Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ

J. van Doorn, A. Ly, M. Marsman & E.-J. Wagenmakers

Published online: 11 Jan 2020.

Submit your article to this journal

Article views: 273

View related articles

View Crossmark data

Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ

J. van Doorn [a], A. Ly [a,b], M. Marsman [a] and E.-J. Wagenmakers[a]

[a]Department of Psychological Methods, University of Amsterdam, Amsterdam, the Netherlands; [b]Centrum voor Wiskunde & Informatica, Amsterdam, the Netherlands

ABSTRACT

Bayesian inference for rank-order problems is frustrated by the absence of an explicit likelihood function. This hurdle can be overcome by assuming a latent normal representation that is consistent with the ordinal information in the data: the observed ranks are conceptualized as an impoverished reflection of an underlying continuous scale, and inference concerns the parameters that govern the latent representation. We apply this generic data-augmentation method to obtain Bayes factors for three popular rank-based tests: the rank sum test, the signed rank test, and Spearman's ρ_s.

1. Introduction

The debate on alternatives to null hypothesis significance tests based on p-values [63] has led to a renewed interest in the Bayesian alternative known as the Bayes factor. Advantages of such Bayesian tests include the ability to provide evidence in favor of *both* the null and the alternative hypotheses [12], the ability to straightforwardly synthesize evidence to assess replicability [36], and the ability to monitor the evidence as the data accumulate [51]; see [13,62] for further details on the advantages of Bayesian inference. These advantages are met by the recently proposed Bayes factors for the classical two- and one-sample t-tests [52], as well as for the Bayes factor for Pearson's correlation [37]. These tests have become increasingly popular in the applied sciences. The goal of this paper is to extend these parametric Bayes factors to their rank-based counterparts.

Rank-based statistical procedures offer a range of advantages over their parametric counterparts. First, they are robust to outliers and to violations of distributional assumptions, which occur frequently in many practical applications, such as the analysis of questionnaire data. Second, they are invariant under monotonic transformations, which is desirable when interest concerns a hypothesized concept (e.g. rat intelligence) whose relation to the measurement scale is only weakly specified (e.g. brain volume or log brain volume could be used as a predictor; without a process model that specifies how brain physiology translates to rat intelligence, neither choice is privileged). Third, many data sets are inherently ordinal (e.g. Likert scales, where survey participants are asked to indicate their

---

CONTACT J. van Doorn ✉ JohnnyDoorn@gmail.com 🏛 Department of Psychological Methods, University of Amsterdam, Valckeniersstraat 59, 1018 XA Amsterdam, the Netherlands

opinion on, say, a 7-point scale ranging from 'disagree completely' to 'agree completely'). Finally, rank-based procedures perform better than their fully parametric counterparts when assumptions are violated, with little loss of efficiency when the assumptions do hold [24].

Prominent rank-based tests include the Mann-Whitney-Wilcoxon rank sum test (i.e. the rank-based equivalent of the two-sample t-test), the Wilcoxon signed rank test (i.e. the rank-based equivalent of the paired sample t-test), and Spearman's ρ_s (i.e. a rank-based equivalent of the Pearson correlation coefficient). These ordinal tests were developed within the frequentist statistical paradigm, and Bayesian analogues through Bayes factor hypothesis testing have, to the best of our knowledge, not yet been proposed. We speculate that the main challenge in the development of Bayesian hypothesis tests for ordinal data is the lack of a straightforward likelihood function. As stated by Harold Jeffreys [27, pp. 178–179] for the case of Spearman's ρ_s:

> 'The rank correlation, while certainly useful in practice, is difficult to interpret. It is an estimate, but what is it an estimate of? That is, it is calculated from the observations, but a function of the observations has no relevance beyond the observations unless it is an estimate of a parameter in some law. Now what can this law be? [⋯] the interpretation is not clear.'

This difficulty can be overcome by postulating a latent, normally distributed level for the observed data (i.e. data augmentation). In other words, the rank data are conceptualized to be an impoverished reflection of richer latent data that are governed by a specific likelihood function. The latent normal distribution was chosen for computational convenience and ease of interpretation. This general procedure is widely known as data augmentation [3,56], and Bayesian inference for the parameters of interest (e.g. a location difference parameter δ or an association parameter ρ) can then be achieved using Markov chain Monte Carlo (MCMC) sampling. In other words, we can use the latent normal approach to overcome the lack of a likelihood function, and thus enable a Bayesian approach to rank-based testing.

Below we first outline the general latent normal framework and then develop Bayesian counterparts for three popular frequentist rank-based procedures: the rank sum test, the signed rank test, and Spearman's rank correlation. Each of these developed Bayesian tests is accompanied by a simulation study that assesses the behavior of the test and a data example that highlights the desirable properties of rank-based inference, as well as the applicability of our proposed tests.

2. General methodology

In the Bayesian framework, the posterior distribution of the parameter of interest θ is often used for hypothesis testing and parameter estimation. The posterior distribution is proportional to the likelihood, i.e. $f(\text{data} \mid \theta)$, times the prior, i.e. $\pi(\theta)$, that is,

$$\pi(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) \times \pi(\theta). \tag{1}$$

In the parametric case, this is often straightforward. For rank-based procedures, however, $f(\text{data} \mid \theta)$ is unavailable and to overcome this complication, we can use a latent normal framework.

2.1. Latent normal models

Latent normal models were first introduced by [48] as a means of modeling data from a 2×2 cross-classification table. The method was later extended by [49] to accommodate $r \times s$ tables. Instead of modeling the count data directly for the 2×2 case, Pearson assumed a latent bivariate normal level with certain governing parameters. In the case of cross-classification tables, the governing parameter is the *polychoric correlation coefficient* (PCC) and refers to Pearson's correlation on the bivariate, latent normal level.

A maximum likelihood estimator for the PCC was developed by [46,47], and a Bayesian framework for the PCC was later introduced by [2]. This idea was extended by [50] to rank likelihood models, where the latent boundaries are not estimated but determined directly by the latent scores (see also [22,23]). For the two-sample location problem, a similar approach has been discussed by [4,5,53], where a continuous distribution is assumed to be underlying the observed data. Further models for ordinal data are given in [15,16,39,41]. However, these methods omit Bayesian hypothesis testing through Bayes factors and/or lack a straightforward interpretation of the model parameters.

In general, the latent normal methodology allows one to transform ordinal problems to parametric problems. The resulting models that are discussed here have a data-generating process, are governed by easily interpretable parameters on the latent level, and enable Bayes factor hypothesis testing. A detailed sampling algorithm of the general methodology is presented in the next section.

2.2. Posterior distribution

We elaborate the main idea of the latent normal approach with data consisting of two groups of samples. Let (r^x, r^y) be two vectors of ranked data, and z^x, z^y be the vectors of associated latent normal scores which depend on a model parameter θ. The latent normal posterior is then proportional to

$$\pi(z^x, z^y, \theta \mid x, y) \propto f(r^x, r^y \mid z^x, z^y) \times f(z^x, z^y \mid \theta) \times \pi(\theta) \tag{2}$$

Note how the parametric likelihood in (1) is now replaced by the product $f(r^x, r^y \mid z^x, z^y) \times f(z^x, z^y \mid \theta)$. As before, the third term on the right-hand side refers to the prior $\pi(\theta)$. The second term refers to the latent normal structure. For instance, in the two-sample case, we replace the generic θ by the population difference δ and take for $f(z^x, z^y \mid \theta)$ the product of two normal densities with unit variances, but a mean depending on δ, see below for further details. On the other hand, for inference on Spearman's ρ_s, we replace the generic θ by ρ, and take for $f(z^x, z^y \mid \theta)$ the centered bivariate normal density with unit variances, and correlation ρ.

The first term on the right-hand side of (2), i.e. $f(r^x, r^y \mid z^x, z^y)$ consists of a set of indicator functions, presented below, that connect the observed ranks to the unobserved latent normal scores, z^x, z^y such that the ordinal information (i.e. the ranking function) in the observations r^x, r^y is preserved. This is similar to the approach of [1,3], who sampled latent scores for binary or polytomous response data from a normal distribution that was truncated with respect to the ordinal information of the data.

With (2) in hand, we have the specified the link between the data, the latent normal scores and parameters, and an MCMC sampler can be constructed in order to obtain the

joint posterior distribution. This sampler takes as input the ordinal information of the observed data, and iteratively generates random parameter values θ as well as random latent scores z^x, z^y. The indicator function $f(r^x, r^y \mid z^x, z^y)$ ensures that the latent scores z^x, z^y retain the ordinal information in the data by truncating the latent normal likelihood $f(z^x, z^y \mid \theta)$. For the latent value z_i^x this means that its range is truncated by the lower and upper thresholds that are respectively defined as:

$$a_i^x = \max_{j:r_j^x < r_i^x} \left(z_j^x \right) \tag{3}$$

$$b_i^x = \min_{j:r_j^x > r_i^x} \left(z_j^x \right). \tag{4}$$

For example, suppose that on a particular MCMC iteration we wish to augment the observed ordinal value r_i^x to a latent z_i^x; on the latent scale, the lower threshold a_i^x is given by the maximum latent value associated with all r^x lower than r_i^x, whereas the upper threshold b_i^x is determined by the minimum latent value associated with all r^x higher than r_i^x. This dependence between the scores can make the sampler inefficient. In order to remedy the high degree of autocorrelation that data augmentation can induce [60], we included an additive decorrelating step documented by [35,44].

2.3. Estimation and testing

After obtaining the joint posterior distribution through the MCMC sampling algorithm outlined above, we can either focus on *estimation* and present the marginal posterior distribution for the parameter of interest θ, or we can conduct a Bayes factor *hypothesis test* and compare the predictive performance of a point-null hypothesis $\mathcal{H}_0$ (in which the parameter of interest is fixed at a predefined value θ_0) against that of an alternative hypothesis $\mathcal{H}_1$ (in which θ is free to vary; [27,29,38]). The Bayes factor can be interpreted as a predictive updating factor, that is, degree to which the observed data drive a change from prior to posterior odds for the hypothesis of interest:

$$\underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\text{Bayes factor}_{10}} = \underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\text{Posterior odds}} \tag{5}$$

For instance, a Bayes factor $\text{BF}_{10} = 7$ implies that the data are seven times more likely under $\mathcal{H}_1$ then under $\mathcal{H}_0$, whereas $\text{BF}_{10} = 1/9$ indicates that the data are 9 times more likely under the null than under the alternative.

For nested models, the Bayes factor be easily obtained using the Savage-Dickey density ratio [11,61], that is, the ratio of the posterior and prior ordinate for the parameter of interest θ, under $\mathcal{H}_1$, evaluated at the point of testing θ_0 specified under $\mathcal{H}_0$:

$$\text{BF}_{10} = \frac{p(\theta_0 \mid \mathcal{H}_0)}{p(\theta_0 \mid \text{data}, \mathcal{H}_1)}. \tag{6}$$

3. Case 1: Wilcoxon rank sum test

3.1. Background

The ordinal counterpart to the two-sample t-test is known as the Wilcoxon rank sum test (or as the Mann-Whitney-Wilcoxon U test). It was introduced by [64] and further developed by [40], who worked out the statistical properties of the test. Let $x = (x_1, \ldots, x_{n_1})$ and $y = (y_1, \ldots, y_{n_2})$ be two data vectors that contain measurements of n_1 and n_2 units, respectively. The aggregated ranks r^x, r^y (i.e. the ranking of x and y together) are defined as:

$$r_i^x = \text{rank of } x_i \text{ among } (x_1, \ldots, x_{n_1}, y_1 \cdots y_{n_2}),$$

$$r_i^y = \text{rank of } y_i \text{ among } (x_1, \ldots, x_{n_1}, y_1 \cdots y_{n_2}).$$

The test statistic U is then given by summing over either r^x or r^y, and subtracting $n_x(n_x + 1)/2$ or $n_y(n_y + 1)/2$, respectively. In order to test for a difference between the two groups, the observed value of U can be compared to the value of U that corresponds to no difference. This point of testing is defined as $n_1 n_2/2$.

To illustrate the procedure, consider the following hypothetical example. In the movie review section of a newspaper, three action movies and three comedy movies are each assigned a star rating between 0 and 5. Let $X = (4, 3, 1)$ be the star ratings for the action movies, and let $Y = (2, 3, 5)$ be the star ratings for the comedy movies. The corresponding aggregated ranks are $R^x = (5, 3.5, 1)$ and $R^y = (2, 3.5, 6)$. The test statistic U is then obtained by summing over either R^x or R^y, and subtracting $\frac{3(3+1)}{2} = 6$, yielding 3.5 or 5.5, respectively. Either of these values can then be compared to the null point which is equal to $n_1 n_2/2 = 4.5$.

The range of U depends on the sample sizes and to avoid this dependence, we consider the rank-biserial correlation, which is a standardized effect size of U instead. The rank-biserial correlation, denoted ρ_{rb}, is the correlation coefficient used as a measure of association between a nominal dichotomous variable and an ordinal variable. The transformation is as follows:

$$\rho_{rb} = 1 - \frac{2U}{n_1 n_2}. \tag{7}$$

When $\rho_{rb} = 1$ we now know that U is at its maximum. The rank-biserial correlation can also be expressed as the difference between the proportion of data pairs where $x_i > y_j$ versus $x_i < y_j$ [10,30]:

$$\rho_{rb} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} Q(x_i - y_j)}{n_1 n_2}, \tag{8}$$

where $Q(d_i)$ is the sign indicator function defined as

$$Q(d_i) = \begin{cases} -1 & \text{if } d_i < 0 \\ +1 & \text{if } d_i > 0 \end{cases}. \tag{9}$$

This provides an intuitive interpretation of the test procedure: each data point in x is compared to each data point in y and scored -1 or 1 if it is lower or higher, respectively. In the movie ratings data example, there are three pairs for which $x_i > y_j$, five pairs for which

$x_i < y_j$, and one pair for which $x_i = y_j$, yielding an observed rank-biserial correlation coefficient of $\frac{3-5}{9} = -0.22$, which is an indication that comedy movies receive slightly more positive reviews.

One argument to favor the Wilcoxon rank sum test over its parametric counterpart is provided by Pitman's asymptotic relative efficiency (ARE); that is, the ratio of the number of observations necessary to achieve the same level of power [33].[1] If ARE > 1 then we require fewer samples for U than for its parametric counterpart [58].

When the data are normally distributed as assumed under the parametric setting, then the rank sum test performs slightly poorer to the parametric two-sample t-test as ARE of $3/\pi \approx 0.955$ [21,32]. Thus, even when the distributional assumption of the t-test holds, the loss of the rank sum test in terms of sample sizes is about 4.5%. The ARE increases as the data distribution grows more heavy-tailed, with a maximum value of infinity. In addition, results for other distributions include the logistic distribution (ARE $= \pi^2/9 \approx$ 1.097), the Laplace distribution (ARE $= 1.5$), and the exponential distribution (ARE $= 3$). Hence, relatively little is lost when using the Wilcoxon rank sum tests as compared to the parametric two-sample t-test when the parametric assumptions are met, but a lot is gained when the assumptions are violated.

3.2. Sampling algorithm

For the Bayesian counterpart of the Wilcoxon rank sum test, we use the latent normal framework as elaborated on above. Specifically, the Bayesian data augmentation algorithm for the rank sum test follows the graphical model outlined in Figure 1. The ordinal information contained in the aggregated ranking constrains the corresponding values for the latent normal parameters Z^x and Z^y to lie within certain intervals (i.e. the ordinal information imposes truncation). The parameter of interest here is the effect size δ, the difference in location of the distributions for Z^x and Z^y. We follow [28] and assign δ a Cauchy prior with scale parameter γ. For computational simplicity, this prior is implemented as a normal distribution with an inverse gamma prior on the variance, where the shape parameter is set to 0.5 and the scale parameter is set to $\gamma^2/2$ [34,52]. The difference with earlier work is that we set the latent normal variances σ to 1, as the rank data contain no information about the variance and the inclusion of σ in the sampling algorithm becomes redundant.

In order to sample from the posterior distributions of δ, Z^x and Z^y, we used Gibbs sampling [17]. Specifically, the sampling algorithm takes the aggregated ranks r^x, r^y as input and iteratively generates the latent δ, Z^x, and Z^y as follows, at sampling time point s:

(1) For each i in $(1, \ldots, n_x)$, sample Z_i^x from a truncated normal distribution, where the lower threshold is a_i^x given in (3) and the upper threshold is b_i^x given in (4):

$$(Z_i^x \mid z_{i'}^x, z_i^y, \delta) \sim \mathcal{N}_{(a_i^x, b_i^x)}\left(-\tfrac{1}{2}\delta, 1\right),$$

where the subscripts of $\mathcal{N}$ indicate the interval that is sampled from.

(2) For each i in $(1, \ldots, n_y)$, the sampling procedure for Z_i^y is analogous to step 1, with

$$(Z_i^y \mid z_{i'}^y, z_i^x, \delta) \sim \mathcal{N}_{(a_i^y, b_i^y)}\left(\tfrac{1}{2}\delta, 1\right).$$
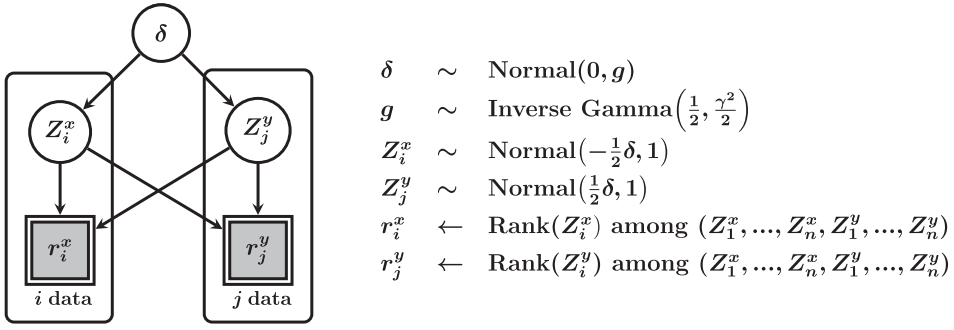
$$
\begin{aligned}
\delta &\sim \text{Normal}(0, g) \\
g &\sim \text{Inverse Gamma}\left(\tfrac{1}{2}, \tfrac{\gamma^2}{2}\right) \\
Z_i^x &\sim \text{Normal}\left(-\tfrac{1}{2}\delta, 1\right) \\
Z_j^y &\sim \text{Normal}\left(\tfrac{1}{2}\delta, 1\right) \\
r_i^x &\leftarrow \text{Rank}(Z_i^x) \text{ among } (Z_1^x, ..., Z_n^x, Z_1^y, ..., Z_n^y) \\
r_j^y &\leftarrow \text{Rank}(Z_i^y) \text{ among } (Z_1^x, ..., Z_n^x, Z_1^y, ..., Z_n^y)
\end{aligned}
$$

Figure 1. The graphical model underlying the Bayesian rank sum test. The latent, continuous scores are denoted by Z_i^x and Z_i^y, and their manifest rank values are denoted by x_i and y_j. The latent scores are assumed to follow a normal distribution governed by the parameter δ. This parameter is assigned a Cauchy prior distribution, which for computational convenience is reparameterized to a normal distribution with variance g (which is then assigned an inverse gamma distribution).

(3) Sample δ from

$$
(\delta \mid z^x, z^y, g) \sim \mathcal{N}(\mu_\delta, \sigma_\delta),
$$

where

$$
\mu_\delta = \frac{2g(n_y \overline{z^y} - n_x \overline{z^x})}{g(n_x + n_y) + 4}
$$

$$
\sigma_\delta^2 = \frac{4g}{g(n_x + n_y) + 4}.
$$

(4) Sample g from

$$
(G \mid \delta) \sim \text{Inverse Gamma}\left(1, \frac{\delta^2 + \gamma^2}{2}\right),
$$

where γ determines the scale (i.e. width) of the Cauchy prior on δ.

Repeating the algorithm a sufficient number of times yields samples from the posterior distributions of Z^x, Z^y, and δ. The posterior distribution of δ can then be used to obtain a Bayes factor through the Savage-Dickey density ratio given in (6).

3.3. Simulation study

In order to provide insight into the behavior of the inferential framework, a simulation study was performed. For three values of difference in location parameters, Δ (0, 0.5, 1.5), and three values of n (10, 20, 50), 1000 data sets were generated under various distributions: skew-normal, Cauchy, logistic, and uniform distributions. In one scenario, both groups have the same distributional shape (e.g. both follow a logistic distribution), and in a second scenario, one group follows the normal distribution and one group follows one of the aforementioned distributions.

First, the relationship between the observed rank statistic U and the latent normal Bayes factor was analyzed. Figure 2 illustrates this relationship, fitted with a cubic smoothing
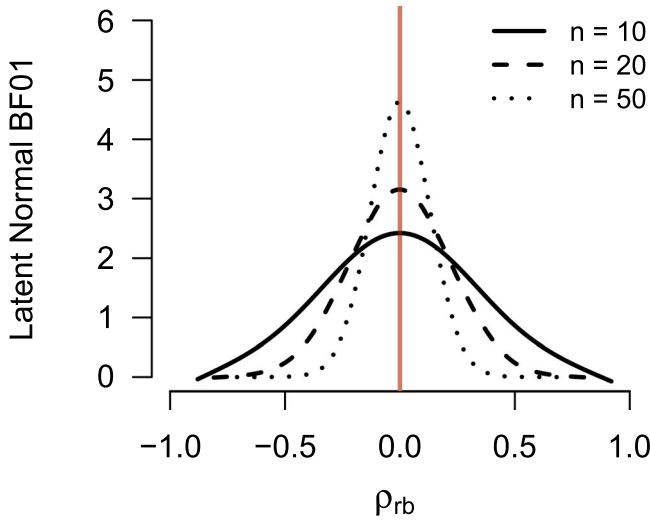
Figure 2. The relationship between the latent normal Bayes factor and the observed rank-based test statistic is illustrated for logistic data. Because U is dependent on n, the rank biserial correlation coefficient is plotted on the x-axis instead of U. The relationship is clearly defined, and maximum evidence in favor of $\mathcal{H}_0$ is attained when $\rho_{rb} = 0$. The further ρ_{rb} deviates from 0, the stronger the evidence in favor of $\mathcal{H}_1$ becomes. The lines depict smoothing splines fitted to the observed Bayes factors.

spline [6], for two logistic distributions ($\alpha = 20$). To show results for multiple values of n in one figure, the rank biserial correlation coefficient ρ_{rb} is plotted instead of U. The figure shows a clear relationship: when $\rho_{rb} = 0$, thus, U corresponds to the test value $n_1 n_2 / 2$, then the evidence in favor of $\mathcal{H}_0$ is at its maximum as one would expect. Similarly, when $|U|$ is maximal, that is, $|\rho_{rb}| = 1$, one has the most evidence against the null, which is apparent from the curves getting closer to 0. This relationship grows more decisive as n increases: both the peak at $\rho_{rb} = 0$ and the decay at $|\rho_{rb}| = 1$ are more prominent as n grows. The results are highly similar for the other distributions that were considered (see the online supplementary material at https://osf.io/gny35/ for the results of these scenarios). Since both statistics, ρ_{rb} and BF_{01}, depend solely on the ordinal information in the data, the observed relationship is not surprising. This result highlights and illustrates the robustness of the latent normal Bayes factor to violations of the assumptions of the parametric test: it illustrates the same robustness as the traditional W test statistic.

Second, the relationship between the latent normal Bayes factor and the parametric Bayes factor [52] was analyzed. For both the parametric and rank-based Bayes factor, a default Cauchy prior with scale $1/\sqrt{2}$ is used. Figure 3 illustrates this relationship for all values of n and Δ that were used, again in the scenario with two logistic distributions. Generally, the two Bayes factors are in agreement. In cases where Δ deviates from 0, the parametric Bayes factor becomes more decisive (i.e. deviates from 1) compared to the latent normal Bayes factor. For distributions of data that violate the assumptions of the parametric test, such as the Cauchy distribution, the relationship between the two Bayes factors is notably less defined. In this case, the results of the rank-based Bayes factor are more reliable, which is expected based on the ARE results as the Cauchy is a heavy-tailed distribution. The parametric test greatly overestimates the variance and is no longer able to detect differences in location parameters (see the supplementary material), whereas the latent normal
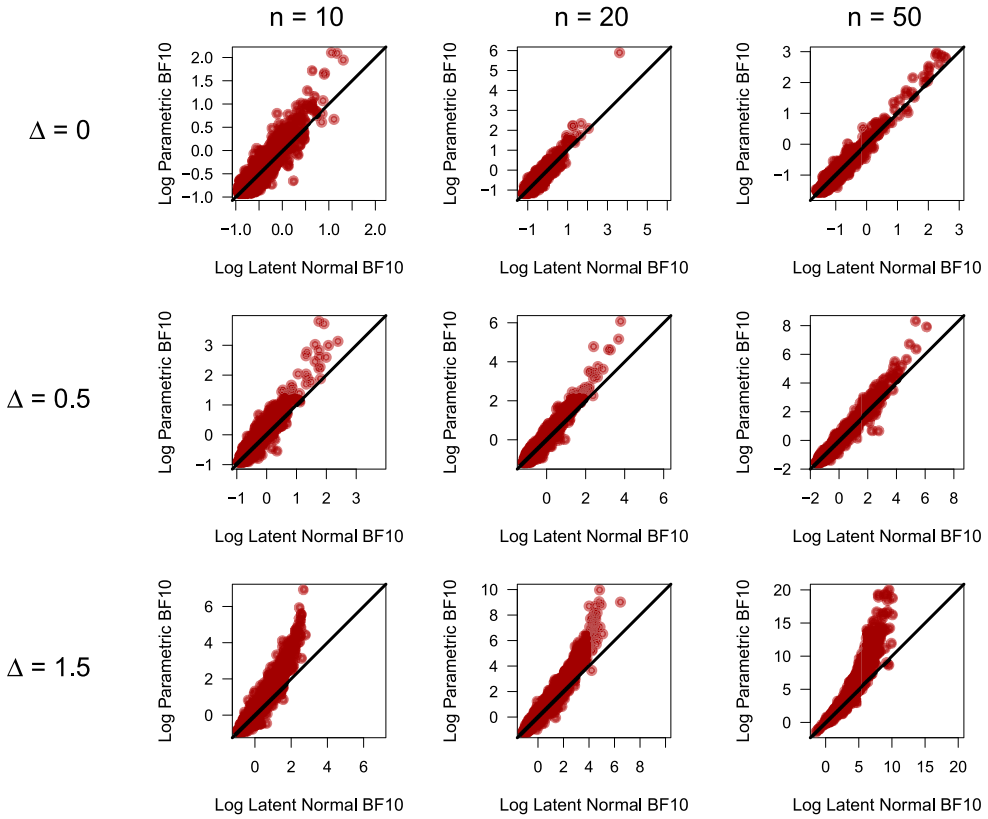
Figure 3. For all combinations of difference in location parameters Δ, and n, the relationship between the latent normal Bayes factor and the parametric Bayes factor is shown for logistic data. The black lines indicate the point of equivalence. The two Bayes factors are generally in agreement, as suggested by the ARE results in [58].

Bayes factor is unaffected by this. Note that the difference in performance is due to the use of the latent normal framework and not due to the prior, as both the parametric and rank-based Bayes factor use the same Cauchy prior.

3.4. Data example

Cortez and Silva [9] gathered data from 395 students concerning their math performance (scored between 1 and 20) and their level of alcohol intake (self-rated on a Likert scale between 1 and 5). Students passed the course if they scored ≥ 10, and we will test whether students who failed the course ($n_1 = 130$) had a higher self-reported alcohol intake than their peers who passed ($n_2 = 265$).

As alcohol intake was measured on a Likert scale, the data contain many ties and show extreme non-normality. These properties make this data set particularly suitable for the latent-normal rank sum test. The hypotheses are $\mathcal{H}_0 : \delta = 0$ which is pitted against $\mathcal{H}_1 : \delta \neq 0$. For the rank-based Bayes factor we use the prior Cauchy prior with scale $1/\sqrt{2}$, that is, $\delta \sim \text{Cauchy}\left(0, \frac{1}{\sqrt{2}}\right)$. The null hypothesis posits that alcohol intake does not differ
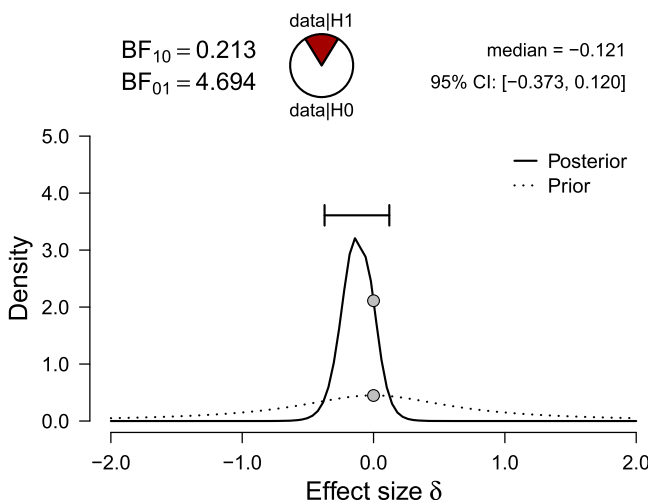
Figure 4. Do students who flunk a math course report drinking more alcohol? Results for the Bayesian rank sum test as applied to the data set from [9]. The dashed line indicates the Cauchy prior with scale $\frac{1}{\sqrt{2}}$. The solid line indicates the posterior distribution. The two grey dots indicate the prior and posterior ordinate at the point under test, in this case $\delta = 0$. The ratio of the ordinates gives the Bayes factor.

between the students who passed the course and those who failed. The alternative hypothesis posits the presence of an effect and assigns effect size a Cauchy distribution with scale parameter set to $\frac{1}{\sqrt{2}}$, as advocated by [43]. Figure 4 shows the resulting posterior distribution for δ under $\mathcal{H}_1$ and the associated Bayes factor. The posterior median for δ equals -0.121, with a 95% credible interval that ranges from -0.373 to 0.120. The corresponding Bayes factor indicates that the data are about 4.694 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, indicating moderate evidence against the hypothesis that self-reported alcohol intake differentiates between students who did and who did not pass the math exam. As a reference, the parametric t-test yields a Bayes factor of 7.138 in favor of $\mathcal{H}_1$, which is less conservative. However, due to the violated assumptions of the parametric t-test model, this result is meaningless.

4. Case 2: Wilcoxon signed rank test

4.1. Background

The rank-based counterpart to the paired samples t-test was proposed by [64], who termed it the *signed rank test*. The test procedure involves taking the difference scores between the two samples under consideration and ranking the absolute values. The procedure may also be applied to one-sample scenarios by ranking the differences between the observed sample and the point of testing. These ranks are then multiplied by the sign of the respective difference scores and summed to produce the test statistic W. For the paired samples signed rank test, let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two data vectors each containing measurements of the same n units, and let $d = (d_1, \ldots, d_n)$ denote the difference scores. For the one-sample signed rank test, this process is analogous, except y is replaced by the

Table 1. The scores, difference scores, ranks of the absolute difference scores, and the sign indicator function Q for the hypothetical scenario where $X = (5, 8, 4)$ are the initial scores on a math exam and $Y = (6, 7, 7)$ are the scores on the exam after a tutoring session.

i	$(y_i - x_i)$	d_i	rank($|d_i|$)	$Q(d_i)$
1	6−5	1	1.5	1
2	7−8	−1	1.5	−1
3	7−4	3	3	1

test value. The test statistic is then defined as:

$$W = \sum_{1}^{n} \left[\text{rank}(|d_i|) \times Q(d_i) \right],$$

where Q is the sign indicator function given in (9).

To illustrate the procedure, consider the following hypothetical data example. Three students take a math exam, graded between 0 and 10, before and after receiving a tutoring session. Let $X = (5, 8, 4)$ be their scores on the exam before the session, and let $Y = (6, 7, 7)$ be their scores on the exam after the session. The difference scores, the ranks of the absolute difference scores, and the sign indicator function are presented in Table 1. In order to have a positive test statistic indicate an increase in scores, the difference scores are defined here as $(y_i - x_i)$. The test statistic W is then calculated by summing over the product of the fourth and fifth column: $1.5 - 1.5 + 3 = 3$. This value indicates a slight increase in math scores after the tutoring session.

An often used standardized effect size for W is the matched-pairs rank-biserial correlation, denoted ρ_{mrb}, which is the correlation coefficient used as a within subjects measure of association between a nominal dichotomous variable and an ordinal variable [10,30]. The transformation is as follows:

$$\rho_{mrb} = 1 - \frac{4W}{n(n+1)}. \tag{10}$$

The matched-pairs rank-biserial correlation can also be expressed as the difference between the proportion of data pairs where $x_i > y_i$ versus $x_i < y_i$. For the grades example, there is one pair for which $x_i > y_i$, and two pairs for which $x_i < y_i$, yielding a matched-pairs rank-biserial correlation coefficient of $\frac{2-1}{3} = \frac{2}{3}$, which is an indication that the tutoring session has increased students' math ability.

The signed rank test is similar to the sign test, where the procedure is to sum over the sign indicator function. The difference here is that the output of the sign indicator function is weighted by the ranked magnitude of the absolute differences. The signed rank test has a higher ARE than the sign test: a relative efficiency of $\frac{3}{2}$ for all distributions [8]. For the one-sample scenario, the Pitman ARE of the signed rank test (compared to the fully parametric t-test) is similar to the ARE of the rank sum test for the unpaired two-sample scenario; for example, when the data follow a normal distribution the ARE equals $3/\pi$. For other distributions, especially when these are heavy-tailed, the signed rank test outperforms the t-test [33,58].

4.2. Sampling algorithm

The data augmentation algorithm is similar to that of the rank sum test and is outlined in Figure 5. Here d denotes the difference scores as ordinal manifestations of latent, normally distributed values Z^d. The parameter of interest is again the standardized location parameter δ, which is assigned a Cauchy prior distribution with scale parameter γ. Similar to the rank sum sampling procedure, the variance of Z^d is set to 1, as the ranked data contain no information about the variance. The computational complexity of sampling from the posterior distribution of δ is again reduced by introducing the parameter g. The Gibbs algorithm for the data augmentation and sampling δ is as follows, at sampling time point s:

(1) For each value of i in $(1, \ldots, n)$, sample Z_i^d from a truncated normal distribution, where the lower threshold is a_i^d given in (3) and the upper threshold is b_i^d given in (4):

$$(Z_i^d \mid z_{i'}^d, \delta) \sim \mathcal{N}_{\left(a_i^d, b_i^d\right)}(\delta, 1)$$

(2) Sample δ from

$$(\delta \mid z^d, g) \sim \mathcal{N}\left(\mu_\delta, \sigma_\delta^2\right),$$

where

$$\mu_\delta = \frac{gn\overline{z^d}}{gn + 1}$$

$$\sigma_\delta^2 = \frac{g}{gn + 1}$$

(3) Sample g from

$$(g \mid \delta) \sim \text{InverseGamma}\left(1, \frac{\delta^2 + \gamma^2}{2}\right),$$

where γ determines the scale (i.e. width) of the Cauchy prior on δ.

Repeating the algorithm a sufficient number of times yields samples from the posterior distributions of Z^d and δ. The posterior distribution of δ can then be used to obtain a Bayes factor through the Savage-Dickey density ratio given in (6).

4.3. Simulation study

Similar to the Wilcoxon rank sum test, a simulation study was performed to illustrate the behavior of the Bayesian signed rank test. For three values of difference in location parameters, Δ (0, 0.5, 1.5), and three values of n (10, 20, 50), 1000 data sets were generated under various distributions: skew-normal, Cauchy, logistic, and uniform distributions. In one scenario, both groups have the same distributional shape, and in a second scenario, one group follows the normal distribution and one group follows one of the aforementioned distributions. After the data were generated, the difference scores between the two groups were calculated, and used as input for the Bayesian latent normal test.
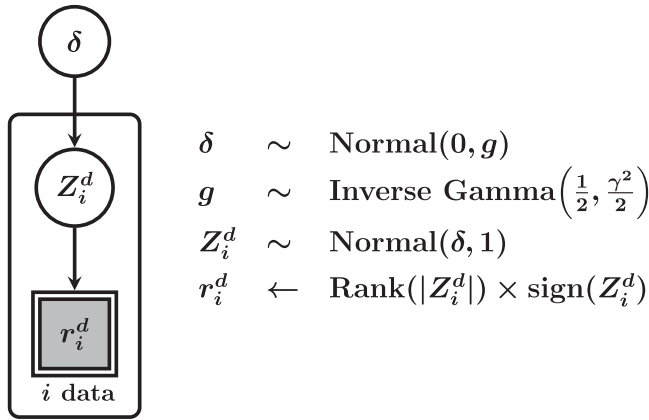
Figure 5. The graphical model underlying the Bayesian signed rank test. The latent, continuous difference scores are denoted by Z_i^d, and their manifest signed rank values are denoted by d_i. The latent scores are assumed to follow a normal distribution governed by parameter δ. This parameter is assigned a Cauchy prior distribution, which for computational convenience is reparameterized to a normal distribution with variance g (which is then assigned an inverse gamma distribution).
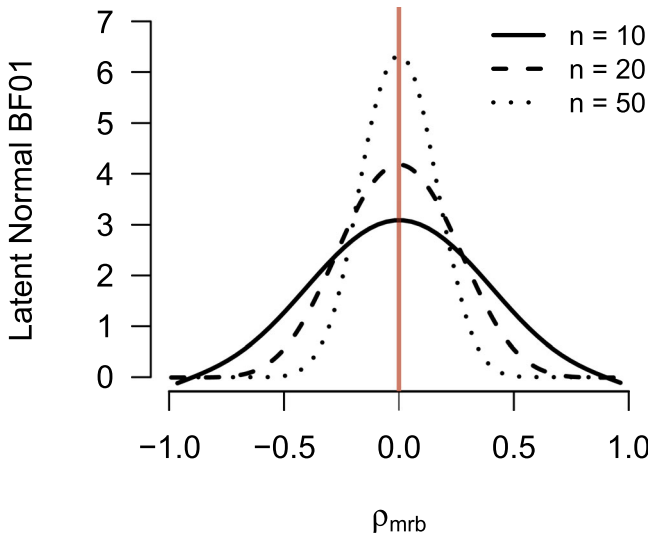

Figure 6. The relationship between the latent normal Bayes factor and the observed rank-based test statistic is illustrated for logistic data. Because W is dependent on n, the matched-pairs rank-biserial correlation coefficient is plotted on the x-axis instead of W. The relationship is clearly defined, and maximum evidence in favor of $\mathcal{H}_0$ is attained when $\rho_{mrb} = 0$. The further ρ_{mrb} deviates from 0, the stronger the evidence in favor of $\mathcal{H}_1$ becomes. The lines are smoothing splines fitted to the observed Bayes factors.

The same analyses were performed as for the Wilcoxon rank sum test. First, the relationship between the observed rank statistic W and the latent normal Bayes Factor was analyzed. Figure 6 illustrates this relationship, fitted with a cubic smoothing spline [6], when the difference scores were taken for two logistic distributions. To show results for
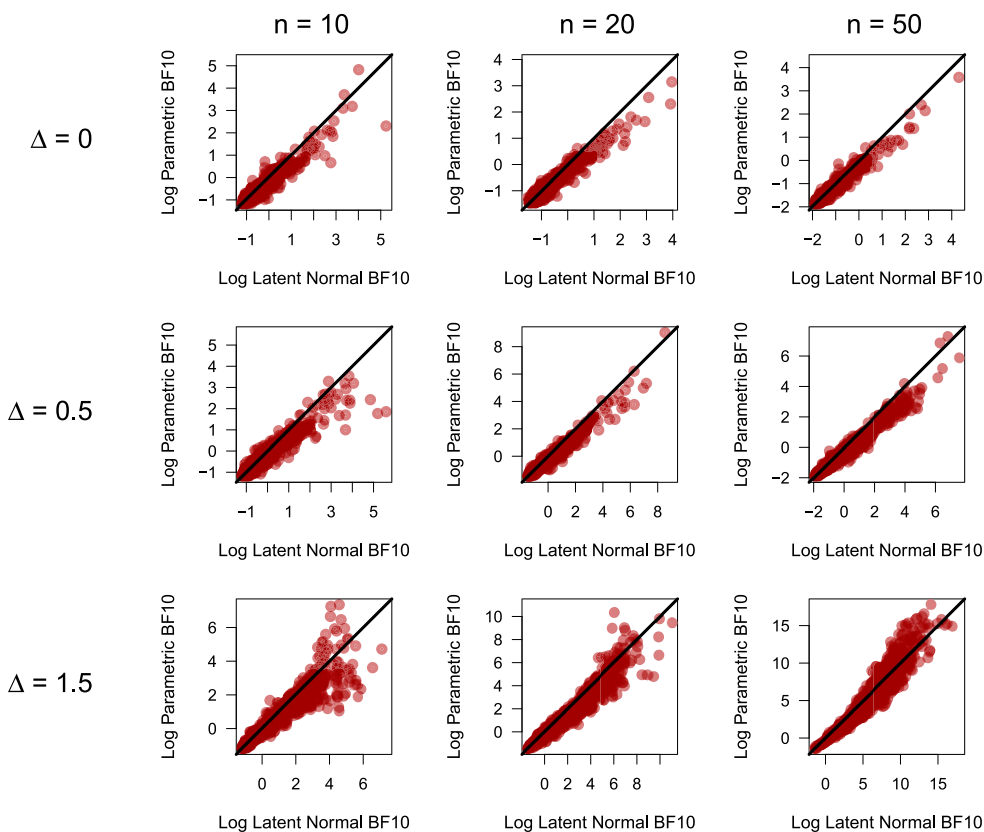
Figure 7. For all combinations of difference in location parameters Δ, and n, the relationship between the latent normal Bayes factor and the parametric Bayes factor is shown for logistic data. The black lines indicate the point of equivalence. The two Bayes factors are generally in agreement, with the latent normal Bayes factor accumulating evidence in favor of the true model faster.

multiple values of n in one figure, the matched-pairs rank-biserial correlation coefficient ρ_{mrb} is plotted instead of W. The Bayes factor shows a clear relationship with the rank-based test statistic, where the maximum evidence in favor of $\mathcal{H}_0$ is obtained when this statistic equals 0. Furthermore, the obtained Bayes factor grows more decisive as n increases. For other distributions of the data, highly similar results were obtained (see the online supplementary material at https://osf.io/gny35/ for the results of these scenarios).

Next to the relationship between W and the latent normal Bayes factor, the relationship between the latent normal Bayes factor and the parametric Bayes factor [52] was analyzed. Figure 7 illustrates the results for all combinations of n and the difference in location parameters, Δ. Note that differences in performance are due to the use of the latent normal framework and not due to the prior specification, as both the parametric and rank-based Bayes factor were based on the same Cauchy prior with scale $1/\sqrt{2}$. The two Bayes factors are generally in agreement, with the parametric Bayes factor accumulating evidence in favor of $\mathcal{H}_1$ faster when this is the true model. The latent normal Bayes factor demonstrates more instability, due to only using the ordinal information in the data. For distributions of

the data that violate the assumptions of the parametric test, such as the Cauchy distribution, the parametric test greatly overestimates the variance and is no longer able to detect differences in location parameters (see the supplementary material). This misspecification does not affect the latent normal Bayes factor, underscoring its robustness.

4.4. Data example

Thall and Vail [57] investigated a data set obtained by D. S. Salsburg concerning the effects of the drug progabide on the occurrence of epileptic seizures. During an initial eight week baseline period, the number of epileptic seizures was recorded in a sample of 31 epileptics. Next, the patients were given progabide, and the number of epileptic seizures was recorded for another eight weeks. In order to accommodate the discreteness and non-normality of the data, Thall and Vail [57] applied a log-transformation on the counts.

This log-transformation has a clear impact on the outcome of a parametric Bayesian t-test [43]: $BF_{10} \approx 0.2$ for the raw data, whereas $BF_{10} \approx 2.95$ for the log-transformed data. Here we analyze the data with the signed rank test; because this test is invariant under monotonic transformations, the same inference will result regardless of whether or not the data are log-transformed.

The hypothesis specification here is similar to that of the setup of the rank sum example: $\mathcal{H}_0 : \delta = 0$ which is pitted against $\mathcal{H}_1 : \delta \neq 0$ and prior $1/\sqrt{2}$, that is, $\delta \sim$ Cauchy$\left(0, \frac{1}{\sqrt{2}}\right)$. Figure 8 shows the resulting posterior distribution for δ under $\mathcal{H}_1$ and the associated Bayes factor. The posterior median for δ equals 0.207, with a 95% credible interval that ranges from -0.138 to 0.549. The corresponding Bayes factor indicates that the data are about 2.513 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, indicating that, for the purpose of discriminating $\mathcal{H}_0$ from $\mathcal{H}_1$, the data are almost perfectly uninformative.

5. Case 3: Spearman's ρ_s

5.1. Background

Spearman [55] introduced the rank correlation coefficient ρ in order to overcome the main shortcoming of Pearson's product moment correlation, namely its inability to capture monotonic but non-linear associations between variables. Spearman's method first applies the rank transformation on the data and then computes the product-moment correlation on the ranks. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two data vectors each containing measurements of the same n units, and let $r^x = (r_1^x, \ldots, r_n^x)$ and $r^y = (r_1^y, \ldots, r_n^y)$ denote their rank-transformed values, where each value is assigned a ranking within its variable. This then leads to the following formula for Spearman's ρ_s:

$$\rho_s = \frac{\text{Cov}_{r^x r^y}}{\sigma_{r^x} \sigma_{r^y}}.$$

The Pitman ARE of Spearman's ρ compared to parametric Pearson's ρ displays a similar pattern to the ARE's discussed before. When the data follow a bivariate normal distribution, the ARE equals $9/\pi^2$ [25]. Thus, under optimal conditions for the parametric test, it is marginally more efficienct compared to Spearman's ρ. As the data depart from normality, the rank-based test outperforms its parametric counterpart.
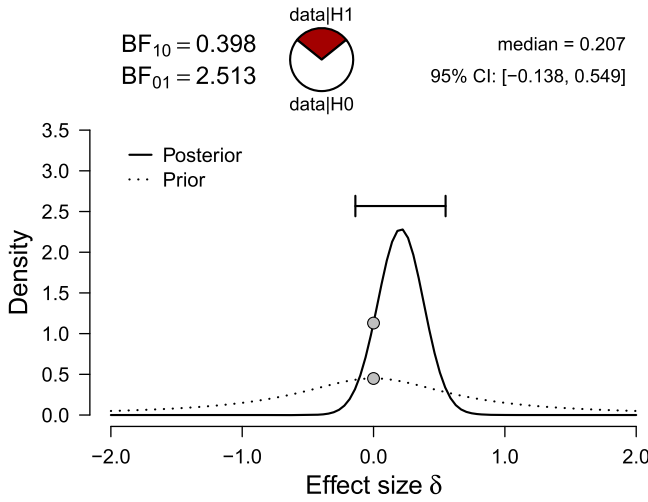
Figure 8. Does progabide reduce the frequency of epileptic seizures? Results for the Bayesian signed rank test as applied to the data set presented in [57]. The dashed line indicates the Cauchy prior with scale $\frac{1}{\sqrt{2}}$. The solid line indicates the posterior distribution. The two grey dots indicate the prior and posterior ordinate at the point under test, in this case $\delta = 0$. The ratio of the ordinates gives the Bayes factor.
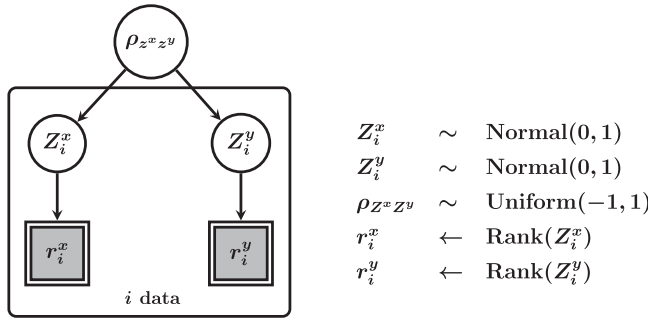

Figure 9. The graphical model underlying the Bayesian test for Spearman's ρ_s. The latent, continuous scores are denoted by Z_i^x and Z_i^y, and their manifest rank values are denoted by r_i^x and r_j^y. The latent scores are assumed to follow a normal distribution governed by parameter ρ (which is assigned a uniform prior distribution).

5.2. Sampling algorithm

The graphical model in Figure 9 illustrates the data augmentation setup for inference on the latent correlation parameter ρ. The sampling method is a Metropolis-within-Gibbs algorithm, where data augmentation is conducted with a Gibbs sampling algorithm as before, but combined with a random walk Metropolis-Hastings sampling algorithm [20,42] to sample from the posterior distribution of ρ (see also [59]).

The sampling algorithm for the latent correlation is as follows, at sampling time point s:

(1) For each i in $(1, \ldots, n_x)$, sample Z_i^x from a truncated normal distribution, where the lower threshold is a_i^x given in (3) and the upper threshold is b_i^x given in (4):

$$(Z_i^x \mid z_{i'}^x, z_i^y, \rho_{z^x,z^y}) \sim \mathcal{N}_{(a_i^x, b_i^x)} \left(z_i^y \rho_{z^x,z^y}, \sqrt{1 - \rho_{z^x,z^y}^2} \right)$$

(2) For each i in $(1, \ldots, n_y)$, the sampling procedure for Z_i^y is analogous to step 1.

(3) Sample a new proposal for ρ_{z^x,z^y}, denoted ρ^*, from the asymptotic normal approximation to the sampling distribution of Fisher's z-transform of ρ [14]:

$$\tanh^{-1}(\rho^*) \sim \mathcal{N} \left(\tanh^{-1}(\rho^{s-1}), \frac{1}{\sqrt{(n-3)}} \right).$$

The acceptance rate α is determined by the likelihood ratio of $(z^x, z^y \mid \rho^*)$ and $(z^x, z^y \mid \rho^{s-1})$, where each likelihood is determined by the centered bivariate normal density with unit variances, and correlation ρ:

$$\alpha = \min \left(1, \frac{P(z^x, z^y \mid \rho^*)}{P(z^x, z^y \mid \rho^{s-1})} \right).$$

Repeating the algorithm a sufficient number of times yields samples from the posterior distributions of z^x, z^y, and ρ_{z^x,z^y}.

5.3. Transforming parameters

The transition from Pearson's ρ to Spearman's ρ_s can be made using a statistical relation described in [31]. This relation, defined as

$$\rho_s = \frac{6}{\pi} \sin^{-1} \left(\frac{\rho}{2} \right).$$

enables the transformation of Pearson's ρ to Spearman's ρ_s when the data follow a bivariate normal distribution. Since the latent data are assumed to be normally distributed, this means that the posterior samples for Pearson's ρ can be easily transformed to posterior samples for Spearman's ρ_s. The posterior distribution of ρ_s can then be used to obtain a Bayes factor through the Savage-Dickey density ratio given in (6).

5.4. Simulation study

Similar to the previous tests, the behavior of the latent normal correlation test was assessed with a simulation study. For four values of Spearman's ρ_s (0, 0.3, 0.8) and three values of n (10, 20, 50), 1000 data sets were generated under four copula models: Clayton, Gumbel, Frank, and Gaussian [7,18,45,54]. Using Sklar's theorem, copula models decompose a joint distribution into univariate marginal distributions and a dependence structure (i.e. the copula). This decomposition enables the generation of data for specific values of Spearman's ρ_s. Furthermore, the copula is independent of the marginal distributions of the data and can therefore encompass a wide range of distributions.

Similar to the previous tests, the relationship between the latent normal Bayes factor and the observed rank-based statistic was analyzed. Figure 10 illustrates this relationship,
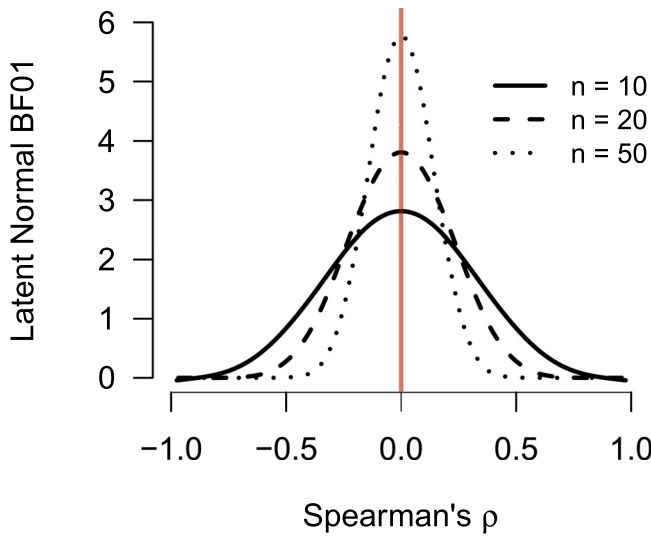
Figure 10. The relationship between the latent normal Bayes factor and the observed rank-based test statistic is illustrated for data generated with the Clayton copula. The relationship is clearly defined, and maximum evidence in favor of $\mathcal{H}_0$ is attained when Spearman's $\rho_s = 0$. The further Spearman's ρ_s deviates from 0, the stronger the evidence in favor of $\mathcal{H}_1$ becomes. The lines are smoothing splines fitted to the observed Bayes factors.

fitted with a cubic smoothing spline [6], for various values of n, for data generated with the Clayton copula. The relationship is similar to those shown for the previous tests: maximum evidence in favor of $\mathcal{H}_0$ is attained when the observed Spearman's ρ_s equals 0. The further the observed test statistic deviates from 0, the more evidence is accumulated in favor of $\mathcal{H}_1$. Furthermore, the obtained Bayes factor grows more decisive as n increases. Highly similar results were obtained for the other copulas that were considered (see the online supplementary material at https://osf.io/gny35/ for the results of these scenarios).

Secondly, the relationship between the latent normal Bayes factor and the parametric Bayes factor [37] for testing correlations was analyzed. For both Bayes factors, a uniform prior between -1 and 1 was used, such that differences in performance are due to the use of the latent normal framework and not due to the prior. Figure 11 shows the results for all combinations of n and ρ that were used, for the Clayton copula. The two Bayes factors are generally in agreement. An important remark here is that the marginal distributions of the data are not taken into account. The data generated with the copula method are located on the unit square, and if so desired, can then be transformed with the inverse cdf to follow any desired distribution. These transformations are monotonic, and therefore do not affect the rank-based Bayes factor, whereas the parametric Bayes factor can be heavily affected by this. This underscores an important property of the rank-based Bayes factor: it solely depends on the copula (i.e. the only component of the data that pertains to the dependence structure), and not on the marginal distribution of the data.

5.5. Data example

We return to the data set from [9] and examine the possibility that math grades (ranging from 0 to 20) are associated with the quality of family relations (self-reported on a Likert
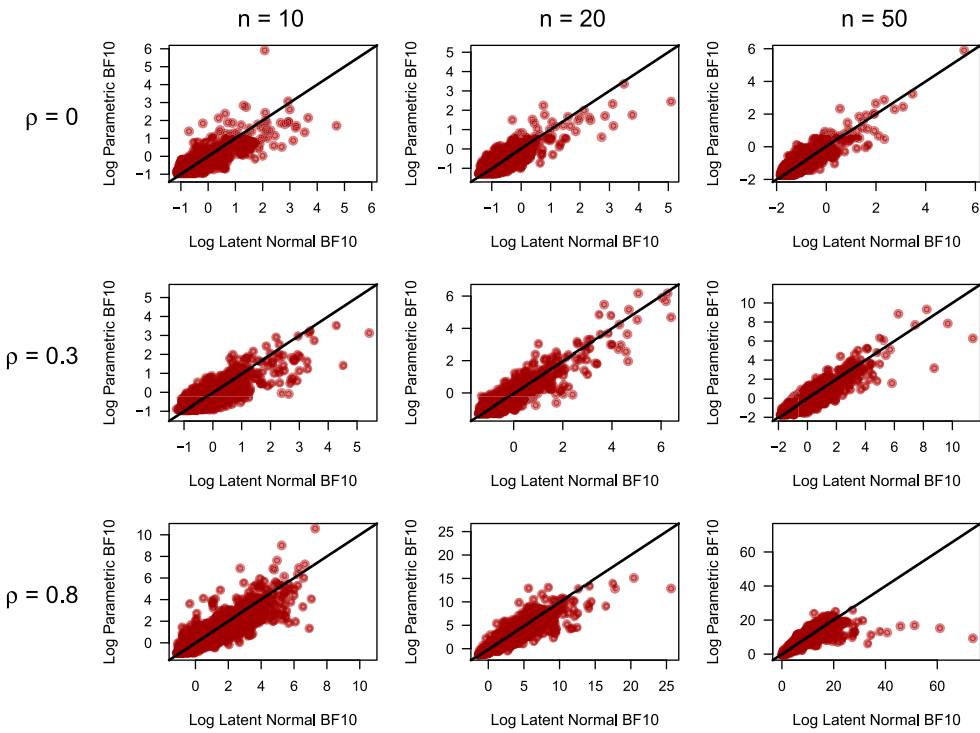
Figure 11. For all combinations of Spearman's ρ_s and n, the relationship between the latent normal Bayes factor and the parametric Bayes factor is shown for data generated with the Clayton copula. The black lines indicate the point of equivalence. The two Bayes factors are generally in agreement.
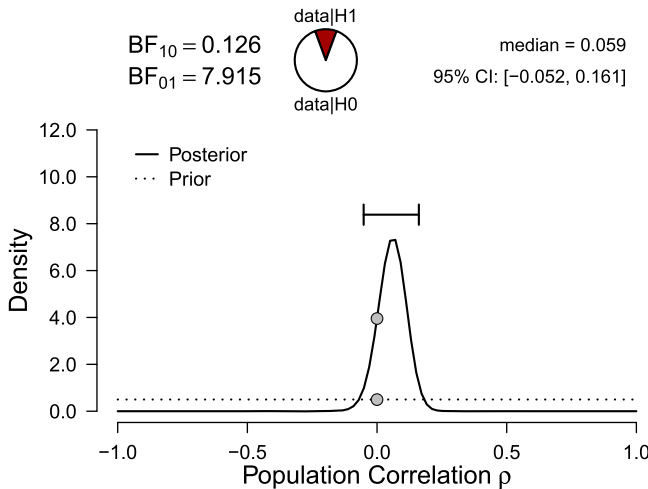

Figure 12. Is performance on a math exam associated with the quality of family relations? Results for the Bayesian version of Spearman's ρ_s as applied to the data set from [9]. The dashed line indicates the uniform prior distribution, and the solid line indicates the posterior distribution. The two grey dots indicate the prior and posterior ordinate at the point under test, in this case $\rho = 0$. The ratio of the ordinates gives the Bayes factor.

scale that ranges from $1-5$). The hypotheses are $\mathcal{H}_0 : \rho = 0$ which is pitted against $\mathcal{H}_1 : \rho \neq 0$. For the Bayes factor we use the uniform prior, that is, $\rho \sim \text{Uniform}[-1, 1]$. Thus, the null hypothesis specifies the lack of an association between the two variables and the alternative hypothesis assigns the degree of association a uniform prior distribution (e.g. [28]). The parametric correlation test [37] yields a Bayes factor of 9.467, but since the data are ordinal measures and not normally distributed, the parametric correlation model is severely misspecified. Thus, conducting the rank-based analysis is more applicable and prudent here.

Figure 12 shows the resulting posterior distribution for ρ_s under $\mathcal{H}_1$ and the associated Bayes factor. The posterior median for ρ_s equals 0.059, with a 95% credible interval that ranges from -0.052 to 0.161. The corresponding Bayes factor indicates that the data are about 7.915 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, indicating moderate evidence against an association between math performance and the quality of family ties.

6. Concluding comments

This article outlined a general methodology for applying conventional Bayesian inference procedures to ordinal data problems. Latent normal distributions are assumed to generate impoverished rank-based observations, and inference is done on the model parameters that govern the latent normal level. This idea, first proposed by [48], yields all the advantages of ordinal inference including robustness to outliers and invariance to monotonic transformations. Moreover, the methodology also handles ties in a natural fashion, which is important for coarse data such as provided by popular Likert scales. Furthermore, the robustness of the latent normal method is underscored by the simulation studies performed for each test. These results illustrate that the method provides accurate inference, even if the data are not normally distributed.

By postulating a latent normal level for the observed rank data, the advantages of ordinal inference can be combined with the advantages of Bayesian inference such as the ability to update uncertainty as the data accumulate, the ability to quantify evidence in favor of either hypothesis being tested, and the ability to incorporate prior information. It should be stressed that, even though our examples used default prior distributions, the proposed methodology is entirely general in the sense that it also applies to informed or subjective prior distributions [19].

For computational convenience and ease of interpretation, our framework used latent normal distributions. This is not a principled limitation, however, and the methodology would work for other families of latent distributions as well (e.g. [2]).

In sum, we have presented a general methodology to conduct Bayesian inference for ordinal problems, and illustrated its potential by developing Bayesian counterparts to three popular ordinal tests: the rank sum test, the signed rank test, and Spearman's ρ_s. Supplementary material, including simulation study results, R-code for each method and the example data used, is available at https:https://osf.io/gny35/. In the near future we intend to make these tests available in the open-source software package JASP (e.g. [26]; jasp-stats.org), which we hope will further increase the possibility that the tests are used to analyze ordinal data sets for which the traditional parametric approach is questionable.

Note

1. More precisely, let θ be a true parameter value and $\alpha, \beta \in (0, 1)$ fixed, then we denote by $N_T(\alpha, \beta, \theta)$ the number of samples necessary for a generic test statistic T at level α to reach the desired power of $1 - \beta$ under θ computed using the asymptotical variance of the test statistic. The ARE of the parametric test over the Mann-Whitney-Wilcoxon U test is defined as $\text{ARE} = N_{\text{par}}(\alpha, \beta, \theta)/N_U(\alpha, \beta, \theta)$.

ORCID

J. van Doorn http://orcid.org/0000-0003-0270-096X
A. Ly http://orcid.org/0000-0003-3925-3833
M. Marsman http://orcid.org/0000-0001-5309-7502

References

[1] J.H. Albert, *Bayesian estimation of normal ogive item response curves using gibbs sampling*, J. Educ. Stat. 17 (1992), pp. 251–269 .
[2] J.H. Albert, *Bayesian estimation of the polychoric correlation coefficient*, J. Stat. Comput. Simul. 44 (1992), pp. 47–61.
[3] J.H. Albert and S. Chib, *Bayesian analysis of binary and polychotomous response data*, J. Am. Stat. Assoc. 88 (1993), pp. 669–679.
[4] R.J. Brooks, *Bayesian analysis of the two-sample problem under the Lehmann alternatives*, Biometrika 61 (1974), pp. 501–507.
[5] R.J. Brooks, *Bayesian analysis of a two-sample problem based on the rank order statistic*, J. R. Stat. Soc. Ser. B 40 (1978), pp. 50–57.
[6] J. Chambers and T. Hastie, *Statistical Models in S*, Wadsworth & Brooks/Cole Computer Science Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, 1992.
[7] H. Colonius, *An invitation to coupling and copulas: With applications to multisensory modeling*, J. Math. Psychol. 74 (2016), pp. 2–10.
[8] W. Conover, *Practical Nonparametric Statistics*, 3rd ed., Wiley, New York, 1999.
[9] P. Cortez and A.M.G. Silva, *Using data mining to predict secondary school student performance*, in *Proceedings of 5th Annual Future Business Technology Conference*, A. Brito and J. Teixeira, eds., EUROSIS, Porto, 2008, pp. 5–12.
[10] E. Cureton, *Rank-biserial correlation*, Psychometrika 21 (1956), pp. 287–290.
[11] J.M. Dickey and B.P. Lientz, *The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain*, Ann. Math. Statist. 41 (1970), pp. 214–226.

[12] Z. Dienes, *Using Bayes to get the most out of non-significant results*, Front. Psychol. 5 (2014), pp. 781.

[13] Z. Dienes and N. McLatchie, *Four reasons to prefer Bayesian analyses over significance testing*, Psychon. Bull. Rev. 25 (2018), pp. 207–218.

[14] R.A. Fisher, *Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population*, Biometrika 10 (1915), pp. 507–521.

[15] M.A. Fligner and J.S. Verducci, *Distance based ranking models*, R. Stat. Soc. Ser. B 48 (1986), pp. 359–369.

[16] M.A. Fligner and J.S. Verducci, *Multistage ranking models*, J. Am. Stat. Assoc. 83 (1988), pp. 892–901.

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

[18] C. Genest and A.C. Favre, *Everything you always wanted to know about copula modeling but were afraid to ask*, J. Hydrol. Eng. 12 (2007), pp. 347–368.

[19] Q.F. Gronau, A. Ly, and E.J. Wagenmakers, *Informed Bayesian t-tests*, preprint (2018), manuscript submitted for publication. Available at arXiv:1704.02479.

[20] W. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika 57 (1970), pp. 97–109.

[21] J. Hodges and E. Lehmann, *The efficiency of some nonparametric competitors of the t-test*, Ann. Math. Statist. 27 (1956), pp. 324–335.

[22] P. Hoff, *Extending the rank likelihood for semiparametric copula estimation*, Ann. Appl. Statist. 1 (2007), pp. 265–283.

[23] P. Hoff, *A First Course in Bayesian Statistical Methods*, Springer-Verlag, New York, 2009.

[24] M. Hollander and D. Wolfe, *Nonparametric Statistical Methods*, 3rd ed., Wiley, New York, 1973.

[25] H. Hotelling and M. Pabst, *Rank correlation and tests of significance involving no assumption of normality*, Ann. Math. Statist. 7 (1936), pp. 29–43.

[26] JASP Team, *JASP (Version 0.11)[Computer software]* (2019). Available at https://jasp-stats.org/.

[27] H. Jeffreys, *Theory of Probability*, 1st ed., Oxford University Press, Oxford, UK, 1939.

[28] H. Jeffreys, *Theory of Probability*, 3rd ed., Oxford University Press, Oxford, UK, 1961.

[29] R.E. Kass and A.E. Raftery, *Bayes factors*, J. Am. Stat. Assoc. 90 (1995), pp. 773–795.

[30] D. Kerby, *The simple difference formula: An approach to teaching nonparametric correlation*, Compr. Psychol. 3 (2014), pp. 1–9.

[31] W. Kruskal, *Ordinal measures of association*, J. Am. Stat. Assoc. 53 (1958), pp. 814–861.

[32] E. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, 1st ed., Holden-Day Series in Probability and Statistics, Holden-Day, Inc., London, New York, 1975.

[33] E. Lehmann, *Elements of Large Sample Theory*, Springer, New York, 1999.

[34] F. Liang, R.P. German, A. Clyde, and J. Berger, *Mixtures of g priors for Bayesian variable selection*, J. Am. Stat. Assoc. 103 (2008), pp. 410–423.

[35] S. Liu and C. Sabatti, *Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation*, Biometrika 87 (2000), pp. 353–369.

[36] A. Ly, A. Etz, M. Marsman, and E.J. Wagenmakers, *Replication Bayes factors from evidence updating*, Behav. Res. Methods 51 (2018), pp. 1–11.

[37] A. Ly, M. Marsman, and E.J. Wagenmakers, *Analytic posteriors for Pearson's correlation coefficient*, Stat. Neerl. 72 (2018), pp. 4–13.

[38] A. Ly, A.J. Verhagen, and E.J. Wagenmakers, *Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology*, J. Math. Psychol. 72 (2016), pp. 19–32.

[39] C.L. Mallows, *Non-null ranking models*, Biometrika 44 (1957), pp. 114–130.

[40] H. Mann and D. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*, Ann. Math. Statist. 18 (1947), pp. 50–60.

[41] J.I. Marden, *Analyzing and Modeling Rank Data*, 1st ed., Monographs on Statistics and Applied Probability, Chapman & Hall, London, New York, 1995.

[42] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys. 21 (1953), pp. 1087–1092.

[43] R.D. Morey and J.N. Rouder, *BayesFactor 0.9.11-1*, Comprehensive R Archive Network (2015). Available at http://cran.r-project.org/web/packages/BayesFactor/index.html.

[44] R.D. Morey, J.N. Rouder, and P.L. Speckman, *A statistical model for discriminating between subliminal and near-liminal performance*, J. Math. Psychol. 52 (2008), pp. 21–36.

[45] R. Nelsen, *An Introduction to Copulas*, 2nd ed., Springer-Verlag, New York, 2006.

[46] U. Olssen, *Maximum likelihood estimation of the polychoric correlation coefficient*, Psychometrika 44 (1979), pp. 443–460.

[47] U. Olssen, F. Drasgow, and N. Dorans, *The polyserial correlation coefficient*, Psychometrika 47 (1982), pp. 443–460.

[48] K. Pearson, *Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable*, Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci. 195 (1900), pp. 1–405.

[49] K. Pearson and E. Pearson, *On polychoric coefficients of correlation*, Biometrika 14 (1922), pp. 127–156.

[50] A. Pettitt, *Inference for the linear model using a likelihood based on ranks*, J. R. Stat. Soc. Ser. B 44 (1982), pp. 234–243.

[51] J.N. Rouder, *Optional stopping: No problem for Bayesians*, Psychon. Bull. Rev. 21 (2014), pp. 301–308.

[52] J.N. Rouder, P.L. Speckman, D. Sun, R.D. Morey, and G. Iverson, *Bayesian t tests for accepting and rejecting the null hypothesis*, Psychon. Bull. Rev. 16 (2009), pp. 225–237.

[53] I. Savage, *Contributions to the theory of rank order statistics-the two-sample case*, Ann. Math. Statist. 27 (1956), pp. 590–615.

[54] A. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publications de l'Institut de Statistique de L'Université de Paris 8 (1959), pp. 229–231.

[55] C. Spearman, *The proof and measurement of association between two things*, Am. J. Psychol. 15 (1904), pp. 72–101.

[56] M.A. Tanner and W.H. Wong, *The calculation of posterior distributions by data augmentation*, J. Am. Stat. Assoc. 82 (1987), pp. 528–540.

[57] P.F. Thall and S.C. Vail, *Some covariance models for longitudinal count data with overdispersion*, Biometrics 46 (1990), pp. 657–671.

[58] A. van der Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2000.

[59] J. van Doorn, A. Ly, M. Marsman, and E.J. Wagenmakers, *Bayesian estimation of Kendall's tau using a latent normal approach*, Stat. Probab. Lett. 145 (2019), pp. 268–272.

[60] D.A. van Dyk and X.L. Meng, *The art of data augmentation*, J. Comput. Graph. Stat. 10 (2001), pp. 1–50.

[61] E.J. Wagenmakers, T. Lodewyckx, H. Kuriyal, and R. Grasman, *Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method*, Cogn. Psychol. 60 (2010), pp. 158–189.

[62] E.J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, J. Verhagen, J. Love, R. Selker, Q.F. Gronau, M. Šmíra, S. Epskamp, D. Matzke, J.N. Rouder, and R.D. Morey, *Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications*, Psychon. Bull. Rev. 25 (2018), pp. 35–57.

[63] R. Wasserstein and N. Lazar, *The ASA's statement on p-values: Context, process, and purpose*, Am. Stat. 70 (2016), pp. 129–133.

[64] F. Wilcoxon, *Individual comparisons by ranking methods*, Biometrics Bull. 1 (1945), pp. 80–83.