The Annals of Statistics
2021, Vol. 49, No. 4, 1836–1849
https://doi.org/10.1214/21-AOS2056

STEIN 1956: EFFICIENT NONPARAMETRIC TESTING AND ESTIMATION

BY A. W. VAN DER VAART[1] AND J. A. WELLNER[2]

[1]*Mathematical Institute, Leiden University, avdvaart@math.leidenuniv.nl*
[2]*Department of Statistics, University of Washington, jaw@stat.washington.edu*

We revisit a paper by Charles Stein, and discuss its follow-up.

1. Introduction. Stein's paper [52] is less than 9 pages in length, but can be viewed as the seminal paper for semiparametric statistics, making the transition from the asymptotic theory for regular parametric models to the estimation of smooth functionals on infinite-dimensional models.

Stein is modest about his achievements and writes in the Introduction:

> Very few results are obtained here, and, with the exception of the lemma of Section 3, they are not rigorous. Also, even for the example of Section 4, where a definite procedure is given, the results are not of immediate practical value. The computations required are excessive, and the procedure is not efficient for sample sizes likely to occur in practice.

It is true that the paper is mostly about information calculations that suggest that certain statistical procedures may exist. However, not only did the paper have much follow-up in constructing these procedures, it also motivated a theory of information calculus on infinite-dimensional models.

We start by revisiting Stein's paper, and next discuss some follow-up.

2. Stein's paper. For a smoothly parametrised finite-dimensional statistical model, the best possible quality of an estimator of a function of the parameter can be characterised by the Fisher information matrix. If θ is the parameter and p_θ, the density of a single observation X, then the *Fisher information matrix* is

$$I_\theta = \mathrm{E}_\theta \dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^T = \mathrm{Cov}_\theta\big(\dot{\ell}_\theta(X)\big),$$

where the gradient (i.e., vector of partial derivatives)

$$\dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta}\log p_\theta(x),$$

is the *score function* of the model. The smallest possible mean square error (in a certain asymptotic sense) of an estimator of a parameter $\phi(\theta) \in \mathbb{R}$ based on a sample of n observations) is then (for $\nabla\phi$ the gradient of ϕ)

$$\frac{1}{n}\nabla\phi(\theta)^T I_\theta^{-1}\nabla\phi(\theta). \tag{1}$$

Under some regularity conditions (Stein refers to [33]), the maximum likelihood estimator $\hat{\theta}_n$ based on an i.i.d. sample $X_1, \ldots, X_n$ from p_θ satisfies that $\sqrt{n}(\hat{\theta}_n - \theta)$ tends to a normal distribution with mean zero and variance $\nabla\phi(\theta)^T I_\theta^{-1}\nabla\phi(\theta)$, and hence attains this bound.

Even if the parameter set Θ may be multi-dimensional, this lower bound for estimation of a real-valued parameter $\phi(\theta)$ can already be obtained from considering a one-dimensional

Received November 2020; revised January 2021.
MSC2020 subject classifications. Primary 62F12; secondary 62G05, 62G20.
Key words and phrases. Semiparametric efficiency, efficient score function.

submodel. Specifically, for given θ_0 and $v_0 = I_{\theta_0}^{-1}\nabla\phi(\theta_0)$, one might consider the submodel $p_{\theta_0+tv_0}$ parameterised by t in a neighbourhood of $0 \in \mathbb{R}$. Then estimating $\phi(\theta_0 + tv_0)$ for unknown t is easier than estimating $\phi(\theta_0 + h)$ for $h = \theta - \theta_0 \in \Theta$, but one can readily work out that the information bounds are the same.

Stein continues to explain the main idea of the paper:

> When Θ is infinite-dimensional, that is, in nonparametric problems, the maximum likelihood method often breaks down. Frequently, the maximum likelihood estimate is undefined, [...], and it is not clear that it is good when it exists. However, the existence of a one-dimensional subproblem asymptotically as difficult as the original problem (of a single real-valued function) often persists, at least formally.

Stein's paper works this out for three models, which we shall review in detail. His method is based on clever parameterisations of these models, and a purely algebraic lemma on the inverse of a partitioned matrix, derived in Section 3 of his paper (the "only rigorous" part of the paper).

Stein partitions the parameters of this model in three parts, as (θ, η, γ), where θ is the parameter of interest, η an additional parametric component, knowledge of which is important for estimating θ, and γ a parameter that is not relevant to the problem of estimating θ. In modern language both η and γ are nuisance parameters. Uncertainty about η has a negative influence on our ability of estimating θ, but to the unknown value of γ we can *adapt* without losing quality. Because quality can be read off from the Fisher information matrix (although Stein points out that this may not be relevant "for sample sizes likely to occur in practice"), the classification of the three parts of the parameters can be made precise in terms of properties of this matrix.

The Fisher information matrix for a partitioned parameter $\tau = (\theta, \eta, \gamma)$ can be partitioned as[1]

$$I_\tau = \begin{pmatrix} I_{\theta,\theta} & I_{\theta,\eta} & I_{\theta,\gamma} \\ I_{\eta,\theta} & I_{\eta,\eta} & I_{\eta,\gamma} \\ I_{\gamma,\theta} & I_{\gamma,\eta} & I_{\gamma,\gamma} \end{pmatrix}.$$

The optimal quality of estimation is determined by the inverse of this matrix, as shown in (1). Specifically, for a function $\phi(\theta, \eta, \gamma) = \chi(\theta)$ that depends only on θ, the gradient $\nabla\phi(\theta, \eta, \gamma) = (\nabla\chi(\theta), 0, 0)$ depends only on the first coordinate, and hence (1) depends only on the (θ, θ)-submatrix of the inverse. The point now is that this submatrix is not equal to $(I_{\theta,\theta})^{-1}$, in general. In fact, in general,

$$\left(I_\tau^{-1}\right)_{\theta,\theta} \geq (I_{\theta,\theta})^{-1}, \tag{2}$$

in the sense that the difference of the two matrices is nonnegative definite. The inequality is a property of general nonnegative-definite matrices, but in this case has a statistical interpretation. The matrix on the right would be the relevant inverse if η and γ were known, and hence the full information matrix of the model would be $I_{\theta,\theta}$. The matrix on the left is bigger, because it is harder to estimate θ when η and γ are unknown.

In Stein's examples, γ plays the role of a nuisance parameter that does not make the estimation of θ harder when it is unknown, whereas η may increase the information bound, that is, in his examples the three parameters are chosen so that

$$(I_\tau^{-1})_{\theta,\theta} = \left(\begin{pmatrix} I_{\theta,\theta} & I_{\theta,\eta} \\ I_{\eta,\theta} & I_{\eta,\eta} \end{pmatrix}^{-1}\right)_{\theta,\theta}. \tag{3}$$

[1] In general, each of the matrices on the right depends on the full parameter τ. Here, and when writing score functions, we alleviate, but abuse, notation by not showing this dependence.

The irrelevant parameter γ is taken to be Euclidean, but may index any parametric submodel within an encompassing nonparametric model. It is thus that the paper concerns "efficient *nonparametric* testing and estimation", as promised in its title.

Clearly, the parameter η must be chosen carefully to render the preceding display correct. It should index what in modern language would be called a *least favourable submodel* for estimating θ. In two of Stein's three examples, the parameterisation is natural, albeit that the adaptive nature is not obvious. In the third example, he employs a clever reparameterisation, and the least favourable submodel seems to come somewhat from the air.

Stein's way of proving the correctness of his choices is based on verifying the preceding display (3). In a lemma, he first reduces (3), by a page of elementary calculations, to the identity $I_{\theta,\gamma} = I_{\theta,\eta} I_{\eta,\eta}^{-1} I_{\eta,\gamma}$, another result valid for general nonnegative-definite matrices.

In the next sections, we revisit Stein's three examples. A modern perspective is to present information identities not through matrices, but, indirectly, through a calculus of score functions [1, 59]. For a partitioned parameter $\tau = (\theta, \eta, \gamma)$, the score function can be partitioned as well, as

$$\dot{\ell}_\tau(x) = \begin{pmatrix} \dot{\ell}_\theta(x) \\ \dot{\ell}_\eta(x) \\ \dot{\ell}_\gamma(x) \end{pmatrix} = \begin{pmatrix} \dfrac{\partial}{\partial\theta} \log p_{\theta,\eta,\gamma}(x) \\ \dfrac{\partial}{\partial\eta} \log p_{\theta,\eta,\gamma}(x) \\ \dfrac{\partial}{\partial\gamma} \log p_{\theta,\eta,\gamma}(x) \end{pmatrix}.$$

The information matrix I_τ is the covariance matrix of the vector $\dot{\ell}_\tau(X)$, and partitions in the covariance and cross-covariance matrices of the three component scores (which are vectors of lengths the dimensions of θ, η, γ).

It turns out that the inverse information matrix I_τ^{-1} can be found as a partitioned matrix of covariances as well, but of projected score functions. Each coordinate function of $\dot{\ell}_\tau$ is an element of the Hilbert space $L_2(p_\tau)$ of square-integrable functions. Let $\Pi_{\eta,\gamma}$ be the orthogonal projection[2] onto the closure of the subspace of $L_2(p_\tau)$ spanned by all coordinate functions of $\dot{\ell}_\eta$ and $\dot{\ell}_\gamma$, and let

$$\tilde{\ell}_\theta(x) = \dot{\ell}_\theta(x) - \Pi_{\eta,\gamma}\dot{\ell}_\theta.$$

LEMMA 2.1. *If* $\tilde{I}_{\theta,\theta} = \mathrm{Cov}_\tau(\tilde{\ell}_\theta(X))$, *then* $(I_\tau^{-1})_{\theta,\theta} = (\tilde{I}_{\theta,\theta})^{-1}$.

PROOF. For a given random vector $(U^T, V^T)^T$ with mean zero and covariance matrix J, the decomposition $U = (U - \Lambda V) + \Lambda V$ is orthogonal for the matrix $\Lambda = J_{U,V} J_{V,V}^{-1}$, if J is partitioned in the submatrices $J_{U,U}$, $J_{U,V}$, $J_{V,U}$ and $J_{V,V}$. The covariance matrix of $U - \Lambda V$ is $J_{U,U} - \Lambda J_{V,V} \Lambda^T = J_{U,U} - J_{U,V} J_{V,V}^{-1} J_{V,U}$. By the formula for the inverse of a partitioned matrix, the inverse of the last matrix is $(J^{-1})_{U,U}$. □

The interpretation is that the information loss due to not knowing the nuisance parameter (η, γ) is coded in the loss of the part of the score function $\dot{\ell}_\theta(x)$ that also arises as a score function for (η, γ). The remaining part $\tilde{\ell}_\theta$ of the score function is called the *efficient score*

[2]By definition the *orthogonal projection* $\Pi_{\eta,\gamma}\ell$ of $\ell \in L_2(p_\tau)$ minimizes the map $g \mapsto \int(\ell - g)^2 p_\tau \, d\mu$ over $g \in \overline{\mathrm{lin}}(\dot{\ell}_\eta, \dot{\ell}_\gamma)$. If θ is multi-dimensional, it is here applied coordinatewise to the coordinates of $\dot{\ell}_\theta$. The projection is characterised by the orthogonality relationship $\langle \tilde{\ell}_\theta, g \rangle_\tau = 0$, for every $g \in \overline{\mathrm{lin}}(\dot{\ell}_\eta, \dot{\ell}_\gamma)$. Here, $\langle g_1, g_2 \rangle_\tau = \int g_1(x) g_2(x) p_\tau(x) \, d\mu(x)$, where μ is the dominating measure for the densities p_τ, is the inner product.)

function for θ, and its covariance matrix the *efficient information* for θ. The inverse of this matrix is the lower bound for estimating θ in the presence of the unknown nuisance parameters (η, γ). Since the right-hand side of inequality (2) is the covariance matrix of $\dot{\ell}_\theta(X)$, this inequality is a consequence of the fact that projections decrease variance.

Equation (3) allows an information loss for estimating θ due to the parameter η being unknown, but expresses that there is no loss through not knowing γ. In terms of score functions, it means that the projection of $\dot{\ell}_\theta$ onto $\overline{\text{lin}}(\dot{\ell}_\eta, \dot{\ell}_\gamma)$ is no different than the projection onto $\overline{\text{lin}}(\dot{\ell}_\eta)$, or, equivalently, if Π_η is the orthogonal projection onto the latter space, then $\dot{\ell}_\theta - \Pi_\eta \dot{\ell}_\theta$ is orthogonal to $\overline{\text{lin}}(\dot{\ell}_\gamma)$. The interpretation is that after having corrected for the fact that η is unknown, further corrections for γ being unknown are unnecessary.

An optimal estimation procedure for θ should thus be able to "adapt" to the unknown parameter γ: it should have the same (asymptotic) mean square error as the best procedure that may depend on the true value of γ. It is of historical interest that Stein presents his examples within the preceding framework with three component parameters. In later literature, the term *adaptation* became specialised to the situation of parameters (θ, γ) with two components, with the same requirement that estimation of θ is as difficult with or without knowing γ. In this usage, γ is typically allowed to be "nonparametric", rather than finite-dimensional as in Stein's paper, but this is not essentially more general, as information bounds are always suprema over finite-dimensional submodels. Allowing a third component and considering parameters (θ, η, γ) does make a big difference. For instance, Stein's third example, the errors-in-variables model would not be considered "adaptive" in present terminology. Stein's second example, the two-sample problem, is said to satisfy the condition for adaptive estimation in Proposition 1 on page 101 of [8], but strictly speaking is also not adaptive, in the sense of, for instance, [6].

For a k-dimensional parameter θ, there are k components to the score function $\dot{\ell}_\theta$, and then equally many projections on the score space of other unknown parameters. If it were possible to choose "directions" in the set of other parameters along which the score functions are the k projections of $\dot{\ell}_\theta$, then this might give a submodel indexed by a k-dimensional parameter η such that the efficient score function for θ in this submodel would agree with the efficient score function in the full model. The remaining part of the parameter, possibly infinite-dimensional, could then be marked γ, yielding exactly the structure as considered by Stein. Such *least favourable submodels* are known for many situations. It seems that in this sense, Stein's paper is much more general than "adaptive models".

While least favourable submodels often exist, there are exceptions. The worst type of failure comes from the fact that the projection $\Pi_{\eta,\gamma}$ defining $\tilde{\ell}_\theta$ is onto the *linear span* of the nuisance scores, whence $\Pi_{\eta,\gamma}\dot{\ell}_\theta$ may not be a nuisance score itself. In that situation, there may not exist a submodel in the sense of Stein and his heuristic will be overly optimistic. The linearity arises within the context of the Cramér–Rao bound from the fact that this is based on covariance. The covariance between an estimator and a set of score functions imposed by unbiasedness of the estimator, automatically determines the covariance between the estimator and elements from the linear span of the score functions. This leads to a lower bound on the variance not only for every score (and hence submodel in the sense of Stein), but also for elements in the linear span of the scores. The asymptotic version of this phenomenon was investigated in [58, 62]. If one imposes restrictions on estimators such as unbiasedness or local uniformity ("regularity"), again the linear span of the score vectors drives the lower bound. For the local minimax criterion, this remains true (only) if the set of score vectors is convex.

If the efficient score does not correspond to a least favourable submodel at parameter values chosen by maximum likelihood estimators, then the efficient score may not provide a likelihood equation. To analyse maximum likelihood estimators, [41] introduced "approximate" least favourable models. Further observations on least favourable submodels were made by [51, 70] and [12, 13].

2.1. *Symmetric densities.* The example of estimating the median of symmetrically distributed observations, discussed in Section 4 of Stein's paper, was followed up by many authors, who in various ways completed Stein's program. It became perhaps the most famous example of a semiparametric model, competing, or ex aequo, with the Cox model.

In his understated style, Stein starts his discussion with the remark:

> The problem we discuss in this chapter is not one which often arises in practice. However, it is so simple that we can almost treat it satisfactorily without introducing any really new methods [...].

The score function for θ of the one-dimensional location model $p(x-\theta)$ is given by

$$\dot{\ell}_\theta(x) = -\frac{p'}{p}(x-\theta).$$

If the probability density p is symmetric about zero, that is, $p(x) = p(-x)$ for all $x \in \mathbb{R}$, then this is an odd function around θ. On the other hand, perturbing the shape p, leads to symmetric functions: if $\gamma \mapsto p_\gamma$ is a smooth curve through the symmetric densities, then

$$\frac{\partial}{\partial \gamma} \log p_\gamma(x-\theta) = b\big(|x-\theta|\big),$$

for some function b. By the (anti-)symmetry $\int \dot{\ell}_\theta(x) b(|x-\theta|)\, p(x-\theta)\, dx = 0$, and hence $\dot{\ell}_\theta$ is orthogonal to all nuisance scores. No projection is needed, the efficient score function is equal to the ordinary score, and no additional parameter η need be involved. The remarkable message is that estimating θ when p has a completely unknown shape should not be more difficult than estimating θ when p is known to be a particular density.

This promise was fulfilled in steps, under increasingly mild regularity conditions [2, 50, 54, 56, 65]. Eventually it was proved that there exist estimators $\hat{\theta}_n$, based on a sample of n observations and not on the shape p, such that $\sqrt{n}(\hat{\theta}_n - \theta)$ tends to a normal distribution $N(0, I_p^{-1})$, for any p that is absolutely continuous, where $I_p = \int (p'/p)^2(x)\, p(x)\, dx$ is the Fisher information for location. (If the latter is infinite, then the normal limit is understood to be degenerate. Infinite values of I_p are characteristic of "irregular" or "singular" densities; see [23], Chapters V and VI for a detailed treatment.)

The simplest way to construct such estimators is to estimate the density p using the sample absolute values $|X_1 - \tilde{\theta}_n|, \ldots, |X_n - \tilde{\theta}_n|$ for a preliminary $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$, and use a single iteration of the Newton method to solve the estimated likelihood equation $\sum_{i=1}^n (\hat{p}'/\hat{p})(X_i - \theta) = 0$, starting from the same initial estimator $\tilde{\theta}_n$. Estimators based on rank statistics, with estimated score function, have also been considered in the literature; see, for example, [2] and [21]. The second reference studies and exploits asymptotic equivalence of tangent space projections and projections onto residual ranks via conditioning on order statistics in settings (such as the one and two-sample shift models considered by Stein) in which ranks and order statistics are available. A mathematical proof of existence of efficient estimators under the minimal condition is relatively straightforward (see, e.g., [64], Section 25.8.1), but practical implementation is as tricky as in the 1950s, due to the necessity of tuning the density estimator.

In his paper, Stein considers the problems of testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, rather than the estimation problem, which similarly is adaptive to the shape of the density p. Without giving a detailed proof, he seems to give a complete solution of this problem. He proposes test statistics of the form

$$\sum_{i=1}^n Z_i 1_{Z_i^2 \le \varepsilon_n \sum_{i=1}^n Z_i^2} \operatorname{sign}(X_i),$$

where $\varepsilon_n \to 0$ and the Z_i are functions of the absolute values $|X_1|, \ldots, |X_n|$, given by, for $\mathbb{S}_n$ the empirical distribution function of the absolute values and $a > 0$ some constant,

$$Z_i = \frac{2n[\mathbb{S}_n(|X_i| + a/\sqrt{n}) - 2\mathbb{S}_n(|X_i|) + \mathbb{S}_n(|X_i| - a/\sqrt{n})]}{a\sqrt{n}[\mathbb{S}_n(|X_i| + a/\sqrt{n}) - \mathbb{S}_n(|X_i| - a/\sqrt{n})]}.$$

Under the null hypothesis, $\mathbb{S}_n(t) \to \mathrm{P}(|X| \le t) = \int_0^t 2p(x)\,dx$, whence the scaled second-order and first-order differences in the numerator and denominator of the quotient Z_i estimate $4a^2p'(|X_i|)$ and $4a^2p(|X_i|)$. It follows that $Z_i \operatorname{sign}(X_i) \approx (p'/p)(X_i)$ and the test statistic should behave similarly to $\sum_{i=1}^n \dot{\ell}_0(X_i)$, the optimal test statistic when p is known. Given that, conditionally on the the absolute values, the signs are independent variables with values in $\{-1, 1\}$, a central limit theorem can be invoked to make this precise.

The theory of asymptotically optimal tests was being developed in the same era, for example, in the paper [29] presented at the same Berkeley Symposium as Stein's paper, or [30] and [18].

2.2. *Two sample location-scale.* Given two independent samples from the same location-scale family, defined by a single density shape p, but with different location-scale parameters for the two samples, the location and scale are confounded with the parameter p, but the difference of the locations of the two samples and the quotient of the scales are identifiable. In Section 5 of his paper, Stein shows essentially that the least favourable submodels for these relative parameters are within the submodels indexed by the two pairs of location-scale parameters, and do not involve the shape density. To facilitate his calculations, he reparameterises the two location parameters in terms of their half difference and average, and the two scale parameters by the square roots of their quotient and product.

For our calculus of scores, it is slightly easier to use a different parameterisation. For simplicity, we also assume that the sample sizes are equal and pair the two samples to a single sample from the distribution of a pair (X, Y). This latter pair has density

$$(x, y) \mapsto p(x)p\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

The parameter of interest is $\theta = (\mu, \sigma)$, while the nuisance parameters η and γ jointly correspond to p.

The score function for θ takes the form

$$\dot{\ell}_\theta(x, y) = \begin{pmatrix} -\dfrac{p'}{p}\left(\dfrac{y-\mu}{\sigma}\right)\dfrac{1}{\sigma} \\ -\left[1 + \dfrac{y-\mu}{\sigma}\dfrac{p'}{p}\left(\dfrac{y-\mu}{\sigma}\right)\right]\dfrac{1}{\sigma} \end{pmatrix}.$$

The score function corresponding to the perturbation $p_t(x) = p(x)(1 + tb(x))$, for $t \approx 0$ and a given (bounded) function b with $\int b(x)p(x)\,dx = 0$, is

$$\frac{\partial}{\partial t}_{|t=0} \log\left[p_t(x)p_t\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma}\right] = b(x) + b\left(\frac{y-\mu}{\sigma}\right).$$

To find the efficient score function for θ, we must project $\dot{\ell}_\theta$ onto the set of all functions of the latter type, within the L_2-space corresponding to the distribution of (X, Y). Since $(Y - \mu)/\sigma$ has density p, this is equivalent to projecting

$$\begin{pmatrix} -\dfrac{p'}{p}(y)\dfrac{1}{\sigma} \\ -\left[1 + y\dfrac{p'}{p}(y)\right]\dfrac{1}{\sigma} \end{pmatrix}.$$

Onto the set of functions $b(x) + b(y)$ in the space $L_2(p \times p)$, for freely varying b. Now orthogonality of $b(x) + b(y)$ to $\ell(x) + \ell(y)$ in $L_2(p \times p)$ is equivalent to orthogonality of b and ℓ in $L_2(p)$. This readily shows that the projection of a function $\ell(y)$ onto the set of functions $b(x) + b(y)$ is the function $(\ell(x) + \ell(y))/2$. Consequently, the efficient score function for θ is given by

$$\tilde{\ell}_\theta(x) = \begin{pmatrix} \frac{1}{2}\left[\frac{p'}{p}(x) - \frac{p'}{p}\left(\frac{y-\mu}{\sigma}\right)\right]\frac{1}{\sigma} \\ \frac{1}{2}\left[1 + x\frac{p'}{p}(x) - 1 - \frac{y-\mu}{\sigma}\frac{p'}{p}\left(\frac{y-\mu}{\sigma}\right)\right]\frac{1}{\sigma} \end{pmatrix}.$$

We recognise the projections as scores arising from a location-scale model $p_{\eta,\gamma}(x) = p_\gamma((x-\eta_1)/\eta_2)/\eta_2$. Thus the irrelevant nuisance parameter γ corresponds to a shape p_γ with a standardised location and scale; the relevant nuisance parameter $\eta = (\eta_1, \eta_2)$ parametrizes the location and shape of X within the location-scale family; the parameter of interest shifts the location and scale of Y relative to the location scale of X.

Following [44], Sections 18.3 and 18.5, Theorem 1 on page 99 of [8] extend this type of "adaptive" structure to any two-sample group model.

Efficient estimators fulfilling Stein's program can be constructed using a preliminary estimator of the unknown density p, along similar lines as for the symmetric location model [2, 43, 54, 65, 67, 69].

2.3. *Errors-in-variables.* In Section 6, Stein considers observing a sample from the distribution of a two-dimensional vector $X = Y + Z$, for independent vectors Y and Z, where Y has an unspecified distribution that is concentrated on an unspecified line $L = \{a + \lambda b : \lambda \in \mathbb{R}\} \subset \mathbb{R}^2$, and Z is bivariate normally distributed with mean zero and unknown covariance matrix Σ. The parameter of interest is the slope of the line L, or, equivalently, a normalised version of the vector b.

This is a version of the errors-in-variables model: the first coordinate X_1 is the sum of an "independent variable" $Y_1 = a_1 + \Lambda b_1$ and a Gaussian error Z_1, and, provided $b_1 \neq 0$, the second coordinate X_2 is equal to the sum of the linear function $a_2 + b_2(Y_1 - a_1)/b_1$ of Y_1 and the Gaussian error Z_2. The slope parameter b_2/b_1 had been proven to be identifiable if the distribution of Y is not normal (and not degenerate) in [46].

Stein's remarkable insight in the information structure of the model is based on splitting the error vector Z into a component on the line L and an independent remainder. Fix any nonzero vector $c \in \mathbb{R}^2$ such that the vectors b and c form an orthogonal basis relative to the inner product generated by Σ^{-1}, that is, $b^T\Sigma^{-1}c = 0$. Then $Z = Ub + Vc$, for the univariate variables $U = b^T\Sigma^{-1}Z/b^T\Sigma^{-1}b$ and $V = c^T\Sigma^{-1}Z/c^T\Sigma^{-1}c$, which can be seen to be Gaussian and independent. If $Y = a + \Lambda b$ for a univariate variable Λ, then this leads to $X = a + (\Lambda + U)b + Vc$. Choosing b and c of unit length and decomposing the intercept as $a = \nu b + \mu c$, we can further rewrite the equation as

$$X = R\begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix} + S\begin{pmatrix}\sin\phi \\ \cos\phi\end{pmatrix} = \begin{pmatrix}\cos\theta & \sin\phi \\ \sin\theta & \cos\phi\end{pmatrix}\begin{pmatrix}R \\ S\end{pmatrix}, \tag{4}$$

where θ, ϕ are unknown parameters, and R and S are independent univariate random variables, with the distribution of R unspecified but not normal, and S normal with unknown mean μ and unknown variance σ^2. The slope of the line L is given by the parameter θ.

Stein presents a (7×7) information matrix for the parameters θ, ϕ, μ, σ, the location and scale parameters of R and an additional parameter γ for the shape of the density of R. At first reading, he seems to suggest that the information for the pair (θ, ϕ) does not decrease if any of the other parameters are unknown. At second reading, this situation, which Stein calls

"curious" in the last sentence of his paper, is due to his special choice of true values of the parameters: he fixes $\theta = \phi = 0$, and sets the means and variances of R and S to 0 and 1.[3]

For general parameter values, it is still true that the estimation of (θ, ϕ) does not become harder from not knowing the shape of the density of R. In other words, the least favourable submodel for (θ, ϕ) (together corresponding to θ in the partitioned parameter (θ, η, γ) in our general Introduction) is contained in the four-dimensional model (corresponding to η) generated by (μ, σ) and the location and scale parameters of R, and the problem is adaptive to the shape of the density of R (corresponding to γ). As this seems not obvious, we present details on the proof of this fact.

The defining equation (4) can be inverted to show that, with the dependence of $R = R_{\theta,\phi}$ and $S = S_{\theta,\phi}$ on the parameters made explicit,

$$R_{\theta,\phi} = \frac{X_1 \cos\phi - X_2 \sin\phi}{\cos(\theta + \phi)}, \qquad S_{\theta,\phi} = \frac{-X_1 \sin\theta + X_2 \cos\theta}{\cos(\theta + \phi)}.$$

The Jacobian of the linear transformation from $(R, S)^T$ to X is $\cos(\theta + \phi)$. (If we choose c to be the counterclockwise rotation by $\pi/2$ of $\Sigma^{-1}b$, then $\cos(\theta + \phi) = b^T \Sigma^{-1} b / (\|b\| \|\Sigma^{-1} b\|) > 0$.) Thus we find that a probability density of X is given by

$$p_X(x) = p_R(r_{\theta,\phi}) \varphi\left(\frac{s_{\theta,\phi} - \mu}{\sigma}\right) \frac{1}{\sigma} \frac{1}{\cos(\theta + \phi)}. \tag{5}$$

With a little algebra (especially for the fourth and fifth equations), we see that

$$2\frac{\partial}{\partial\theta} \frac{1}{\cos(\theta + \phi)} = \frac{\tan(\theta + \phi)}{\cos(\theta + \phi)}, \qquad \frac{\partial}{\partial\theta} \log\cos(\theta + \phi) = -\tan(\theta + \phi),$$

$$\frac{\partial}{\partial\theta} R_{\theta,\phi} = R_{\theta,\phi} \tan(\theta + \phi), \qquad \frac{\partial}{\partial\phi} R_{\theta,\phi} = \frac{-S_{\theta,\phi}}{\cos(\theta + \phi)},$$

$$\frac{\partial}{\partial\theta} S_{\theta,\phi} = \frac{-R_{\theta,\phi}}{\cos(\theta + \phi)}, \qquad \frac{\partial}{\partial\phi} S_{\theta,\phi} = S_{\theta,\phi} \tan(\theta + \phi).$$

Using that $\varphi'/\varphi(s) = -s$, we find, again after some algebra, that the score functions for θ, ϕ, μ, σ are given by (for notational convenience we omit the subscripts θ, ϕ from R and S),

$$\dot{\ell}_\theta(X) = \left[1 + R\frac{p_R'}{p_R}(R)\right] \tan(\theta + \phi) + \frac{S - \mu}{\sigma^2} \frac{R}{\cos(\theta + \phi)},$$

$$\dot{\ell}_\phi(X) = \frac{p_R'}{p_R}(R) \frac{-S}{\cos(\theta + \phi)} - \left[\frac{(S - \mu)S}{\sigma^2} - 1\right] \tan(\theta + \phi),$$

$$\dot{\ell}_\mu(X) = \frac{S - \mu}{\sigma^2},$$

$$\dot{\ell}_\sigma(X) = \frac{1}{\sigma}\left[\frac{(S - \mu)^2}{\sigma^2} - 1\right].$$

The remaining scores result from varying p_R along submodels. For p_R unspecified, the possible scores for p_R would include any square-integrable function $b(R)$ of mean zero, and the orthogonal projection onto the score space for R would be the conditional expectation relative to R. Although within the original model $R = \nu + \Lambda + U$ always contains a nontrivial Gaussian component U, and hence cannot have any possible distribution, we adopt as working hypothesis that p_R is unspecified. We shall see that it actually only matters that the set of p_R is closed under location and scale changes.

[3]We understand Stein's explicitly named location and scale parameters η and λ for R to be the mean and standard deviation.

The efficient scores for (θ, ϕ) relative to the other parameters viewed as nuisance parameters are by definition the scores $\dot{\ell}_\theta$ and $\dot{\ell}_\phi$ minus their projections onto the sum of the score space for p_R and $\operatorname{lin}(\dot{\ell}_\mu, \dot{\ell}_\sigma)$. By the product structure in the likelihood, the latter spaces are orthogonal, and hence the projection onto the full nuisance score space is the sum of the conditional expectation relative to R and the projection onto $\operatorname{lin}(\dot{\ell}_\mu, \dot{\ell}_\sigma)$.

The first component of $\dot{\ell}_\theta$ is completely explained by a score for p_R, and drops out when forming the efficient score $\tilde{\ell}_\theta$. As this particular function is the score at $\tau = 1$ for the submodel $\tau \mapsto p_R(r/\tau)/\tau$, the projection would be the same if p_R were known just up to its scale. The second component of $\dot{\ell}_\theta$ is orthogonal to the score space for p_R and its projection onto $\operatorname{lin}(\dot{\ell}_\mu, \dot{\ell}_\sigma)$ changes R into its mean value. It follows that

$$\tilde{\ell}_\theta(X) = \frac{S-\mu}{\sigma^2}\frac{R - \mathrm{E}R}{\cos(\theta+\phi)}. \tag{6}$$

The projection of the first component of $\dot{\ell}_\phi$ onto the score space for R changes the variable S to its mean $\mathrm{E}S = \mu$; the projection is a multiple of the location score of R. The second component of $\dot{\ell}_\phi$ is contained in $\operatorname{lin}(\dot{\ell}_\mu, \dot{\ell}_\sigma)$. It follows that

$$\tilde{\ell}_\phi(X) = \frac{p_R'}{p_R}(R)\frac{\mu - S}{\cos(\theta+\phi)}. \tag{7}$$

Thus we find that the projections of $\dot{\ell}_\theta$ and $\dot{\ell}_\phi$ are within the sum of the score spaces of the location-scale families of R and S.

Incidentally, the nonidentifiability of the θ in the case that R is normally distributed is visible in the efficient information: if R were normal, then the location score p_R'/p_R would be a linear function of R, and hence $\tilde{\ell}_\theta$ and $\tilde{\ell}_\phi$ would be proportional and the efficient Fisher information for θ in the absence of knowledge of ϕ equal to zero.

Efficient estimators for the slope b_2/b_1 (equivalently θ) were first constructed in [9] and [63]. These authors used a different principle, but also factorised the likelihood, based on the sufficiency of the "statistic" $b^T\Sigma^{-1}(X-a)$, which depends on the parameters (a, b, Σ), for the remaining parameter of the model: the distribution of Λ (as readily follows from the sufficiency of this statistic for the univariate parameter λ in the model $X = a + b\lambda + Z$). It can then be argued from completeness of the Gaussian exponential family (at least in the case of a true continuous distribution of Λ) that the projection onto the nonparametric nuisance scores is the conditional expectation relative to $b^T\Sigma^{-1}(X-a)$, and one arrives at similar formulas (6)–(7).

Models with sufficient statistics for the nuisance parameter, but then without dependence on the parameter of interest, had earlier been considered in [25, 44]. Actually, if dependence is allowed, all three models considered by Stein are characterised by a sufficient statistic, and a unified theory of adaptive estimation is possible [63].

The connection to Stein's decomposition is the identity

$$b^T\Sigma^{-1}(X-a) = b^T\Sigma^{-1}b(R-\nu),$$

by the relation $X - a = (R-\nu)b + Vc$ and the orthogonality relation $b^T\Sigma^{-1}c = 0$. Thus the sufficient statistic is a scaled version of $R-\nu$, and the likelihood factorisation (5) is equivalent to the one given by the factorisation theorem for sufficient statistics.

An asymptotically efficient estimator for $(\hat{\theta}, \hat{\phi})$ can be obtained as the (approximate) solution to the efficient score equations (cf. [64], Section 25.8 for this method in general)

$$\sum_{i=1}^n \frac{S_{\theta,\phi,i} - \mu_{\theta,\phi,\Sigma}}{\sigma^2}\frac{R_{\theta,\phi,i} - \hat{\mathrm{E}}R_{\theta,\phi}}{\cos(\theta+\phi)} = 0,$$

$$\sum_{i=1}^n \frac{\hat{p}_R'}{\hat{p}_R}(R_{\theta,\phi,i})\frac{\mu_{\theta,\phi,\Sigma} - S_{\theta,\phi,i}}{\cos(\theta+\phi)} = 0.$$

Here, $\hat{\mathrm{E}}$, $\hat{p}'_R$ and $\hat{p}_R$ refer to estimates that involve the infinite-dimensional parameter. As p_R is the density of the density of $b^T\Sigma^{-1}(X-a)/b^T\Sigma^{-1}b+\nu$, given a, b, Σ it can be estimated nonparametrically, similarly as in the symmetric location and two-sample models. The papers [9] and [63] applied kernel density estimators for this purpose. Because $R=\nu+\Lambda+U$, the distribution of R is the convolution of a Gaussian distribution with the distribution of Λ. By estimating the mixing distribution Λ instead of the density p_R, for instance by the nonparametric maximum likelihood estimator as in [24], the inherent instability of density estimators can be avoided [61, 64].

An alternative is the full maximum likelihood estimator, which maximizes the likelihood

$$(a,b,\Sigma,H)\mapsto\prod_{i=1}^{n}\int\frac{1}{\sqrt{\det\Sigma}}e^{-(X_i-a-\lambda b)^T\Sigma^{-1}(X_i-a-\lambda b)/2}\,dH(\lambda),$$

where the scales of a and b and possibly the shape of Σ are restricted, and H ranges over all probability distributions on $\mathbb{R}$. In [24], this estimator was shown to be consistent, provided identifiability, and in [60], the slope component was shown to be efficient, in a version of the model that identifies the parameters through restrictions on Σ rather than nonnormality of Λ, as assumed by Stein. It appears that Stein's insight in the information structure may be exploited to derive also the asymptotic normality in his version of the model.

Stein discusses the version of the errors-in-variables model in which the independent variables Λ are random and i.i.d.. Another version, considered in the same decade by [42, 45], lets the values $\lambda_1,\ldots,\lambda_n$ attached to the observations $X_1,\ldots,X_n$ be arbitrary constants. For the errror-in-variables model, it was shown in [40] that the maximum likelihood estimator of the slope, as in the preceding display, is asymptotically normal also in this version of the model. Estimators based on kernel density estimators were considered in [62]. In the model with an increasing number of parameters, Stein's heuristic of least favourable submodels does not apply and the notion of asymptotic optimality is unclear, but within a class of "perturbation symmetric" estimators, the heuristic can be applied to show that the limiting distribution is optimal [55].

3. Follow-up. Stein began his paper [52] squarely in the context of the theory of finite-dimensional statistical models as known at that time: under suitable regularity conditions the MLE is asymptotically normal and efficient in the sense of the Cramér–Rao bound. (See Stigler [53] for a fascinating history of maximum likelihood estimation theory up to the mid 1950s.) In the last paragraph of his Section 2, he neatly pivoted to infinite-dimensional models and laid out a program of research, which took 40 or 50 years to fully explore and develop:

> The general theory of the infinite-dimensional case would seem to be technically quite involved. However, a procedure which may work is the following. We can often integrate the field of most difficult directions, thus expressing the parameter space Θ as a union of one-dimensional subproblems, each of which is asymptotically a most difficult one-dimensional problem through each of its points. We then make a crude estimate of the parameter point θ, using this estimate to select one of the one-dimensional subproblems. We then proceed as if the true parameter point lay on this curve, using, for example, the maximum likelihood method to complete the estimation of $\varphi(\theta)$. To prove that this works under fairly general conditions seems to be quite difficult.

During the period 1956–1982, efficient estimators of θ in Stein's examples 1 and 2 (the symmetric location model and the two-sample location-scale model) were constructed using a wide variety of methods, as we noted in Sections 2.1 and 2.2. These methods were also extended to a number of other problems featuring adaptivity in the "narrow sense" discussed in Section 1, culminating in Peter Bickel's Wald lectures [6].

New lower bound theory for finite-dimensional problems was still evolving, with the asymptotic minimax and convolution lower bounds of Hájek [19, 20] and Le Cam [32] (reviewed in [66]), and Le Cam's [31] detailed scrutiny of the conditions for asymptotic normality of maximum likelihood estimators. By the middle of the 1970s, these new tools were developing in the direction of models involving infinite-dimensional parameters: Beran [32] obtained a convolution type lower bound for the problem of estimating a distribution function and Beran [3–5] used Hájek's convolution formulation to establish lower bounds in several related problems. Millar [37–39] developed infinite-dimensional versions of both the asymptotic minimax and convolution theorems.

Parallel developments in Russia by Ibragimov and Has'minskii [22, 23] and their students Koshevnik and Levit [26, 34–36] contributed to these developments from the perspective of differentiable functionals defined on nonparametric (or infinite-dimensional) families of probability distributions. Pfanzagl [44] summarised and extended these developments.

New models in survival analysis began developing rapidly following Cox's [11] introduction of the proportional hazards model. Efron [16] established efficiency of Cox's partial likelihood estimators using a clever (brute force) construction of a least favourable submodel. Efron's paper provided considerable motivation for construction of a lower bound theory for semiparametric models which would handle Stein's examples, and Cox's model, and other semiparametric models of interest. This lead to [1], where the type of models considered were called "parametric-nonparametric". (The term "semiparametric" apparently did not come into common use until slightly later.) And this lead in turn to [7, 8], and a calculus of information bounds for differentiable parameters in [59], which unified the nonparametric functional approach of the Russian school and the "parametric-nonparametric" approach of the Seattle-Berkeley groups.

The differentiability referred to in the last paragraph is Hellinger (pathwise) differentiability along paths in the class of densities describing the model. This leads to local approximations by Gaussian limit experiments in the sense of Le Cam [32]. Unfortunately, this approach to lower bounds breaks down when the classes of distributions involved are too large for existence of (initial) estimators that can localize the problem, or when the parameter of interest is not smooth enough relative to the Hellinger metrics. See Bickel and Ritov (1987) [10], Ritov and Bickel (1990) [71], Donoho and Liu (1991) [14], Donoho and Low (1992) [15], Groeneboom and Wellner (1992) [17], Laurent and Massart (2000) [28] and Tchetgen, Robins and van der Vaart (2009, 2017) [47, 48] for more on this and, for example, Bin Yu (1997) [71] for different methods (Assouad, Fano and Le Cam), for establishing lower bounds on rates of convergence. In some cases, the latter is still possible through considering one-dimensional submodels (and sometimes non-Gaussian limit experiments), whereas in other cases submodels of increasing dimension must be used to bound a minimax risk, even for a one-dimensional parameter of interest. Thus Stein's heuristic does not always work.

For further developments in the general area of semiparametric models since Stein (1956), see [27, 49, 57] and [68]. Stein's insight in the information structure has inspired considerable progress in the construction and study of various types of (partially) efficient estimators and tests.

Funding. The research leading to these results is partly financed by the NWO Spinoza prize awarded to A. W. van der Vaart by the Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452. MR0696057 https://doi.org/10.1214/aos/1176346151

[2] BERAN, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.* **2** 63–74. MR0345295
[3] BERAN, R. (1977). Robust location estimates. *Ann. Statist.* **5** 431–444. MR0448699
[4] BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463. MR0448700
[5] BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313. MR0518885
[6] BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. MR0663424
[7] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941
[8] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559
[9] BICKEL, P. J. and RITOV, Y. (1987). Efficient estimation in the errors in variables model. *Ann. Statist.* **15** 513–540. MR0888423 https://doi.org/10.1214/aos/1176350358
[10] BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550
[11] COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758
[12] COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39. With a discussion. MR0893334
[13] COX, D. R. and REID, N. (1989). On the stability of maximum-likelihood estimators of orthogonal parameters. *Canad. J. Statist.* **17** 229–233. MR1033105 https://doi.org/10.2307/3314851
[14] DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. II, III. *Ann. Statist.* **19** 633–667, 668–701. MR1105839 https://doi.org/10.1214/aos/1176348114
[15] DONOHO, D. L. and LOW, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970. MR1165601 https://doi.org/10.1214/aos/1176348665
[16] EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565. MR0451514
[17] GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation. DMV Seminar* **19**. Birkhäuser, Basel. MR1180321 https://doi.org/10.1007/978-3-0348-8621-5
[18] HÁJEK, J. (1962). Asymptotically most powerful rank-order tests. *Ann. Math. Stat.* **33** 1124–1147. MR0143304 https://doi.org/10.1214/aoms/1177704476
[19] HÁJEK, J. (1969/70). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330. MR0283911 https://doi.org/10.1007/BF00533669
[20] HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I: Theory of Statistics* 175–194. MR0400513
[21] HALLIN, M. and WERKER, B. J. M. (2003). Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli* **9** 137–165. MR1963675 https://doi.org/10.3150/bj/1068129013
[22] HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1983). On asymptotic efficiency in the presence of an infinite-dimensional nuisance parameter. In *Probability Theory and Mathematical Statistics* (*Tbilisi*, 1982). *Lecture Notes in Math.* **1021** 195–229. Springer, Berlin. MR0735986 https://doi.org/10.1007/BFb0072916
[23] IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation. Applications of Mathematics* **16**. Springer, New York-Berlin. Asymptotic theory, Translated from the Russian by Samuel Kotz. MR0620321
[24] KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464 https://doi.org/10.1214/aoms/1177728066
[25] KLAASSEN, C. A. J. and VAN ZWET, W. R. (1985). On estimating a parameter and its score function. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 827–839. Wadsworth, Belmont, CA. MR0822068
[26] KOŠEVNIK, J. A. and LEVIT, B. J. (1976). On a nonparametric analogue of the information matrix. *Teor. Veroyatn. Primen.* **21** 759–774. MR0428578
[27] KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference. Springer Series in Statistics*. Springer, New York. MR2724368 https://doi.org/10.1007/978-0-387-74978-5
[28] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785 https://doi.org/10.1214/aos/1015957395

[29] LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 129–156. Univ. California Press, Berkeley and Los Angeles. MR0084918
[30] LE CAM, L. (1960). Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *Univ. Calif. Publ. Stat.* **3** 37–98. MR0126903
[31] LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Stat.* **41** 802–828. MR0267676 https://doi.org/10.1214/aoms/1177696960
[32] LE CAM, L. (1972). Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 245–261. MR0415819
[33] LECAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Stat.* **1** 277–329. MR0054913
[34] LEVIT, B. J. (1975). Conditional estimation of linear functionals. *Problemy Peredachi Informatsii* **11** 39–54. MR0494664
[35] LEVIT, B. J. (1978). Infinite-dimensional informational bounds. *Teor. Veroyatn. Primen.* **23** 388–394. MR0488446
[36] LEVIT, B. Y. Asymptotically efficient estimation of nonlinear functionals. MR0533450
[37] MILLAR, P. W. (1979). Asymptotic minimax theorems for the sample distribution function. *Z. Wahrsch. Verw. Gebiete* **48** 233–252. MR0537670 https://doi.org/10.1007/BF00537522
[38] MILLAR, P. W. (1983). The minimax principle in asymptotic statistical theory. In *Eleventh Saint Flour Probability Summer School—*1981 (*Saint Flour*, 1981). *Lecture Notes in Math.* **976** 75–265. Springer, Berlin. MR0722983 https://doi.org/10.1007/BFb0067986
[39] MILLAR, P. W. (1985). Nonparametric applications of an infinite-dimensional convolution theorem. *Z. Wahrsch. Verw. Gebiete* **68** 545–556. MR0772198 https://doi.org/10.1007/BF00535344
[40] MURPHY, S. A. and VAN DER VAART, A. W. (1996). Likelihood inference in the errors-in-variables model. *J. Multivariate Anal.* **59** 81–108. MR1424904 https://doi.org/10.1006/jmva.1996.0055
[41] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. With comments and a rejoinder by the authors. MR1803168 https://doi.org/10.2307/2669386
[42] NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113 https://doi.org/10.2307/1914288
[43] PARK, B. U. (1990). Efficient estimation in the two-sample semiparametric location-scale models. *Probab. Theory Related Fields* **86** 21–39. MR1061946 https://doi.org/10.1007/BF01207511
[44] PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. *Lecture Notes in Statistics* **13**. Springer, New York-Berlin. With the assistance of W. Wefelmeyer. MR0675954
[45] PFANZAGL, J. (1993). Incidental versus random nuisance parameters. *Ann. Statist.* **21** 1663–1691. MR1245763 https://doi.org/10.1214/aos/1176349392
[46] REIERSØL, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica* **18** 375–389. MR0038054 https://doi.org/10.2307/1907835
[47] ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479
[48] ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515
[49] RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720 https://doi.org/10.1017/CBO9780511755453
[50] SACKS, J. (1975). An asymptotically efficient sequence of estimators of a location parameter. *Ann. Statist.* **3** 285–298. MR0359174
[51] SLUD, E. V. and VONTA, F. (2005). Efficient semiparametric estimators via modified profile likelihood. *J. Statist. Plann. Inference* **129** 339–367. MR2126854 https://doi.org/10.1016/j.jspi.2004.06.057
[52] STEIN, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 187–195. Univ. California Press, Berkeley and Los Angeles. MR0084921
[53] STIGLER, S. M. (2007). The epic story of maximum likelihood. *Statist. Sci.* **22** 598–620. MR2410255 https://doi.org/10.1214/07-STS249
[54] STONE, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284. MR0362669
[55] STRASSER, H. (1998). Perturbation invariant estimates and incidental nuisance parameters. *Math. Methods Statist.* **7** 1–26. MR1626568

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

[57] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

[58] VAN DER VAART, A. (1989). On the asymptotic information bound. *Ann. Statist.* **17** 1487–1500. MR1026295 https://doi.org/10.1214/aos/1176347377

[59] VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204. MR1091845 https://doi.org/10.1214/aos/1176347976

[60] VAN DER VAART, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.* **24** 862–878. MR1394993 https://doi.org/10.1214/aos/1032894470

[61] VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1999). *Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. MR1915446

[62] VAN DER VAART, A. W. (1988). *Statistical Estimation in Large Parameter Spaces*. *CWI Tract* **44**. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam. MR0927725

[63] VAN DER VAART, A. W. (1988). Estimating a real parameter in a class of semiparametric models. *Ann. Statist.* **16** 1450–1474. MR0964933 https://doi.org/10.1214/aos/1176351048

[64] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[65] VAN EEDEN, C. (1970). Efficiency-robust estimation of location. *Ann. Math. Stat.* **41** 172–181. MR0263194 https://doi.org/10.1214/aoms/1177697197

[66] VAN DER VAART, A. (1991). An asymptotic representation theorem. *Int. Stat. Rev.* **59** 97–121.

[67] WEISS, L. and WOLFOWITZ, J. (1970). Asymptotically efficient non-parametric estimators of location and scale parameters. *Z. Wahrsch. Verw. Gebiete* **16** 134–150. MR0323023 https://doi.org/10.1007/BF00535694

[68] WELLNER, J. A., KLAASSEN, C. A. J. and RITOV, Y. (2006). Semiparametric models: A review of progress since BKRW (1993). In *Frontiers in Statistics* 25–44. Imp. Coll. Press, London. MR2325995 https://doi.org/10.1142/9781860948886_0002

[69] WOLFOWITZ, J. (1974). Asymptotically efficient non-parametric estimators of location and scale parameters. II. *Z. Wahrsch. Verw. Gebiete* **30** 117–128. MR0403054 https://doi.org/10.1007/BF00532264

[70] WONG, W. H. and SEVERINI, T. A. (1991). On maximum likelihood estimation in infinite-dimensional parameter spaces. *Ann. Statist.* **19** 603–632. MR1105838 https://doi.org/10.1214/aos/1176348113

[71] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963