The Annals of Statistics
1996, Vol. 24, No. 5, 2049–2057

MAXIMUM LIKELIHOOD ESTIMATION UNDER A SPATIAL SAMPLING SCHEME

By Aad van der Vaart

Vrije Universiteit

It is shown that the maximum likelihood estimator in a model used in the statistical analysis of computer experiments is asymptotically efficient.

1. Results. Motivated by the modelling of computer experiments, Ying (1993) considers estimation of the parameter (λ, μ, σ^2) based on a matrix-valued observation $X = (X_{i,k})$ (where $i = 1, \dots, m$ and $k = 1, \dots, n$) from a multivariate normal distribution with mean zero and covariances given by

$$\operatorname{cov}\bigl(X_{i,k}, X_{j,l}\bigr) = \sigma^2 \exp\bigl(-\lambda|u_i - u_j| - \mu|v_k - v_l|\bigr).$$

The grids $0 \le u_1 < u_2 < \cdots < u_m \le 1$ and $0 \le v_1 < v_2 < \cdots < v_n \le 1$ are known to the experimenter and the parameters λ and μ are positive. Suppose that the grids become dense in $[0, 1]$ in such a way that

$$\max|u_{i+1} - u_i| = o(m^{-1/2}); \qquad \max|v_{k+1} - v_k| = o(n^{-1/2}). \tag{1.1}$$

Under this condition, Ying (1993) establishes asymptotic normality of the maximum likelihood estimator $(\hat\lambda, \hat\mu, \hat\sigma^2)$ as $m, n \to \infty$ in such a way that $n/m \to \rho \in (0, \infty)$. Precisely,

$$\sqrt{n}\begin{pmatrix} \hat\lambda - \lambda \\ \hat\mu - \mu \\ \hat\sigma^2 - \sigma^2 \end{pmatrix} \to_d \left(0, \begin{pmatrix} \dfrac{2\lambda^2}{1+\lambda} & 0 & \dfrac{-2\sigma^2\lambda}{1+\lambda} \\ 0 & \dfrac{2\mu^2}{1+\mu}\rho & \dfrac{-2\sigma^2\mu}{1+\mu}\rho \\ \dfrac{-2\sigma^2\lambda}{1+\lambda} & \dfrac{-2\sigma^2\mu}{1+\mu}\rho & \dfrac{2\sigma^4}{1+\lambda} + \dfrac{2\sigma^4}{1+\mu}\rho \end{pmatrix}\right). \tag{1.2}$$

Somewhat surprisingly the "usual" theory concerning asymptotic efficiency of the maximum likelihood estimator does not apply. However, Ying (1993) conjectures that the maximum likelihood estimator is nevertheless asymptotically efficient. In this note we show this to be true.

Since the data in the model are dependent, the "usual" theory should refer to the general concept of local asymptotic normality due to Le Cam and Hájek. [See Ibragimov and Hasminskii (1981) for a discussion.] It turns out

Received July 1992; revised August 1993.

AMS 1991 *subject classifications*. Primary 62F12; secondary 60G60, 62M30.

Key words and phrases. Asymptotic efficiency, maximum likelihood, local asymptotic normality.

that the model is not locally asymptotically normal in its natural parametrization, but can be reparametrized so as to make the "usual" techniques apply. Precisely, let $L_{m,n}(\lambda, \mu, \sigma^2)$ be the log likelihood for the model. Local asymptotic normality of the model in its natural parametrization would entail that for $\delta_n = 1/\sqrt{n}$ and every fixed (λ, μ, σ^2) the "local log likelihood"

$$L_{m,n}\big(\lambda + a\delta_n, \mu + b\delta_n, \sigma^2 + c\delta_n\big) \tag{1.3}$$

has a linear expansion of the form

$$L_{m,n}(\lambda, \mu, \sigma^2) + (a,b,c)\Delta_n + (a,b,c)J(a,b,c)^t + o_P(1)$$

for a sequence of random vectors Δ_n that converges to a normal $N_3(0, J)$ distribution. [By "linear" it is understood that the stochastic part of the expansion is linear in (a, b, c).] However, in Section 3 it is shown that this expansion is valid for $\delta_n = 1/n$ (with J nondegenerate), rather than $\delta_n = 1/\sqrt{n}$. The expansion with $\delta_n = 1/n$ would suggest that the rate of the maximum likelihood estimator is suboptimal. That the maximum likelihood estimator is in fact asymptotically efficient can be seen by using a nonlinear "localization." The localized log likelihood

$$L_{m,n}\left(\lambda + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{m}}, \frac{\lambda\mu\sigma^2 + c/\sqrt{mn}}{\big(\lambda + a/\sqrt{n}\big)\big(\mu + b/\sqrt{m}\big)}\right) \tag{1.4}$$

does permit the required expansion. Moreover, this localization leads to convergence to a Gaussian shift experiment, and standard efficiency theory can be applied in a nonstandard manner. The third local parameter in (1.4) is up to order $O(1/n)$ equivalent to

$$\sigma^2 - \frac{\sigma^2 a}{\lambda\sqrt{n}} - \frac{\sigma^2 b}{\mu\sqrt{m}} + \frac{c}{\lambda\mu\sqrt{mn}} + \frac{\sigma^2 ab}{\lambda\mu\sqrt{mn}} + \frac{\sigma^2 a^2}{n\lambda^2} + \frac{\sigma^2 b^2}{m\mu^2}. \tag{1.5}$$

Thus the perturbation of σ^2 due to the free parameter c is of lower order than the perturbation due to the local parameters a and b connected to λ and μ. The explanation is that σ^2 is confounded with (λ, μ). As shown by Ying (1993), the parameter $\lambda\mu\sigma^2$ is estimable at rate n. Thus knowledge of two out of three of the parameters would make it possible to estimate the unknown third parameter at rate n also. If all three parameters are unknown, the rate drops to $\sqrt{n}$. This fact should be incorporated in the localization. Section 3 contains more remarks on this point.

Actually, the given nonlinear localization corresponds to a linear one for the parameter $(\lambda, \mu, \tau) = (\lambda, \mu, \sigma^2\lambda\mu)$. In terms of this new parameter, the sequence of models is locally asymptotically normal in a standard manner.

THEOREM 1.1. *Let* (1.1) *hold. Then for every fixed* (λ, μ, σ^2) *the localized* log *likelihood* (1.4) *behaves as*

$$L_{m,n}(\lambda, \mu, \sigma^2) + (a,b,c)\Delta_n - \tfrac{1}{2}(a,b,c)J(a,b,c)^t + o_P(1),$$

where J is the diagonal matrix $J = \mathrm{diag}(\lambda^{-1} + \lambda^{-2}, \mu^{-1} + \mu^{-2}, \tau^{-2})/2$ *and* Δ_n *is the sequence of random vectors given by* (2.5). *Furthermore, the sequence* Δ_n *converges in distribution to* $N_3(0, J)$.

Given this theorem, standard results such as the convolution and minimax theorem designate estimator sequences such that $(\sqrt{n}(\hat{\lambda} - \lambda), \sqrt{m}(\hat{\mu} - \mu), \sqrt{mn}(\hat{\tau} - \tau))$ tends in distribution to a normal $N(0, J^{-1})$ distribution as asymptotically optimal. In view of Ying's (1993) results,the maximum likelihood estimator is asymptotically optimal. (Marginal limit distributions of the estimator are given in his Theorem 2; the joint limit distribution follows from the expansions in his proof on pages 1582 and 1584–1585.)

Since the original parameter (λ, μ, σ^2) is a smooth function of the new parameter, the optimality of $(\hat{\lambda}, \hat{\mu}, \hat{\tau})$ must somehow carry over onto $(\hat{\lambda}, \hat{\mu}, \hat{\sigma}^2)$, where $\hat{\sigma}^2 = \hat{\tau}/(\hat{\lambda}\hat{\mu})$. This is not entirely trivial. The following two results make this concrete.

COROLLARY 1.2. *Under the conditions of the theorem, let* $T_n = (T_{n1}, T_{n2}, T_{n3})$ *be estimators such that the sequence* $(\sqrt{n}(T_{n1} - \lambda), \sqrt{n}(T_{n2} - \mu), \sqrt{n}(T_{n3} - \sigma^2))$ *possesses a limit distribution under* (λ, μ, σ^2) *for almost every* (λ, μ, σ^2) *in an open set. Then this limit distribution is the convolution of the normal distribution in* (1.2) *and some other probability distribution for Lebesgue almost all* (λ, μ, σ^2).

COROLLARY 1.3. *Under the conditions of the theorem, for every estimator sequence* T_n *and subconvex loss function* $l: \mathbb{R}^3 \to [0, \infty)$ *and every* (λ, μ, σ^2), *the local minimax risk*

$$\sup_I \liminf_{n\to\infty} \sup_{(a,b)\in I} \mathrm{E}_{\lambda + a/\sqrt{n},\, \mu + b/\sqrt{m},\, \sigma^2_{m,n}(a,b)} l\left(\sqrt{n}\begin{pmatrix} T_{n1} - \lambda - a/\sqrt{n} \\ T_{n2} - \mu - b/\sqrt{m} \\ T_{n3} - \sigma^2_{m,n}(a,b) \end{pmatrix}\right)$$

is bounded below by $\int l\, dN(0, \Sigma)$ *where* $N(0, \Sigma)$ *is the normal distribution in* (1.2). *Here* $\sigma^2_{m,n}(a,b) = \tau(\lambda + a/\sqrt{n})^{-1}(\mu + b/\sqrt{m})^{-1}$ *and the first supremum is taken over all finite subsets* $I \subset \mathbb{R}^2$.

PROOFS. According to the theorem the sequence of models is locally asymptotically normal for the parameter (λ, μ, τ). For fixed (λ, μ, τ) let $\mathscr{E}_{m,n}(\lambda, \mu, \tau)$ be the experiment with parameter $(a, b, c) \in \mathbb{R}^3$ corresponding to observing an $m \times n$ matrix X distributed according to the normal distribution with parameters $(\lambda + a/\sqrt{n}, \mu + b/\sqrt{m}, \tau + c/\sqrt{mn})$. Then the sequence $\mathscr{E}_{m,n}(\lambda, \mu, \tau)$ converges to the experiment of observing one observation Δ from the $N_3((a, b, c), J^{-1})$ distribution. Convergence of experiments was introduced by Le Cam (1972). van der Vaart (1991) gives a review that is appropriate in the present context.

Define functionals $\kappa_{m,n}(\lambda, \mu, \tau) = (\lambda, \mu, \tau/\lambda\mu)$. These functionals are differentiable with respect to the localization at every (λ, μ, τ) in the sense that

$$\sqrt{n}\left(\kappa_{m,n}\left(\lambda + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{m}}, \tau + \frac{c}{\sqrt{mn}}\right) - \kappa_{m,n}(\lambda, \mu, \tau)\right)$$
$$\to \left(a, \sqrt{\rho}\, b, -\frac{\sigma^2}{\lambda}a - \frac{\sigma^2}{\mu}\sqrt{\rho}\, b\right) = \kappa'(a, b, c).$$

Le Cam's theory of convergence of experiments now implies that estimating the functionals $\kappa_{m,n}(\lambda + a/\sqrt{n}, \mu + b/\sqrt{m}, \tau + c/\sqrt{mn})$ in $\mathscr{E}_{m,n}(\lambda, \mu, \sigma^2)$ is asymptotically not easier than estimating $\kappa'(a, b, c)$ in the limit experiment. The two corollaries make this concrete.

By assumption, the sequence $\sqrt{n}(T_n - \kappa_{m,n}(\lambda, \mu, \tau))$ possesses a limit distribution $L_{\lambda,\mu,\tau}$ under almost every (λ, μ, τ). This implies that the sequence

$$\sqrt{n}\left(T_n - \kappa_{m,n}\left(\lambda + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{m}}, \tau + \frac{c}{\sqrt{mn}}\right)\right)$$

possesses the same limit distribution under the parameter $\lambda + a/\sqrt{n}, \mu + b/\sqrt{m}, \tau + c/\sqrt{mn}$ for almost every $(a, b, c) \in \mathbb{R}^3$, for almost every $(\lambda, \mu, \tau) \in \mathbb{R}^3$. The argument is similar to the argument given by Jeganathan (1981). [Also see Le Cam (1986) and van der Vaart (1996).] At (λ, μ, τ) such that this is true, the sequence T_n is "almost regular" in the sense of Hájek (1970) at almost all (λ, μ, τ). By Hájek's convolution theorem, $L_{\lambda,\mu,\tau}$ is the distribution of the sum of $\kappa'(\Delta)$ [under $(a, b, c) = 0$] and an independent variable. Since $\kappa'(\Delta)$ is distributed as the right side of (1.2), this gives the first corollary.

The display in the second corollary gives the limiting local minimax risk for estimating the functionals $\kappa_{m,n}$ in the experiments $\mathscr{E}_{m,n}(\lambda, \mu, \tau)$. This is bounded below by the minimax risk in the limit experiment for estimating $\kappa'(a, b, c)$. The minimax estimator in the limit experiment is $\kappa'(\Delta)$ and this has constant risk $\int l\, dN(0, \Sigma)$. □

Although Corollary 1.3 defines the local minimax risk in an unusual manner (it is based on a two-dimensional local submodel in a three-dimensional parameter space), its interpretation is as usual. For instance, the expression displayed in Corollary 1.3 is a lower bound for

$$\liminf_{n\to\infty} \sup_{|\tilde{\lambda}-\lambda|+|\tilde{\mu}-\mu|+|\tilde{\sigma}^2-\sigma^2|<\varepsilon} \mathrm{E}_{\tilde{\lambda},\tilde{\mu},\tilde{\sigma}^2} l\left(\sqrt{n}\left(T_n - \left(\tilde{\lambda}, \tilde{\mu}, \tilde{\sigma}^2\right)\right)\right)$$

for every $\varepsilon > 0$. The preceding two results are only two examples of how efficiency of the maximum likelihood estimator might be expressed.

2. Proof of Theorem 1.1. The proof of Theorem 1.1 consists of somewhat tedious Taylor expansions. Below we describe the general structure.

Form a vector from the matrix X by putting the second column under the first, next the third under the first two, etcetera. Let $Y_{1,1}$ be zero and let $Y_{i,k}$ be the "innovation" defined as $X_{i,k}$ minus the conditional expectation of $X_{i,k}$ given the preceding $X_{j,l}$ (the ones above $X_{i,k}$) for the other indices. Let

$v_{i,k} = \mathrm{E}Y_{i,k}^2$. Since the innovations are orthogonal and form a multivariate normal vector, they are independent. Furthermore, the original $X_{i,k}$ can be regained from the innovations as $X = LY$, where L is an $(mn \times mn)$ matrix with zeros above and ones on the diagonal. Thus $\mathrm{cov}(X) = L\,\mathrm{diag}(v_{i,k})L^t$ and -2 log likelihood equals

$$\begin{aligned} -2L_{m,n}(\lambda, \mu, \sigma^2) & \\ &= mn\log(2\pi) + \log\det\mathrm{cov}(X) + X^t\,\mathrm{cov}(X)^{-1}X \\ &= mn\log(2\pi) + \sum\sum\left(\frac{Y_{i,k}^2}{v_{i,k}} + \log v_{i,k}\right). \end{aligned} \tag{2.1}$$

This is well known. In the present case the innovations are surprisingly simple. Let $\xi_i = u_i - u_{i-1}$, $\zeta_k = v_k - v_{k-1}$ and

$$\begin{aligned} U_{i,k} &= X_{i,k} - e^{-\mu\zeta_k}X_{i,k-1}, \\ V_{i,k} &= X_{i,k} - e^{-\lambda\xi_i}X_{i-1,k}, \\ Y_{i,k} &= X_{i,k} - e^{-\lambda\xi_i}X_{i-1,k} - e^{-\mu\zeta_k}X_{i,k-1} + e^{-\lambda\xi_i-\mu\zeta_k}X_{i-1,k-1}, \end{aligned}$$

where $X_{i,k}$ is defined as zero if $i = 0$ or $k = 0$. It can be checked that the $Y_{i,k}$ are indeed the innovations. Thus the projection of $X_{i,k}$ on all the preceding ones is actually the same as the projection of $X_{i,k}$ on just three preceding ones. It turns out that $Y_{i,k}$ is also orthogonal to all $X_{j,l}$ with $j < i$. Similarly $U_{i,k}$ is orthogonal to all $X_{j,l}$ with $l < k$ and $V_{i,k}$ is orthogonal to all $X_{j,l}$ with $j < i$. Figure 1 illustrates these facts, which imply many useful independences and zero correlations. It is clear that the $Y_{i,k}$ in the first column or row are special. Even though there are relatively few of them, they give nonzero contributions in the Taylor expansion. Setting $\psi(x) = (1 - e^{-2x})/x$, we calculate

$$\begin{aligned} v_{i,k} = \mathrm{E}Y_{i,k}^2 &= \sigma^2(1 - e^{-2\lambda\xi_i})(1 - e^{-2\mu\zeta_k}) \\ &= \tau\xi_i\zeta_k\psi(\lambda\xi_i)\psi(\mu\zeta_k), \qquad i, k \geq 2, \\ \mathrm{E}U_{i,k}^2 &= \sigma^2(1 - e^{-2\mu\zeta_k}) = \tau\zeta_k\psi(\mu\zeta_k)/\lambda, \qquad k \geq 2, \\ \mathrm{E}V_{i,k}^2 &= \sigma^2(1 - e^{-2\lambda\xi_i}) = \tau\xi_i\psi(\lambda\xi_i)/\mu, \quad i \geq 2. \end{aligned}$$

Note also that $Y_{1,k} = U_{1,k}$ and $Y_{i,1} = V_{i,1}$.

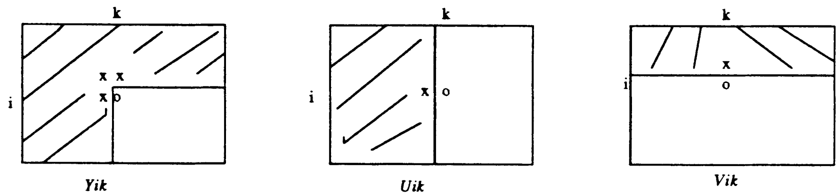

FIG. 1. *The $X_{i,k}$ at the "o" is projected on the $X_{j,l}$ in the shaded area. The projection happens to sit at the crosses.*

Write $(\tilde{\lambda}, \tilde{\mu}, \tilde{\tau})$ for $(\lambda + a/\sqrt{n}, \mu + b/\sqrt{m}, \tau + c/\sqrt{mn})$ and also use tildes for corresponding functions of the parameters, such as $\tilde{v}_{i,k}$ for the value of $v_{i,k}$ at $(\tilde{\lambda}, \tilde{\mu}, \tilde{\tau})$ and $\tilde{Y}_{i,k}$ for $Y_{i,k}$ with the parameters replaced by their versions with tildes. We need to expand the difference of the expression in (2.1) evaluated with and without tildes. We first show the effect of replacing $\tilde{v}_{i,k}$ by $v_{i,k}$. Since $\psi(x) = 2(1 - x + 2x^2/3 + O(x^3))$ as $x \to 0$, we have for $i, k \geq 2$,

$$\frac{v_{i,k}}{\tilde{v}_{i,k}} = 1 - \frac{c}{\tau\sqrt{mn}} + o(n^{-1}),$$

uniformly in (i, k), in view of (1.1). Using the expansion $\log(1 + v) = v - v^2/2 + O(v^3)$, we obtain, with the double sums understood to be over the indices $i, k \geq 2$,

$$\begin{aligned} &\sum\sum\left(\frac{\tilde{Y}_{i,k}^2}{\tilde{v}_{i,k}} + \log \tilde{v}_{i,k}\right) - \sum\sum\left(\frac{\tilde{Y}_{i,k}^2}{v_{i,k}} + \log v_{i,k}\right) \\ &\quad = \sum\sum\left(\frac{\tilde{Y}_{i,k}^2}{v_{i,k}} - 1\right)\left(\frac{v_{i,k}}{\tilde{v}_{i,k}} - 1\right) + \sum\sum\frac{1}{2}\left(\frac{v_{i,k}}{\tilde{v}_{i,k}} - 1\right)^2 + o(1) \\ &\quad = \frac{-c}{\tau\sqrt{mn}}\sum\sum\left(\frac{\tilde{Y}_{i,k}^2}{v_{i,k}} - 1\right) + \frac{c^2}{2\tau^2} + o_P(1), \end{aligned} \tag{2.2}$$

under the assumption that the random element in the last line is $O_P(1)$. The latter is shown below and also that the $\tilde{Y}_{i,k}$ can be replaced by $Y_{i,k}$ at the cost of a further $o_P(1)$ term.

The analogous sums over the first column and row are not negligible. By the same method as before we obtain, for the first row,

$$\begin{aligned} &\sum\left(\frac{\tilde{Y}_{1,k}^2}{\tilde{v}_{1,k}} + \log \tilde{v}_{1,k}\right) - \sum\left(\frac{\tilde{Y}_{1,k}^2}{v_{1,k}} + \log v_{1,k}\right) \\ &\quad = \frac{a}{\lambda\sqrt{n}}\sum\left(\frac{\tilde{Y}_{1,k}^2}{v_{1,k}} - 1\right) + \frac{1}{2}\frac{a^2}{\lambda^2} + o_P(1), \end{aligned} \tag{2.3}$$

under the assumption that the random element in the last line is $O_P(1)$. The first column gives a similar contribution.

Next consider expansion of $\tilde{Y}_{i,k}^2$ around $Y_{i,k}^2$. Some algebra yields

$$\begin{aligned} \tilde{Y}_{i,k} - Y_{i,k} &= (e^{-\lambda\xi_i} - e^{-\tilde{\lambda}\xi_i})U_{i-1,k} + (e^{-\mu\zeta_k} - e^{-\tilde{\mu}\zeta_k})V_{i,k-1} \\ &\quad + (e^{-\lambda\xi_i} - e^{-\tilde{\lambda}\xi_i})(e^{-\mu\zeta_k} - e^{-\tilde{\mu}\zeta_k})X_{i-1,k-1} \\ &= A_{i,k} + B_{i,k} + C_{i,k} \quad \text{(say)}. \end{aligned}$$

Thus with $R = 2YC + 2AB + 2BC + 2AC + C^2$, we obtain

$$\tilde{Y}_{i,k}^2 = Y_{i,k}^2 + 2Y_{i,k}(A_{i,k} + B_{i,k}) + A_{i,k}^2 + B_{i,k}^2 + R_{i,k} \quad (\text{say}).$$

With the double sums again understood to be over $i, k \geq 2$, we have

$$\sum\sum \frac{A_{i,k}^2}{v_{i,k}} \to_P \frac{a^2}{2\lambda}; \qquad \sum\sum \frac{B_{i,k}^2}{v_{i,k}} \to_P \frac{b^2}{2\mu}; \qquad \sum\sum \frac{R_{i,k}}{v_{i,k}} \to_P 0. \tag{2.4}$$

The first two statements can be proved by calculating means and variances of the sums, where we use the projection properties illustrated in Figure 1. For instance, since $U_{i,k}$ and $U_{j,l}$ are independent for $k \neq l$,

$$\begin{aligned}\operatorname{var}\left(\sum\sum \frac{A_{i,k}^2}{v_{i,k}}\right) &= O\left(\frac{1}{n^2}\right)\sum_i\sum_j\sum_k \frac{\xi_i^2\xi_j^2}{v_{i,k}v_{j,k}}\operatorname{cov}\left(U_{i-1,k}^2, U_{j-1,k}^2\right)\\ &\leq O\left(\frac{1}{n^2}\right)\sum_i\sum_j\sum_k \frac{\xi_i\,\xi_j}{\zeta_k^2}\operatorname{var}\left(U_{i-1,k}^2\right)\operatorname{var}\left(U_{j-1,k}^2\right) = O\left(\frac{1}{n}\right).\end{aligned}$$

The $R_{i,k}$ in the third statement of (2.4) can be broken into two parts. The part of R corresponding to $YC + AB$ can be handled by computing second moments, where we can use that $A_{i,k}B_{i,k}$ and $A_{j,l}B_{j,l}$ are uncorrelated and have mean zero for $(i,k) \neq (j,l)$. Thus

$$\mathrm{E}\left(\sum\sum \frac{A_{i,k}B_{i,k}}{v_{i,k}}\right)^2 = O\left(\frac{1}{n^2}\right)\sum\sum \frac{\zeta_i^2\zeta_k^2}{v_{i,k}^2}\mathrm{E}U_{i-1,k}^2\mathrm{E}V_{i,k-1}^2 = O\left(\frac{1}{n^2}\right),$$

where we use that $A_{i,k}$ and $B_{i,k}$ are also uncorrelated, hence independent. The part of R corresponding to $2BC + 2AC + C^2$ can be handled with the inequality $|C_{i,k}| \leq |\tilde{\lambda} - \lambda|\xi_i|\tilde{\mu} - \mu|\zeta_k|X_{i-1,k-1}|$ followed by computing means. For instance,

$$\mathrm{E}\left|\sum\sum \frac{A_{i,k}C_{i,k}}{v_{i,k}}\right| = O\left(\frac{1}{n\sqrt{n}}\right)\sum\sum \xi_i \mathrm{E}|U_{i-1,k}||X_{i-1,k-1}| = O\left(\frac{1}{n}\right).$$

Considering (2.4) proved, we obtain with the double sums being over $i, k \geq 2$,

$$\sum\sum \frac{\tilde{Y}_{i,k}^2}{v_{i,k}} = \sum\sum \frac{Y_{i,k}^2}{v_{i,k}} + \sum\sum 2\frac{Y_{i,k}}{v_{i,k}}(A_{i,k} + B_{i,k}) + \frac{a^2}{2\lambda} + \frac{b^2}{2\mu} + o_P(1).$$

Here the $A_{i,k}$ and $B_{i,k}$ can be replaced by their expansions $3(\tilde{\lambda} - \lambda)\xi_i U_{i-1,k}$ and $(\tilde{\mu} - \mu)\zeta_k V_{i,k-1}$, respectively. In the corresponding sums over the first row and column the $\tilde{Y}_{i,k}$ can be replaced by the $Y_{i,k}$ at the cost of only a $o_P(1)$ term.

Combination of the last display with (2.1), (2.2) and (2.3) yields the expansion of the theorem with

$$\Delta_n = \begin{pmatrix} -\frac{1}{\sqrt{n}}\Sigma_{i\geq 2}\Sigma_{k\geq 2}\frac{Y_{i,k}}{v_{i,k}}\xi_i U_{i-1,k} - \frac{1}{\sqrt{n}}\frac{1}{2\lambda}\Sigma_{k\geq 2}\left(\frac{Y_{1,k}^2}{v_{1,k}} - 1\right) \\ -\frac{1}{\sqrt{m}}\Sigma_{i\geq 2}\Sigma_{k\geq 2}\frac{Y_{i,k}}{v_{i,k}}\zeta_k V_{i,k-1} - \frac{1}{\sqrt{m}}\frac{1}{2\mu}\Sigma_{i\geq 2}\left(\frac{Y_{i,1}^2}{v_{i,1}} - 1\right) \\ \frac{1}{\sqrt{mn}}\frac{1}{2\tau}\Sigma_{i\geq 2}\Sigma_{k\geq 2}\left(\frac{Y_{i,k}^2}{v_{i,k}} - 1\right) \end{pmatrix}. \tag{2.5}$$

The sequence Δ_n can be seen to be asymptotically normal by a martingale central limit theorem. See page 1585 of Ying (1993) for more details. □

3. Invalidity of standard local asymptotic normality. This section contains a number of additional results meant to illustrate that the local parametrization chosen in Theorem 1.1 is the right one.

Inspection of the proof of Theorem 1.1 shows that the expansion for the local likelihood (1.4) is valid uniformly in (a, b, c) ranging over compacta. This allows us to obtain expansions for several other localizations. First the third parameter in (1.4) can be replaced by its expansion (1.5) at the cost of a $o_P(1)$ term. Thus the quadratic localization

$$L_{m,n}\Bigg(\lambda + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{m}},$$
$$\sigma^2 - \frac{\sigma^2 a}{\lambda\sqrt{n}} - \frac{\sigma^2 b}{\mu\sqrt{m}} + \frac{c}{\lambda\mu\sqrt{mn}} + \frac{\sigma^2 ab}{\lambda\mu\sqrt{mn}} + \frac{\sigma^2 a^2}{n\lambda^2} + \frac{\sigma^2 b^2}{m\mu^2}\Bigg)$$

possesses the linear expansion given by Theorem 1.1. Next the terms involving a^2, b^2 and ab could be absorbed into c. This would yield a *quadratic* expansion for the linear localization

$$L_{m,n}\left(\lambda + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{m}}, \sigma^2 - \frac{\sigma^2 a}{\lambda\sqrt{n}} - \frac{\sigma^2 b}{\mu\sqrt{m}} + \frac{c}{\lambda\mu\sqrt{mn}}\right).$$

This expansion contains the same information as the expansion of Theorem 1.1. It leads to convergence of the sequence of local experiments to a Gaussian experiment, but not a Gaussian location experiment. Since it is easier to work with a location experiment, the parametrization of Theorem 1.1 seems preferable.

It is also possible to replace the pair (a, b) in the first display of this section by $(a/\sqrt{n}, b/\sqrt{m})$ and next absorb in c the terms involving a and b

that arise in the third parameter. This gives the expansion

$$L_{m,n}\left(\lambda + \frac{a}{n}, \mu + \frac{b}{m}, \sigma^2 + \frac{c}{\sqrt{mn}}\right)$$
$$= L_{m,n}(\lambda, \mu, \sigma^2) + \left(\frac{\sigma^2 a\mu}{\sqrt{\rho}} + \sigma^2 b\lambda\sqrt{\rho} + c\lambda\mu\right)\Delta_{n3}$$
$$- \frac{1}{2}\left(\frac{\sigma^2 a\mu}{\sqrt{\rho}} + \sigma^2 b\lambda\sqrt{\rho} + c\lambda\mu\right)^2 J_{33} + o_P(1).$$

This expansion is less informative than the expansion of Theorem 1.1. However, it is of interest, because it shows that the original model is locally asymptotically normal in its standard parametrization with rate $1/n$. It seems intuitively clear that it cannot be locally asymptotically normal with rate $1/\sqrt{n}$ at the same time. For instance, taking $b = c = 0$ in the preceding display, we can see that the models of the observations under $(\lambda + a/n, \mu, \sigma^2)$ and (λ, μ, σ^2) are contiguous with Hellinger distance bounded away from zero. It is reasonable to expect that the models under $(\lambda + a/\sqrt{n}, \mu, \sigma^2)$ and (λ, μ, σ^2) will be asymptotically further apart, indeed, will be asymptotically orthogonal, contradicting local asymptotic normality with rate $1/\sqrt{n}$. We shall not prove this orthogonality, but note that it can be seen without calculations that local asymptotic normality with rate $1/\sqrt{n}$ is not valid. If it were valid, then the parameter λ would not be estimable at rate faster than $1/\sqrt{n}$, not even when μ and σ^2 are known. However, the parameter $\tau = \lambda\mu\sigma^2$ is always estimable at rate $1/n$ [as shown by Ying (1993)], so that, given knowledge of μ and σ^2, λ is estimable at rate n as well.

REFERENCES

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimators. *Z. Wahrsch. Verw. Gebiete* **14** 323–330.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

JEGANATHAN, P. (1981). On a decomposition of the limit distribution of a sequence of estimators. *Sankhyā Ser. A* **43** 26–36.

LE CAM, L. (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 245–261. Univ. California Press, Berkeley.

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

VAN DER VAART, A. (1991). An asymptotic representation theorem. *Internat. Statist. Rev.* **59** 97–121.

VAN DER VAART, A. W. (1996). On superefficiency. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. Yang, eds.). To appear.

YING, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Ann. Statist.* **21** 1567–1590.

DEPT MATHEMATICS
VRIJE UNIVERSITEIT
DE BODELAAN 1081A
1081 HV AMSTERDAM
NETHERLANDS
E-MAIL: aad.van-der-vaart@math.u-psud.fr