Statistical Science
2016, Vol. 31, No. 3, 325–347
DOI: 10.1214/16-STS549

Rank Tests from Partially Ordered Data Using Importance and MCMC Sampling Methods

Debashis Mondal and Nina Hinrichs

Abstract. We discuss distribution-free exact rank tests from partially ordered data that arise in various biological and other applications where the primary objective is to conduct testing of significance to assess the linear dependence or to compare different groups. The tests here are obtained by treating the usual rank statistics, based on the completely ordered data as "latent" or missing, and conceptualizing the "latent" p-value as the random probability under the null hypothesis of a test statistic that is as extreme, or more extreme, than the latent test statistics based on the completely ordered data. The latent p-value is then predicted by sampling linear extensions or the complete orderings that are consistent with the observed partially ordered data. The sampling methods explored here include importance sampling methods based on randomized topological sorting algorithms, Gibbs sampling methods, random-walk based Metropolis–Hasting sampling methods and random-walk based modern perfect Markov chain Monte Carlo sampling methods. We discuss running times of these sampling methods and their strength and weaknesses. A simulation experiment and three data examples are given. The simulation experiment illustrates how the exact rank tests from partially ordered data work when the desired result is known. The first data example concerns the light preference behavior of fruit flies and tests whether heterogeneity observed in average light-preference behavior can be explained by manipulations in serotonin signaling. The second one is a reanalysis of the lead absorption data in children of employees who worked in a lead battery factory and consolidates the results reported in Rosenbaum [*Ann. Statist.* **19** (1991) 1091–1097]. The third one reexamines the breast cosmesis data from Finkelstein [*Biometrics* **42** (1986) 845–854].

Key words and phrases: Exact tests, fuzzy p-values, Gibbs sampling, interval censoring, linear extensions, linear rank statistics, perfect MCMC, proportional hazard model, topological sorting.

1. INTRODUCTION

Classical nonparametric methods based on permutation tests and linear rank statistics are central to various studies in biological and other sciences in which the primary objective is to conduct testing of significance to assess the presence of a trend with the serial order of data, to establish a linear dependence on some explanatory variables or to compare different groups of observations. Extensive work on these topics can be found in Page (1963), Hájek (1968), Zar (1972), Cox and Hinkley (1979), Prentice (1978), Hájek, Šidák and Sen (1999) and in many subsequent references. Suppose the data comprise a stream of in-

Debashis Mondal is Assistant Professor, Department of Statistics, Oregon State University, Corvallis, Oregon 97330, USA (e-mail: debashis@stat.oregonstate.edu). Nina Hinrichs is Assistant Professor, Department of Computer Science, University of Chicago, Chicago, Illinois 60637, USA (e-mail: nshinrichs@cs.uchicago.edu).

dependent and identically (i.i.d.) distributed bivariate observations $(y_1, z_1), \dots, (y_n, z_n)$ where the probability distribution of a pair (y_i, z_i) is unknown, and we wish to test the null hypothesis of no monotonic association between $\mathbf{y}$ and $\mathbf{z}$. Then, with monotonic association between $\mathbf{y}$ and $\mathbf{z}$ as alternative, the Spearman's rank correlation statistics ρ and its standardized version t_1 (after taking the Fisher's Z-transformation) are respectively given by

$$\begin{aligned} \rho &= 1 - 6 \sum_{j=1}^{n} (r_{y,j} - r_{z,j})^2 / \{n(n^2 - 1)\}, \\ t_1 &= \frac{1}{2} n^{\frac{1}{2}} \log \frac{1+\rho}{1-\rho}, \end{aligned} \tag{1.1}$$

where $r_{y,j}$ denotes the rank of y_j in the y-sequence $\mathbf{y} = (y_1, \dots, y_n)$ and $r_{z,j}$ is the rank of z_j in the z-sequence $\mathbf{z} = (z_1, \dots, z_n)$. If no monotonic association between $\mathbf{y}$ and $\mathbf{z}$ exists, the finite sampling distributions of ρ and t_1 are free of the unknown probability distribution of a pair (y_i, z_i), and can be evaluated exactly by enumerating all possible permutations of the data. Let Φ_n denote the null distribution of t_1. Then the exact p-value for a two-sided test is equal to $2(1 - \Phi_n(|t_1|))$. If n is moderately large, asymptotic analysis applies and the sampling distribution of t_1 becomes approximately Gaussian with mean 0 and variance 1 and the asymptotic p-value for a two-sided test comes to $2(1 - \Phi(|t_1|))$, where Φ is the cumulative distribution function of a standard Gaussian random variable.

Similarly, in testing a location shift between two independent samples of observations, namely, $\mathbf{y} = (y_1, \dots, y_n)$ and $\mathbf{z} = (z_1, \dots, z_{n'})$, the classical Wilcoxon rank sum statistic takes the form of

$$t_2 = \sum_{j=1}^{n} w(r_{x,j}), \tag{1.2}$$

where $\mathbf{x}$ is the combined sample $(y_1, \dots, y_n, z_1, \dots, z_{n'})$, $r_{x,j}$, for $j = 1, 2, \dots, n$, is the rank of y_j among $y_1, \dots, y_n, z_1, \dots, z_{n'}$ and w is a weight function such that

$$w(j) = -1 + 2j/(n + n' + 1). \tag{1.3}$$

If no location shift occurs, the finite sampling distribution of t_2 is also free of the unknown marginal probability distribution of x_i and can be evaluated exactly by enumerating all possible permutations of the data. Furthermore, it is well known that the standardized version of t_2 is asymptotically Gaussian again providing an easy way to compute asymptotic p-values when both n and n' are moderately large. More generally, linear rank statistics and other permutation test statistics are formulated in a fashion similar to t_1 or t_2 and, in principle, it is not just the asymptotic p-values, but also the exact distributions of the test statistics and the exact p-values are of interest and can often be obtained by enumerating all possible permutations of the data.

However, when only partially ordered data are observed, the above mentioned classical test procedures based on rank statistics fail because of our inability to compute the test statistics such as t_1 or t_2 which require the knowledge of the complete ordering of the data.

As a specific example, consider Figure 1 that provides a partial order of serotonin signaling in 14 different experimental manipulations of fruit flies (Drosophila). These serotonin signalings are outcomes of one of the randomized experiments conducted at The Evolution of Behavior Group (EBG) in the Rowland Institute at Harvard University. In this experiment, starting with 8 different elementary manipulations $a_1, a_2, \dots, a_8$, EBG designed $n = 14$ different combinations of manipulation, each of which gives rise to a variant of fruit flies. We call these combinations $\mathcal{V} = \{v_1, \dots, v_{14}\}$. The full details of the phototactic personality behavior data for these flies are displayed in Table 3 of Section 4.2. The response $\mathbf{y}$ is a vector of measurements of the heterogeneity in light-preference behavior, which essentially summarize an average run of individual flies toward light or darkness when startled and are actually determined by the preference of individual flies in choosing either left or right turns in branching mazes. The biology here allows us to identify some of the directions in which each experimental manipulation is going to affect serotonin signaling. However, the actual magnitude of the change in serotonin signaling is not known. Thus, 14 different combinations of genetic mutations give rise to the set of partially ordered treatments shown in Figure 1. In particular, an arrow from i to j implies the serotonin signaling (v_i) in the fruit fly type i is less than that (v_j) in the fruit fly type j. Thus, $v_4 < v_1$. However, we do not know whether $v_5 < v_4$. Furthermore, under the null hypothesis that serotonin has no association with the phototactic personality, we expect that the rank correlation between y_is and v_is will be 0. A positive association, however, would imply a different story. Thus, there is interest in knowing whether the partially ordered v_is is positively associated with the ordering of the experimentally observed measurement y_is. However, classical test procedures based on rank statistics

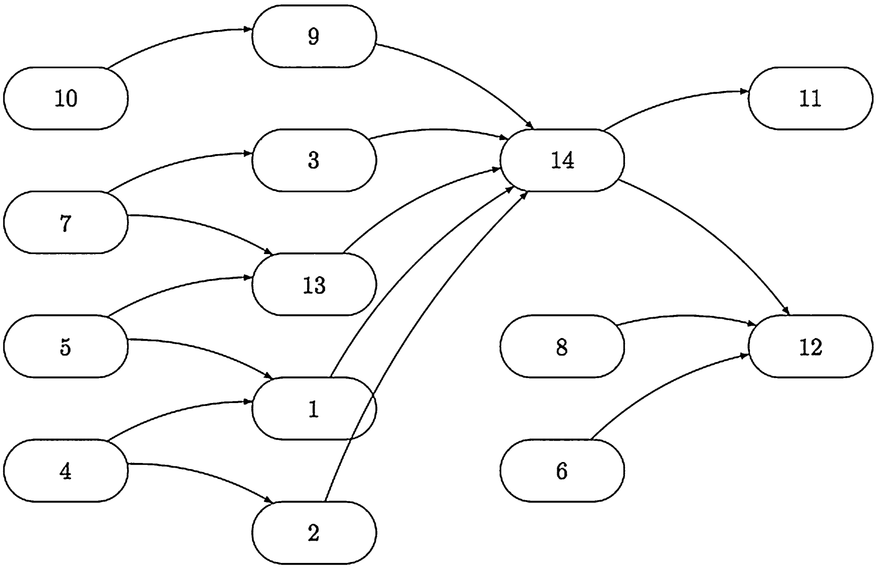

FIG. 1. *Partial order $(\mathcal{V}, \mathcal{E})$ of Serotonin signaling in* 14 *different experimental manipulations of fruit flies.*

t_1 fails here, as we do not know the complete ordering of the serotonin signaling.

In this paper, we focus on exact permutation and other rank tests from partially ordered data. These exact tests make use of the seminal work of Geyer and Meeden (2005), and Thompson and Geyer (2007). However, they differ from the testing approach in Rosenbaum (1991) which relies on the asymptotic null distributions of certain statistics or pivotal quantities of the partially ordered data. Here, we treat the usual test statistics based on the completely ordered data as latent or missing and conceptualize the latent p-value as the random probability under the null hypothesis of a statistic that is as extreme, or more extreme, than the latent test statistics based on the completely ordered data. We then obtain the exact "fuzzy p-value" or the exact predictive distribution of the latent p-value as the exact conditional distribution of the latent p-value given the partially ordered data. If the sample size is large, we also focus on their asymptotic versions. Thus, consider, for example, testing no monotonic association between **y** and **z** based on partially ordered data in that we only know $z_{j_i} < z_{k_i}$ for certain values of $i = 1, 2, \ldots, l$, and the full rank vector of y-values. Obviously, we cannot compute t_1 because in this situation we do not know all the $r_{z,j}$s. Nevertheless, we can respectively conceptualize the latent exact and the latent asymptotic p-values by

$$\begin{aligned}\theta_E(t_1) &= 2\bigl(1 - \Phi_n(|t_1|)\bigr),\\ \theta_A(t_1) &= 2\bigl(1 - \Phi(|t_1|)\bigr)\end{aligned} \tag{1.4}$$

and compute the exact and the asymptotic predictive distributions of the latent p-value using the conditional distributions

$$\begin{aligned}Q_E(\alpha) = \Pr\bigl(&\theta_E(t_1) \le \alpha \mid \mathbf{r_y}, r_{z_{j_i}} < r_{z_{k_i}},\\ &i = 1, 2, \ldots, l\bigr)\end{aligned} \tag{1.5}$$

and

$$\begin{aligned}Q_A(\alpha) = \Pr\bigl(&\theta_A(t_1) \le \alpha \mid \mathbf{r_y}, r_{z_{j_i}} < r_{z_{k_i}},\\ &i = 1, 2, \ldots, l\bigr).\end{aligned} \tag{1.6}$$

Similarly, we can derive the exact and the asymptotic predictive distributions of the latent p-values based on the Wilcoxon rank sum statistics from partially ordered data by making use of the equations (1.2) and (1.3).

There are several advantages of adopting the exact and the asymptotic predictive distributions of latent p-values as an inferential tool. As noted in Thompson and Geyer (2007), these predictive distributions generalize the conventional definition of p-values and derive their validity from the fact that they as random variables (i.e., as functions of the data) have uniform(0, 1) distribution under the null hypotheses, and thus, they achieve their exact nominal significance level like any other randomized tests. Thus, just like the conventional setup, the test statistics here are sensitive to certain departures of interest from the null distribution and these predictive distributions summarize the weight of evidence against the null hypothesis. In particular, just like a low conventional p-value, a predictive mass function (or a density function) of the latent p-value that concentrates largely around a small neighborhood of 0 provides certain evidence against the null hypothesis. However, further caution is required regarding how we quantify these “low values” and interpret the weight of evidence. For example, if the predictive mass function of a discrete latent p-value concentrates almost all of its mass within $(0, 0.01)$ [or within $(0.1, 1)$], we can interpret using the standards of conventional significance testing that there is evidence against (or lack of evidence against) the null hypothesis. But, by the same standards, the weight of evidence is uneven if the predictive mass function of a discrete latent p-value concentrates almost all of its mass uniformly on $\{0.00, 0.01, \dots, 0.15\}$. In practice, this unevenness may appear to make the test procedures more conservative, but they actually point to a more enriching collection of evidence where some evidence speak against the null hypothesis and some does not. They neither suggest outright rejection of the null hypothesis nor complete ignorance of the evidence against it. We shall also see through examples that predictive distributions of latent p-values can provide very strong and enriching evidence against null hypotheses. Finally, it is worthwhile to point out that most rank test statistics that we shall consider here are distribution-free meaning their finite sample distributions do not depend on any nuisance quantity such as the underlying marginal distributions of the data unit. Thus, there would be little advantage in pursuing related Bayesian methods such as computing posterior predictive p-values (see e.g., Meng, 1994; Gelman, Meng and Stern, 1996 and Bayarri and Berger, 2000, 2004). In fact, such Bayesian methods would provide the same numerical answers that we obtain here, no matter what prior we choose for nuisance quantities.

However, the above approach of computing exact predictive distributions of latent p-values using equations (1.5) and (1.6) presents a new challenge, as such computations require complete enumeration of the set of all possible linear extensions (or the complete orderings; see Section 3 for their formal definitions) of the observed finite partial ordering, which is known to be a very difficult problem with $\#P$ computational complexity (Brightwell and Winkler, 1991). For example, there are approximately 2.75×10^6 linear extensions for the partial order shown in Figure 1. Thus, enlisting all possible linear extensions is quite a task even for such a small example. Instead, we resort to sampling based methods and explore various schemes for generating linear extensions of a finite partial order. First, we explore a novel importance sampling scheme based on the topological sorting algorithms introduced in Kahn (1962) and discussed in Cormen et al. (2001). This algorithm generates one random linear extension in linear time in the total number of observations and the total number of partial order constraints. We also discuss modifications to this importance sampling scheme using various “look ahead” schemes. However, as we shall see, these importance sampling schemes can fail when the distribution from which the random linear extensions are generated gets too far apart from the target distribution. Thus, we also consider Markov chain Monte Carlo (MCMC) procedures to generate random samples of linear extensions and advance rank tests from partially ordered data. To this end, we derive novel Gibbs sampling methods that generate uniform samples of linear extensions of a finite partial order efficiently and works very well in small to moderately large sample sizes. Interestingly, there is a substantial, and hitherto unutilized body of literature in probability and discrete mathematics on generating random samples of linear extensions of a finite partial order using the traditional random-walk based MCMC and the modern perfect MCMC algorithms. These include Matthews (1991), Karzanov and Khachiyan (1991), Bubley and Dyer (1999), Wilson (2004), Huber (2006) and many subsequent references. Thus, we also adapt the ideas presented by Karzanov and Khachiyan (1991) and Huber (2006) in deriving random-walk based sampling schemes for generating linear extensions of a given partial order. It must be emphasized that the expected running time of the random-walk based perfect MCMC (and also the mixing time of the random-walk based MCMC) sampling is $O(n^3 \log n)$ for n number of observations. Thus, for small to moderately large sample sizes, we can apply these MCMC procedures to compute the exact predictive distributions of latent p-values with great accuracy, allowing the asymptotic theory to prevail only for very large data sets.

The rest of the paper is laid out as follows. In Section 2, we provide further details of some of the important rank tests that arise in partially ordered data and discuss details of approximating the predictive distribution of the latent p-value using Monte Carlo, MCMC, and importance sampling methods. In Section 3, we introduce the basic topological sorting algorithm to generate linear extensions to the partial order efficiently. We next develop the importance sampling scheme to generate linear extensions. We further provide various “look ahead” schemes that bring the probabilities of generating each linear extension closer to the uniform distribution. Next, we present Gibbs sampling methods to generate uniform linear extensions to the partial order. We then discuss random-walk based MCMC procedures. At the end of Section 3, we devote attention to the sophisticated random-walk based perfect MCMC computations. In Section 4, we provide a simulation experiment and three data examples. The simulation experiment illustrates how the exact rank tests from partially ordered data work when the desired result is known. The first data example examines the association between the phototactic behavior of fruit flies and different combinations of serotonin manipulations. The second one reanalyzes the lead absorption levels in children of workers in a lead battery factory and provides a new way to look into the results obtained in Rosenbaum (1991, 2002). The third one focuses on the interval censored breast cosmesis data and provides a basic comparison test between two groups. Through these examples, we highlight various strength and weaknesses of our approach. Finally, in Section 5, we provide a summary and briefly indicate some future research directions.

2. FURTHER RANK TESTS AND p-VALUES

2.1 Further Rank Tests from Completely Ordered Data

In this section, we review additional rank tests that have proved useful in various statistical applications. As ranks are maximum invariant under monotonic transformations of the original data, it needs to be emphasized that rank tests are particularly appropriate when the scale of measurement is somewhat arbitrary and there is a little loss if we restrict ourselves to just ordering the different response values. In this regard, we have already seen useful rank tests based on the Spearman's rank correlation and the Wilcoxon rank sum statistic. To consolidate these ideas further and to derive other possible test statistics, suppose that we have two sets of data, namely $\mathbf{y} = (y_1, y_2, \dots, y_n)$ and $\mathbf{z} = (z_1, z_2, \dots, z_{n'})$, with one coming from probability density $f(x-\mu)$ and the other from $f(x)$. Furthermore, we are interested in testing $\mu = 0$ against the one-sided alternative $\mu > 0$. It then follows that (see Cox and Hinkley, 1979) the locally most powerful rank test leads to the test statistic

$$t_3 = \sum_{i=1}^{n} w(r_{x,i}), \tag{2.1}$$

where, as in equation (1.2), $\mathbf{x} = (y_1, \dots, y_n, z_1, \dots, z_{n'})$ is the combined sample, $r_{x,i}$, for $i = 1, 2, \dots, n$, is the rank of y_i among $y_1, \dots, y_n$ and $z_1, \dots, z_{n'}$, $x_{(i)}$, for $i = 1, 2, \dots, n$, equals to the ith smallest value of the combined sample $\mathbf{x}$ and

$$w(i) = -\mathrm{E}\left[\frac{f'(x_{(i)})}{f(x_{(i)})}\right]. \tag{2.2}$$

It is well known that, when $\mu = 0$, t_3 is asymptotically Gaussian with

$$\mathrm{E}t_3 = \frac{n}{n+n'} \sum_{j=1}^{n} w(j) = \frac{n}{n+n'}\bar{w}, \tag{2.3}$$

$$\begin{aligned} \operatorname{var} t_3 = {} & \frac{nn'}{(n+n')(n+n'-1)} \\ & \cdot \sum_{j=1}^{n} (w(j) - \bar{w})^2. \end{aligned} \tag{2.4}$$

In general, it may be difficult to compute $w(i)$ in equations (2.1) and (2.2) exactly, but one can obtain useful approximations. Typically, the quantity

$$\begin{aligned} & -f'\{F^{-1}(i/(n+n'+1))\} \\ & /f\{F^{-1}(i/(n+n'+1))\}, \end{aligned} \tag{2.5}$$

where F is the cumulative distribution function of f, provides a good approximation for $w(i)$. Several important rank tests follow from equations (2.1) to (2.5). When f represents the logistic distribution, $w(i) = -1 + 2i/(n+n'+1)$ and the corresponding test statistic t_3 is the traditional Wilcoxon or Mann–Whitney statistic in equation (1.2). However, when f is Gaussian, $w(i) = \mathrm{E}x_{(i)}$ and t_3 approximates the van der Waerden test. An extreme minimum value density $f(x) = \exp(x - e^x)$ gives the exponential score and the log rank statistic. Last but not the least, the double exponential density gives rise to the median test.

The above discussion of rank tests extends to more complicated problems such as comparing multi-samples, testing regression relationships and detecting serial correlations. In such instances, statistical derivations often lead to a general rank test statistic of the form

$$t_4 = \sum_{i=1}^{n} c_i w(r_{x,i}), \tag{2.6}$$

where the constants $c_1, c_2, \dots, c_n$ and weight function $w(i)$ depend the null distribution of the specific probabilistic set-up. Among others, Cox and Hinkley (1979) provide several examples of such test statistic and formulas for their asymptotic distributions. See also the famous work of Kruskal and Wallis (1952) and the book by Puri and Sen (1971).

Furthermore, it must be emphasized that, in certain applications, nonlinear rank test statistics are used and are very different than what we have described through equation (2.6). For example, Gordon (1979a, 1979b) provides different nonstandard measures of agreement between two rankings of a set of n objects which lead to various nonstandard rank tests. One such measure is the size of the maximal subset of objects that gives rise to a perfect agreement between the "reduced" rankings. Thus, the corresponding test statistic here is the length of a longest increasing subsequence, of not necessarily consecutive numbers, in the second sequence of ranks when the first sequence of ranks is sorted in the natural order. This test statistic can be computed in $O(n \log n)$ operations using binary searches (see, e.g., Schensted, 1961 and Hunt and Szymanski, 1977) and its asymptotic null distribution can be found using the well-known Tracy–Widom law (see, e.g., Baik, Deift and Johansson, 1999).

When only partial information is known, we may not observe all ranks of the underlying variables, making it difficult to compute any of the above mentioned rank test statistics. However, in such instances, we shall use methods developed in Section 3 to generate the set of possible full/complete ranks from the observed partial ranks and pursue statistical inference by computing the predictive distributions of latent p-values. In Section 4, we shall see that such rank tests from partially ordered data arise naturally in many practical applications.

2.2 Rank Tests for Censored Data

Censored data is one area in which both partial orderings of observations and various rank tests have had an important and developing role in statistical inference. Typically, for censored data ranks are the possible rank vectors of the underlying uncensored values. However, as Crowley (1974) points out, the distribution of the underlying rank vector often depends in a complicated manner on the censoring mechanism. Thus, even for the simplest situation of the right censored data, such complications has led to the development of linear rank tests based on the score statistics of the marginal probability of the rank vector of the observed uncensored values only; see, for example, Prentice (1978) and Kalbfleisch and Prentice (1980). These complications and many others may not always be easily amenable to the approach we have taken in this paper. Nevertheless, we can still derive rank tests and compute the predictive distribution of latent p-values for certain types of censored data. Below we provide two examples of such rank test procedures.

Our first example is on the most powerful local rank test from interval-censored data which, following Self and Grossman (1986), have received much attention in biostatistics literature. The basic set-up is as follows. We have a regression model

$$y_i = \mu + \mathbf{z}_i^T \boldsymbol{\beta} + \kappa \varepsilon_i,$$

where y_i is the actual time of occurrence of an event of interest for the ith individual, $\mathbf{z}_i$ represents covariate information, $\boldsymbol{\beta}$ denotes regression coefficients, μ is the mean parameter, κ is the scale parameter and ε_i indicates the residual errors. We further assume that f is the probability density of ε_i. Since actual data are interval censored, we do not observe y_i directly. Instead, we observe an interval $I_i = (y_{1,i}, y_{2,i}]$ on which the actual observation y_i falls. In addition, we assume that the censoring mechanism is independent of the response variable Y_i. This assumption implies that the knowledge that an observation y_i is censored into a given interval I_i provides no further information about the distribution of the response beyond that information conveyed by knowledge of the interval's endpoints. A typical hypothesis of interest here is whether $\beta = 0$, in other words, whether the response is independent of the covariates. In such a setup, the marginal likelihood becomes

$$L(\boldsymbol{\gamma}) = \sum_{\mathbf{r} \in \mathcal{R}} \int_{\mathcal{A}(\mathbf{r})} \cdots \int \prod_{i=1}^{n} f(y_{0,i} - \mathbf{z}_i^T \boldsymbol{\gamma})\, dy_{0,i},$$

where $y_{0,i} = (y_i - \mu)/\kappa$, $\boldsymbol{\gamma} = \boldsymbol{\beta}/\kappa$, $\mathcal{R}$ is the set of possible rank vectors $\mathbf{r}$ of $\mathbf{y}$s that are consistent with the observed interval data I_i, and $\mathcal{A}(\mathbf{r})$ is the set of possible values of $y_{0,i}$ that preserves a ranking vector $\mathbf{r}$. For testing $\boldsymbol{\gamma} = 0$, Self and Grossman (1986) then proposed the test statistic

$$\mathbf{t}_5 = \sum_i \mathbf{z}_i c_i, \tag{2.7}$$

where the constant c_i is given by

$$c_i = \sum_k w_i(k) \mathrm{E}\left[\frac{d}{dy_{0,(k)}} \log f(y_{0,(k)})\right]. \tag{2.8}$$

In the above, the summation is over values of k from $\min_{\mathbf{r} \in \mathcal{R}} r_i$ to $\max_{\mathbf{r} \in \mathcal{R}} r_i$. Furthermore, in the above, $y_{0,(k)}$ denotes the kth order statistic of $\mathbf{y}_0$, and $w_i(k)$ is the proportion of rank vectors in $\mathcal{R}$ for which the rank of individual i equals k. Now note that the censored data rank scores are weighted averages of uncensored data rank scores and we can approximate $\mathrm{E}d\{\log f(y_{0,(k)})\}/dy_{0,(k)}$ as done in equations (2.2) and (2.5). Furthermore, we can compute $w_i(k)$ by generating a sample of complete ranks (or linear extensions) from the observed partial ranks using our algorithm developed in Section 3. In principle, we can thus conduct the test developed by Self and Grossman (1986). See also Vandal and Gentleman (1998) and Vandal, Conder and Gentleman (2009) that derive another algorithm, albeit a complicated one, for performing this test.

The second and perhaps the most important example is the interval censored data for the proportional hazard model (Cox, 1972), for which the hazard function takes the form of

$$\lambda(t;\mathbf{z}) = \lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}. \tag{2.9}$$

In the above, λ_0 is the baseline hazard and $\boldsymbol{\beta}$ is the $k \times 1$ vector of regression coefficients. Assuming that the censoring mechanism is independent of the failure times, Kalbfleisch and Prentice (1980) show that the marginal likelihood of $\boldsymbol{\beta}$ based on the distribution of the rank vector $\mathbf{r}$ of the failures times is

$$L(\boldsymbol{\beta}) = \Pr(\mathbf{r} \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{\mathbf{z}_i^T\boldsymbol{\beta}}}{\sum_{j\in\mathcal{T}_i} e^{\mathbf{z}_j^T\boldsymbol{\beta}}}, \tag{2.10}$$

where $\mathcal{T}_i$ is the set of individuals at risk of failure at the time of the ith failure. See also Satten (1996) and Goggins et al. (1998) that made alternative use of this likelihood for interval-censored data.

Of particular interest here are the tests for covariate effects, for which several asymptotic procedures have been used in the past. For brevity of discussion, we only focus on the Wald test for the significance of the covariate effect. Let $\widehat{\boldsymbol{\beta}}$ be the value of $\boldsymbol{\beta}$ that maximizes the marginal likelihood $L(\boldsymbol{\beta})$. Let $\boldsymbol{\Sigma}$ be the dispersion matrix of $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\Sigma}}$ be a consistent estimator of $\boldsymbol{\Sigma}$. Typically, $\widehat{\boldsymbol{\beta}}$ is obtained as the solution of the score equation and $\widehat{\boldsymbol{\Sigma}}$ as the inverse of the Hessian matrix of the logarithm of $L(\boldsymbol{\beta})$. Then, under the null hypothesis $\boldsymbol{\beta} = \mathbf{0}$, the Wald test statistics

$$t_6 = \widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\beta}} \tag{2.11}$$

is asymptotically χ^2 with k degrees of freedom, and the asymptotic p-value is $1 - D(t_6 \mid k)$, where $D(x \mid k)$ is the cumulative distribution function of a χ_k^2 random variable.

It is worthwhile to point out that, if observations are exchangeable, under the null hypothesis $\boldsymbol{\beta} = \mathbf{0}$ the distribution of the rank vector $\mathbf{r}$ is uniform over all permutations of the ranks. Thus, if observations are exchangeable, we can actually pursue an exact test by computing the exact null distribution D_n of t_6 by recalculating the test statistic for each of the possible reorderings of $\mathbf{y}$ with the order of values of $\mathbf{z}$ remaining fixed. This permutation test gives the exact p-value $1 - D_n(t_6)$.

However, if observations fall into overlapping intervals, we only know the partial ranks rather than the complete rank $\mathbf{r}$. The set of complete ranks then is precisely the set of linear extensions of observed partial ranks. In that case, we cannot compute t_6 and so we rather treat $1 - D_n(t_6)$ as the latent exact p-value and $1 - D(t_6 \mid k)$ as latent asymptotic p-value. The predictive distributions of these latent p-values then are given by the conditional distributions

$$Q_E(\alpha) = \Pr\bigl(1 - D_n(t_6) \le \alpha \mid \text{observed partial ranks}\bigr) \tag{2.12}$$

and

$$Q_A(\alpha) = \Pr\bigl(1 - D(t_6 \mid k) \le \alpha \mid \text{observed partial ranks}\bigr). \tag{2.13}$$

2.3 Exact Predictive Distributions of Latent p-Values

Let t denote an arbitrary rank test statistic under consideration. Since ranks are discrete random variables, often t is also a discrete random variable and thus, it is typical that the exact latent p-value is a discrete random variable. One consequence of this is that certain quantiles of t may not be unique and, therefore, any testing of significance based on t demands a randomized version. Thompson and Geyer (2007) thus define the exact latent p-value in the following way. Let ϕ denote the critical function for a test based on t. If the test is two-sided such as the ones discussed in Sections 1 and 2.1, the critical function at the α level of significance takes the form of

$$\phi(t,\alpha) = \begin{cases} 0 & \text{if } a_1 < t < a_2, \\ \eta & \text{if } t = a_1 \text{ or } t = a_2, \\ 1 & \text{if } t < a_1 \text{ or } t > a_2, \end{cases}$$

where a_1, a_2 and η are functions of t and α such that

$$\mathrm{E}\phi(t,\alpha) = \alpha$$

for all $0 \le \alpha \le 1$. In other words, we can respectively take a_1 and a_2 to be any $\alpha/2$ and $1 - \alpha/2$ quantiles of the null distribution of t and η must satisfy

$$\eta \Pr(t = a_1 \text{ or } t = a_2) = \bigl(\alpha - \Pr(t < a_1 \text{ or } t > a_2)\bigr).$$

Since α gives the probability that a randomized test rejects the null model, the predictive distribution of the latent p-value for a discrete test statistic t becomes the random conditional distribution

$$(2.14) \qquad Q(\alpha) = \mathrm{E}\big(\phi(t, \alpha) \mid \text{partially ordered data}\big)$$

generalizing equations such as (1.5) or (2.12).

2.4 Monte Carlo and MCMC Approximations

Typically, in most practical situations, it is extremely difficult to analytically compute the exact predictive distribution of the latent p-value from equation (2.14). The same applies to computing asymptotic predictive distribution of the latent p-value from equations such as (1.6) and (2.13), and thus their Monte Carlo or MCMC approximations are desirable. To this end, suppose that G is the cumulative null distribution function of the test statistics t. Let $d_1, d_2, \dots, d_N$ be a Monte Carlo sample from the uniform(0, 1) distribution and let $t^{(1)}, t^{(2)}, \dots, t^{(M)}$ be a Monte Carlo sample from the conditional distribution of t given the partially ordered observed data. Then the numbers

$$p\big(t^{(i)}\big) = \frac{1}{N} \sum_{j=1}^{N} \big(1_{\{t^{(i)} < G^{-1}(d_j/2) \text{ or } t^{(i)} > G^{-1}(1-d_j/2)\}} + \eta 1_{\{t^{(i)} = G^{-1}(d_j/2) \text{ or } G^{-1}(1-d_j/2)\}}\big)$$

constitute a sample from the predictive distribution of the latent p-value in equation (2.14). In practice, G or G^{-1} may not be known analytically. In such situations, we need to draw a Monte Carlo sample $s_1, s_2, \dots, s_N$ from the null distribution of t and approximate G^{-1} from the empirical distribution of $s_1, s_2, \dots, s_N$. In practice, the empirical distribution function of $p(t^{(i)})$, box plot or some other empirical summary statistic can be used to further assess, describe and graphically display the exact nature of this predictive distribution.

Furthermore, in practice, one would usually adopt a slightly simpler procedure for obtaining a sample from the predictive distribution, based on either the asymptotic analysis or the fact that the latent p-value is often the probability of obtaining a test statistic at least as large as the one that was actually observed. Thus, for the test based on Spearman's rank correlation in equation (1.1), suppose that $t_1^{(1)}, t_1^{(2)}, \dots, t_1^{(M)}$ constitute a Monte Carlo sample from the conditional distribution of t_1 given the partially ordered observed data. It is then immediate that the numbers $2(1 - \Phi(|t_1^{(1)}|)), 2(1 - \Phi(|t_1^{(2)}|)), \dots, 2(1 - \Phi(|t_1^{(M)}|))$ provide a sample from the predictive distribution of the asymptotic latent p-value in equation (1.5). Similarly, for testing linear dependence in the proportional hazard model using the rank based test in equation (2.11), if we have a Monte Carlo sample $t_6^{(1)}, t_6^{(2)}, \dots, t_6^{(M)}$ from the conditional distribution of t_6 given the partially ordered observed ranks, the numbers $1 - D(t_6^{(1)} \mid k), 1 - D(t_6^{(2)} \mid k), \dots, 1 - D(t_6^{(M)} \mid k)$ form a sample from the predictive distribution of the latent p-value in equation (2.13). When the asymptotic χ^2 null distribution is inappropriate for (2.11) due to small sample size, we can alternatively obtain a sample from the exact predictive distribution in equation (2.12) by computing

$$p\big(t_6^{(i)}\big) = \frac{1}{N} \sum_{j=1}^{N} 1_{\{s_j \geq t_6^{(i)}\}},$$

where $s_1, s_2, \dots, s_N$ constitute a random sample from the null distribution of t_6.

In Section 3, we shall discuss standard and sophisticated perfect Markov chain Monte Carlo methods for sampling linear extensions of a finite partial order. While any MCMC sample can provide an approximation to the predictive distribution, it is worthwhile to point out that there is further benefit if we obtain an exchangeable sample $t^{(1)}, t^{(2)}, \dots, t^{(M)}$, by running a running a Markov chain backward as well as forward in time as proposed in Besag and Clifford (1989) and implemented in Besag and Mondal (2013). This is primarily because exchangeable random variables enjoy certain symmetry and better distributional properties of the sample quantiles.

2.5 Approximations via Importance Sampling

We have already discussed how to apply Monte Carlo or MCMC methods to approximate the predictive distribution of the latent p-value. In principle, an alternative to these methods can be importance sampling methods and they deserve some discussion. In particular, any importance sampling methods are considered to have the ability to bridge gaps between Monte Carlo and MCMC samplings, as they allow us to approximate $\mathrm{E}_\pi g$ for a distribution π that is close to another distribution π_0 from which we can easily draw random samples. To consolidate this idea, suppose that we have a discrete random variable X with possible values $\mathcal{X}$, and a probability mass function π on $\mathcal{X}$. The expected value of a function g of X, with respect to the probability function π is

$$\mathrm{E}_\pi g(X) = \sum_{x \in \mathcal{X}} g(x)\pi(x).$$

Suppose further that it is difficult to sample directly from the probability distribution π. Instead, we can sample from another distribution π_0 on $\mathcal{X}$. We can then calculate $\mathrm{E}_\pi g(X)$ as follows:

$$\mathrm{E}_\pi g(x) = \sum_{x\in\mathcal{X}} g(x)\frac{\pi(x)}{\pi_0(x)}\pi_0(x) \tag{2.15}$$
$$\approx \frac{1}{M}\sum_{i=1}^{M} g(x_i)\frac{\pi(x_i)}{\pi_0(x_i)},$$

where $x_1, x_2, \ldots, x_M$ are M random draws from the distribution π_0.

In most situations, we will only know the distributions π and π_0 up to respective normalizing constants. Specifically, we can write $\pi(x) = \pi^*(x)/\varpi$ and $\pi_0(x) = \pi_0^*(x)/\varpi_0$, where the height functions $\pi^*(x)$ and $\pi_0^*(x)$ are known and the normalizing constants ϖ and ϖ_0 are computationally intractable. For example, if we are interested in generating a sample from the uniform distribution over all possible linear extensions of a partial order, we know $\pi(x)$ up to a normalizing constant, namely, $\pi(x) = 1/\varpi$, where ϖ is the total number of linear extensions and is intractably complex. In this case, we can approximate $\mathrm{E}_\pi g$ by adopting the ratio estimate,

$$\widehat{\mathrm{E}}_\pi g(X) \approx \sum_{i=1}^{M} \ell(x_i) g(x_i), \tag{2.16}$$

where

$$\ell(x_i) = \frac{\pi^*(x_i)/\pi_0^*(x_i)}{\sum_{j=1}^{M}\pi^*(x_j)/\pi_0^*(x_j)},$$

which is asymptotically unbiased as M increases to infinity. Thus, in principle we can apply importance sampling estimates in equation (2.16) to compute the predictive distribution of the latent p-values by setting $g(x) = \phi(t(x), \alpha)$. We can also draw a reweighed histogram of this predictive distribution by calculating the expected number of observations in a bin (b, b') using the indicator function $g(x) = 1_{t(x)\in(b,b')}$. Typically, equation (2.16) provides an useful estimate when there are no large weights among the $\ell(x_i)$. In practice, the latter condition holds if π and π_0 are not too far apart; see, for example, Ferrenberg, Landau and Swendsen (1995), Besag (2004) and Liu (2008).

Finally, it is worth pointing out that one can use sampling importance-resampling to draw an approximate random sample from the target distribution π. The basic idea is as follows. Once an independent random sample $x_1, x_2, \ldots, x_M$ is drawn from π_0, we draw a smaller sample $x_1', x_2', \ldots, x_{M'}'$ with or without replacement from $\{x_1, x_2, \ldots, x_M\}$ with probability proportional to weights that can computed from the ratios $\pi^*(x_i)/\pi_0^*(x_i)$, for $i = 1, 2, \ldots, M$; see, for example, Rubin (1987), Smith and Gelfand (1992) and Skare, Bølviken and Holden (2003).

3. SAMPLING LINEAR EXTENSIONS OF A PARTIAL ORDER

3.1 Posets and Linear Extensions

Let $\mathcal{V}$ denote a finite set with elements $v_1, v_2, \ldots, v_n$: in practice, elements of $\mathcal{V}$ might represent, for example, bivariate values, open intervals, or more complex quantities. Additionally, let $\prec$ denote a *partial order* over $\mathcal{V}$. In other words, $\prec$ describes a binary relation that is both antisymmetric [$(v \prec v'$ and $v' \prec v) \Rightarrow v = v'$ for all v, v' in $\mathcal{V}$] and transitive [$(v \prec v'$ and $v' \prec v'') \Rightarrow v \prec v''$ for all $v, v', v'' \in \mathcal{V}$]. Typically, we read $v \prec v'$ as "v precedes v'". It is often convenient to think of the partial order as a directed acyclic graph $(\mathcal{V}, \mathcal{E})$ such that the vertices of the graph are the elements $v_1, v_2, \ldots, v_n$, and a directed edge $(v, v') \in \mathcal{E}$ indicates $v \prec v'$. Next, we denote a *linear extension* of the partial order $(\mathcal{V}, \prec)$ by $(\mathcal{V}, <)$. A linear extension describes a complete order of all elements of $\mathcal{V}$ that is consistent with the original partial order. In other words, if $v \prec v'$ in the partial order, then $v < v'$ in the linear extension. Let $\mathbf{A}$ be the adjacency matrix of the graph $(\mathcal{V}, \mathcal{E})$. In other words, the (v, v')th entry of $\mathbf{A}$, namely, $\mathbf{A}[v, v'] = 1$ if $v \prec v'$ and is 0 otherwise.

3.2 A Randomized Topological Sorting Algorithm

A *topological sorting* of a directed acyclic graph $(\mathcal{V}, \mathcal{E})$ is a complete ordering of its vertices such that the element v comes before element v' in the ordering if there is a direct edge from v to v' in $\mathcal{E}$. Thus, a topological sorting of $(\mathcal{V}, \mathcal{E})$ is the same thing as a linear extension of the partial order $(\mathcal{V}, \prec)$. Various algorithms are available to obtain a topological sorting of a directed acyclic graph $(\mathcal{V}, \mathcal{E})$. These include the sequential algorithm of Kahn (1962) that picks one element of $\mathcal{V}$ at a time in the same order as the eventual topological sort, and the recursive depth first search algorithm of Tarjan (1976) (see also Cormen et al., 2001) that provides an eventual topological sort of the elements of $\mathcal{V}$ in the opposite order of their last visit. These algorithms have a running time linear in the number of elements of $\mathcal{V}$ plus the number of edges in $\mathcal{E}$.

Here, following Kahn (1962), we pursue a randomized topological sorting algorithm that offers an efficient way to produce linear extensions of a partial order. The basic iterative algorithm maintains, in addition to the current list of vertices $\mathcal{V}$ and edges $\mathcal{E}$, a list $\mathcal{S}$ which contains the subject of $\mathcal{V}$ with no incoming edges. At each iteration, the algorithm outputs a random vertex v from $\mathcal{S}$, removes v from $\mathcal{S}$ and $\mathcal{V}$, and removes all edges (v, v') from $\mathcal{E}$. The algorithm terminates when $\mathcal{S}$ becomes empty.

To efficiently implement the algorithm, we maintain adjacency lists for each vertex v such that v' is in the adjacency list of v exactly when there is an edge from v to v'. By iterating over the edges, we calculate the number of incoming edges for each vertex and add to $\mathcal{S}$ those vertices with no incoming edges. When a vertex v is deleted, we simply decrement the number of incoming edges for all v' in the adjacency list of v and add to $\mathcal{S}$ those which equal 0.

3.3 Importance Sampling via Randomized Topological Sorting

The above algorithm, just like the algorithm in Kahn (1962), generates one random linear extension in $O(|\mathcal{V}| + |\mathcal{E}|)$ running time and eventually generates the set of linear extensions of the partial order. However, it does not generate all linear extensions with equal probability. Nevertheless, once a linear extension has been generated from the algorithm, it is straightforward to calculate the probability of having generated it. Call $u_1, u_2, \ldots, u_n$ the vertices output in order by one run of the algorithm. While running the algorithm, the set of vertices in the list $\mathcal{S}$ (i.e., those with no incoming edges) depends on the vertices already chosen. At each iteration of the algorithm, the probability of selecting a particular vertex from $\mathcal{S}$ is $1/|\mathcal{S}|$. The overall probability of generating the final linear extension is the product of these probabilities. Therefore, denote by $\mathcal{S}_{i,\{u_{1:i}\}}$ the set of vertices in the list $\mathcal{S}$ after selecting the i vertices $u_1, u_2, \ldots, u_i$. The probability of generating a specific linear extension $u_1, u_2, \ldots, u_n$ from the algorithm is then

$$\pi_0(\mathbf{u}) = \prod_{i=0}^{n-1} \frac{1}{|\mathcal{S}_{i,\{u_{1:i}\}}|}. \tag{3.1}$$

We now include an illustrative example and also comment in passing on the limitations of such simple importance sampling schemes and on certain improvements that are possible via further modifications of randomized topological sorting. As for an illustrative example, suppose that we have a partial order $a \prec b \prec c \prec d \prec e$ over the set $\{a, b, c, d, e, f\}$. The list initially is $\mathcal{S}_{0,\{\}} = \{a, f\}$. If we output f as the first element, then the only element left in the list will be $\mathcal{S}_{1,\{f\}} = \{a\}$. If we output a as the first element, then the list will contain $\mathcal{S}_{1,\{a\}} = \{b, f\}$. We can thus calculate the probabilities of all the possible linear extensions as shown in Table 1. While every linear extension can be generated from this algorithm, we see that the probabilities can be far from the uniform distribution.

TABLE 1
Probability computations using equation (3.1) for an illustrative example

Linear extension	$\prod_{i=0}^{n-1} \lvert\mathcal{S}_{i,\{u_{1:i}\}}\rvert^{-1}$	**Probability $\pi_0(x)$**
f, a, b, c, d, e	$\frac{1}{2}\frac{1}{1}\frac{1}{1}\frac{1}{1}\frac{1}{1}$	$\frac{1}{2}$
a, f, b, c, d, e	$\frac{1}{2}\frac{1}{2}\frac{1}{1}\frac{1}{1}\frac{1}{1}$	$\frac{1}{4}$
a, b, f, c, d, e	$\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{1}\frac{1}{1}$	$\frac{1}{8}$
a, b, c, f, d, e	$\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{1}$	$\frac{1}{16}$
a, b, c, d, f, e	$\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}$	$\frac{1}{32}$
a, b, c, d, e, f	$\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}$	$\frac{1}{32}$

3.4 Importance Sampling via Look Ahead Schemes

The algorithm developed in the previous subsection can be modified with a "look-ahead" strategy to bring the probabilities of generating each linear extension closer to the uniform distribution. When selecting a vertex from the list $\mathcal{S}$, instead of simply selecting each vertex uniformly at random, we can try sampling a vertex with probability proportional to the number of linear extensions possible if we select that vertex. Typically, computing the number of linear extensions is a $\#P$-complete counting problem, and so we cannot solve this exactly. However, when the size of $\mathcal{S}$ is large, there will be more linear extensions on average. Thus, we can locally look ahead to approximate the number of the linear extensions based on the size of the list $\mathcal{S}$ at future iterations and use this information in sampling vertices from $\mathcal{S}$.

In the one-step look ahead scheme, once we have already selected vertices $u_1, u_2, \ldots, u_i$, we calculate the size of the queue $\mathcal{S}_{i+1,\{u_{1:i},s\}}$, for each vertex $s \in \mathcal{S}_{i,\{u_{1:i}\}}$ and select a vertex with probability proportional to that size. It then follows that the probability of generating a linear extension $u_1, u_2, \ldots, u_n$ using the one-step look ahead scheme is equal to

$$\pi_0(\mathbf{u}) = \prod_{i=0}^{n-1} \frac{|\mathcal{S}_{i+1,\{u_{1:i+1}\}}|}{\sum_{s \in \mathcal{S}_{i,\{u_{1:i}\}}} |\mathcal{S}_{i+1,\{u_{1:i},s\}}|}. \tag{3.2}$$

For our illustrative example in Section 3.3, the probability of selecting a as the first element in a one step

TABLE 2
Probability computations using equation (3.2) for an illustrative example

Linear extension	**Equation (3.2)**	**Probability $\pi_0(x)$**
f, a, b, c, d, e	$\frac{1}{3}\frac{1}{1}\frac{1}{1}\frac{1}{1}\frac{1}{1}$	$\frac{1}{3}$
a, f, b, c, d, e	$\frac{2}{3}\frac{1}{3}\frac{1}{1}\frac{1}{1}\frac{1}{1}$	$\frac{2}{9}$
a, b, f, c, d, e	$\frac{2}{3}\frac{2}{3}\frac{1}{3}\frac{1}{1}\frac{1}{1}$	$\frac{4}{27}$
a, b, c, f, d, e	$\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{1}{3}\frac{1}{1}$	$\frac{8}{81}$
a, b, c, d, f, e	$\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{1}{2}$	$\frac{8}{81}$
a, b, c, d, e, f	$\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{1}{2}$	$\frac{8}{81}$

look ahead scheme is now

$$\frac{|\mathcal{S}_{1,\{a\}}|}{\sum_{s\in\{a,f\}}|\mathcal{S}_{1,\{s\}}|}=\frac{2}{3}.$$

Table 2 continues with such calculations and displays the probabilities of selecting each of the linear extensions according to this one-step look ahead algorithm. This probabilities are significantly closer to the uniform distribution than those displayed in Table 1.

We now discuss implementation details and running time of the "look ahead" scheme. First, note that, for a one-step look ahead algorithm, for each element $s \in \mathcal{S}$, we need to calculate the number of elements in the list after that element were chosen. In particular, any element in the adjacency list of $\mathcal{S}$ will be added to $\mathcal{S}$ if its current number of incoming edges is equal to 1. This requires looking at each edge at most once. Then, instead of picking a vertex from the list at random, we pick it proportional to the number of vertices in the list if that vertex were selected. Each edge may now be processed multiple times in the algorithm, and the running time will be $O(|\mathcal{V}| + |\mathcal{E}|^2)$. Similarly, if we look ahead multiple steps, we would need additional storage to keep track of the different lists at different look-aheads, and the running time will be at least $O(|\mathcal{V}|+|\mathcal{E}|^k)$ for k look aheads. Thus, the running time for the algorithm grows polynomially with the number of look-ahead steps. In practice, one would need to balance the convergence speed up from sampling closer to the target uniform distribution with the increased running time of the sampling algorithm.

In general, the maximum number of edges in a directed acyclic graph is the number of forward edges in the total order, which is $O(|\mathcal{V}|^2)$. Another way we can improve the efficiency of the algorithm is by reducing the number of edges in the partial order. Thus, for large graphs, we could first build the *transitive reduction* of the graph, that is, a graph with as few edges as possible but that has the same reachability relation as the given graph. The transitive reduction consists of a minimal set of edges such that the transitive closure is equal to the original relation; see, for example, Aho, Garey and Ullman (1972). For acyclic graphs, the time to build the transitive reduction is $O(|\mathcal{E}||\mathcal{V}|)$, which is larger than the topological sorting algorithm. The maximum number of edges in the reduced graph can still be $O(|\mathcal{V}|^2)$, but in practice, it will probably be worthwhile to pay an upfront cost of $O(|\mathcal{E}||\mathcal{V}|)$ for the per sample savings.

3.5 Gibbs Sampling Methods

Gibbs sampling (see, e.g., Geman and Geman, 1984; Besag et al., 1995) is applicable when the joint distribution of a set of random variables is difficult to sample from directly, but the conditional distributions of each variable given the others are known and are easier to sample from. Let $\mathcal{U}$ be the set of all linear extensions $\mathbf{u} = \{u_1 < u_2 < \cdots < u_n\}$ of the given partial order $(\mathcal{V}, \prec)$. Let $\mathbf{A}$ denote the adjacency matrix corresponding to the given partial order $(\mathcal{V}, \prec)$. Now, for each $\mathbf{u} \in \mathcal{U}$, let $\mathbf{r_u}$ denote the rank vector of the components of $\mathbf{u}$. We then define $\mathcal{R}$ to be the set of all possible rank vectors: $\{\mathbf{r_u} : \mathbf{u} \in \mathcal{U}\}$. The basic idea of this Gibbs sampling method is as follows. Rather than sampling the linear extensions from $\mathcal{U}$, we sample either the conditional distribution of ranks from $\mathcal{R}$, or the conditional distribution of each element of a random vector $\mathbf{h} = (h_1, \dots, h_n)^T$ that has a uniform distribution on the set

$$\Delta = \{\mathbf{h} : h_i \in (0, 1), h_i < h_j \text{ iff } \mathbf{A}[i, j] = 1\}.$$

In the later case, it trivially follows that the rank vectors $\mathbf{r}$ of $\mathbf{h}$ are uniform on $\mathcal{R}$. To see this, note that any realization of the rank vector $\mathbf{r}$ corresponds to the event $\{h_{\sigma(1)} < \cdots < h_{\sigma(n)}\}$, where $\{\sigma(1), \dots, \sigma(n)\}$ is a permutation of $\{1, \dots, n\}$ such that $\sigma(r_i) = i$ for $i = 1, \dots, n$. In other words, σ is the inverse permutation of $\mathbf{r}$. However, since $\mathbf{h}$ is uniform on Δ, the event $\{h_{\sigma(1)} < \cdots < h_{\sigma(n)}\}$ is equally likely for any permissible permutations $\{\sigma(1), \dots, \sigma(n)\}$ of $\{1, \dots, n\}$. This implies

$$\Pr(\mathbf{r}) = \Pr(h_{\sigma(1)} < \cdots < h_{\sigma(n)}) = 1/|\mathcal{R}|.$$

The above generalizes the fact that ranks of n i.i.d. uniform random variable generate a uniform permutation of $\{1, \dots, n\}$. However, unlike i.i.d. uniform random variables on $(0, 1)$, there is no easy way to draw random vectors $\mathbf{h}$ uniformly from Δ. To overcome this challenge, we now derive the full conditional density of h_i given h_j, $j \neq i$ and develop the following Gibbs sampling method. First, define

$$\mathcal{I}_i = \{j : j \neq i, \mathbf{A}[i, j] = 1, j = 1, 2, \dots, n\}$$

and

$$\mathcal{J}_j = \{i : i \neq j, \mathbf{A}[i, j] = 1, i = 1, 2, \dots, n\}.$$

It then follows that

$$h_i < h_j \quad \text{if } j \in \mathcal{I}_i, \quad h_i > h_j \quad \text{if } j \in \mathcal{J}_i.$$

Thus, it is trivial that

$$h_i \mid h_j, \quad j \neq i \sim \text{uniform}(h_i^-, h_i^+),$$

where

$$h_i^+ = \min\{\{h_j : j \in \mathcal{I}_i\} \cup \{1\}\},$$
$$h_i^- = \max\{\{h_j : j \in \mathcal{J}_i\} \cup \{0\}\}.$$

We can thus update the values of h_i sequentially using a random permutation to ensure that resulting Gibbs sample gives rise to a time-reversible Markov chain with limiting distribution uniform on Δ.

On the other hand, if the rank vector $\mathbf{r}$ is distributed uniformly over $\mathcal{R}$, it follows trivially that the full conditional distributions are

$$r_i \mid r_j, \quad j \neq i \sim \text{uniform on}\{r_i^-, r_i^- + 1, \dots, r_i^+\},$$

where

$$r_i^+ = \min\{\{r_j : j \in \mathcal{I}_i\} \cup \{n\}\},$$
$$r_i^- = \max\{\{r_j : j \in \mathcal{J}_i\} \cup \{1\}\}.$$

We can thus update the values of r_i sequentially using a random permutation to ensure that resulting Gibbs sample gives rise to a time-reversible Markov chain with limiting distribution uniform on $\mathcal{R}$.

Overall, the above Gibbs sampling methods are very easy to implement, as samples from the full conditionals can be drawn using i.i.d. uniform random variables. The computational costs of deriving $\mathcal{I}_i$ and $\mathcal{J}_i$ is at most $O(\max\{n, |\mathcal{E}|\})$. Since maxima and minima can be computed efficiently using the binary search tree, it is not difficult to see that one cycle of this Gibbs sampling algorithm would require at most $O(\max\{n \log n, |\mathcal{E}| \log |\mathcal{E}|\})$ steps. The derivation of mixing time for the above Gibbs sampling will be a matter of future work.

3.6 Random Walks on Linear Extensions

We now describe specific random walk moves that can be used in the Metropolis–Hastings algorithm to generate uniform linear extensions in quick succession. The primary focus here is to obtain via forward and backward sampling, an exchangeable random sample $t^{(1)}, \dots, t^{(B)}$ from the conditional distribution of a test statistic t given the partially observed data. To this end, we call two linear extensions, namely, $\mathbf{u}$ and $\mathbf{u}'$, of $(\mathcal{V}, \prec)$ neighbors if $\mathbf{u}$ can be obtained by a single transposition of two consecutive elements in $\mathbf{u}'$. We can then construct Markov chain sampling by implementing the simple random walk on linear extensions as follows. When the Markov chain is at a linear extension $\mathbf{u} = \{u_1 < u_2 < \cdots < u_n\}$ of the given partial order $(\mathcal{V}, \prec)$, it chooses an integer i between 1 and $2n - 2$ at random. If $i \leq n - 1$ and there is no directed edge from u_i to u_{i+1} in $\mathcal{E}$, then it swaps u_i and u_{i+1} to move to a new linear extension $\mathbf{u}'$. Otherwise, it stays at $\mathbf{u}$. Let $\{U_k\}$ denote this Markov chain on $\mathcal{U}$. Karzanov and Khachiyan (1991) showed that U_k is an irreducible, aperiodic time-reversible doubly stochastic random walk on the set of all linear extensions of $(\mathcal{V}, \prec)$. Following Wilson (2004), it is well established that the above Karzanov–Khachiyan Markov chain mixes in time $O(n^3 \log n)$. In fact $(4/\pi^2 + o(1))n^3 \log n$ provides an upper bound for the separation distance in terms of total variation norm.

3.7 Random-Walk Based Perfect MCMC Sampling

We now make use of the above construction of random walks on linear extensions and present a perfect MCMC sampling algorithm to obtain genuine i.i.d. uniform draws from the set of possible linear extensions. To this end, Propp and Wilson's (1996) basic idea of the perfect MCMC sampling is as follows. In order to sample from the target distribution $\pi(x)$, $x \in \mathcal{X}$, where $\mathcal{X}$ is finite, we first construct an ergodic Markov chain X_k with state space $\mathcal{X}$ transition probability matrix $\mathbf{P}$, and limiting stationary distribution π. The construction is done by using a deterministic mapping ψ and a stream of i.i.d. random variables ξ_k or, equivalently, by an i.i.d. sequence of random mappings Γ_k such that

$$X_{k+1} = \psi(X_k, \xi_{k+1}) = \Gamma_{k+1}(X_k).$$

As for example, if $\mathcal{X} = \{1, 2, \dots, n\}$, we can take ξ_k to be uniform(0, 1) and

$$\psi(x, \xi_{k+1}) = x'$$
$$\iff \quad \sum_{i=1}^{x'-1} \mathbf{P}[x, i] < \xi_{k+1} \leq \sum_{i=1}^{x'} \mathbf{P}[x, i].$$

The primary advantage of this construction is that it allows us to mathematically tract the trajectories of the Markov chain when run from different initial states and helps us determine when entire state space collapses to a single state. The perfect sampling algorithm embraces running X_k backward in time till coalescence occurs. Specifically, let

$$W_k = \Gamma_0 \circ \Gamma_{-1} \circ \cdots \circ \Gamma_{-k},$$

and let τ denote the backward coalescence time

$$\tau = \min\{j \geq 1 : W_j \text{ maps } \mathcal{X} \text{ to a single state}\}.$$

It then follows that

$$\Pr(\tau < \infty) = 1, \quad W_\tau \sim \pi.$$

Although the above description of a perfect MCMC algorithm looks simple and is easy to understand, its numerical implementation is not so easy. In particular, if the state space $\mathcal{X}$ is large and complex, as is in our case, checking whether coalescence has occurred presents an enormous computational challenge. Thus, further steps such as the construction of monotone couplings or smart bounding chains are required to manage computations; see, for example, Propp and Wilson (1996) and Huber (2004). Here, we briefly present the bounding chain methods that detect coalescence without any condition of monotonicity and that usually work in general complex spaces. The basic idea is again very simple. Applying the same stream of random variables ξ_k that were used in X_k, we construct another Markov chain B_k such that

$$B_{k+1} = \Psi(B_k, \xi_{k+1}),$$

for some deterministic mapping Ψ. Furthermore, the construction is such that B_k takes values on the subsets of $\mathcal{X}$, $X_{k+1} \in B_{k+1}$ for all initial states in $\mathcal{X}$ whenever $X_k \in B_k$ for all initial states in $\mathcal{X}$, and it is easy to verify whether B_k has cardinality one. The success of this method depends on finding a simplified form of Ψ so that it is easy to run bounding chains B_k than running X_k from all states and detect coalescence.

Following Huber (2006), we now present a way to construct random mappings Γ_k and bounding chains B_k for the random walk moves U_k on the linear extensions of a given finite partial order. First, note that U_k can be constructed as

$$U_{k+1} = \psi(U_k, I_{k+1}, \delta_{k+1}),$$

where I_k are i.i.d. Uniform random variables on $\{1, 2, \ldots, n-1\}$, δ_k are a Bernoulli process obtained by tossing a fair coin, and

$$\psi(\mathbf{u}, i, \delta) = \begin{cases} \mathbf{u} & \text{if } \delta = 0 \text{ or } u_i \prec u_{i+1}, \\ \mathbf{E}_{i,i+1}\mathbf{u} & \text{otherwise,} \end{cases}$$

where $\mathbf{E}_{i,i'}$ is the permutation matrix that swaps entries i and i' in a vector of length n.

The construction of the bounding chain for U_k is more involved. Let R_k denote the rankings of the nodes in $\mathcal{V}$ according to U_k. Furthermore, the nodes of $(\mathcal{V}, \mathcal{E})$ are so arranged that $\{1, \ldots, n\}$ is a valid ranking of the nodes in $\mathcal{V}$. It is trivial that there is a one-to-one and onto relationship between R_k and U_k. In particular, if we can index and sort, we can rank and vice-versa. However, Huber (2006) noted that it is easier to bound and bookkeep the ranking vector R_k rather than the chain U_k. The corresponding bounding chain B_k takes the form of $\{B_{1,k}, B_{2,k}\}$, where $B_{1,k}$ is a cursor and $B_{2,k}$ is a list of n numbers that maintains the upper bounds of the rankings R_k when U_k is run from all possible initial states [i.e., all possible linear extensions of $(\mathcal{V}, \mathcal{E})$]. The algorithm starts with initial states $B_{1,0} = 1$ and $B_{2,0} = \{n, n, \ldots, n\}$. It then updates B_k using the following probability rules:

$$\begin{aligned} &\Pr(B_{1,k+1}, B_{2,k+1} \mid B_{1,k}, B_{2,k}) \\ &\quad = \Pr(B_{2,k+1} \mid B_{1,t}, B_{2,k}) \Pr(B_{1,k+1} \mid B_{1,k}, B_{2,k+1}), \end{aligned}$$

where

$$B_{2,k+1} = \Psi(B_{1,k}, B_{2,k}, I_{k+1}, \delta_{k+1})$$

with

$$\begin{aligned} &\Psi(l, \mathbf{b}, i, 1) \\ &\quad = \begin{cases} \mathbf{E}_{j,j'}\mathbf{b} \\ \quad \text{if } \exists j, j' \leq l : r_j = i, r_{j'} = i+1, (j, j') \notin \mathcal{E}, \\ \mathbf{b} + \mathbf{e}_j \\ \quad \text{if } \exists j \leq l : r_j = i, \nexists j' \leq i : r_{j'} = i+1, \\ \mathbf{b} - \mathbf{e}_{j'} \\ \quad \text{if } \exists j' \leq l : r_j = i+1, \nexists j \leq i : r_j = i, \end{cases} \\ &\Psi(l, \mathbf{r}, i, 0) = \mathbf{r}, \end{aligned}$$

and

$$\Pr(B_{1,k+1} = l+1 \mid B_{1,k} = l, B_{2,k+1} = \mathbf{r}) = 1$$
$$\text{if } r_j < n \text{ for all } j \leq l.$$

In the equation involving $\Psi(l, \mathbf{b}, i, 1)$, the expression $\exists j \leq l : r_j = i, \nexists j' \leq i : r_{j'} = i+1$ means there exists a $j \leq l$ such that $r_j = i$ and there does not exist a $j' \leq i$ such that $r_{j'} = i+1$ and so on. Furthermore, $\mathbf{e}_j$ denote the vector with a 1 in the jth entry and 0's elsewhere. The cursor $B_{1,k}$ is such that for all $1 \leq j, j' \leq B_{1,k}$, we always have

$$\begin{aligned} B_{2,k}[j] &\neq B_{2,k}[j'] \quad \text{if } j \neq j', \quad \text{and} \\ B_{2,k}[j] &< B_{2,k}[j'] \quad \text{if } v_j \prec v_{j'}. \end{aligned}$$

Thus, Ψ allows us to bookkeep the upper bounds of the ranking R_k of each node in $\mathcal{V}$ and the cursor $B_{1,k}$ allows us to tighten these upper bounds and helps us detect coalescence by checking when $B_{1,k}$ becomes n. Overall the algorithm is run backward in time and run until coalescence is detected. Huber (2006) proved that the expected running time of this perfect MCMC algorithm is at most $O(n^3 \log n)$. In fact, $(16/\pi^2 + o(1))n^3 \log n$ provides an upper bound for this expected running time.

Finally, the above perfect MCMC algorithm also allows us to draw exact uniform sample from Δ. To see this, let $\mathbf{r}$ be drawn uniformly from $\mathcal{R}$ and let $h_{(1)} < h_{(2)} < \cdots < h_{(n)}$ be the order statistics of n i.i.d. uniform(0, 1) random variables. It is then immediate that the random variables

$$h'_i = h_{(r_i)}, \quad i = 1, \dots, n,$$

are a uniform draw from Δ. Thus, as a rudimentary MCMC procedure that does not require an initial "burn-in" part, we can first run the perfect MCMC sampling and generate one random vector $\mathbf{h}$ uniformly from the set Δ; subsequently, we can perform Gibbs sample method discussed in Section 3.5 to draw a valid MCMC sample of rank vectors from $\mathcal{R}$.

4. A SIMULATION EXPERIMENT AND DATA EXAMPLES

4.1 A Simulation Experiment

The purpose of this simulation study is to illustrate how the methods developed in Sections 2 and 3 work in practice. In particular, we consider simulated partially ordered data for which the desired result is known in advance. We then evaluate our methods by constructing an exact rank test and summarizing the results of the test using the predictive distribution of the latent p-value. Specifically, we consider testing the monotonic association between two variables y and z. We take the sample size $n = 20$. Furthermore, we assume that we only know a partial ordering of $z_1, \dots, z_n$. This partial ordering is constructed as follows. For each $i = 1, 2, \dots, n$, we first independently generate two uniform numbers $z_{1,i}$ and $z_{2,i}$ between 0 and 1. We then assume that

$$\min\{z_{1,i}, z_{2,i}\} < z_i \le \max\{z_{1,i}, z_{2,i}\}.$$

The above gives rise to a partial ordering on $\{z_1, \dots, z_n\}$ and the left panel of Figure 2 displays this partial ordering using a directed acrylic graph. Furthermore, there are approximately $O(1.25 \times 10^{12})$ linear extensions to this partial order.

Next, we generate the y variable under the null hypothesis of no monotonic association between two variables y and z. Specifically, we generate $y_1, \dots, y_n$ as independent and identically distributed standard Gaussian random variables. To derive an exact test, we next generate $M = 10{,}000$ uniform linear extensions $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(M)}$ of $\{z_1, \dots, z_n\}$ using the sampling algorithms discussed in Section 3. For each $\mathbf{u}^{(i)}$, we then compute the sample rank correlation $\rho^{(i)}$ between $\mathbf{y}$ and $\mathbf{u}^{(i)}$. We also compute its standardized version $t_1^{(i)}$ and latent p-value $p(t_1^{(i)}) = 2(1 - \Phi(t_1^{(i)}))$. Using $p(t_1^{(1)}), \dots, p(t_1^{(M)})$, we finally compute the exact empirical cumulative predictive distribution of the latent p-value.

We repeat the above procedure $K = 100$ times. The resulting empirical cumulative predictive distributions

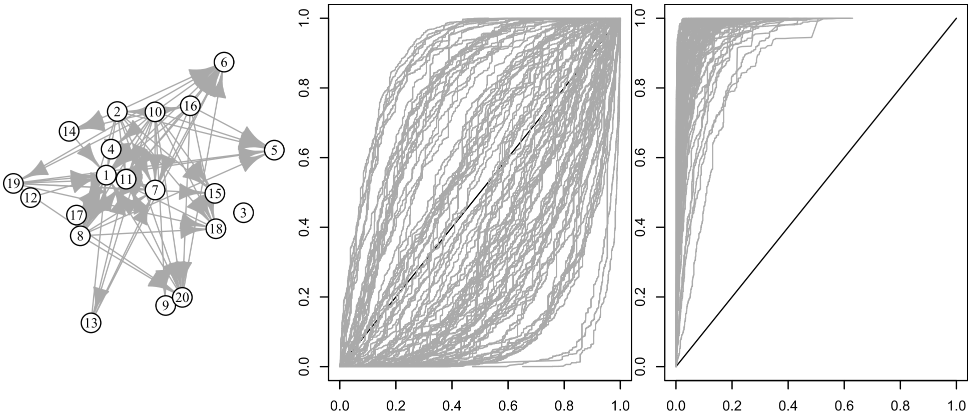

FIG. 2. (a) *Partial order of* $z_1, \dots, z_n$, (b) *a random draws of cumulative predictive distributions of latent* p*-values under a null model and* (c) *a random draws of cumulative predictive null distributions of latent* p*-values under an alternative model.*

are plotted in the middle panel of Figure 2. Since these simulations are repeated under the null hypothesis, we expect the average of these resulting cumulative predictive distributions to match the cumulative distribution of a uniform$(0, 1)$ random variable. Indeed, we find that there is an excellent agreement between the average of the cumulative predictive distributions and the diagonal line $y = x$. For example, at $x = 0.05, 0.25, 0.50$ and 0.75, these averages are respectively 0.04, 0.23, 0.45 and 0.70, which agree well with the theoretical average values within the statistical error. Furthermore, it can be seen that an overwhelming fraction of these cumulative distributions put almost all of their mass within $(0.1, 1)$, and only an extremely small fraction of these cumulative distributions add some mass within $(0, 0.01)$. Thus, these predictive distributions of the latent p-values are correctly indicating a lack of evidence against the null hypothesis.

Next, we examine how the method works when the data is generated under an alternative model. To this end, we generate $y_1, \dots, y_n$ as independent Gaussian random variables with mean $\frac{1}{2}\omega(z_{1,i} + z_{2,i} - 1)/\kappa$ and variance $1 - \omega^2$, where we set $\omega = 0.75$ and $\kappa^2 = \mathrm{var}(z_{1,i}) = 1/12$. As in the previous case, we then generate $M = 10{,}000$ uniform linear extensions of $\{z_1, \dots, z_n\}$, compute the sample rank correlation between $\mathbf{y}$ and generated linear extensions, compute the corresponding latent p-values and obtain the exact empirical cumulative predictive distribution of these latent p-values. (This procedure is the repeated $K = 100$ times.) The resulting empirical cumulative predictive distributions are then plotted in the right panel of Figure 2. Here, simulations are repeated under the alternative hypothesis. Thus, we expect the resulting cumulative predictive distributions to be stochastically much smaller than the cumulative distribution of a uniform$(0, 1)$ random variable. Indeed, we see that about 75% of these cumulative distributions are now putting, at least, 85% of their mass below 0.05. Thus, these cumulative predictive distributions are correctly summarizing the evidence against the null hypothesis. This provides further credence that the exact tests based on predictive distributions of latent p-values are credible and reliable.

4.2 Light Preference Behavior in Fruit Flies

We now return to the fruit fly example introduced in Section 1. The purpose of the fruit flies experiments (see, e.g., Kain, Stokes and de Bivort, 2012 for further details) was to study the behavioral variability or the personality in fruit flies. Typically, fruit flies move toward the light when startled, but different fruit flies also exhibit variations in their respective mean phototactic behaviors. These idiosyncrasies are believed to be noninheritable, but last the lifetime of the flies, and constitute a form of personality. One specific objective was to identify and understand the underlying neurobiological factors such as serotonin that contribute to these individual-to-individual noninheritable behavioral differences, including those from identically reared, isogenic strains.

TABLE 3
Experimental manipulations and measurements on heterogeneity in light-preference behavior. Here, we have 14 different combinations of manipulations, with a_0 representing a baseline serotonin signal, and $a_1, a_2, \dots, a_8$ indicating eight different elementary manipulations

Type (v_i)	**Serotonin manipulations**	**Measurements on light-preference**
1	$a_0 - a_1$	−0.4007
2	$a_0 - a_2$	−0.2325
3	$a_0 - a_3$	−0.1599
4	$a_0 - a_1 - a_2$	−0.1341
5	$a_0 - a_1 - a_4$	−0.3976
6	$a_0 - a_1 + a_5$	−0.2857
7	$a_0 - a_3 - a_4$	−0.1105
8	$a_0 - a_3 + a_5$	−0.0755
9	$a_0 - a_6$	−0.3271
10	$a_0 - a_6 - a_7$	−0.4000
11	$a_0 + a_8$	0.0758
12	$a_0 + a_5$	0.0073
13	$a_0 - a_4$	0.1785
14	a_0	0.0000

Here, we focus on analyzing a set of phototactic personality behavior data displayed in Table 3 and collected at EBG through a randomized experiment. The first column gives the type (genetic mutation) of fruit flies. The second column indicates the actual combinations of serotonin manipulations that are done to the fruit flies. The third column provides the values of the response variable which summarizes an average run of individual flies toward light or darkness when startled and are actually determined by the preference of individual flies in choosing either left or right turns in branching mazes. Our objective is to collect empirical evidence on whether there is a positive association between the serotonin signaling and the phototactic personality.

As mentioned in Section 1, we do not know the actual magnitude of the change in serotonin signaling, but

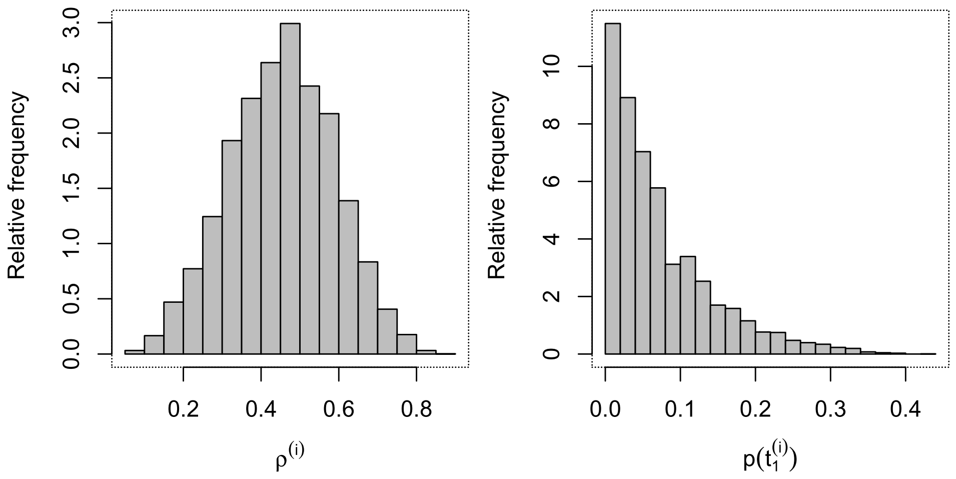

FIG. 3. *Left*: *histogram of Spearman's rank correlation between heterogeneity in light-preference behavior and the experimental manipulations on serotonin signals. Right*: *conditional distribution of latent p-values for testing zero Spearman's rank correlation.*

only some of the directions to which each experimental manipulation affects the serotonin signaling. Furthermore, Figure 1 provides these directions in terms of a partial order $(\mathcal{V}, \mathcal{E})$. As a consequence, we cannot directly apply the sample rank correlation t_1 and compute a standard a p-value here. Instead, we draw uniform linear extensions $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(M)}$, $M = 10{,}000$, of $(\mathcal{V}, \mathcal{E})$ using algorithms discussed in Section 3. For each $\mathbf{u}^{(i)}$, we then compute the sample rank correlation $\rho^{(i)}$ between $\mathbf{y}$ and $\mathbf{u}^{(i)}$. We also compute its standardized version $t_1^{(i)}$ and the corresponding exact empirical p-value $p(t_1^{(i)}) = 1 - \Phi_n(t_1^{(i)})$. Figure 3 plots the histograms of $\rho^{(i)}$ and $p(t_1^{(i)})$ based on a backward-forward Karzanov–Khachiyan MCMC sample. Specifically, this MCMC sample is collected as follows. Using the perfect MCMC method, we draw an initial state $\mathbf{u}^{(1)}$; the Markov chain is then run backward 5000 times to obtain $\mathbf{u}^{(0)}$. From $\mathbf{u}^{(0)}$, we then separately run the chain 5000 steps forward in time to obtain $\mathbf{u}^{(i)}$, for $i = 2, 3, \dots, M$. Since $4n^3 \log n/\pi^2 \approx 2935$, this procedure would approximately draw a uniform sample from $\mathcal{U}$. This sample provides the following statistics. The rank correlation $\rho^{(i)}$s range from 0.042 to 0.842, with median 0.451, first quartile 0.354 and third quartile 0.547. Thus, the overall correlation seems to be strictly positive. However, since the sample size is here very small ($n = 14$), this is likely not significantly strong evidence to reject the null hypothesis of no correlation. Indeed, about 47.5% of $p(t_1^{(i)})$ fall below 0.05, another 25.5% fall between 0.05 and 0.10, whereas only 14.05% of $p(t_1^{(i)})$ fall above 0.15. Overall, these p-values suggest a weak evidence against the null hypothesis. In other words, even though these p-values do not add up to a case for outright rejection, the weight of evidence against the null hypothesis here should not be completely ignored either.

Next, we compare the above MCMC latent p-values with those obtained from the importance sampling, the Gibbs sampling, and the perfect MCMC sampling methods. To this end, Table 4 provides summary statistics of $p(t_1^{(i)})$ obtained from these different sampling methods. Note that the importance sampling algorithm does not draw the linear extensions uniformly, but with probability $\pi_0(\mathbf{u})$ for $\mathbf{u} \in \mathcal{U}$. Therefore, the calculated

TABLE 4
Summary statistics of the conditional distribution of the latent p-value from the fly data for different sampling methods

	Min	**1st quartile**	**Median**	**Mean**	**3rd quartile**	**Max**
Importance sampling	0.0001	0.025	0.054	0.076	0.109	0.441
Random-walk MCMC	0.0001	0.022	0.054	0.074	0.105	0.443
Gibbs sampling	0.0000	0.022	0.054	0.074	0.105	0.422
Perfect MCMC	0.0001	0.022	0.049	0.071	0.098	0.451

p-values based on importance sampling have a corresponding probability distribution associated with them. We thus apply sampling importance resampling to obtain the quantities of the exact predictive distribution of the latent p-values. On the other hand, the Gibbs sampling is implemented as follows. First, we apply perfect MCMC method to draw a random vector $\mathbf{h}^{(1)}$ uniformly on Δ. We then run the Gibbs sampling 1000 steps backward in time to obtain $\mathbf{h}^{(0)}$. From $\mathbf{h}^{(0)}$, we then separately run the chain 1000 steps forward in time to obtain $\mathbf{h}^{(i)}$, for $i = 2, 3, \dots, M$. Following that, we compute ranks and linear extensions based on the exchangeable sample $\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(M)}$. Overall, we see that there is an excellent match among these different sampling methods. The median $p(t_1^{(i)})$ from the importance sampling is 0.054. Random-walk based MCMC sampling, Gibbs sampling and the perfect MCMC sampling, respectively, put this median at 0.54, 0.054 and 0.049. The first quartiles of $p(t_1^{(i)})$ obtained from the importance sampling, the random-walk based MCMC sampling, the Gibbs sampling and the perfect MCMC sampling are all equal to 0.022. However, the third quartiles of $p(t_1^{(i)})$ obtained from these four methods are respectively 0.109, 0.105, 0.105 and 0.098. Overall, the importance sampling method gave a slightly thicker right tail in the distribution of $p(t_1^{(i)})$.

Finally, it is worthwhile to point out that the above exact test results are remarkable given that the sample size is so small, the information is so scant and none of the conventional p-value computations either via certain statistics or pivotal quantities of the partially ordered data would have yielded any better numbers. Although caution is required, these results may be worth pursuing with further well replicated randomized experiments. Note that such replicated experiment would also allow us to estimate the contrasts of different treatment effects and help understand the cross effect of different elementary manipulations.

4.3 Lead in Children's Blood

Rosenbaum (1991, 2002) (see also Morton et al., 1982) provides statistical analysis of lead levels ($\mathbf{y}$) in $n = 33$ children, whose parents were exposed to lead on the job in a battery factory and transmitted to their children an unknown quantity of lead through shared living conditions. They are concerned with the hypothesis that lower lead levels in children's blood are associated with lower parental exposures to lead on the job and higher hygiene standards. In their study, the parental exposure ($\mathbf{z}_1$) is reported into three groups, namely, low, medium and high, and parental hygiene ($\mathbf{z}_2$), up to leaving the factory, is marked into three categories, namely, good, moderately good and poor. The bivariate explanatory variable $(\mathbf{z}_1, \mathbf{z}_2)$ offers indirect information on the quality of lead the parents inadvertently brought home over a long period of time, and the statistical problem is to measure the strength of relationship between $\mathbf{y}$ and $(\mathbf{z}_1, \mathbf{z}_2)$. However, $(\mathbf{z}_1, \mathbf{z}_2)$ assumes a partial ordering in that $(z_{1,j} z_{2,j}) \prec (z_{1,j'} z_{2,j'})$ whenever $z_{1,j} \le z_{1,j'}$ and $z_{2,j} \le z_{2,j'}$. Using certain statistics of the partially ordered data, Rosenbaum (1991, 2002) first uncovered a strong agreement between partially ordered samples of $\mathbf{y}$ and $(\mathbf{z}_1, \mathbf{z}_2)$. In particular, applying the central limit theorem to these statistics, he found that an approximate one-sided significance level was about 0.0007.

Here, we pursue exact test and predict corresponding exact latent p-values for an one sided test. We found that there are approximately 5.67×10^5 linear extensions of $(\mathbf{z}_1, \mathbf{z}_2)$. To derive an exact test, we randomly generate $M = 1000$ linear extensions $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(M)}$ of $(\mathbf{z}_1, \mathbf{z}_2)$ that are consistent with the observed partial order. We then compute the sample Spearman's rank correlation $\rho^{(i)}$ between the children's blood lead level $\mathbf{y}$ and the complete ordering $\mathbf{u}^{(i)}$ of $(\mathbf{z}_1, \mathbf{z}_2)$. We also compute its standardized version $t_1^{(i)}$ and the corresponding exact empirical p-value $p(t_1^{(i)}) = 1 - \Phi_n(t_1^{(i)})$. Figure 4 plots the histograms of $\rho^{(i)}$ and $p(t_1^{(i)})$ based on a backward-forward Karzanov–Khachiyan MCMC sample. Here, the MCMC sample is collected as follows. Using a perfect MCMC sample, we draw an initial state $\mathbf{u}^{(1)}$; the Markov chain is first run backward 100,000 times to obtain $\mathbf{u}^{(0)}$. From $\mathbf{u}^{(0)}$, we then separately run the chain 100,000 steps forward in time to obtain $\mathbf{u}^{(i)}$, for $i = 2, 3, \dots, M$. Note that the mixing time here is roughly $4n^3 \log n/\pi^2 = 50{,}925.64$, which is the reason we run the chain 100,000 time steps in forward and backward directions. In this example, the rank correlations $\rho^{(i)}$ take values between 0.495 and 0.780, with median 0.644, first quartile 0.615 and third quartile 0.675. The corresponding p-values $p(t_1^{(i)})$ range from 10^{-6} to 0.002, with median 4×10^{-5}, first quartile 3×10^{-5} and third quartile 1×10^{-4}. Overall, an overwhelming fraction of $p(t_1^{(i)})$s are less than 0.0007. Thus, the exact predictive distribution of the latent p-value suggests a stronger evidence against the null hypothesis than what was reported by Rosenbaum (1991, 2002).

We also compare random-walk based MCMC latent p-values with those obtained from the importance sam-

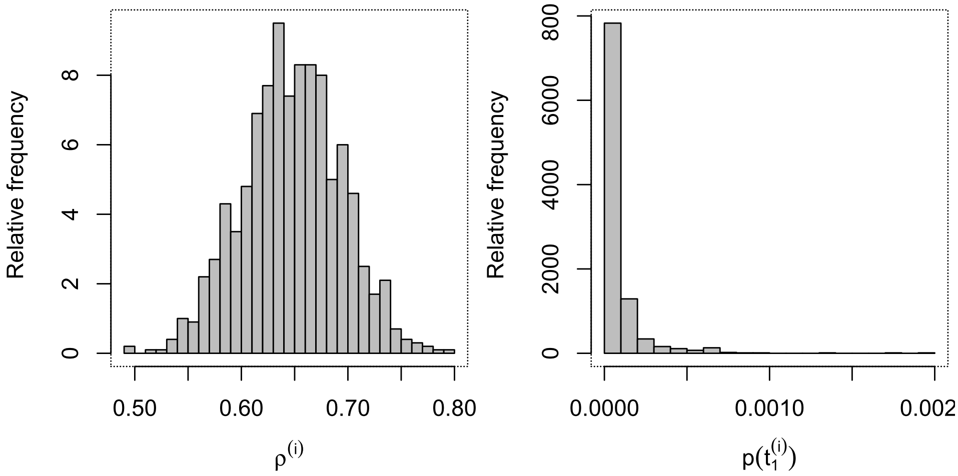

FIG. 4. *Left*: *histogram of Spearman's rank correlation between children's blood lead level and the complete ordering of the categories. Right*: *conditional distribution of latent p-values for testing zero Spearman's rank correlation.*

pling, Gibbs sampling and the perfect MCMC sampling methods. In particular, Table 5 has the same format as Table 4 and it provides summary statistics of $p(t_1^{(i)})$ obtained from these different sampling methods. Overall, we see that there is a good agreement among the four different sampling methods. The median $p(t_1^{(i)})$ from the MCMC sampling is 4×10^{-5}. The importance sampling, the Gibbs sampling and the perfect MCMC sampling methods respectively give this median 6×10^{-5}, 3×10^{-5} and 6×10^{-5}. The first quartiles of $p(t_1^{(i)})$ obtained from the importance sampling, random-walk MCMC, Gibbs and perfect MCMC sampling are respectively equal to 3×10^{-5}, 3×10^{-5}, 10^{-5} and 4×10^{-5}. On the other hand, the third quartiles of $p(t_1^{(i)})$ obtained from the these methods are respectively 10^{-4}, 10^{-4}, 8×10^{-5} and 9×10^{-5}. Overall, there is excellent agreement among different sampling methods.

4.4 Breast Cosmesis Study

This dataset, analyzed by Finkelstein (1986) and others, records the time until the appearance of breast retraction for two groups of early breast cancer patients. The first group consists of 46 patients who received radiotherapy alone, whereas the second group with 48 patients received both the radiotherapy and the adjuvant chemotherapy. Each patient was monitored for the cosmetic effects of their therapy and underwent medical follow-ups for breast retraction once in every 4 to 6 months for a period of time. Missed visits and returning with a changed status took place for certain patients who were monitored weekly or monthly for a response, but like others, we also assume that the effect of such missed visits is negligible. The primary focus is to compare the patients who received adjuvant chemotherapy to those who did not and to determine whether chemotherapy affects the rate of deterioration of the cosmetic state. To this end, we assume that failure times follow a Cox promotional hazard model (2.9) with $z_i = 0$ if the ith patient belongs to the first group and 1 otherwise, and test whether $\beta = 0$ against the alternative $\beta > 0$.

From the observed interval data, we first compute the partial order of the failure times. We found that

TABLE 5
Summary statistics of the conditional distribution of the latent p-value from the lead data for different sampling methods

	1st quartile	Median	3rd quartile	Max
Importance sampling	3×10^{-5}	6×10^{-5}	1×10^{-4}	1.53×10^{-3}
Random-walk MCMC	3×10^{-5}	4×10^{-5}	1×10^{-4}	1.93×10^{-3}
Gibbs sampling	1×10^{-5}	3×10^{-5}	8×10^{-5}	2.61×10^{-3}
Perfect MCMC	4×10^{-5}	6×10^{-5}	9×10^{-5}	1.52×10^{-3}

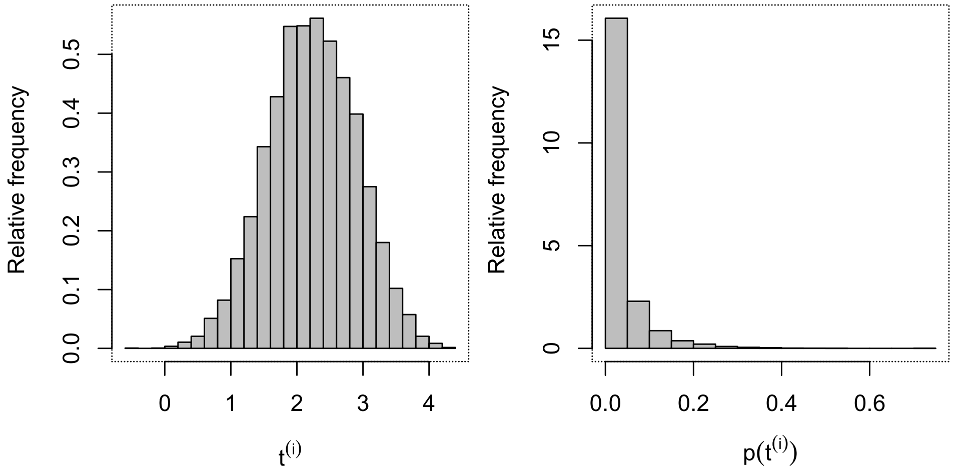

FIG. 5. *Left: histogram of the test statistics t from breast cosmesis data. Right: histogram of the conditional distribution of latent p-values based on the statistics t.*

there are approximately 2.4×10^{92} linear extensions that are consistent with the observed partial order. Using MCMC, we then generate $M = 10{,}000$ of these linear extensions to the observed partial order. The MCMC sample is obtained as follows. First, we apply the perfect MCMC sampling method to obtain a starting state $\mathbf{u}^{(1)}$. In this example, $n = 94$ and the expected running time of the perfect MCMC sampling method is about $16n^3 \log n/\pi^2$ is about 6,117,510 time steps. We then run the random-walk MCMC starting from $\mathbf{u}^{(1)}$ 100,000 time steps backward to obtain $\mathbf{u}^{(0)}$. From $\mathbf{u}^{(0)}$, we then separately run the random walk 100,000 steps forward in time to obtain $\mathbf{u}^{(i)}$, for $i = 2, \dots, M = 10{,}000$. These linear extensions then give rise to M random full rank vectors $\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \dots, \mathbf{r}^{(M)}$ of failure times. Applying the Survival R Cran package function "coxph" to the full rank vector $\mathbf{r}^{(i)}$, we next maximize (2.10), obtain the corresponding $\widehat{\beta}^{(i)}$, the standard error of $\widehat{\beta}^{(i)}$ and finally the value of the test statistic $t^{(i)} = \widehat{\beta}^{(i)}/se(\widehat{\beta}^{(i)})$. Figure 5 provides two histograms. The first one is based on the predicted values of the test statistics $t^{(i)}$. The second one corresponds to the histogram of one-sided latent exact p-values $p(t^{(i)})$.

Following Finkelstein (1986), several approximate rank test procedures have been developed for this problem and applied to this particular data set. See, for example, Fay and Shaw (2010) who report various two-sample z-scores and the corresponding p-values for this set of breast cosmesis data. Most of these z-scores range between 2.17 to 2.70 resulting in various conventional p-values between 0.03 and 0.007. In the MCMC run, we found that the test statistics $t^{(i)}$ range from -0.556 to 4.373 with median 2.231 and the first and the third quartile respectively equal to 1.772 and 2.703. The corresponding latent p-values $p(t^{(i)})$ range from 10^{-6} to 0.707. Furthermore, the first, second and the third quartiles of $p(t^{(i)})$ are respectively 0.003, 0.013 and 0.039. Thus, about 80.34% of the latent p-values are below 0.05, whereas only 3.87% fall above 0.15. Overall, these numbers consolidate the evidence obtained by Finkelstein (1986) and others on the effect of adjuvant chemotherapy on the rate of deterioration of the cosmetic state of breast cancer patients. At the same time, we must also recognize that the predictive distribution of the latent p-value here is well spread out and not concentrated around a point. Thus, it cannot be approximated by a single number. In other words, in this example, a conventional p-value based on the asymptotic null distribution of a test statistic is not appropriate (and can be misleading) in assessing the strength of the evidence against the null hypothesis.

As was the case with other examples, different MCMC sampling methods also provide consistent results here. These results are summarized in Table 6. It can be seen that, in the perfect MCMC run, the latent p-values $p(t^{(i)})$ ranges from 10^{-5} to 0.55, with the first, second and the third quartiles of $p(t^{(i)})$ being respectively 0.0035, 0.0130 and 0.0393. Furthermore, in the perfect MCMC run, about 80.04% of the latent p-values fall below 0.05 and only 3.52% fall above 0.15. We also get very similar numbers using the Gibbs sampling methods. However, it is useful to mention that the basic importance sampling algorithms in Section 3 did not work in this example. In particular, every time we draw an importance sample, we found that one of the weights is unusually large (i.e., very close to one)

TABLE 6

Summary statistics of the exact conditional distribution of the latent p-value from the breast cosmoses data using different sampling methods

	1st quartile	Median	Mean	3rd quartile	Max
Random-walk MCMC	0.0036	0.0133	0.0327	0.0389	0.7073
Gibbs sampling	0.0037	0.0137	0.0337	0.0405	0.5846
Perfect MCMC	0.0035	0.0130	0.0324	0.0393	0.5487

among the $\ell(\mathbf{u}^{(i)})$. This problem of few unusually large weights did not mitigate when we tried to implement the one-step look ahead scheme. To rectify this situation, we thus need the sampling distribution $\pi_0(\mathbf{u})$ to be much closer to the target distribution $\pi(\mathbf{u})$. It is likely that a k-steps look ahead scheme would work for some suitably chosen value of k, but we did not pursue its implementation because of the increase in computational cost.

5. DISCUSSION

To summarize, we present some rank test procedures from partially ordered data that arise in various biological and environmental applications. The procedures are distribution-free in that we assume no additional knowledge on our part regarding the form of the underlying probability distribution function of the random variables. The exact null distributions of these rank tests often lead to complex combinatorial problems. However, we present several sampling based methods that allow us to draw samples from the exact predictive distributions of latent p-values. The procedures apply for small to moderately large sample sizes and this adds practical importance when fewer data are available, and asymptotic results are inaccurate. Finally, one simulation study and three examples, two on rank correlations and one on interval censored data demonstrate the potential use of our methods.

In the above context, we must reiterate the practical benefits of collecting empirical evidence through computing predictive distribution of latent p values. In classical statistics, empirical evidence is often summarized through conventional p-values. These p-values are probability under the null distribution of obtaining a test statistic that equals to or is more extreme than what was actually observed. Thus, conventional p-values are just numbers. In contrast, the predictive distributions of the latent p-value are probability distributions. However, in simulations and data examples, we have seen that the histograms of these predictive distributions are all well spread out and not concentrated around a point. This suggests that we better not approximate them by single numbers. In other words, it is important to recognize that there is a qualitative difference in the answers that we obtained in the simulations and the data examples through computing the predictive distributions of the latent p-values. Put another way, in these examples, conventional p-values are no match for predictive distributions of the latent p-values. Furthermore, under regularity conditions, as the sample size increases to infinity, we expect the predictive distributions of a latent p-value to converge to a conventional p-value. Thus, had the sample size been really large, we would be seeing more agreement between predictive p values and corresponding conventional p-values. However, in many biological and environmental applications, the sample size may not be very large and asymptotic results can be inaccurate. Thus, in a range of applications where the sample size is small or moderately large, we will greatly benefit by computing the predictive distributions of the latent p-values.

As statisticians our rule of computations is simple. We should pursue exact computations and obtain best possible answers when possible. When we cannot pursue exact computations directly (e.g., if the sample size is large or storage requirements are enormous), we can look for alternatives that will give us answers that are as good as exact answers. In this regard, we have already seen that a basic importance sampling can fail in some instances. Thus, we must focus on MCMC methods and, whenever possible, we must pursue perfect MCMC sampling. In cases when we cannot afford to draw a large sample of linear extensions using the perfect MCMC method (because of the computing time or storage restrictions), we can at least try to use perfect MCMC algorithm (or run the MCMC longer than the mixing time) to generate one linear extension and then use either Gibbs or random walk Metropolis MCMC and compute the predictive distributions of latent p-values. We have already seen in simulations and data examples that there is a little loss if we draw one linear extension using the perfect MCMC and then draw the rest using Gibbs or random walk Metropolis MCMC methods. However, this will still take $O(n^3 \log n)$ computations. Note that typically matrix determinant computations take $O(n^3)$ time steps, and with modern computing power, we can routinely compute the determinant of a 10,000 × 10,000 matrix. Thus, these MCMC computations can also be done in a routine way if the sample size is less than 10,000. In this regard, the preliminary R codes provided in the supplement to the paper (Mondal and Hinrichs, 2016) can be fine-tuned using C/C++ language and with distributed systems to achieve various algorithm efficiency. Finally, if the sample size is large (say, $n \gg 10{,}000$), we will need to look for alternatives such as computations based on asymptotic results.

There are further opportunities to develop and implement our statistical procedures, particularly when we are interested in the estimation of regression parameters and statistical inference of interval censored data. A typical framework is described in Satten (1996) in that ranks are generated using Gibbs samples and score equations are computed using stochastic approximations. Thus, instead of Gibbs sampling, we can adapt the perfect MCMC sampling scheme and construct approximations to marginal likelihood score equations and to the inverse of the Fisher's information matrix, without specifying the baseline hazard. These computations are also extensible for making a statistical inference from the multivariate interval-censored data; see, for example Goggins and Finkelstein (2000). The sampling procedures discussed here can also be useful in many other contexts, as incomplete ranking arises in many other applications. One context is that of constructing the average ranks from partially ordered data that arise in chemistry and environmental science; see, for example, Lerche et al. (2002), Patil and Taillie (2004) and subsequent references. Another context is the estimation of coherency in observational studies (see, e.g., Rosenbaum, 2002). Some of the sampling schemes derived here are also applicable even when further restrictions are imposed in the partial orders, for example, in elementary ranking conditions or antimatroids in linguistic studies (Riggle, 2009) where we have directed edges from not just one vertex to another, but also from certain subsets of vertices to some other subset of vertices. Finally, development of an R package with an optimized code will greatly facilitate the use of our methods to different scientific applications and will be a matter of future research.

ACKNOWLEDGMENTS

We thank the Editor, the Associate Editor and two anonymous reviewers for their encouraging and constructive comments. Benjamin de Bivort, a fellow at the Rowland Institute at Harvard University, provided the fruit flies data in Section 4.2. The work of Debashis Mondal is supported by the NSF Career award DMS-1519890.

SUPPLEMENTARY MATERIAL

Supplement to "Rank Tests from Partially Ordered Data Using Importance and MCMC Sampling Methods" (DOI: 10.1214/16-STS549SUPP; .zip). Data and some basic R codes are available in the Supplementary files.

REFERENCES

AHO, A. V., GAREY, M. R. and ULLMAN, J. D. (1972). The transitive reduction of a directed graph. *SIAM J. Comput.* **1** 131–137. MR0306032

BAIK, J., DEIFT, P. and JOHANSSON, K. (1999). On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.* **12** 1119–1178. MR1682248

BAYARRI, M. J. and BERGER, J. O. (2000). *p* values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142. MR1804239

BAYARRI, M. J. and BERGER, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Sci.* **19** 58–80. MR2082147

BESAG, J. (2004). *Markov Chain Monte Carlo Methods for Statistical Inference*. Dept. Statistics, Univ. Washington, Seattle.

BESAG, J. and CLIFFORD, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76** 633–642. MR1041408

BESAG, J. and MONDAL, D. (2013). Exact goodness-of-fit tests for Markov chains. *Biometrics* **69** 488–496. MR3071067

BESAG, J., GREEN, P., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems. *Statist. Sci.* **10** 3–41. MR1349818

BRIGHTWELL, G. and WINKLER, P. (1991). Counting linear extensions. *Order* **8** 225–242. MR1154926

BUBLEY, R. and DYER, M. (1999). Faster random generation of linear extensions. *Discrete Math.* **201** 81–88. MR1687870

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. and STEIN, C. (2001). *Introduction to Algorithms*, 2nd ed. MIT Press, Cambridge, MA. MR1848805

COX, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc. Ser. B* **34** 187–220. MR0341758

COX, D. R. and HINKLEY, D. V. (1979). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

CROWLEY, J. (1974). A note on some recent likelihoods leading to the log rank test. *Biometrika* **61** 533–538. MR0378229

FAY, M. P. and SHAW, P. A. (2010). Exact and asymptotic weighted log-rank tests for interval censored data: The interval R package. *J. Stat. Software* **36** 1–34.

FERRENBERG, A. M., LANDAU, D. P. and SWENDSEN, R. H. (1995). Statistical errors in histogram reweighting. *Phys. Rev. E* **51** 5092–5100.

FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42** 845–854. MR0872963

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GEYER, C. J. and MEEDEN, G. D. (2005). Fuzzy and randomized confidence intervals and P-values. *Statist. Sci.* **20** 358–366. MR2210225

GOGGINS, W. B. and FINKELSTEIN, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* **56** 940–943.

GOGGINS, W. B., FINKELSTEIN, D. M., SCHOENFELD, D. A. and ZASLAVSKY, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* **54** 1498–1507.

GORDON, A. D. (1979a). A measure of the agreement between rankings. *Biometrika* **66** 7–15. MR0529142

GORDON, A. D. (1979b). Another measure of the agreement between rankings. *Biometrika* **66** 327–332. MR0548201

HÁJEK, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Stat.* **39** 325–346. MR0222988

HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. Academic Press, San Diego, CA. MR1680991

HUBER, M. (2004). Perfect sampling using bounding chains. *Ann. Appl. Probab.* **14** 734–753. MR2052900

HUBER, M. (2006). Fast perfect sampling from linear extensions. *Discrete Math.* **306** 420–428. MR2211010

HUNT, J. W. and SZYMANSKI, T. G. (1977). A fast algorithm for computing longest common subsequences. *Commun. ACM* **20** 350–353. MR0436655

KAHN, A. B. (1962). Topological sorting of large networks. *Commun. ACM* **5** 558562.

KAIN, J. S., STOKES, C. and DE BIVORT, B. L. (2012). Phototactic personality in fruit flies and its suppression by serotonin and white. *Proc. Natl. Acad. Sci. USA* **109** 19834–19839.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York. MR0570114

KARZANOV, A. and KHACHIYAN, L. (1991). On the conductance of order Markov chains. *Order* **8** 7–15. MR1129609

KRUSKAL, W. H. and WALLIS, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* **47** 583–621.

LERCHE, D., BRÜGGEMANN, R., SØRENSEN, P., CARLSEN, L. and NIELSEN, O. J. (2002). A comparison of partial order technique with three methods of multi-criteria analysis for ranking of chemical substances. *J. Chem. Inf. Comput. Sci.* **42** 1086–1098.

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR2401592

MATTHEWS, P. (1991). Generating a random linear extension of a partial order. *Ann. Probab.* **19** 1367–1392. MR1112421

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160. MR1311969

MONDAL, D. and HINRICHS, N. (2016). Supplement to "Rank tests from partially ordered data using importance and MCMC sampling methods." DOI:10.1214/16-STS549SUPP.

MORTON, D. E., SAAH, A. J., SILBERG, S. L., OWENS, W. L., ROBERTS, M. A. and SAAH, M. D. (1982). Lead absorption in children of employees in a lead-related industry. *Am. J. Epidemiol.* **115** 549–555.

PAGE, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *J. Amer. Statist. Assoc.* **58** 216–230. MR0145623

PATIL, G. P. and TAILLIE, C. (2004). Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. *Environ. Ecol. Stat.* **11** 199–228. MR2086395

PRENTICE, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65** 167–179. MR0497517

PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252. MR1611693

PURI, M. L. and SEN, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York. MR0298844

RIGGLE, J. (2009). The complexity of ranking hypotheses in optimality theory. *Comput. Linguist.* **35** 47–59. MR2511114

ROSENBAUM, P. R. (1991). Some poset statistics. *Ann. Statist.* **19** 1091–1097. MR1105865

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

RUBIN, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *J. Amer. Statist. Assoc.* **82** 543–546.

SATTEN, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83** 355–370.

SCHENSTED, C. (1961). Longest increasing and decreasing subsequences. *Canad. J. Math.* **13** 179–191. MR0121305

SELF, S. G. and GROSSMAN, E. A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics* **42** 521–530.

SKARE, Ø., BØLVIKEN, E. and HOLDEN, L. (2003). Improved sampling-importance resampling and reduced bias importance sampling. *Scand. J. Stat.* **30** 719–737. MR2155479

SMITH, A. F. M. and GELFAND, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *Amer. Statist.* **46** 84–88. MR1165566

TARJAN, R. E. (1976). Edge-disjoint spanning trees and depth-first search. *Acta Inform.* **6** 171–185. MR0424598

THOMPSON, E. A. and GEYER, C. J. (2007). Fuzzy p-values in latent variable problems. *Biometrika* **94** 49–60. MR2307899

VANDAL, A. C., CONDER, M. D. E. and GENTLEMAN, R. (2009). Minimal covers of maximal cliques for interval graphs. *Ars Combin.* **92** 97–129. MR2532568

VANDAL, A. C. and GENTLEMAN, R. (1998). Weak order partitioning of interval orders with applications to interval censored data Technical Report STAT9702, Dept. Statistics, Univ. Auckland.

WILSON, D. B. (2004). Mixing times of Lozenge tiling and card shuffling Markov chains. *Ann. Appl. Probab.* **14** 274–325. MR2023023

ZAR, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. *J. Amer. Statist. Assoc.* **67** 578–580.