and $\alpha > 1/2$ close to $1/2$ the right hand side is finite and bounded in n by the assumption that the $7+\delta$-moment of $\bar{\eta}_n$ is bounded. Furthermore, by Lemma 7.1 the class $\mathcal{F}$ has envelope function $F(x, y) = \bigl(1+|x|+|y|\bigr)^2$.

It follows that $\mathcal{F}$ satisfies the first two conditions of Proposition 3.2. The third condition of this proposition concerns the expressions

$$\int \Big[\int \big\| \tilde{\ell}_{\theta,\gamma} - \tilde{\ell}_{\theta',\gamma'} \big\|^2 \, p_{\theta_0,\sigma_0}(\cdot|\, z)\, d\lambda^2 \Big] \, d(\bar{\eta}_n - \eta_0)(z).$$

Here write $p_{\theta,\sigma}(x, y|\, z)$ for the bivariate Gaussian density of (X, Y) given $Z = z$. The functions in square brackets can be bounded by

$$\iint 4F^2 \, p_{\theta_0,\sigma_0}(\cdot|\, z)\, d\lambda^2 \lesssim 1 + |z|^4.$$

Their derivatives with respect to z can be bounded similarly by a multiple of $1+|z|^5$. It now follows by Lemma 7.4 that $\mathcal{F}$ satisfies also the third condition of Proposition 3.2.

Thus the process G_n converges in distribution in the space $\ell^\infty(U, \mathbb{R}^2)$ to a tight Gaussian process, that can be identified with a P_{0,η_0}-brownian bridge process. The sample paths of the limit process are uniformly continuous with respect to the semi-metric with square

$$d^2\big((\theta,\gamma),(\theta',\gamma')\big) = \iint \big\| \tilde{\ell}_{\theta,\gamma} - \tilde{\ell}_{\theta',\gamma'} \big\|^2 \, p_{\theta_0,\sigma_0}(\cdot|\, z)\, d\lambda^2 \, d\eta_0(z).$$

By the dominated convergence theorem and Theorem 2.1 the distance between $(\hat{\theta}, \hat{\gamma})$ and (θ_0, γ_0) converges to zero in probability. Conclude that

$$G_n(\hat{\theta}, \hat{\gamma}) - G_n(\theta_0, \gamma_0) \xrightarrow{\mathrm{P}} 0.$$

In view of the efficient score equation (3.4) and the unbiasedness (3.3) of the efficient score functions this is equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{\theta_0,\gamma_0}(X_i, Y_i) + \sqrt{n} \int \tilde{\ell}_{\hat{\theta},\hat{\gamma}} \, (p_{\theta_0,\sigma_0,\bar{\eta}_n} - p_{\hat{\theta},\sigma_0,\bar{\eta}_n})\, d\lambda^2 \xrightarrow{\mathrm{P}} 0.$$

The final step is to linearize the integral in $\hat{\theta} - \theta_0$. More precisely, the theorem follows if it can be shown that

$$\int \tilde{\ell}_{\hat{\theta},\hat{\gamma}} \, \big[p_{\hat{\theta},\sigma_0,\bar{\eta}_n} - p_{\theta_0,\sigma_0,\bar{\eta}_n} - (\hat{\theta} - \theta_0)' \dot{\ell}_{\theta_0,\sigma_0,\bar{\eta}_n} \, p_{\theta_0,\sigma_0,\bar{\eta}_n} \big] \, d\lambda^2 = o_P\big(\|\hat{\theta} - \theta_0\|\big),$$

$$\int \tilde{\ell}_{\hat{\theta},\hat{\gamma}} \, \dot{\ell}'_{\theta_0,\sigma_0,\bar{\eta}_n} \, p_{\theta_0,\sigma_0,\bar{\eta}_n} \, d\lambda^2 = \tilde{I}_{\theta_0,\sigma_0,\eta_0} + o_P(1).$$

This follows by standard arguments, where for the second line we note that the inner product of the efficient score function with the ordinary score function for θ equals the efficient information matrix by the projection property of an efficient score function.

5. Proof of Theorem 1.2

We shall derive the limit distribution of the sequence $L_n(\theta_0)$. The arguments for the sequence $K_n(\beta_0)$ are similar and easier.

Assume without loss of generality that the sequence $\bar{\eta}_n$ converges weakly to a limit η_0; otherwise argue along subsequences. It suffices to show that both the left and the right side of (3.5) converge in distribution to a chi-squared distribution with two degrees of freedom. We show this by expanding the left side in a two-term Taylor expansion in $\hat{\theta}-\theta_0$ around θ_0 and, similarly, the right side around $\hat{\theta}$. In the expansion of the right side the linear term vanishes in view of the efficient score equation (3.4) and it suffices to consider the quadratic term. In the expansion of the left side both the linear term and the quadratic term contribute to the limit distribution. We shall only give the details for this side, the details for the right side being simpler.

The expansion of the left side of (3.5) takes the form

$$\begin{aligned} 2(\hat{\theta}-\theta_0)' n\mathbb{P}_n \frac{\partial}{\partial\theta}\log p_{\theta,\hat{\sigma}_{00},\eta_\theta(\theta_0,\hat{\eta}_{00})|\theta=\theta_0} \\ + (\hat{\theta}-\theta_0)' n\mathbb{P}_n \frac{\partial^2}{\partial\theta^2}\log p_{\theta,\hat{\sigma}_{00},\eta_\theta(\theta_0,\hat{\eta}_{00})|\theta=\tilde{\theta}}(\hat{\theta}-\theta_0), \end{aligned} \tag{5.1}$$

for a point $\tilde{\theta}$ between θ_0 and $\hat{\theta}$. By construction of the least favorable submodel the linear term equals

$$2\sqrt{n}(\hat{\theta}-\theta_0)'\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta_0,\hat{\gamma}_{00}} = 2\sqrt{n}(\hat{\theta}-\theta_0)'G_n(\theta_0,\hat{\gamma}_{00}),$$

with G_n as defined by (4.1) in the proof of Theorem 1.1, in view of the unbiasedness (3.3). According to the proof of Theorem 1.1 this can be further rewritten as

$$2\big(I^{-1}_{\theta_0,\gamma_0}G_n(\theta_0,\gamma_0)+o_P(1)\big)'\big(G_n(\theta_0,\gamma_0)+o_P(1)\big). \tag{5.2}$$

The second order term in (5.1) is a quadratic form in $\sqrt{n}(\hat{\theta}_n-\theta_0)$ with matrix of coefficients

$$\mathbb{P}_n\left(\frac{\frac{\partial^2}{\partial\theta^2}p_{\theta,\hat{\sigma}_{00},\eta_\theta(\theta_0,\hat{\eta}_{00})}}{p_{\theta,\hat{\sigma}_{00},\eta_\theta(\theta_0,\hat{\eta}_{00})}}\right)_{\theta=\tilde{\theta}} - \mathbb{P}_n\tilde{\ell}_{\tilde{\theta},\hat{\gamma}_{00}}\tilde{\ell}'_{\tilde{\theta},\hat{\gamma}_{00}}. \tag{5.3}$$

We show that the first term on the right converges in probability to zero and the second term to $-\tilde{I}_{\theta_0,\eta_0}$.

The (2,2)-elements of the matrix $\tilde{\ell}_{\theta,\gamma}\tilde{\ell}'_{\theta,\gamma}$ involve the functions

$$\Big(\frac{-\beta X+Y-\alpha}{\sigma^2(1+\beta)^2}\Big)^2\left(\frac{\int z p_{\theta,\sigma}(x,y|\,z)\,d\eta(z)}{\int p_{\theta,\sigma}(x,y|\,z)\,d\eta(z)}\right)^2.$$

By Lemma 7.3 there exists a neighbourhood of (θ_0,γ_0) such that the class $\mathcal{F}$ of all such functions with (θ,γ) ranging over this neighbourhood has bracketing numbers satisfying

$$\log N_{[]}\big(\varepsilon,\mathcal{F},L_1(P_{0,\bar{\eta}_n})\big) \lesssim \Big(\frac{1}{\varepsilon}\Big)^V\Big(P_{0,\bar{\eta}_n}\Big(1+|x|+|y|\Big)^{5+\alpha+1/V+\delta}\Big)^V.$$

The right side is bounded in n for e.g. $\alpha = V = 1$, whence $\mathcal{F}$ satisfies the first condition of Proposition 3.1. By Lemma 7.1 the class $\mathcal{F}$ has envelope function $\big(1 + |x| + |y|\big)^4$, so that $\mathcal{F}$ satisfies the second condition as well. In view of Theorem 2.2 and similar arguments applied to the other elements of the matrix $\tilde{\ell}_{\theta,\gamma}\tilde{\ell}'_{\theta,\gamma}$ we obtain

$$(\mathbb{P}_n - P_{0,\bar{\eta}_n})\tilde{\ell}_{\tilde{\theta},\hat{\gamma}_{00}}\tilde{\ell}'_{\tilde{\theta},\hat{\gamma}_{00}} \xrightarrow{\mathrm{P}} 0.$$

Apply the dominated convergence theorem to conclude that

$$\mathbb{P}_n\tilde{\ell}_{\tilde{\theta},\hat{\gamma}_{00}}\tilde{\ell}'_{\tilde{\theta},\hat{\gamma}_{00}} \xrightarrow{\mathrm{P}} \tilde{I}_{\theta_0,\gamma_0}.$$

This concludes the proof of convergence of the second term in (5.3).

By explicit calculations the first term in (5.3) can be seen to involve functions of the type in Lemma 7.3 with $k_0 + \sum ik_i \leq 6$. Thus we can apply again Proposition 3.1 and next the dominated convergence theorem to show that this term converges to zero. This concludes the proof of (5.3).

Finally combine (5.1), (5.2) and (5.3) to see that the left side of (3.5) is asymptotically equivalent to $G_n(\theta_0,\gamma_0)'I^{-1}_{\theta_0,\gamma_0}G_n(\theta_0,\gamma_0)$. This sequence is asymptotically chi-squared with two degrees of freedom.

6. Efficiency

In this section we discuss the asymptotic efficiency of our estimator and test statistics and compare our proposals to the standard procedures. We are particularly interested in efficiency under the incidental version of model, but throughout the section we assume the more general model parametrized by $(\alpha, \beta, \sigma, \eta_1, \eta_2, \ldots)$ as given in the introduction. For simplicity we shall concentrate on the slope parameter β alone. Our estimator for the intercept α gives no improvement over the usual procedures. We conjecture that similar results are valid for our estimator for σ. However the results of this paper do not even show that our estimator for σ converges at $\sqrt{n}$-rate. This remains to be investigated.

6.1. Estimating the slope

The standard procedure for estimating the slope parameter β, which we shall denote by $\hat{\beta}_{LS}$, can be described in (at least) three different ways. First it is the β-component of the maximum likelihood estimator for the parameter $(\alpha, \beta, \sigma, z_1, \ldots, z_n)$ in the functional version of the model, found by maximizing

$$\prod_{i=1}^{n} \frac{1}{\sigma}\phi\Big(\frac{X_i - z_i}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{Y_i - \alpha - \beta z_i}{\sigma}\Big).$$

Second $\hat{\beta}_{LS}$ is the β-component of the maximum likelihood estimator for the parameter $(\alpha, \beta, \sigma, \eta)$ in the mixture model restricted by the a-priori knowledge that the mixing distribution η belongs to the normal location scale family. In this case the observations have a bivariate Gaussian distribution depending on five unknown parameters: α, β, σ and the location and scale of the mixing distribution. Third, and motivating our notation, $\hat{\beta}_{LS}$ is the β-component of the least squares estimator for (α, β) found by minimizing

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

This is the sum of the squared (true and not vertical) distances of the points (X_i, Y_i) to the line

$$\begin{pmatrix} 0 \\ \alpha \end{pmatrix} + z \begin{pmatrix} 1 \\ \beta \end{pmatrix}.$$

For a discussion of these characterizations see e.g. Kendall and Stuart (1979), Chapter 29, Fuller (1987), Chapter 1 or Gleser (1981). From any of the three characterizations $\hat{\beta}_{LS}$ can be solved explicitly to give

$$\hat{\beta}_{LS} = \frac{S_Y^2 - S_X^2 + \sqrt{(S_Y^2 - S_X^2)^2 + 4S_{XY}^2}}{2S_{XY}},$$

where S_X^2, S_Y^2 and S_{XY} are the sample variances and covariances of the vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$. The limit distribution of $\hat{\beta}_{LS}$ can easily be obtained from this formula by means of the delta-method. Under our model as described in Section 1 under the conditions that $\bar{\eta}_n \rightsquigarrow \eta_0$ and $\int z^{2+\delta}\, d\bar{\eta}_n(z) = O(1)$ for some $\delta > 0$, we have

$$\sqrt{n}(\hat{\beta}_{LS} - \beta) \rightsquigarrow N\Big(0, \frac{\sigma^2(1+\beta^2)}{\operatorname{var} \eta_0} + \frac{\sigma^4}{\operatorname{var}^2 \eta_0}\Big). \tag{6.1}$$

(The conditions on the second plus delta moments and the convergence of $\bar{\eta}_n$ could be relaxed, but are certainly satisfied in the context of Theorem 1.1.) Theorem 4.2 in Gleser (1981) and Theorem 1.3.1 in Fuller (1989) imply this result for the incidental version of the model and the model with η Gaussian, respectively.

We wish to compare the asymptotic variance of $\hat{\beta}_{LS}$ to the asymptotic variance of our estimator, which is given by the (2,2)-element of the inverse of the matrix I_{θ,σ,η_0} given in (3.2). Alternatively $(I^{-1}_{\theta,\sigma,\eta_0})_{2,2}$ is the inverse of the second moment of the efficient influence function for β given in (3.6), with $\eta = \eta_0$, computed relatively to the mixture model with η_0.

A number of qualitative comparisons are possible without calculations. First, since $\hat{\beta}_{LS}$ is the maximum likelihood estimator in the mixture model restricted by the a-priori knowledge that η is Gaussian, it follows that the asymptotic variance of $\hat{\beta}_{LS}$ is not larger than that of our estimator $\hat{\beta}$ for η_0 a normal distribution. Second, that the variances are actually equal in this case follows from the fact that the least

favourable model in Section 3 is a location-scale model. Thus for η_0 Gaussian, the least favourable submodel remains within the Gaussian family. Since our estimator is efficient in the least favourable model, its asymptotic variance is least possible for η_0 Gaussian, hence equals the asymptotic variance of $\hat{\beta}_{LS}$. This was already noted by Bickel and Ritov (1987). Third, the asymptotic variance of our proposal is never larger than the asymptotic variance of the usual estimator. The distributional result (6.1) can be extended to the assertion that the sequence $\sqrt{n}(\hat{\beta}_{LS} - \beta - h/\sqrt{n})$ has the same normal limit distribution in the mixture model under every sequence of parameters $(\alpha + g/\sqrt{n}, \beta + h/\sqrt{n}, \sigma + \gamma/\sqrt{n}, \eta_n)$ such that for some function k

$$\int \left[\sqrt{n}(d\eta_n^{1/2} - d\eta_0^{1/2}) - \tfrac{1}{2} k \, d\eta_0^{1/2}\right]^2 \to 0. \tag{6.2}$$

(Under these conditions we have local asymptotic normality. See e.g. Theorem 5.13 and its proof in Van der Vaart (1988).) Thus the sequence $\hat{\beta}_{LS}$ is regular in the mixture model at $(\theta, \beta, \sigma, \eta_0)$, so that its asymptotic variance cannot be smaller than the inverse of the efficient information for β, by the convolution theorem. (cf. Begun, Hall, Huang, Wellner (1983).)

The following theorem asserts strict improvement of our estimator whenever η_0 is not normal.

THEOREM 6.1. *For any θ, σ and η_0 we have*

$$(\tilde{I}^{-1}_{\theta,\sigma,\eta_0})_{2,2} \leq \frac{\sigma^2(1+\beta^2)}{\operatorname{var} \eta_0} + \frac{\sigma^4}{\operatorname{var}^2 \eta_0}, \tag{6.3}$$

with equality if and only if η_0 is a normal distribution.

Proof. By the delta-method the least squares estimator can be shown to be asymptotically linear under $(\theta, \sigma, \eta_1, \eta_2, \ldots)$ in the sense that

$$\sqrt{n}(\hat{\beta}_{LS} - \beta) = n^{-1/2} \sum_{i=1}^{n} \ell_{LS}(X_i, Y_i) + o_P(1),$$

for the 'asymptotic influence function' ℓ_{LS} given by

$$\begin{aligned} \ell_{LS}(x,y) = \frac{1}{(1+\beta^2)\operatorname{var} \eta_0} \Big[& -\beta\big((x - \mathrm{E}X)^2 - \operatorname{var} X\big) \\ & + \beta\big((y - \mathrm{E}Y)^2 - \operatorname{var} Y\big) + (1-\beta^2)\big((x - \mathrm{E}X)(y - \mathrm{E}Y) - \operatorname{cov}(X,Y)\big)\Big]. \end{aligned}$$

Here the expectations and covariances are computed for (X, Y) distributed according to the mixture model with parameters (θ, σ, η_0). Equality in (6.3) would mean that the least squares estimator is asymptotically efficient in the mixture model at (θ, σ, η_0). Since it is regular in the sense of Hájek, the convolution theorem would show that its asymptotic influence function coincides almost surely with the efficient influence

function for β given in (3.6), relative to the mixture model. This means that the function

$$(-\beta x + y - \alpha)\Big(\frac{\int z\phi\Big(\frac{x-z}{\sigma}\Big)\phi\Big(\frac{y-\alpha-\beta z}{\sigma}\Big)\,\frac{1}{\sigma^2}\,d\eta_0(z)}{\int \phi\Big(\frac{x-z}{\sigma}\Big)\phi\Big(\frac{y-\alpha-\beta z}{\sigma}\Big)\,\frac{1}{\sigma^2}\,d\eta_0(z)} - \int z\,d\eta_0\Big)$$

is almost surely equal to a polynomial in (x, y) of degree at most 2. By continuity we have equality for all (x, y). Setting y equal to α we conclude that the function

$$\frac{\sigma^{-2}\int z e^{zx/\sigma^2}\, e^{-\frac{1}{2}z^2(1+\beta^2)/\sigma^2}\,d\eta_0(z)}{\int e^{zx/\sigma^2}\, e^{-\frac{1}{2}z^2(1+\beta^2)/\sigma^2}\,d\eta_0(z)}$$

is a polynomial of degree 1 in x. Integrate with respect to x to conclude that there exist constants a, b and c such that for every $x \in \mathbb{R}$

$$\int e^{zx}\, e^{-\frac{1}{2}z^2(1+\beta^2)/\sigma^2}\,d\eta_0(z) = e^{ax^2+bx+c}.$$

The left side is the Laplace transform of the measure μ with density $e^{-\frac{1}{2}z^2(1+\beta^2)/\sigma^2}$ with respect to η_0. It is finite for all $x \in \mathbb{R}$, hence analytic in $x \in \mathbb{C}$. By analytic continuation the identity remains true for $x \in \mathbb{C}$. Set $x = it$ to see that μ has characteristic function $\exp(-at^2 + bit + c)$. Conclude that $a \geq 0$ and that μ is a Gaussian measure. So is η_0. ■

The preceding theorem is encouraging, since it shows that the usual estimator sequence can be improved globally, at least under the condition of Theorem 1.1 that $\int z^{7+\delta}\,d\bar{\eta}_n(z)$ remains bounded. It does not show the size of the improvement. Since it does not seem feasible to evaluate the relative efficiency analytically (except for Gaussian η_0) we estimated the relative efficiency for β by a combination of an explicit and a Monte Carlo integration. Table 1 shows that the gain may be between 0 and 20% for a variety of design distributions and depending on the error variance σ^2. (Somewhat disappointingly the gain in the case of a uniform design appears less than 10 %.) The gain in efficiency in the incidental model is perhaps a surprising fact. For a discussion in a similar situation from an empirical Bayes perspective, see Lindsay (1985).

6.2. Testing the slope

The most pupular procedure to test the hypothesis $H_0{:}\beta = \beta_0$ appears to be the test suggested by Creasy (1956). It is based on the sample correlation coefficient $r_n = r_n(\beta_0)$ of the vectors $(V_1, W_1), \ldots, (V_n, W_n)$ defined by

$$V_i = X_i + \beta_0 Y_i; \qquad W_i = Y_i - \beta_0 X_i.$$

η_0	exp(1)	exp(1)	exp(1)	exp(1)	D_1	D_2	D_2	D_3	D_3
β	1	1	0	0.5	1	1	1	1	1
σ	1	2	2	0.5	1	1	2	1	4
ARE	91	79	79	95	92	97	93	99	97

Table 1. Asymptotic relative efficiencies of the least squares estimator for the slope relative to $\hat{\beta}$ for some distributions and values of β. The distribution η_0 is the limit of the sequence $\bar{\eta}_n$. The distributions coded D_1, D_2 and D_3 are the discrete distributions with masses $\frac{1}{2}, \frac{1}{2}$ on $\{1,2\}$, masses $10/18, 1/18, \ldots, 1/18$ on $\{0,2,3,\ldots,9\}$ and masses $1/10, \ldots, 1/10$ on $\{1,\ldots,10\}$, respectively. (The numbers are based on Monte-Carlo integration of the efficient score function using 1000 000 samples.)

The null hypothesis is rejected for large values of $\left|r_n(\beta_0)\right|$. Under the null hypothesis the statistic $r_n(\beta_0)$ possesses the same distribution as the sample correlation of n vectors from a bivariate standard normal distribution. Thus the procedure can be exact in the sense that the critical value of the test can be chosen such that the level is exactly equal to a given nominal value α. A disadvantage of the test is that it is really testing the hypothesis that the correlation between V and W is zero and this is equivalent to the hypothesis $H_0'\colon \beta = \beta_0$ or $\beta = -1/\beta_0$, rather than $H_0\colon \beta = \beta_0$, in the case that $\beta_0 \neq 0$. (Note that $\operatorname{cov}_\beta(V_i, W_i) = (1+\beta\beta_0)(\beta-\beta_0)\operatorname{var}\eta_i$.) Similarly, since $\left|r_n(\beta_0)\right| = \left|r_n(-1/\beta_0)\right|$ a confidence interval based on Creasy's test will contain the value $-1/\beta_0$ whenever it contains β_0. Several approaches have been suggested to remedy this situation. See e.g. Fuller (1987), Section 1.3.4, Kendall and Stuart (1979) or Zhang (1994). From an asymptotic perspective the problem is negligible, for the confidence set could just be intersected with an interval $(\hat{\beta}-\delta, \hat{\beta}+\delta)$ for any consistent estimator $\hat{\beta}$ and $\delta > 0$.

The asymptotic power of the test based on r_n can be investigated using the delta-method. Define functions

$$\phi_n(\beta) = \frac{(1+\beta\beta_0)(\beta-\beta_0)\operatorname{var}\bar{\eta}_n}{\sqrt{(1+\beta_0\beta)^2\operatorname{var}\bar{\eta}_n + (1+\beta_0^2)\sigma^2}\sqrt{(\beta-\beta_0)^2\operatorname{var}\bar{\eta}_n + \sigma^2(1+\beta_0^2)}}.$$

Then the sequence $\sqrt{n}\big(r_n(\beta_0) - \phi_n(\beta)\big)$ converges to a mean zero normal distribution under every sequence of parameters $(\alpha, \beta, \sigma, \eta_1, \eta_2, \ldots)$ such that $\int |z|^4\, d\bar{\eta}_n(z) = O(1)$ and $\bar{\eta}_n \rightsquigarrow \eta_0$. For $\beta = \beta_0$ its asymptotic variance is equal to one and the convergence is uniform and continuous in β ranging through a neighbourhood of β_0. It follows that the test that rejects the null hypothesis if $\left|r_n(\beta_0)\right| > \chi^2_{1,\alpha}$ possesses asymptotic level α. Its asymptotic power under the sequence of alternatives $\beta_n = \beta_0 + h/\sqrt{n}$ equals

$$\mathrm{P}_{\beta_n}\Big(\big|r_n(\beta_0)\big|^2 > \chi^2_{1,\alpha}\Big) \to \mathrm{P}\Big(N\left(sh,1\right)^2 > \chi^2_{1,\alpha}\Big).$$

where $s = \lim \sqrt{n}\big(\phi_n(\beta_n) - \phi_n(\beta_0)\big)$ is the 'slope of the test' and has square

$$s^2 = \frac{\operatorname{var}^2\eta_0}{\sigma^2(1+\beta_0^2)\operatorname{var}\eta_0 + \sigma^4}.$$

The relative efficiency (in the sense of Pitman) of two sequences of tests with an asymptotic power of the form as given can be defined as the squared quotient of the slopes of the tests.

One competitor of the test based on r_n is the likelihood ratio test in the incidental version of the model. Two times this log likelihood ratio statistic takes the form

$$2\log\frac{\sup_{\alpha,\beta,\sigma,z_1,\ldots,z_n}\prod_{i=1}^n\frac{1}{\sigma^2}\phi\Big(\frac{X_i-z_i}{\sigma}\Big)\phi\Big(\frac{Y_i-\alpha-\beta z_i}{\sigma}\Big)}{\sup_{\alpha,\sigma,z_1,\ldots,z_n}\prod_{i=1}^n\frac{1}{\sigma^2}\phi\Big(\frac{X_i-z_i}{\sigma}\Big)\phi\Big(\frac{Y_i-\alpha-\beta_0 z_i}{\sigma}\Big)}$$
$$= -2n\log\frac{\min_\beta(\beta^2S_X^2-2\beta S_{XY}+S_Y^2)/(1+\beta^2)}{(\beta_0^2S_X^2-2\beta_0S_{XY}+S_Y^2)/(1+\beta_0^2)}.$$

(Cf. Zhang (1994).) The minimum in the numerator is taken for the least squares estimator $\hat\beta_{LS}$. By standard arguments the likelihood ratio statistic can be expanded and be shown to be asymptotically equivalent to

$$n(\hat\beta_{LS}-\beta_0)^2/\tau^2,$$

for τ^2 the asymptotic variance under $\beta=\beta_0$ of $\hat\beta_{LS}$ given in (6.1) with $\beta=\beta_0$. It follows that the likelihood ratio statistic is asymptotically chi-squared with one degree of freedom under the null hypothesis, as usual. Furthermore, the asymptotic slope of the test that rejects the null hypothesis for values of the log likelihood statistics bigger than $\chi^2_{1,\alpha}$ is equal to τ^{-1}. Inspection of the formulas shows that τ^{-1} and s are identical, so that Creasy's test and the likelihood ratio test are asymptotically equivalent. (Zhang (1994) shows an interesting nonasymptotic connection between the two tests: the likelihood ratio test conditioned on the variables $\sum_{i=1}^n(X_i^2+Y_i^2), V_1,\ldots,V_n, \bar W_n$ is exactly the test based on r_n. The conditioning removes the dependence of the distribution of the likelihood ratio on the nuisance parameters $\alpha,\sigma,z_1,z_2,\ldots$.)

Finally consider the test based on the log likelihood ratio $K_n(\beta_0)$ defined in Section 1. According to Theorem 1.2 the sequence $K_n(\beta_0)$ is asymptotically chi-squared with one degree of freedom under the null hypothesis. Inspection of the proofs of Theorems 1.1 and Theorem 1.2 shows that

$$K_n(\beta_0)=n(\hat\beta-\beta_0)^2(\tilde I^{-1}_{\theta,\sigma,\eta_0})^{-1}_{2,2}+o_P(1),$$

for $\hat\beta$ the estimator of the slope suggested in this paper. Thus the squared slope of the test based on $K_n(\beta_0)$ is equal to $(\tilde I^{-1}_{\theta,\sigma,\eta_0})^{-1}_{2,2}$.

We conclude that the relative efficiency of the usual test and the test based on $K_n(\beta_0)$ is equal to

$$(\tilde I^{-1}_{\theta,\sigma,\eta_0})_{2,2}\Big(\frac{\sigma^2(1+\beta_0^2)}{\operatorname{var}\eta_0}+\frac{\sigma^4}{\operatorname{var}^2\eta_0}\Big)^{-1}.$$

Hence the situation for testing β is exactly the same as the situation for estimating β: in view of Theorem 6.1 the best based on $K_n(\beta_0)$ is strictly more efficient than Creasy's test or the likelihood ratio test, unless η_0 is Gaussian. Table 1 gives some insight in the relative efficiencies.

7. Some Technical Lemmas

LEMMA 7.1. *For every probability measure η_0 on $\mathbb{R}$ and compact set $K \subset (0,\infty)$ there exists a neighbourhood U of η_0 in the weak topology such that*

$$\sup_{\eta\in U, c\in K} \frac{\int |z|^j\, e^{zs}\, e^{-cz^2}\, d\eta(z)}{\int e^{zs}\, e^{-cz^2}\, d\eta(z)} \le C\Big(1+|s|\Big)^j,$$

for all $s \in \mathbb{R}$ and a constant C depending on j, U, η_0 and K only.

Proof. It suffices to show that the functions

$$h_{c,\eta}(s) = \frac{\int_0^\infty z^j\, e^{zs}\, e^{-cz^2}\, d\eta(z)}{\int e^{zs}\, e^{-cz^2}\, d\eta(z)} \quad \text{and} \quad \frac{\int_{-\infty}^0 |z|^j\, e^{zs}\, e^{-cz^2}\, d\eta(z)}{\int e^{zs}\, e^{-cz^2}\, d\eta(z)}$$

both can be bounded appropriately. We shall give the proof for the first; the second can be handled similarly.

Since the function $z \mapsto z^j 1\{z>0\}$ is nondecreasing on $\mathbb{R}$, the functions $h_{c,\eta}$ are nondecreasing in s. For $s \le 1$ they can be bounded by their value at 1

$$\frac{\int_0^\infty z^j\, e^z\, e^{-cz^2}\, d\eta(z)}{\int e^z\, e^{-cz^2}\, d\eta(z)} \to \frac{\int_0^\infty z^j\, e^z\, e^{-c_0 z^2}\, d\eta_0(z)}{\int e^z\, e^{-c_0 z^2}\, d\eta_0(z)},$$

as (c,η) converges to (c_0,η_0) with $c_0 > 0$.

Choose $z_0 \le 0$ such that $\eta_0(z_0,\infty) > 0$. For $s > 1$ and $z \ge r = (4s/c) \vee \sqrt{-2z_0 s/c}$ we have

$$zs - cz^2 = zs - \tfrac{1}{2}cz^2 - \tfrac{1}{2}cz^2 \le -zs + z_0 s \le -z + z_0 s.$$

Therefore, the function $h_{c,\eta}$ can for $s > 1$ be bounded by

$$r^j + \frac{\int_r^\infty z^j\, e^{-z}\, d\eta(z)}{\int e^{(z-z_0)s}\, e^{-cz^2}\, d\eta(z)}.$$

For c close to c_0 and η sufficiently close to η_0 there exists a constant L (depending on c_0 and z_0) such that this is bounded by

$$L(1+|s|^j) + \frac{\int_0^\infty z^j\, e^{-z}\, d\eta_0(z)}{\int_{z_0}^\infty e^{-c_0 z^2}\, d\eta_0(z)} + 1.$$

Conclude that for every c_0 there exist open neighbourhoods V of c_0 and U of η_0 and a constant C such that $h_{c,\eta}(s) \le C\bigl(1+|s|\bigr)^j$ for every s and every $\eta \in U$ and $c \in V$. The compact K is covered by the neighbourhoods V as c_0 ranges over K. For a finite subcover $K \subset \cup_i V_i$ take $U = \cap_i U_i$ and $C = \sup_i C_i$ to satisfy the requirements of the lemma. ■

LEMMA 7.2. *Let $0 < \alpha \le 1$ and k_1, k_2, k_3, k_4 be given integers. For every probability distribution η_0 on $\mathbb{R}$ and compact $K \subset (0,\infty)$ there exists a neighbourhood U of η_0 in the weak topology such that the class $\mathcal{F}$ of all functions*

$$s \mapsto \prod_{i=1}^{4} \Bigl(\frac{\int z^i e^{zs}\, e^{-cz^2}\, d\eta(z)}{\int e^{zs}\, e^{-cz^2}\, d\eta(z)} \Bigr)^{k_i},$$

with η ranging over U and c ranging over K, satisfies

$$\log N_{[]}\bigl(\varepsilon, \mathcal{F}, L_r(Q)\bigr) \le C\Bigl(\frac{1}{\varepsilon}\Bigr)^V \left(\sum_{j=-\infty}^{\infty} \bigl[(1+|j|^{\sum_i ik_i r + \alpha r})\, Q(j,j+1]\bigr]^{\frac{V}{V+r}} \right)^{\frac{V+r}{r}},$$

for every $r \ge 1$ and $V \ge 1/\alpha$ and measure Q on $\mathbb{R}$, and a constant C depending only on η_0, U, α, V, r, K and k_1, k_2, k_3, k_4.

Proof. Write $h_{c,\eta}(s)$ for the function in the display. In view of the preceding lemma we can find a neighbourhood U and a constant C_1 such that

$$\begin{aligned} \bigl|h_{c,\eta}(s)\bigr| &\le C_1\bigl(1+|s|\bigr)^{\sum ik_i} \\ \bigl|h'_{c,\eta}(s)\bigr| &\le C_1\bigl(1+|s|\bigr)^{1+\sum ik_i}. \end{aligned}$$

It follows that the restrictions of the functions $h_{c,\eta}$ to the interval $(j, j+1]$ are uniformly bounded by a multiple of $1+|j|^{\sum ik_i}$ and Lipschitz of order α with Lipschitz constant $1+|j|^{\sum ik_i + \alpha}$. The lemma now follows from Theorem 2.1 of Van der Vaart (1994). ■

LEMMA 7.3. *Let $0 < \alpha \le 1$ and k_0, k_1, k_2, k_3, k_4 be given integers. For every probability distribution η_0 on $\mathbb{R}$ and compact $K \subset (0,\infty)$ there exists an open neighbourhood U of η_0 in the weak topology such that the class $\mathcal{F}$ of all functions*

$$(x,y) \mapsto (a_0 + a_1 x + a_2 y)^{k_0} \prod_{i=1}^{4} \Bigl(\frac{\int z^i e^{z(b_0+b_1x+b_2y)}\, e^{-cz^2}\, d\eta(z)}{\int e^{z(b_0+b_1x+b_2y)}\, e^{-cz^2}\, d\eta(z)} \Bigr)^{k_i},$$

with η ranging over U, c ranging over K and a and b ranging over compacta in $\mathbb{R}^3$, satisfies

$$\log N_{[]}\bigl(\varepsilon, \mathcal{F}, L_r(P)\bigr) \le C\Bigl(\frac{1}{\varepsilon}\Bigr)^V \left(P\bigl(1+|x|+|y|\bigr)^r \sum_i ik_i + \alpha r + (V+r)/V + k_0 r + \delta \right)^{V/r},$$

for every $r \geq 1$ and $V \geq 1/\alpha$ and measure P on $\mathbb{R}^2$ and $\delta > 0$, and a constant C depending only on η_0, U, α, V, r, the compacta, δ and k_0, k_1, k_2, k_3, k_4.

Proof. Let U be the neighbourhood of the preceding lemma. Let $\mathcal{F}_{a,b}$ be the class of functions with a and b fixed and only η and c varying. Set $f(x,y) = (a_0 + a_1 x + a_2 y)^{k_0}$ and let $h_{c,\eta}(s)$ be as in the preceding lemma. A bracket $[l,u]$ for $h_{c,\eta}$ yields a bracket

$$\begin{aligned}\big[&f^+(x,y)l(b_0+b_1x+b_2y) - f^-(x,y)u(b_0+b_1x+b_2y),\\ &f^+(x,y)u(b_0+b_1x+b_2y) - f^-(x,y)l(b_0+b_1x+b_2y)\big]\end{aligned}$$

for the function $f(x,y)h_{c,\eta}(b_0+b_1x+b_2y)$. Its size in $L_r(P)$ is equal to the size of the bracket $[l,u]$ in $L_r(Q)$ for the measure Q defined by

$$Q(B) = \int 1_B(b_0+b_1x+b_2y)\,|f|^r(x,y)\,dP(x,y).$$

It follows that the bracketing numbers of the class $\mathcal{F}_{a,b}$ in $L_r(P)$ are bounded by the bracketing numbers of the class of functions $h_{c,\eta}$ in $L_r(Q)$. By Markov's inequality

$$Q(j,j+1] \leq \frac{Q|s|^p}{|j|^p} = \frac{P|b_0+b_1x+b_2y|^p|a_0+a_1x+a_2y|^{k_0r}}{|j|^p}.$$

Choose $(p - r(\sum ik_i + \alpha))V/(V+r) > 1$ and apply the preceding lemma to obtain the bound of the lemma on the bracketing numbers of the class $\mathcal{F}_{a,b}$ for every fixed (a,b), where the constant C can be chosen independently of (a,b).

In view of Lemma 1 the partial derivatives of the functions in $\mathcal{F}_{a,b}$ with respect to a and b are bounded by a multiple of the function $\big(1+|x|+|y|\big)^{k_0+2+\sum ik_i}$ and the functions themselves are bounded by $\big(1+|x|+|y|\big)^{k_0+\sum ik_i}$. Conclude that for any (a,b) and (a',b')

$$|g_{a,b} - g_{a',b'}| \leq \big\|(a,b)-(a',b')\big\|^\beta\, G,$$

for any $0 < \beta \leq 1$ and the function G defined by

$$G(x,y) = \big(1+|x|+|y|\big)^{k_0+2\beta+\sum ik_i}.$$

For $\beta = \alpha/2$ the $L_r(P)$-norm of this function is finite, whenever the right side of the lemma is finite. We can assume this without loss of generality. Construct brackets over the class $\mathcal{F}$ by first choosing an $\varepsilon^{1/\beta}/\|G\|_{P,r}$-net over the set of all (a,b). The number of elements in this net can be chosen bounded by $(C_2/\varepsilon)^{6/\beta}$ for some constant C_2. Next for every (a_i,b_i) in this net choose a minimal number of brackets $[l,u]$ over $\mathcal{F}_{a_i,b_i}$ and finally form the brackets $[l - \varepsilon G/\|G\|_{P,r}, u + \varepsilon G/\|G\|_{P,r}]$. These brackets cover $\mathcal{F}$ and have size proportional to ε. The total number of brackets obtained in this manner is bounded by

$$\Big(\frac{C_2}{\varepsilon}\Big)^{6/\beta} \sup_{a,b} N_{[]}\big(\varepsilon, \mathcal{F}_{a,b}, L_r(P)\big).$$

The logarithm of this expression is bounded by the right side of the lemma. ■

LEMMA 7.4. *Suppose that $\mathcal{F}$ is a class of functions $f:\mathbb{R} \to \mathbb{R}$ such that $|f|(z) \leq 1 + |z|^k$ for some k and such that the restrictions of the functions in $\mathcal{F}$ to a fixed interval $[-M, M]$ are equi-continuous for every M. Then $\int f\, d\eta_n \to \int f\, d\eta$ uniformly in f for every weakly convergent sequence of probability measures $\eta_n \rightsquigarrow \eta$ with $\int |z|^{k+\delta}\, d\eta_n(z) = O(1)$ for some $\delta > 0$.*

Proof. For every constant M we have

$$\Big|\int f\, d(\eta_n - \eta)\Big| \leq \sup_f \Big|\int_{-M}^{M} f\, d(\eta_n - \eta)\Big| + \int_{|z|>M} \big(1 + |z|^k\big)\, d(\eta_n + \eta)(z).$$

The lim sup of the second term on the right side can be made arbitrarily small by choice of M. Since the functions $f1_{[-M,M]}$ are uniformly bounded and equicontinuous on a set of η-probability one whenever $-M$ and M are continuity points of η, the first term converges to zero for almost every M. See e.g. Dudley (1976) or Van der Vaart and Wellner (1995), Theorem 1.12.1. ■

REFERENCES

[1] Andersen, N.T., Giné, E., Ossiander, M. and Zinn, J., (1988). The central limit theorem and the law of iterated logarithm for empirical processes under local conditions. *Probability Theory and Related Fields* **77**, 271–305.

[2] Anderson, T.W., (1984). Estimation of linear statistical relationships. *Annals of Statistics* **12**, 1–45.

[3] Begun, J.M., Hall, W.J., Huang, W. and Wellner, J.A., (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics* **11**, 432–452.

[4] Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J., (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore.

[5] Bickel, P.J. and Ritov, Y., (1987). Efficient estimation in the errors-in-variables model. *Annals of Statistics* **15**, 513–540.

[6] Creasy, M.A., (1956). Confidence limits for the gradient in the linear functional relationship. *Journal of the Royal Statistical Society b* **18**, 65–69.

[7] Dudley, R.M., (1976). *Probabilities and Metrics: Convergence of Laws on Metric Spaces. Mathematics Institute Lecture Note Series* **45**. Aarhus University.

[8] Fuller, W.A., (1987). *Measurement error models*. John Wiley and Sons, New York.

[9] Gleser, L.J., (1981). Estimation in a multivariate errors in variables regression model: large sample results. *Annals of Statistics* **9**, 24–44.

[10] Gleser, L.J. and Hwang, J.T., (1987). The nonexistence of 100(1-α)% confidence sets of finite expected diameter in errors-in-varables and related models. *Annals of Statistics* **15**, 1351–1362.

[11] Groeneboom, P.J., (1991). Nonparametric maximum likelihood estimators for interval censoring and deconvolution. Report **91-53**. Delft University of Technology.

[12] Jongbloed, G., (1995). *Three Statistical Inverse Problems*. Department of Mathematics, Delft University.

[13] Kendall, M.G. and Stuart, A., (1979). *The advanced theory of statistics 2* . Hafner, New York.

[14] Kiefer, J. and Wolfowitz, J., (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Nuisance Parameters. *Annals of Mathematical Statistics* **27**, 887–906.

[15] Lesperance, M.L. and Kalbfleisch, J.D., (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association* **87**, 120–126.

[16] Lindsay, B.G., (1983a). Efficiency of the conditional score in a mixture setting. *Annals of Statistics* **11**, 486–497.

[17] Lindsay, B.G., (1983b). The geometry of mixture likelihoods. *Annals of Statistics* **11**, 86-94.

[18] Lindsay, B.G., (1985). Using empirical Bayes inference for increased efficiency. *Annals of Statistics* **13**, 914–931.

[19] Murphy, S.A. and van der Vaart, A.W., (1995). Semiparametric Likelihood ratio inference. *preprint*.

[20] Neyman, J. and Scott, E.L., (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

[21] Pfanzagl, J., (1990). *Estimation in Semiparametric models. Lecture Notes in Statistics* **63**. Springer-Verlag, New York.

[22] Pfanzagl, J., (1993). Incidental versus random nuisance parameters. *Annals of Statistics* **21**, 1663–1691.

[23] Reiersøl, O., (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica* **18**, 375–389.

[24] Spiegelman, C., (1979). On estimating the slope of a straight line, when both variables are subject to error. *Annals of Statistics* **7**, 201–206.

[25] Van der Vaart, A.W., (1988a). *Statistical Estimation in Large Parameter Spaces. CWI tract* **44**. CWI, Amsterdam.

[26] Van der Vaart, A.W., (1988b). Estimating a parameter in incidental and structural models by approximate maximum likelihood. Report **139**. Dept. Statistics, University of Washington, Seattle.

[27] Van der Vaart, A.W., (1994). Bracketing smooth functions. *Stochastic Processes and Applications* **52**, 93–105.

[28] Van der Vaart, A.W., (1996). Efficient estimation in semiparametric models. *Annals of Statistics* **24**, to appear.

[29] Van der Vaart, A.W. and Wellner, J.A., (1995). *Weak Convergence and Empirical Processes*. Springer Verlag, New York.

[30] Wald, A., (1949). Note on the consistency of the maximum likelihood estimate. *Annals Math. Statist.* **20**, 595–601.

[31] Zhang, H., (1994). Confidence regions in linear functional relationships. *Annals of Statistics* **22**, 49–66.