AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR

Journal of Geophysical Research: Solid Earth

RESEARCH ARTICLE

10.1029/2018JB016081

**Special Section:**
Magnetism in the Geosciences - Advances and Perspectives

**Key Points:**

- Tests for a common mean paleomagnetic direction are placed into a Bayesian framework
- Test results naturally incorporate uncertainty when only small sets of observations are available
- The Bayesian test for a common mean direction is illustrated with a number of case studies

**Supporting Information:**

- Supporting Information S1
- Data Set S1
- Data Set S2
- Data Set S3
- Data Set S4
- Data Set S5
- Data Set S6

**Correspondence to:**
D. Heslop,
david.heslop@anu.edu.au

**Citation:**
Heslop, D., & Roberts, A. P. (2018). Revisiting the paleomagnetic reversal test: A Bayesian hypothesis testing framework for a common mean direction. *Journal of Geophysical Research: Solid Earth*, *123*, 7225–7236. https://doi.org/10.1029/2018JB016081

Received 9 MAY 2018
Accepted 15 AUG 2018
Accepted article online 22 AUG 2018
Published online 13 SEP 2018



# Revisiting the Paleomagnetic Reversal Test: A Bayesian Hypothesis Testing Framework for a Common Mean Direction

**D. Heslop[1] and A. P. Roberts[1]**

[1]Research School of Earth Sciences, The Australian National University, Canberra, ACT, Australia

**Abstract** For over 60 years a suite of field tests has provided paleomagnetists with a means to evaluate the timing of magnetic remanence acquisition and subsequent remanence stability. These tests are crucial for detecting potential overprinting by secondary remanences and for assessing the fidelity of the geological information carried by paleomagnetic signals. The reversal test was developed to detect secondary remanent magnetizations that could potentially bias paleomagnetic reconstructions.
More recently, the reversal test has been applied to a broader set of problems, which require statistical assessment of common or antipodal paleomagnetic directions. From a statistical standpoint, the reversal test must distinguish whether two sets of paleomagnetic directions originate from populations with a common mean. However, earlier work has demonstrated that the reversal test may be ambiguous for small numbers of observations because insufficient information is available to reject the null hypothesis of a common mean direction. Here we develop a Bayesian framework to estimate directly the probability that two Fisher-distributed sets of directions originate from populations with a common mean. This framework can be used to consider data sets with common or different precisions and, thus, provides a fully probabilistic version of the parametric reversal test. Additionally, adoption of a Bayesian framework means that ambiguity associated with the lack of information provided by small numbers of observations is incorporated into the final probability estimate in a natural way. Our new Bayesian test is demonstrated with numerical examples and case studies.

## 1. Introduction

Paleomagnetic field tests provide a framework to assess the origin and geological stability of magnetic remanences preserved in rocks and sediments. Within this framework, the reversal test was developed to detect the presence of secondary magnetizations that could bias paleomagnetic directions (Cox & Doell, 1960). Specifically, if a collection of paleomagnetic samples spans a sufficiently long time interval to have recorded both normal and reversed polarities, the mean directions of the groups of normal and reversed polarity directions are expected to have antipodal means if the field is dominantly axisymmetric, such as in the case of a dipole. In the presence of a secondary magnetization, however, the normal and reversed polarity mean directions will no longer be antipodal (Cox & Doell, 1960). Thus, the reversal test provides a framework to detect secondary magnetizations by testing for antipodal mean directions. Beyond this original purpose, application of the reversal test has been extended to identification of polarity reversals (Strik et al., 2003) and hypotheses testing in paleogeographic reconstructions (Swanson-Hysell et al., 2009).

Rather than testing for antipodal directions directly, the reversal test involves inverting one group of paleomagnetic directions (e.g., with either normal or reversed polarities) and testing if the directions come from populations with a common mean. Under the assumption that directions are Fisher (1953) distributed, McFadden and Lowes (1981) and Watson (1983) developed parametric null hypothesis significance tests for a common mean direction for populations with common and different precisions, respectively.

Experimental studies and empirical geomagnetic field models have demonstrated that paleomagnetic directions may not be Fisher (1953) distributed, whereas virtual geomagnetic poles (VGPs) may conform more closely to a Fisher (1953) distribution (Beck Jr., 1999; Cox, 1970; Creer et al., 1959; Creer, 1962; Deenen et al., 2011; Tauxe & Kent, 2004). In such cases, VGPs can be employed directly in the parametric tests of McFadden and Lowes (1981) and Watson (1983). It is important to use standard statistical techniques, such as quantile-quantile plots (Fisher et al., 1987), to assess whether palaeomagnetic directions (or their corresponding VGPs) conform with a Fisher (1953) distribution before undertaking parametric analysis underpinned by an assumption of Fisher (1953) distributed data.

Tauxe et al. (1991) proposed an alternative bootstrap approach, which removed the requirement to assume a Fisher (1953) distribution of directions. This approach was used to develop a reversal test where bootstrap-derived confidence intervals for the means of two sets of directions could be compared to assess if they are statistically indistinguishable (Tauxe, 2010). While the bootstrap approach of Tauxe (2010) removes the need to assume a Fisher (1953) distribution, it assumes that the data sets are representative of their underlying populations, which may require more specimens than parametric approaches. Furthermore, the overlap of confidence regions about a mean direction may not provide a rigorous basis to test for common mean directions (Ryan & Leadbetter, 2002).

The parametric tests of McFadden and Lowes (1981) and Watson (1983) adopt a null hypothesis ($H_0$) of populations with a common mean direction. If the observed data are sufficiently inconsistent with $H_0$, the hypothesis is rejected in favor of an alternative hypothesis ($H_1$) that the mean directions are different. This decision is made based on comparing the angle ($\gamma_o$) between the mean directions of the two groups of observations and a critical angle ($\gamma_c$) at which the null hypothesis of common mean directions can be rejected at a specified significance level. Subsequently, McFadden and McElhinny (1990) showed that such a null hypothesis significance testing framework can be inappropriate because an inability to reject the null hypothesis may result from a lack of data. Therefore, a spurious inference of antipodal directions may be due to insufficient information rather than due to truly antipodal directions. This constraint is particularly relevant in paleomagnetism, where only a small number of directions may be available due to logistical constraints. To quantify this ambiguity, McFadden and McElhinny (1990) developed a reversal test classification scheme based on the critical angle at which the null hypothesis would be rejected. In this way, $\gamma_c$ is used as a measure of the amount of information included in a reversal test, with lower $\gamma_c$ values indicating the presence of more information (i.e., more data or observations from distributions with greater precisions, or both). McFadden and McElhinny (1990) proposed that for a *positive* or *passed* test where the null hypothesis of a common mean direction cannot be rejected, the result should be classified as *A* if $\gamma_c \leq 5°$, *B* if $5° < \gamma_c \leq 10°$, *C* if $10° < \gamma_c \leq 20°$, and *indeterminate* (I) if $\gamma_c > 20°$. If $\gamma_0 > \gamma_c$, the test is negative and is given an F classification (failed).

Null hypothesis significance tests estimate the probability of obtaining a test statistic as extreme as that produced by the data under the assumption that the null hypothesis is correct. The tests of McFadden and Lowes (1981) and Watson (1983) are, therefore, based on estimating the probability of the observations occurring under the assumption that $H_0$ is true. We propose, however, that a more powerful approach is to estimate directly the probability that $H_0$ is true given the observations. In this work, we use the rule of Bayes (1763) to determine the relative plausibility of competing hypotheses corresponding to populations of paleomagnetic directions with a common or different mean directions. This approach complements the statistical tests of McFadden and Lowes (1981) and Watson (1983) and the classification scheme of McFadden and McElhinny (1990) by providing an estimate of the probability that two sets of observed directions share a common mean. Such a scheme provides a probabilistic framework for interpreting the reversal test, which provides more quantitative insight than the binary *pass* or *fail* outcomes of existing tests. Additionally, our proposed Bayesian approach naturally incorporates the amount of available information into the probability estimate.

## 2. A Bayesian Hypothesis Testing Framework

Following the theorem of Bayes (1763), the posterior probability of a hypothesis, $H$, given a collection of observations, $X$, is

$$p(H|X) = \frac{p(H)p(X|H)}{p(X)}, \tag{1}$$

where $p(X)$ is the probability of observing $X$ independently of any specific hypothesis and $p(H)$ is the probability of hypothesis $H$ being true prior to any observations being made. The bar in the probability terms, such as $p(H|X)$ and $p(X|H)$, indicates conditional dependence, where quantities on the left and right are variable and

**Table 1**
*Categories of Support Corresponding to Different Probabilities, $p(H_A|X)$, of Hypothesis $H_A$ That Two Sets of Directions Are Drawn From Populations With a Common Mean Versus a Hypothesis $H_B$ of Different Means (Modified From Raftery, 1995)*

| BF | $p(H_A\|X)$ | Support |
|---|---|---|
| <1/99 | <0.01 | Different means: very strong support |
| 1/99–<1/19 | 0.01–<0.05 | Different means: strong support |
| 1/19–<1/3 | 0.05–<0.25 | Different means: positive support |
| 1/3–<3 | 0.25–<0.75 | Ambiguous: weak support |
| 3–<19 | 0.75–<0.95 | Common mean: positive support |
| 19–<99 | 0.95–<0.99 | Common mean: strong support |
| >99 | ≥0.99 | Common mean: very strong support |

*Note*. BF = Bayes factor.

fixed, respectively. This can be extended to estimate the relative probabilities of two competing hypotheses, $H_A$ and $H_B$ (Gallagher et al., 2009; Sambridge et al., 2006):

$$\frac{p(H_A|X)}{p(H_B|X)} = \frac{p(H_A)}{p(H_B)} \cdot \frac{p(X|H_A)}{p(X|H_B)}. \tag{2}$$

The final term in equation (2) is the Bayes factor (BF), which is a measure of the extent to which the observations in $X$ change the prior probabilities: $p(H_A)$ and $p(H_B)$, of the hypotheses. The BF is given by (Burnham & Anderson, 2002):

$$\mathrm{BF}(H_A, H_B) = \frac{p(X|H_A)}{p(X|H_B)} = \frac{p(H_A|X)}{p(H_B|X)} \Big/ \frac{p(H_A)}{p(H_B)}. \tag{3}$$

BF consists of the marginal likelihoods of the individual hypotheses, which are obtained by integrating over the corresponding parameter space (Kass & Raftery, 1995):

$$p(X|H_i) = \int p(X|\theta_i, H_i)p(\theta_i|H_i)\mathrm{d}\theta_i \qquad (i = A, B), \tag{4}$$

where $\theta_i$ represents the model parameters of hypothesis $i$, $p(\theta_i|H_i)$ is the parameter prior densities, and $p(X|\theta_i, H_i)$ is the probability density of $X$ given $\theta_i$. When the two hypotheses are assigned equal prior probabilities, $p(H_A) = p(H_B) = 0.5$, the posterior probabilities of the hypotheses become

$$p(H_A|X) = \frac{\mathrm{BF}}{\mathrm{BF}+1} \qquad \text{and} \qquad p(H_B|X) = \frac{1}{\mathrm{BF}+1}. \tag{5}$$

Categories that describe the level of *support* the data provide for each hypothesis can then be assigned based on a modified version of the scheme of Raftery, (1995; Table 1). It is important to note that the categories of support proposed by Raftery (1995) were constructed in a social science context based on the *rules of thumbs* developed by Jeffreys (1961). Therefore, while these categories provide an indicative framework to interpret BF, they do not represent a collection of fundamental boundaries.

In the following section we show how a test for a common mean direction can be constructed within the presented Bayesian framework. We develop separate approaches to compare samples from Fisher (1953) distributed populations with common and different precisions. Although null hypothesis approaches are available to test for a common precision (Watson, 1956), we recommend that the decision of which approach to use should be based on a priori geological information. The examples in section 3 demonstrate this, whereby comparisons involving short time scales over which the geomagnetic field is not expected to have fundamentally changed its configuration and tectonic effects are expected to be minimal can be compared under an assumption of a common precision. In the final case concerning the Osler Volcanic Group, it was considered that either the geomagnetic field or paleolatitudinal position was changing rapidly. Given this setting, a test that does not assume common precisions should be adopted.

## 3. Bayesian Tests for a Common Mean Direction

The parametric tests of McFadden and Lowes (1981) and Watson (1983) for a common mean direction are based on the Fisher (1953) distribution, with the probability density of a unit vector, $x$, given by

$$F(x|\mu,\kappa) = \frac{\kappa}{4\pi \sinh \kappa} \exp\left(\kappa \mu^T x\right), \tag{6}$$

where $\kappa$ is the distribution precision parameter and $\mu$ is a unit vector corresponding to the mean direction. For a collection of $n$ unit vectors, $\mathbf{x}$, their likelihood is given by

$$\mathcal{L} = \prod_{i=1}^{n} F(x_i|\mu,\kappa). \tag{7}$$

Consider the case where $\mu$ is unknown, while $\kappa$ is known and fixed. An appropriate prior for the unknown mean direction is a Fisher (1953) distribution with a mean and precision of $\mu_0$ and $\kappa_0$, respectively (Mardia & El-Atoum, 1976). For a given $\mu$ the combination of the prior and the data yields

$$p(\mu|\mathbf{x},\kappa,\mu_0,\kappa_0) \propto F(\mu|\mu_0,\kappa_0) \prod_{i=1}^{n} F(x_i|\mu,\kappa). \tag{8}$$

We employ a prior with $\kappa_0 = 0$, which corresponds to a uniform distribution across the surface of a unit sphere. This implies that we have no a priori information concerning the mean direction of the population from which the observations in $\mathbf{x}$ were drawn. For compactness we employ the notation $p(\mu|\mathbf{x},\kappa) = p(\mu|\mathbf{x},\kappa,\mu_0,0)$.

Determination of the marginal likelihood requires integration across the parameter space of $\mu$ (i.e., the surface of a unit sphere). For simplicity we denote integration over the surface of a unit sphere with polar angle $\theta$ and azimuthal angle $\phi$:

$$\int_{S^2} f(\mu)\,\mathrm{d}\mu = \int_0^{2\pi} \int_0^{\pi} f(\theta,\phi) \sin\theta \,\mathrm{d}\theta\,\mathrm{d}\phi. \tag{9}$$

Thus, the marginal likelihood when $\kappa$ is known is written as

$$\int_{S^2} p(\mu|\mathbf{x},\kappa)\,\mathrm{d}\mu. \tag{10}$$

In paleomagnetic investigations the distribution precision is typically unknown and, therefore, we adopt the prior for $\kappa$ recommended by Dowe et al. (1996):

$$h(\kappa) = \frac{4\kappa^2}{\pi(1+\kappa^2)^2}. \tag{11}$$

With inclusion of $\kappa$ as an unknown, the marginal likelihood becomes

$$\int_{\kappa=0}^{\infty} \int_{S^2} p(\mu|\mathbf{x},\kappa)\,h(\kappa)\,\mathrm{d}\mu\,\mathrm{d}\kappa. \tag{12}$$

Using this framework we develop Bayesian tests for Fisher (1953) distributed populations with common and different $\kappa$.

### 3.1. Comparing Populations With a Common Precision

If two sets of directions, $\mathbf{x}_1$ and $\mathbf{x}_2$, are drawn from a single Fisher (1953)-distributed population (i.e., they share a common $\mu$ and $\kappa$), then the pooled directions (i.e., all the directions combined into a single group), $\mathbf{x}_{12}$, will correspond to the same population. Therefore, a BF-based test can be constructed by comparing the marginal likelihoods of the separate and pooled data sets. The hypothesis of a common mean direction ($H_A$) is represented by the marginal likelihood of the pooled directions (numerator of BF):

$$\int_{\kappa=0}^{\infty} \int_{S^2} p(\mu|\mathbf{x}_{12},\kappa)\,\mathrm{d}\mu\,h(\kappa)\,\mathrm{d}\kappa. \tag{13}$$

10.1029/2018JB016081

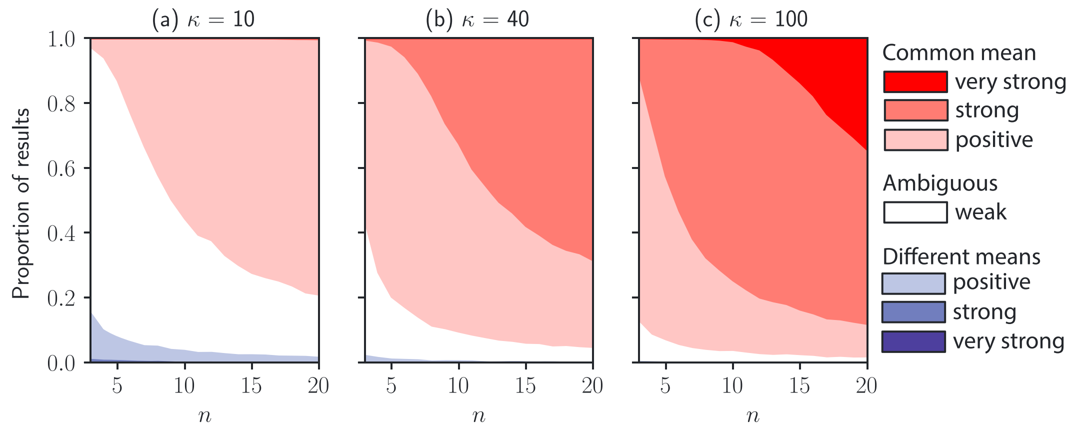


**Figure 1.** Demonstration of the manner in which information is incorporated into the proposed Bayesian test. Samples of size $n$ were drawn from a Fisher (1953)-distributed population with a fixed $\mu$ and $P(H_A|X)$ was estimated using the proposed procedure for common precisions. This process was repeated $10^4$ times for given precisions ($\kappa = 10$ in panel [a], $\kappa = 40$ in panel [b], and $\kappa = 100$ in panel [c]) and the results are classified according to the modified scheme of Raftery (1995; Table 1). Colors represent the categories of support, with darker shades corresponding to stronger support for $H_A$ (red) or $H_B$ (blue). As sample size increases, the information provided by the data overcomes the initial prior that the two hypotheses are equally probable. This effect depends on the precision of the distributions under consideration, with higher (lower) precisions requiring lower (higher) $n$ to achieve the same level of support for a given hypothesis.

A hypothesis of populations with different $\mu$, but common $\kappa$ ($H_B$) is given by the marginal likelihood (denominator of BF):

$$\int_{\kappa=0}^{\infty} \int_{S^2} p(\mu_1|\mathbf{x}_1, \kappa)\,\mathrm{d}\mu_1 \int_{S^2} p(\mu_2|\mathbf{x}_2, \kappa)\,\mathrm{d}\mu_2\, h(\kappa)\, d\kappa, \tag{14}$$

where $\mu_1$ and $\mu_2$ correspond to the population mean directions of $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, and integration takes place for a common $\kappa$. Description of the calculation of equations (13) and (14) is provided in Appendix A and a Jupyter Notebook to perform the calculation in Python (Version 3) is provided as supporting information.

We present a simple numerical example to demonstrate how our proposed approach represents the amount of available information in a natural way. Two sets of $n$ directions were drawn at random from the same Fisher (1953) distribution with prespecified $\mu$ and $\kappa$ (thus, both samples contain directions originating from the same population). Equations (13) and (14) were used to estimate $p(H_A|X)$, and the result was categorized according to the modified scheme of Raftery (1995; Table 1). This procedure was repeated $10^4$ times for $\kappa = 10$, 40, and 100 (a range of precisions broadly typical of paleomagnetic data sets). When $\kappa$ is low (i.e., scattered directions) and $n$ is small (Figure 1a) the majority of tests produce an ambiguous result because there is insufficient information to support either $H_A$ or $H_B$ and the initial $p(H_A) = p(H_B) = 0.5$ prior plays a strong role. As $n$ increases, the information available to the test increases and the proportion of results corresponding to positive support for a common mean direction increases. However, because of the low concentration, cases corresponding to strong and very strong support remain absent at $n = 20$. When $\kappa = 40$ (Figure 1b), the sample directions are less scattered and greater support for a common mean direction is achieved at lower $n$. As expected, as $n$ increases the level of support for a common mean direction increases with a growing number of results achieving strong support. For $\kappa = 100$ (Figure 1c), directions are concentrated around the mean and positive and strong support dominates even when $n$ is small. Once $n > 9$, results appear with very strong support for a common mean direction and their frequency increases with $n$. These examples provide a clear demonstration of the ability of our proposed approach to incorporate the amount of available information into the estimate of $p(H_A|X)$.

To further demonstrate that our proposed Bayesian approach naturally incorporates the level of available information, we employed a Monte Carlo simulation to compare estimated $p(H_A|X)$ to the reversal test classifications of McFadden and McElhinny (1990). This involved drawing two sets of random directions from the same (Fisher, 1953) distribution to generate samples for comparison that are known a priori to originate from the same population. For each iteration of the simulation a precision was selected for the population with $\kappa$ drawn randomly from a uniform distribution between 10 and 100 (i.e., a range precisions that are typical of

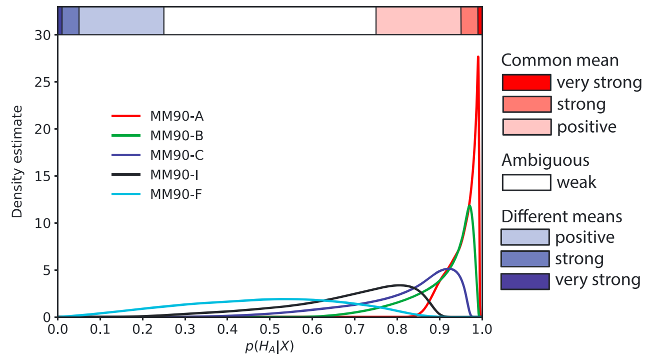


**Figure 2.** Pairs of differently sized samples of directions were drawn from a Fisher (1953)-distributed population with a randomly selected $\kappa$. The reversal test of McFadden and Lowes (1981) was performed on each pair of samples and the result classified according to the scheme of McFadden and McElhinny (1990). An estimate of $P(H_A|X)$ was then made based on the same samples assuming a common precision. Probability density estimates were made based on the distributions of $P(H_A|X)$ given their corresponding McFadden and McElhinny (1990) classification. The colored bar along the top of the figure indicates the categories of support corresponding to different $P(H_A|X)$ values according to the modified scheme of Raftery (1995; Table 1).

paleomagnetic data sets). Samples consisting of $n_1$ and $n_2$ directions were then drawn from the distribution, where $n_1$ and $n_2$ are random integers between 3 and 20 (i.e., sample sizes that are typical of paleomagnetic data sets). The angle between the mean sample directions ($\gamma_0$) was then compared to the critical angle ($\gamma_c$) and assigned a reversal test classification according to the scheme of McFadden and McElhinny (1990). An estimate of $p(H_A|X)$ was then made for the same sample directions. This process was repeated $10^6$ times, and the distribution of $p(H_A|X)$ values for each McFadden and McElhinny (1990) classification was evaluated using kernel density estimation (Silverman, 1986).

The $p(H_A|X)$ distributions are consistent with the McFadden and McElhinny (1990) classification scheme (Figure 2). Cases that result in a failed test (F classification) have broadly distributed values about $p(H_A|X) \approx 0.5$, with only a small number of the cases achieving levels of support beyond the ambiguous range. Approximately 40% of the I-classified tests achieved positive support for a common mean, but the remainder were ambiguous with a small number ($< 1\%$) corresponding to positive support for different mean directions. The majority (69%) of the C-classified tests correspond to positive support for a common mean direction, with a small proportion (5%) achieving strong support. The majority (58%) of the B-classified tests correspond to positive support for a common mean direction, with 36% achieving strong support. Finally, 54% and 9% of the A-classified tests achieve strong and very strong support for a common mean direction, respectively. Thus, estimates of $p(H_A|X)$ are consistent with the McFadden and McElhinny (1990) classification scheme and demonstrate that the level of information available in the analysis is naturally incorporated into our Bayesian approach. It is important to reiterate that the null hypothesis test and our proposed Bayesian approach estimate different things. The null hypothesis test estimates the probability of observing the data assuming that the null hypothesis of a common mean direction is true, while the Bayesian approach estimates the probability of a common mean direction based on the observed data.

To demonstrate the use of our proposed test on real data we consider two published case studies that employed parametric reversal tests. These examples are presented solely for illustrative purpose and are not presented as critiques of the original studies in which they were reported. Jupyter Notebooks documenting these examples are provided as supporting information.

#### 3.1.1. Pleistocene Magnetostratigraphy of the Nihewan Basin

Hong et al. (2013) developed a detailed magnetostratigraphy for fluvio-lacustrine sequences of the Nihewan Basin, North China, to date Paleolithic sites and fauna preserved in the sediments. Thermal demagnetization of the natural remanent magnetization revealed a low-temperature secondary component and a higher-temperature univectoral magnetization that is directed toward the origin of a Zijderveld (1967) orthogonal plot. The high-temperature component was deemed to be the primary remanent magnetization and was attributed to (potentially partially oxidized) magnetite and hematite. Primary paleomagnetic directions were estimated by principal component analysis (Kirschvink, 1980) and considered to be reliable if they

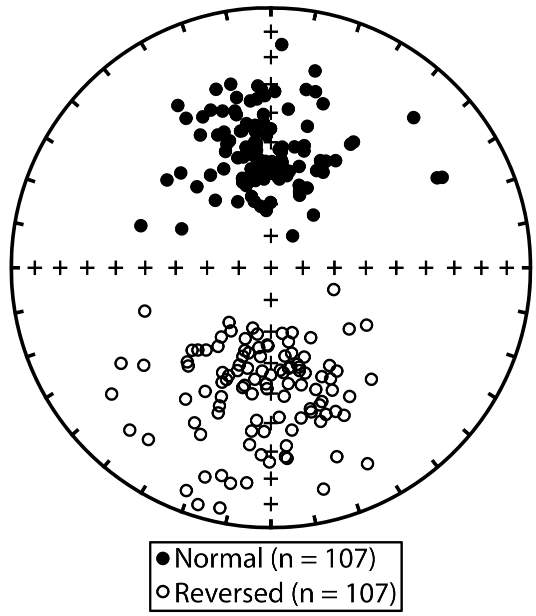


**Figure 3.** Normal and reversed polarity paleomagnetic directions from the study of Hong et al. (2013). The data pass the reversal test of McFadden and Lowes (1981) with a *B* classification (McFadden & McElhinny, 1990). Under an assumption of common $\kappa$, the described Bayesian test estimates a probability of 0.93 that the normal and reversed polarity directions share a common mean. This corresponds to positive support in the scheme of Raftery (1995).

were based on at least four consecutive demagnetization steps at $\geq$ 250°C with a maximum angular deviation of $\leq$ 15° and VGP latitudes $>$ 30° or $< -30°$ . These selection criteria were met by 107 samples with normal polarity directions and 107 samples with reversed polarity directions (Figure 3). Hong et al. (2013) performed the parametric reversal test of McFadden and Lowes (1981) under an assumption of common $\kappa$ to test if the normal and reversed polarity directions are antipodal. This corresponds to the original purpose of the reversal test (Cox & Doell, 1960) and the data pass with a *B* classification according to the scheme of McFadden and McElhinny (1990).

Estimation of BF for the data of Hong et al. (2013) is based on equations (13) and (14) and results in BF $\approx$ 12.74. This implies that $H_A$ is supported ~12.74 times more by the data than $H_B$ and corresponds to $P(H_A|X) = 0.93$, which indicates positive support for antipodal mean directions (see Table 1). This example demonstrates that even though the groups of normal and reversed polarity paleomagnetic directions are dispersed ($\kappa < 20$), positive support can be achieved because of the large number of observations involved in the test.

### 3.1.2. Late Archean Paleomagnetic Reversals

In a study of Late Archean flood basalts from the Pilbara Craton in Western Australia, Strik et al. (2003) inferred the presence of two paleomagnetic reversals within the sampled zones of a 60-million-year succession. High-temperature remanence components isolated through progressive thermal demagnetization were interpreted as primary magnetizations. In a sequence of lava packages through the succession these high-temperature components were found to be approximately antipodal. Package mean directions (Figure 4) pass the reversal test of McFadden and Lowes (1981) with a B classification, which was interpreted as indicating the oldest known geomagnetic field reversal (Strik et al., 2003).

Following the approach of Strik et al. (2003), it is reasonable to assume when identifying a potential geomagnetic reversal that the normal and reversed polarity directions should share a common $\kappa$. For the Strik et al. (2003) flood basalt site mean directions, BF $\approx$ 3.48 (equations (13) and (14), so that $H_A$ is supported ~3.48 times more by the data than $H_B$. This corresponds to $P(H_A|X) = 0.78$, which indicates positive support for antipodal mean directions (see Table 1). In this example, the groups have high precisions ($\kappa >150$), but only eight observations are included in the test. While the logistical constraints involved in such a study are clear, the small size of the data set limits the information available in the test, which results in a $P(H_A|X)$ that only just reaches positive support.

10.1029/2018JB016081

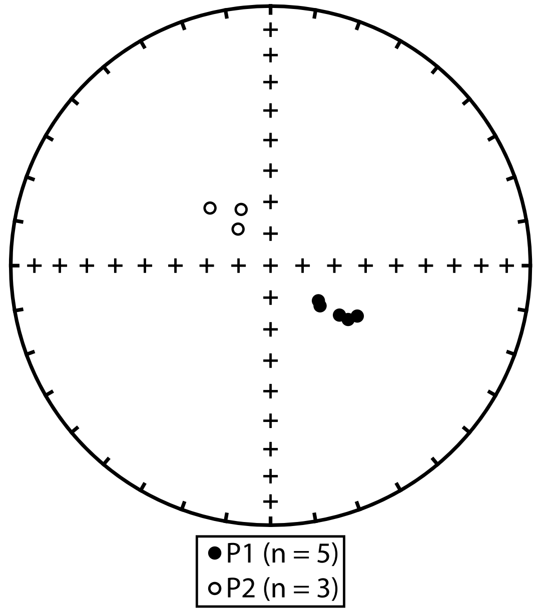


**Figure 4.** Site mean directions for Package 1 and Package 2 of the Nullagine Synclinorium from Strik et al. (2003). These data pass the reversal test of McFadden and Lowes (1981) with a *B* classification (McFadden & McElhinny, 1990). Under an assumption of common $\kappa$ the described Bayesian test estimates a probability of 0.78 that the Package 1 and Package 2 data sets have antipodal mean directions. This corresponds to positive support in the modified scheme of Raftery (1995).

### 3.2. Comparing Populations With Different Precisions

When comparing sets of directions that are assumed to come from populations with different precisions, hypothesis $H_A$ corresponds to sets of directions $\mathbf{x}_1$ and $\mathbf{x}_2$ that originate from Fisher (1953)-distributed populations with a common mean but different $\kappa$. If $\mu$ represents the common mean direction, while $\kappa_1$ and $\kappa_2$ represent the population precisions of $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, the marginal likelihood (numerator of BF) is

$$\int_{\kappa_2=0}^{\infty}\int_{\kappa_1=0}^{\infty}\int_{S^2} p(\mu|\mathbf{x}_1,\kappa_1)\, p(\mu|\mathbf{x}_2,\kappa_2)\,\mathrm{d}\mu\, h(\kappa_1)\, h(\kappa_2)\,\mathrm{d}\kappa_1\,\mathrm{d}\kappa_2. \tag{15}$$

Hypothesis $H_B$ corresponds to populations with different means (denoted as $\mu_1$ and $\mu_2$) and different precisions ($\kappa_1$ and $\kappa_2$). Thus, the denominator of BF is given by the marginal likelihood:

$$\int_{\kappa_1=0}^{\infty}\int_{S^2} p(\mu_1|\mathbf{x}_1,\kappa)\,\mathrm{d}\mu_1\, h(\kappa_1)\,\mathrm{d}\kappa_1 \int_{\kappa_2=0}^{\infty}\int_{S^2} p(\mu_2|\mathbf{x}_2,\kappa)\,\mathrm{d}\mu_2\, h(\kappa_2)\,\mathrm{d}\kappa_2. \tag{16}$$

A description of the calculation of equations (15) and (16) is provided in Appendix B, and a Jupyter Notebook for performing the described calculation in Python is provided as supporting information. A demonstration of this test is given in the following case study, with an accompanying Jupyter Notebook provided in the supporting information.

#### 3.2.1. Osler Volcanic Group, Ontario, Canada

Paleomagnetic investigations of North American Midcontinent Rift volcanics and intrusives reveals a significant directional difference between steep (predominantly reversed polarity) inclinations in the oldest units and shallower (predominantly normal polarity) inclinations in the younger units (Halls & Pesonen, 1982). Different studies attribute this inclination inconsistency to either significant nondipole field contributions (Pesonen & Nevanlinna, 1981) or to rapid plate motion (Davis & Green, 1997; Robertson & Fahrig, 1971) approximately 1.1 billion years ago. A detailed paleomagnetic study by Swanson-Hysell et al. (2009) revealed a sustained inclination decrease across multiple field reversals through a succession of lavas, which was interpreted to indicate that the observed paleomagnetic signal is attributable to paleolatitudinal change rather than to nondipole field components. To further test the hypothesis of progressive paleolatitudinal change, Swanson-Hysell et al. (2014) studied 84 lava flows of the Olser Volcanic Group and partitioned flow paleomagnetic directions based on stratigraphic position. Three groups were used for the lower (0–1,041 m above

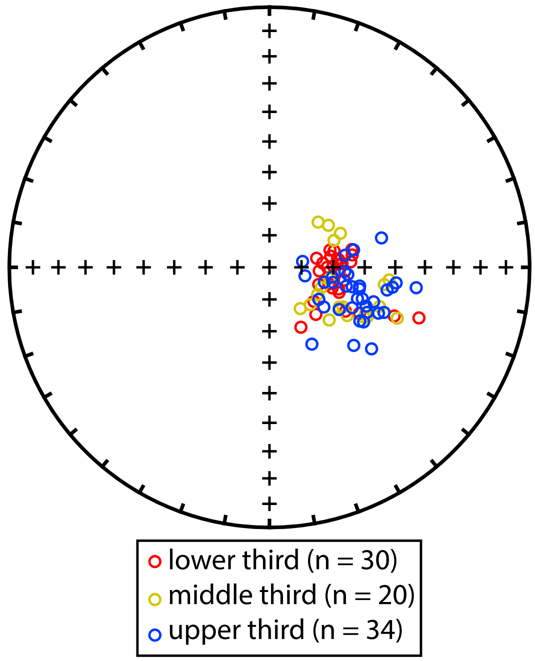


**Figure 5.** Paleomagnetic directions for the lower, middle, and upper thirds of the Osler Volcanic Group from Swanson-Hysell et al. (2014). Tests comparing data for the lower to middle third of the sequence and the middle to upper third of the sequence both pass the test of Watson (1983) with *B* classifications (McFadden & McElhinny, 1990). A Watson (1983) test that compares data for the lower and upper thirds of the sequence results in a fail. Under an assumption of populations with different $\kappa$ values, the outlined Bayesian test estimates a probability of 0.96 that data for the lower and middle thirds of the sequence share a common mean direction, which corresponds to strong support. The probability that data for the middle and upper thirds of the sequence share a common mean direction is estimated as 0.97, which also corresponds to strong support. Finally, the probability that data for the lower and upper thirds of the sequence share a common mean direction is estimated as 0.15, which corresponds to positive support for different mean directions.

the base of the studied sequence, $n = 30$), middle (1,041–2,083 m, $n = 20$), and upper (2,083–3,124 m, $n = 34$) thirds of the sequence (Figure 5). Watson (1983) tests for a common mean were then performed to assess whether the group mean directions are statistically indistinguishable assuming that they do not share a common $\kappa$. The test results demonstrate that while mean directions for the lower and middle groups are indistinguishable and that the mean directions for the middle and upper groups are indistinguishable, the mean directions for the lower and upper groups are significantly different. This result was taken as support for a paleolatitudinal, rather than nondipole field, origin for the inclination change through the Osler Volcanic group (Swanson-Hysell et al., 2014).

We have repeated the analysis of Swanson-Hysell et al. (2014) using equations (15) and (16) to estimate BF for populations assumed to not share a common $\kappa$. Comparison of directions from the lower and middle thirds of the sequence yields BF $\approx$ 21.5, which corresponds to $P(H_A|X) = 0.96$, which indicates that there is strong support for a common mean direction. Comparison of directions for the middle and upper thirds of the sequence yields BF $\approx$ 30.1, which corresponds to $P(H_A|X) = 0.97$, which again indicates strong support for a common mean direction. In contrast, comparison of data for the lower and upper thirds of the sequence result in BF $\approx$ 0.18, with $P(H_A|X) = 0.15$, which corresponds to positive support for different mean directions. In this example, all three groups are moderately sized and have reasonably high precisions ($\kappa > 35$). In combination, the properties provide sufficient information for the series of three tests to produce conclusive results.

## 4. Discussion and Conclusions

An ability to compare mean paleomagnetic directions or VGPs is a key inferential step in many paleomagnetic studies. McFadden and McElhinny (1990) demonstrated a shortcoming in reversal tests based on a null hypothesis significance testing framework. For small or dispersed data sets, insufficient information may be available to reject the null hypothesis of a common mean direction. This would lead to a spurious pass of the reversal test that could result in an incorrect paleomagnetic interpretation. The classification system of McFadden and McElhinny (1990) goes some way to resolving this issue, but ambiguity remains as to how to interpret the A, B, and C classifications of positive reversal tests in a probabilistic manner.

We have demonstrated how a Bayesian framework can be used to compare hypotheses that represent common and different population mean directions for paleomagnetic directions, or their corresponding VGPs, that are Fisher (1953) distributed. Rather than the binary pass or fail outcome of null hypothesis-based reversal tests, our approach provides a probability that two data sets share a common mean direction. Additionally, by assigning equal prior probabilities, $p(H_A) = p(H_B) = 0.5$, to the two hypotheses implies that sufficient information (i.e., data) must be available to achieve higher levels of support for either hypothesis. When insufficient information is available, the $p(H_A) = p(H_B) = 0.5$ prior will dominate and will result in an inconclusive outcome. In this way, inclusion of information as employed in the classification system of McFadden and McElhinny (1990) is represented automatically in the Bayesian framework in a natural way.

Our proposed approach is underpinned by an assumption of Fisher (1953) distributed data, which may not be appropriate in all cases. As discussed by Deenen et al. (2011) there are different situations when paleomagnetic directions can and cannot be assumed to Fisher (1953) distributed. For example, Deenen et al. (2011) noted that spot readings of the paleomagnetic field recorded by a lava will include random errors that can be expected to conform to a Fisher (1953) distribution, whereas directional scatter resulting from paleosecular variation of the geomagnetic field is known to deviate from a Fisher (1953) distribution. Thus, before applying the proposed Bayesian approach, investigators should employ standard techniques (Fisher et al., 1987) to assess whether their data are approximately Fisher (1953) distributed. If this is not the case, and $n$ is sufficiently large, then a nonparametric bootstrap approach should be adopted (Tauxe et al., 1991). With this caveat in mind, the flexibility of the framework outlined in section 2 should also be emphasized. This approach is not limited to Fisher (1953) distributions and an alternative spherical probability distribution that is deemed to be more appropriate to represent paleomagnetic data could be substituted into the same Bayesian hypothesis testing framework.

Our new approach complements the work of McFadden and Lowes (1981), Watson (1983), and McFadden and McElhinny (1990), but it also provides an advance by enabling a fully probabilistic analysis to support quantitative paleomagnetic reconstructions.

## Appendix A: Populations With a Common Precision

As discussed in section 3.1, representation of the hypothesis of a common population mean requires calculation of the marginal likelihood of the pooled directions ($\mathbf{x}_{12}$):

$$\int_{\kappa=0}^{\infty} \int_{S^2} p(\mu|\mathbf{x}_{12}, \kappa)\, \mathrm{d}\mu\, h(\kappa)\, \mathrm{d}\kappa. \tag{A1}$$

For a Fisher (1953) distribution, if $R$ represents the resultant vector length of the $n$ unit vectors that define $\mathbf{x}_{12}$, for a given value of $\kappa$:

$$\int_{S^2} p(\mu|\mathbf{x}_{12}, \kappa)\mathrm{d}\mu = \frac{4\pi \sinh(\kappa R)}{\kappa R} \left[\frac{\kappa}{4\pi \sinh(\kappa)}\right]^n. \tag{A2}$$

Thus, equation (A1) can be evaluated by numerical quadrature, which integrates the product of equation (A2) and the prior on $\kappa$ (equation (11) for $0 \leq \kappa \leq \infty$. This integration was performed using the scipy.integrate.quad algorithm, which is part of the SciPy package.

The marginal likelihood for sets of directions with different population mean directions and common $\kappa$ is given by

$$\int_{\kappa=0}^{\infty} \int_{S^2} p(\mu_1|\mathbf{x}_1, \kappa)\mathrm{d}\mu_1 \int_{S^2} p(\mu_2|\mathbf{x}_2, \kappa)\, \mathrm{d}\mu_2\, h(\kappa)\, \mathrm{d}\kappa. \tag{A3}$$

Calculation of equation (A3) follows a similar strategy to equation (A1). Specifically, with reference to equation (A2), for a given value of $\kappa$:

$$\int_{S^2} p(\mu_1|\mathbf{x}_1, \kappa)\mathrm{d}\mu_1 \int_{S^2} p(\mu_2|\mathbf{x}_2, \kappa)\mathrm{d}\mu_2 = \frac{16\pi^2 \sinh(\kappa R_1)\sinh(\kappa R_2)}{\kappa^2 R_1 R_2} \left[\frac{\kappa}{4\pi \sinh(\kappa)}\right]^{n_1+n_2}, \tag{A4}$$

where $R_1$ and $R_2$ are the resultant vector lengths of the $n_1$ and $n_2$ unit vectors in $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. As before, equation (A3) can be evaluated by numerical quadrature, which integrates the product of equation (A4) and the prior on $\kappa$ (equation (11) for $0 \leq \kappa \leq \infty$.

## Appendix B: Populations With Different Precisions

As shown in section 3.2, the BF numerator corresponds to the marginal likelihood of two populations with different precisions ($\kappa_1$ and $\kappa_2$) and a common mean ($\mu$):

$$\int_{\kappa_2=0}^{\infty}\int_{\kappa_1=0}^{\infty}\int_{S^2} p(\mu|\mathbf{x}_1,\kappa_1)\,p(\mu|\mathbf{x}_2,\kappa_2)\,\mathrm{d}\mu\,h(\kappa_1)\,h(\kappa_2)\,\mathrm{d}\kappa_1\,\mathrm{d}\kappa_2. \tag{B1}$$

This integral is estimated numerically. The surface integral of the function over a unit sphere is achieved using the Lebedev and Laikov (1999) quadrature scheme (Schlömer, 2018) and integration with respect to the precisions is performed using the scipy.integrate.quad algorithm. The BF denominator is given by the marginal likelihood of two populations with different mean directions ($\mu_1$ and $\mu_2$) and different precisions ($\kappa_1$ and $\kappa_2$):

$$\int_{\kappa_1=0}^{\infty}\int_{S^2} p(\mu_1|\mathbf{x}_1,\kappa)\,\mathrm{d}\mu_1\,h(\kappa_1)\,\mathrm{d}\kappa_1 \int_{\kappa_2=0}^{\infty}\int_{S^2} p(\mu_2|\mathbf{x}_2,\kappa)\,\mathrm{d}\mu_2\,h(\kappa_2)\,\mathrm{d}\kappa_2. \tag{B2}$$

Evaluation of equation (B2) follows the approach outlined in Appendix A. If $R_1$ is the resultant vector length of the $n_1$ unit vectors in $\mathbf{x}_1$ and $R_2$ is the resultant vector length of the $n_2$ unit vectors in $\mathbf{x}_2$, equation (B2) is calculated as

$$\int_{\kappa_1=0}^{\infty}\frac{4\pi\sinh(\kappa_1R_1)}{\kappa_1R_1}\left[\frac{\kappa_1}{4\pi\sinh(\kappa_1)}\right]^{n_1}h(\kappa_1)\,\mathrm{d}\kappa_1\int_{\kappa_2=0}^{\infty}\frac{4\pi\sinh(\kappa_2R_2)}{\kappa_2R_2}\left[\frac{\kappa_2}{4\pi\sinh(\kappa_2)}\right]^{n_2}h(\kappa_2)\,\mathrm{d}\kappa_2. \tag{B3}$$

**Acknowledgments**
This work was supported by the Australian Research Council (through grant DP160100805). We are grateful to Nick Swanson-Hysell and Cor Langereis who provided thoughtful and constructive reviews that improved the paper. All data from the presented examples are available as supporting information.

## References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.
Beck Jr, M. E. (1999). On the shape of paleomagnetic data sets. *Journal of Geophysical Research*, *104*, 25,427–25,441.
Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Berlin: Springer-Verlag Inc.
Cox, A. (1970). Latitude dependence of the angular dispersion of the geomagnetic field. *Geophysical Journal of the Royal Astronomical Society*, *20*, 253–269.
Cox, A., & Doell, R. R. (1960). Review of paleomagnetism. *Bulletin of the Geological Society of America*, *71*, 645–768.
Creer, K. M. (1962). The dispersion of the geomagnetic field due to secular variation and its determination for remote times from paleomagnetic data. *Journal of Geophysical Research*, *67*, 3461–3476.
Creer, K. M., Irving, E., & Nairn, A. E. M. (1959). Paleomagnetism of the Great Whin Sill. *Geophysical Journal of the Royal Astronomical Society*, *2*, 306–323.
Davis, D., & Green, J. (1997). Geochronology of the North America Midcontinent rift in western Lake Superior and implications for its geodynamic evolution. *Canadian Journal of Earth Sciences*, *34*, 476–488.
Deenen, M. H. L., Langereis, C. G., van Hinsbergen, D. J. J., & Biggin, A. J. (2011). Geomagnetic secular variation and the statistics of palaeomagnetic directions. *Geophysical Journal International*, *186*, 509–520.
Dowe, D. L., Oliver, J. J., & Wallace, C. S. (1996). *Algorithmic Learning Theory: 7th International Workshop, ALT '96 Sydney, Australia, October 23–25, 1996 Proceedings* (pp. 213–227). Berlin: Springer. https://doi.org/10.1007/3-540-61863-5_48
Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London A*, *217*, 295–305.
Fisher, N. I., Lewis, T., & Embleton, B. J. J. (1987). *Statistical Analysis of Spherical Data* (p. 329). Cambridge: Cambridge University Press.
Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M., & Stephenson, J. (2009). Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems. *Marine and Petroleum Geology*, *26*, 525–535.
Halls, H., & Pesonen, L. (1982). Paleomagnetism of Keweenawan rocks. *Geological Society of America, Memoir*, *156*, 173–201.
Hong, A., An, Z. S., Dekkers, M. J., Li, Y., Xiao, G., Zhao, H., & Qiang, X. (2013). Pleistocene magnetochronology of the fauna and Paleolithic sites in the Nihewan Basin: Significance for environmental and hominin evolution in North China. *Quaternary Geochronology*, *18*, 78–92.
Jeffreys, H. (1961). *The Theory of Probability* (pp. 447). Oxford: Oxford University Press.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
Kirschvink, J. L. (1980). The least-squares line and plane and the analysis of palaeomagnetic data. *Geophysical Journal of the Royal Astronomical Society*, *62*, 699–718.
Lebedev, V. I., & Laikov, D. N. (1999). A quadrature formula for the sphere of the 131st algebraic order of accuracy. *Doklady Mathematics*, *59*, 477–481.
Mardia, K. V., & El-Atoum, S. A. M. (1976). Bayesian inference for the von Mises-Fisher distribution. *Biometrika*, *63*, 203–206.
McFadden, P. L., & Lowes, F. J. (1981). The discrimination of mean directions drawn from Fisher distributions. *Geophysical Journal of the Royal Astronomical Society*, *67*, 19–33.
McFadden, P. L., & McElhinny, M. W. (1990). Classification of the reversal test in palaeomagnetism. *Geophysical Journal International*, *103*, 725–729.
Pesonen, L., & Nevanlinna, H. (1981). Late Precambrian Keweenawan asymmetric reversals. *Nature*, *294*, 436–439.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.
Robertson, W., & Fahrig, W. (1971). The great Logan Loop–The polar wander path from Canadian shield rocks during the Neohelikian era. *Canadian Journal of Earth Sciences*, *8*, 1355–1372.
Ryan, G. W., & Leadbetter, S. D. (2002). On the misuse of confidence intervals for two means in testing for the significance of the difference between the means. *Journal of Modern Applied Statistical Methods*, *1*, 473–478.
Sambridge, M., Gallagher, K., Jackson, A., & Rickwood, P. (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International*, *167*, 528–542.
Schlömer, N. (2018). QuadPy: Numerical integration, quadrature for various domains. https://doi.org/10.5281/zenodo.1173132
Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.
Strik, G., Blake, T. S., Zegers, T. E., White, S. H., & Langereis, C. G. (2003). Palaeomagnetism of flood basalts in the Pilbara Craton, Western Australia: Late Archaean continental drift and the oldest known reversal of the geomagnetic field. *Journal of Geophysical Research*, *108*(B12), 2551. https://doi.org/10.1029/2003JB002475
Swanson-Hysell, N. L., Maloof, A. C., Weiss, B. P., & Evans, D. A. D. (2009). No asymmetry in geomagnetic reversals recorded by 1.1-billion-year-old Keweenawan basalts. *Nature Geoscience*, *42*, 5. https://doi.org/10.1130/G35271.1
Swanson-Hysell, N. L., Vaughan, A. V., Mustain, M. R., & Asp, K. E. (2014). Confirmation of progressive plate motion during the Midcontinent Rift's early magmatic stage from the Osler Volcanic Group, Ontario, Canada. *Geochemistry, Geophysics, Geosystems*, *15*, 2039–2047. https://doi.org/10.1002/2013GC005180
Tauxe, L. (2010). *Essentials of paleomagnetism*. Berkeley: University of California Press.
Tauxe, L., & Kent, D. V. (2004). *A Simplified Statistical Model for the Geomagnetic Field and the Setection of Shallow Bias in Paleomagnetic Inclinations: Was the Ancient Magnetic Field Dipolar? Timescales of the Paleomagnetic Field, Geophysical Monographs* (pp. 101–116). Washington, DC: American Geophysical Union.
Tauxe, L., Klystra, N., & Constable, C. (1991). Bootstrap statistics for paleomagnetic data. *Journal of Geophysical Research*, *96*, 11,723–11,740.
Watson, G. S. (1956). Analysis of dispersion on a sphere. *Monthly Notices of the Royal Astronomical Society*, *7*, 153–159.
Watson, G. S. (1983). Large sample theory of the Langevin distributions. *Journal of Statistical Planning and Inference*, *8*, 245–256.
Zijderveld, J. D. A. (1967). Methods in palaeomagnetism. In D. W. Collinson, K. M. Creer, & S. K. Runcorn (Eds.), *A. C. Demagnetization of Rocks: Analysis of Results* (pp. 254–286). Amsterdam: Elsevier.