MONASH University

MONASH
BUSINESS
SCHOOL

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

Familial Inference

Ryan Thompson
Catherine S. Forbes

Working Paper No.2/22
MONTH 2022

Familial Inference

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*

February 25, 2022

Abstract

Statistical hypotheses are translations of scientific hypotheses into statements about one or more distributions, often concerning their center. Tests that assess statistical hypotheses of center implicitly assume a specific center, e.g., the mean or median. Yet, scientific hypotheses do not always specify a particular center. 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 center, this paper proposes testing a family of plausible centers, such as that induced by the Huber loss function (the "Huber family"). Each center 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 pathwise optimization routine to fit the Huber family. The favorable properties of the new test are verified through numerical simulation in one- and two-sample settings. Two experiments 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 the center μ of one or more distributions. In the simplest one-sample setting, these statements take the form

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

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 center. When μ is the mean, the most well-known of

*Corresponding author. Email: ryan.thompson@monash.edu

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

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 centers such as the mean and median raises the question of what qualifies as a "center." We posit that a center of a random variable X should satisfy at least two criteria: (1) a reflection of X about the center should preserve the center, and (2) a shift in X by a constant should move the center by that same constant. This definition is purposefully broad to accommodate the many notions of center 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 centers under these criteria as they are not preserved by reflection in general. Still, the fact that there are many possibilities for center can complicate hypothesis testing in science.

In certain applied areas (e.g., psychology and medicine), *scientific hypotheses* are often silent about a specific center 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 center 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) center 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 center 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. Transparent statistical tools are needed to instill 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 center, methods for familial inference test hypotheses about a *family* of plausible centers, with the ultimate goal of strengthening any claims of significance. More specifically, consider a family of centers $\{\mu(\lambda) : \lambda \in \Lambda\}$ where λ indexes each member (center). 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 (center) of the family is contained in the null set $\mathcal{M}_0$. The alternative hypothesis is that no member is in $\mathcal{M}_0$. This paper studies the family of centers 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 centers that includes the mean and median as limit points. All members of this "Huber family" satisfy our criteria for center.

Familial inference is more sophisticated than inference for a single center and requires new tools developed in this paper. Our first methodological development is a Bayesian nonparametric procedure for one- and two-sample testing. The procedure is based on the

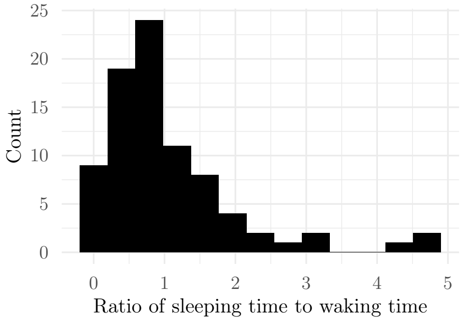

Figure 1: Histogram of the mammalian sleep data.

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 centers) 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 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.

Using our procedure, we estimate the posterior Huber family via 1000 Bayesian bootstraps,

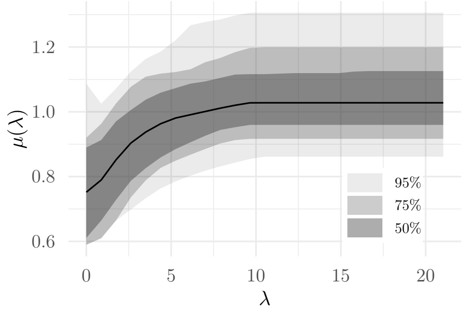

Figure 2: Functional boxplot of the posterior density of the Huber family for the mammalian sleep data. Shading indicates different central regions of the posterior.

summarized in Figure 2 by a functional boxplot. As the Huber parameter $\lambda \to \infty$, the 50% 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 center 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 favor of the alternative. Because no specific choice was made about the center, 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 center.

1.1 Organization

This paper is structured as follows. Section 2 describes the Bayesian nonparametric testing procedure. Section 3 details the pathwise algorithm for fitting the Huber family. Section 4 addresses the two-sample problem. Section 5 discusses the relation between familial testing and intersection-union testing. Section 6 presents results from numerical simulations. Section 7 illustrates the new test in two real-world case studies. Section 8 closes the paper. Proofs are available in the appendix.

2 Bayesian Nonparametric Test

This section presents our Bayesian nonparametric procedure for familial testing.

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

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

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.[1] The population

[1] Invariance of λ to spread is particularly important for testing independent samples, addressed later.

center $\mu_0(\lambda)$ 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 centers therebetween comprises the Huber family. Though our focus is the full Huber family corresponding to $\Lambda = (0, \infty)$, 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

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

where Dirichlet$(1, \dots, 1)$ 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 centers. To estimate the posterior for a given family we propose Algorithm 1. Simulating random numbers from Dirichlet$(1, \dots, 1)$ in step one is straightforward: take n iid draws from an exponential distribution with rate parameter one and rescale these draws such that their sum is one. Solving the minimization problem in step two for all $\lambda \in \Lambda$ is more complex, with the exact complexity depending 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 $w_1^{(b)}, \dots, w_n^{(b)}$). The standard deviation of the bootstrap sample could also be used.

Algorithm 1: Bayesian bootstrap for familial inference

```
  input  : $(x_1, \dots, x_n)$
  for $b = 1, \dots, B$ do
1     Sample $(w_1^{(b)}, \dots, w_n^{(b)})$ from Dirichlet$(1, \dots, 1)$
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 $\lambda \in \Lambda$
  end
  output : $\{\mu^{(b)}(\lambda) : \lambda \in \Lambda\}_{b=1}^{B}$
```

From the output of Algorithm 1, the posterior probabilities $p_{\mathrm{H}_0} := \mathrm{P}(\mathrm{H}_0 \,|\, x_1, \dots, x_n)$ of $\mathrm{H}_0 : \mu \in \mathcal{M}_0$ and $p_{\mathrm{H}_1} := \mathrm{P}(\mathrm{H}_1 \,|\, x_1, \dots, x_n)$ of $\mathrm{H}_1 : \mu \in \mathcal{M}_1$ 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, $\hat{p}_{\mathrm{H}_0} + \hat{p}_{\mathrm{H}_1} = 1$.

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

$$L := \begin{pmatrix} l_{\mathrm{H}_0|\mathrm{H}_0} & l_{\mathrm{H}_0|\mathrm{H}_1} \\ l_{\mathrm{H}_1|\mathrm{H}_0} & l_{\mathrm{H}_1|\mathrm{H}_1} \\ l_{\mathrm{I}|\mathrm{H}_0} & l_{\mathrm{I}|\mathrm{H}_1} \end{pmatrix} = \begin{array}{c} \\ \scriptstyle\mathrm{H}_0 \\ \scriptstyle\mathrm{H}_1 \\ \scriptstyle\mathrm{I} \end{array} \!\!\begin{array}{c} \begin{array}{cc} \scriptstyle\mathrm{H}_0 & \scriptstyle\mathrm{H}_1 \end{array} \\ \begin{pmatrix} 0 & 20 \\ 20 & 0 \\ 1 & 1 \end{pmatrix} \end{array}, \tag{2.1}$$

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.

3 Huber Family

To implement the testing procedure of the preceding section, we require a method for fitting the family of centers to each distribution drawn from the posterior—i.e., for solving the optimization problems in step two of Algorithm 1 given fixed bootstrap weights $w_1^{(b)}, \ldots, w_n^{(b)}$. This section develops a method for optimization with the Huber function.

3.1 Optimization Problem

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

$$\mu(\lambda) := \arg\min_{\mu \in \mathbb{R}} \mathcal{L}_\lambda(\mu) \tag{3.1}$$

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

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

and the weighted median by

$$\tilde{\mu} := \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} w_i |x_i - \mu|.$$

There is no analytical solution for the weighted median. In fact, the weighted mean is the only Huber center that admits an analytical solution in general.

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 1000, 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 chain rule gives

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

which, after evaluating gradients and rearranging terms, leads to

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

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 $0 < \lambda < \lambda^\star$, there 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.

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.

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 center and then iteratively

Algorithm 2: Pathwise optimization for the Huber family

```
input      : $(x_1, \dots, x_n)$ and $(w_1, \dots, w_n)$
initialize : $\mu^{(1)} = \sum_{i=1}^{n} w_i x_i$ and $\lambda^{(1)} = \max_i(|x_i - \mu^{(1)}|)$
Calculate the sign $s_i = \text{sign}(x_i - \tilde{\mu})$ for $i = 1, \dots, n$
for $m = 1, \dots, n-1$ do
    if $\{i : |x_i - \mu^{(m)}| < \lambda^{(m)}\} = \emptyset$ then $m = m - 1$ break
    Calculate the gradient
        $\eta = -\frac{\sum_{i:|x_i-\mu^{(m)}| \geq \lambda^{(m)}} w_i \,\text{sign}(x_i - \mu^{(m)})}{\sum_{i:|x_i-\mu^{(m)}| < \lambda^{(m)}} w_i}$
    Calculate the step size
        $\gamma = \min_{i:|x_i-\mu^{(m)}| < \lambda^{(m)}} \left( \frac{\lambda^{(m)} - s_i(x_i - \mu^{(m)})}{1 - s_i\eta} \right)$
    Perform the updates $\lambda^{(m+1)} = \lambda^{(m)} - \gamma$ and $\mu^{(m+1)} = \mu^{(m)} + \gamma\eta$.
end
output     : $(\lambda^{(1)}, \dots, \lambda^{(m+1)})$ and $(\mu^{(1)}, \dots, \mu^{(m+1)})$
```

decreases λ. The final λ in this sequence of iterates is sufficiently small to induce the weighted median as the center.

Thus far the spread has been fixed at $\sigma = 1$. To recover the solution path for $\sigma \neq 1$, we scale the output $(\lambda^{(1)}, \dots, \lambda^{(m+1)})$ from Algorithm 2 by multiplying it by σ. The centers $(\mu^{(1)}, \dots, \mu^{(m+1)})$ 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 is important for making the solution path scale-free, desirable for testing independent samples (addressed in Section 4).

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 $1/(2\lambda)(z - \beta)^2$ over $\beta \in \mathbb{R}$ (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:

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

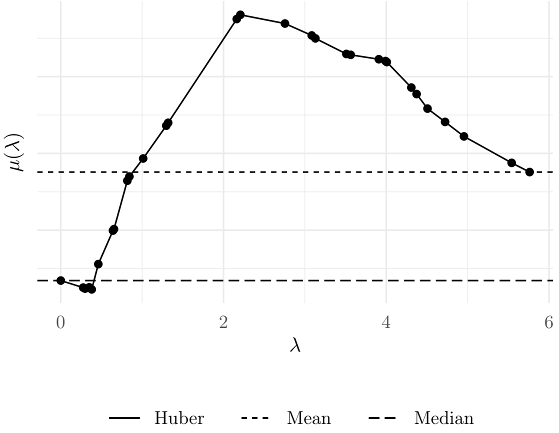

Figure 3: Algorithm 2 applied with $x_1, \ldots, x_n$ drawn from a standard normal distribution and $w_1, \ldots, w_n$ drawn from a flat Dirichlet distribution with $n = 30$. The algorithm starts at the weighted mean on the right at large λ and progresses towards the weighted median on the left at small λ. The solid points are iterates (knots) from the algorithm. Centers between iterates are linearly interpolated. Observe that the path is piecewise linear and continuous.

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 Two-Sample Problem

The discussion up to now has focused on the one-sample setting. This section addresses the two-sample setting with samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. Both paired samples and independent samples are covered.

4.1 Paired Samples

The two samples are paired if X_i and Y_i are meaningfully coupled together (e.g., measurements on the same subject before and after treatment), in which case the sample sizes n_1 and n_2 are equal. 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 center of Z. The test and algorithms of the previous sections apply directly to the sample $Z_1, \ldots, Z_n$ with $n = n_1 = n_2$.

4.2 Independent Samples

At least two different types of hypotheses are possible with independent samples. The first type is

$$\mathrm{H}_0 : \begin{array}{c} \mu_X(\lambda) - \mu_Y(\lambda) \in \mathcal{M}_0 \\ \text{for some } \lambda \in \Lambda \end{array} \quad \text{vs.} \quad \mathrm{H}_1 : \begin{array}{c} \mu_X(\lambda) - \mu_Y(\lambda) \in \mathcal{M}_1 \\ \text{for all } \lambda \in \Lambda. \end{array} \tag{4.1}$$

Here, the same center of X is compared with the same center of Y, i.e., the mean is compared with the mean, the median with the median, and so on. For the majority of situations we envisage, these hypotheses are a sensible choice. Another type of hypotheses is

$$\mathrm{H}_0 : \begin{array}{c} \mu_X(\lambda_1) - \mu_Y(\lambda_2) \in \mathcal{M}_0 \\ \text{for some } (\lambda_1, \lambda_2) \in \Lambda^2 \end{array} \quad \text{vs.} \quad \mathrm{H}_1 : \begin{array}{c} \mu_X(\lambda_1) - \mu_Y(\lambda_2) \in \mathcal{M}_1 \\ \text{for all } (\lambda_1, \lambda_2) \in \Lambda^2. \end{array} \tag{4.2}$$

With this type, every center of X is compared with every center of Y, i.e., the mean is compared with the median, the mean with the mean, etc. A test of hypotheses (4.2) is necessarily more conservative than a test of (4.1) since the former null encompasses the latter. We do not pursue (4.2) further in this paper and leave it as the subject of future work.

Testing either of these hypotheses requires bootstrapping the families of X and Y with independently drawn weights. For a test of (4.1), each center of Y is subtracted from the same center of X. These differences are recorded within each bootstrap iteration. 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 $\mathcal{M}_0$.

5 Relation to Intersection-Union Testing

The familial test we propose may be considered to have an *intersection-union* (IU) test format. Introduced by Berger (1982), an IU 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 IU test evaluates the hypotheses

$$\mathrm{H}_0 : \theta \in \cup_{j=1}^{k} \Theta_j \quad \text{vs.} \quad \mathrm{H}_1 : \theta \in \cap_{j=1}^{k} \Theta_j^c, \tag{5.1}$$

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 IU 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 IU 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 IU 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 IU 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 center 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 IU 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 IU test. Of course, the IU hypotheses (5.1) 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 IU 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.

6 Simulations

This section reports numerical simulations designed to evaluate the finite sample properties of our test. 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 1000 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 parallelized, number of processor cores.

6.1 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 analyzed are:

Test	Null hypothesis (H_0)	Alternative hypothesis (H_1)	Center
Huber familial	$\exists\,\lambda \in \Lambda : \mu(\lambda) = \mu_0$	$\forall\,\lambda \in \Lambda : \mu(\lambda) \neq \mu_0$	Huber
Student t	$\mu = \mu_0$	$\mu \neq \mu_0$	Mean
Fisher sign	$\mu = \mu_0$	$\mu \neq \mu_0$	Median
Wilcoxon signed-rank	$\mu = \mu_0$	$\mu \neq \mu_0$	Median*

* Provided X is symmetric

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

- $X \sim \text{Normal}(0, 1)$;
- $X \sim \text{Exponential}(1)$;
- $X \sim \text{Lognormal}(0, 1)$; and
- $X \sim \text{Poisson}(1)$.

These distributions cover different support types, skewness levels, and tail behaviors. Figure 4 visualizes the distributions and their Huber families. For the normal, the family is a singleton. For the 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 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 μ_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 $B = 1000$.

Figure 5 reports rejection frequencies for different values of μ_0 as averaged over 1000 simulations. The sample size is fixed at $n = 200$. The shaded region indicates values of μ_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 $\mu_0 = 0$ 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 exponential and lognormal distributions. Here, the curves for the sign and t tests attain their minima at different values of μ_0 since the null

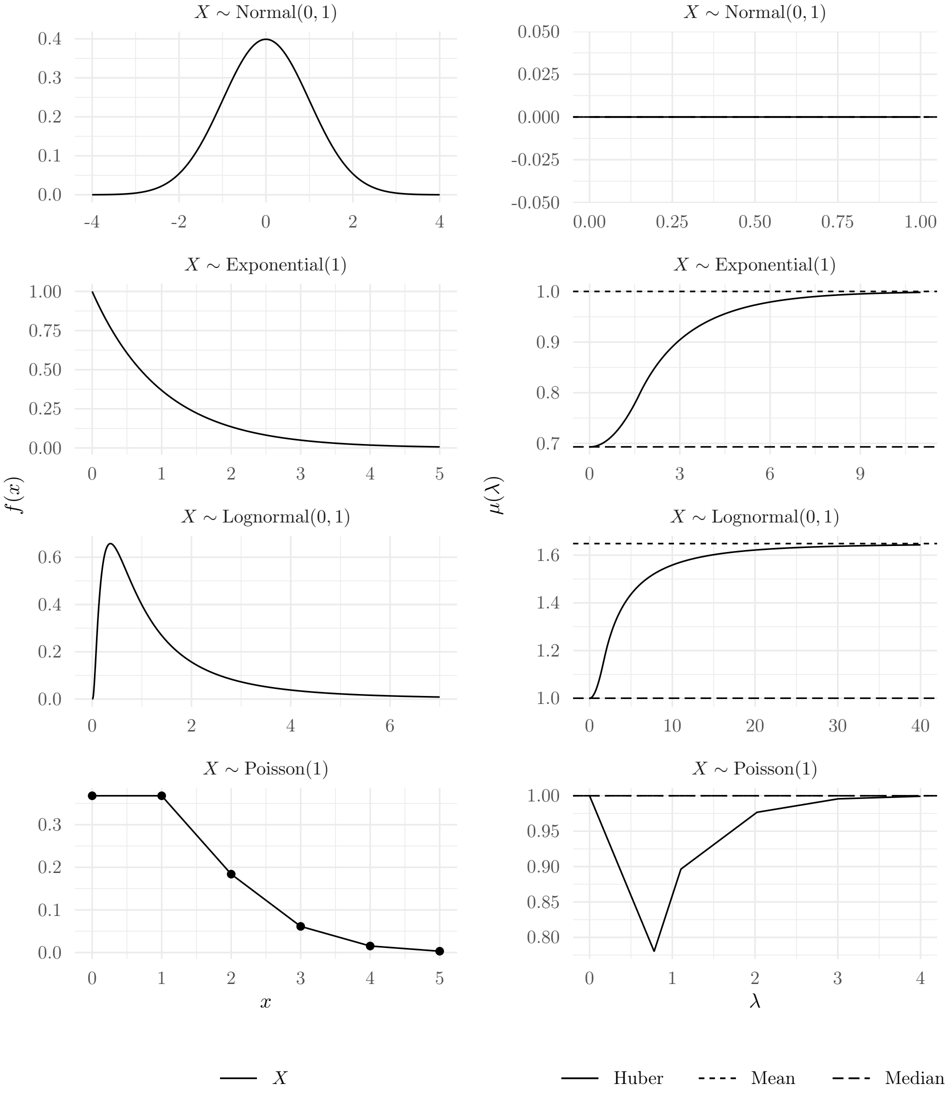

Figure 4: Distributions analyzed in the one-sample (paired samples) setting. The plots in the left column depict the density or mass function for the population. The plots in the right column depict the corresponding Huber family.

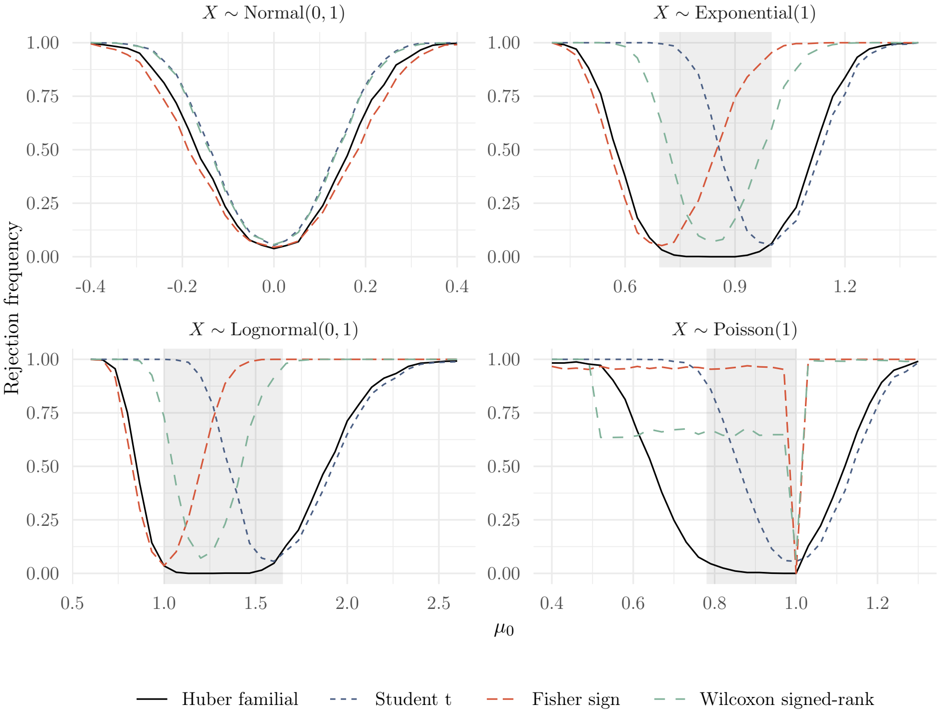

Figure 5: Rejection frequency as a function of the null value μ_0 in the one-sample (paired samples) setting. The sample size $n = 200$. The shaded region indicates values of μ_0 consistent with the familial null. Rejection frequency inside this region is size according to the familial null, and rejection frequency outside this region is power according to the familial alternative.

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 center is equal to μ_0. In regions with no Huber center equal to μ_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 $\mu_0 > 1$. It is also more powerful than the sign test when $\mu_0 < 0.7$.

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 distributions, the curve of the sign test does not straddle the lower boundary of the familial null—due to the lower boundary being some center other than the median. The familial test respects its null and has good power for $\mu_0 > 1$ compared with the t test.

6.2 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 analyzed are:

Test	Null hypothesis (H_0)	Alternative hypothesis (H_1)	Center
Huber familial	$\exists\, \lambda \in \Lambda : \mu_X(\lambda) - \mu_Y(\lambda) = \mu_0$	$\forall\, \lambda \in \Lambda : \mu_X(\lambda) - \mu_Y(\lambda) \neq \mu_0$	Huber
Welch t	$\mu_X - \mu_Y = \mu_0$	$\mu_X - \mu_Y \neq \mu_0$	Mean
Mood median	$\mu_X - \mu_Y = \mu_0$	$\mu_X - \mu_Y \neq \mu_0$	Median
Wilcoxon rank-sum	$\mu_X - \mu_Y = \mu_0$	$\mu_X - \mu_Y \neq \mu_0$	Median*

* Provided X and Y only differ in location

Table 2: Tests evaluated in the independent samples setting.

- $X \sim \text{Normal}(0, 1)$, $Y \sim \text{Normal}(1, 1)$;
- $X \sim \text{Exponential}(1)$, $Y \sim \text{Exponential}(2)$;
- $X \sim \text{Lognormal}(0, 1)$, $Y \sim \text{Lognormal}(0, 0.5)$; 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 6 plots the distributions and corresponding differences in Huber families. For the normal, Y is a location 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 1000 simulations are reported in Figure 7. The sample sizes are fixed at $n_1 = n_2 = 200$. The shaded region again represents values of μ_0 consistent with the familial null. The power curves for the normal distribution are not too different from the one-sample setting. The Huber center 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 exponential distribution, 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 μ_0. There remains another Huber center, not equal to the mean or median, for which the difference in centers is equal to μ_0. The familial test accounts for this center and correctly accepts the null with high probability.

The median test applied to the Poisson distribution does not have the correct size at $\mu_0 = 0$ 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 $\mu_0 = 0$.

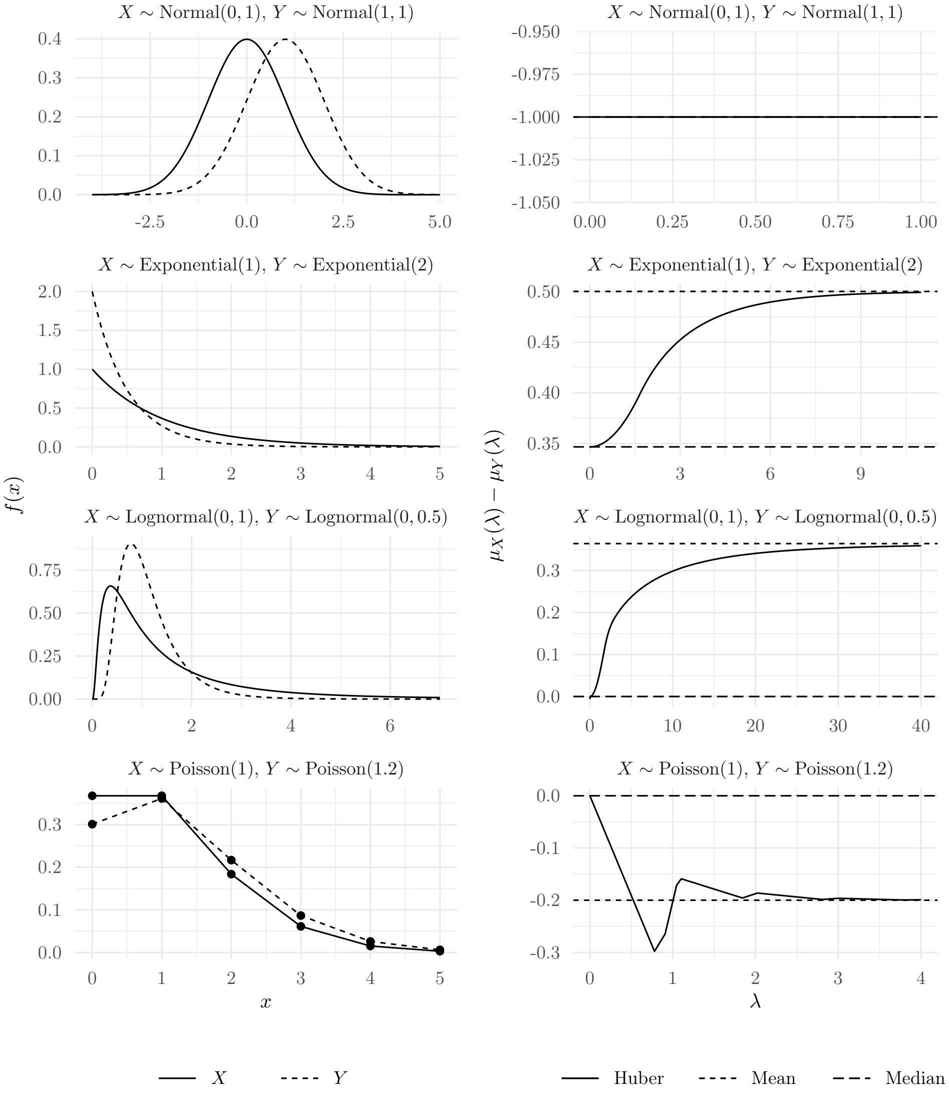

Figure 6: Distributions analyzed in the independent samples setting. The plots in the left column depict the density or mass function for the populations. The plots in the right column depict the corresponding difference in Huber families.

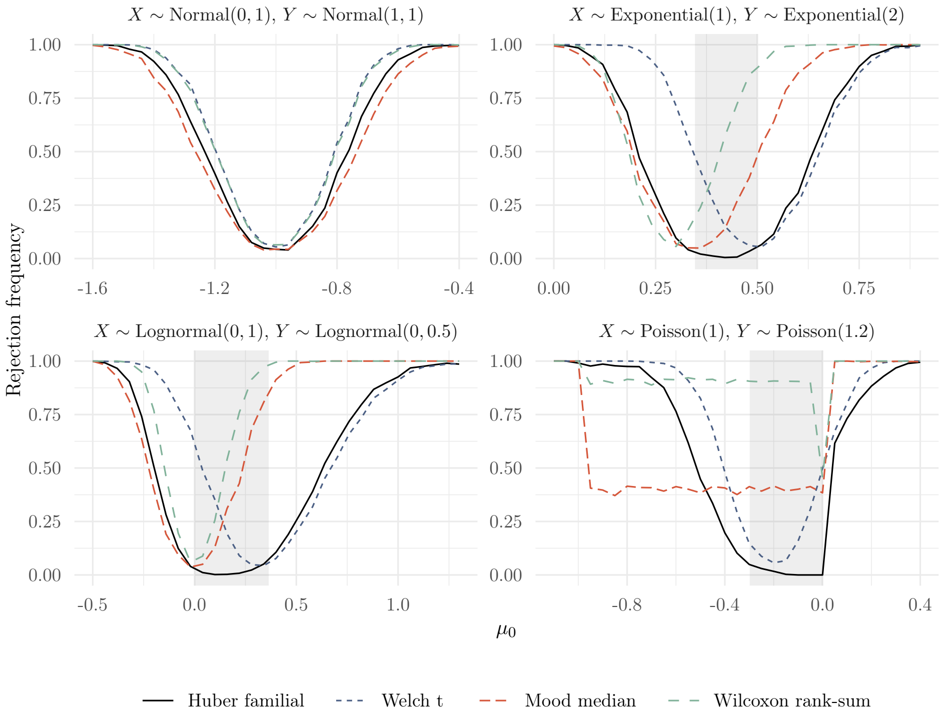

Figure 7: Rejection frequency as a function of the null value μ_0 in the independent samples setting. The sample sizes $n_1 = n_2 = 200$. The shaded region indicates values of μ_0 consistent with the familial null. Rejection frequency inside this region is size according to the familial null, and rejection frequency outside this region is power according to the familial alternative.

7 Case Studies

This section illustrates the application of familial inference to two psychology experiments. The first illustration concerns paired samples. The second illustration addresses independent samples. All tests are conducted with the same configurations and rejection criteria used in the simulations.

7.1 Body Posture Study

Rosenbaum, Mama, and Algom (2017) conducted an experiment to ascertain the effect of body posture on selective attention.[2] The experiment employed the Stroop test, where subjects are asked to announce colors 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-color 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.

[2] Refer to Experiment 3 in that paper.

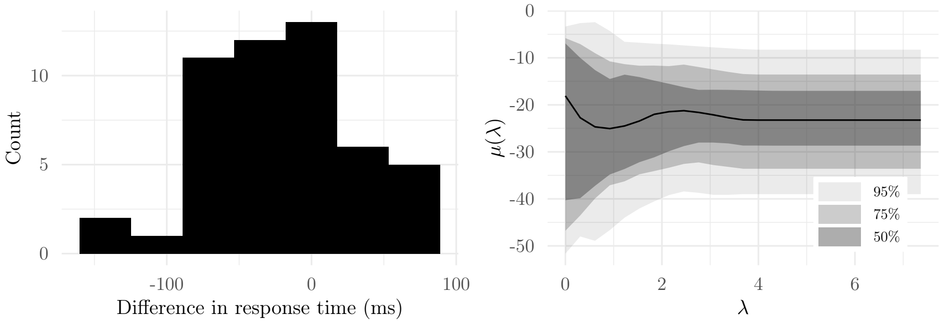

Figure 8: Body posture data. The left plot is a histogram of the data. The right plot is a functional boxplot of the posterior density of the Huber family. Shading indicates different central regions of the posterior.

The dataset contains paired observations on response times of $n = 50$ subjects. Figure 8 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 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 center tested.

7.2 Multi-Task Perception Study

Srna, Schrift, and Zauberman (2018) ran an experiment to investigate if human performance at certain activities is affected by whether the activity is perceived as multi-tasking.[3] 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 contains $n_1 = 82$ subjects in the treatment group and $n_2 = 80$ in the control group; see Figure 9. 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

[3] Refer to Study 1a in that paper.

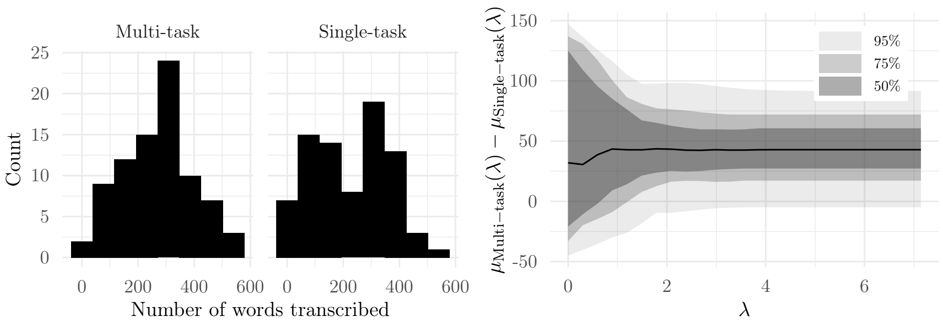

Figure 9: Multi-task perception data. The left plot is a histogram of the data by control and treatment. The right plot is a functional boxplot of the posterior density of the difference in Huber families. Shading indicates different central regions of the posterior.

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 Huber familial test returns 0.170 as the probability of the null. In contrast to the t test, the familial, sign, and rank-sum 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 remains up to subject-matter experts to decide.

8 Concluding Remarks

It has become standard practice to translate scientific hypotheses into statistical hypotheses about a specific center for the underlying distribution(s). Despite the ubiquity of this approach, there can be a lack of consensus about which center 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 centers. The contribution of this paper is to study familial inference for centers belonging to the Huber family. A natural next step in this line of work is to develop familial inference for other statistical parameters such as conditional centers (regressions). Frequentist tests can be developed along these lines as well.

Our package `familial` implements the tools developed in this paper and is publicly available on `CRAN`.

Acknowledgments

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.

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.

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.

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.

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.

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.

Srna, S., Schrift, R. Y., and Zauberman, G. (2018). “The Illusion of Multitasking and Its Positive Effect on Performance”. *Psychological Science* 29.12, pp. 1942–1955.

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

$$-\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i \lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i (x_i - \mu(\lambda))} > \left| -\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i \lambda \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i \lambda} \right|$$

$$= \left| -\frac{\sum_{i:|x_i-\mu(\lambda)|\geq\lambda} w_i \operatorname{sign}(x_i - \mu(\lambda))}{\sum_{i:|x_i-\mu(\lambda)|<\lambda} w_i} \right|$$

$$= \left| \frac{\partial \mu(\lambda)}{\partial \lambda} \right| .$$

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* $s_i := \operatorname{sign}(x_i - \tilde{\mu})$*, 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* $\operatorname{sign}(x_0 - \mu(\lambda^\star)) = s_0$*.*

Proof. We proceed using proof by contradiction and suppose $\operatorname{sign}(x_0 - \mu(\lambda^\star)) \neq s_0$. 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 $\operatorname{sign}(x_0 - \mu(\lambda^\star)) = \operatorname{sign}(x_0 - \mu(\lambda))$ 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

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

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. □