MONASH University

MONASH BUSINESS SCHOOL

DEPARTMENT OF ECONOMETRCS AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

Familial inference: Tests for hypotheses on a family of centres.

Ryan Thompson, Catherine S. Forbes, Steven N. MacEachern and Mario Peruggia

Working Paper No. 16/23
June 2023

Familial inference: Tests for hypotheses on a family of centres

Ryan Thompson*[1], Catherine S. Forbes[1], Steven N. MacEachern[2], Mario Peruggia[2]

[1]*Department of Econometrics and Business Statistics, Monash University*
[2]*Department of Statistics, The Ohio State University*

June 21, 2023

Abstract

Statistical hypotheses are translations of scientific hypotheses into statements about one or more distributions, often concerning their centre. Tests that assess statistical hypotheses of centre implicitly assume a specific centre, e.g., the mean or median. Yet, scientific hypotheses do not always specify a particular centre. This ambiguity leaves the possibility for a gap between scientific theory and statistical practice that can lead to rejection of a true null. In the face of replicability crises in many scientific disciplines, significant results of this kind are concerning. Rather than testing a single centre, this paper proposes testing a family of plausible centres, such as that induced by the Huber loss function (the Huber family). Each centre in the family generates a testing problem, and the resulting family of hypotheses constitutes a familial hypothesis. A Bayesian nonparametric procedure is devised to test familial hypotheses, enabled by a novel pathwise optimization routine to fit the Huber family. The favourable properties of the new test are demonstrated theoretically and experimentally. Two examples from psychology serve as real-world case studies.

1 Introduction

Hypothesis testing is one of statistics' most important contributions to the scientific method. Testing helps advance diverse lines of inquiry, from evaluating the efficacy of experimental drugs to assessing the validity of psychological theories. Researchers working on these problems often characterize their questions as competing statements about a centre μ of one or more distributions. In the simplest one-sample setting, these statements take the form

$$\mathrm{H}_0 : \mu \in \mathcal{M}_0 \quad \text{vs.} \quad \mathrm{H}_1 : \mu \in \mathcal{M}_1,$$

*Corresponding author. Now at School of Mathematics and Statistics, University of New South Wales and Data61, Commonwealth Scientific and Industrial Research Organisation. Email: ryan.thompson1@unsw.edu.au

Keywords: Bayesian bootstrap, Dirichlet process, Huber loss, hypothesis testing, pathwise optimization

where $\mathcal{M}_0$ and $\mathcal{M}_1$ are a partition of the support $\mathcal{M}$ of μ. There are myriads of classical tests for one- and two-sample hypotheses of centre. When μ is the mean, the most well-known of these is the t test (Student 1908), and its extension to independent samples from populations with differing variances (Welch 1947). When μ is the median, the sign test (Fisher 1925) is available, as is the median test for independent samples (Mood 1950). The signed-rank test, or rank-sum test for independent samples, are also tests of medians under certain assumptions (Wilcoxon 1945; Mann and Whitney 1947).

The possibility to test different centres such as the mean and median raises the question of what qualifies as a centre. We posit that a centre of a random variable X should satisfy at least two criteria: (1) a reflection of X about the centre should preserve the centre, and (2) a shift in X by a constant should move the centre by that same constant. This definition is purposefully broad to accommodate the many notions of centre used throughout statistics. The mean and median trivially satisfy these criteria, as do other popular notions such as the mode, trimmed mean, and Winsorized mean. Quantiles other than the median, and by extension order statistics such as the minimum and maximum, are not centres under these criteria as they are not preserved by reflection in general. Still, the fact that there are many possibilities for centre can complicate hypothesis testing in science.

In certain applied areas, e.g., psychology and medicine, *scientific hypotheses* are often silent about a specific centre and instead tend to be statistically vague, e.g., treatment A is more efficacious than treatment B. This ambiguity makes translation to *statistical hypotheses* inherently subjective and can leave researchers questioning which centre to use. See Blakely and Kawachi (2001), Ben-Aharon et al. (2019), and Rousselet and Wilcox (2020) for discussions of this issue in epidemiology, medicine, and psychology. Moreover, ambiguity about the correct or best centre leaves the possibility for a gap between scientific theory and statistical practice that can lead to rejection of a true null, threatening the validity of findings. Sometimes H_0 can be rejected just by switching from one centre to another, say from the mean to the median. In the face of replicability crises in various disciplines (see, e.g., Ioannidis 2005; Open Science Collaboration 2015; Christensen and Miguel 2018), the possibility for significant results of this sort is concerning. The fact that statistical experts often have no input on the statistical aspects of scientific research only aggravates the issue (see, e.g., Strasak et al. 2007; Hardwicke and Goodman 2020). Transparent statistical tools are needed to instil confidence in scientific claims.

Motivated by the preceding discussion, this paper proposes a new approach to hypothesis testing: familial inference. Unlike existing inferential methods, which test hypotheses about a single centre, methods for familial inference test hypotheses about a *family* of plausible centres, with the ultimate goal of strengthening any claims of significance. More specifically, consider a family of centres $\{\mu(\lambda) : \lambda \in \Lambda\}$ where λ indexes each member (centre). The familial testing problem is to decide which hypothesis concerning this family is correct:

$$\mathrm{H}_0 : \mu(\lambda) \in \mathcal{M}_0 \text{ for some } \lambda \in \Lambda \quad \text{vs.} \quad \mathrm{H}_1 : \mu(\lambda) \in \mathcal{M}_1 \text{ for all } \lambda \in \Lambda.$$

The familial null hypothesis states that at least one member (centre) of the family is contained in the null set $\mathcal{M}_0$. The alternative hypothesis is that no member is in $\mathcal{M}_0$.[1] This paper

[1] This style of testing may be considered to have an intersection-union test format. See the discussion in Appendix C.

studies the family of centres induced by the Huber loss function (Huber 1964). The Huber function is a mixture of square and absolute loss, where λ controls the mixture. By sweeping λ between 0 and infinity, one obtains a family of centres that includes the mean and median as limit points. All members of this Huber family satisfy our criteria for centre. While this more conservative testing approach could potentially overlook a useful treatment, it reduces the risk of promoting an ineffective (or harmful) treatment, wasting limited resources, and misdirecting future research. Given the well-reported scientific reproducibility crises, the benefits may far outweigh the cost.

Familial inference is more sophisticated than inference for a single centre and requires new tools developed in this paper. Our first methodological development is a Bayesian nonparametric procedure for one- and two-sample testing with continuous and discrete random variables. The procedure is based on the limit of a Dirichlet process prior (Ferguson 1973), sometimes referred to as the Bayesian bootstrap (Rubin 1981). Bayesian tests have several advantages over frequentist tests, including that they measure the probability of H_0. Frequentist approaches only deliver p-values that are at best a proxy for this probability. We refer the reader to Kruschke (2013) and Benavoli et al. (2017) for discussions on the merits of Bayesian testing. Besides the advantages of a Bayesian approach, the nonparametric nature of our test ameliorates concern about model misspecification. Though numerous existing works address Bayesian nonparametric testing (Ma and Wong 2011; Benavoli et al. 2014; Huang and Ghosh 2014; Benavoli et al. 2015; Holmes et al. 2015; Filippi and Holmes 2017; Gutiérrez et al. 2019; Pereira, Taylor-Rodríguez, and Gutiérrez 2020), these treat hypotheses about single statistical parameters or entire distributions, distinct from the familial hypotheses treated in this paper.

Our second methodological development is an algorithm for fitting the Huber family, necessary to implement the new test. The algorithm is a pathwise optimization routine that exploits piecewise linearity of the Huber solution path to fit the family, containing infinitely many centres, in a single pass over the data. It has low computational complexity and terminates in at most $n - 1$ steps, where n is the sample size. We elucidate the connection between our algorithm and least angle regression (Efron et al. 2004; Rosset and Zhu 2007), popularly used for fitting the lasso regularization path (Tibshirani 1996). The algorithms devised in this paper are made available in the open-source `R` package `familial`, designed with a standard interface similar to that of existing tests in the `stats` package. Methods for visualizing the posterior family via functional boxplots (Sun and Genton 2011) are provided. `familial` is publicly available on the `R` repository `CRAN`.

To illustrate our proposal, we consider data from a study of mammalian sleep patterns in Savage and West (2007). The data contains sleep times for $n = 83$ species of mammals. A histogram of the ratio of sleeping hours to waking hours is plotted in Figure 1. The data are heavily right-skewed, suggesting the mean and median are probably far separated. Suppose we ask whether mammals tend to spend as much time sleeping as they do awake, i.e., whether $\mu = 1$. A t test that the mean is one yields a p-value of 0.698. A bootstrap test that the mean is one, conducted as a robustness check, produces only a marginally smaller p-value of 0.668. A sign test that the median is one gives a p-value of 0.028. At a conventional 0.05 significance level, these tests do not yield the same answer to our scientific question. This inconsistency raises the question of how exactly to proceed in the absence of a guiding scientific theory.

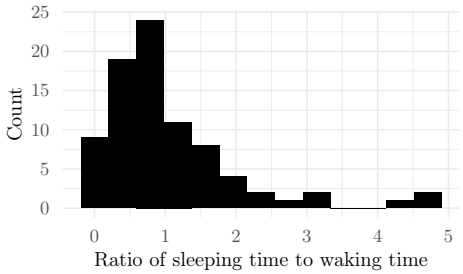

Figure 1: Histogram of the mammalian sleep data.

Using our procedure, we estimate the posterior Huber family via 1,000 Bayesian bootstraps, summarized in Figure 2 by a functional boxplot. As the Huber parameter $\lambda \to \infty$, the 50%

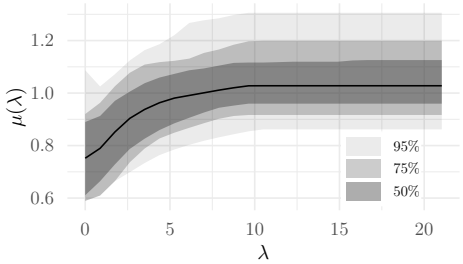

Figure 2: Functional boxplot of the posterior Huber family for the mammalian sleep data. Shading indicates different central regions of the posterior.

central region of the posterior encloses the null value (recall the mean is attained in the limit). By querying the posterior, we find a probability of 0.633 that at least one centre in the family equals one. Under zero-one loss configured analogously to using a 0.05 frequentist significance level (detailed later), the familial test finds insufficient evidence to reject the null in favour of the alternative. Because no specific choice was made about the centre, the problem of choosing between conflicting tests does not arise. Most importantly, we do not arrive at a result that would hold only under a certain centre.

Before delving into the details of our approach, we pause to ask whether a familial test of centres might be less useful than a more general test of distributions. Earlier Bayesian nonparametric tests of distributions, such as those of Holmes et al. (2015) and Pereira, Taylor-Rodríguez, and Gutiérrez (2020), consider a larger class of alternatives than ours. The usefulness of these tests depends, however, upon the scientific question. A change in scale, e.g., can cause these tests to reject even with equal centres. If such changes are not scientifically pertinent, these tests are unsuitable.

2 Bayesian nonparametric test

2.1 Inference problem

Let $X_1, \ldots, X_n$ be an iid sample according to a distribution P_0. Our goal is to carry out inference on the set $\{\mu_0(\lambda) : \lambda \in \Lambda\}$, where

$$\mu_0(\lambda) := \operatorname*{arg\,min}_{\mu \in \mathcal{M}} \mathrm{E}\left[\ell_\lambda\left(\frac{X-\mu}{\sigma}\right)\right] = \operatorname*{arg\,min}_{\mu \in \mathcal{M}} \int \ell_\lambda\left(\frac{x-\mu}{\sigma}\right) dP_0(x).$$

Here, $\ell_\lambda : \mathbb{R} \to \mathbb{R}_+$ is a loss function controlled by the parameter λ. The constant $\sigma > 0$ is necessary in certain loss functions to make λ invariant to the spread of X. Invariance of λ to spread is relevant for testing independent samples, addressed later. The population centre $\mu_0(\lambda)$ minimizes the expectation of the loss configured by λ under P_0. To maintain generality throughout this section, we do not specify a particular loss function. However, to give a concrete example that will be the focus of subsequent sections, one may consider the Huber function

$$\ell_\lambda(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| < \lambda, \\ \lambda|z| - \frac{1}{2}\lambda^2, & \text{if } |z| \geq \lambda. \end{cases}$$

The support of λ is $\Lambda = (0, \infty)$. The mean of P_0 is the limiting solution as $\lambda \to \infty$. The median is the limiting solution in the other direction. The continuum of centres therebetween comprises the Huber family. The approach we propose can accommodate the restriction to any subset of the family given by $\Lambda = [a, b]$ for $0 < a < b < \infty$.

If the true generative model P_0 were known, we would immediately have access to the family $\{\mu_0(\lambda) : \lambda \in \Lambda\}$. Of course, this is not the case in practice, P_0 is unknown. The traditional parametric Bayesian approach to this problem proceeds by means of a prior on parameters for a class of models for P. A valid criticism of this approach is the implicit assumption that P_0 is contained in the model class. Misspecified models can lead to false conclusions, which is troubling in the context of hypothesis testing. To this end, the Bayesian nonparametric approach is an appealing alternative. Rather than placing a prior on the parameters governing a distribution for P, one places a prior directly on the distribution itself. The Dirichlet process, a probability distribution on the space of probability distributions, is a natural candidate for this task. Since Dirichlet processes have support on a large class of distributions, they are a popular prior in Bayesian nonparametrics. The reader is referred to MacEachern (2016) for a recent and accessible overview of their properties.

2.2 Bayesian bootstrap

We denote by $\mathrm{DP}(cP_\pi)$ a Dirichlet process with base distribution P_π and concentration parameter $c > 0$. The concentration parameter is used to impart confidence in P_π. With a Dirichlet process as a prior on P, our Bayesian model is

$$X_1, \ldots, X_n \,|\, P \overset{\text{iid}}{\sim} P, \quad P \sim \mathrm{DP}(cP_\pi).$$

Ferguson (1973) shows the posterior corresponding to this model is also a Dirichlet process:

$$P \,|\, X_1 = x_1, \ldots, X_n = x_n \sim \mathrm{DP}\left(cP_\pi + \sum_{i=1}^{n} \delta_{x_i}\right),$$

where δ_{x_i} is the Dirac measure at x_i. The Dirichlet process is a conjugate prior for iid sampling under P, and the posterior is the base distribution P_π with added point masses at the sample realizations $x_1, \ldots, x_n$. A base distribution P_π must be chosen to operationalize this model. If one wishes to minimize the impact of the choice of P_π, it is sensible to consider the limiting case where the concentration parameter $c \to 0$, which leads to the posterior

$$P \,|\, X_1 = x_1, \ldots, X_n = x_n \sim \mathrm{DP}\left(\sum_{i=1}^{n} \delta_{x_i}\right).$$

Gasparini (1995) shows that this posterior exactly matches the Bayesian bootstrap, proposed by Rubin (1981) as the Bayesian analog of the frequentist bootstrap (Efron 1979). MacEachern (1993) also establishes a unique connection of this posterior to the empirical distribution of the data. The Bayesian bootstrap places support only on the observed data and is equivalent to

$$P(\cdot) = \sum_{i=1}^{n} w_i \delta_{x_i}(\cdot), \quad (w_1, \ldots, w_n) \sim \mathrm{Dirichlet}(1, \ldots, 1),$$

where $\mathrm{Dirichlet}(1, \ldots, 1)$ is the n-dimensional Dirichlet distribution with all concentration parameters equal to one. Sometimes this distribution is referred to as flat or uniform. The first- and second-order asymptotic properties of the Bayesian bootstrap are described in Lo (1987) and Weng (1989). As well as being theoretically well-understood, the Bayesian bootstrap admits scalable sampling algorithms that are trivially parallelizable, making posterior exploration highly tractable. See Fong, Lyddon, and Holmes (2019), Lyddon, Holmes, and Walker (2019), and Barrientos and Peña (2020) for recent applications of the Bayesian bootstrap to complex models and data. As with those applications, tractability is key here.

We now have a posterior for P, and consequently also a posterior on any summaries of P (see, e.g., Lee and MacEachern 2014), including those of interest: families of centres. To estimate the posterior for a given family we propose Algorithm 1. The complexity of solving

Algorithm 1: Bayesian bootstrap for familial inference

Input $(x_1, \ldots, x_n)$

For $b = 1, \ldots, B$:

1. Sample $(w_1^{(b)}, \ldots, w_n^{(b)})$ from $\mathrm{Dirichlet}(1, \ldots, 1)$
2. Compute $\mu^{(b)}(\lambda) = \arg\min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} w_i^{(b)} \ell_\lambda([x_i - \mu]/\sigma^{(b)})$ for all $\lambda \in \Lambda$

Output $\{\mu^{(b)}(\lambda) : \lambda \in \Lambda\}_{b=1}^{B}$

the minimization problem in step two for all $\lambda \in \Lambda$ depends on the loss function. In the next section, we present a numerical routine that addresses the case where the loss function is the Huber function. Since λ in the Huber function is sensitive to changes in spread, we configure $\sigma^{(b)}$ to be the median absolute deviation of the bootstrap sample, i.e., the weighted median absolute deviation with weights $w_1^{(b)}, \ldots, w_n^{(b)}$. The standard deviation of the bootstrap sample could also be used. In our experience, switching between standard deviation and median absolute deviation usually does not lead to materially different outcomes, and the choice is only relevant for independent samples testing (see Section 2.4).

From the output of Algorithm 1, the posterior probabilities $p_{\mathrm{H}_0} := \Pr(\mathrm{H}_0 \,|\, x_1, \ldots, x_n)$ and $p_{\mathrm{H}_1} := \Pr(\mathrm{H}_1 \,|\, x_1, \ldots, x_n)$ are estimable as

$$\hat{p}_{\mathrm{H}_0} := \frac{1}{B} \sum_{b=1}^{B} 1(\exists\, \lambda \in \Lambda : \mu^{(b)}(\lambda) \in \mathcal{M}_0)$$

and

$$\hat{p}_{\mathrm{H}_1} := \frac{1}{B} \sum_{b=1}^{B} 1(\forall\, \lambda \in \Lambda : \mu^{(b)}(\lambda) \in \mathcal{M}_1).$$

Since H_0 and H_1 are mutually exclusive and collectively exhaustive, $p_{\mathrm{H}_0} + p_{\mathrm{H}_1} = 1$ and, for any B, $\hat{p}_{\mathrm{H}_0} + \hat{p}_{\mathrm{H}_1} = 1$.

Though our focus is the Bayesian bootstrap (i.e., the Dirichlet process prior with concentration parameter $c \to 0$), one can extend Algorithm 1 to handle the prior with $c > 0$ via the stick-breaking construction of Sethuraman (1994). The interested reader may refer to Chapter 2 of Müller et al. (2015) for details of this approach.

2.3 Decision rule

To map the estimated posterior probabilities $\hat{p}_{\mathrm{H}_0}$ and $\hat{p}_{\mathrm{H}_1}$ to a decision, we assign a loss to each possible decision. Specifically, given the posterior probability vector $\hat{p} = (\hat{p}_{\mathrm{H}_0}, \hat{p}_{\mathrm{H}_1})^\top$, we make the decision giving lowest posterior expected loss $L\hat{p}$, where L is loss matrix with rows corresponding to the decision to accept H_0, accept H_1, or accept neither (an *indeterminate* decision). We use

$$L := \begin{pmatrix} l_{\mathrm{H}_0|\mathrm{H}_0} & l_{\mathrm{H}_0|\mathrm{H}_1} \\ l_{\mathrm{H}_1|\mathrm{H}_0} & l_{\mathrm{H}_1|\mathrm{H}_1} \\ l_{\mathrm{I}|\mathrm{H}_0} & l_{\mathrm{I}|\mathrm{H}_1} \end{pmatrix} = \begin{matrix} \\ \mathrm{H}_0 \\ \mathrm{H}_1 \\ \mathrm{I} \end{matrix} \overset{\begin{matrix} \mathrm{H}_0 & \mathrm{H}_1 \end{matrix}}{\begin{pmatrix} 0 & 20 \\ 20 & 0 \\ 1 & 1 \end{pmatrix}}, \tag{2.1}$$

where $l_{\mathrm{H}_j|\mathrm{H}_k}$ denotes the loss incurred in accepting H_j when H_k is true for $j, k = 0, 1$, and where $l_{\mathrm{I}|\mathrm{H}_k}$ denotes the loss from an indeterminate decision for $k = 0, 1$. Under the above configuration of L, either H_0 or H_1 is accepted depending on whether $\hat{p}_{\mathrm{H}_0}$ or $\hat{p}_{\mathrm{H}_1}$ is greater than 0.95, analogous to a 0.05 level frequentist test. When both probabilities are less than 0.95 the decision is indeterminate.

2.4 Two-sample problem

The discussion up to now has focused on the one-sample setting. Consider now the two-sample setting with samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. If X_i and Y_i are meaningfully coupled together, e.g., measurements on the same subject before and after treatment, the two samples are paired and $n_1 = n_2$. Define the random variable Z_i as the difference $X_i - Y_i$. Then the familial hypotheses are

$$\mathrm{H}_0 : \mu_Z(\lambda) \in \mathcal{M}_0 \text{ for some } \lambda \in \Lambda \quad \text{vs.} \quad \mathrm{H}_1 : \mu_Z(\lambda) \in \mathcal{M}_1 \text{ for all } \lambda \in \Lambda,$$

where $\mu_Z(\lambda)$ is a centre of Z. Algorithm 1 applies directly to the sample $Z_1, \ldots, Z_n$ with $n = n_1 = n_2$.

When X_i and Y_i are not coupled together the samples are independent. The familial hypotheses are then

$$\mathrm{H}_0 : \underset{\text{for some } \lambda \in \Lambda}{\mu_X(\lambda) - \mu_Y(\lambda) \in \mathcal{M}_0} \quad \text{vs.} \quad \mathrm{H}_1 : \underset{\text{for all } \lambda \in \Lambda.}{\mu_X(\lambda) - \mu_Y(\lambda) \in \mathcal{M}_1}$$

Here, the same centre of X is compared with the same centre of Y, i.e., the mean is compared with the mean, the median with the median, and so on. Testing these hypotheses requires bootstrapping the families of X and Y with independently drawn weights. When independent weights are drawn at a bootstrap iteration, each centre of Y is subtracted from the same centre of X. The posterior probability of H_0 is estimated by the proportion of times across bootstrap iterations that the set of differences intersects the null set $\mathcal{M}_0$.

3 Huber family

3.1 Optimization problem

To implement the testing procedure of the preceding section, we require a method for fitting the family of centres to each distribution drawn from the posterior, i.e., for solving the optimization problems in step two of Algorithm 1 given fixed bootstrap weights $w_1^{(b)}, \ldots, w_n^{(b)}$. For simplicity of exposition, we drop the bootstrap iteration superscript (b) and fix $\sigma = 1$ without loss of generality. The Huber function as a function of the residual $x - \mu$ can then be expressed as

$$\ell_\lambda(x - \mu) = \begin{cases} \frac{1}{2}(x-\mu)^2, & \text{if } |x-\mu| < \lambda, \\ \lambda|x-\mu| - \frac{1}{2}\lambda^2, & \text{if } |x-\mu| \geq \lambda. \end{cases}$$

We denote the loss over the weighted (bootstrap) sample by

$$\mathcal{L}_\lambda(\mu) := \sum_{i=1}^n w_i \ell_\lambda(x_i - \mu).$$

Our goal is to devise an algorithm for computing the set $\{\mu(\lambda) : \lambda \in \Lambda\}$, where

$$\mu(\lambda) := \operatorname*{arg\,min}_{\mu \in \mathbb{R}} \mathcal{L}_\lambda(\mu) \tag{3.1}$$

and $\Lambda = (0, \infty)$. For an equally weighted sample, (3.1) includes as limiting cases the sample mean and sample median. When the weights are unequal, the limit points become the *weighted mean* and *weighted median*, interpretable as the mean and median of the bootstrap sample. The weighted mean is defined by

$$\bar{\mu} := \operatorname*{arg\,min}_{\mu \in \mathbb{R}} \sum_{i=1}^n w_i (x_i - \mu)^2 = \sum_{i=1}^n w_i x_i,$$

and the weighted median by

$$\tilde{\mu} := \operatorname*{arg\,min}_{\mu \in \mathbb{R}} \sum_{i=1}^n w_i |x_i - \mu|.$$

There is no analytical solution for the weighted median. In fact, the weighted mean is the only Huber centre that admits an analytical solution in general. For general λ, the existence of a solution to the minimization (2) requires only that the realized sample $x_1, \ldots, x_n$ and weights $w_1, \ldots, w_n$ be finite. Hence, successful computation of a solution does not hinge upon the existence of a solution in the population, e.g., if $x_1, \ldots, x_n$ are realizations of a Cauchy.

If Λ were a finite set, it would be possible to solve the optimization problem (3.1) for each of its elements. For given λ, the one-dimensional problem (3.1) is convex, and although it does not admit an analytical solution, it is amenable to simple numerical routines (Huber and Ronchetti 2009). Even if Λ is not finite, one might try approximating it using a fine grid and then proceed to solve each minimization individually. Recall though each set of minimization problems needs to be solved B times in the Bayesian bootstrap, where B might be 1,000, 10,000, or larger. Thus, even with an efficient algorithm, total cumulative runtime can be prohibitive. Notwithstanding runtime considerations, such an approach still only yields an approximation. Instead of an approximation, we propose a fast and exact pathwise algorithm that optimizes (3.1) for all values of λ.

3.2 Pathwise optimization routine

Our approach exploits piecewise linearity of the solution path $\mu(\lambda)$ for $\lambda \in (0, \infty)$, a property we now demonstrate. The gradient of the Huber function with respect to μ is

$$\frac{\partial \ell_\lambda(x-\mu)}{\partial \mu} = \begin{cases} -(x-\mu), & \text{if } |x-\mu| < \lambda, \\ -\lambda \operatorname{sign}(x_i - \mu), & \text{if } |x-\mu| \geq \lambda. \end{cases}$$

Hence, the gradient of the loss over the weighted sample is

$$\frac{\partial \mathcal{L}_\lambda(\mu)}{\partial \mu} = \sum_{i=1}^{n} w_i \frac{\partial \ell_\lambda(x_i - \mu)}{\partial \mu} = - \sum_{i:|x_i-\mu|<\lambda} w_i (x_i - \mu) - \sum_{i:|x_i-\mu|\geq\lambda} w_i \lambda \operatorname{sign}(x_i - \mu).$$

We denote the above gradient by $\mathcal{L}'(\mu)$, suppressing the explicit dependency on λ. The first-order condition for optimality of $\mu(\lambda)$ requires that $\mathcal{L}'(\mu(\lambda)) = 0$. The implicit function theorem then gives

$$\frac{\partial \mu(\lambda)}{\partial \lambda} = -\frac{\partial \mathcal{L}'(\mu(\lambda))}{\partial \lambda} \Big/ \left. \frac{\partial \mathcal{L}'(\mu)}{\partial \mu} \right|_{\mu=\mu(\lambda)},$$

which, after evaluating gradients on the right-hand side, yields

$$\frac{\partial \mu(\lambda)}{\partial \lambda} = \frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i}. \tag{3.2}$$

Observe that the gradient of the solution path $\partial\mu(\lambda)/\partial\lambda$ is piecewise constant as a function of λ, implying that $\mu(\lambda)$ is piecewise linear. It follows that $\mu(\lambda)$ is also piecewise continuous with left and right limits. It can be verified that the left and right limits at any knot $\lambda^\star$ equal $\mu(\lambda^\star)$, and hence that $\mu(\lambda)$ is continuous.

Since the solution path is piecewise linear, it is composed of a sequence of knots, i.e., certain values of λ at which $|x_i - \mu(\lambda)| = \lambda$ for one or more sample points. These knots

correspond to crossing events, where sample points transition between the square and absolute pieces of the Huber function. Lemma 1 characterizes a useful property in relation to these crossing events.

Lemma 1. *Suppose sample point x_0 satisfies $|x_0 - \mu(\lambda^\star)| \geq \lambda^\star$ for some $\lambda^\star > 0$. Then, for all $0 < \lambda < \lambda^\star$, it holds $|x_0 - \mu(\lambda)| \geq \lambda$.*

Lemma 1 implies that, for a decreasing sequence of λ, once a sample point has crossed to the absolute piece of the Huber function, it remains there. This property guarantees the existence of at most n knots along the solution path. The lemma is proven in Appendix A.

To trace out the solution path, we need only fit μ at each λ in the sequence of knots since any solution between knots is linearly interpolable. A method to efficiently determine the location and solution at each knot is required. Suppose we are at an arbitrary point (λ, μ) along the solution path. Then, thanks to piecewise linearity, the closest knot point (λ^+, μ^+) to the left of (λ, μ) is computable by taking a step $\gamma > 0$, of a certain size, as follows:

$$\lambda^+ = \lambda - \gamma \tag{3.3}$$

and

$$\mu^+ = \mu - \gamma \frac{\partial \mu(\lambda)}{\partial \lambda}. \tag{3.4}$$

Equation (3.2) provides an analytical expression for the gradient $\partial \mu(\lambda)/\partial \lambda$. An expression for the required step size γ is still needed. To this end, we present Proposition 1.

Proposition 1. *Let (λ, μ) be any point along the solution path such that $|x_i - \mu| < \lambda$ for at least one $i = 1, \ldots, n$. Then the largest positive step size before the solution path reaches a knot point (λ^+, μ^+) to the left of (λ, μ) is*

$$\gamma = \min_{i:|x_i - \mu| < \lambda} \left(\frac{\lambda - s_i(x_i - \mu)}{1 + s_i \partial \mu(\lambda)/\partial \lambda} \right),$$

where $s_i = \operatorname{sign}(x_i - \tilde{\mu})$ and $\tilde{\mu}$ is the weighted median.

The requirement $|x_i - \mu| < \lambda$ for at least one $i = 1, \ldots, n$ guarantees the existence of at least one more unexplored knot along the solution path. Beyond the first and last knots, the solution path is flat. The proposition is proven in Appendix A.

Putting together the above ingredients and letting $\lambda = \lambda^{(m)}$, $\lambda^+ = \lambda^{(m+1)}$, $\mu = \mu^{(m)}$, and $\mu^+ = \mu^{(m+1)}$ we arrive at Algorithm 2. Starting at the rightmost knot point $(\lambda^{(1)}, \mu^{(1)})$, which corresponds to the weighted mean, the algorithm forges a path step-by-step to the leftmost knot point, which corresponds to the weighted median. Figure 3 illustrates this process on $n = 30$ iid draws from a standard normal distribution. The algorithm begins at a value of λ large enough to induce the weighted mean as the centre and then iteratively decreases λ. The final λ in this sequence of iterates is sufficiently small to induce the weighted median as the centre. Observe that the path is piecewise linear and continuous.

Thus far the spread has been fixed at $\sigma = 1$. To recover the solution path for $\sigma \neq 1$, we scale the output $(\lambda^{(1)}, \ldots, \lambda^{(m+1)})$ from Algorithm 2 by multiplying it by σ. The centres $(\mu^{(1)}, \ldots, \mu^{(m+1)})$ do not change. This scaling has the intended effect of using the scaled residual $(x - \mu)/\sigma$ in the Huber function instead of $x - \mu$. We remind the reader that this scaling makes the solution path scale-free, relevant for testing independent samples.

Algorithm 2: Pathwise optimization for the Huber family

Input $(x_1, \ldots, x_n)$ and $(w_1, \ldots, w_n)$
Initialize $\mu^{(1)} = \sum_{i=1}^n w_i x_i$ and $\lambda^{(1)} = \max_i(|x_i - \mu^{(1)}|)$
1. Calculate the sign $s_i = \text{sign}(x_i - \tilde{\mu})$ for $i = 1, \ldots, n$
For $m = 1, \ldots, n-1$:
2. If $\{i : |x_i - \mu^{(m)}| < \lambda^{(m)}\} = \emptyset$ then $m = m - 1$ and break
3. Calculate the gradient

$$\eta = \frac{\sum_{i:|x_i - \mu^{(m)}| \geq \lambda^{(m)}} w_i \, \text{sign}(x_i - \mu^{(m)})}{\sum_{i:|x_i - \mu^{(m)}| < \lambda^{(m)}} w_i}$$

4. Calculate the step size

$$\gamma = \min_{i:|x_i - \mu^{(m)}| < \lambda^{(m)}} \left(\frac{\lambda^{(m)} - s_i(x_i - \mu^{(m)})}{1 + s_i \eta} \right)$$

5. Perform the updates $\lambda^{(m+1)} = \lambda^{(m)} - \gamma$ and $\mu^{(m+1)} = \mu^{(m)} - \gamma\eta$
Output $(\lambda^{(1)}, \ldots, \lambda^{(m+1)})$ and $(\mu^{(1)}, \ldots, \mu^{(m+1)})$

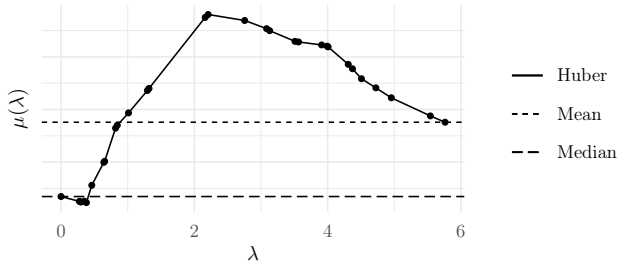

Figure 3: Algorithm 2 applied with $x_1, \ldots, x_n$ drawn from a standard normal distribution and $w_1, \ldots, w_n$ drawn from a flat Dirichlet distribution with $n = 30$. The solid points are iterates (knots) from the algorithm. Centres between iterates are linearly interpolated.

3.3 Relation to least angle regression

Algorithm 2 bears similarity to least angle regression (Efron et al. 2004; Rosset and Zhu 2007), a pathwise optimization routine that traces the solution path of lasso regression coefficients. To clarify this similarity, first recall the Moreau envelope $f_\lambda(z)$ of a real-valued function $f(z)$, which is the infimal convolution of $f(z)$ and $g(z) = 1/(2\lambda)z^2$ (see, e.g., Polson, Scott, and Willard 2015). When $f(z) = |z|$, there is a precise relation between the Huber function $\ell_\lambda(z)$ and the Moreau envelope:

$$f_\lambda(z) := \inf_{\beta \in \mathbb{R}} \left(|\beta| + \frac{1}{2\lambda}(z - \beta)^2 \right) = \begin{cases} \frac{1}{2\lambda} z^2, & \text{if } |z| < \lambda, \\ |z| - \frac{1}{2}\lambda, & \text{if } |z| \geq \lambda. \end{cases}$$

The right-hand side is equal to $\ell_\lambda(z)/\lambda$. In words, multiplying the Moreau envelope of the absolute value function by λ yields the Huber function, a known result from convex analysis

(Beck 2017). Hence, we have the chain of equalities

$$\begin{aligned}\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} w_i \ell_\lambda(x_i - \mu) &= \min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} w_i \lambda f_\lambda(x_i - \mu) \\ &= \min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} w_i \inf_{\beta_i\in\mathbb{R}} \left(\frac{1}{2}(x_i - \mu - \beta_i)^2 + \lambda|\beta_i| \right) \\ &= \min_{\mu,\beta_1,\ldots,\beta_n\in\mathbb{R}} \sum_{i=1}^{n} w_i \left(\frac{1}{2}(x_i - \mu - \beta_i)^2 + \lambda|\beta_i| \right).\end{aligned}$$

The infimum can be written as a minimum since the absolute value function is closed convex. The final line is a weighted lasso regression of $x_1, \ldots, x_n$ on an identity design matrix of dimensions $n \times n$, showing that the Huber problem (3.1) can be recast as a weighted lasso problem. Thus, applying least angle regression, configured with weights, to an identity design matrix yields a path identical to that produced by Algorithm 2. Despite this equivalence, the development of Algorithm 2 remains essential. Least angle regression is designed for general design matrices and, as such, does not exploit the structure of regression with an identity design, i.e., the Huber problem. Algorithm 2, on the other hand, takes full advantage of this structure. In numerical experimentation, we observed that Algorithm 2 is typically an order of magnitude faster than least angle regression. Without this speedup, the Bayesian bootstrap would remain computationally burdensome.

4 Consistency

The asymptotic properties of the familial test are now established. We focus on a real-valued null m_0 in the one-sample (and paired samples) setting, which serves as a fundamental scenario, though our result is adaptable to a set-valued null $\mathcal{M}_0$ and independent samples. We consider two cases: (1) the null hypothesis $\mathrm{H}_0 : \mu_0(\lambda) = m_0$ for some $\lambda \in \Lambda$ is true and (2) the alternative hypothesis $\mathrm{H}_1 : \mu_0(\lambda) \neq m_0$ for all $\lambda \in \Lambda$ is true.

Theorem 1 states the familial test is asymptotically consistent under both H_0 and H_1.

Theorem 1. *Let $X_1, \ldots, X_n$ be iid random variables with distribution P_0 and, independently, let $(w_1, \ldots, w_n) \sim \mathrm{Dirichlet}(1, \ldots, 1)$. Let $\Lambda = [0, \bar{\lambda}]$ for some $0 < \bar{\lambda} < \infty$. Define the functionals*

$$\hat{\mu}(\lambda) := \arg\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} w_i \ell_\lambda(X_i - \mu)$$

and

$$\mu_0(\lambda) := \arg\min_{\mu\in\mathbb{R}} \int \ell_\lambda(x - \mu)\, dP_0(x).$$

Then the following results hold.

1. *If $\mathrm{H}_0 : \mu_0(\lambda) = m_0$ for some $\lambda \in \Lambda$ is true with $\mu_0(\lambda)$ strictly crossing through m_0, then*

$$\lim_{n\to\infty} \Pr(\exists\, \lambda \in \Lambda : \hat{\mu}(\lambda) = m_0) = 1.$$

2. *If* $\mathrm{H}_1 : \mu_0(\lambda) \neq m_0$ *for all* $\lambda \in \Lambda$ *is true with* $\mu_0(\lambda)$ *bounded away from* m_0*, then*

$$\lim_{n\to\infty} \Pr(\forall\, \lambda \in \Lambda : \hat{\mu}(\lambda) \neq m_0) = 1.$$

A proof is available in Appendix B. The theorem requires only a single Bayesian bootstrap distribution, i.e., $B = 1$, for consistency of the test. Having $B > 1$ presents no technical complication, but it is unnecessary for consistency when $n \to \infty$.

At a high level, the proof of Theorem 1 involves showing the estimator $\hat{\mu}(\lambda)$ (actually, its gradient) under a Bayesian bootstrap distribution can be made arbitrarily close to that under the true distribution P_0, uniformly over all $\lambda \in \Lambda$. This result allows us to argue if H_0 is true and $\mu_0(\lambda)$ strictly crosses through m_0, then so does $\hat{\mu}(\lambda)$ in the limit. Alternatively, if H_0 is false and $\mu_0(\lambda)$ is bounded above or below away from m_0, then so is $\hat{\mu}(\lambda)$ in the limit.

The case where the null is true and $\mu_0(\lambda)$ does not strictly cross through m_0 is not covered by Theorem 1. This case would occur, say, when P_0 is a normal distribution with mean m_0. However, symmetric distributions like the normal are uncommon in practice and not the focus of our test.

5 Experiments

5.1 Familial package

This section reports experiments on real and synthetic data. To enable these exercises, the test and algorithms described in the preceding sections are implemented in the `R` package `familial`. For a sample of size $n = 200$, `familial` takes about half a second to perform 1,000 bootstraps for a single sample on one core of a modern processor. Parallelism is also supported. Run time scales linearly with the sample size, number of bootstraps, and, if parallelised, number of processor cores.

5.2 Body posture study

Rosenbaum, Mama, and Algom (2017) conducted an experiment to ascertain the effect of body posture on selective attention (refer to Experiment 3 in that paper). The experiment employed the Stroop test, where subjects are asked to announce colours of a sequence of words and not the words themselves, e.g., announce blue when the word red is printed in blue. The difference in response times between congruent word-colour pairs and incongruent pairs is the Stroop effect. Experimental subjects took the test once while sitting and once while standing. The study found standing lowered the Stroop effect compared with sitting, indicating improved selective attention while standing.

The dataset (Mama 2018) contains paired observations on response times of $n = 50$ subjects. Figure 4 presents a histogram of differences in response time alongside a functional boxplot of the posterior Huber family. The response times do not deviate markedly from a normal distribution, though they are slightly left-skewed. The posterior concentrates well below zero, suggesting standing might reduce the Stroop effect.

The study reported a p-value of 0.004 from an F test of the interaction between congruency and posture in a repeated-measures ANOVA, equivalent to a Student t test that the mean difference in response times is zero. The Fisher sign test and Wilcoxon signed-rank test

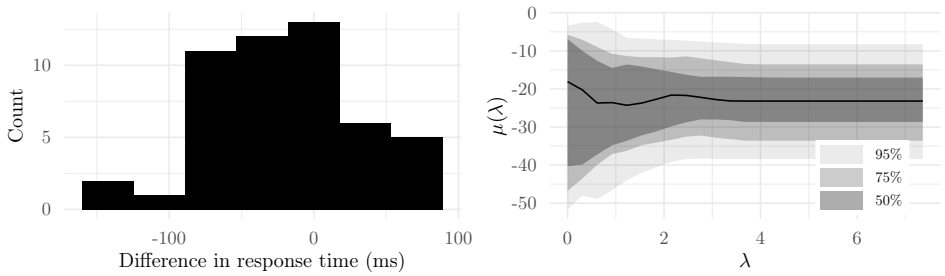

Figure 4: Body posture data. The left plot is a histogram of the data. The right plot is a functional boxplot of the posterior Huber family. Shading indicates different central regions of the posterior.

produce p-values of 0.007 and 0.006, respectively. The Huber familial test finds that the probability of the null is 0.005. All tests reject the null that body posture does not affect the Stroop effect. This result confirms that the original finding is not sensitive to the centre tested. As additional benchmarks, we ran the maximum mean discrepancy (MMD) test of Gretton et al. (2012) and the Bayesian Pólya tree test of Holmes et al. (2015), both tests for equality of distributions. The p-value and null probability is 0.499 and 0.223, respectively. These tests assume the samples are independent, so they are likely underpowered here.

5.3 Multi-task perception study

Srna, Schrift, and Zauberman (2018a) ran an experiment to investigate if human performance at certain activities is affected by whether the activity is perceived as multi-tasking (refer to Study 1a in that paper). Experimental subjects were required to watch a video and transcribe the audio. This activity was framed as multi-tasking to a treatment group and single-tasking to a control group. Assignment to either group was random. The study found that subjects in the treatment group transcribed more words than those in the control group and the accuracy of their transcriptions was higher, suggesting perceiving an activity as multi-tasking improves performance at that activity.

We focus on the number of words transcribed. The dataset (Srna, Schrift, and Zauberman 2018b) contains $n_1 = 82$ subjects in the treatment group and $n_2 = 80$ in the control group; see Figure 5. The groups are dissimilar in distribution, with the multi-task group being unimodal and the single-task group being multimodal. For small values of λ, the 50% central region of the posterior includes zero, indicating the null might be plausible.

The study reported a p-value of 0.033 from an F test of the multi-task condition in a one-way ANOVA, identical to a two-sample Student t test with equal variance that the mean number of words transcribed is equal between groups. The Mood median test yields a p-value of 0.271. The p-value from a Wilcoxon rank-sum test is 0.072. The MMD and Pólya tree tests produce a p-value and null probability of 0.590 and 0.978, respectively. The Huber familial test returns 0.170 as the probability of the null. In contrast to the t test, the familial, sign, rank-sum, MMD, and Pólya tree tests do not find the multi-task condition to affect performance. In particular, the familial test fails to find sufficient support for either hypothesis and returns an indeterminate result. Whether this is a meaningful discrepancy

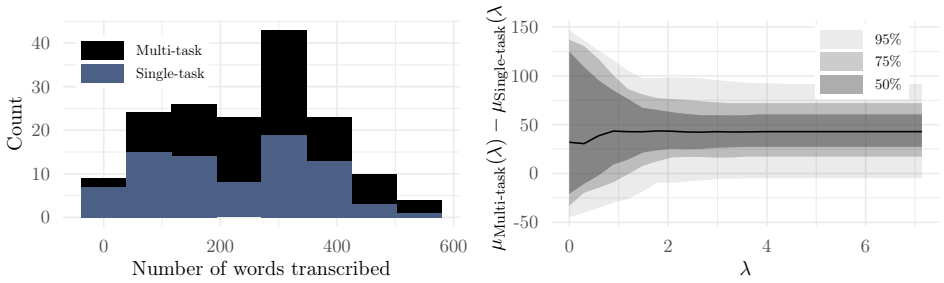

Figure 5: Multi-task perception data. The left plot is a histogram of the data by control and treatment. The right plot is a functional boxplot of the posterior difference in Huber families. Shading indicates different central regions of the posterior.

remains up to subject-matter experts to decide.

5.4 Simulations

Appendix D contains extensive results on synthetic datasets that verify the finite-sample properties of the familial test in one- and two-sample settings. The results cover a variety of distributions, both continuous and discrete, and validate that the familial test is well-behaved. The test has good frequentist size and, for a sufficiently large departure from the null or a reasonably large sample size, high power. In fact, its power remains highly competitive with existing frequentist and Bayesian tests in regions where the familial null fails to hold.

6 Concluding remarks

It has become standard practice to translate scientific hypotheses into statistical hypotheses about a specific centre for the underlying distribution(s). Despite the ubiquity of this approach, there can be a lack of consensus about which centre bests reflects the original scientific hypotheses. When there is ambiguity, we argue one should adopt familial inference, which formulates hypotheses via a family of plausible centres. Our package `familial` implements the tools developed in this paper and is publicly available on `CRAN`.

A natural next step in this line of work is to develop familial inference for other statistical parameters and models. For instance, we found in ongoing work that our tools extend gracefully to linear models. The Huber family of linear models constitutes a continuum of models for conditional centres, from the conditional mean to the conditional median. The pathwise algorithm extends to this setting because the solution path remains piecewise linear under Huber loss.

Another intriguing direction is familial inference for multivariate random variables. The univariate Huber function $\ell_\lambda(z)$ studied in this paper has a lesser-known multivariate counterpart (Hampel et al., 1984), which composes the ℓ_2-norm with the univariate Huber function, i.e., $\ell_\lambda(\|\mathbf{z}\|)$. It remains to be determined whether the multivariate problem remains computationally tractable.

It is also interesting to consider other loss functions for the univariate, multivariate, and linear model problems. For instance, the univariate trimmed square loss (i.e., the trimmed

mean), is also piecewise linear, and readily admits a pathwise algorithm. More generally, fundamentally different algorithms are required.

A frequentist version of the familial test is likewise an appealing avenue of future research. A major challenge is deriving the finite-sample, or asymptotic, distribution of the Huber family under the null hypothesis. A bootstrap test is also possible, but it remains to determine the appropriate way of imposing the null hypothesis on the observed sample to attain the correct size.

Acknowledgements

Thompson acknowledges financial support by an Australian Government Research Training Program (RTP) Scholarship. Forbes, MacEachern, and Peruggia acknowledge financial support by the National Science Foundation (NSF) Grant SES-1921523. MacEachern also acknowledges financial support by the NSF Grant DMS-2015552.

References

Barrientos, A. F. and Peña, V. (2020). 'Bayesian bootstraps for massive data'. *Bayesian Analysis* 15.2, pp. 363–388.

Beck, A. (2017). *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics and Mathematical Optimization Society.

Ben-Aharon, O., Magnezi, R., Leshno, M., and Goldstein, D. A. (2019). 'Median survival or mean survival: Which measure is the most appropriate for patients, physicians, and policymakers?' *Oncologist* 24.11, pp. 1469–1478.

Benavoli, A., Mangili, F., Corani, G., Zaffalon, M., and Ruggeri, F. (2014). 'A Bayesian Wilcoxon signed-rank test based on the Dirichlet process'. *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32, pp. 1026–1034.

Benavoli, A., Corani, G., Demšar, J., and Zaffalon, M. (2017). 'Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis'. *Journal of Machine Learning Research* 18, pp. 1–36.

Benavoli, A., Mangili, F., Ruggeri, F., and Zaffalon, M. (2015). 'Imprecise Dirichlet process with application to the hypothesis test on the probability that $X \leq Y$'. *Journal of Statistical Theory and Practice* 9.3, pp. 658–684.

Berger, R. L. (1982). 'Multiparameter hypothesis testing and acceptance sampling'. *Technometrics* 24.4, pp. 295–300.

Berger, R. L. and Hsu, J. C. (1996). 'Bioequivalence trials, intersection–union tests and equivalence confidence sets'. *Statistical Science* 11.4, pp. 283–319.

Blakely, T. A. and Kawachi, I. (2001). 'What is the difference between controlling for mean versus median income in analyses of income inequality?' *Journal of Epidemiology and Community Health* 55.5, pp. 352–353.

Christensen, G. and Miguel, E. (2018). 'Transparency, reproducibility, and the credibility of economics research'. *Journal of Economic Literature* 56.3, pp. 920–980.

Efron, B. (1979). 'Bootstrap methods: Another look at the jackknife'. *Annals of Statistics* 7.1, pp. 1–26.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). 'Least angle regression'. *Annals of Statistics* 32.2, pp. 407–499.

Ferguson, T. S. (1973). 'A Bayesian analysis of some nonparametric problems'. *Annals of Statistics* 1.2, pp. 209–230.

Filippi, S. and Holmes, C. C. (2017). 'A Bayesian nonparametric approach to testing for dependence between random variables'. *Bayesian Analysis* 12.4, pp. 919–938.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. London, UK: Oliver and Boyd.

Fong, D. Y. T., Kwan, C. W., Lam, K. F., and Lam, K. S. L. (2003). 'Use of the sign test for the median in the presence of ties'. *American Statistician* 57.4, pp. 237–240.

Fong, E., Lyddon, S., and Holmes, C. (2019). 'Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap'. *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97, pp. 1952–1962.

Gasparini, M. (1995). 'Exact multivariate Bayesian bootstrap distributions of moments'. *Annals of Statistics* 23.3, pp. 762–768.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). 'A kernel two-sample test'. *Journal of Machine Learning Research* 13.25, pp. 723–773.

Gutiérrez, L., Barrientos, A. F., González, J., and Taylor-Rodríguez, D. (2019). 'A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control'. *Bayesian Analysis* 14.2, pp. 649–675.

Hardwicke, T. E. and Goodman, S. N. (2020). 'How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey'. *PLoS ONE* 15.10, e0239598.

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). 'Two-sample Bayesian nonparametric hypothesis testing'. *Bayesian Analysis* 10.2, pp. 297–320.

Huang, L. and Ghosh, M. (2014). 'Two-sample hypothesis testing under Lehmann alternatives and Polya tree priors'. *Statistica Sinica* 24.4, pp. 1717–1733.

Huber, P. J. (1964). 'Robust estimation of a location parameter'. *Annals of Mathematical Statistics* 35.1, pp. 73–101.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons.

Ioannidis, J. P. A. (2005). 'Why most published research findings are false'. *PLoS Medicine* 2.8, pp. 696–701.

Kruschke, J. K. (2013). 'Bayesian estimation supersedes the t test'. *Journal of Experimental Psychology: General* 142.2, pp. 573–603.

Lee, J. and MacEachern, S. N. (2014). 'Inference functions in high dimensional Bayesian inference'. *Statistics and Its Interface* 7.4, pp. 477–486.

Li, D., Cao, J., and Zhang, S. (2020). 'Power analysis for cluster randomized trials with multiple binary co-primary endpoints'. *Biometrics* 76.4, pp. 1064–1074.

Lo, A. Y. (1987). 'A large sample study of the Bayesian bootstrap'. *Annals of Statistics* 15.1, pp. 360–375.

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). 'General Bayesian updating and the loss-likelihood bootstrap'. *Biometrika* 106.2, pp. 465–478.

Ma, L. and Wong, W. H. (2011). 'Coupling optional Pólya trees and the two sample problem'. *Journal of the American Statistical Association* 106.496, pp. 1553–1565.
MacEachern, S. (1993). 'An evaluation of Bayes posterior probability regions for a survival curve'. *Journal of Nonparametric Statistics* 3.2, pp. 175–186.
MacEachern, S. N. (2016). 'Nonparametric Bayesian methods: A gentle introduction and overview'. *Communications for Statistical Applications and Methods* 23.6, pp. 445–466.
Mama, Y. (2018). *Project data*. OSF Registries. https://doi.org/10.17605/OSF.IO/T2ED9.
Mann, H. B. and Whitney, D. R. (1947). 'On a test of whether one of two random variables is stochastically larger than the other'. *Annals of Mathematical Statistics* 18.1, pp. 50–60.
Mood, A. M. (1950). *Introduction to the Theory of Statistics*. McGraw-Hill Series in Probability and Statistics. New York, NY, USA: McGraw-Hill.
Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics. Cham, Switzerland: Springer.
Open Science Collaboration (2015). 'Estimating the reproducibility of psychological science'. *Science* 349.6251, aac4716.
Pereira, L. A., Taylor-Rodríguez, D., and Gutiérrez, L. (2020). 'A Bayesian nonparametric testing procedure for paired samples'. *Biometrics* 76.4, pp. 1133–1146.
Polson, N. G., Scott, J. G., and Willard, B. T. (2015). 'Proximal algorithms in statistics and machine learning'. *Statistical Science* 30.4, pp. 559–581.
Rosenbaum, D., Mama, Y., and Algom, D. (2017). 'Stand by your Stroop: Standing up enhances selective attention and cognitive control'. *Psychological Science* 28.12, pp. 1864–1867.
Rosset, S. and Zhu, J. (2007). 'Piecewise linear regularized solution paths'. *Annals of Statistics* 35.3, pp. 1012–1030.
Rousselet, G. A. and Wilcox, R. R. (2020). 'Reaction times and other skewed distributions: Problems with the mean and the median'. *Meta-Psychology* 4.
Roy, S. N. (1953). 'On a heuristic method of test construction and its use in multivariate analysis'. *Annals of Mathematical Statistics* 24.2, pp. 220–238.
Rubin, D. B. (1981). 'The Bayesian bootstrap'. *Annals of Statistics* 9.1, pp. 130–134.
Savage, V. M. and West, G. B. (2007). 'A quantitative, theoretical framework for understanding mammalian sleep'. *Proceedings of the National Academy of Sciences of the United States of America* 104.3, pp. 1051–1056.
Sethuraman, J. (1994). 'A constructive definition of Dirichlet priors'. *Statistica Sinica* 4.2, pp. 639–650.
Srna, S., Schrift, R. Y., and Zauberman, G. (2018a). 'The illusion of multitasking and its positive effect on performance'. *Psychological Science* 29.12, pp. 1942–1955.
— (2018b). *The illusion of multitasking and its positive effect on performance materials*. OSF Registries. https://doi.org/10.17605/OSF.IO/9UNGB.
Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P., and Ulmer, H. (2007). 'The use of statistics in medical research: A comparison of The New England Journal of Medicine and Nature Medicine'. *American Statistician* 61.1, pp. 47–55.
Student (1908). 'The probable error of a mean'. *Biometrika* 6.1, pp. 1–25.
Sun, Y. and Genton, M. G. (2011). 'Functional boxplots'. *Journal of Computational and Graphical Statistics* 20.2, pp. 316–334.

Tibshirani, R. (1996). 'Regression shrinkage and selection via the lasso'. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.
Welch, B. L. (1947). 'The generalization of 'Student's' problem when several different population variances are involved'. *Biometrika* 34.1/2, pp. 28–35.
Weng, C.-S. (1989). 'On a second-order asymptotic property of the Bayesian bootstrap mean'. *Annals of Statistics* 17.2, pp. 705–710.
Wilcoxon, F. (1945). 'Individual comparisons by ranking methods'. *Biometrics Bulletin* 1.6, pp. 80–83.
Yin, J., Mutiso, F., and Tian, L. (2021). 'Joint hypothesis testing of the area under the receiver operating characteristic curve and the Youden index'. *Pharmaceutical Statistics* 20.3, pp. 657–674.

Appendix A Huber family

A.1 Proof of Lemma 1

Proof. A sufficient condition for the result of the lemma is for $\lambda - |x_0 - \mu(\lambda)|$ to be increasing as a function of λ. To establish the function is increasing, consider its gradient:

$$\frac{\partial}{\partial \lambda}\left(\lambda - |x_0 - \mu(\lambda)|\right) = 1 + \operatorname{sign}(x_0 - \mu(\lambda))\frac{\partial \mu(\lambda)}{\partial \lambda}. \tag{A.1}$$

A sufficient condition for the gradient to be positive, and hence for $\lambda - |x_0 - \mu(\lambda)|$ to be increasing, is that $|\partial\mu(\lambda)/\partial\lambda| < 1$. The first-order condition for optimality of $\mu(\lambda)$ is

$$\mathcal{L}'(\mu(\lambda)) = -\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i(x_i - \mu(\lambda)) - \sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda)) = 0,$$

which gives

$$-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i(x_i - \mu(\lambda))} = 1. \tag{A.2}$$

Using the bound $|x_i - \mu(\lambda)| < \lambda$ in the denominator of (A.2) yields

$$\begin{aligned}
-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i(x_i - \mu(\lambda))} &> \left|-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i\lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i\lambda}\right| \\
&= \left|-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i}\right| \\
&= \left|\frac{\partial \mu(\lambda)}{\partial \lambda}\right|.
\end{aligned}$$

Together with (A.2), the above bound shows $|\partial\mu(\lambda)/\partial\lambda| < 1$, and hence the gradient (A.1) is positive. We conclude $\lambda - |x_0 - \mu(\lambda)|$ is increasing, thereby establishing the result of the lemma. □

A.2 Proof of Proposition 1

The proof of Proposition 1 requires the following lemma.

Lemma 2. *Let* $s_i := \text{sign}(x_i - \tilde{\mu})$*, where* $\tilde{\mu}$ *is the weighted median. Suppose sample point* x_0 *satisfies* $|x_0 - \mu(\lambda^\star)| \geq \lambda^\star$ *for some* $\lambda^\star > 0$*. Then* $\text{sign}(x_0 - \mu(\lambda^\star)) = s_0$*.*

Proof. We proceed using proof by contradiction and suppose $\text{sign}(x_0 - \mu(\lambda^\star)) \neq s_0$. This event can only occur if there exists a $0 < \lambda < \lambda^\star$ such that $|x_0 - \mu(\lambda)| < \lambda$, since for the sign of $x_0 - \mu(\lambda)$ to change the residual must cross through zero. But the existence of such a λ contradicts Lemma 1 since $|x_0 - \mu(\lambda^\star)| \geq \lambda^\star$. Hence, it must be the case that $\text{sign}(x_0 - \mu(\lambda^\star)) = \text{sign}(x_0 - \mu(\lambda))$ for all $0 < \lambda < \lambda^\star$. The result of the lemma immediately follows from the fact that $\lim_{\lambda \to 0} \mu(\lambda) = \tilde{\mu}$. □

We are now ready to prove Proposition 1.

Proof. By equation (3.3), $\gamma = \lambda - \lambda^+$. Since (λ^+, μ^+) is a knot point, one or more sample points cross from the square piece of the Huber function to the absolute piece and satisfy $|x_i - \mu^+| = \lambda^+$. Among all sample points eligible to cross (i.e., all i satisfying $|x_i - \mu| < \lambda$), those with with maximal absolute deviation from μ^+ cross:

$$\lambda^+ = \max_{i:|x_i - \mu| < \lambda} \left(|x_i - \mu^+| \right).$$

Together, the above expressions for γ and λ^+ give

$$\gamma = \lambda - \max_{i:|x_i - \mu| < \lambda} \left(|x_i - \mu^+| \right) = \min_{i:|x_i - \mu| < \lambda} \left(\lambda - |x_i - \mu^+| \right).$$

Since $|x_i - \mu^+| = \lambda^+$ for i satisfying the above equalities, we can invoke Lemma 2 to get

$$\gamma = \min_{i:|x_i - \mu| < \lambda} \left(\lambda - s_i(x_i - \mu^+) \right).$$

Now, making the substitution $\mu^+ = \mu - \gamma \partial \mu(\lambda)/\partial \lambda$ per equation (3.4) and rearranging terms leads to

$$0 = \min_{i:|x_i - \mu| < \lambda} \left(\lambda - s_i(x_i - \mu) - \gamma \left(1 + s_i \frac{\partial \mu(\lambda)}{\partial \lambda} \right) \right). \tag{A.3}$$

We have $1 + s_i \partial \mu(\lambda)/\partial \lambda > 0$ since $|\partial \mu(\lambda)/\partial \lambda| < 1$, as established in the proof of Lemma 1. Hence, equality (A.3) remains valid after division by $1 + s_i \partial \mu(\lambda)/\partial \lambda$ inside the minimization. Performing the division and isolating γ yields

$$\gamma = \min_{i:|x_i - \mu| < \lambda} \left(\frac{\lambda - s_i(x_i - \mu)}{1 + s_i \partial \mu(\lambda)/\partial \lambda} \right),$$

as per the result of the proposition. □

Appendix B Consistency

B.1 Proof of Theorem 1

We begin by introducing some notation. Denote the empirical distribution function by

$$P_n(x) := \sum_{i=1}^{n} \frac{1}{n} 1(X_i \leq x)$$

and the random Bayesian bootstrap distribution function by

$$G_n(x) := \sum_{i=1}^{n} w_i 1(X_i \leq x).$$

Denote the integral of the gradient of the Huber function under some distribution P by

$$A(P, \lambda, \mu) := \int \ell'_\lambda(X_i - \mu)\, dP(x).$$

The gradient $\ell'_\lambda(x - \mu)$ is Lipschitz continuous in λ, x, and μ with Lipschitz constant 1.

Our work focuses on the gradient of the Huber function, since its roots define the Huber family. The proof of the theorem requires several technical lemmas, which we now state and prove before arriving at the main argument.

We begin with Lemma 3, which provides a probabilistic finite-sample bound for the distance between $A(G_n, \lambda, \mu)$—the integral under a Bayesian bootstrap distribution—and $A(P_n, \lambda, \mu)$—the integral under the empirical distribution. This bound holds for fixed $\lambda \in \Lambda$.

Lemma 3. *Let $X_1, \ldots, X_n$ be iid random variables with distribution P_0 and, independently, let $(w_1, \ldots, w_n) \sim \text{Dirichlet}(1, \ldots, 1)$. Let $\epsilon > 0$ and $\lambda \geq 0$. Then, for any $\mu \in \mathbb{R}$, it holds*

$$\Pr(|A(G_n, \lambda, \mu) - A(P_n, \lambda, \mu)| \geq \epsilon) \leq \frac{\lambda^2}{\epsilon^2} \frac{n-1}{n+1} \frac{1}{n}.$$

Proof. Since $(w_1, \ldots, w_n)$ are Dirichlet$(1, ..., 1)$, we have

$$\begin{aligned}
\mathrm{E}(w_i) &= \frac{1}{n}, \\
\mathrm{Var}(w_i) &= \frac{1}{n+1} \frac{1}{n} \frac{n-1}{n}, \text{ and} \\
\mathrm{Cov}(w_i, w_j) &= -\frac{1}{n+1} \frac{1}{n^2} \text{ for } i \neq j.
\end{aligned}$$

First, observe that

$$\begin{aligned}
\mathrm{E}\left[A(G_n, \lambda, \mu) - A(P_n, \lambda, \mu)\right] &= \sum_{i=1}^{n} \mathrm{E}\left[\left(w_i - \frac{1}{n}\right) \ell'_\lambda(X_i - \mu)\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left\{\mathrm{E}\left[\left(w_i - \frac{1}{n}\right) \ell'_\lambda(X_i - \mu)\right] \middle| w_i\right\} \\
&= \sum_{i=1}^{n} \mathrm{E}\left[\left(w_i - \frac{1}{n}\right)\right] \mathrm{E}\left[\ell'_\lambda(X_1 - \mu)\right] \\
&= 0.
\end{aligned}$$

For the variance, we have

$$\begin{aligned}
&\operatorname{Var}\left[A(G_n,\lambda,\mu)-A(P_n,\lambda,\mu)\right] \\
&\qquad=\operatorname{Var}\left[\sum_{i=1}^{n}\left(w_i-\frac{1}{n}\right)\ell'_\lambda(X_i-\mu)\right] \\
&\qquad=\operatorname{Var}\left\{\operatorname{E}\left[\sum_{i=1}^{n}\left(w_i-\frac{1}{n}\right)\ell'_\lambda(X_i-\mu)\,\middle|\,(w_1,\ldots,w_n)\right]\right\} \\
&\qquad\quad+\operatorname{E}\left\{\operatorname{Var}\left[\sum_{i=1}^{n}\left(w_i-\frac{1}{n}\right)\ell'_\lambda(X_i-\mu)\,\middle|\,(w_1,\ldots,w_n)\right]\right\}.
\end{aligned} \tag{B.1}$$

The first-term on the right-hand side of (B.1) is zero, since

$$\begin{aligned}
&\operatorname{Var}\left\{\operatorname{E}\left[\sum_{i=1}^{n}\left(w_i-\frac{1}{n}\right)\ell'_\lambda(X_i-\mu)\,\middle|\,(w_1,\ldots,w_n)\right]\right\} \\
&\qquad=\operatorname{Var}\left\{\sum_{i=1}^{n}\left(w_i-\frac{1}{n}\right)\operatorname{E}\left[\ell'_\lambda(X_i-\mu)\right]\right\} \\
&\qquad=\operatorname{Var}(0)=0.
\end{aligned}$$

The second term on the right-hand side of (B.1) satisfies

$$\begin{aligned}
&\operatorname{E}\left\{\operatorname{Var}\left[\sum_{i=1}^{n}\left(w_i-\frac{1}{n}\right)\ell'_\lambda(X_i-\mu)\,\middle|\,(w_1,\ldots,w_n)\right]\right\} \\
&\qquad=\operatorname{E}\left\{\operatorname{Var}\left[\ell'_\lambda(X_1-\mu)\right]\sum_{i=1}^{n}\left(w_i-\frac{1}{n}\right)^2\right\} \\
&\qquad=\operatorname{Var}\left[\ell'_\lambda(X_1-\mu)\right]n\frac{1}{n+1}\frac{1}{n}\frac{n-1}{n} \\
&\qquad\leq\lambda^2\frac{n-1}{n+1}\frac{1}{n},
\end{aligned}$$

where the inequality follows from the fact that, for all $\mu\in\mathbb{R}$, the gradient $\ell'_\lambda(X_i-\mu)$ is bounded between $-\lambda$ and λ. The result of the lemma now follows directly from Chebyshev's inequality. □

Lemma 3 holds for fixed $\lambda\in\Lambda$. We now state and prove Lemma 4, which extends the bound in Lemma 3 to hold uniformly for all $\lambda\in\Lambda$.

Lemma 4. *Let $X_1,\ldots,X_n$ be iid random variables with distribution P_0 and, independently, let $(w_1,\ldots,w_n)\sim\operatorname{Dirichlet}(1,\ldots,1)$. Let $\epsilon>0$ and $\Lambda=[0,\bar{\lambda}]$ with $0<\bar{\lambda}<\infty$. Then, for any $\mu\in\mathbb{R}$, it holds*

$$\Pr\left(\sup_{\lambda\in\Lambda}\left|A(G_n,\lambda,\mu)-A(P_n,\lambda,\mu)\right|\geq\epsilon\right)\leq\left(\left\lceil\frac{4\bar{\lambda}}{\epsilon}\right\rceil+1\right)\frac{4\bar{\lambda}^2}{\epsilon^2}\frac{n-1}{n+1}\frac{1}{n}.$$

Proof. We proceed by bounding the right-hand side using an epsilon-net argument. First, we require a useful inequality. For any $\lambda \neq \lambda'$, we have

$$\begin{aligned}
|A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| &= \big|A(G_n,\lambda,\mu) - A(G_n,\lambda',\mu) \\
&\qquad + A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu) + \\
&\qquad\quad A(P_n,\lambda',\mu) - A(P_n,\lambda,\mu)\big| \\
&\leq \big|A(G_n,\lambda,\mu) - A(G_n,\lambda',\mu)\big| \\
&\qquad + \big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| \\
&\qquad\quad + \big|A(P_n,\lambda',\mu) - A(P_n,\lambda,\mu)\big| \\
&\leq |\lambda - \lambda'| + \big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| + |\lambda' - \lambda| \\
&= \big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| + 2|\lambda - \lambda'|.
\end{aligned}$$

The last inequality follows from the fact that $\ell'_\lambda(X-\mu)$ is Lipschitz in λ with constant 1 (also, at $\lambda = 0$ the left-hand side will be zero). Now, let $\mathcal{E}$ be an δ-net of Λ. For any $\lambda \in \Lambda$, there exists a $\lambda' \in \mathcal{E}$ such that $|\lambda - \lambda'| \leq \delta$. This result in combination with the previous bound gives

$$|A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| \leq \big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| + 2\delta.$$

Taking the maximum of the right-hand side and the supremum of the left-hand side yields

$$\sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| \leq \max_{\lambda'\in\mathcal{E}} \big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| + 2\delta.$$

It then follows

$$\begin{aligned}
&\Pr\left(\sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| \geq \epsilon\right) \\
&\qquad\qquad \leq \Pr\left(\max_{\lambda'\in\mathcal{E}} \big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| + 2\delta \geq \epsilon\right).
\end{aligned}$$

Next, setting $\delta = \epsilon/4$ and applying a union bound on the right-hand side, gives

$$\begin{aligned}
&\Pr\left(\max_{\lambda'\in\mathcal{E}} \big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| \geq \frac{1}{2}\epsilon\right) \\
&\qquad\qquad \leq \sum_{\lambda'\in\mathcal{E}} \Pr\left(\big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| \geq \frac{1}{2}\epsilon\right).
\end{aligned}$$

We can bound elements of the sum by Lemma 3 as

$$\Pr\left(\big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| \geq \frac{1}{2}\epsilon\right) \leq \frac{4\lambda'^2}{\epsilon^2}\frac{n-1}{n+1}\frac{1}{n} \leq \frac{4\bar{\lambda}^2}{\epsilon^2}\frac{n-1}{n+1}\frac{1}{n}.$$

The set $\mathcal{E}$ has cardinality at most $\lceil \bar{\lambda}/\delta \rceil + 1$, so

$$\sum_{\lambda'\in\mathcal{E}} \Pr\left(\big|A(G_n,\lambda',\mu) - A(P_n,\lambda',\mu)\big| \geq \frac{1}{2}\epsilon\right) \leq \left(\left\lceil \frac{4\bar{\lambda}}{\epsilon} \right\rceil + 1\right)\frac{4\bar{\lambda}^2}{\epsilon^2}\frac{n-1}{n+1}\frac{1}{n}.$$

The statement of the theorem now follows. □

We now turn to Lemma 5. Recall the Levy distance between two distributions G and F is defined as

$$d_L(G, F) := \inf\{\epsilon \mid F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon \text{ for all } x \in \mathbb{R}\}.$$

The following lemma provides a non-probabilistic bound on the distance between $A(G, \lambda, \mu)$ and $A(F, \lambda, \mu)$ when G and F are in an ϵ-neighbourhood under the Levy metric.

Lemma 5. *Let F be a distribution and $\epsilon > 0$. Let $\Lambda = [0, \bar{\lambda}]$ for some $0 < \bar{\lambda} < \infty$. Consider an open neighbourhood of F, given by $U = \{G : d_L(F, G) < \epsilon/2\}$. Then, for all $G \in U$, it holds*

$$\sup_{\lambda \in \Lambda} |A(G, \lambda, \mu) - A(F, \lambda, \mu)| < \epsilon.$$

Proof. We begin with a fixed $\lambda \in \Lambda$ and then extend to a uniform bound for all $\lambda \in \Lambda$. Let $\delta = \epsilon/2$. Then, for any λ, we have

$$\begin{aligned}
|A(G, \lambda, \mu) - A(F, \lambda, \mu)| &= \left| \int \ell'_\lambda(x-\mu)\, dG(x) - \int \ell'_\lambda(x-\mu)\, dF(x+\delta) \right. \\
&\qquad \left. + \int \ell'_\lambda(x-\mu)\, dF(x+\delta) - \int \ell'_\lambda(x-\mu)\, dF(x) \right| \\
&\leq \left| \delta + \int \ell'_\lambda(x-\mu)\, dF(x+\delta) \; - \int \ell'_\lambda(x-\mu)\, dF(x) \right| \qquad \text{(B.2)} \\
&\leq \delta + \int \left| \ell'_\lambda(x-\delta-\mu) - \ell'_\lambda(x-\mu) \right| dF(x) \\
&\leq \delta + \delta = 2\delta.
\end{aligned}$$

The first inequality is due to $d_L(G, F) < \delta$ and the third due to $\ell'_\lambda(x-\mu)$ being Lipschitz in x with constant 1. We now extend this result to hold over all $\lambda \in \Lambda$ by taking the supremum of the left-hand side of (B.2), yielding

$$\sup_{\lambda \in \Lambda} |A(G, \lambda, \mu) - A(F, \lambda, \mu)| \leq 2\delta.$$

The statement of the lemma now follows from the fact that $\delta = \epsilon/2$. □

Lemma 6 is the last piece required before establishing the result of the theorem. It draws on the preceding lemmas to prove that $A(G_n, \lambda, \mu)$ converges in probability to $A(P_0, \lambda, \mu)$ uniformly for all $\lambda \in \Lambda$.

Lemma 6. *Let $X_1, \ldots, X_n$ be iid random variables with distribution P_0 and, independently, let $(w_1, \ldots, w_n) \sim \text{Dirichlet}(1, \ldots, 1)$. Let $\epsilon > 0$ and $\Lambda = [0, \bar{\lambda}]$ with $0 < \bar{\lambda} < \infty$. Then, for any $\mu \in \mathbb{R}$, it holds*

$$\lim_{n \to \infty} \Pr\left(\sup_{\lambda \in \Lambda} |A(G_n, \lambda, \mu) - A(P_0, \lambda, \mu)| \geq \epsilon \right) = 0.$$

Proof. We begin with the inequality

$$\begin{aligned}|A(G_n,\lambda,\mu) - A(P_0,\lambda,\mu)| &= |A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu) + \\ &\qquad A(P_n,\lambda,\mu) - A(P_0,\lambda,\mu)| \\ &\leq |A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| + \\ &\qquad |A(P_n,\lambda,\mu) - A(P_0,\lambda,\mu)|.\end{aligned}$$

Taking the supremum over both sides gives

$$\begin{aligned}\sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_0,\lambda,\mu)| &\leq \sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| + \\ &\qquad \sup_{\lambda\in\Lambda} |A(P_n,\lambda,\mu) - A(P_0,\lambda,\mu)|.\end{aligned}$$

It follows

$$\begin{aligned}&\Pr\left(\sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_0,\lambda,\mu)| \geq \epsilon\right) \\ &\qquad \leq \Pr\left(\sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| \geq \epsilon\right) + \\ &\qquad\qquad \Pr\left(\sup_{\lambda\in\Lambda} |A(P_n,\lambda,\mu) - A(P_0,\lambda,\mu)| \geq \epsilon\right).\end{aligned} \tag{B.3}$$

We now bound terms on the right-hand side of (B.3). For the first term, we can apply Lemma 4. Given $\epsilon > 0$, there exists an N_1 such that for all $n \geq N_1$, it holds

$$\Pr\left(\sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_n,\lambda,\mu)| \geq \epsilon\right) \leq \frac{\epsilon}{2}.$$

Consider the second term on the right-hand side of (B.3). We have that P_n converges almost surely to P_0 by the Glivenko-Cantelli theorem and hence to a δ-neighbourhood of P_0 under the Levy metric. Set $\delta = \epsilon/2$. This result and Lemma 5 provide there exists an N_2 such that for all $n \geq N_2$, it holds

$$\Pr\left(\sup_{\lambda\in\Lambda} |A(P_n,\lambda,\mu) - A(P_0,\lambda,\mu)| \geq \epsilon\right) \leq \frac{\epsilon}{2}.$$

Set $N = \max(N_1, N_2)$. Then, for all $n \geq N$, we can bound the left-hand side of (B.3) as

$$\Pr\left(\sup_{\lambda\in\Lambda} |A(G_n,\lambda,\mu) - A(P_0,\lambda,\mu)| \geq \epsilon\right) \leq \epsilon$$

The result of the lemma now follows. □

We now have everything necessary to prove the two claims of Theorem 1.

Proof. We consider case 1 (H_0 is true) and case 2 (H_1 is true) in turn. To simplify the proof's presentation, we assume the hypothesized null value $m_0 = 0$ without loss of generality.

Case 1: H_0 is true

Suppose the null hypothesis is true and $\mu_0(\lambda)$ strictly crosses through zero. In this case, it must hold that $A(P_0, \lambda, 0) > c > 0$ for some $\lambda \in \Lambda$ and $A(P_0, \lambda, 0) < -c < 0$ for some $\lambda \in \Lambda$. It follows that both

$$\sup_{\lambda \in \Lambda} A(P_0, \lambda, 0) \geq c > 0$$

and

$$\inf_{\lambda \in \Lambda} A(P_0, \lambda, 0) \leq -c < 0$$

hold simultaneously. Our task is to show that $\sup_{\lambda \in \Lambda} A(G_n, \lambda, 0)$ is bounded below away from zero and $\inf_{\lambda \in \Lambda} A(G_n, \lambda, 0)$ is bounded above away from zero, respectively, with probability tending to one. For the supremum, we have

$$\begin{aligned}
\sup_{\lambda \in \Lambda} A(G_n, \lambda, 0) &= \sup_{\lambda \in \Lambda} (A(P_0, \lambda, 0) + A(G_n, \lambda, 0) - A(P_0, \lambda, 0)) \\
&\geq \sup_{\lambda \in \Lambda} (A(P_0, \lambda, 0) - |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|) \\
&\geq \sup_{\lambda \in \Lambda} A(P_0, \lambda, 0) - \sup_{\lambda \in \Lambda} |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|.
\end{aligned}$$

Likewise, for the infimum, it holds

$$\begin{aligned}
\inf_{\lambda \in \Lambda} A(G_n, \lambda, 0) &= \inf_{\lambda \in \Lambda} (A(P_0, \lambda, 0) + A(G_n, \lambda, 0) - A(P_0, \lambda, 0)) \\
&\leq \inf_{\lambda \in \Lambda} (A(P_0, \lambda, 0) + |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|) \\
&\leq \inf_{\lambda \in \Lambda} A(P_0, \lambda, 0) + \sup_{\lambda \in \Lambda} |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|.
\end{aligned}$$

Now, fix any $0 < \epsilon < c$. Then, by Lemma 6, there exists an N such that for all $n \geq N$, it holds with probability at least $1 - \epsilon$ that

$$\begin{aligned}
\sup_{\lambda \in \Lambda} A(G_n, \lambda, 0) &\geq \sup_{\lambda \in \Lambda} A(P_0, \lambda, 0) - \epsilon \\
&\geq c - \epsilon > 0
\end{aligned}$$

and

$$\begin{aligned}
\inf_{\lambda \in \Lambda} A(G_n, \lambda, 0) &\leq \inf_{\lambda \in \Lambda} A(P_0, \lambda, 0) + \epsilon \\
&\leq -c + \epsilon < 0.
\end{aligned}$$

Hence, the supremum is bounded below by zero and the infimum is bounded above by zero, with high probability. Thus, with probability tending to one, it must hold that $A(G_n, \lambda, 0) = 0$ for some $\lambda \in \Lambda$, and hence (by continuity of the solution path) that $\hat{\mu}(\lambda) = 0$ for some $\lambda \in \Lambda$. This result concludes the first claim of the theorem.

Case 2: H_1 is true

Suppose the alternative hypothesis is true and $\mu_0(\lambda)$ is bounded away from zero for all $\lambda \in \Lambda$. In this case, it must hold that (a) $A(P_0, \lambda, 0) \leq -c_1 < 0$ for all $\lambda \in \Lambda$ or (b) $A(P_0, \lambda, 0) \geq c_2 > 0$ for all $\lambda \in \Lambda$. For (a), it follows that

$$\sup_{\lambda \in \Lambda} A(P_0, \lambda, 0) \leq -c_1 < 0,$$

while (b) corresponds to

$$\inf_{\lambda \in \Lambda} A(P_0, \lambda, 0) \geq c_2 > 0.$$

Our task for (a) is to show that $\sup_{\lambda \in \Lambda} A(G_n, \lambda, 0)$ is also bounded above away from zero, and for (b) that $\inf_{\lambda \in \Lambda} A(G_n, \lambda, 0)$ is also bounded below away from zero, in both situations with probability tending to one. Consider (a) first. We have

$$\begin{aligned}
\sup_{\lambda \in \Lambda} A(G_n, \lambda, 0) &= \sup_{\lambda \in \Lambda} (A(P_0, \lambda, 0) + A(G_n, \lambda, 0) - A(P_0, \lambda, 0)) \\
&\leq \sup_{\lambda \in \Lambda} (A(P_0, \lambda, 0) + |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|) \\
&\leq \sup_{\lambda \in \Lambda} A(P_0, \lambda, 0) + \sup_{\lambda \in \Lambda} |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|.
\end{aligned}$$

Fix any $0 < \epsilon_1 < c_1$. By Lemma 6, there exists an N_1 such that for all $n \geq N_1$, it holds with probability at least $1 - \epsilon_1$ that

$$\begin{aligned}
\sup_{\lambda \in \Lambda} A(G_n, \lambda, 0) &\leq \sup_{\lambda \in \Lambda} A(P_0, \lambda, 0) + \epsilon_1 \\
&\leq -c_1 + \epsilon_1 < 0.
\end{aligned}$$

Hence, $\sup_{\lambda \in \Lambda} A(G_n, \lambda, 0)$ is bounded above away from zero with high probability. Turning now to (b), we have

$$\begin{aligned}
\inf_{\lambda \in \Lambda} A(G_n, \lambda, 0) &= \inf_{\lambda \in \Lambda} (A(P_0, \lambda, 0) + A(G_n, \lambda, 0) - A(P_0, \lambda, 0)) \\
&\geq \inf_{\lambda \in \Lambda} (A(P_0, \lambda, 0) - |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|) \\
&\geq \inf_{\lambda \in \Lambda} A(P_0, \lambda, 0) - \sup_{\lambda \in \Lambda} |A(G_n, \lambda, 0) - A(P_0, \lambda, 0)|.
\end{aligned}$$

Again, fix any $0 < \epsilon_2 < c_2$. Then there exists an N_2 such that for all $n \geq N_2$, it holds with probability at least $1 - \epsilon_2$ that

$$\begin{aligned}
\inf_{\lambda \in \Lambda} A(G_n, \lambda, 0) &\geq \inf_{\lambda \in \Lambda} A(P_0, \lambda, 0) - \epsilon_2 \\
&\geq c_2 - \epsilon_2 > 0.
\end{aligned}$$

Hence, $\inf_{\lambda \in \Lambda} A(G_n, \lambda, 0)$ is bounded below away from zero with high probability. Thus, with probability tending it one, it must hold that (a) $A(G_n, \lambda, 0) > 0$ for all $\lambda \in \Lambda$ and hence that $\hat{\mu}(\lambda) > 0$ for all $\lambda \in \Lambda$ or (b) $A(G_n, \lambda, 0) < 0$ for all $\lambda \in \Lambda$ and hence that $\hat{\mu}(\lambda) < 0$ for all $\lambda \in \Lambda$. This result concludes the second claim of the theorem.

□

Appendix C Intersection-union testing

The familial test we propose may be considered to have an *intersection-union* test format. Introduced by Berger (1982), an intersection-union test for a parameter $\theta \in \Theta$ is a test involving a null hypothesis that is a union of sets and an alternative hypothesis that is an intersection of sets. Specifically, letting Θ_j denote a subset of Θ for $j = 1, 2, \ldots, k$, an intersection-union test evaluates the hypotheses

$$\mathrm{H}_0 : \theta \in \cup_{j=1}^{k} \Theta_j \quad \text{vs.} \quad \mathrm{H}_1 : \theta \in \cap_{j=1}^{k} \Theta_j^c, \tag{C.1}$$

where Θ_j^c is the complement of Θ_j. If H_0 is true, θ must be contained in at least one of the Θ_j subsets. Hence, to conduct an intersection-union test, it suffices to perform k separate tests of

$$\mathrm{H}_{0j} : \theta \in \Theta_j \quad \text{vs.} \quad \mathrm{H}_{1j} : \theta \in \Theta_j^c$$

and then reject the overall null hypothesis H_0 if and only if all k individual null hypotheses H_{0j} are rejected. Berger (1982) proves the overall type I error rate of this procedure is no bigger than α if the individual tests are conducted with level α. Berger (1982) also states conditions under which intersection-union tests have size exactly equal to α, since they are generally conservative with type I error rate less than α. Berger and Hsu (1996) generalize these conditions and also provide an example where an initially conservative intersection-union test can be modified to improve its frequentist power characteristics. Li, Cao, and Zhang (2020) and Yin, Mutiso, and Tian (2021) contain some recent applications.

The connection between intersection-union tests and our familial test arises from the fact that the familial null and alternative can be broken down into a collection of individual hypotheses, each concerning a different centre indexed by λ:

$$\mathrm{H}_{0\lambda} : \mu(\lambda) \in \mathcal{M}_0 \quad \text{vs.} \quad \mathrm{H}_{1\lambda} : \mu(\lambda) \in \mathcal{M}_1.$$

Recall that $\mathcal{M}_0$ and $\mathcal{M}_1$ are a partition of the parameter space, so $\mathcal{M}_1 = \mathcal{M}_0^c$. Similar to an intersection-union test, the familial test rejects if and only if the individual null hypotheses $\mathrm{H}_{0\lambda}$ are rejected for all $\lambda \in \Lambda$. Consequently, the overall hypotheses can be expressed using a union and intersection:

$$\mathrm{H}_0 : \cup_{\lambda \in \Lambda} \{\mu(\lambda) \in \mathcal{M}_0\} \quad \text{vs.} \quad \mathrm{H}_1 : \cap_{\lambda \in \Lambda} \{\mu(\lambda) \in \mathcal{M}_1\}.$$

Here, the union and intersection are with respect to the individual events $\{\mu(\lambda) \in \mathcal{M}_0\}$ and $\{\mu(\lambda) \in \mathcal{M}_1\}$, rather than subsets of the parameter spaces as with the intersection-union test. Of course, the intersection-union hypotheses (C.1) can also be expressed in terms of individual events as $\mathrm{H}_0 : \cup_{j=1}^{k} \{\theta \in \Theta_j\}$ vs. $\mathrm{H}_1 : \cap_{j=1}^{k} \{\theta \in \Theta_j^c\}$. A key difference between the tests, however, is that the familial test involves an uncountable number of events, whereas the number of events k is typically finite in an intersection-union test. Though the Bayesian nonparametric procedure outlined in Section 2 does not formally control the size of the test, it is insightful to consider its size and power properties in repeated sampling experiments, an exercise undertaken next.

Intersection-union tests are the opposite of union-intersection tests (Roy 1953), where the null is an intersection of sets and the alternative is a union of sets. Though the familial

hypotheses we consider do not conform to this format, the analogous null hypothesis would impose that all familial centres lie in the null set and the alternative that at least one centre does not. For the Huber family, these hypotheses would correspond to a test of symmetry when $\mathcal{M}_0$ is a singleton.

Appendix D Simulations

D.1 One-sample and paired samples

We first study the one-sample setting with $X_1, \ldots, X_n$. This setting can also be interpreted as the paired samples setting where X_i is the difference of random variables. The distributions analysed are: $X \sim \text{Normal}(0, 1)$; $X \sim 0.8 \cdot \text{Normal}(0, 1) + 0.2 \cdot \text{Normal}(5, 1)$; $X \sim \text{Exponential}(1)$; $X \sim \text{Lognormal}(0, 1)$; and $X \sim \text{Poisson}(1)$. These distributions cover different types of support, modality, skewness, and tail behaviour. Figure 6 visualizes the distributions and their Huber families. For the normal, the family is a singleton. For the

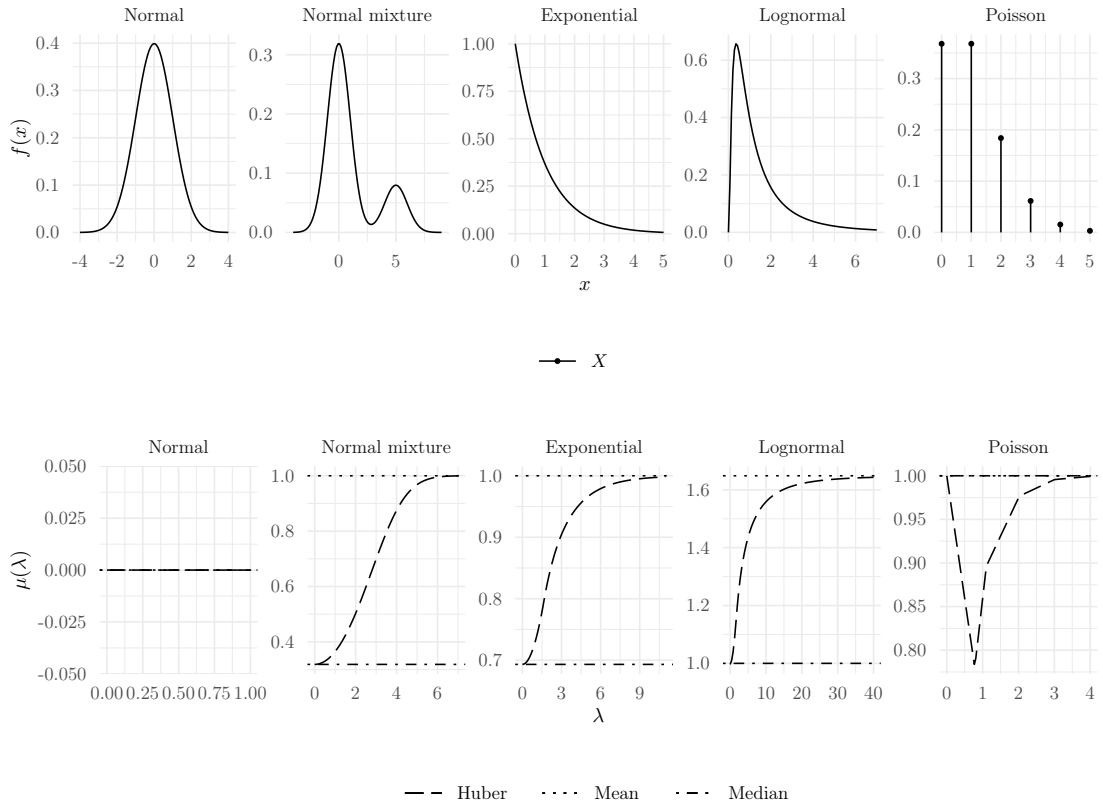

Figure 6: Distributions analysed in the one-sample (paired samples) setting. The plots in the top row depict the density or mass function for the population. The plots in the bottom row depict the corresponding Huber family.

normal mixture, exponential, and lognormal, the family is an interval with the mean and

median as its endpoints. As the Poisson demonstrates, the family need not be bounded by the mean and median.

Table 1 summarizes the tests evaluated and their associated hypotheses. A Bayesian

Test	Null hypothesis (H_0)	Alternative hypothesis (H_1)	Center
Huber familial	$\exists\, \lambda \in \Lambda : \mu(\lambda) = m_0$	$\forall\, \lambda \in \Lambda : \mu(\lambda) \neq m_0$	Huber
Student t	$\mu = m_0$	$\mu \neq m_0$	Mean
Fisher sign	$\mu = m_0$	$\mu \neq m_0$	Median
Wilcoxon signed-rank	$\mu = m_0$	$\mu \neq m_0$	Median*

* Provided X is symmetric

Table 1: Tests evaluated in the one-sample (paired samples) setting.

adaptation of the signed-rank test developed by Benavoli et al. (2014) is also evaluated. The results from this Bayesian test are not included as they are practically indistinguishable from those for the regular signed-rank test. The t and signed-rank tests are performed using `t.test` and `wilcox.test` from the `stats` package in `R`. The sign test is a special case of the binomial test, performed using `binom.test` from the same package. To handle data points equal to the null value m_0 (so-called ties) that can arise in testing discrete distributions, a modification to the sign test due to Fong et al. (2003) is used. The `wilcox.test` function uses a normal approximation, which is capable of dealing with ties. The Bayesian bootstrap used in the familial test has the advantage of being insensitive to ties. The number of Bayesian bootstraps is fixed at $B = 1,000$.

Figure 7 reports rejection frequencies for different values of m_0 as averaged over 1,000 simulations. The sample size is varied between $n = 50$ and $n = 500$. The shaded region indicates values of m_0 for which the familial (Huber) null is true. Rejection frequency inside this region indicates the size of a test according to the familial null. Power of a test according to the familial alternative is the rejection frequency outside this region. The frequentist tests are carried out at the 0.05 level. The familial test is conducted using loss matrix (2.1), which rejects when the null has posterior probability less than 0.05.

For the normal distribution, the familial test behaves similarly to the other tests. It has size no greater than 0.05 at $m_0 = 0$ and rejects sufficiently large departures from zero with high probability. Its power curve sits between those of the sign test and the signed-rank and t tests. The t test is well known to have optimal power here.

The story is more interesting for the normal mixture, exponential, and lognormal distributions. Here, the curves for the sign and t tests attain their minima at different values of m_0 since the null of each test is true at different locations. The signed-rank test fails as a test of the median due to X being asymmetric. The familial test behaves more conservatively than all three of these tests. It respects the familial null by rejecting with probability at most 0.05 in regions where some Huber centre is equal to m_0. In regions with no Huber centre equal to m_0, the familial test can be more powerful than the t or sign tests. For instance, it is more powerful than the t test for the exponential distribution when $m_0 > 1$. It is also more powerful than the sign test when $m_0 < 0.7$.

The Poisson distribution also tells an intriguing story. Since the Poisson is discrete, the power curves of the sign and signed-rank tests are step functions. In contrast to the other

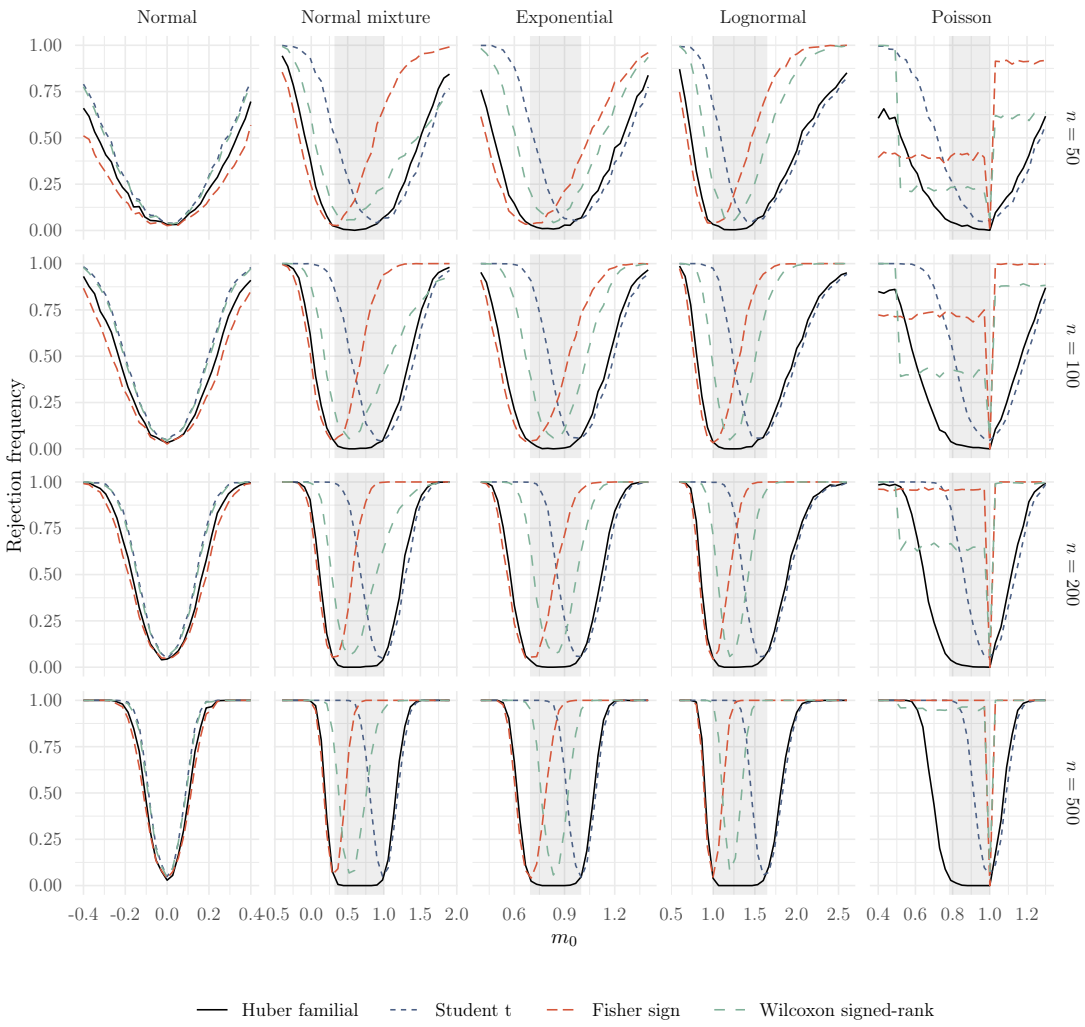

Figure 7: Rejection frequency as a function of the null value m_0 in the one-sample (paired samples) setting. The sample size $n = 200$. The shaded region indicates values of m_0 consistent with the familial null. Rejection frequency inside this region is size according to the familial null, and rejection frequency outside this region is power according to the familial alternative.

distributions, the curve of the sign test does not straddle the lower boundary of the familial null due to the lower boundary being some centre other than the median. The familial test respects its null and has good power for $m_0 > 1$ compared with the t test.

D.2 Independent samples

We now consider the independent samples setting with $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. The distributions analysed are: $X \sim \text{Normal}(0, 1)$, $Y \sim \text{Normal}(1, 1)$; $X \sim 0.8 \cdot \text{Normal}(0, 1) +$

$0.2 \cdot \text{Normal}(5, 1)$, $Y \sim \text{Normal}(0, 1)$; $X \sim \text{Exponential}(1)$, $Y \sim \text{Exponential}(2)$; $X \sim \text{Lognormal}(0, 1)$, $Y \sim \text{Lognormal}(0, 0.5)$; and $X \sim \text{Poisson}(1)$, $Y \sim \text{Poisson}(1.2)$. The distributions for X are the same as those in the one-sample setting. Figure 8 plots the distributions and corresponding differences in Huber families. For the normal, Y is a location

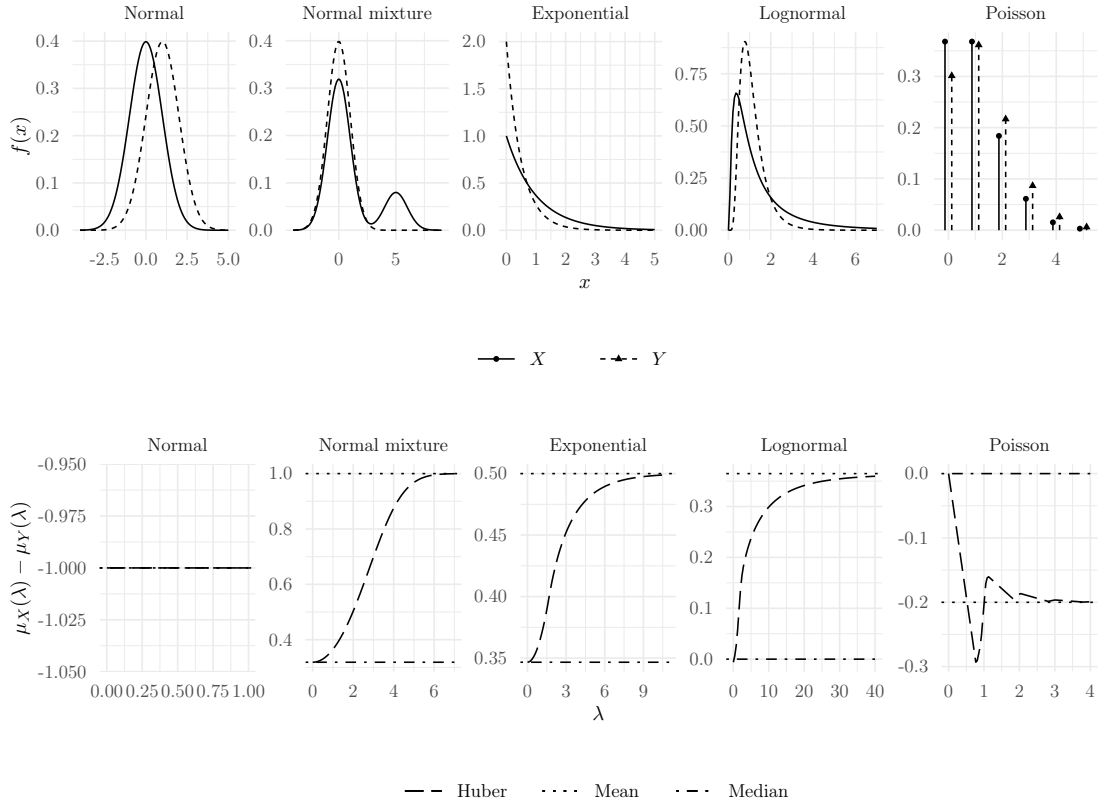

Figure 8: Distributions analysed in the independent samples setting. The plots in the top row depict the density or mass function for the populations. The plots in the bottom row depict the corresponding difference in Huber families.

shift on X, so the difference in families is a singleton. For the remaining distributions, Y has different skew and tailedness than X, so the difference in families are intervals. The Poisson is an example where the lower endpoint of the interval is not equal to the difference of means or medians.

We evaluate independent sample versions of the tests studied previously, summarized in Table 2. A Bayesian version of the rank-sum test by Benavoli et al. (2015) is also evaluated. The results from that test are not materially different from those for the regular rank-sum test, so they are not reported. The t and rank-sum tests are performed using `t.test` and `wilcox.test`. The median test is a special case of the chi-square test, performed using `chisq.test` from `stats`. Ties are again handled by `wilcox.test` via the normal approximation. For the median test, ties are discarded when calculating the test statistic.

Results from 1,000 simulations are reported in Figure 9. The shaded region again

Test	Null hypothesis (H_0)	Alternative hypothesis (H_1)	Center
Huber familial	$\exists\, \lambda \in \Lambda : \mu_X(\lambda) - \mu_Y(\lambda) = m_0$	$\forall\, \lambda \in \Lambda : \mu_X(\lambda) - \mu_Y(\lambda) \neq m_0$	Huber
Welch t	$\mu_X - \mu_Y = m_0$	$\mu_X - \mu_Y \neq m_0$	Mean
Mood median	$\mu_X - \mu_Y = m_0$	$\mu_X - \mu_Y \neq m_0$	Median
Wilcoxon rank-sum	$\mu_X - \mu_Y = m_0$	$\mu_X - \mu_Y \neq m_0$	Median*

* Provided X and Y only differ in location

Table 2: Tests evaluated in the independent samples setting.

represents values of m_0 consistent with the familial null.

The power curves for the normal distribution are not too different from the one-sample setting. The Huber centre for Y in the population is a point, so an independent samples test is not substantially different from a one-sample test with a point null.

For the normal mixture and exponential distributions, the rank-sum test fails as a test of medians since X and Y differ in scale, though curiously, it does not fail as a test of medians for the lognormal distribution. The t and median tests reject at rates above 0.05 in the middle of the familial null region, where the difference in means and difference in medians are both far from m_0. There remains another Huber centre, not equal to the mean or median, for which the difference in centres is equal to m_0. The familial test accounts for this centre and correctly accepts the null with high probability.

The median test applied to the Poisson distribution does not have the correct size at $m_0 = 0$ due to ties in the data. Likewise, the rank-sum test fails as a test of medians for the Poisson due to X and Y differing in shape and scale. The power curve of the t test does not straddle a boundary of the familial null. Unlike the median and rank-sum tests, the familial test succeeds as a test of medians, having zero size at $m_0 = 0$.

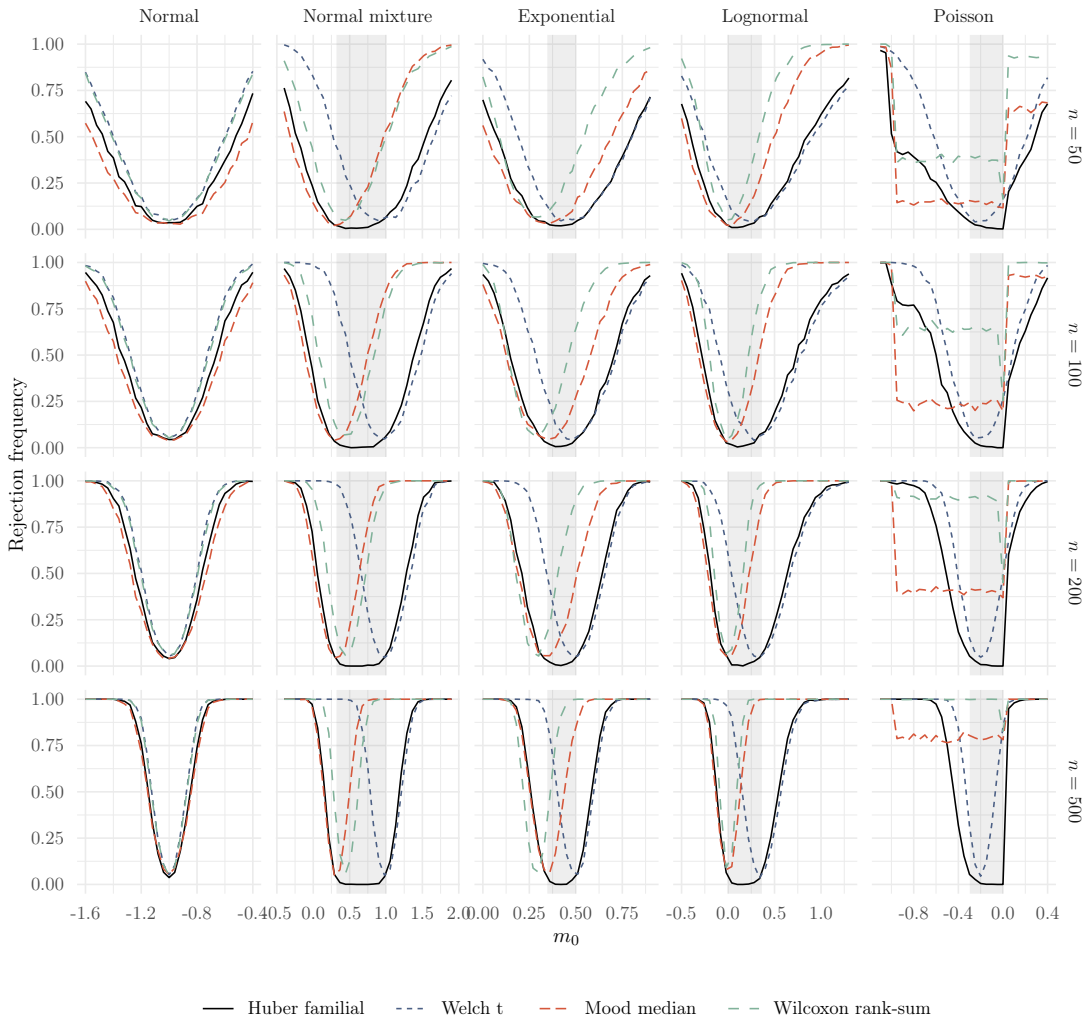

Figure 9: Rejection frequency as a function of the null value m_0 in the independent samples setting. The sample sizes $n = 200$. The shaded region indicates values of m_0 consistent with the familial null. Rejection frequency inside this region is size according to the familial null, and rejection frequency outside this region is power according to the familial alternative.