Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 332–345
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Permutation procedures with censored data

Johan Lim*

Department of Statistics, Texas A&M University, RM 429, 3143 TAMU, College Station 778433143, USA

Received 30 April 2003; received in revised form 9 September 2004; accepted 9 September 2004
Available online 2 October 2004

Abstract

This paper extends the permutation procedures for truncated data in Diaconis et al. (http://www-stat.stanford.edu/~susan/.) to doubly censored data. As in Diaconis et al. (http://www-stat.stanford.edu/~susan/.), the proposed procedure is based on samples from the conditional distribution of rank statistics which is uniformly distributed on a set of permutations. Subsequently, our procedure is applied to testing independence with bivariate censored data and estimating a regression coefficient with doubly censored data. Also, when estimating a regression coefficient with doubly censored data, simulation studies show that the proposed procedure is superior to that of Akritas et al. (J. Amer. Statist. Assoc. 90 (1995) 170).

Keywords: Permutation; Bivariate Censored Data; Kendalls' tau statistics; Testing independence; Theil–Sen estimator

1. Introduction

Statistical procedures based on rank statistics have been popular because of their robustness to distributional assumptions. Among many rank based procedures, Kendall's tau (KT) statistics has been widely used as a measure for sample correlation. Two most important applications of KT statistics are testing independence and estimating the regression coefficient between two samples. For observations $\{(X_i, Y_i)\}_{i=1}^n$, the KT statistics is defined as $\tau = 2S/n(n-1)$, where S is the difference between the number of concordant pairs and that of disconcordant pairs. Since τ is asymptotically normally distributed with a mean of 0 and the variance of $(4n+10)/(9 \cdot n(n-1))$ (Kendall, 1948), the independence between X and Y can be tested by test statistics $t = \tau/\sqrt{\text{var}(\tau)}$. The KT statistics can also be used

* Tel.: +1 979 8621576; fax: +1 979 8453144.
E-mail address: johanlim@stat.tamu.edu (J. Lim).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.09.002

for estimating a regression coefficient. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are from $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$. Then, one important regression assumption is the independence between X_i and $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$, or, equivalently, the independence between X_i and $Y_i - \beta_1 X_i$. Based on the observation, Theil (1950) and Sen (1968) propose to estimate β_1 using β which makes KT statistics between X_i and $Y_i - \beta X_i$ as 0; we denote this estimator as the Theil–Sen estimator.

In practice, however, incompleteness in data (e.g. truncation, censoring, missing) produces bias or inefficiency in existing procedures. Among various types of incompleteness, this paper focuses on doubly censored data. Here, doubly censoring means censoring occurs in both X and Y. Several modifications of the KT statistics have been proposed in the previous literature. First, Oakes (1982) proposed to compute KT statistics only based on known (dis)con-cordance pairs and derived its' asymptotic properties. Akritas et al. (1995) proposed a modification of Theil–Sen slope estimator, where the KT statistics between residuals and covariates was computed using only known (dis)con-cordance pairs as in Oakes (1982). In contrast, Weier and Basu (1980) proposed to use the information from unknown pairs relying on the Kaplan–Meier estimator when testing independence.

It has been known that, when X and Y are independent, the KT statistics is equivalent to the graphical distance of the observed point from the center in a suitable permutation space (Diaconis et al., 1999). Accordingly, the independence can be tested by examining how much the KT statistics of the observations are deviate from the center in permutation space. Like other procedures, some modifications for incomplete data (mostly for truncated data) have been proposed in literature. In particular, when data is truncated, the permutation space is restricted and Diaconis et al. (1999) discussed the general frameworks for permutation procedures with restricted position.

Censoring is quite different from truncation in the sense that censored subjects are "observed" and have some information on the underlying distribution. Accordingly, censoring produces uncertainty in the KT statistics of the observation and, equivalently, produces the restriction on the permutation space. It is worth noting that the truncated subjects are "not observed" and also result in the restriction of the permutation space.

This paper extends the permutation procedures in Diaconis et al. (1999) to doubly censored data. In Section 2, we prove that the KT statistics are distributed uniformly on a specific set when the data is doubly censored. Subsequently, procedures are proposed for generating samples from the uniform distribution. In Section 3, we apply the proposed procedure in Section 2 to testing independence between doubly censored pairs. In Section 4, we discuss the estimation of a regression coefficient with doubly censored data. Simulation studies show the superiority of the permutation procedure to the modified estimator by Akritas et al. (1995). Finally, Section 5 concludes the paper. Random censoring for both X and Y is assumed in testing independence, whereas the conditional independence between censoring distribution and uncensored Y, given uncensored X, is assumed in doubly censored regression.

2. Permutation procedures for censored data

In this section, we introduce the notations which will be used in the remainder of the paper and propose a general permutation procedure for the censored data.

2.1. Notations

Suppose the following n-pairs of data are observed:

$$\{(x_1, u_1), (y_1, v_1)\}, \{(x_2, u_2), (y_2, v_2)\}, \ldots, \{(x_n, u_n), (y_n, v_n)\},$$

where $\{u_i\}_{i=1}^n$ and $\{v_i\}_{i=1}^n$ takes 0 or 1 depending on whether the corresponding component is censored or not. Then, the following notations will be used in the remainder of the paper:

- The set $\{1, 2, \ldots, n\}$ is denoted by $[n]$.
- $\pi = (\pi(1), \ldots, \pi(n))$ is a permutation of $[n] = \{1, 2, \ldots, n\}$, where $\pi(i)$ is the label at position i of the permutation π.
- $\pi_0 = (\pi_0(1), \ldots, \pi_0(n)) = (1, \ldots, n)$.
- The graphical distance (GD) between two permutations π_1 and π_2, denoted by $GD(\pi_1, \pi_2)$, is defined as the minimum number of pairwise adjacent transpositions required to bring π_1 to π_2. For example, the graphical distance between $\pi_1 = (1, 4, 3, 2)$ and $\pi_2 = (1, 2, 3, 4)$ is 2, since sequential transpositions of 4 and 3 and that of 4 and 2 in π_1 gives π_2; there is no way to reach from π_1 to π_2 with one transposition. In higher dimension, the distance is often hard to evaluate, but it is well defined distance in a permutation space.
- Let S_n be the set of all possible permutations on $\{1, 2, \ldots, n\}$ equipped with the graphical distance $GD(\cdot, \cdot)$.

Beyond the above conventional notations from combinatorics, we define more notations for this paper. Let

$$(X, U) = \{(x_1, u_1), \ldots, (x_n, u_n)\} \quad \text{and} \quad (Y, V) = \{(y_1, v_1), (y_2, v_2), \ldots, (y_n, v_n)\}.$$

- Let $R_X(k)$ be the all possible ranks of X_k in the permutation and $R_X = (R_X(1), R_X(2), \ldots, R_X(n))$. $R_Y(k)$ and R_Y are similarly defined.
- The *compatible* set to the sequence (X, U) and (Y, V), say A_{XY}, is defined to be a class of permutation pairs (π_X, π_Y) satisfying $\pi_X(k) \in R_X(k)$ and $\pi_Y(k) \in R_Y(k)$. For example, when

$$(X, U) = \{(1, 0), (3, 1), (4, 1)\} \quad \text{and} \quad (Y, V) = \{(2, 1), (3, 0), (4, 1)\}$$

and the right censoring is assumed ($u_i = 0$ implies true observation is larger than x_i),

$$\begin{aligned} R_X &= \{(1, 2, 3), (2, 1, 3), (3, 1, 2)\}, \\ R_Y &= \{(1, 2, 3), (1, 3, 2)\}. \end{aligned} \tag{1}$$

The compatible set A_{XY} becomes the set of

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \ldots, \begin{pmatrix} 3 & 1 & 2 \\ 1 & 3 & 2 \end{pmatrix}.$$

- The projection $\Sigma(\pi_X, \pi_Y)$ (from $S_n \times S_n$ to S_n) maps the permutation pair $(\pi_X, \pi_Y) \in S_n \times S_n$ representing

$$\begin{pmatrix} \pi_X(1) & \pi_X(2) & \pi_X(3) & \cdots & \pi_X(n) \\ \pi_Y(1) & \pi_Y(2) & \pi_Y(3) & \cdots & \pi_Y(n) \end{pmatrix}$$

  to an element in S_n whose kth component is $\pi_Y(\pi_X^{-1}(k))$. For example, the projection map Σ maps both

$$\begin{pmatrix} 1 & 3 & 2 \\ 1 & 2 & 3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 3 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix}$$

  to (1, 3, 2).
- Given the observation (X, U), let I be an $n \times n$ zero–one matrix, where 1 in the (i, j)th component indicates that the ith subject can be matched (transposed) to the jth subject. Further, let S_I be the set of all permitted permutations corresponding to a zero–one matrix I. For example, (Y, V) in (1) has the incidence matrix

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

  where 1 in (2, 2) represents the second component can have rank 2.
- Let the set of permutations A be connected if any pair of two permutations in A can be accessible to each other using transposition operation.

2.2. General procedures

Both testing independence and estimating a regression coefficient with doubly censored data heavily depend on the KT statistics, which is equivalent to the graphical distance (GD) between a permutation and the pre-determined center π_0 (Diaconis et al., 1999). Hence, in a censored pair of permutations (π_X, π_Y), the graphical distance between $\Sigma(A_{XY})$ and π_0, say $GD(\Sigma(A_{XY}), \pi_0)$, is of interest in this paper. When no censoring occurs, the uncertainty does not appear in the compatible set A_{XY} and $GD(\Sigma(A_{XY}), \pi_0)$.

The difficulties in analyzing censored data stem from complexity of the conditional distribution of $GD(\Sigma(A_{XY}), \pi_0)$ (given the observations), which cannot be expressed as a simple formula. However, it can be shown that $P((\pi_X, \pi_Y)|(\underline{X}, U), (Y, V))$ is distributed uniformly on A_{XY}. Accordingly, to generate a sample from the distribution of $GD(\Sigma(A_{XY}), \pi_0)$, it suffices to obtain a sample from the uniform distribution on A_{XY}.

Theorem 2.1. *Suppose there exist measures μ_X and μ_Y such that*

$$P(X_i \in A_i, Y_i \in B_i, 1 \leqslant i \leqslant n) = \prod_{i=1}^{n} \mu_X(A_i) \cdot \mu_Y(B_i). \tag{2}$$

Then, any pair of permutations π_1 and π_2 in R_X satisfies

$$\begin{aligned} &P\left(X_{\pi_1(1)} \leqslant X_{\pi_1(2)} \leqslant \cdots \leqslant X_{\pi_1(n)} | (X, U)\right) \\ &\quad = P\left(X_{\pi_2(1)} \leqslant X_{\pi_2(2)} \leqslant \cdots \leqslant X_{\pi_2(n)} | (X, U)\right). \end{aligned} \tag{3}$$

Also the same is true for any permutation pair (π_1, π_2) in R_Y.

Proof. Lemma 3.3 in Diaconis et al. (1999) proved that the graph G—having a vertex set S_I and the edges between σ and τ (where σ and τ differ by a transposition of label)—is connected if the ones in each row of I (zero–one restriction matrix) lie in an interval. Hence, it suffices to show that Eq. (2) holds for the pair π_1 and π_2, whose graphical distance is one (i.e. for some $i, j \in [n]$, $(\pi_1(i), \pi_1(j)) = (\pi_2(j), \pi_2(j))$ and $\pi_1(k) = \pi_2(k)$ for every other k). Events $E_{-\{i,j\}}(\pi)$, $E_{i,j}(\pi)$, and $E_{j,i}(\pi)$ are defined as $\{X_{\pi(1)} \leqslant \cdots \leqslant X_{\pi(i-1)} \leqslant X_{\pi(i+1)} \leqslant \cdots \leqslant X_{\pi(j-1)} \leqslant X_{\pi(j+1)} \leqslant X_{\pi(n)}\}$, $\{X_{\pi(i-1)} \leqslant X_{\pi(i)} \leqslant X_{\pi(i+1)}$ and $X_{\pi(j-1)} \leqslant X_{\pi(j)} \leqslant X_{\pi(j+1)}\}$ and $\{X_{\pi(i-1)} \leqslant X_{\pi(j)} \leqslant X_{\pi(i+1)}$ and $X_{\pi(j-1)} \leqslant X_{\pi(i)} \leqslant X_{\pi(j+1)}\}$,respectively. Then, the left-hand side of Eq. (2) is

$$\begin{aligned}
(LHS) &= P\left(X_{\pi_1(1)} \leqslant X_{\pi_1(2)} \leqslant \cdots \leqslant X_{\pi_1(n)} | (X, U)\right) \\
&= P\left(E_{i,j}(\pi_1) | E_{-\{i,j\}}(\pi_1), (X, U)\right) \cdot P\left(E_{-\{i,j\}}(\pi_1) | (X, U)\right) \\
&= P\left(E_{j,i}(\pi_1) | E_{-\{i,j\}}(\pi_1), (X, U)\right) \cdot P\left(E_{-\{i,j\}}(\pi_1) | (X, U)\right) \qquad (4) \\
&= P\left(E_{i,j}(\pi_2) | E_{-\{i,j\}}(\pi_2), (X, U)\right) \cdot P\left(E_{-\{i,j\}}(\pi_2) | (X, U)\right) \qquad (5) \\
&= P\left(X_{\pi_2(1)} \leqslant X_{\pi_2(2)} \leqslant \cdots \leqslant X_{\pi_2(n)} | (X, U)\right) = (RHS). \qquad \square
\end{aligned}$$

Generating a pair of permutation from the $P(\Sigma(A_{XY})|(X, U), (Y, V))$ can be achieved by sequentially sampling from uniform distributions $P(\pi_X|(X, U))$ and $P(\pi_Y|(Y, V))$. As in the remarks for Lemma 3.3 in Diaconis et al. (1999), when the set S_I is connected, we can run a Markov chain on S_I having a uniform stationary distribution on S_I. The Markov chain can be described as follows: (i) from $\sigma \in S_I$, choose one of $\binom{n}{2}$ transpositions uniformly at random and transform σ by switching those two chosen labels and (ii) the chain moves the new permutation if it is in S_I and it stays in the current position otherwise. When I has one-sided restrictions, the procedure is much simpler. Let S_b be a set of permutations π satisfying $\pi(i) \geqslant b_i$ for all $i \in [n]$. Without loss of generality, $b_1 \leqslant b_2 \leqslant \cdots \leqslant b_n$ is assumed. The uniform choice from S_b could be achieved by

1. choosing $\pi(1)$ uniformly from $J_1 = \{j : j \geqslant b_1\}$
2. choosing $\pi(2)$ from the set $J_2 = \{j : j \geqslant b_2\} - \{\pi(1)\}$, and
3. repeat (i) and (ii).

3. Testing independence with bivariate censored data

One simple random effects model is matched pair data model, where individuals in the ith pair share a common random effect (heterogeneity). This common random effect yields a positive correlation between the individuals in each pair; hence, testing the existence of random effects can be implemented by testing the independence between pairs.

The KT statistics has provided a simple test for independence in a bivariate distribution. Furthermore, as pointed out in the Introduction, much research has been done on modifying the KT statistics to adapting censored data. In this section, the proposed permutation procedure is applied to testing independence as a potential remedy for existing conventional procedures (Weier and Basu, 1980; Oakes, 1982). Three data sets—the leukemia remission times data from Oakes (1982), the data on times to tumor occurrences from

Mantel and Ciminera (1979), and the kidney patients data in McGilchrist and Aisbett (1991)—are analyzed. In all three examples, only right censoring occurs and it results in one-sided restrictions in permutations.

3.1. Example in Oakes (1982)

The leukemia remission times data in Oakes (1982) was analyzed. As in Oakes (1982), it was assumed that death at a given time always precedes censoring at the same time and other ties were broken down randomly. In the permutation procedure, $N = 2000$ samples were generated from the uniform distribution on the compatible set. From each permutation, the KT statistics was evaluated. The mean and the variance of the sampled KT statistics were -0.0690 and 0.0072, respectively. The permutation based estimator is given by $\widehat{\tau} = \sum_{i=1}^{N} \tau_i / N$ and the variance of $\widehat{\tau}$ can be approximated to the sum of within variation, $(4n + 1)/(9n(n - 1))$, and between variation, $\sum_{i=1}^{N} (\tau_i - \widehat{\tau})^2/(N - 1)$, as in Xie and Paik (1997). Accordingly, $\widehat{\tau} = -0.0690$ and $\mathrm{var}(\widehat{\tau}) = 0.3206$. The corresponding t-value was -0.3853, which indicated no evidence against the independence assumption. In Oakes (1982), the standardized statistics was -0.84 and the same conclusion was reached.

3.2. Example in Clayton (1991)

The example in this section is the time to tumor occurrence as described in Mantel and Ciminera (1979) and Clayton (1991). Three rats were from each of 50 litters where one rat was treated with putative carcinogen and the other two served as control cases. The time to tumor occurrence or censoring was recorded to the nearest week. In the analysis, the first (the exposed) and the third column (the control) of Table 1 in Clayton (1991) were used. In the permutation procedure, $N = 10\,000$ samples were generated from the uniform distribution on the compatible set. In each permutation sample, the KT statistics was evaluated. The mean and the variance of sampled τs were 0.0756 and 0.0065, respectively. According to the asymptotic results, $\widehat{\tau} = 0.0756$, $\mathrm{var}(\widehat{\tau}) = 0.01606$, and t-value was 0.5972. Hence, it can be concluded that there was no positive correlation between subjects in a pair, which implies that random effects may not be present in the model. This was also apparent from the negligible difference between the maximum partial likelihood estimator $\left(\widehat{\beta}^{\mathrm{Cox}} = 0.907\right)$ and the estimator from the frailty model $\left(\widehat{\beta}^{\mathrm{Clayton}} = 0.919\right)$. It should be noted that Oakes' procedure cannot be applicable to this data because of heavy censoring as pointed out in Oakes (1982).

3.3. Example in McGilchrist and Aisbett (1991)

The last example is times to infection for kidney patients using portable dialyze. For each patient two such recurrence times were observed. McGilchrist and Aisbett (1991) assumed that the only correlation between recurrence intervals for failures was due to common random effects, called frailties. For more explanation on data, see McGilchrist and Aisbett (1991). The data showed a large difference between two recurrence times of some individuals, and the assumption of positive correlation was quite spurious. Since

Table 1
Comparison with Theil–Sen estimator with $n = 50$, $n = 30$, and $n = 20$

($n = 50$)		Parallel		Type I	
		Perm.	T–S	Perm.	T–S
Nocensored X	mean (std)	1.5039 (0.0422)	1.4971 (0.0355)	1.4635 (0.0421)	1.3910 (0.0509)
	$\sqrt{MSE}$	0.04229	0.0355	0.0556	0.1202
Censored X	mean (std)	1.5076 (0.0384)	1.5001 (0.0369)	1.4655 (0.0470)	1.3935 (0.0543)
	$\sqrt{MSE}$	0.0390	0.0368	0.0582	0.1195
($n = 30$)					
Nocensored X	mean (std)	1.4950 (0.0410)	1.4959 (0.0496)	1.4938 (0.0497)	1.2897 (0.0896)
	$\sqrt{MSE}$	0.0411	0.0495	0.0499	0.2284
Censored X	mean (std)	1.4962 (0.0460)	1.4963 (0.0543)	1.4611 (0.0612)	1.3873 (0.0670)
	$\sqrt{MSE}$	0.0459	0.0542	0.0722	0.1310
($n = 20$)					
Nocensored X	mean (std)	1.4899 (0.0655)	1.4942 (0.0635)	1.4947 (0.0647)	1.2710 (0.1066)
	$\sqrt{MSE}$	0.0659	0.0634	0.0646	0.2523
Censored X	mean (std)	1.4930 (0.0789)	1.4971 (0.0681)	1.4581 (0.0867)	1.3842 (0.0842)
	$\sqrt{MSE}$	0.0788	0.0678	0.0959	0.1429

"Perm." and "T–S" represents our permutation procedure and Theil–Sen estimator, respectively.

each pair also shared other common factors causing strong dependency such as age and gender (besides the common heterogeneity), the rejection of independence assumption could provide a strong basis against the common frailty model. The permutation procedure was applied with $N = 2000$ samples from the uniform distribution on $\Sigma(\pi_X, \pi_Y)$. As in other examples, the KT statistics was evaluated for each sample. The mean and the variance of sampled τs were 0.17943 and 0.00391, respectively. Accordingly, $\widehat{\tau} = 0.1794$, $\text{var}(\widehat{\tau}) = 0.0167$, and $t = 1.38796$ (p-value$=0.1651$). Here, the result implied that the common frailty model may not be appropriate. On the other hand, Oakes' procedure yielded $\widehat{\tau} = 0.0448$ and $\text{var}(\widehat{\tau}) = 0.00482$, and the standardized statistics was 0.6455. Hence, the results were consistent with those obtained by Oakes' procedure.

4. Theil–Sen estimator with doubly censored data

In this section, we apply the permutation procedure to the regression with doubly censored data. Simulation studies are implemented to compare the proposed procedures with the modified Theil–Sen estimator by Akritas et al. (1995). Hereafter, we denote the above modified Theil–Sen estimator as the Theil–Sen estimator for notational simplicity.

Suppose $\{X_i, Y_i\}_{i=1}^{\infty}$ are observed from the model, $Y_i = \beta \cdot X_i + \varepsilon_i$, where ε_i are independently and identically distributed with a mean of 0 and a variance of σ^2. To estimate regression coefficients, Theil (1950) and Sen (1968) proposed the Theil–Sen estimator defined as the value of b that makes the KT statistics between the residual and the covariate

be 0. To be specific,

$$\widehat{b}^{\mathrm{TS}} = \left\{ b : \sum_{i<j} \left[I(X_i < X_j) - I(X_j < X_i) \right] \times \left[I\left(r_i(b) < r_j(b)\right) - I\left(r_j(b) < r_j(b)\right) \right] = 0 \right\},$$

where $r_i(b) = y_i - b \cdot x_i$, for every i. The Theil–Sen estimator can also be interpreted as the median of slopes $(y_i - y_j)/(x_i - x_j)$. Akritas et al. (1995) proposed a modification of the Theil–Sen estimator for doubly censored data, which is defined as the solution b of the equation

$$T_n(b) = \sum_{i<j} \delta_i^x \cdot \delta_j^x \cdot \left[I(X_i < X_j) - I(X_j < X_i) \right] \times \left[\delta_i^y I\left(r_i(b) < r_j(b)\right) - \delta_j^y I\left(r_j(b) < r_j(b)\right) \right] = 0.$$

In this paper, $r_i(b)$ is written as r_i for simplicity.

The permutation procedure in Section 2 can also be applied to estimating the regression coefficient with doubly censored data. Here, the residuals could be right censored, left censored or both and the permutation procedures for interval restrictions are employed.

Suppose $R = \{r_i\}_{i=1}^n$ is ordered as follows: (i) the interval-censored observations appear first, the left- and the right-censored observations are in the middle, then the uncensored follow in the last position, (ii) among left-censored (or right-censored) observations, the smaller (or larger) $r_i(b)$s come earlier. For example, when observing the following data (with $b = 1$)

$$(0.5, 2), (1^{\mathrm{R}}, 2), (1.5, 2^{\mathrm{R}}), (1.5^{\mathrm{R}}, 2), (0.5^{\mathrm{R}}, 1.5^{\mathrm{R}}),$$

their residuals $(1.5, 1^{\mathrm{L}}, 0.5^{\mathrm{R}}, 0.5^{\mathrm{L}}, 1^{\mathrm{B}})$ are ordered as $(1^{\mathrm{B}}, 0.5^{\mathrm{L}}, 1^{\mathrm{L}}, 0.5^{\mathrm{R}}, 1.5)$. Here, L, R and B represent the censoring direction. Let A_{XR} be the compatible space corresponding to $\{r_i(b)\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$. As in Section 3, $J(r_i)$ is all possible ranks of r_i compatible with the censored observations. Under the assumption that it cannot be stopped, the following algorithm results in a uniform choice from A_{XR}. First, choose r_1^* uniformly from $J(r_1)$. Subsequently, choose r_k^* uniformly from $J(r_k) - \{r_1^*, r_2^*, \ldots, r_{k-1}^*\}$. After sampling the ranks of censored observations, fill in the remainder of ranks with uncensored observations. Finally, repeat this procedures with respect to $X = \{X_k\}_{k=1}^n$. The proof of the above algorithm is similar to that of Lemma 3.3 in Diaconis et al. (1999). If there is a chance that the proposed procedure stops, the Markov chain method introduced in Section 2.2 can be used.

In next section, we implement simulation studies to show the superiority of the proposed procedure to the modified Theil–Sen estimator by Akritas et al. (1995) in finite sample. Subsequently, the permutation procedure is applied to analyze the astronomical data from Heckman et al. (1989).

4.1. Simulation study

In this section, we implemented several simulation studies to investigate the performance of the proposed procedure. In the following simulation, we restricted our interest into the cases considered in Akritas et al. (1995); parallel and Type I censoring with one-sided restrictions (right or left censoring). In our rank based procedure, right censoring and left censoring were symmetric to each other along with the transformation $Y' = -Y$. Hence, we only consider right censoring in this paper.

Simulations were implemented with small sample sizes ($n = 20$, 30, and 50) and with a considerable amount of censoring in both X and Y. As in Akritas et al. (1995), X^{T}s were from the exponential distribution and the Y^{T}s were from the linear regression model with a Gaussian error. Four different cases were considered depending on (i) parallel censoring or Type-I censoring and (ii) the covariates are censored or not.

The true covariates X_i^{T}s were generated from the exponential distribution with a mean of 10, and the true response Y_i^{T}s were from $Y_i^{\mathrm{T}} = \beta_1 \cdot X_i^{\mathrm{T}} + N(0, \sigma^2)$. On the other hand, the censoring covariates X_i^{C}s were from the exponential distribution with a mean of 2, and the censoring responses Y_i^{C}s were from $Y_i^{\mathrm{C}} = \beta_0 + \beta_1 \cdot X_i^{\mathrm{T}} + N(0, 3)$. First, in Type I censoring with uncensored covariates, the minimum between Y_i^{T} and 10 was recorded. Second, in Type I censoring with censored covariates, the minimum between Y_i^{T} and 5 was recorded and the observed covariate was the maximum of X_i^{T} and X_i^{C}. Third, in parallel censoring with uncensored covariates, β_0 was chosen as -0.5 and the maximum between Y_i^{T} and Y_i^{C} was recorded. Finally, in parallel censoring with censored covariates, β_0 was chosen as -0.2 and the maximum between Y_i^{T} and Y_i^{C} was recorded, where the observed covariates were same with that in Type-I censoring.

First, to investigate small sample properties of the proposed estimator, we set $\beta_1 = 1.5$ and $\sigma^2 = 3$, and simulated 100 data sets with $n = 20$, 30, and 50 for the above four cases. When X was not censored, the censoring rates of Y was approximately 50%; when X was censored, rates of X and Y was approximately 20% and 35% for X and Y, respectively. In Type I censoring, our procedure was superior to the Theil–Sen estimator in terms of both bias and the mean squared error (MSE). In parallel censoring, both the permutation procedure and the Theil–Sen procedure performed quite well. The results were reported in Table 1. We also presented the density plots of the estimates by both methods in Fig. 1.

Second, we implemented a simulation study to investigate the performance of our procedure under heterogeneity and to compare it with the Theil–Sen estimator. We generated data sets with $n = 50$. The first 25 samples in each data set were generated from the model with Gaussian noise $N(0, \sigma_1^2)$, whereas the second 25 samples were from the model with Gaussian noise $N(0, \sigma_2^2)$ with $\sigma_1^2 < \sigma_2^2$. We considered three different cases depending on $R = \sigma_2^2/\sigma_1^2$ ($R = 1$, 2, and 3) and, in each case, 100 data sets were generated. In parallel censoring type, our procedures were performed as well as the Theil–Sen estimator did. In Type-I censoring, the superiority of the permutation procedure to the Theil–Sen estimator did not hurt by heterogeneity (Table 2).

Third, we implemented a simulation study for different values of regression slopes $\beta_1 = 0.5$, 1.5, and 2.5. Table 3 showed that, in parallel censoring, both the permutation based estimator and the Theil–Sen estimator performed well regardless of the magni-

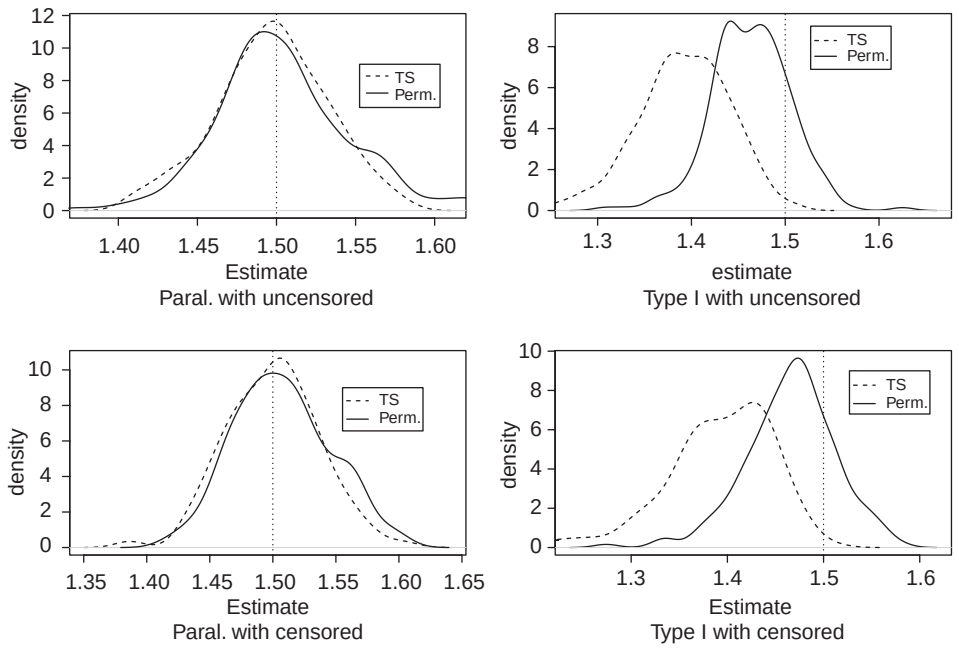

Fig. 1. Density plot of the estimates from two different procedures; the proposed permutation procedure and the modified Theil–Sen procedure by Akritas et al. (1995).

Table 2
The results are based on 100 data sets with $n = 50$. The first 25 observations have Gaussian errors with a variance of σ_1^2 and the second half observations have that of σ_2^2. Ratio is the ratio of those standard deviations σ_2^2/σ_1^2

		Parallel		Type I	
		Perm.	T–S	Perm.	T–S
Ratio = 1	mean (std)	1.5076 (0.0384)	1.4970 (0.0368)	1.4655 (0.0470)	1.3935 (0.0543)
	$\sqrt{MSE}$	0.0390	0.0368	0.0582	0.1195
Ratio = 2	mean (std)	1.5034 (0.0357)	1.5013 (0.0348)	1.4653 (0.0446)	1.3895 (0.0524)
	$\sqrt{MSE}$	0.0355	0.0346	0.0563	0.1221
Ratio = 3	mean (std)	1.4986 (0.0624)	1.5039 (0.0466)	1.4507 (0.0592)	1.3500 (0.0695)
	$\sqrt{MSE}$	0.0621	0.0465	0.0768	0.1652

"Perm." and "T–S" represents our permutation procedure and Theil–Sen estimator, respectively.

tude of the regression slope. However, in Type I censoring, our estimator performed better than the Theil–Sen estimator in terms of both bias and MSE. In particular, the Theil–Sen estimate was significantly downward biased in Type I censoring when β was small ($\beta = 0.5$ or 1.5).

Table 3
Comparison with Theil–Sen estimator for different slope values, $\beta = 0.5$, $\beta = 1.5$, and $\beta = 2.5$

$\beta = 0.5$		Parallel		Type I	
		Perm.	T–S	Perm.	T–S
Nocensored X	mean (std)	0.4995 (0.0180)	0.5006 (0.0178)	0.4812 (0.0398)	0.1622 (0.0504)
	$\sqrt{MSE}$	0.0180	0.0178	0.0439	0.3415
Censored X	mean (std)	0.4992 (0.0207)	0.4997 (0.0200)	0.3261 (0.0537)	0.2154 (0.0584)
	$\sqrt{MSE}$	0.0207	0.0200	0.1819	0.2905
$\beta = 1.5$					
Nocensored X	mean (std)	1.5039 (0.0422)	1.4971 (0.0355)	1.4635 (0.0421)	1.3910 (0.0509)
	$\sqrt{MSE}$	0.04229	0.0355	0.0556	0.1202
Censored X	mean (std)	1.5076 (0.0384)	1.5001 (0.0369)	1.4655 (0.0470)	1.3935 (0.0543)
	$\sqrt{MSE}$	0.0390	0.0368	0.0582	0.1195
$\beta = 2.5$					
Nocensored X	mean (std)	2.5014 (0.0181)	2.5044 (0.0192)	2.5008 (0.0188)	2.4062 (0.0536)
	$\sqrt{MSE}$	0.0181	0.0197	0.0187	0.1079
Censored X	mean (std)	2.5082 (0.0233)	2.5042 (0.0183)	2.4980 (0.0261)	2.4672 (0.0305)
	$\sqrt{MSE}$	0.0247	0.0188	0.0260	0.0447

"Perm." and "T–S" represents our permutation procedure and Theil–Sen estimator, respectively. Each values are based on 100 simulated data sets.

Finally, we discuss the confidence interval (C.I). For the Theil–Sen estimation with complete data, Sen (1968) proposed the $100 \cdot (1-\alpha)\%$ C.I. as (β_L^*, β_U^*), where

$$\beta_L^* = \inf\left\{b : \tau_n(b) \geqslant -z_{\alpha/2}\sqrt{(4n+10)/(9 \cdot n(n-1))}\right\} \tag{6}$$

$$\beta_U^* = \sup\left\{b : \tau_n(b) \geqslant z_{\alpha/2}\sqrt{(4n+10)/(9 \cdot n(n-1))}\right\}, \tag{7}$$

and $\tau_n(b) = 2 \cdot T_n(b)/n \cdot (n-1)$. Here, the variability associated with the permutation estimate, say $\overline{\tau_n}(b)$, is the sum of two components:

$$\mathrm{var}(\overline{\tau_n}(b_0)|Obs.) \approx \frac{4n+10}{9 \cdot n(n-1)} + \mathrm{var}(\tau_n(b)|Obs.), \tag{8}$$

where the first term is the contribution of the random data itself and the second term is the contribution of the censoring, roughly. The above variance formula of the estimator based on Monte Carlo samples is not new from this paper, but has often been used in the literature of multiple imputation (see p. 257 in Little and Rubin, 1987; Xie and Paik, 1997). In practice, $\mathrm{var}(\tau_n(b_0)|Obs.)$ is estimated using the samples from the conditional distribution of ranks given the observed values. Accordingly, we propose to approximate the $100 \times (1-\alpha)$ C.I. to

$$\beta_L^* = \inf\left\{b : \tau_n(b) \geqslant -z_{\alpha/2}\sqrt{(4n+10)/(9 \cdot n(n-1)) + \mathrm{var}(\tau_n(b)|Obs.)}\right\} \tag{9}$$

$$\beta_U^* = \sup\left\{b : \tau_n(b) \geqslant z_{\alpha/2}\sqrt{(4n+10)/(9 \cdot n(n-1)) + \mathrm{var}(\tau_n(b)|Obs.)}\right\}. \tag{10}$$

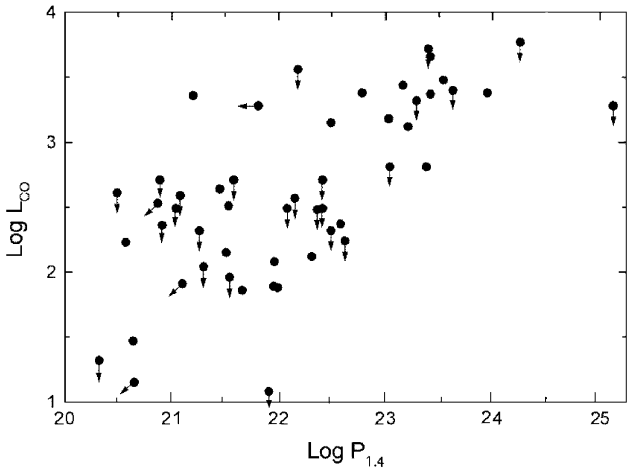

Fig. 2. The X-axis is log $P_{1.4}$, which is the radio power at 1.4 GHz in units of W/Hz and the Y-axis is log L_{CO}, the luminosity of carbon monoxide emission in kps^2 km/s.

To investigate the accuracy of the proposed C.I., we computed the coverage probabilities for the cases of $\beta = 1.5$ and $n = 50$ in Table 1. In each case, 1000 data sets with $n = 50$ were generated and the coverage probabilities of 95% C.I. were computed. For the case of Type I censoring with uncensored covariate, the coverage probability was 99.5%; for the case of Type I censoring with censored covariate, it was 96.2%; for the case of parallel censoring with uncensored covariate, it was 98.4%; finally, for the case of parallel censoring with censored covariate, it was 99.3%. In overall, the proposed C.I. did not provide an accurate coverage probability except a few cases. It may be because the proposed C.I. again relied on large sample theory and can be inaccurate with small number of samples.

4.2. Example in astronomical data

In this section, the permutation procedure was applied to the astronomical data used for constructing Table 2 in Heckman et al. (1989). They were interested in whether or not the radio emission produced by Seyfert galaxies was related to its' rate of star formulation as in normal spiral galaxies. As explained in Akritas et al. (1995), the star formulation rate scales with the quantity of dense molecular gas in the galaxy, which is measured by the line emission of carbon monoxide molecule. Thus, they regressed the radio power on the luminosity of carbon monoxide emission (see Fig. 2).

We computed the regression slope for the Seyfert galaxy sample of 52 observations of which 23 were uncensored; 25 were censored in log L_{CO}; 1 was censored in log $P_{1,4}$; and 3 were censored in both variables. As in Akritas et al. (1995), the permutation procedure was applied to log-transformed data and obtained the following estimates (for more detail on other estimators, see Akritas et al., 1995):

Laan median of pairwise slopes 0.5553.

Laan weighted least squares slope 0.6075.

Nearest-neighbor median of pairwise slopes 0.5179.
Nearest-neighbor weighted least squares slope 0.5970.
Theil–Sen slope 0.5176.
Permutation estimate 0.4963.

The estimated slope was slightly lower than that of the modified Theil–Sen estimator. The proposed permutation procedure may use more information on the censored pairs than the Theil–Sen estimator relying only known (dis)concordant pairs. In this example, most of response variables were left censored (toward to 0). Accordingly, if the censoring information was not fully considered in the analysis, the slope would be overestimated. In this sense, it was not surprising that the estimated slope which was smaller than that of the Theil–Sen estimator. According to Eq. (6) and (7), the 95% C.I. was approximately (0.2045, 0.7615) for this example. As pointed out in Akritas et al. (1995), these slopes were significantly smaller than 0.74 obtained by Heckman et al. (1989). This confirmed their conclusion that Seyfert galaxies had an excess radio emission when compared to normal galaxies.

5. Conclusion

This paper proposed a permutation procedure which is applicable to a wide class of censoring problems. The proposed permutation procedure provided a modification of the KT statistics to censored data, using samples from the conditional distribution $P(\Sigma(A_{XY})|(X, U), (Y, V))$. Applications of the permutation procedure to two well known problems, testing independence in bivariate censored data and estimating regression slope with doubly censored data, were throughly discussed. Especially, when estimating a regression coefficient with doubly censored data, the permutation procedure was superior to the modified Theil–Sen estimator proposed by Akritas et al. (1995) in terms of MSE.

Acknowledgements

We are grateful to referees for many helpful suggestions and to SungIm Lee at DanKook University for providing Fig. 2.

References

Akritas, M.G., Murphy, S.A., LaValley, M.P., 1995. The Theil–Sen estimator with doubly censored data and application to astronomy. J. Amer. Statist. Assoc. 90, 170–177.
Clayton, D.G., 1991. A Monte Carlo method for Bayesian inference in frailty model. Biometrics 47, 467–485.
Diaconis, P., Graham, R., Holmes, S., 1999. Statistical problems involving permutations with restricted position. http://www-stat.stanford.edu/∼susan/.
Heckman, T.M., Vlitz, L., Wilson, A.S., Arumus, L., Miley, G.K., 1989. A millimeter wave survey of CO emission in Seyfert galaxies. Astrophys. J. 342, 735–758.
Kendall, M.G., 1948. The Advanced Theory of Statistics, Charles Griffin & Company,
Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data, Wiley, New York.
Mantel, N., Ciminera, J.L., 1979. Mantel–Haenszel analysis of linear matched time to response data with modifications of recovery inter-little information. Cancer Res. 37, 3863–3868.

McGilchrist, C.A., Aisbett, C.W., 1991. Regression with frailty in survival analysis. Biometrics 47, 461–466.
Oakes, D., 1982. A concordance test for independence in the presence of censoring. Biometrics 38, 451–455.
Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. J. Amer. Statist. Assoc. 63, 1379–1389.
Theil, H., 1950. A rank invariant method of linear and polynomial regression analysis. Koninklijke Nederlandse Akademie var Wetenschappen Proc. 53, 386–392.
Weier, D.R., Basu, A.P., 1980. An investigation of Kendall tau-modified for censored data with applications. J. Statist. Plann. Inference 4, 381–390.
Xie, F., Paik, M.C., 1997. Multiple imputation methods for the missing covariates in generalized estimating equation. Biometrics 53, 1538–1546.