JOURNAL OF STATISTICAL THEORY AND PRACTICE
https://doi.org/10.1080/15598608.2017.1399840

Check for updates

# Parametric embedding of nonparametric inference problems

Mayer Alvo[a], Tze Leung Lai[b], and Philip L. H. Yu [c]

Q1

[a]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada; [b]Department of Statistics, Stanford University, Stanford, California, USA; [c]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

Q2

**ABSTRACT**

In 1937, Neyman introduced the notion of smooth tests of the null hypothesis that the sample data come from a uniform distribution on the interval (0,1) against alternatives in a smooth parametric family. This idea can be used to embed various nonparametric inference problems in a parametric family. Focusing on nonparametric rank tests, we show how to derive traditional rank tests by applying this approach. We also show how to use it to obtain simplifying insights and optimality results in complicated settings that involve censored and truncated data, for which it is more convenient to use hazard functions to define the embedded family. We describe an application of the embedding approach to the problem of testing for trend in environmental studies.

**ARTICLE HISTORY**
Received 15 September 2016
Accepted 30 October 2017

**KEYWORDS**

Q3

**AMS SUBJECT CLASSIFICATION**
62G



## 1. Introduction and background

We have maintained repeated research interactions with Bimal Sinha since the 1980s. One of our shared common interests with him is statistical inference with rank data, and with his twin brother Bikas a common interest is generalized linear models. Although the former is inherently nonparametric and the latter is intrinsically parametric, parametric embedding of nonparametric inference problems bridges the apparent gap between them. This is the major theme of the present article, which shows that the key to deriving fundamental results on nonparametric inference with the embedding approach lies in appropriate choice of the parametric family. We give in this section a review of this idea dating back to Neyman (1937). In section 2, we revisit important developments in semiparametric inference from censored and truncated data using this parametric embedding approach as a versatile tool that provides simplifying insights into complicated settings and extends optimality arguments from parametric to nonparametric and semiparametric problems. Section 3 provides some concluding remarks, including an application to environmetrics which is one of Bimal's major research areas.

### 1.1. *Smooth tests of uniform null against embedded alternatives in exponential family and applications to rank data*

Neyman (1937) introduced smooth tests of the null hypothesis that the sample data are generated from a uniform distribution on the interval (0,1) so that the tests have good

**CONTACT** Philip Yu plhyu@hku.hk Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong.

power against alternatives whose probability densities depart smoothly from the null hypothesis. Smooth changes include shifts in mean, variance, skewness, and kurtosis, and the smooth alternative density has the form

$$g(y,\theta) = \exp\left\{\sum_{i=1}^{k} \theta_i h_i(y) - K(\theta)\right\}, 0 < y < 1, \tag{1}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^{\mathrm{T}}$ is a set of unknown parameters, $K(\theta)$ is a normalizing constant, and the $h_i(y)$ are the Legendre polynomials that are orthonormal with respect to the uniform distribution on $(0, 1)$. The null hypothesis can be expressed as $H_0 : \theta = 0$.

A parallel parametric family for inference from a sample consisting of ranking data of $t$ objects can be described as follows. Let $\mathcal{P}$ be the space of $t!$ permutations of the integers $1, 2, \ldots, t$, and let

$$p = (p_1, \ldots, p_{t!})^{\mathrm{T}}$$

with $p_j = p(\omega_j), \omega_j \in \mathcal{P}$, be a probability mass distribution defined on $\mathcal{P}$. Consider the null hypothesis that all rankings are equally likely, that is,

$$H_0 : p = p_0 \triangleq \frac{1}{t!}1$$

against the alternative $H_1 : p \neq p_0$. Proceeding in the spirit of Neyman, let $X$ be a random vector of dimension $k$ defined over $\mathcal{P}$ and let its probability mass function be given by

$$\pi_j(\theta) = \exp\{\theta^{\mathrm{T}} x_j - K(\theta)\} p_{0j}, j = 1, \ldots, t!, \tag{2}$$

where $\theta = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ is a $k$-dimensional vector of unknown parameters, $K(\theta)$ is a normalizing constant, and $X(\omega_j) = x_j$. The $p_{0j}$ represent the values of the probabilities prescribed by the null hypothesis, which reduces to $p_{0j} \equiv \frac{1}{t!}$ for the preceding $H_0$. We can rewrite $H_0$ as $\theta = 0$, or equivalently, $H_0 : \sum_{i=1}^{k} \theta_i^2 = 0$. Since $\sum_{j=1}^{t!} \pi_j(\theta) = 1$, the expectation of the vector $X$ is

$$\eta(\theta) \triangleq E_\theta X = \sum x_j \pi_j(\theta) = \left(\frac{\partial K(\theta)}{\partial \theta_r}\right),$$

and the variance–covariance matrix is

$$Cov_\theta(X) = \left(\frac{\partial^2 K(\theta)}{\partial \theta_r \partial \theta_s}\right).$$

Suppose we take a random sample of $n$ observations. Let $n_j$ denote the frequency of occurrence of the ranking $\omega_j$ with $\Sigma_j n_j = n$. The likelihood function is given by the multinomial distribution and is proportional to

$$L(\theta) = \pi_1^{n_1}(\theta) \pi_2^{n_2}(\theta) \ldots \pi_{t!}^{n_{t!}}(\theta).$$

Taking logs, we have $\log L(\theta) = n\left[\theta^{\mathrm{T}} \hat{\eta} - K(\theta)\right] + C$, where $C$ does not depend on $\theta$ and $\hat{\eta}$ is the usual sufficient statistic given by

$$\hat{\eta} = \left[\sum_{j=1}^{t!} x_j \left(n_j/n\right)\right].$$

The score test calculates the score statistic

$$U(\theta) = \left(\frac{\partial \log L(\theta)}{\partial \theta_r}\right) \tag{3}$$

and rejects for large values of $S_k = [U(\theta_0)]^{\mathrm{T}}[I(\theta_0)]^{-1}[U(\theta_0)]$, where $\theta_0$ is the value set by the null hypothesis, which is 0 in the present case, and

$$I(\theta) = \left(-E_\theta \frac{\partial^2 \log L(\theta)}{\partial \theta_r \partial \theta_s}\right)$$

is the Fisher information matrix. Note that the score statistic does not require the calculation of the maximum likelihood estimate but does require the calculation of the inverse (or generalized inverse) of the information matrix. It can be shown that for large $n$, $S_k \Rightarrow_{\mathcal{L}} \chi_f^2$, where $f$ is the rank of $I(\theta_0)$. The Neyman smooth tests of fit provide a blueprint for deriving various tests involving the use of ranks, as shown by Alvo (2016). This application of the blueprint dated back to the Nobel Laureate (in economics) Milton Friedman in 1937 when he studied statistics and economics as a PhD student at Columbia University.

**Example 1**. Suppose $n$ judges rank $t$ objects in accordance with some criterion. Let $X$ be the $t$-dimensional random vector of adjusted ranks for which

$$X\left(\omega_j\right) = \left(\omega_j(1) - \frac{t+1}{2}, \ldots, \omega_j(t) - \frac{t+1}{2}\right)^{\mathrm{T}}, j = 1, \ldots, t!.$$

Under the null hypothesis $H_0 : \theta = 0$, or equivalently that the $t!$ possible rankings have the same probability of being chosen,

$$Cov_0(X) = \frac{(t+1)}{12}[tI - J_t],$$

and where $J_t$ is a matrix of 1's. Hence, the score test statistic becomes

$$S_k = [U(0)]^{\mathrm{T}}[I(0)]^{-1}[U(0)] = \frac{12n}{t(t+1)} \sum_{i=0}^{t} \left(\bar{R}_i - \frac{(t+1)}{2}\right)^2,$$

where $\bar{R}_i$ is mean rank for object $i$. This is the test of Friedman (1937), who showed that $S_k$ has a chi-squared distribution with $(t-1)$ degrees of freedom under the null hypothesis. In section 1.2, we use the parametric embedding to derive some optimum properties of this test.

### 1.2. Locally most powerful rank tests

Lehmann and Stein (1949) and Hoeffding (1951) pioneered the development of an optimality theory for nonparametric tests, parallel to that of Neyman and Pearson (1933) and Wald (1949) for parametric testing. They considered nonparametric hypotheses that are invariant under permutations of the variables in $k$-sample problems[1] so that rank statistics are the maximal invariants, and extended the Neyman–Pearson and Wald theories for independent observations to the joint density function of the maximal invariants. Terry (1952) and others subsequently implemented and refined Hoeffding's approach to show that a number of rank tests are locally most powerful at certain alternatives near the null hypothesis. In particular, for $k=2$, let $R_1 < \cdots < R_m$ denote the ranks of sample 1 (with sample size $m$) in the combined sample of $n$ independent observations. Suppose sample 1 is generated from a distribution with density function $g$. Let $V_{(1)}, \ldots, V_{(n)}$ denote the order statistics of the combined sample; Hoeffding (1951) introduced the change-of-measure formula

$$P\{R_1 = r_1, \ldots, R_m = r_m\} = E_g\left[\frac{f(V_{(r_1)})}{g(V_{(r_1)})} \cdots \frac{f(V_{(r_m)})}{g(V_{(r_m)})}\right] \Big/ \binom{n}{m}, \tag{4}$$

where $E_g$ denotes expectation with respect to the probability measure under which the $n$ observations are i.i.d. with common density function $g$, assuming that $g$ is positive whenever $f$ is. In particular, consider testing $H_0 : f = g$ versus the location alternative $f(x) = g(x-\theta)$ for small positive values of $\theta$. In this case, differentiating both sides of Eq. (4) with respect to $\theta$ and letting $\theta \downarrow 0$ yield

$$\frac{\partial}{\partial\theta} P\{R_1 = r_1, \ldots, R_m = r_m\}|_{\theta=0} = -\sum_{i=1}^{m} E_g\left[\frac{g'(V_{(r_i)})}{g(V_{(r_i)})}\right] \Big/ \binom{n}{m}. \tag{5}$$

Hence by an extension of the Neyman–Pearson lemma, the derivative of the power function at $\theta = 0$ is maximized by a test that rejects $H_0$ when the right-hand side of Eq. (4) exceeds some threshold $C$, which is chosen so that the test has type I error $\alpha$ when $\theta = 0$. This test, therefore, is locally most powerful, for testing alternatives of the form $f(x) = g(x-\theta)$, with $\theta \downarrow 0$, and examples include the Fisher–Yates test when $g$ is standard normal and the Wilcoxon test when $g(x) = e^x/(1+e^x)^2$ is the logistic density.

A parametric embedding argument similar to the second paragraph of section 1.1 can be used to give an alternative derivation of the local optimality of the Fisher–Yates and Wilcoxon tests. Generalize Eq. (2) from the case $k=1$ to $k=2$ by defining

$$\pi_j(\theta_1, \theta_2) = \exp\left\{\sum_{\ell=1}^{2} \left[\theta_\ell^{\mathrm{T}} x_{\ell j} - K(\theta_\ell)\right]\right\} p_{0j}, j = 1, \ldots, n!, \tag{6}$$

where $\theta_\ell = (\theta_{\ell 1}, \ldots, \theta_{\ell k})^{\mathrm{T}}$ represents the parameter vector for sample $\ell(=1,2)$ and $x_{1j}, x_{2j}$ are the data from sample 1 and sample 2 with respective sizes $m$ and $n-m$ that are associated with the ranking (permutation) $\omega_j, j = 1, \ldots, n!$. Under the null hypothesis $H_0 : \theta_1 = \theta_2$, we can assume without loss of generality that the underlying $V_1, \ldots, V_n$ from the combined sample are i.i.d. uniform (by considering $G(V_i)$, where $G$ is the common

[1]Lehmann and Stein considered the case $k=2$ and Hoeffding general $k$, including $k=1$.

distribution function, assumed to be continuous, of the $V_i$) and that all rankings of the $V_i$ are equally likely. Hence Eq. (6) represents an exponential family constructed by exponential tilting of the baseline measure (i.e., corresponding to $H_0$) on the rank-order data. This has the same spirit as Neyman's smooth test of the null hypothesis that the data are i.i.d. uniform against alternatives in the exponential family of Eq. (1). Neyman and Pearson (1936, 1938) applied the Neyman–Pearson lemma to show that the score tests based on the statistics Eq. (3) have maximum local power at the alternatives in Eq. (1) that are near $\theta = 0$. The parametric embedding of Eqs. (2) or (6) makes these results directly applicable to the rank-order statistics. In particular, this shows that the two-sample Wilcoxon test of $H_0$ is locally powerful for testing the uniform distribution against the truncated exponential distribution for which the $x_{\ell j}$ are constrained to lie in the range $(0, 1)$ of the uniform distribution. Note that these exponential tilting alternatives differ from the location alternatives in the preceding paragraph not only in their distributional form (truncated exponential instead of logistic) but also in avoiding the strong assumption of the preceding paragraph that the data have to be generated from the logistic distribution even under the null hypothesis.

### 1.3. *Lecam's local asymptotic normality and hajek-lecam theory*

The local alternatives in section 1.2 refer to $\theta$ near the value(s) $\theta_0$ assumed by the null hypothesis. The sample size $n$ is not involved in the analysis of local power. On the other hand, the central limit theorem has played a major role in the development of rank tests, as asymptotic normality is used to provide approximate critical values under the null hypothesis and to approximate the power function under alternatives within $O(n^{-1/2})$ from $\theta_0$. LeCam (1960) introduced a fundamental concept, which he called *contiguity* of a sequence $Q_n$ of probability measures to another sequence $P_n$, written $Q_n \triangleleft P_n$, defined by the property that the likelihood ratio $dQ_n/dP_n$ is bounded in probability (under $P_n$) as $n \to \infty$.[2] He proved three key results that have been called LeCam's first, second, and third lemmas, related to the log-likelihood ratio $\log(dQ_n/dP_n)$; see Hajek et al. (1999, Section 7.1) and van der Vaart (1998, Section 6.2). Hajek (1962) applied this theory to rank tests of the null hypothesis $H_0 : \beta = 0$ in the simple regression model $Y_i = \alpha + \beta u_i + \varepsilon_i$, in which $\varepsilon_i$ are i.i.d. with common density function $f$, using linear rank statistics of the form

$$S_n = \sum_{i=1}^{n} (u_i - \bar{u}) \varphi\left(\frac{R_i}{n+1}\right). \tag{7}$$

He derived the asymptotic normality of $S_n$ under the null hypothesis and contiguous alternatives, and showed the test to have asymptotically maximum power uniformly for these alternatives if $\varphi = -(f' \circ F^{-1})/(f \circ F^{-1})$, where $F$ is the distribution function with derivative $f$. Note that this result is consistent with the choice of the score function given by Eq. (5) for locally most powerful tests in section 1.2. Hajek (1968) subsequently introduced the projection method to extend these results to local alternatives that need not be contiguous to the null.

[2]This is also equivalent to $P_n(A_n) \to 0$ if and only if $Q_n(A_n) \to 0$ for any sequence of events $A_n$.

The rank tests in the preceding paragraph deal with the regression setting, which is related to the location alternatives in the first paragraph of section 1.2. If we focus on $k$-sample problems, then parametric embedding as in the second paragraph of that section can be applied and the idea of *local asymptotic normality* (LAN), which was also introduced by LeCam (1960) in conjunction with contiguity, can be applied to derive the LAN property of the embedded family. As pointed out by Van Der Vaart (1998, Chap. 7), a sequence of parametric models is LAN if asymptotically (as $n \to \infty$) their likelihood ratio processes behave like those for the normal mean model via a quadratic expansion of the log-likelihood function. Hajek (1970; 1972) and LeCam (1972) made use of the LAN property to derive asymptotic optimality in parametric estimation and testing via convolution theorems and local asymptotic minimax bounds; see Van Der Vaart (1998, Chap. 8). In the next section we discuss these results further and generalize them in the much more complicated setting of censored and truncated data. We also apply the generalization of parametric embedding to revisit a number of major developments for these data.

## 2. Parametric embedding approach to rank tests with censored and truncated data

Extension of rank tests to censored data began with Gehan's (1965) extension of the Wilcoxon test and Mantel's (1966) logrank test. An idea similar to Gehan's was extended to truncated data by Bhattacharya, Chernoff, and Yang (1983). Lai and Ying (1991; 1992) gave a unified treatment of rank statistics for left-truncated and right-censored (LTRC) data. Section 2.2 gives an overview of the developments of rank tests for these incomplete data, highlighting the difficulties caused by ranking incomplete data and describing important landmarks in overcoming these difficulties. In section 2.1 we generalize the parametric embedding approach to give a new derivation of what these landmarks have finally led to. More importantly, coupled with the LAN and local minimaxity results of section 2.3, the approach introduced in section 2.1 yields asymptotically optimal tests for local alternatives in the embedded parametric family. Since the actual alternatives are unknown, the problem of adaptive (data-dependent) choice of the score function for rank tests has witnessed important developments. Section 2.3 gives a brief review of this topic and its implications on the choice of the parametric family in parametric embedding.

### 2.1. *Extension of parametric embedding to censored data*

We begin with the right-censored case for which our basic idea of using the hazard function instead of the density function for exponential tilting can become transparent. For complete data $V_1, \ldots, V_n$, the parametric embedding of Eq. (2) or Eq. (6) assumes (a) equally likely rankings that give rise to $p_{0j}$ and i.i.d. uniform $G(V_1), \ldots, G(V_n)$ under the null hypothesis, and (b) exponential tilting via distinct values of $x_j$ that are functions of the ranks as in Example 1. The $V_i$ are not completely observable when the data are censored so that the observations are $(\tilde{V}_i, \delta_i)$, where $\tilde{V}_i = \min(V_i, c_i)$ and $\delta_i = I_{\{V_i \le c_i\}}$. Since the rank assigned to $V_i$ for complete data is the empirical distribution function evaluated to $V_i$, the analog for censored data is $\hat{G}(\tilde{V}_i)$, where $\hat{G}$ is the Kaplan–Meier estimator, which is the nonparametric maximum likelihood estimator (MLE) of $G$ for censored data. Hence the model under the null hypothesis is that of i.i.d. uniform random variables censored by $G(c_i)$, providing a partial analog of (a). Since $\hat{G}$ puts all its mass at the uncensored observations (with $\delta = 1$), this causes some difficulty in generalizing (b) because the sample also contains censored observations. Note that at each uncensored observation $\tilde{V}_i$, the information in the ordered sample conveys not only the value of $V_i$ but also how many observations $\tilde{V}_j$ in the sample are $\geq \tilde{V}_i$. When the $V_i$ denote failure times in survival analysis, this means the size of the risk set, that is, the number of subjects who are at risk at an observed failure time $V_i$. This resolves the inherent difficulty of ordering the censored observations for which the actual failure times are unknown except for their exceedance over $c_i$. To rank the data, we need to have a total order of the sample space, but the subset consisting of censored observations cannot be totally ordered because the underlying failure times are unknown. Using the observed failure time and the risk set size at each uncensored observations gives a partial analog of the ranking for complete data. To be at risk at an observed failure time $V_i$, the subject cannot fail prior to $V_i$. The jump $\Delta\hat{G}(\tilde{V}_i)$ basically measures the conditional probability of failing in an infinitesimal interval around $\tilde{V}_i$ given that failure has not occurred prior to $\tilde{V}_i$. This means that we should think of hazard functions instead of density functions and perform exponential tilting using the hazard functions rather density functions. The hazard function is related to the density function $f$ via

$$\lambda(t) = \frac{f(t)}{\int_t^\infty f(u)\,du}. \tag{8}$$

Consider the two-sample problem with censored data. Let $V_{(1)} < \cdots < V_{(k)}$ denote the ordered uncensored observations of the combined sample, $N_j$ (resp. $M_j$) denote the number of observations in the combined sample (resp. in sample 1) that are $\geq V_{(j)}$, and $u_j = 1$ (resp. 0) if $V_{(j)}$ comes from sample 1 (resp. sample 2). Note that $\{(1), \ldots, (k), M_1, N_1, \ldots, M_k, N_k\}$ is invariant under the group of strictly increasing transformations on the testing problem. The model under the null hypothesis is described in the preceding paragraph, and we now introduce embedding of the null model into a smooth parametric family that also consists of alternatives. Instead of tilting the density functions as in Eq. (2) or Eq. (6), we define the change of measures via intensity (hazard) functions, as in Section II.7 of Andersen et al. (1993). Because the normalizing constant $e^{-K(\theta)}$ in Eq. (2) gets canceled in the numerator and denominator of Eq. (8), it does not appear in the likelihood ratio statistic. On the other hand, the denominator of Eq. (8) will induce a function $\lambda_0(t)$, which can be chosen as the baseline (or null hypothesis) hazard function, in the likelihood ratio. An analog of Eq. (2) or Eq. (6) therefore takes the proportional hazards form

$$\lambda_j(t) = \lambda_0(t) \exp\left(\theta^{\mathrm{T}} x_j\right). \tag{9}$$

We discuss in the following the choice of $x_j$ that extends $x_j = X(\omega_j)$ in Eq. (2) to LTRC data, for which we also define the hazard-induced rank statistics.

## 2.2. *From Gehan and Bhattacharya et al. to hazard-induced rank tests*

In this section we first focus on some landmark developments of two-sample rank tests for censored data in the literature and then show how $x_j$ in Eq. (9) can be chosen on the basis of the insights provided by these developments. We next show how these two-sample rank statistics can be extended to the $k$-sample and regression settings, and then further extend them for left-truncated and LTRC data.

The first landmark development was Gehan's extension of the Mann–Whitney version of the Wilcoxon test to censored data. Let $T_i$ (resp. $T_j'$) denote the actual failure times of sample 1 (resp. sample 2), and $(\tilde{T}_i, \delta_i)$ and $\left(\tilde{T}_j', \delta_j'\right)$ be the corresponding observations. For complete data, the Mann–Whitney statistic is $W = \sum_{i=1}^{m} \sum_{j=1}^{n-m} w(T_i, T_j')$, where $w(t, t') = 1$ (resp. $-1$) if $t > t'$ (resp. $t < t'$), and $w(t, t') = 0$ if $t = t'$. For censored data, Gehan replaced $w(T_i, T_j')$ by

$$w\left(\tilde{T}_i, \delta_i; \tilde{T}_j', \delta_j'\right) = \begin{pmatrix} -1 & \text{if } \tilde{T}_i \le \tilde{T}_j' \text{and } \delta_i = 1 \\ 1 & \text{if } \tilde{T}_i \ge \tilde{T}_j' \text{and } \delta_j' = 1 \\ 0 & \textit{otherwise}, \end{pmatrix} \tag{10}$$

noting that comparisons can be made if the smaller of $\tilde{T}_i$ and $\tilde{T}_j'$ is uncensored.[3] Breslow (1970) subsequently extended this to the $k$-sample case and expressed $W$ in the counting process form

$$W = \int Y(s)\, dN'(s) - \int Y'(s)\, dN(s), \tag{11}$$

where $N(s) = \sum_{i=1}^{m} I_{\{\tilde{T}_i \le s, \delta_i = 1\}}$, $N'(s) = \sum_{j=1}^{n-m} I_{\{\tilde{T}_j' \le s, \delta_j' = 1\}}$, and $Y(s) = \sum_{i=1}^{m} I_{\{\tilde{T}_i \ge s\}}$ and $Y'(s) = \sum_{j=1}^{n-m} I_{\{\tilde{T}_j' \ge s\}}$ are the corresponding risk set sizes.

Instead of the weight processes $Y$ and $Y'$ that depend on both failures and censoring, Prentice (1978) suggested that a better alternative should depend on the survival experience in the combined sample. For complete data the classical two-sample rank statistics have the form $S_n = \sum_{i=1}^{m} a_n(R_i)$, where the scores $a_n(j)$ are obtained from a score function $\varphi$ on $(0, 1]$ by $a_n(j) = \varphi(j/n)$ so that $S_n = \sum_{i=1}^{m} \varphi(G_n(T_i))$, where $G_n$ is the distribution of the combined sample, or by some asymptotically equivalent variant such as the expected value of $\varphi$ evaluated at the $j$th uniform order statistic from a sample of size $n$. As pointed out in the first paragraph of section 2.1, the counterpart of $G_n(T_i)$ for censored data is $\hat{G}_n(\tilde{T}_i)$, where $\hat{G}_n$ is the Kaplan–Meier estimate based on the combined sample. If $\delta_i = 1$, $\tilde{T}_i$ is the actual failure time and has score $\varphi(\hat{G}_n(\tilde{T}_i))$. On the other hand, if $\delta_i = 0$, then the actual failure time $T_i$ is unknown, other than that it exceeds $\tilde{T}_i$ and therefore has score $\Phi(\hat{G}_n(\tilde{T}_i))$, where

$$\Phi(t) = \int_t^1 \varphi(u)\, du/(1 - t), \quad 0 \le t < 1, \tag{12}$$

[3] In fact, Gehan introduced a further refinement depending on whether the larger observation is censored or not.

represents the average of scores $\varphi(u)$ with $u \geq t$. This leads to the following extension of the classical rank statistic $\sum_{i=1}^{m} \varphi(G_n(T_i))$ to censored data:

$$S_n^* = \sum_{i=1}^{m} \{\delta_i \varphi(\tilde{T}_i) + (1-\delta_i)\Phi(\tilde{T}_i)\}. \tag{13}$$

Prentice (1978) conjectured the asymptotic equivalence of Eq. (13) to another class of rank statistics that he proposed for censored data based on the generalized rank vector, which is a permutation of $\{1, \ldots, n\}$ of the form

$$R = \Big[(1), \ldots, (k); \{(i\,1), \ldots, (i\,\nu_i)\}_{i=0,\ldots,k}\Big], \tag{14}$$

where $V_{(1)} < \cdots < V_{(k)}$ are the ordered uncensored observations of the combined sample (as in section 2.1) and $\{\tilde{V}_{(i\,1)}, \ldots, \tilde{V}_{(i\,\nu_i)}\}$ is the unordered set of censored observations between $V_{(i)}$ and $V_{(i+1)}$, setting $V_{(0)} = 0$. Cuzick (1985) proved this conjecture under some smoothness assumptions on $\varphi$ and also extended the proof to show in his Section 3 the asymptotic equivalence of Eq. (13) and

$$S_n = \sum_{j=1}^{k} \psi\big(\hat{G}_n(V_{(j)})\big)\left(u_j - \frac{M_j}{N_j}\right), \quad \textit{where}\ \psi = \varphi - \Phi. \tag{15}$$

This form of rank statistics for censored data dated back to Mantel (1966) with $\psi = 1$. As shown by Gu, Lai, and Lan (1991), there is a one-to-one correspondence between $\varphi$ and $\psi$:

$$\varphi(t) = \psi(t) - \int_0^t \frac{\psi(s)}{1-s}\, ds, \quad 0 < t < 1,$$

and rank statistics of the form in Eq. (15) can be expressed in the form of generalized Mann–Whitney statistics $W = \sum_{i=1}^{m} \sum_{j=1}^{n-m} w\left(\tilde{T}_i, \delta_i; \tilde{T}_j', \delta_j'\right)$ with

$$w\left(\tilde{T}_i, \delta_i; \tilde{T}_j', \delta_j'\right) = \begin{pmatrix} -n\omega\big(\hat{G}_n(\tilde{T}_i)\big)/Y.(\tilde{T}_i) & \textit{if}\ \tilde{T}_i \leq \tilde{T}_j' \text{ and } \delta_i = 1 \\ n\omega\big(\hat{G}_n(\tilde{T}_j)\big)/Y.(\tilde{T}_j) & \textit{if}\ \tilde{T}_i \geq \tilde{T}_j' \text{ and } \delta_j' = 1 \\ 0 & \textit{otherwise}, \end{pmatrix} \tag{16}$$

where $Y.(s) = \sum_{i=1}^{m} I\{\tilde{T}_i \geq s\} + \sum_{j=1}^{n-m} I\{\tilde{T}_i \geq s\}$ is the risk set size of the combined sample at $s$.

The representation (15) is convenient for extensions from two-sample to the regression settting in which $u_j$are the covariates in the regression model $V_i = \beta u_i + \varepsilon_i$, as in Eq. (7). The $M_j/N_j$ in Eq. (15) is now generalized to

$$\bar{u}_j = \left(\sum_{i=1}^{n} u_i I_{\{\tilde{V}_i \geq \tilde{V}_{(j)}\}}\right) / N_j, \tag{17}$$

which is the average value of the covariate associated with the risk set at the uncensored observation $\tilde{V}_{(j)}$. Lai and Ying (1991, Theorem 1) established the asymptotic normality of these rank statistics under the null hypothesis $H_0 : \beta = 0$ and under local alternatives. Analogous to the complete data case, these tests are asymptotically efficient when $\psi = (\lambda' \circ F^{-1})/(\lambda \circ F^{-1})$, where $F$ is the common distribution function and $\lambda$ the hazard function of the $\varepsilon_i$. They proved this result when the data can also be subject to left truncation.

Suppose $(u_i, V_i, \delta_i)$ can be observed only when $\tilde{V}_i = \min(V_i, c_i) \geq \tau_i$, where $(\tau_i, c_i, u_i)$ are independent random vectors that are independent of the $\varepsilon_i$. The $\tau_i$ are left truncation variables and $V_i$ is also subject to right censoring by $c_i$. The case $c_i \equiv \infty$ corresponds to the left-truncated model, for which multiplication of $V_i$ and $\tau_i$ by $-1$ converts it into a right-truncated model. Motivated by a controversy in cosmology involving Hubble's law and chronometric theory, Bhattacharya, Chernoff, and Yang (1983) introduced a Mann–Whitney-type statistic $W_n(\beta) = \sum\sum_{i \neq j} w_{ij}(\beta)$ in the regression model $V_i = \beta u_i + \varepsilon_i$, in which $u_i$ represents log velocity and $V_i$ the negative log of luminosity; moreover, $(u_i, V_i)$ can only be observed if $V_i \leq v_0$. This is a right-truncated model with truncation variables $\tau_i \equiv v_0$, and letting $(V_i^*, u_i^*), i = 1, \ldots, n$, denote the observations, they defined $e_i(\beta) = V_i^* - \beta u_i^*$ and

$$w_{ij}(\beta) = \begin{cases} u_i^* - u_j^* & \text{if } e_j(\beta) < e_i(\beta) \leq v_0 - \beta u_j^* \\ u_j^* - u_i^* & \text{if } e_i(\beta) < e_j(\beta) \leq v_0 - \beta u_i^* \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

since it is impossible to compare $e_i(\beta)$ and $e_j(\beta)$ if $V_i^* - \beta u_j^* > v_0 - \beta u_j^*$ or $V_j^* - \beta u_i^* > v_0 - \beta u_j^*$. Note the similarity of this idea to Eq. (10) proposed by Gehan for censored data, and again it has the same drawbacks as Eq. (10). In fact, as shown by Lai and Ying (1991), what we discussed in the preceding paragraph for censored data can be readily extended to LTRC data $(u_i^*, \tilde{V}_i^*, \delta_i^*), i = 1, \ldots, n$, that are generated from the larger sample consisting of $(V_i, u_i), i = 1, \ldots, m(n) \triangleq \inf\left\{m : \sum_{i=1}^{m} I_{\{\tau_i \leq \min(V_i, c_i)\}} = n\right\}$, with $(\tilde{V}_i, \delta_i)$ observable only when $\tilde{V}_i \geq \tau_i$. The risk set size at $t$ in this case is $Y(t) = \sum_{i=1}^{m(n)} I_{\{\tau_i - \beta u_i \leq t \leq \tilde{V}_i - \beta u_i\}}$ and the nonparametric MLE of the common distribution function $G$ of $\varepsilon_i$ is the product-limit estimator

$$\hat{G}_n(t) = 1 - \prod_{s \leq t}\left(1 - \frac{\Delta N(s)}{Y(s)}\right),$$

where $N(s) = \sum_{i=1}^{m(n)} I_{\{\tau_i - \beta u_i \leq \tilde{V}_i - \beta u_i \leq s, \delta_i = 1\}}$ and $\Delta N(s) = N(s) - N(s-)$ when the value of $\beta$ is specified (e.g., $\beta = 0$ under the null hypothesis). The counting process $N(s)$ plays a fundamental role in the martingale theory underlying the analysis of rank tests via $N(s)$ and $Y(s)$ by Aalen (1978), Gill (1980), and Andersen et al. (1993, Chap. 5) for censored data, and by Lai and Ying (1991; 1992) for LTRC data.

As pointed out in section 1.3, the parametric embedding associated with these regression models is that of a location shift family. Parametric embedding via exponential tilting as in Eq. (9) is associated with another kind of regression models, called hazard regression models, which model how the hazard functions (rather than the means) of $V_i$ vary with the covariates $u_i$. Seminal contributions to this problem were made by Cox (1972), who introduced the model (9) for censored survival data. Kalbfleisch and Prentice (1973) derived the marginal likelihood $L(\theta)$ of the rank vector $R$ given by Eq. (14) for this model:

$$L(\theta) = \prod_{j=1}^{k} \left\{ e^{\theta^{\mathrm{T}} x_{(j)}} \Big/ \left( \sum_{i \in I_j} e^{\theta^{\mathrm{T}} x_{(i)}} \right) \right\}, \tag{19}$$

where $I_j = \{i : \tilde{V}_i \geq \tilde{V}_{(j)}\}$ is the risk set at the ordered uncensored observation $\tilde{V}_{(j)}$, which is the same as that given by Cox using conditional arguments and later by Cox (1975) using partial likelihood. This can be readily extended to LTRC data by redefining the risk set at $\tilde{V}_{(j)}$ as $\{i : \tilde{V}_i \geq \tilde{V}_{(j)} \geq \tau_i\}$. Basically, the regression model in the preceding paragraph considers the residuals $\tilde{V}_i - \beta u_i$, whereas for hazard regression we consider $\tilde{V}_i$ instead.

### 2.3. LAN, least favorable parametric submodels and semiparametric efficiency

The LAN property for the embedded families (exponential tilting of Eq. (6) and location shifts) associated with rank tests for complete data in section 1.3 can be extended to those for LTRC data discussed in the preceding two sections; see Chapter 8 of Andersen et al. (1993) for censored data and Lai and Ying (1992) for LTRC data in the regression setting. For the embedded family (9), the well-known arguments for Cox regression extend readily to LTRC data if the $x_i$ in Eq. (9) are the vector of covariates $u_i$. For the two-sample problem in which $x_i$ depends on the generalized rank vector, we can choose $x_j = \psi\left(\hat{G}_n\left(\tilde{V}^*_{(j)}\right)\right)$ to devise an asymptotically efficient rank statistic as in the censored case, where $\psi = \varphi - \Phi$. The asymptotic efficiency of the rank tests depends on the class of alternatives in the embedded parametric family, which may not contain the actual alternative.

The problem of finding the parametric family that gives the best asymptotic minimax bound has been an active area of research since the seminal paper of Stein (1956) that described a basic idea inherently related to the theme of our article, as follows.

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 there is, through each state of nature, a one-dimensional problem that 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.

The implication of Stein's idea on our parametric embedding theme is the possibility of establishing full asymptotic efficiency of a nonparametric/semiparametric test by using a "least favorable" parametric family of densities for parametric embedding. Lai and Ying (1992, section 2) have shown how this can be done for regression models with i.i.d. additive noise $\varepsilon_i$. The least favorable parametric family has hazard functions of the form $\lambda(t) + \theta\eta(t)$, where $\eta$ is an approximation to $-\lambda'\Gamma_1/\Gamma_0$, $\lambda$ is the hazard function of $\varepsilon_i$, and it is assumed that for $h = 0, 1, 2$, $\Gamma_h(s) = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} E\{u_i^h I_{\{\tau_i - \beta u_i \leq s \leq c_i - \beta u_i\}} / (1 - F(\tau_i - \beta u_i))\}$ exists for every $s$ with $F(s) < 1$, where $F$ is the distribution function of $\varepsilon_i$. In particular, the technical details underlying the approximation are given in (2.26 a, b, c) of that paper. Lai and Ying (1991; 1992) have also shown how these semiparametric information bounds can be attained by using a score function that incorporates adaptive estimation of $\lambda$. For a comprehensive overview of semiparametric efficiency and adaptive estimation in other contexts, see Bickel et al. (1993).

## 3. Applications and discussion

Although "it is customary to treat nonparametric statistical theory as a subject completely different from parametric theory," Stein (1956) developed the least favorable parametric subfamilies for nonparametric testing and estimation as "one of the more obvious connections between the two subjects." Our recent work on a parametric likelihood approach to the analysis of rank tests described in section 1.1 and an ongoing research project on nonparametric tests and estimation of multiple change points have led us to formulate herein a general parametric embedding approach to nonparametric inference problems, focusing on rank data. We have shown how major developments in the theory of rank tests can be related to this approach. In particular, for censored or truncated data, it is relatively straightforward to write down the parametric likelihood ratio, which shows the advantages of using the hazard instead of the joint density function. Section 2 uses parametric embedding via hazard functions to elucidate some landmark developments in rank tests based on LTRC data and to derive optimality results in these complex settings.

Censored and truncated data arise in many applications in astronomy, biomedicine, econometrics and industrial engineering, as pointed out by Lai and Ying (1991; 1992). We conclude with another application of the parametric embedding approach. In environmental studies, it is often of interest to test for trend in monitoring data as for example in the study of lake pH. The usual distribution-free tests based on either the Spearman or Kendall correlation statistics require that data be collected at regularly spaced intervals in time. Alvo and Cabilio (1994) developed a test of trend when data are collected at irregular time intervals. Suppose that observations are taken at $T$ regularly spaced points but recorded only at the $k$ time points $1 \leq t_1 < t_2 < \ldots < t_k \leq T$. Denote the ranks from smallest to largest of the recorded observations by $R(t_i), i = 1, \ldots, k$. Alvo and Cabilio (1994) introduced the test statistic

$$A_k = \left(\frac{T+1}{k+1}\right) \sum_{i=1}^{k} \left(t_i - \frac{T+1}{2}\right)\left(R(t_i) - \frac{k+1}{2}\right), \tag{20}$$

which corresponds to the sample covariance between a string in which time forms a complete ranking $(1, \ldots, T)$ and a string of the ranked observations from an incomplete ranking with $T - k$ blanks and ranks $R = (R(t_1), \ldots, R(t_k))$ at the time points $(t_1, t_2 \ldots, t_k)$.

We now show that the test statistic for the null hypothesis of no trend, $A_k$, may be obtained from the following parametric model. In the complete data situation, let $S = \left(1 - \frac{T+1}{2}, 2 - \frac{T+1}{2}, \ldots, T - \frac{T+1}{2}\right)^{\mathrm{T}}$ be the centered permutation representing time and similarly let $R = \left(R(1) - \frac{T+1}{2}, \ldots, R(T) - \frac{T+1}{2}\right)^{\mathrm{T}}$ be the centered permutation representing the ranked observations. A parametric distance-based model described in Alvo and Yu (2014) for the distribution of $R$ has the form $\pi(R|\theta) = e^{\theta\left(S^{\mathrm{T}}R\right) - K(\theta)}/T!$. Note that the score statistic for this sample is

$$\frac{\partial \log \pi(R|\theta)}{\partial \theta} = S^{\mathrm{T}}R - K'(\theta)$$

evaluated at $\theta = 0$. The asymptotics as $T \to \infty$ lead to the well-known Spearman test of trend. In the situation when data is recorded only at $k$ time points, we may make use of the notion of compatibility as defined in Alvo and Yu (2014) whereby we project the rankings on the class of complete order-preserving permutations. Specifically, if $C(R)$ denotes the compatibility class corresponding to $R$, then

$$E\left[\left(R(i) - \frac{T+1}{2}\right)|C(R)\right] = \left(\frac{T+1}{k+1}\right)\left(R(t_i) - \frac{k+1}{2}\right)\delta(i),$$

where $\delta(i) = 0$ or 1 according to whether the observation at the $i$th regularly spaced time is or is not recorded. This yields the statistic of Eq. (20), which has been shown to be more efficient than the naive test, which ignores missing data. The parametric distance-based model can thus be seen to provide a justification for this trend statistic.

## Funding

This research is supported by Natural Sciences and Engineering Research Council of Canada OGP0009068 and the Research Grants Council of the Hong Kong Special Administrative Region, China (project 17303515). It is also supported by the National Science Foundation, grant DMS-1407828, and supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (project 17303515).

## ORCID

Philip L. H. Yu http://orcid.org/0000-0002-9449-0420

## References

Aalen, O. 1978. Nonparametric estimation of partial transition probabilities in multiple decrement models. *Annals of Statistics* 6 (3):534–45. doi:10.1214/aos/1176344198.

Alvo, M. 2016. Bridging the gap: A likelihood function approach for the analysis of ranking data. *Communications in Statistics—Theory and Methods* 45 (19).

Alvo, M., and P. Cabilio. 1994. Rank test of trend when data are incomplete. *Environmetrics* 5:21–27. doi:10.1002/(ISSN)1099-095X.

Alvo, M., and P. L. H. Yu. 2014. *Statistical methods for ranking data*. Springer.

Andersen, P., O. Borgan, R. Gill, and N. Keiding. 1993. *Statistical models based on counting processes*. New York, NY: Springer.

Bhattacharya, P. K., H. Chernoff, and S. S. Yang. 1983. Nonparametric estimation of the slope of a truncated regression. *Annals of Statistics* 11 (2):505–14. doi:10.1214/aos/1176346157.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. 1993. *Efficient and adaptive estimation for semiparametric models*. Baltimore, MD: Johns Hopkins University Press.

Breslow, N. 1970. A generalized Kruskai-Wallis test for comparing K samples subject to unequal ppattern of censorship. *Biometrika* 57:579–94. doi:10.1093/biomet/57.3.579.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* 34 (2):187–220.

Cox, D. R. 1975. Partial likelihood. *Biometrika* 62 (2):269–76. doi:10.1093/biomet/62.2.269.

Cuzick, J. 1985. Asymptotic properties of censored linear rank tests. *Annals of Statistics* 13 (1):133–41. doi:10.1214/aos/1176346581.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32:675–701. doi:10.1080/01621459.1937.10503522.

Gehan, E. A. 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52 (1/2):203–23. doi:10.1093/biomet/52.1-2.203.

Gill, R. D. 1980. *Censoring and stochastic integrals*. Amsterdam, The Netherlands: Mathematical Centre.

Gu, M. G., T. L. Lai, and K. K. G. Lan. 1991. Rank tests based on censored data and their sequential analogues. *American Journal of Mathematical & Management Sciences* 11 (1–2):147–76.

Hajek, J. 1962. Asymptotically most powerful rank-order tests. *Annals of Mathematical Statistics* 33 (3):1124–47. doi:10.1214/aoms/1177704476.

Hajek, J. 1968. Asymptotic normality of simple linear rank statistics under alternatives. *Annals of Mathematical Statistics* 39:325–46. doi:10.1214/aoms/1177698394.

Hajek, J. 1970. A characterization of limiting distributions of regular estimates. *Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 14:323–30. doi:10.1007/BF00533669.

Hajek, J. 1972. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, ed J. N. L. LeCam and E. Scott, Vol. 1, 175–94. Berkeley, CA: University of California Press.

Hajek, J., Z. Sidak, and P. K. Sen. 1999. *Theory of statistics*, 2nd ed. London, UK: Academic Press.

Hoeffding, W. 1951. Optimum non-parametric tests. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, 83–92. Berkeley, CA: University of California Press.

Kalbfleisch, J. D., and R. L. Prentice. 1973. Marginal likelihoods based on Cox's regression and life model. *Biometrika* 60 (2):267–78. doi:10.1093/biomet/60.2.267.

Lai, T. L., and Z. Ying. 1991. Rank regression methods for left-truncated and right-censored data. *Annals of Statistics* 19 (2):531–56. doi:10.1214/aos/1176348110.

Lai, T. L., and Z. Ying. 1992. Asymptotically efficient estimation in censored and truncated regression models. *Statistica Sinica* 2 (1):17–46.

LeCam, L. 1960. Locally asymptotically normal families of distributions. *University California Publication Statistics* 3:37–98.

LeCam, L. 1972. Limits of experiments. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed J. N. L. LeCam and E. Scott, Vol. 1, 245–61. Berkeley, CA: University of California Press.

Lehmann, E., and C. Stein. 1949. On the theory of some non-parametric hypotheses. *Annals of Mathematical Statistics* 20 (1):28–45. doi:10.1214/aoms/1177730089.

Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 50 (3):163–70.

Neyman, J. 1937. Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift* 20:149–99.

Neyman, J., and E. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Social A* 231:289–337. doi:10.1098/rsta.1933.0009.

Neyman, J., and E. Pearson. 1936, 1938. Contributions to the theory of testing statistical hypotheses. *Statistics Researcher Memory* 1 (1–37):2, 25–57.

Prentice, R. 1978. Linear rank tests with right censored data. *Biometrika* 65:167–79. doi:10.1093/biomet/65.1.167.

Stein, C. 1956. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Vol. 1, 187–95. Berkeley, CA: University of California Press.

Terry, M. 1952. Some rank order tests which are most powerful against specific parametric alternatives. *Annals of Mathematical Statistics* 23 (3):346–66. doi:10.1214/aoms/1177729381.

Van Der Vaart, A. 1998. *Asymptotic statistics*. New York, NY: Cambridge University Press.

Wald, A. 1949. Statistical decision functions. *Annals of Mathematical Statistics* 22 (2):165–205. doi:10.1214/aoms/1177730030.

# Parametric embedding of nonparametric inference problems

Mayer Alvo, Tze Leung Lai, and Philip L. H. Yu

## QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left can be found in the text of the paper for reference. In addition, please review your paper as a whole for correctness.

**Q1:** Au: Please note that the ORCID for Yu Philip L.H. has been created from information provided through CATS. Please correct if this is inaccurate.
**Q2:** Au: As per journal style, reference citation is not allowed in abstract, so first sentence has been rephrased.
**Q3:** Au: Please provide 3 to 5 keywords for the article.
**Q4:** Au: Please provide missing [page number] for [Alvo, 2016].
**Q5:** Au: Please provide missing [Publisher location] for [Alvo and Yu, 2014].
**Q6:** Au: please confirm publication year for Neyman and Pearson reference



## TABLE OF CONTENTS LISTING

The table of contents for the journal will list your paper exactly as it appears below: