EFFICIENT NONPARAMETRIC TESTING AND ESTIMATION

CHARLES STEIN
STANFORD UNIVERSITY

1. Introduction

It is customary to treat nonparametric statistical theory as a subject completely different from parametric theory. In this paper, I try to study one of the more obvious connections between the two subjects. Clearly a nonparametric problem is at least as difficult as any of the parametric problems obtained by assuming we have enough knowledge of the unknown state of nature to restrict it to a finite-dimensional set. For a problem in which one wants to estimate a single real-valued function of the unknown state of nature it frequently happens that (in a sense made somewhat more precise in section 2 and, for special cases, in later sections) there is, through each state of nature, a one-dimensional problem which is, for large samples, at least as difficult (to a first approximation) as any other finite-dimensional problem at that point. If a procedure does essentially as well, for large samples, as one could do for each such one-dimensional problem, one is justified in considering the procedure efficient for large samples. If there is no such procedure, one may be forced to adopt a less severe definition of efficiency, as suggested by Wolfowitz [1].

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.

2. A review of the finite-dimensional case

Let Θ be an open subset of a finite-dimensional Euclidean space. For each $\theta \in \Theta$, let p_θ be a probability density with respect to a σ-finite measure μ on a σ-algebra $\mathcal{B}$ of subsets of a space $\mathcal{X}$. Subject to certain differentiability conditions and other regularity conditions (see for example [2]), the maximum likelihood estimate $\hat{\theta}$ of θ, based on a large sample $X_1, \cdots, X_n$ independently distributed according to p_θ for some unknown $\theta \in \Theta$, has certain desirable properties.

We define Fisher's information matrix $I(\theta)$ at θ by

$$I_{ij}(\theta) = E_\theta \frac{\partial \log p_\theta(X)}{\partial \theta_i} \frac{\partial \log p_\theta(X)}{\partial \theta_j}. \tag{1}$$

If φ is a continuously differentiable real-valued function on Θ, then the asymptotic mean-squared error of $\varphi(\hat{\theta})$ as an estimate of $\varphi(\theta)$ is

$$\frac{1}{n}(\nabla\varphi)'I^{-1}(\nabla\varphi) \tag{2}$$

Research carried out partly under sponsorship of the Office of Naval Research.

where $\nabla\varphi$, the gradient of φ at θ is the column vector whose ith coordinate is $\partial\varphi/\partial\theta_i$. In a sense this is asymptotically the best possible estimate of $\varphi(\theta)$; roughly speaking, the mean-squared error cannot be improved appreciably throughout an open subset of Θ.

Of course, the expression (2) is independent of the (curvilinear) coordinate system used. More precisely if we introduce new coordinates related to the old in a 1-1 manner, continuously differentiable in both directions, the value of (2) at a given point θ of the set Θ considered geometrically is unchanged. The vector $v = I^{-1}(\nabla\varphi)$ determines a direction through θ which is also independent of the coordinate system used. This determines a one-dimensional subproblem which at θ is asymptotically as difficult as the original multidimensional problem. To see this, consider a problem in which the parameter space is an open subset of the real line containing 0 and the density at the point τ is $\pi_\tau(x) = p_{\theta+\tau v}(x)$, where v is evaluated at θ. The information matrix for this problem at $\tau = 0$ reduces to the real number

$$(3)\quad E_\theta\left[\frac{d \log p_{\theta+\tau v}(X)}{d\tau}\right]^2 = E_\theta\left[\sum v_i \frac{\partial \log p_\theta(X)}{\partial\theta_i}\right]^2$$
$$= v'Iv = (\nabla\varphi)'I^{-1}II^{-1}(\nabla\varphi) = (\nabla\varphi)'I^{-1}(\nabla\varphi).$$

Also the gradient of $\varphi(\theta + \tau v)$ as a function of τ reduces to the real number

$$(4)\quad \frac{d\varphi(\theta+\tau v)}{d\tau} = \sum v_i \frac{\partial\varphi(\theta)}{\partial\theta_i} = v'(\nabla\varphi)$$
$$= (\nabla\varphi)'I^{-1}(\nabla\varphi).$$

Evaluating (2) for this one-dimensional problem we see that it has the same value as for the original problem.

When Θ is infinite-dimensional, that is, in nonparametric problems, the maximum likelihood method often breaks down. Frequently the maximum likelihood estimate is undefined (as it is for the problems of sections 4 to 6), 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 estimating a single real-valued function) often persists, at least formally. In the remainder of this paper we examine a small number of special cases, of which only one (that of section 4) is treated with any sort of completeness.

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.

3. An algebraic lemma

The somewhat trivial lemma obtained in this section is often useful in proving that an estimation problem is not made more difficult asymptotically by introducing additional unknown parameters in a certain way. After stating and proving the lemma we indicate how it is used. Applications occur in sections 5 and 6. In section 4, a trivial application could have been made, but it seemed pointless because a more complete solution can be given.

LEMMA. *Let* $G = \begin{bmatrix} A & B & C \\ B' & D & E \\ C' & E' & F \end{bmatrix}$ *be a partitioned positive definite symmetric matrix. In order that the upper left submatrix (of the same size as* A*) of* G^{-1} *be the same as that of* $\begin{bmatrix} A & B \\ B' & D \end{bmatrix}^{-1}$, *it is necessary and sufficient that* $C = BD^{-1}E$.

PROOF. The inverse of $\begin{bmatrix} A & B \\ B' & D \end{bmatrix}$ is a matrix $\begin{bmatrix} \alpha & \beta \\ \beta' & \delta \end{bmatrix}$ which satisfies

$$\alpha A + \beta B' = I \tag{5}$$

$$\alpha B + \beta D = 0 \tag{6}$$

$$\beta' B + \delta D = I \tag{7}$$

where the I's are identity matrices of appropriate size and 0 is a matrix of 0's. It follows that

$$\beta = -\alpha B D^{-1}, \tag{8}$$

and

$$\alpha(A - BD^{-1}B') = I\,. \tag{9}$$

Applying the same operations to G, we see that for the upper left submatrix of G^{-1} to be the same as that of $\begin{bmatrix} A & B \\ B' & D \end{bmatrix}^{-1}$ it is necessary and sufficient that

$$A - BD^{-1}B' = A - (B, C)\begin{pmatrix} D & E \\ E' & F \end{pmatrix}^{-1}\begin{pmatrix} B' \\ C' \end{pmatrix}. \tag{10}$$

Letting

$$H = C - BD^{-1}E\,, \tag{11}$$

we shall show that (10) is equivalent to $H = 0$. We obtain the following equation equivalent to (10).

$$\begin{aligned} BD^{-1}B' &= [BD^{-1}(D, E) + (0, H)]\begin{pmatrix} D & E \\ E' & F \end{pmatrix}^{-1}\left[\begin{pmatrix} D \\ E' \end{pmatrix} D^{-1}B' + \begin{pmatrix} 0 \\ H' \end{pmatrix}\right] \\ &= BD^{-1}(I, 0)\begin{pmatrix} D \\ E' \end{pmatrix} D^{-1}B' + BD^{-1}(I, 0)\begin{pmatrix} 0 \\ H' \end{pmatrix} \\ &\qquad + (0, H)\begin{pmatrix} I \\ 0 \end{pmatrix} D^{-1}B' + (0, H)\begin{pmatrix} D & E \\ E' & F \end{pmatrix}^{-1}\begin{pmatrix} 0 \\ H' \end{pmatrix} \\ &= BD^{-1}B' + (0, H)\begin{pmatrix} D & E \\ E' & F \end{pmatrix}^{-1}(0, H)'. \end{aligned} \tag{12}$$

Since $\begin{bmatrix} D & E \\ E' & F \end{bmatrix}^{-1}$ is positive definite, this is equivalent to $H = 0$, which completes the proof.

In the applications we shall make, G is an information matrix. A is obtained from the partial derivatives with respect to the parameters we are interested in estimating, D from a set of parameters which (in general) make the problem more difficult, and F from those parameters which we wish to show to be superfluous. The upper left submatrix of G^{-1} tells us the asymptotic covariance matrix of the best estimates of the parameters in which we are interested. Thus the lemma gives us a condition for the problem with the unknown parameters added in F to be asymptotically no harder than the problem in which these parameters are known.

4. Testing for the median of a symmetric density

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, and intuitively it is likely that the results obtained in this case also hold for the more difficult problem considered in the following section.

Let p be an unknown density with respect to Lebesgue measure on the real line, symmetrical about 0, that is

$$p(x) = p(-x) \tag{13}$$

for all x. Let $X_1, \cdots, X_n$ be real random variables such that for some unknown ξ, $X_1 - \xi, \cdots, X_n - \xi$ are distributed according to the density p. We want to test H_0: $\xi = 0$ against H_1: $\xi > 0$ for large n. We shall suppose

$$I = \int \frac{p'^2(x)}{p(x)} dx < \infty , \tag{14}$$

where p' is the derivative of p, so that if we knew p, we should have a regular problem. In order to avoid inessential complications, we shall suppose p continuously differentiable, although (14) could be given a more general interpretation.

If p is known, the best asymptotic mean-squared error attainable for estimating ξ is $1/nI$ and the best asymptotic power for testing $\xi = 0$ against $\xi > 0$ at the level of significance α is

$$\beta(\xi) = \frac{1}{\sqrt{2\pi}} \int_{c_\alpha - \xi\sqrt{nI}}^{\infty} e^{-x^2/2} dx \tag{15}$$

where

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{c_\alpha}^{\infty} e^{-x^2/2} dx . \tag{16}$$

This asymptotic power is achieved (at least for small ξ) by the test which rejects H_0 if $-\sum [p'(X_i)/p(X_i)]$ is larger than a certain constant depending on n. One gets approximately the right significance level by taking this constant to be $c_\alpha\sqrt{nI}$. We shall show that the same asymptotic power is attainable if p is unknown.

In this case it seems plausible to estimate $p'(|X_i|)/p(|X_i|)$, and, the estimate being Z_i, to reject H_0 if

$$\sum Z_i \operatorname{sgn} X_i \tag{17}$$

is in the upper proportion α of the 2^n numbers $\sum \pm Z_i$ with the signs assigned in all possible ways. Here we are influenced by the identity

$$\frac{p'(-x)}{p(-x)} = -\frac{p'(x)}{p(x)} . \tag{18}$$

It happens that this works provided

$$Z_i = \varphi_i(|X_1|, \cdots, |X_n|), \tag{19}$$

that is, the estimates depend only on the absolute values of the observations, and certain conditions of consistency [see equations (24) and (25)] are satisfied. Let us look at the conditional distribution of $\sum Z_i \operatorname{sgn} X_i$ given $|X_1|, \cdots, |X_n|$ when the X_i are distributed according to $p(X - \xi)$. For given $|X_1|, \cdots, |X_n|$ the sgn X_i are conditionally independent with

$$P\{\operatorname{sgn} X_i = 1 \mid |X_1|, \cdots, |X_n|\} = \frac{p(|X_i| - \xi)}{p(|X_i| - \xi) + p(|X_i| + \xi)}. \tag{20}$$

By the central limit theorem (see, for example, section 21 in Loève [3]) the conditional distribution of $\sum Z_i \operatorname{sgn} X_i$ given $|X_1|, \cdots, |X_n|$ nearly normal with mean

$$\mu = \sum \frac{p(|X_i| - \xi) - p(|X_i| + \xi)}{p(|X_i| - \xi) + p(|X_i| + \xi)} Z_i \tag{21}$$

and variance

$$\sigma^2 = \sum Z_i^2 \left[1 - \left(\frac{p(|X_i| - \xi) - p(|X_i| + \xi)}{p(|X_i| - \xi) + p(|X_i| + \xi)}\right)^2\right] \tag{22}$$

provided

$$\frac{\max Z_i^2}{\sigma^2} \tag{23}$$

is small.

It is intuitively clear that the Z_i can be chosen in such a way that (at least if p is sufficiently regular)

$$\mu + n\xi \int \frac{p'^2(x)}{p(x)}\, dx = o_p(\sqrt{n}) \tag{24}$$

and with high probability

$$\frac{\sigma^2}{n \int \frac{p'^2(x)}{p(x)}\, dx} \approx 1 \tag{25}$$

for $\xi = O(n^{-1/2})$, $n \to \infty$. This will insure that the power of the test is nearly that attainable with known p. For example we may take

$$Z_i = \frac{2\left\{S_n\left(|X_i| + \frac{a}{\sqrt{n}}\right) - 2S_n(|X_i|) + S_n\left(|X_i| - \frac{a}{\sqrt{n}}\right)\right\}}{\frac{a}{\sqrt{n}}\left\{S_n\left(|X_i| + \frac{a}{\sqrt{n}}\right) - S_n\left(|X_i| - \frac{a}{\sqrt{n}}\right)\right\}} \tag{26}$$

where S_n is the empirical cumulative distribution function of the $|X_i|$, and a is a positive constant independent of n. However, in order to satisfy (23), any Z_i for which $Z_i^2 \big/ \sum Z_j^2 > \epsilon_n$ (where $\epsilon_n \downarrow 0$ as $n \to \infty$, but not too rapidly) must be cut down so that this is no longer the case.

5. Two distributions differing only in location and scale

For this problem I have not been able to obtain a complete solution. However, a simple computation will show that formally the problem of estimating (or testing) the difference of location parameters or ratio of scale parameters is as difficult when the form of the distribution is known as it is when the form depends in a regular way on an unknown parameter. I shall also indicate a method which seems likely to yield efficient tests for the hypotheses of equal location parameters or equal scale parameters.

Let p be an unknown density with respect to Lebesgue measure on the real line, and ξ, η, σ, τ unknown real numbers with σ, $\tau > 0$. With m and n large, $X_1, \cdots, X_m$ are independently distributed with density $p(x - \xi/\sigma)/\sigma$ and $Y_1, \cdots, Y_n$ independently distributed with density $p(y - \eta/\tau)/\tau$. We are interested in estimating $\eta - \xi$ and τ/σ and in testing hypotheses concerning these quantities. We shall suppose p continuously differentiable and

$$\int \frac{p'^2(x)}{p(x)}\,dx < \infty\,, \qquad \int \frac{x^2 p'^2(x)}{p(x)}\,dx < \infty \tag{27}$$

which will insure that the problem would be a regular one if we knew p.

Now let us imagine that we know p except for a single unknown parameter γ, so that we shall write p_γ for p. We suppose

$$\int \left(\frac{\partial \log p_\gamma(x)}{\partial \gamma}\right)^2 p_\gamma(x)\,dx < \infty\,. \tag{28}$$

We shall give essentially the information matrix for the five unknown parameters ξ, η, σ, τ, γ. However, since we are particularly interested in $\eta - \xi$ and τ/σ, it is more convenient to use as coordinates λ_1, λ_2, λ_3, λ_4, and γ where

$$\lambda_1 = \sqrt{\frac{\tau}{\sigma}} \tag{29}$$

$$\lambda_2 = \frac{\eta - \xi}{2} \tag{30}$$

$$\lambda_3 = \sqrt{\sigma\tau} \tag{31}$$

$$\lambda_4 = \frac{\xi + \eta}{2}\,. \tag{32}$$

The information matrix, evaluated at $\lambda_1 = \lambda_3 = 1$, $\lambda_2 = \lambda_4 = 0$ (that is, $\xi = \eta = 0$, $\sigma = \tau = 1$) is

$$\begin{pmatrix} (m+n)\,a, & (m+n)\,b, & (n-m)\,a, & (n-m)\,b, & (m-n)\,c \\ (m+n)\,b, & (m+n)\,d, & (n-m)\,b, & (n-m)\,d, & (m-n)\,e \\ (n-m)\,a, & (n-m)\,b, & (m+n)\,a, & (m+n)\,b, & -(m+n)\,c \\ (n-m)\,b, & (n-m)\,d, & (m+n)\,b, & (m+n)\,d, & -(m+n)\,e \\ (m-n)\,c, & (m-n)\,e, & -(m+n)\,c, & -(m+n)\,e, & (m+n)\,f \end{pmatrix} \tag{33}$$

where

$$a = E[1 + X\varphi_1(X, \gamma)]^2 \tag{34}$$

$$b = EX\varphi_1^2(X, \gamma) \tag{35}$$

$$c = EX\varphi_1(X, \gamma)\varphi_2(X, \gamma) \tag{36}$$

$$d = E\varphi_1^2(X, \gamma) \tag{37}$$

$$e = E\varphi_1(X, \gamma)\varphi_2(X, \gamma) \tag{38}$$

$$f = E\varphi_2^2(X, \gamma) \tag{39}$$

$$\varphi_1(x, \gamma) = \frac{\partial \log p_\gamma(x)}{\partial x} \tag{40}$$

$$\varphi_2(x, \gamma) = \frac{\partial \log p_\gamma(x)}{\partial \gamma} \tag{41}$$

and the expectations are for X distributed according to p_γ.

By the lemma of section 2, in order that the addition of the unknown parameter γ should not make the problem of estimating $\lambda_1 = \sqrt{\tau/\sigma}$ and $\lambda_2 = (\eta - \xi)/2$ more difficult, it is necessary and sufficient that

$$\begin{pmatrix} (m-n)\,c \\ (m-n)\,e \end{pmatrix} = \begin{pmatrix} (n-m)\,a, & (n-m)\,b \\ (n-m)\,b, & (n-m)\,d \end{pmatrix} \begin{pmatrix} (m+n)\,a, & (m+n)\,b \\ (m+n)\,b, & (m+n)\,d \end{pmatrix}^{-1} \begin{pmatrix} -(m+n)\,c \\ -(m+n)\,e \end{pmatrix}, \tag{42}$$

which is true. Since any parameter point ξ, η, σ, τ can be reduced to $\xi = \eta = 0$, $\sigma = \tau = 1$ by affine transformations of the X's and Y's separately, this result must hold at every parameter point.

In the comparatively simple case where it is known that $\sigma = \tau$ so that we may as well take their common value to be 1, it seems reasonable to test the hypothesis H_0: $\eta = \xi$ against $\eta > \xi$ in the following way. Much as in section 4, we estimate $p'(X_i)/p(X_i)$, $p'(Y_j)/p(Y_j)$. Instead of requiring the estimate to depend only on the absolute values, we require the estimate to ignore the labelling as X's and Y's, that is, to depend only on the numerical values of the $m + n$ observations. Let the resulting estimate of $p'(X_i)/p(X_i)$ be Z_i, and of $p'(Y_j)/p(Y_j)$ be Z_j'. Reject H_0 if $\sum Z_i - \sum Z_j'$ is large among the $\binom{m+n}{m}$ numbers $\sum_{k=1}^{m+n} \pm U_k$ where the U_k are the Z_i and Z_j' and the signs are assigned in all possible ways subject to the condition that exactly m of them are positive. However, it seems difficult to prove the asymptotic normality of $\sum Z_i - \sum Z_j'$ under the alternative hypothesis.

6. Estimation of a linear relation

Let $X_1, \cdots, X_n$ be independently distributed random points in a two-dimensional real linear space. The X_i are observed and it is assumed that their distribution can be described with the aid of unobservable random points Y_i, Z_i in the following way.

$$X_i = Y_i + Z_i, \tag{43}$$

where $Y_1, \cdots, Y_n, Z_1, \cdots, Z_n$ are independently distributed, the Y_i with common unknown distribution concentrated on an unknown line L, and the Z_i with a common unknown bivariate normal distribution. We are interested in estimating the slope of L. Reiersøl [4] has shown that the slope is identifiable if the distribution of the Y_i is required to be nonnormal. (A distribution concentrated at one point will be considered a special case of a normal distribution.) Neyman [5] has given a consistent estimate in this case. Further work has been done by Jeeves [6] and Wolfowitz.

We shall consider the parametric problem in which, except for one unknown parameter entering in an arbitrary way, the distribution is known to within an affine transformation of the plane. However, we find it convenient to modify the specification of the problem in an inessential way. First we assume the Z_i have mean 0, thus shifting their contribution to the mean to the Y_i. Then we can represent the Z_i as $U_i + V_i$ where the U_i lie on a line parallel to L and the V_i on another line M, and $U_1, \cdots, U_n, V_1, \cdots, V_n$ are independently normally distributed.

Letting

$$Z_i' = V_i \tag{44}$$

$$Y_i' = Y_i + U_i , \tag{45}$$

we have

$$X_i = Y_i' + Z_i' \tag{46}$$

where the Y_i' and Z_i' satisfy the same type of conditions as before, and in addition the Z_i' are concentrated on a line M through the origin. We now drop the primes.

Suppose that in a particular coordinate system X_i has coordinates X_{1i}, X_{2i}. There are real random variables $R_1, \cdots, R_n, S_1, \cdots, S_n$ (which are merely the Y_i and Z_i measured in some arbitrary units along the lines on which they are concentrated) such that

$$X_{1i} = (\lambda R_i + \eta) \cos \theta - (\mu S_i + \zeta) \sin \varphi \tag{47}$$

$$X_{2i} = (\lambda R_i + \eta) \sin \theta - (\mu S_i + \zeta) \cos \varphi \tag{48}$$

where θ is the angle from the first coordinate axis to L, φ the angle from the second coordinate axis to M, λ and μ are positive real numbers, and η and ζ are real numbers. The R_i are independently distributed with nonnormal density p_γ and the Z_i are independent unit normal random variables. We suppose

$$\int x^2 p_\gamma(x)\, dx < \infty , \tag{49}$$

$$\int \frac{p_\gamma'^2(x)}{p_\gamma(x)}\, dx < \infty , \tag{50}$$

and

$$\int \left(\frac{\partial \log p_\gamma(x)}{\partial \gamma}\right)^2 p_\gamma(x)\, dx < \infty , \tag{51}$$

where $p_\gamma'(x)$ is the derivative of p_γ with respect to x.

Then the information matrix (with one observation) for θ, φ, η, ζ, λ, μ, γ at the point $\theta = \varphi = 0$, $\eta = \zeta = 0$, $\lambda = \mu = 1$ is

$$
(52) \quad \begin{array}{ccccccc}
1, & -ERa(R), & 0, & 0, & 0, & 0, & 0 \\
-ERa(R), & Ea^2(R), & 0, & 0, & 0, & 0, & 0 \\
0, & 0, & Ea^2(R), & 0, & -ERa^2(R), & 0, & -Ea(R)\beta(R) \\
0, & 0, & 0, & 1, & 0, & 0, & 0 \\
0, & 0, & -ERa^2(R), & 0, & E[1+Ra(R)]^2, & 0, & ERa(R)\beta(R) \\
0, & 0, & 0, & 0, & 0, & 2, & 0 \\
0, & 0, & -Ea(R)\beta(R), & 0, & ERa(R)\beta(R) & 0, & E\beta^2(R)
\end{array}
$$

where

$$
(53) \qquad a(r) = \frac{\partial \log p_\gamma(r)}{\partial r}
$$

$$
(54) \qquad \beta(r) = \frac{\partial \log p_\gamma(r)}{\partial \gamma}.
$$

The expectations are taken with R distributed according to p_γ.

Applying the lemma of section 3 we see that the addition of γ as an unknown parameter does not make the problem of estimating θ and φ asymptotically more difficult. Curiously, even the presence of η, ζ, λ, μ as unknown parameters does not make the problem asymptotically more difficult.

REFERENCES

[1] J. Wolfowitz, "Estimation of structural parameters when the number of incidental parameters is unbounded" (abstract), *Annals of Math. Stat.*, Vol. 25 (1954), p. 811.

[2] L. Le Cam, "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates," *Univ. of California Publ. in Stat.*, Vol. 1 (1953), pp. 277–330.

[3] M. Loève, *Probability Theory*, New York, D. Van Nostrand Co., 1955.

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

[5] J. Neyman, "Existence of consistent estimates of the directional parameter in a linear structural relation between two variables," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 497–512.

[6] T. A. Jeeves, "Identification and estimation of linear manifolds in n-dimensions," *Annals of Math. Stat.*, Vol. 25 (1954), pp. 714–723.