Doctoral theses at NTNU, 2021:20

Rasmus Erlemann

Contributions to the Theory of Goodness-of-Fit Testing and Change Point Detection

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Rasmus Erlemann

Contributions to the Theory of Goodness-of-Fit Testing and Change Point Detection

Thesis for the Degree of Philosophiae Doctor

Trondheim, January 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-471-9838-4 (printed ver.)
ISBN 978-82-471-9354-9 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2021:20

Printed by NTNU Grafisk senter

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree Philosophiae Doctor (PhD) at the Norwegian University of Science and Technology (NTNU). The work has mainly been carried out at the Department of Mathematical Sciences at NTNU. Part of it was also done at Department of Statistics and Actuarial Sciences at Simon Fraser University.

First of all, I would like to thank my main supervisor Bo Henry Lindqvist for support and leading by example. He has played a major role in writing this thesis and I am greatly thankful for his support.

Secondly, Richard Lockhart at SFU gave a big contribution to this thesis. I would also like to thank my second supervisor Gunnar Taraldsen. He offered his deep understanding in the subject and creative solutions to our mathematical problems. The Department of Mathematical Sciences at NTNU was very helpful and supported the research throughout my studies. International scientific collaboration is the product of our team work and I am very grateful for all the work. I would like to thank my friends Lars Simon, Mads Adrian Simonsen and Kristo Väljako for support and interesting discussions throughout the years. I am grateful to Susan Anyosa and Hannah Elissa Conway for proof reading this thesis. Most importantly, this thesis would have never came to be without my family: Birgit, Jaanus, Robin and Linda. Their unconditional support throughout the years has been significant.

Rasmus Erlemann
Trondheim, October 2020

Contents

Introduction

This thesis consists of 4 chapters. The first chapter introduces the fundamental theory used in the following articles. It is written as a graduate level text. We assume the reader is familiar with the basics of probability theory and statistics. We created examples which connect the introduced theory to the articles later on. Their purpose is to support reading the papers.

The introduction consists of 5 sections. In the first section we introduce conditional distributions and show how they can be used in hypothesis testing and hierarchical distributions. These are connected to the first and second paper. The next section is about hypothesis testing. We briefly introduce how hypothesis testing rises from decision theory. In the third section we take a look at a specific hypothesis testing problem: goodness-of-fit testing. The Cramér-von Mises test statistic is used to illustrate it. Similar test statistics are also considered in the second and third paper. In the end of this section, we introduce how conditional distributions are used in goodness-of-fit testing. The fourth section is about change point detection, which is another problem in hypothesis testing. In change point detection we use the two-sample Cramér-von Mises test to define a new test statistic. Later we explain how exact p-values are calculated in this setting and also focus on its large sample theory to calculate asymptotic p-values. In the last section, we give a summary of the introduction.

In the last 3 chapters we introduce the articles.

Vectors are denoted by a bold letter, for instance a random vector would be denoted as $\mathbf{X} = (X_1, \ldots, X_n)$. Abbreviation IID is short for independent and identically distributed.

R code that we used in this thesis is available at `https://github.com/rasmuserlemann`

1.1 Conditional Distributions

Let X be a random variable with a sample space S. We can impose a condition on it and study how its properties change. For example, we can condition on $S_{X>1} = \{s \in S : X(s) > 1\}$. This means, we define a subset of S, for which $X > 1$ holds and it is denoted by $S_{X>1}$. Let $S_{X<2}$ and $S_{X=0}$ also be events. How are the probabilities of those events affected by conditioning?

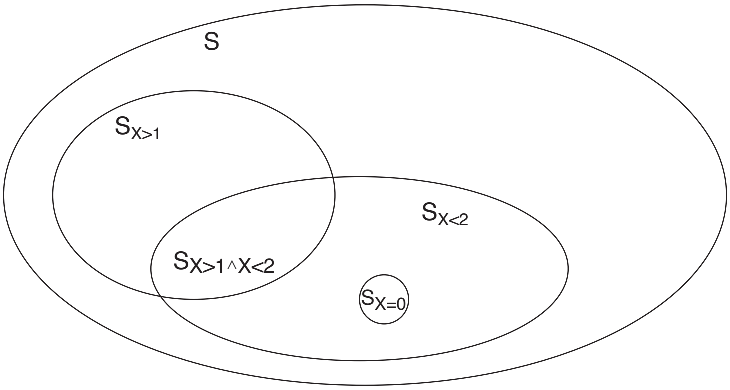

Figure 1.1: Sample space with event subsets $S_{X>1}$, $S_{X>1 \wedge X<2}$, $S_{x=0}$, $S_{X<2}$.

As we can see, $S_{X=0} \cap S_{X>1} = \emptyset$. We can conclude that the conditional probability $P(X = 0 \mid X > 1) = 0$. In this section we look at how to calculate the probability for more general cases, such as $P(X < 2 \mid X > 1)$. We define the conditional probability separately for continuous and discrete distributions in the next subsections.

Conditional distributions play an important role in statistics. For example, Bayesian statistics is built on this concept. More specifically, simulations from conditional distributions play an important role in eliminating nuisance parameters, reducing variance in Monte Carlo methods etc. There are also direct applications in other disciplines like economics [2] and finance [12].

Statistic is a function of the data **x**. Any statistic $T(\mathbf{x}) = \mathbf{t}$, defines a form of data reduction or data summary. In this thesis we focus on sufficient statistics. They capture information about the underlying parameters, while reducing the sample size.

Definition 1.1. Let **X** be a vector of IID random variables with its distribution characterized by a parameter vector $\boldsymbol{\theta}$. Statistic T is sufficient if and only if the conditional distribution $\mathbf{X} \mid T(\mathbf{X}) = \mathbf{t}$ does not depend on $\boldsymbol{\theta}$.

If we condition on the sufficient statistic's value, the resulting distribution does not depend on the parameters any more. This property is useful in eliminating underlying parameters, which we cover further on.

It is inconvenient to have to compute the conditional distribution of **X** given $T(\mathbf{X}) = \mathbf{t}$ to determine if it is sufficient or not. A simple check can be done by the so-called factorization criterion.

Theorem 1.1 (Factorization theorem)**.** *Let $X_1, \dots, X_n$ be IID random variables with joint density $f(x_1, \dots, x_n \mid \boldsymbol{\theta})$. A statistic $T(X_1, \dots, X_n) = \mathbf{t}$ is sufficient if and only if the joint density can be factorized as follows*

$$f(x_1, \dots, x_n \mid \boldsymbol{\theta}) = u(x_1, \dots, x_n) v(T(x_1, \dots, x_n), \boldsymbol{\theta}).$$

Functions u and v are non-negative. The function u only depends on the whole data $x_1, \dots, x_n$ and the function v depends on the data only through T. If the random vector $\mathbf{X}$ is discrete, we exchange f for the joint probability mass function.

In this thesis we focus on both continuous and discrete conditional distributions. The main focus is on the geometric, Gamma, uniform, inverse Gaussian and normal distributions. These distributions are widely used in practice. For example the geometric distribution is the memoryless discrete distribution and is the go to choice when the memoryless property is needed in modeling.

Often it is not possible to draw simulations from a conditional distribution because the analytical form of the probability density function or cumulative distribution function is not known. Even if we have the analytical form, applying methods such as inverse transform sampling, Metropolis-Hastings algorithm or rejection sampling can be unsuccessful. Another option is to use the Gibbs algorithm for drawing simulations [9]. In this thesis we also cover the naive sampler, which is very versatile and easy to use. However the sample outcome is only approximately from the specified distribution (in continuous case) and the computational time might be too extensive.

Conditional distributions are also used to construct hierarchical models. It means that there is some sort of hierarchical structure to their parameters. This allows us to model situations where the independence property is violated, letting samples come from the same family of distribution but with different parameters. The sample may consist of independent clusters. Samples which come from the same cluster will be more similar to each other than they will be to samples from the other clusters. In this thesis we use two-stage hierarchies.

Example 1.1. *Let us consider a two-stage hierarchical model where $X \mid p \sim Bin(n, p)$ and $p \sim Beta(\alpha, \beta)$. The given conditional probability mass function is*

$$P(X = x \mid p, n) = \binom{n}{x} p^x (1-p)^{1-x}, \quad x = 0, 1, \dots, n,$$

and the beta distribution density is

$$f_p(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)},$$

where B is the beta function,

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-t)^{\beta-1}dt.$$

We will integrate the joint probability of X and p to find the unconditional distribution. It is given by

$$\begin{aligned} P(X = x \mid n,\alpha,\beta) &= \int_0^1 f_{X,p}(x,y)dy \\ &= \int_0^1 P(X = x \mid p = y, n) f_p(y)dy \\ &= \int_0^1 \binom{n}{x} y^x (1-y)^{1-x} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} dy \\ &= \binom{n}{x} \frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}, \qquad x = 0,1,\ldots,n. \end{aligned}$$

With this we have derived the unconditional probability mass function for $X \sim betabin(n,\alpha,\beta)$.

We use the beta-geometric distribution in the second article as an alternative in the likelihood ratio test statistic. It is defined analogously as the beta-binomial distribution. Instead of the binomial distribution, there is the geometric distribution.

1.1.1 Conditional Discrete Distributions

For discrete distributions, random variables only take up to a countable number of values. It changes the way we find conditional distributions compared to the continuous case. Methods for drawing simulations are also fundamentally different.

In this thesis we focus on drawing simulations from conditional distributions, conditioned on a sufficient statistic. There are different reasons why this can be a difficult task. The probability of the event we condition on can be unknown. We might lack methods to draw simulations on the restricted support with given probabilities. For the discrete case, there is an algorithm that satisfies the first two conditions. It is called the naive sampler.

Let us look at the case in which we want to draw simulations from $\mathbf{X} \mid T(\mathbf{X}) = \mathbf{t}$, where $\mathbf{X}$ is a vector of discrete random variables and T is a sufficient statistic. Let us also assume we know how to draw simulations from $\mathbf{X}$. We would draw independent samples from $\mathbf{X}$, with freely chosen parameters and check if the condition $T(\mathbf{X}) = \mathbf{t}$ is met. If it is met, we accept it and if it is not met, we reject it. We propose a new method in this thesis to facilitate the fact that the naive sampler's acceptance rate can be very low, making it a slow method. The acceptance rate depends mostly on the value **t**. Some values are in a sense rarer and they occur less often.

1.1.2 Conditional Continuous Distributions

Continuous distributions are used in cases where the random variable can take on any real value. Given a vector of data **x**, in the continuous case, we assume that each element can be any number in $\mathbb{R}$. Simplest continuous distribution is the uniform distribution. Random variable is said to be uniformly distributed over an interval $[a,b]$, if its density is constant over $[a,b]$.

In the following example we calculate the analytical density for a conditional continuous distribution. The naive sampler is then used to estimate the marginal analytical density of X_1. We can see that the variability of the drawn samples depend highly on the chosen error rate.

Example 1.2. *Let X_1 and X_2 be IID random variables from exp(λ) and $\lambda > 0$ is the rate parameter. We will find the marginal distribution $f_{X_1 \mid X_1+X_2}$ and condition on $X_1 + X_2 = 2$. The joint density is*

$$f_{X_1,X_2}(x_1,x_2) = 2\lambda e^{-\lambda(x_1+x_2)}.$$

We can see that the joint density can be expressed through the sum $x_1 + x_2$. Theorem 1.1 indicates that the sum is a sufficient statistic. This means the conditional density does not depend on the rate parameter λ. Density of the sum can be found with marginalization to be

$$f_{X_1+X_2}(x) = \lambda^2 x e^{-\lambda x}.$$

We can find the conditional density with the Bayes formula

$$\begin{aligned} f_{X_1 \mid X_1+X_2}(x_1 \mid x_1 + x_2 = x) &= \frac{f_{X_1}(x_1) f_{X_1+X_2 \mid X_1}(x \mid X_1 = x_1)}{f_{X_1+X_2}(x)} \\ &= \frac{f_{X_1}(x_1) f_{X_2}(x - x_1)}{f_{X_1+X_2}(x)} \\ &= \frac{e^{-\lambda x_1} e^{-\lambda(x-x_1)}}{x e^{-\lambda x}} \\ &= \frac{1}{x}, \qquad\qquad 0 \le x_1 \le 2. \end{aligned}$$

We conditioned on the sum of X_1 and X_2 to be 2, so the conditional density becomes $1/2$. As the conditional distribution does not depend on x_1, it is uniform over the support $[0,2]$.

For the next figure we use the naive sampler to draw conditional samples. We chose the error rate to be $\varepsilon = 10^{-3}$. We sample X_1, X_2 independently from the exponential distribution with the maximum likelihood estimate as the parameter. If the drawn samples satisfy the conditioning statement within ε ball around $x = 2$, we accept them as valid conditional samples. In other words, the drawn samples x_1, x_2 are accepted if $2 - \varepsilon \le |x_1 + x_2| \le 2 + \varepsilon$. If the condition is not met, we disregard them and draw new ones.

The following figure shows how the samples drawn by the naive sampler approximate the analytically found density. We plot the density estimate of the X_1 marginal distribution.

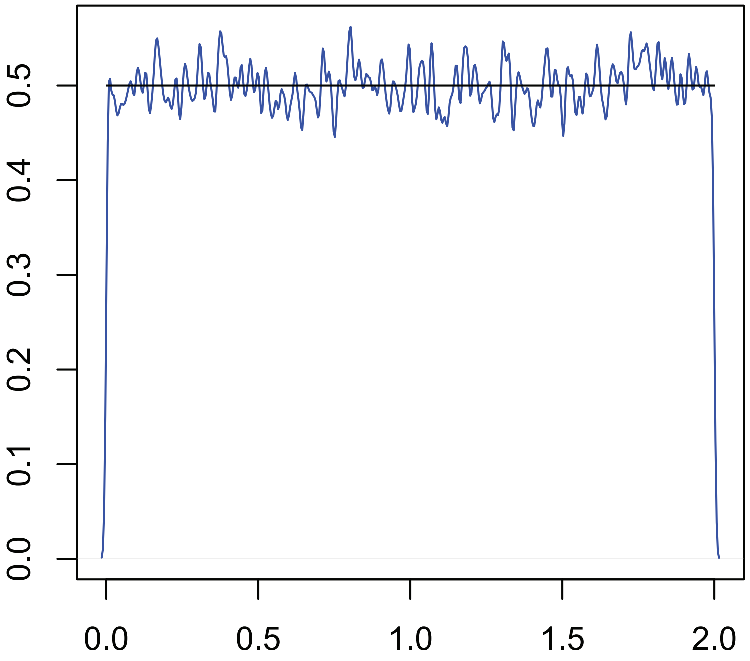

Figure 1.2: Blue line represents the naive sampler with error rate $\varepsilon = 10^{-3}$ and 10000 sample points were found to construct the density. Black line represents the analytical conditional density.

The conditional density is uniform on $[0,2]$. *The naive sampler empirial density is approximately the same, but the error term creates the difference. We accept samples for which the sum satisfies* $2-\varepsilon < x_1 + x_2 < 2+\varepsilon$. *Depending on the distribution, these sums are often skewed in one way. This means we might accept more samples where* $x_1 + x_2 > 2$ *instead of* $x_1 + x_2 < 2$, *depending on how we chose the parameters.*

1.2 Hypothesis Testing

Hypothesis testing is a branch in a more general field called decision theory [7]. In hypothesis testing we formulate a hypothesis and use statistical modeling to replicate it in a mathematical way. The general idea is that, as we are given data and we use a statistical model to compare how unlikely the data we observed is. For example, if we throw a coin 10 times and we get heads 9 out of 10 times. Is that unlikely enough to say that the coin is not fair? We can calculate the probability of observing 9 heads when throwing a fair coin. It comes out to be less than 0.01. So, can we say that the coin is not fair based on that probability?

One wishes to decide whether or not some hypothesis that has been formulated is correct. The choice lies between two decisions: accepting or rejecting the hypothesis. The decision is based on the value of a certain random vector $\mathbf{X}$ and its distribution $P_{\boldsymbol{\theta}}$ which belongs to a class $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega\}$. We want to decide on whether to accept or reject the hypothesis based on what $\boldsymbol{\theta}$ is associated with the random vector. The distributions in $\mathcal{P}$ can be classified to classes for which we accept or reject the hypothesis. The resulting mutually exclusive classes are denoted as H_0 and H_1 and the corresponding subsets of Ω are Ω_{H_0} and Ω_{H_1}. Mathematically a hypothesis is equivalent to the statement that $P_{\boldsymbol{\theta}}$ is an element of H_0. Analogously we call the distributions in H_1 the alternatives to H_0. Let the decision of accepting or rejecting H_0 be denoted by d_0 and d_1 respectively. A nonrandomized test procedure assigns a decision to each possible value $\mathbf{x}$ of $\mathbf{X}$. This means the sample space of $\mathbf{X}$ can be divided into two complementary regions: S_0 for which the hypothesis is accepted and S_1 for which the hypothesis is rejected.

Definition 1.2. Significance level α is chosen to be a real number between 0 and 1. It imposes a condition that

$$P_{\boldsymbol{\theta}}(\mathbf{X} \in S_1) \leq \alpha, \quad \boldsymbol{\theta} \in \Omega_{H_0}.$$

In other words, we impose a condition such that the probability of falsely rejecting the null hypothesis is less than the chosen significance level α. Obviously we want to keep the significance level as low as possible. Standard value for this is 0.01, 0.05 or 0.1. At the same time we want to maximize the probability of correctly rejecting the null hypothesis.

Definition 1.3. The probability

$$\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\mathbf{X} \in S_1), \quad \boldsymbol{\theta} \in \Omega_{H_1},$$

is called the power of a test against the alternative hypothesis H_1.

Throughout the thesis we also mention type I and type II errors. When performing a hypothesis test one may arrive at the correct decision or make one of two errors: rejecting the hypothesis when it is true (error of the first kind) or accepting it when it is false (error of the second kind). It is important to distinguish between these two types. For example, if we tested for the presence of a disease, incorrectly deciding on the necessity of treatment may cause the patient discomfort or financial loss but failure to diagnose the disease may lead to death. In practice, type I error is controlled by choosing the significance level α and type II error is controlled by choosing the sample size. Type II error β is closely related to the power. In fact, it is $1-\beta$.

Instead of fixing a significance level to either accept or reject the hypothesis, a popular method is to report the p-value, which leaves the choice to the reader. We chose the p-value definition from [4] because in this thesis we use tests W with the following property.

Definition 1.4. Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that the alternative hypothesis is true. For each sample point $\mathbf{x}$, the p-value is defined as

$$p(\mathbf{x}) = \sup_{\boldsymbol{\theta} \in \Theta_{H_0}} P_{\boldsymbol{\theta}}(W(\mathbf{X}) \geq W(\mathbf{x})).$$

We say the p-value is exact if the assumption of the null hypothesis is fully met. For example, when using asymptotic results or parameter estimates we make additional assumptions with the null hypothesis. As a result, the p-value we calculate is not exact. We calculate exact p-values with conditioning on a sufficient statistic. If we condition on a statistic T which is sufficient for $\mathcal{P}$, the p-value becomes

$$\begin{aligned} p(\mathbf{x}) &= \sup_{\boldsymbol{\theta} \in \Theta_{H_0}} P_{\boldsymbol{\theta}}(W(\mathbf{X}) \geq W(\mathbf{x}) \mid T(\mathbf{X}) = T(\mathbf{x}) \\ &= P(W(\mathbf{X}) \geq W(\mathbf{x})) \mid T(\mathbf{X}) = T(\mathbf{x})). \end{aligned} \tag{1.2.1}$$

We shall denote the conditional p-value with p_{cond} and it is equal to (1.2.1). The argument $\mathbf{x}$ is often omitted if we are dealing with a specific sample and it is understood from the context.

Next, we introduce the procedure of calculating p-values via parametric bootstrapping [14]. This procedure is iterated $j = 1, \ldots, M$ times.

1. Given data $\mathbf{x}$ and its parameter estimates $\hat{\boldsymbol{\theta}}$ under the null hypothesis, generate a new IID sample $\mathbf{x}_{\hat{\boldsymbol{\theta}}}^{j}$ under the null hypothesis with $\hat{\boldsymbol{\theta}}$ as parameters.

2. The sample is used to calculate the test statistic value $W\left(\mathbf{x}_{\hat{\boldsymbol{\theta}}}^{j}\right)$.

The p-value is then approximated as

$$p \approx \frac{1}{M}\sum_{i=1}^{M} I\left(W\left(\mathbf{x}_{\hat{\boldsymbol{\theta}}}^{i}\right) \geq W(\mathbf{x})\right).$$

In the following example we calculate the analytical expression of a conditional p-value. It is inspired by Example 8.3.30 from [4].

Example 1.3. *Let X_1 and X_2 be independent random variables from geometric distributions, i.e. $X_1 \sim Geom(p_1)$ and $X_2 \sim Geom(p_2)$. Consider testing $H_0 : p_1 = p_2$ against the alternative $H_1 : p_1 > p_2$. Let p denote the common value of p_1, p_2, then under the null hypothesis, the joint probability mass function is*

$$\begin{aligned} P_{X_1,X_2}(x_1, x_2 \mid p) &= (1-p)^{x_1} p (1-p)^{x_2} p \\ &= (1-p)^{x_1+x_2} p^2. \end{aligned}$$

As we can see, the joint probability can be expressed through the sum $x_1 + x_2$, taking $g_p(x_1 + x_2) = (1-p)^{x_1+x_2} p^2$. Hence, $X_1 + X_2$ is sufficient under the null hypothesis. Let us condition on $X_1 + X_2 = t$ and define a test $W(X_1, X_2) = X_2/(X_1 + X_2)$. With conditioning, the test statistic becomes X_2/t. Large values of X_2 indicate that the underlying parameter of p_2 is small. Also, since we are conditioning on $X_1 + X_2 = t$, as X_2 increases, X_1 must decrease, which suggests p_1 must increase. This explains why large values of $W(X_1, X_2) = X_2/t$ give evidence for the alternative hypothesis.

We know that $X_2 \mid X_1 + X_2 = t$ is uniformly distributed (see chapter 3). Hence the conditional p-value is

$$\begin{aligned} p_{cond} &= P(X_2/t \geq x_2/t \mid X_1 + X_2 = t) \\ &= P(X_2 \geq x_2 \mid X_1 + X_2 = t) \\ &= \sum_{i=x_2}^{t} P_{X_2 \mid X_1+X_2=t}(i) \\ &= \frac{t - x_2 + 1}{t+1}, \end{aligned} \tag{1.2.2}$$

where $P_{X_2 \mid X_1+X_2=t}$ is the conditional probability mass function of $X_2 \mid X_1 + X_2 = t$.

We considered the following 3 *cases and calculated p-values (via parametric bootstrapping), conditional p-values (via Monte Carlo simulations) and* *analytical conditional p-values from* (1.2.2)*. The maximum likelihood estimators for the parametric bootstrapping are given by $\hat{p} = n/(t+n)$.*

x_1, x_2	p-value	conditional p-value	analytical conditional p-value
$2, 30$	0.094	0.092	0.090
$3, 10$	0.278	0.286	0.285
$2, 2$	0.597	0.601	0.6

Table 1.1: Comparison of different p-values. We used 10^5 iterations to calculate the p-values and conditional p-values.

As we can see, the Monte Carlo error is very small between the analytical conditional p-values and conditional p-values. Parametric bootstrapping values are also very close to the analytical conditional p-values.

1.3 Goodness-of-Fit

Goodness-of-fit tests are used to verify a statistical model. In this type of hypothesis test, we determine whether the data fits a particular family of distributions or not. The last few decades have seen a wide range of applications in finance [10], cybersecurity [6], cosmology [15] and various other fields.

Often, the null hypothesis involves fitting a model with parameters estimated from the observed data. For example, estimating the test statistic distribution via parametric bootstrapping uses estimated parameters. We use conditional p-values defined in the previous section, so there is no need to estimate any parameters. We know that p-values found with conditional distributions and parametric bootstrapping are highly correlated [8].

Let $X_1, \ldots, X_n$ be IID random variables with distribution function F. We want to test a family of distributions $\mathcal{P} = \{F_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{k}}\}$ for a fit. In parametric goodness-of-fit testing we test the null hypothesis

$$H_0 : F \in \mathcal{P}$$

against the alternative

$$H_1 : F \notin \mathcal{P}.$$

Cramér-von Mises goodness-of-fit test statistic is one of the classical tests. It is defined in [5] for continuous data as

$$W^2 = n \int_{-\infty}^{\infty} \big(F_n(x) - F(x)\big)^2 \, dF(x),$$

where F_n is the empirical cumulative distribution function of the data and F is the cumulative distribution function under the null hypothesis with maximum likelihood estimates.

The test value describes how far the empirical distribution function is from the theoretical cumulative distribution function, assuming the null hypothesis is true. We are squaring the difference, which makes it a quadratic test. There are other quadratic test statistics, such as the Anderson-Darling test statistic. From the maximal type test statistics, there is the Kolmogorov-Smirnov test statistic.

Goodness-of-fit testing for discrete null hypotheses has been studied before in [3] and [11]. Both articles focus on the geometric distribution.

Example 1.4. *We generated two data sets. Both are size $n = 50$, the first one comes from $\mathcal{N}(0,1)$ and the second one from Gumbel$(2.5,1)$. Normal distribution maximum likelihood estimates for the first case were $\hat{\mu} = -0.002$, $\hat{\sigma} = 0.834$ and for the second case $\hat{\mu} = 1.609$, $\hat{\sigma} = 1.516$. The following plot shows how well the normal distribution density curves with maximum likelihood estimators approximate the data.*

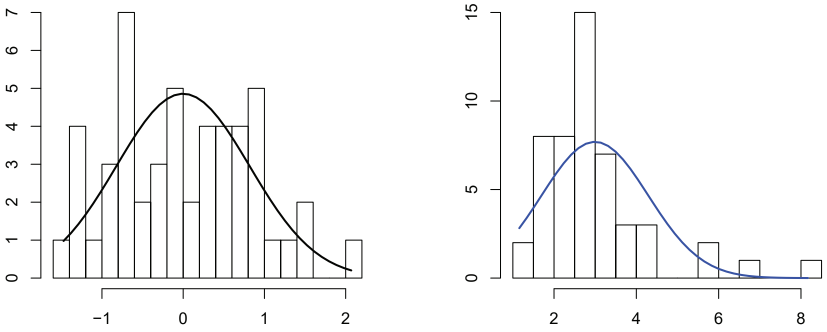

Figure 1.3: Histogram on the left represents data from the normal distribution and on the right from the Gumbel distribution. Black and blue lines are the normal distribution density lines with maximum likelihood estimators.

We calculated Cramér-von Mises test statistic values for these two data sets. These were $W^2 = 0.074$ for the first data set and $W^2 = 0.368$ for the second data set. Parametric bootstrapping p-value for the first data set was 0.726 and 0.087 for the second one.

Calculated p-values suggest that the normal distribution does not fit the second data set, if we set the significance level to be $\alpha = 0.1$. For the first data *set, the p-value is higher and we can not reject the hypothesis, that the data comes from the normal distribution.*

1.3.1 Conditional Test Distributions

Suppose we have chosen the family of distributions for the null hypothesis, we have the data and we calculated the goodness-of-fit test statistic value for the data. In order to give a quantitative assessment of whether the fit is good or not, we need to calculate the p-value. There are various different ways for calculating the p-value in goodness-of-fit testing. It can be done asymptotically, via bootstrapping or by using conditional distributions. When conditioning on a sufficient statistic, we eliminate nuisance parameters and there is no need to assume normality or estimate parameters, for what is the case when using asymptotic theory or parametric bootstrapping. Conditional p-value is found by conditioning on a sufficient statistic value of the data. Statistic is chosen such that it is sufficient with respect to the H_0 family of distributions.

Example 1.5. *Let $X_1, \dots, X_n$ be IID random variables from Geom(0.5) and $Y_1, \dots, Y_n$ are IID random variables from the discrete Weibull distribution of type I, with a probability mass function*

$$P(Y_1 = y) = q^{x^\beta} - q^{(x+1)^\beta},$$

for $x = 0, 1, \dots$ and the parameters are $q = 0.7$ and $\beta = 0.8$. Let us fix $n = 100$ and draw sample sets from both distributions. We used the Cramér-von Mises test statistic for discrete distributions, it is defined [13] as

$$W^2 = \frac{1}{n} \sum_{i=1}^{k} \hat{Z}_i^2 \hat{p}_i,$$

where $\hat{Z}_k = \sum_{j=0}^{k}(o_j - \hat{e}_j)$ and o_j is the observed number of values j in the data and $\hat{e}_j = n\hat{p}_j$ is the expected number of values j. Parameter estimates $\hat{p}_j$ is the probability of j in the geometric distribution with maximum likelihood estimates calculated from the data and $k = \max_{i=1,\dots,n} x_i$. Essentially, the Cramér-von Mises test statistic measures how far the empirical cumulative distribution function of the data is from the theoretical cumulative distribution function, under the null hypothesis with the maximum likelihood estimate.

The test statistic's distribution was calculated by drawing 10^5 samples from the conditional distribution and 10^5 parametric bootstrap samples.

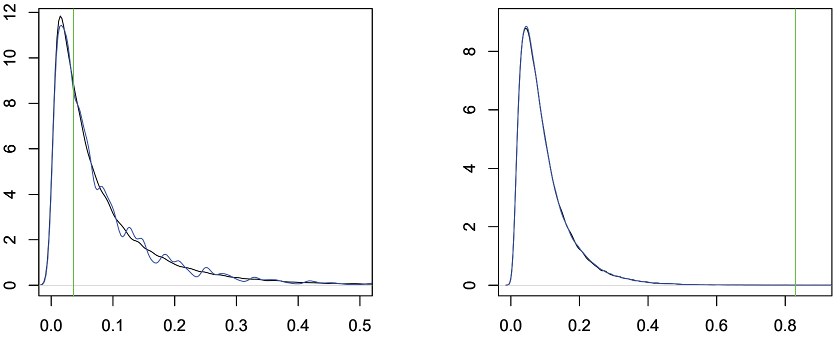

Figure 1.4: On the left we have the test distributions for $x_1,\ldots,x_{100}$ with $t = 91$ and the maximum likelihood estimator for p is 0.518. On the right we have the test statistic's distribution for $y_1,\ldots,y_{100}$ with $t = 354$ and the maximum likelihood estimator is 0.220. The black lines represent parametric bootstrapping, the blue lines conditional simulations and the green lines observed values.

We can see that the test statistic's distributions calculated with conditional simulations and parametric bootstrapping are very similar, especially for the second case. The green lines show clearly that a sample drawn under the null hypothesis is much closer to the value 0 *than the sample drawn from the discrete Weibull distribution of type I, which is on the right. This is because large values of the test statistic indicate deviation from the null hypothesis.*

1.4 Change Points Detection

Change points detection has been an active research area since its launch in the early 1950s. It has been applied in various disciplines. Some of them are economics, finance, medicine, psychology, geology and literature. There are many other works and several approaches in change point detection that are important but not included in this thesis. Our approach is univariate, nonparametric and we focus on detecting a single change point instead of multiple change points.

Let $X_1,\ldots,X_n$ be independent random variables in $\mathbb{R}$ with continuous cumulative distribution functions $F_1,\ldots,F_n$. In nonparametric change point detection we test the null hypothesis

$$H_0 : F_1 = \ldots = F_n$$

against the alternative

$$H_1 : \exists c \in \{1,\ldots,n-1\} : F_1 = \ldots = F_c \neq F_{c+1} = F_{c+2} = \ldots = F_n.$$

This is the so-called at most one change-point (AMOC) model. First c random variables have one distribution function, a change happens and the remaining $n-c$ random variables have a different distribution function. Integer c can be considered as known or unknown. If c is known, we have a two-sample problem and we can test if $X_1,\dots,X_c$ and $X_{c+1},X_{c+2},\dots,X_n$ come from the same distribution. A classical test statistic for this is the Cramér-von Mises two-sample test statistic. Let $x_1,\dots,x_{c-1},x_c,x_{c+1},\dots,x_n$ be the given sample and we want to test if there is a change point at the index c. The two-sample Cramér-von Mises test is [1] defined as

$$W_n(c) = \frac{c(n-c)}{n^2}\sum_{i=1}^{n}\left[F_c(x_i) - G_{n-c}(x_i)\right]^2.$$

Function F_c is the empirical cumulative distribution function of the first c sample elements and G_{n-c} is of the remaining $n-c$.

It is one of the quadratic test statistics. It measures how far either of the empirical cumulative distribution functions are from each other. If the test statistic value $W_n(c)$ is large, it implies that the samples come from different distributions. Large values of $W_n(c)$ give evidence to reject the null hypothesis.

If the change point c is unknown, we can use summation or maximal type test statistics which are based on the two-sample Cramér-von Mises test statistic. Let us define

$$W_{\max} = \max_{c=1,2,\dots,n-1} W_n(c),$$

$$\overline{W}_n = \frac{1}{n-1}\sum_{c=1}^{n-1} W_n(c).$$

Change point estimator is defined as

$$\hat{c} = \operatorname*{arg\,max}_{c=1,2,\dots,n-1} W_n(c).$$

The test statistic $W_n(c)$ is nonparametric and its distribution under the null hypothesis does not depend on the distribution F_1. Exact p-value can be calculated by just letting the samples be IID. The test distribution only depends on the sample size. We calculated the test statistics' distributions by generating sample sets from the standard uniform distribution, calculating test statistic values for each set and finding the cumulative distribution functions $F_{W_{\max}}$ and $F_{\overline{W}_n}$ of those values.

Given a data set $x_1,\dots,x_n$, how would we decide to reject or not to reject the null hypothesis with significance level α? Let us find the test statistic values of the data and denote them $W_{\max}^{\text{obs}}$ and $\overline{W}_n^{\text{obs}}$. Decision to reject or not can be done by comparison against the critical values $F_{W_{\max}}^{-1}(1-\alpha)$ and $F_{\overline{W}_n}^{-1}(1-\alpha)$

asymptotically. Another option is to find the p-values $1-F^{-1}_{W_{\max}}\left(W^{\text{obs}}_{\max}\right)$ and $1-F^{-1}_{\overline{W}_n}\left(\overline{W}^{\text{obs}}_n\right)$ and compare them against the significance level. The following example illustrates where the critical values position and how to calculate the exact p-values.

Example 1.6. *We generated a data set of size* $n = 100$. *First* 30 *sample points are from* $\mathcal{N}(0,1)$ *and the remaining* 70 *are from* $\mathcal{N}(1,1)$. *The second data set is all from* $\mathcal{N}(0,1)$.

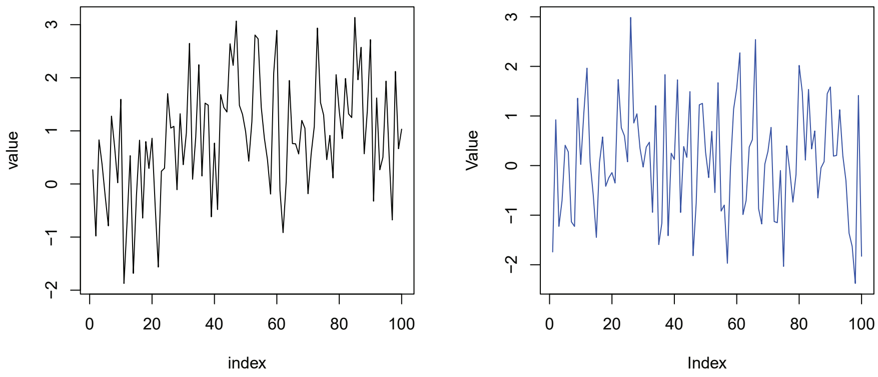

Figure 1.5: Data represented by the black line is generated under the alternative hypothesis and the blue line under the null hypothesis.

Test statistic and estimator values for the first case are

$$W_{\max} = 2.106, \quad \overline{W}_n = 0.696, \quad \hat{c} = 30$$

and for the second case

$$W_{\max} = 0.675, \quad \overline{W}_n = 0.110, \quad \hat{c} = 95.$$

Those values are displayed with the test statistic distributions. 90*th and* 95*th percentiles of* $\overline{W}_n$ *are* 0.265 *and* 0.321. *For the* $W_{\max}$, 0.826 *and* 0.963.

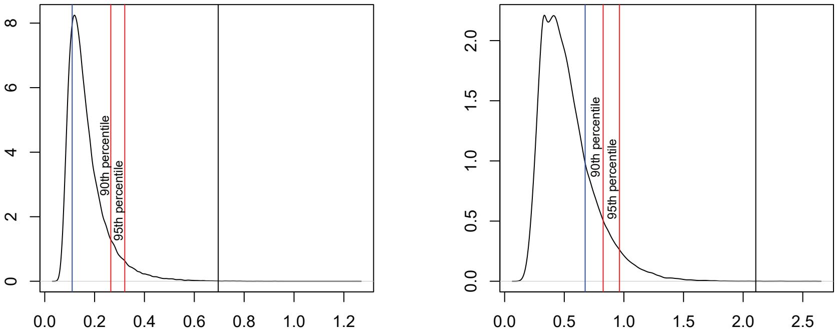

Figure 1.6: $\overline{W}_n$ (left) and $W_{\max}$ (right) distributions for $n = 100$ under the null hypothesis. Blue and black lines represent the test statistic values of the data sets.

The first data set was generated under the alternative hypothesis and the tests successfully rejected the null hypothesis for significance levels $\alpha = 0.1$ *and* $\alpha = 0.05$. *Both test statistic values were higher than the critical values. The second data set was generated under the null hypothesis. Both test statistic values were low enough to not reject the null hypothesis. Existence of a change point is not clearly visible for the first data set. The tests still detected the right index and the test statistic values were high enough to reject the null hypothesis.*

P-values for $\overline{W}_n$ *were* 0.00065 *for the first data set and* 0.645 *for the second data set. For the test statistic* $W_{\max}$*, p-values were* 0.00005 *and* 0.394.

1.4.1 Asymptotic Nonparametric Change Point Detection

The test statistics $W_{\max}$ and $\overline{W}_n$ are nonparametric, i.e. their distributions do not depend on the distribution of the data. As a method of bootstrapping, it allows us to draw sample sets from a freely chosen distribution, like the standard uniform distribution. Then, we calculate the test statistic values of the sample sets and these give us the test statistic's distribution. It only has to be done separately for each sample size n.

Another approach is to develop the large sample theory of those test statistics. The theory is readily available for summation type statistics like $\overline{W}_n$. We only touched on $W_{\max}$ and the asymptotic distribution of it is still unknown. For $\overline{W}_n$, the asymptotic distribution was found and as a result asymptotic p-values can be calculated from it. Our test statistics are based on the two sample Cramér-von Mises statistic and the large sample theory for it is well known [1]. The correlation of asymptotic and exact p-values was also studied. We found that the sample size of just $n = 200$ is large enough for these p-values to be relatively similar.

1.5 Summary

In the above discussion we have given brief introductions to the topics and main concepts that are used in this thesis. In the three papers of this thesis, we developed methods for drawing conditional simulations from both continuous and discrete distributions. These were then used in goodness-of-fit testing to calculate exact p-values. For discrete null hypothesis, we defined new test statistics which were based on the likelihood function. Next, we covered another hypothesis testing problem, change point detection. We defined new test statistics, which were based on the two-sample Cramér-von Mises test statistic. Like in the previous cases, methods for calculating exact p-values were developed. For $\overline{W}_n$ we also found asymptotic p-values.

In the next 3 chapters we introduce the articles.

In the first paper we develop a new method for drawing samples from an arbitrary conditional continuous distribution $\mathbf{X} \mid T(\mathbf{X}) = \mathbf{t}$, where T is a function. We do this by introducing an artificial parametric model, representing the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ within this new model. The key is to provide the parameter of the artificial model by a distribution. The approach is illustrated by several examples. For example, how to sample conditional uniform, normal distribution with unknown mean and two-parameter exponential family random variables. These simulations are then compared to the naive sampler. We also study how the conditional distribution samples can be used in goodness-of-fit setting with some real life data.

The second paper is focused on the same problem but for discrete distributions. The classical "stars and bars" framework is used in developing new methods for drawing samples from the conditional geometric distribution. We also define new likelihood based goodness-of-fit test statistics. The type I Weibull and beta-geometric distributions are used as alternatives in these test distributions. The new test statistics and some standard tests are also used with real life data in goodness-of-fit setting.

In the last paper we define change point detection test statistics $\overline{W}_n$ and $W_{\max}$. These are closely related to goodness-of-fit testing. We study the large sample theory of these test statistics and use it to calculate asymptotic p-values. We also study how the asymptotic and exact p-values are connected and how well the asymptotic techniques perform in comparison to the exact p-value.

References

[1] T. W. Anderson, *On the distribution of the two-sample Cramer-Von Mises criterion*, Annals of Mathematical Statistics, 33 (1962), pp. 1148–1159.

[2] S. A. Bond and K. Patel, *The Conditional Distribution of Real Estate Returns: Are Higher Moments Time Varying?*, The Journal of Real Estate Finance and Economics, 26 (2003), pp. 319–339.

[3] C. Bracquemond, E. Crétois, and O. Gaudoin, *A comparative study of goodness-of-fit tests for the geometric distribution and application to discrete time reliability*, Applied Mathematics and Computer Science, (2002).

[4] G. Casella and R. Berger, *Statistical Inference*, Duxbury Resource Center, June 2001.

[5] D. A. Darling, *The cramer-smirnov test in the parametric case*, Ann. Math. Statist., 26 (1955), pp. 1–20.

[6] M. Eling and N. Loperfido, *Data breaches: Goodness of fit, pricing, and risk measurement*, Insurance: Mathematics and Economics, 75 (2017), pp. 126 – 136.

[7] E. L. E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses E.L. Lehmann, Joseph P. Romano.*, Springer, New York, 3. ed., 2005.

[8] R. A. Lockhart, F. O'reilly, and M. Stephens, *Exact conditional tests and approximate bootstrap tests for the von mises distribution*, Journal of Statistical Theory and Practice, 3 (2009), pp. 543–554.

[9] R. A. Lockhart, F. J. O'Reilly, and M. A. Stephens, *Use of the gibbs sampler to obtain conditional tests, with applications*, Biometrika, 94 (2007), pp. 992–998.

[10] M. Mahdizadeh and E. Zamanzade, *Goodness-of-fit testing for the cauchy distribution with application to financial modeling*, Journal of King Saud University - Science, 31 (2019), pp. 1167 – 1174.

[11] S. Paul, *Testing goodness of fit of the geometric distribution: An application to human fecundability data*, Journal of Modern Applied Statistical Methods, 4 (2005), pp. 425–433.

[12] E. QIAN AND S. GORMAN, *Conditional distribution in portfolio theory*, Financial Analysts Journal, 57 (2001), pp. 44–51.

[13] J. SPINELLI AND M. STEPHENS, *Cramer-von Mises tests of fit for the Poisson distributions*, Canadian Journal of Statistics, 25 (2008), pp. 257 – 268.

[14] W. STUTE, W. G. MANTEIGA, AND M. P. QUINDIMIL, *Bootstrap based goodness-of-fit-tests*, Metrika, 40 (1993), pp. 243–256.

[15] K. TAUSCHER, D. RAPETTI, AND J. BURNS, *A new goodness-of-fit statistic and its application to 21-cm cosmology*, Journal of Cosmology and Astroparticle Physics, 2018 (2018), pp. 015–015.

Conditional Monte Carlo Revisited

Bo Henry Lindqvist, Rasmus Erlemann and Gunnar Taraldsen

Submitted and under revision, available at

`https://arxiv.org/abs/2010.07065`

Conditional Monte Carlo revisited

Bo H. Lindqvist[*], Rasmus Erlemann[†], Gunnar Taraldsen[‡]
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway

Abstract

Conditional Monte Carlo refers to sampling from the conditional distribution of a random vector $\mathbf{X}$ given the value $T(\mathbf{X}) = \mathbf{t}$ for a function $T(\mathbf{X})$. Classical conditional Monte Carlo methods were designed for estimating conditional expectations of functions $\phi(\mathbf{X})$ by sampling from unconditional distributions obtained by certain weighting schemes. The basic ingredients were the use of importance sampling and change of variables. In the present paper we reformulate the problem by introducing an artificial parametric model, representing the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ within this new model. The key is to provide the parameter of the artificial model by a distribution. The approach is illustrated by several examples, which are particularly chosen to illustrate conditional sampling in cases where such sampling is not straightforward. A simulation study and an application to goodness-of-fit testing of real data are also given.

Keywords – change of variable, conditional distribution, exponential family, goodness-of-fit testing, Monte Carlo simulation, sufficiency

[*] bo.lindqvist@ntnu.no

[†] rasmus.erlemann@ntnu.no

[‡] gunnar.taraldsen@ntnu.no

1 Introduction

Suppose we want to sample from the conditional distribution of a random vector $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ for a function $T(\mathbf{X})$ of $\mathbf{X}$. Trotter and Tukey (1956) presented an interesting technique which they named *conditional Monte Carlo*. Their idea was to determine a weight $w_{\mathbf{t}}(\mathbf{X})$ and a modified sample $\mathbf{X}_{\mathbf{t}}$ such that $\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] = \mathrm{E}[\phi(\mathbf{X}_{\mathbf{t}})w_{\mathbf{t}}(\mathbf{X})]$ for any function ϕ, thus replacing conditional expectations by ordinary expectations and allowing Monte Carlo computation.

Although the authors were aware that the method had generalizations, they confined themselves to rather special cases. Hammersley (1956) used their idea in a slightly more general and flexible analytic setting, see also Chapter 6 of the monograph by Hammersley and Handscomb (1964). Wendel et al. (1957) gave an alternative explanation, wherein the group-theoretic aspect of the problem played the dominant role. Later, Dubi and Horowitz (1979) gave an explanation of conditional Monte Carlo in terms of importance sampling and change of variables. Their approach provides a framework by which in principle any conditional sampling problem can be handled, and is the survivor in textbooks (Ripley, 1987; Evans and Swartz, 2000). Conditional Monte Carlo, in the form as introduced in the 1950s and the following nearest decades, has apparently received little attention in the later literature and has seemingly remained theoretically underdeveloped. An interesting recent reference is Feng and Liu (2016) who exploit the change of variables framework of conditional Monte Carlo with application to sensitivity estimation for financial options.

Sampling from conditional distributions has been of particular interest in statistical inference problems involving sufficient statistics (Lehmann and Romano, 2005; Lehmann and Casella, 1998). Typical applications are in construction of optimal estimators, nuisance parameter elimination and goodness-of-fit testing. In some special cases one is able to derive conditional distributions analytically. Typically this is not possible, however, thus leading to the need for approximations or simulation algorithms.

Engen and Lillegård (1997) considered the general problem of Monte

Carlo computation of conditional expectations given a sufficient statistic. Their approach was further studied and generalized by Lindqvist and Taraldsen (2005), see also Lindqvist et al. (2003) and Lindqvist and Taraldsen (2007). Further applications of the technique can be found in Schweder and Hjort (2016), pp. 239, 250. Consider a statistical model where a random vector $\mathbf{X}$ has a distribution indexed by the parameter θ, and suppose the statistic $\mathbf{T}$ is sufficient for θ. The basic assumption is that there is given a random vector $\mathbf{U}$ with a known distribution, such that $(\mathbf{X}, \mathbf{T})$ for a given parameter value θ, say, can be simulated by $(\chi(\mathbf{U}, \theta), \tau(\mathbf{U}, \theta))$ for given functions χ and τ. Let $\mathbf{t}$ be the observed value of $\mathbf{T}$, and suppose that a sample from the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}$ is wanted. Since the conditional distribution by sufficiency does not depend on θ, it seems reasonable that it can be described in some simple way in terms of the distribution of $\mathbf{U}$, and thus enabling Monte Carlo simulation based on $\mathbf{U}$. The main idea of Engen and Lillegård (1997) was to first draw $\mathbf{U} = \mathbf{u}$ from its known distribution, then to determine a parameter value $\hat{\theta}$ such that $\tau(\mathbf{u}, \hat{\theta}) = t$ and finally to use $\chi(\mathbf{u}, \hat{\theta})$ as the desired sample. In this way one indeed gets a sample of $\mathbf{X}$ with the corresponding $\mathbf{T}$ having the correct value $\mathbf{t}$. However, as shown by Lindqvist and Taraldsen (2005), only under a so-called pivotal condition will this be a sample from the true conditional distribution. The clue (Lindqvist and Taraldsen, 2005) is to let the parameter θ be given a suitable distribution, changing it to a random vector Θ, independent of $\mathbf{U}$, and showing that the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}$ equals the conditional distribution of $\chi(\mathbf{U}, \Theta)$ given $\tau(\mathbf{U}, \Theta) = \mathbf{t}$.

In the present paper, motivated by the classical approaches of conditional Monte Carlo, we construct a method for sampling from conditional distributions of $\mathbf{X}$ given $\mathbf{T} \equiv T(\mathbf{X}) = \mathbf{t}$ in general, without reference to a particular statistical model and sufficiency. As was suggested in Lindqvist and Taraldsen (2005), this could in principle be done by embedding the pair $(\mathbf{X}, \mathbf{T})$ in an artificial parametric model where $\mathbf{T}$ is a sufficient statistic, and proceed as above. This may, however, often not be a simple task, if practically doable at all. While the new method is also based on the construction of an artificial parametric statistical model, sufficiency of $\mathbf{T}$ is not part of this construction.

As will be demonstrated in examples, algorithms derived from the present approach will often be more attractive than the ones based on the sufficency approach as described above.

The main idea of the new method is to construct an artificial statistical model for a random vector $\mathbf{U}$ with distribution depending on a parameter θ, such that a "pivot" $\chi(\mathbf{U}, \theta)$ has the same distribution as $\mathbf{X}$ for each θ. Moreover, defining $\tau(\mathbf{U}, \theta) = T(\chi(\mathbf{U}, \theta))$, and considering θ as the realization of a random Θ, it will follow that the pair $(\chi(U, \Theta), \tau(U, \Theta))$ has the same distribution as $(\mathbf{X}, \mathbf{T})$. Consequently, the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}$ equals the conditional distribution of $\chi(\mathbf{U}, \Theta)$ given $\tau(\mathbf{U}, \Theta) = \mathbf{t}$. This end result similar to what was described above for the approach of Lindqvist and Taraldsen (2005), but a crucial difference from the latter approach is that the $\mathbf{U}$ and Θ are no longer independent.

As indicated above, an advantage of the new approach is that it applies to a single distribution for $\mathbf{X}$ instead of a parametric model. Thus, when applied to conditional sampling given a sufficient statistic, the method may be based on the original model only under a conveniently chosen single parameter value, for example using a standard exponential distribution when the model is a two-parameter gamma model as in Section 4.2.1.

We give several examples to demonstrate the approach and illustrate different aspects of the theoretical derivations. In particular, the examples include a new method for sampling of uniformly distributed random variables conditional on their sum, where the method of embedding the distribution into a parametric family and using sufficiency is much less attractive than the new method. Other examples consider conditional sampling given sufficent statistics in the gamma and inverse Gaussian models, as well as a new treatment of a classical example from Trotter and Tukey (1956).

The recent literature contains several other approaches to conditional sampling. For example, Lockhart et al. (2007) and Lockhart et al. (2009) studied the use of Gibbs sampling to generate samples from the conditional distribution given the minimal sufficient statistic for the gamma distribution and the von Mises distribution, respectively. Gracia-Medrano and O'Reilly (2005) and O'Reilly and Gracia-Medrano (2006) constructed corresponding sampling methods based on the Rao-Blackwell theorem, while Santos and Filho (2019) suggested a method using the Metropolis-Hastings algorithm. An older reference for conditional sampling in the inverse Gaussian distribution is Cheng (1984).

The present paper is structured as follows. In Section 2 we explain the main method and prove the basic results underpinning the approach. Specific methods for simulation and computation within the approach are also briefly described. Some further explanations and theoretical extensions are given in Section 3. Section 4 is devoted to examples, in particular for a general two-parameter exponential family of positive variables. Some simulation results which indicate the correctness of the methods are given in Section 5, while an example of goodness-of-fit testing with real data is given in Section 6. Some final remarks are given in Section 7. The paper is concluded by an Appendix containing two lemmas referred to earlier in the paper.

2 The main method

Let $\mathbf{X}$ be a random vector and let $\mathbf{T} = T(\mathbf{X})$ be a function of $\mathbf{X}$. Our aim is to sample from the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}$. As indicated in the Introduction, the idea is to construct a pair $(\mathbf{U}, \Theta)$ of random vectors and functions $\chi(\mathbf{U}, \Theta)$ and $\tau(\mathbf{U}, \Theta)$ such that this conditional distribution equals the one of $\chi(\mathbf{U}, \Theta)$ given $\tau(\mathbf{U}, \Theta) = t$.

Let $\mathbf{U}$ be a random vector with values in $\mathcal{U}$ and distribution P_θ depending on a parameter $\theta \in \Omega$. Assume that there is a function $\chi(\mathbf{u}, \theta)$ defined for $\mathbf{u} \in \mathcal{U}$, $\theta \in \Omega$, such that

$$\chi(\mathbf{U}, \theta) \sim \mathbf{X} \text{ for each } \theta \in \Omega. \tag{1}$$

Here '$\sim$' means 'having the same distribution as', and $\mathbf{U}$ in (1) is assumed to have the distibution P_θ. Note that $\chi(\mathbf{U}, \theta)$ is then a *pivot* in the statistical model defined by $\mathbf{U}$ and P_θ.

The following result is basic to our approach. Let notation and assumptions be as above and let $\tau(\mathbf{u}, \theta) = T(\chi(\mathbf{u}, \theta))$ for $\mathbf{u} \in \mathcal{U}$ and $\theta \in \Omega$.

Theorem 1. *Let Θ be a random vector taking values in Ω and let $\mathbf{U}$ conditional on $\Theta = \theta$ be distributed as P_θ. If χ satisfies (1), then the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = t$ is equal to the conditional distribution of $\chi(\mathbf{U}, \Theta)$ given $\tau(\mathbf{U}, \Theta) = t$.*

Proof. It is enough to prove that $\chi(\mathbf{U}, \Theta) \sim \mathbf{X}$. Then it will follow that $(\chi(\mathbf{U}, \Theta), \tau(\mathbf{U}, \Theta)) \sim (\mathbf{X}, \mathbf{T})$ and the result of the theorem will follow. Now, by (1), for any bounded function ϕ,

$$\mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))] = \mathrm{E}\left[\mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))|\Theta\right] = \mathrm{E}[\phi(\mathbf{X})].$$

Since this holds for all ϕ, we conclude that $\chi(\mathbf{U}, \Theta) \sim \mathbf{X}$. □

The following result shows how $\mathbf{U}$ and P_θ can be constructed from a function $\chi(\mathbf{u}, \theta)$.

Proposition 1. *Let $\mathbf{X}$ be a random vector with density $f_{\mathbf{X}}(\mathbf{x})$ and support $\mathcal{X}$. Let further $\chi(\mathbf{u}, \theta)$ for $\mathbf{u} \in \mathcal{U}$, $\theta \in \Omega$ be such that $\chi(\mathbf{u}, \theta)$ for each fixed $\theta \in \Omega$ has a range that contains $\mathcal{X}$, is differentiable, and is one-to-one with a continuous inverse. Let $\mathbf{U}$ be a random vector taking values in $\mathcal{U}$, with distribution depending on $\theta \in \Omega$ and given by the density*

$$f(\mathbf{u} \mid \theta) = f_{\mathbf{X}}(\chi(\mathbf{u}, \theta)) \left|\det \partial_{\mathbf{u}} \chi(\mathbf{u}, \theta)\right|. \tag{2}$$

Then (1) holds.

Proof. Let ϕ be an arbitrary bounded function and fix a θ. Then by a standard change of variable argument (Rudin, 1987, Theorem 7.26) we have

$$\begin{aligned}
\mathrm{E}[\phi(\chi(\mathbf{U}, \theta))] &= \int \phi(\chi(\mathbf{u}, \theta)) f(\mathbf{u} \mid \theta) d\mathbf{u} \\
&= \int \phi(\chi(\mathbf{u}, \theta)) f_{\mathbf{X}}(\chi(\mathbf{u}, \theta)) \cdot \left|\det \partial_{\mathbf{u}} \chi(\mathbf{u}, \theta)\right| d\mathbf{u} \\
&= \int \phi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \mathrm{E}[\phi(\mathbf{X})].
\end{aligned}$$

The result of the proposition then holds since ϕ was arbitrarily chosen. □

We now introduce the following assumption:

Assumption 1. *For any* $\mathbf{u} \in \mathcal{U}$ *and* $\theta \in \Omega$, *the equation* $\tau(\mathbf{u}, \theta) = t$ *can be uniquely solved for* θ *by* $\theta = \hat{\theta}(\mathbf{u}, t)$.

In order to derive the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$, we will consider conditional expectations of a function ϕ. Under Assumption 1 we have

$$\begin{aligned} \mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] &= \mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))|\tau(\mathbf{U}, \Theta) = \mathbf{t}] \\ &= \mathrm{E}[\phi(\chi(\mathbf{U}, \hat{\theta}(\mathbf{U}, \mathbf{t})))|\tau(\mathbf{U}, \Theta) = \mathbf{t}], \qquad (3) \end{aligned}$$

where we used the substitution principle (Bahadur et al., 1968) noting that $\tau(\mathbf{U}, \Theta) = \mathbf{t} \Leftrightarrow \Theta = \hat{\theta}(\mathbf{U}, \mathbf{t})$. In order to calculate (3) we will hence need the conditional distribution of $\mathbf{U}$ given $\tau(\mathbf{U}, \Theta) = \mathbf{t}$. This distribution is obtained from a standard transformation from $(\mathbf{U}, \Theta)$ to $(\mathbf{U}, \tau(\mathbf{U}, \Theta))$, which gives the joint density $h(\mathbf{u}, \mathbf{t})$ of $(\mathbf{U}, \tau(\mathbf{U}, \Theta)$ as

$$h(\mathbf{u}, \mathbf{t}) = f(\mathbf{u}|\hat{\theta}(\mathbf{u}, \mathbf{t}))w(\mathbf{t}, \mathbf{u}), \qquad (4)$$

where $\mathbf{t} \mapsto w(\mathbf{t}, \mathbf{u})$ is the density of $\tau(\mathbf{u}, \Theta)$ for fixed $\mathbf{u}$. This density is given by

$$w(\mathbf{t}, \mathbf{u}) = \pi(\hat{\theta}(\mathbf{u}, \mathbf{t})) \left|\det \partial_t \hat{\theta}(\mathbf{u}, \mathbf{t})\right| = \left|\frac{\pi(\theta)}{\det \partial_\theta \tau(\mathbf{u}, \theta)}\right|_{\theta = \hat{\theta}(\mathbf{u}, \mathbf{t})}, \qquad (5)$$

where $\pi(\theta)$ is the density of Θ. From this we get the conditional distribution of $\mathbf{U}$ given $\tau(\mathbf{U}, \Theta) = \mathbf{t}$ as $h(\mathbf{u}|\mathbf{t}) = h(\mathbf{u}, \mathbf{t}) / \int h(\mathbf{u}, \mathbf{t}) d\mathbf{u}$, and we are then in a position to complete the calculation of (3):

$$\begin{aligned} \mathrm{E}[\phi(\mathbf{X})|\mathbf{T} = \mathbf{t}] &= \mathrm{E}[\phi(\chi(\mathbf{U}, \hat{\theta}(\mathbf{U}, \mathbf{t})))|\tau(\mathbf{U}, \Theta) = t] \\ &= \int \phi(\chi(\mathbf{u}, \hat{\theta}(\mathbf{u}, \mathbf{t}))h(\mathbf{u}|\mathbf{t}) d\mathbf{u} \\ &= \frac{\int \phi(\chi(\mathbf{u}, \hat{\theta}(\mathbf{u}, \mathbf{t})))h(\mathbf{u}, \mathbf{t}) d\mathbf{u}}{\int h(\mathbf{u}, \mathbf{t}) d\mathbf{u}}. \qquad (6) \end{aligned}$$

2.1 Methods of computation and simulation from the conditional distribution

The integrals in (6) will usually have an intractable form. The calculation of (6) or simulation of samples from $h(\mathbf{u}|\mathbf{t})$, may hence be done by suitable numerical techniques. Some approaches are briefly considered below.

2.1.1 Importance sampling

Importance sampling appears to be the traditional method used in conditional Monte Carlo, see for example Dubi and Horowitz (1979). Consider the computation of (6). If $\mathbf{U}$ is sampled from a density $g(\mathbf{u})$, then (6) can be written

$$\mathrm{E}[\phi(\mathbf{X})|\mathbf{T}=\mathbf{t}] = \frac{\mathrm{E}[\phi(\chi(\mathbf{U},\hat{\theta}(\mathbf{U},\mathbf{t})))h(\mathbf{U},\mathbf{t})/g(\mathbf{U})]}{\mathrm{E}[h(\mathbf{U},\mathbf{t})/g(\mathbf{U})]},$$

which in principle is straightforward to calculate by Monte Carlo simulation.

2.1.2 Rejection sampling

In order to obtain samples from the conditional distribution of $\mathbf{X}$ given $\mathbf{T}=\mathbf{t}$, we need to first sample $\mathbf{U}=\mathbf{u}$ from a density proportional to $h(\mathbf{u},\mathbf{t})$, then solve the equation $\tau(\mathbf{u},\theta)=\mathbf{t}$, and finally return the conditional sample $\hat{\mathbf{x}}=\chi(\mathbf{u},\hat{\theta}(\mathbf{u},t))$. Let $\tilde{h}(\mathbf{u},\mathbf{t})$ be proportional to $h(\mathbf{u},\mathbf{t})$ as a function of $\mathbf{u}$. In rejection sampling (Ripley, 1987, p. 60) one samples from a density $g(\mathbf{u})$ with support which includes the support of $\tilde{h}(\mathbf{u},\mathbf{t})$ and for which we can find a bound $M<\infty$ such that $\tilde{h}/g \le M$. One then samples $\mathbf{u}$ from g and a uniform random number $z \in [0,1]$ until $Mz \le \tilde{h}(\mathbf{u})/g(\mathbf{u})$.

2.1.3 Markov Chain Monte Carlo

A disadvantage of rejection sampling is the need for the bound M which may be difficult to obtain. The Metropolis-Hastings algorithm (Hastings, 1970) needs no such bound but, on the other hand, produces dependent samples. We describe below an approach where proposals of the Metropolis-Hastings algorithm are independent samples $\mathbf{u}$ from a density $g(\mathbf{u})$, where g, as for the rejection sampling method, needs to have a support which includes the support of $\tilde{h}(\mathbf{u}, \mathbf{t})$.

To initialize the algorithm one needs an initial sample $\mathbf{u}^0$ with $\tilde{h}(\mathbf{u}^0, \mathbf{t}) > 0$. Then for each iteration k, one generates (i) a proposal $\mathbf{u}^{'}$ from $g(\cdot)$; (ii) a uniform random number $z \in [0, 1]$. One then accepts the proposal and let $\mathbf{u}^{k+1} = \mathbf{u}^{'}$ if

$$z \leq \frac{\tilde{h}(\mathbf{u}^{'}, \mathbf{t})}{\tilde{h}(\mathbf{u}^k, \mathbf{t})} \cdot \frac{g(\mathbf{u}^k)}{g(\mathbf{u}')}, \tag{7}$$

and otherwise lets $\mathbf{u}^{k+1} = \mathbf{u}^k$. It should be noted that for each new proposal $\mathbf{u}^{'}$ one needs to solve the equations leading to $\hat{\theta}(\mathbf{u}^{'}, \mathbf{t})$. As for rejection sampling, one obtains the desired samples $\hat{\mathbf{x}}^k = \chi(\mathbf{u}^k, \hat{\theta}(\mathbf{u}^k, \mathbf{t}))$.

2.1.4 The naive sampler

In order to check algorithms for conditional sampling, a type of benchmark might be to use a naive sampler as follows. Then $\mathbf{x}$ are sampled from $f_{\mathbf{X}}(\mathbf{x})$ and accepted if and only if $|T(\mathbf{x}) - \mathbf{t}| < \epsilon$ for an apriori chosen (small) $\epsilon > 0$ and an appropriate norm $|\cdot|$. The successive accepted samples $\hat{\mathbf{x}}$ are approximate samples from the desired conditional distribution, see Section 4.2 for examples.

3 Application of the method

As might be clair from the previous section, the choice of the function $\chi(\mathbf{u}, \theta)$ and the marginal distribution for Θ are of crucial importance for the construction of an efficient algorithm.

3.1 The choice of the function $\chi(\mathbf{u}, \theta)$

The choice of $\chi(\mathbf{u}, \theta)$ will obviously depend very much on the application, and we refer to the examples in order to give some advice here. The uniqueness requirement of Assumption 1 of course restricts considerably the choice. An important issue is the requirement that the range of $\chi(\mathbf{u}, \theta)$, for each θ, should include the support of $\mathbf{X}$. A further discussion on the form of the function $\chi(\mathbf{u}, \theta)$ is found in the concluding remarks of Section 7. In particular is considered a possible relaxation of the uniqueness requirement of Assumption 1.

3.2 The choice of distribution of Θ

In Bayesian statistics it is well recognized that prior distributions for parameters may be chosen as improper distributions. Also, in the approach on conditional sampling given sufficient statistics in Lindqvist and Taraldsen (2005) it was argued that a random vector similar to our Θ may sometimes preferably be given an improper distribution. The following argument shows, however, that in the present approach, Θ must be given a proper distribution (i.e., having an integrable density function $\pi(\theta)$).

Suppose namely that Θ is given an improper distribution. In order to condition on $\tau(\mathbf{U}, \Theta)$ it is necessary that its density is finite. (In Bayesian analysis, this is the marginal density of the data which appears in the denominator of Bayes' formula.) This property implies that there is a set A such that $P(\tau(\mathbf{U}, \Theta) \in A) < \infty$, where this set clearly may be chosen so that $P(T(\mathbf{X}) \in A) > 0$. Now for this set A,

$$\begin{aligned} P(\tau(\mathbf{U}, \Theta) \in A) &= \int_{\Omega} P(\tau(\mathbf{U}, \Theta) \in A \mid \Theta = \theta)\pi(\theta)d\theta \\ &= \int_{\Omega} P(\tau(\mathbf{U}, \theta) \in A \mid \Theta = \theta)\pi(\theta)d\theta \\ &= P(T(\mathbf{X}) \in A) \int_{\Omega} \pi(\theta)d\theta, \end{aligned}$$

where the last equality follows from the basic property (1). This clearly implies that $\int_{\Omega} \pi(\theta)d\theta < \infty$.

In the following we shall therefore always assume that Θ is given an integrable density $\pi(\theta)$. (This density may of course be normalized to have integral one, but as we shall see in our examples, the normalizing constant is usually of no concern). Particular choices will depend on the application, but also on certain structural issues of the problem as explained in the next subsection.

3.3 The "pivotal" condition

In some cases it is possible to choose the function $\chi(\mathbf{u}, \theta)$ in such a way that $\tau(\mathbf{u}, \theta)$ depends on $\mathbf{u}$ only through a lower dimensionable function $r(\mathbf{u})$, where the value of $r(\mathbf{u})$ can be uniquely recovered from the equation $\tau(\mathbf{u}, \theta) = \mathbf{t}$ for given θ and $\mathbf{t}$. This means that there is a function $\tilde{\tau}$ such that $\tau(\mathbf{u}, \theta) = \tilde{\tau}(r(\mathbf{u}), \theta)$ for all $(\mathbf{u}, \theta)$, and a function $\tilde{v}$ such that $\tilde{\tau}(r(\mathbf{u}), \theta) = \mathbf{t}$ implies $r(\mathbf{u}) = \tilde{v}(\theta, \mathbf{t})$. The notion of "pivotal" for the present case is borrowed from Lindqvist and Taraldsen (2005), who considered a similar condition in which case $\tilde{v}(\theta, \mathbf{T})$ is a pivotal quantity in the classical meaning of the notion. Although the setting here is different, we shall keep calling this the *pivotal* condition.

Under Assumption 1, the following equivalences hold under the pivotal condition:

$$\hat{\theta}(\mathbf{u}, \mathbf{t}) = \theta \;\Leftrightarrow\; \tau(\mathbf{u}, \theta) = \mathbf{t} \;\Leftrightarrow\; \tilde{\tau}(r(\mathbf{u}), \theta) = \mathbf{t} \;\Leftrightarrow\; r(\mathbf{u}) = \tilde{v}(\theta, \mathbf{t})$$

We hence have the identity

$$r(\mathbf{u}) = \tilde{v}(\hat{\theta}(\mathbf{u}, \mathbf{t}), \mathbf{t}) \text{ for all } \mathbf{u}, \mathbf{t}$$

so that

$$\tau(\mathbf{u}, \theta) = \tilde{\tau}(\tilde{v}(\hat{\theta}(\mathbf{u}, \mathbf{t}), \mathbf{t}), \theta)$$

and hence

$$\det \partial_\theta \tau(\mathbf{u}, \theta) = \det \partial_\theta \tilde{\tau}(\tilde{v}(\hat{\theta}(\mathbf{u}, \mathbf{t}), \mathbf{t}), \theta).$$

Substituting $\hat{\theta}(\mathbf{u}, \mathbf{t})$ for θ it is therefore seen that

$$|\det \partial_\theta \tau(\mathbf{u}, \theta)|_{\theta = \hat{\theta}(\mathbf{u}, \mathbf{t})} = J(\hat{\theta}(\mathbf{u}, \mathbf{t}), \mathbf{t}) \tag{8}$$

where

$$J(\theta, \mathbf{t}) = |\det \partial_\theta \tilde{\tau}(\mathbf{v}, \theta)|_{\mathbf{v} = \tilde{v}(\theta, \mathbf{t})}.$$

Consider first the case where $J(\hat{\theta}(\mathbf{u}, \mathbf{t}), \mathbf{t})$ factors as $K(\hat{\theta}(\mathbf{u}, \mathbf{t}))a(\mathbf{t})$. Suppose also that $f(\mathbf{u}|\theta)$ factors as

$$f(\mathbf{u}|\theta) = \rho(\theta)\tilde{f}(\mathbf{u}|\theta).$$

Then (4) and (5) suggest the choice of $\pi(\theta)$ proportional to $\rho(\theta)^{-1}K(\theta)$, which simplifies the expression for $h(\mathbf{u}, \mathbf{t})$ in (4). In order to ensure that $\pi(\theta)$ has a finite integral, we might in addition restrict the support of π to some bounded set, letting for example

$$\pi(\theta) = \rho(\theta)^{-1}K(\theta)I(\theta \in A),$$

where $I(\cdot)$ is the indicator function of the condition in the parantheses, and A is such that $\int_A \rho(\theta)^{-1}K(\theta)d\theta < \infty$. It follows in this case that

$$h(\mathbf{u}, \mathbf{t}) \propto \tilde{f}(\mathbf{u}|\hat{\theta}(\mathbf{u}, \mathbf{t}))I(\hat{\theta}(\mathbf{u}, \mathbf{t}) \in A). \tag{9}$$

For the general pivotal case, leading to (8), we would have to choose a $\pi(\cdot)$ that depends on $\mathbf{t}$, by replacing $K(\theta)$ by $J(\theta, \mathbf{t})$ in the above. Since $\mathbf{t}$ is fixed when conditioning on $\mathbf{T} = \mathbf{t}$, it is seen that the crucial arguments will go through also in this case, thus still leading to (9). A similar argument was used in Lindqvist and Taraldsen (2005).

As a further refinement, it may happen that $r(\mathbf{U})$ is a sufficient statistic in the model defined by $f(\mathbf{u}|\theta)$. Then by Neyman's factorization criterion (Casella and Berger, 2002, Ch. 6), we can write

$$f(\mathbf{u}|\theta) = p(r(\mathbf{u})|\theta)q(\mathbf{u})$$

for appropriate functions p and q. Hence we can write

$$f(\mathbf{u}|\hat{\theta}(\mathbf{u}, \mathbf{t})) = p(\tilde{v}(\hat{\theta}(\mathbf{u}, \mathbf{t}), \mathbf{t})|\hat{\theta}(\mathbf{u}, \mathbf{t}))q(\mathbf{u})$$

By assimilating the $p(\cdot)$-part of the above into $\pi(\hat{\theta}(\mathbf{u}, \mathbf{t}))$ (where $\pi(\cdot)$ will now possibly depend on $\mathbf{t}$) we get

$$h(\mathbf{u}, \mathbf{t}) \propto q(\mathbf{u})I(\hat{\theta}(\mathbf{u}, \mathbf{t}) \in A).$$

4 Examples

4.1 Two examples involving the pivotal condition

4.1.1 Conditional sampling of uniforms

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be an i.i.d sample from $U[0, 1]$, the uniform distribution on $[0, 1]$, and let $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$. Suppose one wants to sample from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$ where $0 < t < n$. There appears to be no simple expression for this conditional distribution. Lindqvist and Taraldsen (2005) considered an approach where the uniform distribution is embedded in a parametric family involving truncated exponential distributions and utilzed the sufficiency of $T(\mathbf{X})$ in this model. The resulting method is, however, surprisingly complicated in absence of the pivotal condition of Lindqvist and Taraldsen (2005). A Gibbs sampling method was devised by Lindqvist and Rannestad (2011), apparently being much quicker than the former method, and much easier to implement.

We now present a simple solution to the problem using the approach of the previous sections and utilizing the presence of a pivotal condition as studied in Section 3.3. An advantage as compared to the Gibbs sampling algorithm is that the present method produces independent samples.

Let $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ be an i.i.d. sample from $U[0, \theta]$, where $\theta \in (0, 1]$. Then the U_i/θ are i.i.d. from $U[0, 1]$, so condition (1) in Section 2 is satisfied with

$$\chi(\mathbf{u}, \theta) = \left(\frac{u_1}{\theta}, \frac{u_2}{\theta}, \ldots, \frac{u_n}{\theta}\right). \tag{10}$$

defined for $\mathbf{u} \in [0, 1]^n$ and $\theta \in (0, 1]$.

The above is moreover in accordance with Proposition 1, which readily gives

$$f(\mathbf{u} \mid \theta) = \frac{1}{\theta^n} I(\max_i u_i \le \theta),$$

where we used that $f_{\mathbf{X}}(\mathbf{x}) = I(\max x_i \le 1)$. Note that here and below we tacitly assume that we are working with nonnegative variables only.

Now we have

$$\tau(\mathbf{u}, \theta) = \frac{\sum_{i=1}^n u_i}{\theta},$$

and hence there exists a unique solution for θ of the equation $\tau(\mathbf{u}, \theta) = t$, given by

$$\hat{\theta}(\mathbf{u}, t) = \frac{\sum_{i=1}^n u_i}{t}. \tag{11}$$

Clearly, the pivotal condition of Section 3.3 is satisfied with $r(\mathbf{u}) = \sum_{i=1}^n u_i$. It is seen, however, that $r(\mathbf{u})$ does not correspond to a sufficient statistic for the model $f(\mathbf{u}|\theta)$. We should therefore stick to (9), which by choosing

$A = [0,1]$ gives

$$\begin{aligned} h(\mathbf{u}, t) &\propto I(\max_i u_i \leq \hat{\theta}(\mathbf{u}, \mathbf{t})) \cdot I(\hat{\theta}(\mathbf{u}, \mathbf{t}) \leq 1) \\ &= I\left(\max u_i \leq \frac{\sum_{i=1}^n u_i}{t}\right) \cdot I\left(\frac{\sum_{i=1}^n u_i}{t} \leq 1\right) \\ &= I\left(t \cdot \max_i u_i \leq \sum_{i=1}^n u_i \leq t\right). \qquad (12) \end{aligned}$$

Thus the conditional distribution $h(\mathbf{u}|t)$ is uniform on the set of $\mathbf{u} \in [0,1]^n$ satisfying the restriction given by the indicator function in (12). We may hence sample the u_i independently from $U[0,1]$ and accept the sample if and only if the restriction is satisfied. (Note that if $t \leq 1$, then the left inequality in (12) is always satisfied.) Finally, for the accepted samples we conclude from (10) and (11) that the resulting conditional sample is

$$\hat{\mathbf{x}} = \left(t \frac{u_1}{\sum_{i=1}^n u_i}, t \frac{u_2}{\sum_{i=1}^n u_i}, \ldots, t \frac{u_n}{\sum_{i=1}^n u_i}\right). \qquad (13)$$

Figure 1 shows the result of a simulation with $n = 2$ and $t = 0.3$. It is easy to show by direct calculation that the conditional distribution of X_1 (and hence also of X_2) given $X_1 + X_2 = 0.3$ is uniform on $[0, 0.3]$. The left panel of the figure shows the empirical cumulative distribution of X_1 (and X_2) resulting from (13), which is clearly uniform on $[0, 0.3]$ as expected. The right panel of the figure shows, on the other hand, the empirical distribution obtained from (13) when sampling (u_1, u_2) i.i.d. from $U[0,1]$ without ignoring the pairs (u_1, u_2) with $u_1 + u_2 > 0.3$ (which is required by (12)). The discrepancy from a straight line shows that condition (12) is necessary here. Still the algorithm is very simple, and simpler than the corresponding algorithms of Lindqvist and Taraldsen (2005) and Lindqvist and Rannestad (2011) that were mentioned above.

The algorithm may be slow if t is close to 0 or n. In these cases it might be better to use importance sampling by drawing the u_i from a density $g(u) = cu^{c-1}$ for $c > 0$, where c is small (large) if t is close to 0 (close to n). But note that we will then need to sample from a non-uniform density $h(\mathbf{u}|t)$.

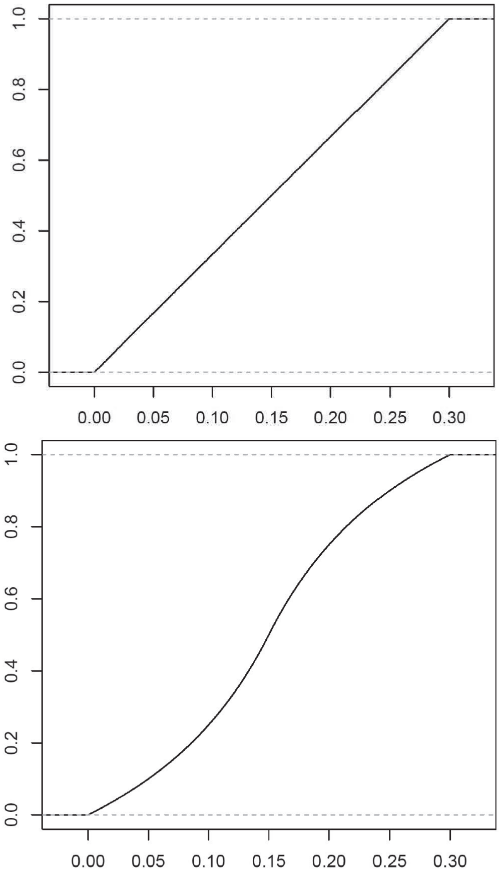

Figure 1: Empirical distribution functions for marginal distributions of conditional samples of uniforms when $n = 2$. Left: Sampling $(u_1, u_2) \sim U[0, 1]$ and using (12) and (13). Right: Sampling $(u_1, u_2) \sim U[0, 1]$ and using (13) only.

As a final remark on this example, suppose instead that we wanted to condition on $\sum_{i=1}^{n} X_i^r = t$ for some given $r > 0$. It is then straightforward to check that only a minor modification of the above derivation is needed. As a result, one should still sample u_i from $U[0,1]$, but change condition (12) into

$$I\left(t \cdot \max_i u_i^r \leq \sum_{i=1}^{n} u_i^r \leq t\right)$$

and use the samples $\hat{\mathbf{x}}$ where

$$\hat{x}_i = t^{1/r}\,\frac{u_i}{(\sum_{\ell=1}^{n} u_\ell^r)^{1/r}}.$$

4.1.2 Conditional sampling of normals

The following is a classical example in conditional Monte Carlo, see e.g. Trotter and Tukey (1956), Hammersley (1956), Granovsky (1981), Ripley (1987). Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be i.i.d from $N(0,1)$ and let $T(\mathbf{X}) = \max_i X_i - \min_i X_i$. We wish to sample from the conditional distribution of $\mathbf{X}$, given $T(\mathbf{X}) = t$ for $t > 0$.

Now let $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ be an i.i.d. sample from $N(0, \theta^2)$. Then condition (1) in Section 2 is clearly satisfied when $\chi(\mathbf{u}, \theta)$ is given by the scale transformation (10) for $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n$ and $\theta \in (0,1]$. It is furthermore seen that the pivotal condition of Section 3.3 is satisfied with $r(\mathbf{u}) = \max_i u_i - \min_i u_i$, and in a similar way as for the uniform distribution case treated above, we arrive at

$$h(\mathbf{u}, t) \propto \exp\left(-\frac{t^2}{2(\max_i u_i - \min_i u_i)^2}\sum_{i=1}^{n} u_i^2\right) I(\max_i u_i - \min_i u_i < t), \quad (14)$$

by letting $\pi(\cdot)$ be supported on the interval $[0,1]$. Actually, we have used $\pi(\theta) = \theta^{n-1} I(0 < \theta \leq 1)$.

Noting that the right hand side of (14) is less than or equal to

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n} u_i^2\right) I(\max_i u_i - \min_i u_i < t)$$

we can use rejection sampling (Section 2.1.2) based on sampling of i.i.d. standard normal variates. If t is small, then in order to increase the acceptance probability of the rejection sampling, it might be beneficial to use as the proposal distribution, a mixture of a standard normal and a normal distribution with small variance.

The resulting conditional samples are now of the form

$$\hat{\mathbf{x}} = \left(t \, \frac{u_1}{\max_i u_i - \min_i u_i}, \ldots, t \, \frac{u_n}{\max_i u_i - \min_i u_i} \right).$$

4.2 Conditional sampling from two-parameter exponential families

Suppose $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is distributed as an i.i.d. sample from a two-parameter exponential family of *positive* random variables, with minimal sufficient statistic

$$\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = \left(\sum_{i=1}^{n} g_1(X_i), \sum_{i=1}^{n} g_2(X_i) \right). \tag{15}$$

Suppose now that $\mathbf{t} = (t_1, t_2)$ is the observed value of $\mathbf{T}(\mathbf{X})$, and that we want to sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ conditionally on $\mathbf{T}(\mathbf{X}) = \mathbf{t}$. By sufficiency, samples from the conditional distribution of $\mathbf{X}$ given $\mathbf{T}(\mathbf{X}) = \mathbf{t}$ can be obtained by choosing any density from the given family as the basic density. Let $f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i)$ be the chosen density.

Let

$$\chi(\mathbf{u}, \theta) = \left(\left(\frac{u_1}{\beta}\right)^{\alpha}, \left(\frac{u_2}{\beta}\right)^{\alpha}, \ldots, \left(\frac{u_n}{\beta}\right)^{\alpha} \right), \tag{16}$$

where $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ is a vector of positive numbers and $\theta = (\alpha, \beta)$ is a pair of positve parameters α, β. Then, using Proposition 1, condition (1) of Section 2 is satisfied if $\mathbf{U}$ for given θ has density

$$f(\mathbf{u} \mid \theta) = \prod_{i=1}^{n} \frac{\alpha}{\beta} \left(\frac{u_i}{\beta}\right)^{\alpha-1} f_X\left(\left(\frac{u_i}{\beta}\right)^{\alpha} \right). \tag{17}$$

Assumption 1 requires that there is a unique solution for θ of the equation

$$\tau(\mathbf{u}, \theta) = \mathbf{t},$$

which here means

$$\sum_{i=1}^{n} g_1\left(\left(\frac{u_i}{\beta}\right)^{\alpha} \right) = t_1,$$

$$\sum_{i=1}^{n} g_2\left(\left(\frac{u_i}{\beta}\right)^{\alpha}\right) = t_2.$$

Assume that there is a unique solution $\hat{\theta}(\mathbf{u},\mathbf{t}) = (\hat{\alpha}(\mathbf{u},\mathbf{t}),\hat{\beta}(\mathbf{u},\mathbf{t}))$ of these equations.

Letting $\pi(\theta) \equiv \pi(\alpha,\beta)$ be the density of Θ, the density $h(\mathbf{u},\mathbf{t})$ is found from (4) and (5), giving

$$\begin{aligned} h(\mathbf{u},\mathbf{t}) &= f(\mathbf{u} \mid \hat{\theta}(\mathbf{u},\mathbf{t}))w(\mathbf{t},\mathbf{u}) \\ &= \frac{(\hat{\alpha}/\hat{\beta})^n \left(\prod_{i=1}^n \hat{x}_i\right)^{1-1/\hat{\alpha}} \left(\prod_{i=1}^n f_X(\hat{x}_i)\right) \pi(\hat{\alpha},\hat{\beta})}{|\det \partial_\theta \tau(\mathbf{u},\theta)|_{\theta=\hat{\theta}(\mathbf{u},\mathbf{t})}}. \end{aligned} \tag{18}$$

where

$$\hat{x}_i = \left(\frac{u_i}{\hat{\beta}}\right)^{\hat{\alpha}} \tag{19}$$

and

$$\begin{aligned} \det \partial_\theta \tau(\mathbf{u},\theta)|_{\theta=\hat{\theta}(\mathbf{u},\mathbf{t})} &= \frac{1}{\hat{\beta}(\mathbf{u},\mathbf{t})}\left[\left(\sum_{i=1}^n g_1'(\hat{x}_i)\hat{x}_i\right)\left(\sum_{i=1}^n g_2'(\hat{x}_i)\hat{x}_i \log(\hat{x}_i)\right)\right. \\ &- \left.\left(\sum_{i=1}^n g_2'(\hat{x}_i)\hat{x}_i\right)\left(\sum_{i=1}^n g_1'(\hat{x}_i)\hat{x}_i \log(\hat{x}_i)\right)\right]. \end{aligned}$$

When sampling from (18) by the Metropolis-Hastings algorithm (Section 2.1.3) it seems to be a good idea to let the proposal distribution $g(\mathbf{u})$ be the distribution of the original exponential familiy with parameter values equal to the maximum likelihood estimates based on the observation $\mathbf{t}$. Then the calculated $\hat{\alpha},\hat{\beta}$ are expected to be around 1, and we therefore suggest to choose the distribution of Θ as $\pi(\alpha,\beta) = I(a_1 \le \alpha \le a_2, b_1 \le \alpha \le b_2)$ for suitably chosen $0 < a_1 < 1 < a_2$, $0 < b_1 < 1 < b_2$, see examples in Section 5.

In a practical application one would usually also have the original data $\mathbf{x} = (x_1,\ldots,x_n)$ which led to the values $t_1 = T_1(\mathbf{x}), t_2 = T_2(\mathbf{x})$. The vector $\mathbf{x}$ may then be used as the initial sample of the Metropolis-Hastings simulation, and will give $\hat{\alpha} = \hat{\beta} = 1$. In this case, the successively simulated accepted conditional samples $\hat{\mathbf{x}} = (\hat{x}_1,\ldots,\hat{x}_n)$ defined by (19) will have the correct distribution, so there is no need for a burn-in period in the Metropolis-Hastings simulations.

4.2.1 Gamma Distribution

The gamma-distribution with shape parameter $k > 0$ and scale parameter $\theta > 0$ has density

$$f(x; k, \theta) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta} \text{ for } x > 0. \tag{20}$$

We suggest using $k = \theta = 1$ to get $f_X(x) = e^{-x}$. Referring to (15), we have for the gamma model, $g_1(x) = x,\ g_2(x) = \log x$, and hence we need to solve the equations

$$\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} = t_1, \tag{21}$$

$$\sum_{i=1}^{n} \log \left(\frac{u_i}{\beta}\right)^{\alpha} = t_2. \tag{22}$$

It is shown in Lemma 1 in Appendix that there is a unique solution $(\hat{\alpha}, \hat{\theta})$ for (α, θ). The actual solution turns out to be easily obtained via a single equation involving α. Now (18) gives

$$h(\mathbf{u}, \mathbf{t}) = \frac{(\hat{\alpha}/\hat{\beta})^n e^{(1-1/\hat{\alpha})t_2} e^{-t_1} \pi(\hat{\alpha}, \hat{\beta})}{(1/\hat{\beta})\left(t_1 t_2 - n \sum_{i=1}^{n} \hat{x}_i \log \hat{x}_i\right)},$$

which is the basis for simulation of conditional samples as outlined above. It is easy to see, however, that the pivotal condition of Section 3.3 is not satisfied here.

4.2.2 Inverse Gaussian Distribution

The Inverse Gaussian distribution has density which can be written as

$$f(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda}{2x} - \frac{\lambda x}{2\mu^2} + \frac{\lambda}{\mu}\right), \quad x > 0. \tag{23}$$

Let now $f_X(x)$ be the density obtained when $\mu = \lambda = 1$, i.e.

$$f_X(x) = \sqrt{\frac{1}{2\pi x^3}} \exp\left(-\frac{1}{2x} - \frac{x}{2} + 1\right), \quad x > 0.$$

Furthermore, for the inverse Gaussian distributions we can choose $g_1(x) = x,\ g_2(x) = 1/x$ (Seshadri, 2012, p. 7), and hence we obtain the equations

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{u_i}{\beta}\right)^{\alpha} &= t_1,\\ \sum_{i=1}^{n}\left(\frac{u_i}{\beta}\right)^{-\alpha} &= t_2.\end{aligned}$$

As for the gamma case, there is a unique solution $(\hat{\alpha}, \hat{\beta})$ for (α, β), see Lemma 2 in the Appendix. Now we get from (18),

$$h(\mathbf{u}, \mathbf{t}) = \frac{(\hat{\alpha}/\hat{\beta})^n \left(\prod_{i=1}^{n} \hat{x}_i\right)^{-1/2-1/\hat{\alpha}} e^{-(1/2)(t_1+t_2)+n} \pi(\hat{\alpha}, \hat{\beta})}{(1/\hat{\beta})\left(t_2 \sum_{i=1}^{n} \hat{x}_i \log \hat{x}_i - t_1 \sum_{i=1}^{n} \log \hat{x}_i/\hat{x}_i\right)}.$$

It was suggested above to use the parametric model itself as a proposal distribution in Metropolis-Hastings simulations, with parameters given by the maximum likelihood estimates from the original data. Following (Seshadri, 2012, p. 7), the maximum likelihood estimates of the parameters in (23) are given from

$$\hat{\mu} = \bar{x}, \quad \hat{\lambda}^{-1} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right).$$

Note also that Michael et al. (1976) presented a nice method of simulating from the inverse Gaussian distribution.

5 A simulation study

A simulation study was performed in order to illustrate the algorithms of Section 4.2 for the gamma and inverse Gaussian distributions, respectively. The setup of the study is summarized in Table 1.

For example, in case 1, a sample $\mathbf{x}$ with $n = 3$ was drawn from a gamma distribution, giving the observed sufficient statistic $(t_1, t_2) = (4.86, 1.02)$. Conditional samples were then simulated using the Metropolis-Hastings algorithm in the way described in Section 4.2. More precisely, the proposal distribution was chosen to be the gamma density (20) using the maximum likelihood estimates $\hat{k} = 3.66,\ \hat{\theta} = 0.44$ as parameters. The density $\pi(\alpha, \beta)$

was chosen to be uniform over $(\alpha, \beta) \in [0.5, 1.5] \times [0.5, 1.5]$. In addition, we applied the naive sampling method described in Section 2.1.4. Values ϵ_1, ϵ_2 (see Table 1) were chosen so that the sampler accepts an i.i.d. sample $\mathbf{x}' = (x_1', x_2', \ldots, x_n')$ from the proposal distribution if and only if

$$|T_1(\mathbf{x}') - t_1| \leq \epsilon_1 \text{ and } |T_2(\mathbf{x}') - t_2| \leq \epsilon_2.$$

In case 1 were used $\epsilon_1 = \epsilon_2 = 10^{-2}$. Both the Metropolis-Hastings algorithm and the naive sampler were ran for enough iterations to produce at least 10^4 samples.

The description is similar for cases 2-4. Figure 2 shows, for each of the four cases in Table 1, the simulated cumulative distribution functions for the sampled $\hat{x}_1$. The closeness of the curves corresponding to the two methods is remarkable. Considering the naive sampler as a "benchmark", although only approximately correct, this closeness can be taken as a confirmation that the algorithms derived in the paper are producing samples from the correct conditional distributions.

Case	t_1, t_2	n	Distribution	Sample sizes	π	ϵ_1, ϵ_2
1	$4.86, 1.02$	3	Gamma	10^4	$I_{[0.5,1.5]^2}$	$10^{-2}, 10^{-2}$
2	$16.49, 2.85$	10	Gamma	10^4	$I_{[0.5,1.5]^2}$	$10^{-1}, 10^{-1}$
3	$3.67, 6.01$	3	Inverse Gaussian	10^4	$I_{[0.5,1.5]^2}$	$10^{-1}, 10^{-1}$
4	$936.36, 0.59$	10	Inverse Gaussian	10^4	$I_{[0.5,1.5]^2}$	$10^{-1}, 10^{-1}$

Table 1: Values used for generating examples.

6 Application to goodness-of-fit testing

As noted in the introduction, a typical use of conditional samples given sufficient statistics is in goodness-of-fit testing.

Consider the null hypothesis H_0 that an observation vector $\mathbf{X}$ comes from a particular distribution indexed by an unknown parameter θ and such that $\mathbf{T} = T(\mathbf{X})$ is sufficient for θ. For a test statistic $W(\mathbf{X})$ we define the conditional p-value by

$$p_{obs}^W = P_{H_0}(W(\mathbf{X}) \geq w_* | \mathbf{T} = \mathbf{t})$$

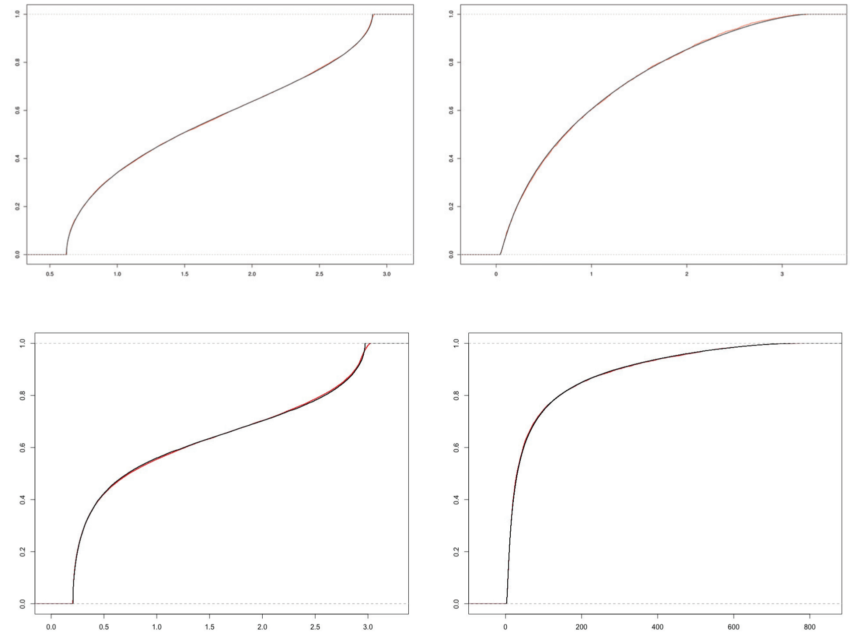

Figure 2: Simulated margi6al cumulative distribution functions for the sampled $\hat{x}_1$ from the conditional samples for the cases of Table 1. Using the Metropolis-Hastings algorithms of Section 4.2 (black); using the naive sampler of Section 2.1.4 (red). Case 1: upper left. Case 2: upper right. Case 3: lower left. Case 4: lower right.

where w_* is the observed value of the test statistic and $\mathbf{t}$ is the observed value of the sufficient statistic. A conditional goodness-of-fit test based on W rejects H_0 at significance level α if $p_{obs}^W \leq \alpha$. Let now $\hat{\mathbf{x}}_j$ for $j = 1, 2, \ldots, k$ be samples from the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}$. Then the observed p-values are approximated by

$$p_{obs}^W \approx \frac{1}{k} \sum_{j=1}^{k} I(W(\mathbf{x}_j) \geq w_*). \tag{24}$$

Consider now data from Best et al. (2012), giving the precipitation from storms in inches at the Jug bridge in Maryland, USA. The observed data are

1.01, 1.11, 1.13, 1.15, 1.16, 1.17, 1.2, 1.52, 1.54, 1.54, 1.57, 1.64,
1.73, 1.79, 2.09, 2.09, 2.57, 2.75, 2.93, 3.19, 3.54, 3.57, 5.11, 5.62

comprising the data vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where $n = 24$. The question is whether the gamma or inverse Gaussian distributions fit the data. Using the setup and notation from Section 4.2 we calculate the sufficient statistics as

$$t_1 = \sum_{i=1}^{n} x_i = 52.72, \quad t_2 = \sum_{i=1}^{n} \log x_i = 15.7815$$

for the gamma distribution and

$$t_1 = \sum_{i=1}^{n} x_i = 52.72, \quad t_2 = \sum_{i=1}^{n} \frac{1}{x_i} = 13.8363$$

for the inverse Gaussian distribution.

Some common test statistics for goodness-of-fit testing are constructed as follows. Let $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ be the order statistic of $\mathbf{x}$. Then define the transformed values $z_i = F(x_{(i)}\ ;\ \hat{\theta}_1, \hat{\theta}_2)$, where $F(\cdot; \theta_1, \theta_2)$ is the cumulative distribution function of the gamma or inverse Gaussian distributions with parameters θ_1, θ_2, while $\hat{\theta}_1, \hat{\theta}_2$ are the maximum likelihood estimates which can be found from the corresponding t_1 and t_2.

From this setup we can write down the following test statistics:

Kolmogorov-Smirnov test (Razali et al., 2011)

$$D = \max_{1 \leq i \leq n} \left(z_i - \frac{i-1}{n}, \frac{i}{n} - z_i \right).$$

Test	Inverse Gaussian distribution	Gamma distribution
A^2	0.094	0.024
ω^2	0.102	0.031
D	0.217	0.061

Table 2: Conditional p-values

Anderson-Darling test (Stephens, 1970)

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\ln z_i + \ln(1 - z_{n-i+1})\right).$$

The Cramér-von Mises test (Stephens, 1970)

$$\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(z_i - \frac{2j-1}{2n}\right)^2.$$

Now let A^2_*, ω_*, D_* denote the observed values of the test statistics as calculated from the observed data $\mathbf{x}$. The approximated conditional p-values p^D_{obs}, $p^{A^2}_{obs}$, $p^{\omega^2}_{obs}$ can now be calculated from (24) for each the null hypotheses of gamma distribution and inverse Gaussian distribution, respectively.

We simulated $k = 10^5$ samples from the conditional distributions and obtained the results of Table 2. The calculated conditional p-values indicate that the fit of the inverse Gaussian is marginal, which agrees with the results of Best et al. (2012). Using significance level $\alpha = 0.05$, the tests based on A^2 and ω^2 suggest that the gamma distribution does not fit the data.

7 Concluding remarks

7.1 Classical conditional Monte Carlo

The method of the present paper can be seen as a reformulation of the main ideas of the classical concept of conditional Monte Carlo as introduced in the 1950s. The basic idea of conditional Monte Carlo was essentially the introduction of new coordinates. Trotter and Tukey (1956) made a point of the “skullduggery” related to such arbitrary new variables which had “nothing to do with the way our samples were drawn". This "trick" was, however, the successful ingredient of the method, and is basically also the way our method works. The main new coordinate of our approach is represented by a parameter in an artificial statistical model.

7.2 Comparison to Lindqvist and Taraldsen (2005)

As indicated in the Introduction, there are some basic differences between the method of Lindqvist and Taraldsen (2005) and the present approach. Still, the methods share several important ingredients, and we have therefore found it useful to adopt much of the notation from Lindqvist and Taraldsen (2005) in the present paper. With this, both methods end up with the goal of calculating conditional distributions of certain functions $\chi(\mathbf{U}, \Theta)$ given related functions $\tau(\mathbf{U}, \Theta) = \mathbf{t}$. While the role of Θ is apparently very similar in the two methods, there is indeed a difference. As shown in Section 3.2, Θ can in the present approach basically be given any bounded distribution, but not an improper distribution. This is in contrast to Lindqvist and Taraldsen (2005), where the distribution of Θ plays a role more in line with Bayesian and fiducial statistics. The framework and methods of Lindqvist and Taraldsen (2005), although tailored for the special situation of conditional sampling under sufficiency, in fact also induces interesting algorithms for calculation of Bayesian posterior distributions as well as fiducial sampling.

7.3 The roles of θ and $\chi(\mathbf{u}, \theta)$

As we have seen, the parameter θ will normally have the same dimension as the statistic $T(\mathbf{X})$. This ensures that the number of equations to solve for obtaining the $\hat{\theta}(\mathbf{u}, \mathbf{t})$ is the same as the number of unknowns (see Assumption 1). In the examples of Sections 4.1.1 and 4.1.2 we considered a one-dimensional $T(\mathbf{X})$, using the "scaling" transformation, $\chi(u, \theta) = u/\theta$. In Section 4.2 we conditioned on a two-dimensional statistic and used the transformation $\chi(u, \theta) = (u/\beta)^\alpha$ which is appropriate for positive variables. This transformation would not be appropriate, however, for models where the observations have support in all of $\mathbb{R}$. In this case, the linear transformation $(u - \alpha)/\beta$ could be used instead. This would for example be a suitable transformation if, in the example of Section 4.1.2, we conditioned on the average $\bar{X}$ in addition to the range $\max X_i - \min X_i$.

7.4 Conditioning on $T(\mathbf{X})$ with dimension $k > 2$

Our initial motivation for the paper came from the conditional sampling given sufficient statistics in two-parameter models like gamma and inverse Gaussian distributions. Still a natural question is, of course, what to do if we want to condition on $T(\mathbf{X})$ with dimension $k > 2$. For the i.i.d. case with X_i having support in all of $\mathbb{R}$, an obvious choice might be to let $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{k-1})$ and

$$\chi(u, \boldsymbol{\theta}) = \sum_{j=0}^{k-1} \theta_j u^j. \tag{25}$$

If we put $k = 2$ in (25), then this is in fact equivalent to the transformation $(u - \alpha)/\beta$ as suggested above.

In the i.i.d. case with positive X_i, a general suggestion might be to use

$$\chi(u, \boldsymbol{\theta}) = \exp\left\{\sum_{j=0}^{k-1} \theta_j u^j\right\}. \tag{26}$$

For $k = 2$ this transformation is in fact equivalent to the transformation used for the two-parameter exponential families of positive variables in Section 4.2, since

$$\left(\frac{u}{\beta}\right)^{\alpha} = \exp\{-\alpha \log \beta + \alpha \log u\}.$$

It follows from this that, in the gamma and inverse Gaussian cases treated in Section 4.2, we could as well have used the transformation (26) with $k = 2$, and still obtained unique solutions for $\hat{\boldsymbol{\theta}}$. In general, however, there might be several solutions for $\boldsymbol{\theta}$ in the equations $\tau(\mathbf{u}, \boldsymbol{\theta}) = \mathbf{t}$. There would therefore be a need for the possibility of relaxing Assumption 1 to allow more than one solution of the equation $\tau(\mathbf{u}, \theta) = \mathbf{t}$. We sketch an approach below.

7.5 Multiple solutions of the equation $\tau(\mathbf{u}, \theta) = t$

In general it might be difficult or impossible to find a suitable function $\chi(\mathbf{u}, \theta)$ such that Assumption 1 holds. In practice, there may be a finite number of solutions, where the number may also depend on the values of $(\mathbf{u}, \mathbf{t})$. Define then

$$\Gamma(\mathbf{u}, \mathbf{t}) = \{\hat{\theta} : \tau(\mathbf{u}, \hat{\theta}) = \mathbf{t}\}.$$

An extension of the arguments leading to (4), taking into account the multiplicity of the roots of the equation $\tau(\mathbf{u}, \theta) = \mathbf{t}$, then gives the following expression for the joint density of $(\mathbf{U}, \Theta, \tau(\mathbf{U}, \Theta))$,

$$h(\mathbf{u}, \hat{\theta}, \mathbf{t}) = f(\mathbf{u}|\hat{\theta}) \left| \frac{\pi(\theta)}{\det \partial_\theta \tau(\mathbf{u}, \theta)} \right|_{\theta = \hat{\theta}} \tag{27}$$

for $\mathbf{u}, \mathbf{t}$ as before, and $\hat{\theta} \in \Gamma(\mathbf{u}, \mathbf{t})$. A similar expression was obtained in Lindqvist and Taraldsen (2007).

The formula (6) for conditional expectations now becomes

$$\mathrm{E}[\phi(\mathbf{X})|\mathbf{T} = \mathbf{t}] = \frac{\int \sum_{\hat{\theta} \in \Gamma(\mathbf{u}, \mathbf{t})} \phi(\chi(\mathbf{u}, \hat{\theta})) h(\mathbf{u}, \hat{\theta}, \mathbf{t}) d\mathbf{u}}{\int \sum_{\hat{\theta} \in \Gamma(\mathbf{u}, \mathbf{t})} h(\mathbf{u}, \hat{\theta}, \mathbf{t}) d\mathbf{u}},$$

while the Metropolis-Hastings method of Section 2.1.3 may proceed as follows. First, propose the $\mathbf{u}$ in the same way as in Section 2.1.3, and then calculate the roots $\hat{\theta} \in \Gamma(\mathbf{u}, \mathbf{t})$. One of these roots, say $\hat{\theta}'$, is then chosen at random according to the conditional distribution of $\hat{\theta}$ given $\mathbf{u}$ and $\mathbf{t}$, as found from (27). A properly modified version of the criterion (7) is then used for acceptance or non-acceptance, using $h(\mathbf{u}', \hat{\theta}', \mathbf{t})$ instead of $\tilde{h}(\mathbf{u}, \mathbf{t})$.

7.6 Using the pivot $\tau(\mathbf{U}, \theta)$ in statistical inference

As noted in Section 2, the random vector $\chi(\mathbf{U}, \theta)$ and hence also $\tau(\mathbf{U}, \theta)$ are pivots in the constructed artificial statistical model. In order to study their possible properties in a statistical inference setting, recall that for the gamma distribution case of Section 4.2.1, we used $f_X(x) = e^{-x}$ and the transformation (16). In this case, (17) is in fact the joint density of n i.i.d. Weibull-distributed random variables with shape parameter α and scale parameter

β. A curious biproduct of our method is therefore the construction of exact confidence sets for the pair (α, β) from observed i.i.d. Weibull-distributed data $\mathbf{u} = (u_1, u_2, \ldots, u_n)$. The basis of the confidence sets would then be to sample vectors $\mathbf{x}$ from the unit exponential distribution, calculate $t_1 = \sum x_i$ and $t_2 = \sum \log x_i$ and solve (21)-(22) for α and β with $\mathbf{u}$ fixed at the observed Weibull-data. The resulting pairs $(\hat{\alpha}, \hat{\beta})$ would then have a joint distribution corresponding to a two-dimensional confidence distribution for (α, β). (This is in some sense exactly the opposite of what we are doing in Section 4.2.1, where t_1 and t_2 are fixed, and we sample the u_i). For exact inference in Weibull models based on the maximum likelihood estimators for (α, β) we refer to Thoman et al. (1969).

References

Bahadur, R., P. Bickel, et al. (1968). Substitution in conditional expectation. *The Annals of Mathematical Statistics 39*(2), 377–378.

Best, D. J., J. C. Rayner, and O. Thas (2012). Comparison of some tests of fit for the inverse Gaussian distribution. *Advances in Decision Sciences 2012*.

Casella, G. and R. L. Berger (2002). *Statistical Inference* (2 ed.). Duxbury, Pacific Grove, CA.

Cheng, R. C. (1984). Generation of inverse Gaussian variates with given sample mean and dispersion. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 33*(3), 309–316.

Dubi, A. and Y. Horowitz (1979). The interpretation of conditional Monte Carlo as a form of importance sampling. *SIAM J.APPL.MATH 36*, 115–122.

Engen, S. and M. Lillegård (1997). Stochastic simulations conditioned on sufficient statistics. *Biometrika 84*(1), 235–240.

Evans, M. and T. Swartz (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press, New York.

Feng, G. and G. Liu (2016). Conditional Monte Carlo: A change-of-variables approach. *arXiv preprint arXiv:1603.06378*.

Gracia-Medrano, L. and F. O'Reilly (2005). Transformations for testing the fit of the inverse-Gaussian distribution. *Communications in Statistics – Theory and Methods 33*(4), 919–924.

Granovsky, B. (1981). Optimal formulae of the conditional Monte Carlo. *SIAM J.Alg.Disc.Meth. 2*, 289–294.

Hammersley, J. (1956). Conditional Monte Carlo. *Journal of the ACM (JACM) 3*(2), 73–76.

Hammersley, J. and D. Handscomb (1964). *Monte Carlo Methods*. Methuen's Monographs on Applied Probability and Statistics. Methuen, London.

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

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2 ed.). Springer Texts in Statistics. Springer-Verlag, New York.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (3 ed.). Springer Science & Business Media, New York.

Lindqvist, B. H. and B. Rannestad (2011). Monte Carlo exact goodness-of-fit tests for nonhomogeneous Poisson processes. *Applied Stochastic Models in Business and Industry 27*(3), 329–341.

Lindqvist, B. H. and G. Taraldsen (2005). Monte Carlo conditioning on a sufficient statistic. *Biometrika 92*(2), 451–464.

Lindqvist, B. H. and G. Taraldsen (2007). Conditional Monte Carlo based on sufficient statistics with applications. In V. Nair (Ed.), *Advances In Statistical Modeling And Inference: Essays in Honor of Kjell A Doksum*, pp. 545–561. World Scientific, Singapore.

Lindqvist, B. H., G. Taraldsen, M. Lillegård, and S. Engen (2003). A counterexample to a claim about stochastic simulations. *Biometrika 90*(2), 489–490.

Lockhart, R. A., F. O'Reilly, and M. Stephens (2009). Exact conditional tests and approximate bootstrap tests for the von Mises distribution. *Journal of Statistical Theory and Practice 3*(3), 543–554.

Lockhart, R. A., F. J. O'Reilly, and M. A. Stephens (2007). Use of the Gibbs sampler to obtain conditional tests, with applications. *Biometrika 94*(4), 992–998.

Michael, J. R., W. R. Schucany, and R. W. Haas (1976). Generating random variates using transformations with multiple roots. *The American Statistician 30*(2), 88–90.

O'Reilly, F. and L. Gracia-Medrano (2006). On the conditional distribution of goodness-of-fit tests. *Communicatiions in Statistics – Theory and Methods 35*(3), 541–549.

Razali, N. M., Y. B. Wah, et al. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics 2*(1), 21–33.

Ripley, B. (1987). *Stochastic Simulation.* Wiley, New York.

Rudin, W. (1987). *Real and Complex Analysis* (3 ed.). McGraw-Hill, Singapore.

Santos, J. D. and N. L. S. Filho (2019). A Metropolis algorithm to obtain co-sufficient samples with applications in conditional tests. *Communications in Statistics-Simulation and Computation 48*(9), 2655–2659.

Schweder, T. and N. L. Hjort (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions.* Cambridge University Press, New York.

Seshadri, V. (2012). *The Inverse Gaussian Distribution: Statistical Theory and Applications*. Springer Science & Business Media, New York.

Stephens, M. A. (1970). Use of the Kolmogorov–Smirnov, Cramer–von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society: Series B (Methodological) 32*(1), 115–122.

Thoman, D. R., L. J. Bain, and C. E. Antle (1969). Inferences on the parameters of the Weibull distribution. *Technometrics 11*(3), 445–460.

Trotter, H. and J. Tukey (1956). Conditional Monte Carlo for normal samples. In H. Meyer (Ed.), *Proc. Symp. on Monte Carlo Methods*, pp. 64–79. John Wiley and Sons, New York.

Wendel, J. et al. (1957). Groups and conditional Monte Carlo. *The Annals of Mathematical Statistics 28*(4), 1048–1052.

Appendix

Lemma 1. *Let $n \in \mathbb{N}$ and $u_1, u_2, \ldots, u_n \in \mathbb{R}^+$, and let for some $v_1, v_2, \ldots, v_n \in \mathbb{R}^+$,*

$$\sum_{i=1}^{n} v_i = t_1,$$

$$\sum_{i=1}^{n} \ln v_i = t_2.$$

Then the system of equations

$$\begin{cases} \sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} & = & t_1 \\ \sum_{i=1}^{n} \ln \left(\frac{u_i}{\beta}\right)^{\alpha} & = & t_2, \end{cases}$$

has a unique solution for $\alpha, \beta \in \mathbb{R}^+$.

Proof. We can transform the system into

$$\begin{cases} \sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} & = & t_1 \\ \frac{\sum_{i=1}^{n} u_i^{\alpha}}{\left(\prod_{i=1}^{n} u_i^{\alpha}\right)^{1/n}} & = & \frac{t_1}{\exp(t_2/n)} \end{cases}$$

If the function

$$p(\alpha) = \frac{\sum_{i=1}^{n} u_i^{\alpha}}{\left(\prod_{i=1}^{n} u_i^{\alpha}\right)^{1/n}}$$

is monotone, then there is a unique solution. The derivative is

$$\begin{aligned} p'(\alpha) &= \left(\sum_{i=1}^{n} \left(\frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}\right)^{\alpha}\right)' \\ &= \sum_{i=1}^{n} \left(\frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}\right)^{\alpha} \ln \frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}. \end{aligned}$$

We note that $\lim_{\alpha \to 0^+} p'(\alpha) = 0$. The second derivative is

$$p''(\alpha) = \sum_{i=1}^{n} \left(\frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}\right)^{\alpha} \ln^2 \frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}} \geq 0.$$

Since the second derivative is positive, the first derivative is increasing. Hence we can conclude that the first derivative is always positive and p is increasing. The solution exists, since

$$\lim_{\alpha \to 0^+} p(\alpha) = n$$

and

$$\frac{t_1}{\exp(t_2/n)} = \frac{\sum_{i=1}^n v_i}{\left(\prod_{i=1}^n v_i\right)^{1/n}} \geq n.$$

The last inequality holds because the arithmetic mean is always larger than or equal to the geometric mean. □

Lemma 2. *Let $n \in \mathbb{N}$ and $u_1, u_2, \ldots, u_n \in \mathbb{R}^+$, and let for some $v_1, v_2, \ldots, v_n \in \mathbb{R}^+$,*

$$\sum_{i=1}^n v_i = t_1,$$

$$\sum_{i=1}^n v_i^{-1} = t_2.$$

Then the system of equations

$$\begin{cases} \sum_{i=1}^n \left(\frac{u_i}{\beta}\right)^{\alpha} & = \quad t_1 \\ \sum_{i=1}^n \left(\frac{u_i}{\beta}\right)^{-\alpha} = t_2, & \end{cases}$$

has a unique solution for $\alpha, \beta \in \mathbb{R}^+$.

Proof. We can transform the system into

$$\begin{cases} \sum_{j=1}^n u_j^{\alpha} \sum_{i=1}^n u_i^{-\alpha} & = \quad t_1 t_2 \\ \\ \sum_{i=1}^n \left(\frac{u_i}{\beta}\right)^{-\alpha} & = \quad t_2 \end{cases}$$

If the function

$$p(\alpha) = \sum_{j=1}^n u_j^{\alpha} \sum_{i=1}^n u_i^{-\alpha}$$

is monotone, then there is a unique solution for α. In order to prove the monotonicity, let $y_{ij} = \frac{u_j}{u_i}$, where $i, j = 1, 2, \ldots, n$, $i \neq j$. The derivative is

$$\begin{aligned}
p'(\alpha) &= \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \left(\frac{u_j}{u_i} \right)^{\alpha} \right)' \\
&= \left(\sum_{j=1}^{n} \sum_{i=1}^{n} y_{ij}^{\alpha} \right)' \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} y_{ij}^{\alpha} \ln y_{ij} \\
&= \sum_{i<j} \ln y_{ij} \left(y_{ij}^{\alpha} - y_{ij}^{-\alpha} \right). \qquad (28)
\end{aligned}$$

Now, if $y_{ij} > 1$, then $\ln y_{ij} > 0$ and $y_{ij}^{\alpha} > y_{ij}^{-\alpha}$, which means that

$$\ln y_{ij} \left(y_{ij}^{\alpha} - y_{ij}^{-\alpha} \right) > 0.$$

If $y_{ij} < 1$, then $\ln y_{ij} < 0$ and $y_{ij}^{\alpha} < y_{ij}^{-\alpha}$, which means that

$$\ln y_{ij} \left(y_{ij}^{\alpha} - y_{ij}^{-\alpha} \right) > 0.$$

Hence, we can conclude that (28) is positive and the function p is increasing. Since

$$\lim_{\alpha \to 0^+} p(\alpha) = n^2$$

and

$$t_1 t_2 = \sum_{i=1}^{n} v_i \sum_{i=1}^{n} v_i^{-1} \geq n^2$$

the solution always exists. □

Conditional Goodness-of-Fit Tests for Discrete Distributions

Rasmus Erlemann and Bo Henry Lindqvist

Submitted and under revision, available at

`https://arxiv.org/abs/2010.03891`

Conditional Goodness-of-Fit Tests for Discrete Distributions

Rasmus Erlemann[*], Bo Henry Lindqvist

Department of Mathematical Sciences, NTNU

October 8, 2020

Abstract

In this paper, we address the problem of testing goodness-of-fit for discrete distributions, where we focus on the geometric distribution. We define new likelihood-based goodness-of-fit tests using the beta-geometric distribution and the type I discrete Weibull distribution as alternative distributions. The tests are compared in a simulation study, where also the classical goodness-of-fit tests are considered for comparison. Throughout the paper we consider conditional testing given a minimal sufficient statistic under the null hypothesis, which enables the calculation of exact p-values. For this purpose, a new method is developed for drawing conditional samples from the geometric distribution and the negative binomial distribution. We also explain briefly how the conditional approach can be modified for the binomial, negative binomial and Poisson distributions. It is finally noted that the simulation method may be extended to other discrete distributions having the same sufficient statistic, by using the Metropolis-Hastings algorithm.

Keywords: Goodness-of-fit; Conditional distribution; Geometric distribution; Monte Carlo simulation; Sufficient statistic; Beta-geometric distribution; Discrete Weibull distribution.

[*]Email: rasmus.erlemann@ntnu.no   Address: Kilu 17, 13516, Tallinn, Estonia

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution F. Goodness-of-fit testing is concerned with how well a family of distributions $\mathcal{F}$ fits the data as a probability model. The null hypothesis is $F \in \mathcal{F}$ and the alternative hypothesis is $F \notin \mathcal{F}$. In-depth literature on this topic includes D'Agostino and Stephens (1986). In the literature of goodness-of-fit testing, most of the work has been focused on continuous distributions, i.e. $\mathcal{F}$ a family of continuous distributions. For discrete distributions, the main interest has been in the Poisson distribution which plays a special role in probability theory. Goodness-of-fit tests for the Poisson distribution go at least back to Fisher (1950) and Rao and Chakravarti (1956). More recent studies of the Poisson distribution are Spinelli and Stephens (1997) and Rueda et al. (1991).

Common alternatives to the Poisson distribution are the negative binomial distribution and its special case, the geometric distribution. The latter distribution is of particular interest since it is the discrete counterpart of the exponential distribution, and is hence an important distribution with various applications, for example in survival analysis, reliability analysis and queuing theory. Bracquemond et al. (2002) presented a comprehensive study of different goodness-of-fit test statistics for the geometric distribution and a comparison between them in a simulation study. Another paper considering tests for the geometric distribution is Ozonur et al. (2013). The present paper will mainly be concerned with goodness-of-fit testing for the geometric distribution, although several ideas considered can easily be modified to cover other discrete distributions.

Traditional goodness-of-fit tests, both in the continuous and discrete distribution cases, are the Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling tests, see for example D'Agostino and Stephens (1986). Various methods are used for finding critical values, typically based on standard asymptotic techniques or parametric bootstrapping. There are, however, also other methods or tricks available, often used to tailor goodness-of-fit testing for specific models.

One such "trick", which will be the main tool in the present paper, is to condition on sufficient statistics under the null hypothesis model to be tested. Such approaches go back to the 1950s. More specifically, Fisher (1950) obtained in this way exact versions of the chi-squared test and an alternative test based on the dispersion for the Poisson distribution, using the fact that the sum of the observations is a sufficient statistic in this case. As a follow-up, Rao and Chakravarti (1956) used the same idea to derive an exact test for the Poisson case based on a likelihood ratio statistic (see Section 3.3.1). Conditioning on sufficient statistics has also been used recently in Beltrán-Beltrán and O'Reilly (2019) and Puig and Weiß (2020). While the just cited papers have considered models with one unknown parameter under the null hypothesis, Heller (1986) did goodness-of-fit testing for the two-parameter negative binomial distribution, assuming both parameters are unknown. Then she conditioned on the sum of the observations in order to eliminate the probability parameter and then using an asymptotic approach having only one unknown parameter.

Often, the sufficient statistic under the null model is easy to find, but still the calculation of critical values or p-values for the conditional tests can be problematic. Usually it will be necessary to sample from the conditional distributions given the sufficient statistic. For goodness-of-fit testing in continuous distributions, and in particular in models where there are more than one parameter, this may however not be straightforward. For possible approaches, see Lindqvist and Taraldsen (2005), Lindqvist et al. (2020), Lockhart et al. (2007).

For the most common discrete distributions, like the binomial and the Poisson distribution, it is straightforward and well known how to do conditional sampling (González-Barrios et al., 2006). How to perform conditional sampling for the geometric distribution and the negative binomial distribution is, apparently, less studied. González-Barrios et al. (2006) derive the conditional distribution for this case, but does not advice a way of simulating from it. In Section 3.1 we show how this can be done by using the so called "bars and stars" framework of Feller (1968). It is believed that the associated algorithm is new in goodness-of-fit studies of the geometric distribution. An extension to the negative binomial distribution is given in the Appendix. Another way of obtaining conditional samples in discrete distributions is suggested by Beltrán-Beltrán and O'Reilly (2019), based on the so called Rao-Blackwell distribution.

The most important ingredient of a goodness-of-fit test is of course the test statistic. The three standard tests, the Kolmogorov-Smirnov, the Cramér-von Mises and the Anderson-Darling test, are already mentioned. These are examples of tests based on the empirical distribution function of the data. While the Kolmogorov-Smirnov statistic considers the maximal difference between the null model and the empirical distribution of the data, the two other tests are based on the corresponding integrated squared difference. A well known fact is that the Anderson-Darling statistic differs from the Cramér-von Mises statistic in that it gives more weight to extreme values of the observations. There are in the literature also considered other test statistics that are known to be large when the null hypothesis model does not hold, but without connection to particular alternative models. Examples are chi-squared tests and tests based on Fisher's index of dispersion, which is the ratio of the variance to the mean, and is well known to be 1 for the Poisson distribution. Closely related to these tests are the tests derived by Kyriakoussis et al. (1998) based on characterizations of the Poisson, binomial and negative binomial distributions by their power-series representations (see Section 3.3.2).

The above tests are essentially not tailored for specific alternative distributions. There might in applications be of importance, however, to have tests that are particularly powerful for given alternative distributions. One purpose of the present paper is to investigate how well the standard goodness-of-fit tests for the geometric distribution will do compared to tests tailored for specific alternatives.

A classical problem is to test the Poisson distribution versus models for over-dispersion, such as the negative binomial distribution. The above cited paper by Puig and Weiß (2020) gives another example. These authors considered testing of the Poisson distribution versus alternatives with log-convex probability generating functions, shown to have important applications in biodosimetry.

Weinberg and Gladen (1986)) considered human fecundability data, using the geometric distribution to model the number of menstrual cycles required to achieve pregnancy. It is then reasonable to believe that the parameter p of the geometric distribution varies between couples. The cited authors showed how to model this variation by means of beta distributions, which leads to counts following a so called beta-geometric distribution. Subsequently, Paul (2005) studied goodness-of-fit testing for the geometric distribution by testing versus the alternative being the beta-geometric distribution, using a score test and a likelihood ratio test. We return to this in Section 3.3.2.

In reliability, a classical problem is to test the null hypothesis of an exponential distribution versus the alternative of a Weibull distribution. This may be done in a straightforward manner using a likelihood ratio test. In the discrete case, this would mean to test the geometric distribution versus some kind of discrete Weibull distribution. We will consider this problem in Section 3.3.3, using the so called type I discrete Weibull distribution, see for example Bracquemond and Gaudoin (2003).

As a final comment on the use of conditional testing, one might ask what is possibly lost in power by such an approach when compared to unconditional ones. We have not pursued this problem, but refer to Lockhart et al. (2009) who concluded from a particular study that calculated p-values from conditional tests are highly correlated with p-values found by parametric bootstrapping. An apparent advantage with the conditional tests is, moreover, that these tests are exact, while the bootstrap based tests are not exact (albeit almost so).

In the second section, we introduce how conditional tests are used in goodness-of-fit testing with discrete null hypothesis. We also cover how the p-values and powers are calculated with Monte Carlo methods in that setting. In the third section we present our method for drawing conditional simulations from the geometric distribution. It is based on the so-called stars and bars representation, introduced by Feller (1968). In the same section, we also introduce some classical test statistics and define new likelihood based tests. In the end of the section, there are two examples where the data is simulated from the beta-geometric and the discrete Weibull distribution of type I and we calculate the conditional p-values for both cases. The fourth section consists of the power study and simulated type I errors. In the fifth section we consider a real life data set and use previously mentioned methods to test if the geometric distribution fits the data. The sixth section consists of conclusions and a brief outline for possible future work in this subject. The last section is the appendix. There are proofs, algorithm descriptions, method for the negative binomial and parameterizations of the distributions we used. References are at the very end of this paper.

In the following we shall let $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. A sample of random variables $Y_1, Y_2, \ldots, Y_n$ is denoted shortly in its vector form by a bold letter, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Bold $\mathbf{P}$ is reserved for the probability function to differentiate it from other functions. Notation for different distributions and the parametrizations are specified in the appendix.

2 Conditional Tests in Goodness-of-Fit Testing for Discrete Distributions

2.1 Calculation of Conditional p-values by Monte Carlo Simulation

For illustration we focus on goodness-of-fit testing for the geometric distribution. Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample from a population with values in $\mathbb{N}_0$. The null and alternative hypotheses are as follows.

H_0 : The random sample is from a population which has the geometric distribution,

H_1 : The random sample is not from a population which has the geometric distribution.

Let $D = D(\mathbf{X})$ be a test statistic such that large values of D are supposed to indicate deviations from H_0. Suppose that, under the null hypothesis, $T = T(\mathbf{X})$ is a sufficient statistic. Algorithm 2 in the Appendix calculates, by Monte Carlo simulation, the conditional p-value of the test from the formula

$$p^{\text{cond}} = \mathbf{P}(D(\mathbf{X}) \geq D(\mathbf{x}_{obs}) \mid T(\mathbf{X}) = t)$$

where $\mathbf{x}_{obs}$ is the observed value of $\mathbf{X}$ and t is the observed value of $T(\mathbf{X})$. For the Monte Carlo simulation one therefore needs a way of simulating from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$. In the next Section we show how this can be done in the case of the geometric distribution.

2.2 Calculation of Test Power of Conditional Tests

To calculate the power of a goodness-of-fit test for a given alternative distribution and for a given significance level α, we proceed as follows. Draw a large number M data sets from the alternative distribution. For the i-th such set, calculate the conditional p-value, p_i^{cond} by Algorithm 2, $i = 1, 2, \ldots, M$. The Monte Carlo power of the test can then be calculated as

$$\beta(\alpha) \approx \frac{\sum_{i=1}^{M} I(p_i^{\text{cond}} \leq \alpha)}{M}.$$

Power calculations require a large number of iterations. Let K be the number of iterations used to calculate each conditional p-value. If M is the number of data sets drawn to calculate the power, then in total we are doing M times K iterations. The number of data sets M is chosen to be as large as possible depending on computational capabilities. We chose $M = 1000$.

It should be noted that since we are dealing with discrete distributions, for a given data set $\mathbf{x}$ with $T(\mathbf{x}) = t$, there are only finitely many possible data sets in the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$. This means that, although we fix a significance level α and are guaranteed a size of the test that is $\leq \alpha$, the size may be strictly less than α. This problem is of course of less concern if n is large.

3 Goodness-of-fit Testing in the Geometric Distribution

3.1 Conditional Sampling from the Geometric Distribution

Let $X_1, X_2, \ldots, X_n$ be iid random variables, such that $X_i \sim \text{Geom}(p)$ for all $i = 1, 2, \ldots, n$, i.e.,

$$\mathbf{P}(X_i = x) = p(1-p)^x \quad \text{for } x = 0, 1, 2, \ldots$$

Then $T(\mathbf{X}) = \sum_{i=1}^{n} X_i = t$ is a sufficient statistic. The conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$ is calculated as

$$\begin{aligned}
\mathbf{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) &= \mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \sum_{i=1}^{n} X_i = t) \\
&= \frac{\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = t)}{\mathbf{P}(\sum_{i=1}^{n} X_i = t)} \\
&= \begin{cases} \dfrac{\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{\mathbf{P}(\sum_{i=1}^{n} X_i = t)}, & \text{if } \sum_{i=1}^{n} X_i = t \\ 0, & \text{if } \sum_{i=1}^{n} X_i \neq t \end{cases}.
\end{aligned}$$

If we restrict the support to be $S = \{(x_1, x_2, \ldots, x_n) : \ \sum_{i=1}^{n} x_i = t, \ x_1, x_2, \ldots, x_n \in \mathbb{N}_0\}$, we get

$$\begin{aligned}
\mathbf{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) &= \frac{p^n (1-p)^{\sum_{i=1}^{n} x_i}}{p^n (1-p)^t \binom{t+n-1}{n-1}} \\
&= \frac{1}{\binom{t+n-1}{n-1}}. \qquad (1)
\end{aligned}$$

As the conditional probability (1) does not depend on $\mathbf{x} \in S$, the distribution of $\mathbf{X} \mid T(\mathbf{X}) = t)$ is uniform on S, i.e., on the set of all possible ways that n non-negative integers sum to t. Feller (1968) introduced a representation of such sums through the so called "stars and bars" framework. To construct one such sum, we lay down t stars and put $n-1$ bars between them. The sum is constructed by counting the number of stars between the bars, letting the first element be the number of stars in front of the first bar, and letting the last one be the number to the right of the last bar. For example, for $t = 8$ and $n = 4$, Figure 1 represents the sum $1 + 0 + 3 + 4 = 8$.

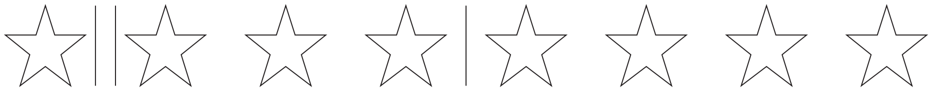

Figure 1: Stars and Bars Representation

The following lemma states that the representation gives rise to any possible sum and conversely. The proof is given in Appendix.

Lemma 1 *For $t, n \in \mathbb{N}$, let*

$$L_1 = \left\{(x_1, x_2, \dots, x_n) : \sum_{i=1}^{n} x_i = t, \ x_1, x_2, \dots, x_n \in \mathbb{N}_0\right\}$$

and

$$L_2 = \{(k_1, k_2, \dots, k_{n-1}) : \ k_1 < k_2 < \dots < k_{n-1} < t + n, \ k_1, \dots, k_{n-1} \in \mathbb{N}\}.$$

Define a transformation ϕ: $L_2 \to L_1$, such that

$$\phi(k_1, k_2, \dots, k_{n-1}) =$$

$$= (k_1 - 1, (k_2 - 2) - (k_1 - 1), \dots, (k_{n-1} - (n-1)) - (k_{n-2} - (n-2)), t - (k_{n-1} - (n-1))).$$

Then ϕ is a bijection between the sets L_2 and L_1.

Example 1 *The k_j in L_2 are the positions of the bars in the stars and bars representation. For example, the stars and bars representation of $n = 4$ and $t = 1+0+3+4 = 8$ in Figure 1, correspond to*

$$k_1 = 2, \ k_2 = 3, \ k_3 = 7.$$

Lemma 2 *Let $t, n \in \mathbb{N}$. The set*

$$\left\{(x_1, x_2, \dots, x_n) : \sum_{i=1}^{n} x_i = t, \ x_1, x_2, \dots, x_n \in \mathbb{N}_0\right\}$$

has $\binom{t+n-1}{n-1}$ elements.

The proof of Lemma 2 can be found in Feller (1968). In fact, it also follows from (1).

Algorithm RandomKSubsets from Wilf (1999) is a method for drawing the so-called bars $k_1, \dots, k_{n-1}$ uniformly from L_2. We have modified it by making recursive calls into iterative ones. This allows the algorithm to be used with large values of n and t more efficiently and there are no issues with recursion depth limitations. It is described by Algorithm 1 in Appendix, where a proof of its correctness is also given (Lemma 3).

Algorithm 1 gives us a sample $k_1, \dots, k_{n-1}$ and the last step is to transform it back into an element from L_1. We use the previously defined function ϕ for it and

$$\phi(k_1, k_2, \dots, k_{n-1}) \sim \mathbf{X} \mid T(\mathbf{X}) = t.$$

3.2 Standard Goodness-of-fit Test Statistics for Discrete Distributions

The following is a general setup for calculation of test statistics for the most common goodness-of-fit tests, with focus on the geometric distribution. The setup essentially follows the one of Spinelli and Stephens (1997) who in particular studied the performance for the Poisson distribution.

Let $x_1, x_2 \ldots, x_n$ be the observed sample and $t = \sum_{i=1}^{n} x_i$ the sufficient statistic. Maximum likelihood estimator for the geometric distribution is given by

$$\hat{p} = \frac{n}{t+n}.$$

In order to avoid trivial cases, we will assume $t > 0$ and hence $0 < \hat{p} < 1$. Now, define for $j = 0, 1, 2, \ldots$,

$$\begin{aligned} o_j &= \#\{i : x_i = j\} = \text{ observed number of values } j \text{ for the sample} \\ \hat{p}_j &= \hat{p}(1-\hat{p})^j = \text{ probability of value } j \text{ in geometric distribution} \\ \hat{e}_j &= n\hat{p}_j = \text{ estimated expected number of values } j \text{ for the sample} \end{aligned}$$

From this define, for $k = 0, 1, 2, \ldots$,

$$\begin{aligned} \hat{Z}_k &= \sum_{j=0}^{k} (o_j - \hat{e}_j) \equiv O_k - \hat{E}_k \\ \hat{H}_k &= \sum_{j=0}^{k} \hat{p}_j \end{aligned}$$

where $O_k = \sum_{j=0}^{k} o_j$ is the observed number of values $\leq k$ in the sample and $\hat{E}_k = n\hat{H}_k$ is its estimated expected value.

Further, define

$$\begin{aligned}
M_0^u &= \min\{j : o_{j'} = 0 \text{ for all } j' > j\} \\
M_1^u &= \min\{j : p_{j'} < 10^{-3}/n \text{ for all } j' > j\} \\
M^u &= \max\{M_0, M_1\} \\
M_0^l &= \max\{j : o_{j'} = 0 \text{ for all } j' < j\} \\
M_1^l &= \max\{j : p_{j'} < 10^{-3}/n \text{ for all } j' < j\} \\
M^l &= \min\{M_0, M_1\}
\end{aligned}$$

3.2.1 The Cramér-von Mises Test

The Cramér-von Mises test statistic is defined by

$$W^2 = \frac{1}{n} \sum_{M^l}^{M^u} \hat{Z}_i^2 \hat{p}_i.$$

3.2.2 The Anderson-Darling Test

The Anderson-Darling test statistic is defined by

$$A^2 = \frac{1}{n} \sum_{M_l}^{M^u} \frac{\hat{Z}_i^2 \hat{p}_i}{\hat{H}_i(1 - \hat{H}_i)}.$$

3.2.3 The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test statistic should ideally be defined as $\max_{k=1,2,\ldots} |Z_k|$. As shown by Bracquemond et al. (2002), the maximum will always occur for a $k \leq M_0^u$, so that we define

$$KS = \max_{k=0,1,2,\ldots,M_0^u} |Z_k|.$$

To see this, recall that $\hat{Z}_k = O_k - \hat{E}_k$. Now $O_k = n$ for $k \geq M_0^u$, while $\hat{E}_k < n$ and $\hat{E}_k$ is increasing in k. Hence, for $k \geq M_0$, $|Z_k| = Z_k$ and is decreasing.

3.3 Likelihood Based Tests

In the present subsection we study tests that are derived with the aim of having high power against given alternative distributions. The main tool is here to consider likelihood functions.

3.3.1 Test Versus Heterogeneous Geometric Observations

In this subsection we follow the idea of Rao and Chakravarti (1956), who considered the Poisson distribution where we consider the geometric distribution.

Suppose $X_1, X_2, \ldots, X_n$ are independent and geometrically distributed, but with different parameters p_i. The log likelihood for data $x_1, x_2, \ldots, x_n$ would then be

$$\ell(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} (\log(p_i) + x_i \log(1 - p_i)),$$

which is maximized by $\hat{p}_i = 1/(1 + x_i)$ for $i = 1, 2, \ldots, n$. The relevant null hypothesis is now

$$H_0 : p_1 = p_2 = \ldots = p_n = p.$$

The log likelihood under the null hypothesis is then $\ell(p, \ldots, p) = n \log(p) + t \log(1 - p)$ where $t = \sum_{i=1}^{n} x_i$, which is maximized by $\hat{p} = n/(n + t)$.

The likelihood ratio statistic can therefore be written

$$\begin{aligned}
& \ell(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n) - \ell(\hat{p}, \ldots, \hat{p}) \\
= & \sum_{i=1}^{n} \left[\log\left(\frac{1}{1 + x_i}\right) + x_i \log\left(\frac{x_i}{1 + x_i}\right) \right] - n \log\left(\frac{n}{n + t}\right) - t \log\left(\frac{t}{n + t}\right) \\
= & \sum_{i=1}^{n} [x_i \log(x_i) - (x_i + 1) \log(x_i + 1)] - n \log\left(\frac{n}{n + t}\right) - t \log\left(\frac{t}{n + t}\right)
\end{aligned}$$

Since we consider conditional tests given $\sum_{i=1}^{n} X_i = t$, we may exclude the last terms above, which after rewriting the first sum gives the test statistic

$$CR = \sum_{i=1}^{n} [x_i \log(x_i) - (x_i + 1) \log(x_i + 1)] = \sum_{j=M_0^l}^{M_0^u} o_j \big(j \log j - (j + 1) \log(j + 1)\big)$$

where we use $0 \log 0 = 0$.

3.3.2 The Beta-Geometric Distribution

In the previous subsection we considered the alternative hypothesis that the observations were geometrically distributed, but with possibly different parameters p_i. Suppose now that these p_i are drawn independently from the beta distribution.

Thus, for a single observation X we assume that it is geometrically distributed with parameter p, where p is generated from the beta-distribution with parameters $\alpha > 0$ and $\beta > 0$. Let $B(\alpha, \beta)$ be the beta function, defined by

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Then it is seen that the unconditional distribution of X is what has been named the beta-geometric distribution,

$$\mathbf{P}(X = x) = \int_0^1 p(1-p)^x \, \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} dp = \frac{B(\alpha+1, \beta+x)}{B(\alpha, \beta)} \tag{2}$$

for $x = 0, 1, 2, \ldots$. As suggested by Paul (2005), a useful reparametrization is

$$\pi = \frac{\alpha}{\alpha+\beta}, \quad \theta = \frac{1}{\alpha+\beta}. \tag{3}$$

With this parametrization it is seen that $\theta = 0$ corresponds to the geometric distribution with $p = \pi$. Tests for the null hypothesis of geometric distribution can hence be derived by testing

$$H_0 : \theta = 0 \text{ vs. } \theta > 0.$$

Using the reparametrization (3), we find from (2), noting that $\alpha = \pi/\theta$, $\beta = (1-\pi)/\theta$ and using properties of the gamma function,

$$\mathbf{P}(X = x) = \frac{\alpha \prod_{j=1}^{x}(\beta + x - j)}{\prod_{j=0}^{x}(\alpha + \beta + x - j)} = \frac{\pi \prod_{j=0}^{x-1}(1 - \pi + j\theta)}{\prod_{j=0}^{x}(1 + j\theta)} \tag{4}$$

for $x = 0, 1, \ldots$. (Note that the formula also holds for $x = 0$, giving $\mathbf{P}(X = 0) = \pi$, since an empty product by convention equals 1.)

The log-likelihood for data $x_1, x_2, \ldots, x_n$ with values in $\mathbb{N}_0$ is hence, see also Paul (2005),

$$\ell(\mathbf{x}; \pi, \theta) = n \log \pi + \sum_{i=1}^{n} \sum_{j=0}^{x_i - 1} \log(1 - \pi + j\theta) - \sum_{i=1}^{n} \sum_{j=0}^{x_i} \log(1 + j\theta) \tag{5}$$

Differentiating with respect to θ and π give, respectively,

$$\begin{aligned}\frac{\partial\ell}{\partial\theta} &= \sum_{i=1}^{n}\sum_{j=0}^{x_i-1}\frac{j}{1-\pi+j\theta}-\sum_{i=1}^{n}\sum_{j=0}^{x_i}\frac{j}{1+j\theta},\\ \frac{\partial\ell}{\partial\pi} &= \frac{n}{\pi}-\sum_{i=1}^{n}\sum_{j=0}^{x_i-1}\frac{1}{1-\pi+j\theta}.\end{aligned}$$

Letting $\theta = 0$ in the expression for $\frac{\partial\ell}{\partial\theta}$ gives the score statistic for testing $H_0 : \theta = 0$, namely

$$S=\sum_{i=1}^{n}\sum_{j=0}^{x_i-1}\frac{j}{1-\pi}-\sum_{i=1}^{n}\sum_{j=0}^{x_i} j \;=\; \frac{\pi\sum_{i=1}^{n}x_i^2-(2-\pi)\sum_{i=1}^{n}x_i}{2(1-\pi)}. \tag{6}$$

The score test in general rejects $H_0 : \theta = 0$ for large values of $|S|$. Indeed, it can be shown from the rightmost expression in (6) that, under H_0 where the X_i are geometrically distributed with probability $p = \pi$, we have $E(S) = 0$. Paul (2005) replaced π by the maximum likelihood estimate $\hat{p}$ under H_0, and divided the expression in (6) by an estimate of the standard deviation of S under H_0, which is $\sqrt{n}/\hat{p}$. The resulting statistic then has an asymptotically standard normal distribution under H_0.

Let now $m_1 = (1/n)\sum_{i=1}^{n} x_i$ and $m_2 = (1/n)\sum_{i=1}^{n} x_i^2$ be the first and second empirical moments, respectively, from the data $\mathbf{x}$. Replacing π by $\hat{p} = n/(t + n) = 1/(1 + m_1)$ we can write the right hand side of (6) as

$$n\frac{\frac{m_2}{1+m_1}-\left(2-\frac{1}{1+m_1}\right)m_1}{2(1-\frac{1}{1+m_1})}=n\frac{m_2-m_1-2m_1^2}{2m_1}.$$

Since we consider conditional testing given t, or equivalently given m_1, we may use the test statistic.

$$SB = m_2 - m_1 - 2m_1^2. \tag{7}$$

Actually, we could also have deleted the other terms involving m_1. We keep them, however, due to the fact that the sign of SB is of some importance, as explained below.

Note now that, since $\theta \geq 0$ is a model restriction, the maximum of (5) may occur at the boundary point where $\theta = 0$. But for $\theta = 0$, (5) is simply the log likelihood of the geometric distribution and is hence maximized by $\pi = \hat{p}$. Thus if $SB > 0$, then we know that the maximum of (5) is not at a boundary point with $\theta = 0$, and must hence be at a point (π, θ) with $\theta > 0$. This point may hence presumably be found by using the partial derivatives derived above. If, instead, $SB < 0$, then the maximum likelihood estimate is likely to be at the point $(\hat{p}, 0)$.

It follows from the above that if $SB < 0$, then the numerical value is uninteresting, because it corresponds to parameter values outside of the parameter set and intuitively to parameters for which we would not reject the null hypothesis. We therefore suggest to replace SB by $SB_0 = \max(0, SB)$ and call this the score test statistic for the null hypothesis $\theta = 0$.

Paul (2005) considered the score test and in addition the likelihood ratio test based on the log likelihood (5) and standard asymptotics (taking into account the fact that the null hypothesis is on the boundary of the parameter space). He further noted that the likelihood ratio test, as well as the score test, are rather liberal (non-conservative) as regards the size. He therefore found that a bootstrap test might be preferable.

Singh et al. (2014) considered both maximum likelihood estimation and moment estimation of α and β. The moment estimators are obtained as follows. First, define

$$\begin{aligned}
\mu_1 \equiv E(X) &= \frac{\beta}{\alpha - 1} \qquad \text{for } \alpha > 1, \\
\mu_2 \equiv E(X^2) &= \frac{\beta(\alpha + 2\beta)}{(\alpha - 1)(\alpha - 2)} \qquad \text{for } \alpha > 2.
\end{aligned}$$

Solving for α and β we get

$$\begin{aligned}
\alpha &= \frac{2(\mu_2 - \mu_1^2)}{\mu_2 - \mu_1 - 2\mu_1^2}, \\
\beta &= \mu_1(\alpha - 1).
\end{aligned}$$

The moment estimators $\tilde{\alpha}$ and $\tilde{\beta}$ for α and β are obtained by substituting the empirical moments m_1 and m_2 for μ_1 and μ_2, respectively. This leads to an estimator for the parameter θ which can be expressed by

$$\tilde{\theta} = (\tilde{\alpha} + \tilde{\beta})^{-1} = \frac{m_2 - m_1 - 2m_1^2}{2m_2 - m_1^2 + m_1 m_2}.$$

It is noticeable that the numerator of $\tilde{\theta}$ equals SB (see (7)). Thus $\tilde{\theta}$ and SB have the same sign (since the denominator above is always positive). As already noted, this sign is of importance for maximum likelihood estimation based on (5). Note, on the other hand, that in a conditional test using $\hat{\theta}$ we cannot ignore the denominator of $\tilde{\theta}$, since it contains m_2.

As a final note in this subsection, the test statistic SB_0 appears to be essentially identical to the one for the geometric distribution which is derived in Kyriakoussis et al. (1998). These authors derived test statistics from characterizations of distributions given by power-series distribution laws, which include Poisson, binomial, and the negative binomial distribution. Their general goodness-of-fit test statistic is

$$\hat{c} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - 1)}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2} = \frac{m_2 - m_1}{m_1^2} \tag{8}$$

and their test for the geometric distribution rejects the null hypothesis if a normalized version of

$$|\hat{c} - 2| = \frac{|m_2 - m_1 - 2m_1^2|}{m_1^2}$$

is large, where the normalization leads to an asymptotically standard normal distribution under H_0. Since the normalization is a function of m_1 only, and the denominator of (8) can be deleted, using their statistic in a conditional testing we in fact end up with the test statistic $|SB|$.

Example

Suppose we have observed the data in Table 1, which are simulated from the beta-geometric distribution with $n = 100, \pi = 0.4, \theta = 0.125$.

Using the described test statistics for testing the null hypothesis of a geometric distribution, we obtained the conditional p-values given in Table 2. We note that also the three standard tests are able to detect the deviation from the geometric distribution here, while p-values are remarkably lower for the tests derived above that are tailored for detecting deviations in the direction of a beta-geometric distribution. In fact, the same low p-values are obtained for tests versus the discrete Weibull distribution that will be studied below.

Maximum likelihood estimates for π and θ in the beta-geometric distribution can be calculated using the R-package VGAM, giving $\hat{\pi} = 0.4274$, $\hat{\theta} = 0.1166$. These estimates are used to calculate the estimated expected counts in Table 1. It is remarkable that these are much closer to the observed values than the ones estimated from the geometric distribution.

j	o_j	$\hat{e}_j^g$	$\hat{e}_j^b$
0	42	35.4	42.7
1	24	22.8	21.9
2	11	14.8	12.2
3	8	9.5	7.3
4	4	6.2	4.6
5	4	4.0	3.0
6	0	2.6	2.1
7	1	1.7	1.4
8	0	1.1	1.0
9	2	0.7	0.8
10	2	0.4	0.6
11	0	0.3	0.4
12	0	0.2	0.3
13	0	0.1	0.3
14	0	0.1	0.2
15	1	0.0	0.2
16	1	0.0	0.1

Table 1: Data simulated from the beta-geometric distribution with $n = 100, \pi = 0.4, \theta = 0.125$. The column o_j gives the number of observations x_i that resulted in $x_i = j$. The two last columns give the estimated expected frequencies under a geometric distribution and beta-geometric distribution, respectively.

Statistic	W^2	A^2	KS	CR	SB	SB_0	$\hat{\theta}$	$\lvert SW\rvert$	SWL	SWU
p^{cond}	0.034	0.028	0.059	0.009	0.004	0.004	0.004	0.005	0.004	0.996

Table 2: Conditional p-values obtained by simulating 10000 data sets from the conditional distribution

3.3.3 The discrete Weibull distribution of type I

Let for $x = 0, 1, 2, \ldots,$

$$\mathbf{P}(X = x) = q^{x^\beta} - q^{(x+1)^\beta} \tag{9}$$

where $0 < q < 1$ and $\beta > 0$. This is the probability distribution of the type I Weibull distribution, which was introduced by Nakagawa and Osaki (1975). We denote it by $\mathcal{W}(q, \beta)$. The geometric distribution with parameter p is now a special case obtained when $q = 1-p$ and $\beta = 1$.

The R-package DiscreteWeibull contains routines for this distribution, including simulation of data and estimation of parameters.

The discrete hazard rate of a random variable with values in the (nonnegative) integers can be defined by (Barlow et al., 1963) $\lambda(x) = \mathbf{P}(X = x \mid X \geq x)$. From (9) we get

$$\lambda(x) = \frac{\mathbf{P}(X = x)}{\mathbf{P}(X \geq x)} = \frac{q^{x_i^\beta} - q^{(x_i+1)^\beta}}{q^{x_i^\beta}} = 1 - q^{(x+1)^\beta - x^\beta}$$

which is seen to be increasing in x if $\beta > 1$ and decreasing in x if $\beta < 1$, and constant equal to p when $\beta = 1$, which corresponds to the geometric distribution.

Suppose now we have data $x_1, x_2, \ldots, x_n$ with values in $\mathbb{N}_0$. Testing the null hypothesis that the data come from the geometric distribution, is now equivalent to testing $H_0 : \beta = 1$ vs. $H_1 : \beta \neq 1$, or possibly the one-sided versions of the alternative. The testing can be done by a likelihood ratio test. It follows from (9) that the log-likelihood for the sample $x_1, x_2, \ldots, x_n$ from the type I Weibull distribution is given by

$$\ell(q, \beta) = \sum_{i=1}^{n} \ln\left(q^{x_i^\beta} - q^{(x_i+1)^\beta}\right).$$

The likelihood ratio test statistic can be computed by calculating the maximum likelihood estimates of q and β, and of p, which is the parameter under the null hypothesis model. Details are given by Vila et al. (2019), while computations can be done using the R-package DiscreteWeibull.

A score test can be derived in a way similar to what we did in Section 3.3.2 for the beta-geometric distribution. First, the partial derivative with respect to β of the log-likelihood function ℓ, is given by

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{q^{x_i^\beta} x_i^\beta \ln(q) \ln(x_i) - q^{(x_i+1)^\beta} (x_i+1)^\beta \ln(q) \ln(x_i+1)}{q^{x_i^\beta} - q^{(x_i+1)^\beta}}$$

The score statistic of H_0 can then be found by letting $\beta = 1$, which leads to

$$\frac{\partial \ell}{\partial \beta}|_{\beta=1} = \frac{\ln(q)}{1-q} \sum_{i=1}^{n} \left(x_i \ln(x_i) - q(x_i+1)\ln(x_i+1) \right). \qquad (10)$$

It can now be checked that if the x_i are from the geometric distribution with parameter p, the expected value of (10) is 0 (noting that $q = 1-p$). The standard approach is now to estimate q by $1-\hat{p}$ (from the geometric distribution) and divide (10) by the estimated standard deviation, in order to obtain a test statistic which is standard normally distributed under the null hypothesis. We shall, however, consider conditional testing, conditioning on $\sum_{i=1}^{n} X_i$ or, equivalently, on $\hat{p}$, and we may hence use the test statistic

$$SW = \sum_{i=1}^{n} \left[(1-\hat{p})(x_i+1)\ln(x_i+1) - x_i \ln(x_i) \right].$$

where $\hat{p} = n/(n + \sum_i x_i)$. If $x_i = 0$, we shall let $x_i \ln(x_i) = 0$. Note that we have changed the order of the terms inside the sum as compared to (10). This is because $\ln(q) < 0$ and will lead to a statistic SW with the same sign as $\frac{\partial \ell}{\partial \beta}|_{\beta=1}$. Then for the two-sided alternative, $\beta \neq 1$, we should use the statistic $|SW|$ as the test statistic. A more powerful test can then be defined for the two one-sided alternatives, by using $SWU = SW$ if the alternative is $\beta > 1$ and $SWL = -SW$ for the alternative $\beta < 1$, and reject in both cases for high values of the test statistic.

The resemblance between the statistics SW and CR is striking. In fact, CR is obtained from SW by letting $\hat{p} = 0$ and switching the sign. Simulations and p-value calculations in the following will indicate the possible difference between their merits.

Example

Suppose we have observed the data in Table 3, which are simulated from a type I discrete Weibull distribution with $n = 50, q = 0.8, \beta = 1.4$ using the R-package DiscreteWeibull.

Using all the tests considered so far in the paper, we obtained the conditional p-values given in Table 4. We note that also the three standard tests are able to detect the deviation from the discrete Weibull distribution here, while p-values are remarkably lower for the tests $|SW|$ and SWL, which are tailored for detecting deviations in the direction of the discrete Weibull distribution. It should be noted, however, that the test based on CR as well as the tests versus the beta-geometric distribution are useless for these data. The reason for this last fact is that the beta-geometric distribution always increases the variance of the data as compared to the geometric distribution, while the discrete Weibull with $\beta > 1$ decreases the variance (a property well known for the continuous Weibull distribution). Thus a beta-geometric distribution would have difficulties fitting these data.

Maximum likelihood estimates for π and θ are calculated as $\hat{q} = 0.7239$, $\hat{\beta} = 1.267$ using the R-package DiscreteWeibull. These estimates are used to calculate the estimated expected counts in Table 3. Again, these are much closer to the observed values than the ones estimated from the geometric distribution.

j	o_j	$\hat{e}_j^g$	$\hat{e}_j^b$
0	13	18.0	13.8
1	14	11.5	13.2
2	10	7.4	9.3
3	8	4.7	5.9
4	1	3.0	3.5
5	1	1.9	2.0
6	0	1.2	1.1
7	2	0.8	0.6
8	1	0.5	0.3

Table 3: Data simulated from a type I Weibull distribution with $n = 50, q = 0.8, \beta = 1.4$. The column o_j gives the number of observations x_i that resulted in $x_i = j$. The two last columns give the estimated expected frequencies under a geometric distribution and type I Weibull distribution, respectively.

Statistic	W^2	A^2	KS	CR	SB	SB_0	$\hat{\theta}$	$\lvert SW\rvert$	SWL	SWU
p^{cond}	0.072	0.078	0.124	0.962	0.890	1.0	0.890	0.083	0.956	0.044

Table 4: Conditional p-values obtained by simulating 10000 data sets from the conditional distribution. Two-sided (one-sided) testing using $|SW|$ (SWU) means that the alternative hypothesis is $\beta \neq 1$ ($\beta > 1$). The statistics SB and SB_0 can only be used to detect an increased variance compared to the geometric distribution. Here the test statistic SB is negative, which makes these tests meaningless.

4 Computer Simulations

4.1 Power Study

We did a power study with various sample sizes, alternative distributions and all previously defined test statistics with significance level $\alpha = 0.1$. In some cases we disregarded sample sizes $n = 5$ or $n = 100$ if the powers were too close to the significance level or 1. We used 1000 iterations to calculate each conditional p-value and another 1000 iterations to calculate the power. These numbers were chosen to make sure each power calculation takes less than half an hour of computation time.

Alternative	Sample size	W^2	A^2	KS	CR	SB_0	$\lvert SW \rvert$	SWL	SWU
Pois(0.5)	$n = 25$	0.231	0.238	0.208	0.002	0.001	0.225	0.002	0.329
	$n = 100$	0.736	0.734	0.705	0.000	0.000	0.763	0.000	0.851
Pois(1)	$n = 5$	0.119	0.119	0.110	0.008	0.008	0.122	0.008	0.134
	$n = 25$	0.613	0.605	0.543	0.000	0.001	0.618	0.000	0.730
	$n = 100$	0.996	0.996	0.992	0.000	0.000	0.998	0.000	0.999
Pois(2)	$n = 5$	0.332	0.321	0.163	0.003	0.003	0.339	0.003	0.395
	$n = 25$	0.963	0.965	0.914	0.000	0.000	0.966	0.000	0.985
$\mathrm{Bin}(5, 0.3)$	$n = 5$	0.418	0.410	0.294	0.000	0.000	0.403	0.000	0.432
	$n = 25$	0.986	0.985	0.972	0.000	0.000	0.990	0.000	0.996
$\mathrm{NB}(5, 0.5)$	$n = 5$	0.531	0.466	0.434	0.000	0.001	0.538	0.000	0.672
	$n = 25$	0.997	0.998	0.986	0.000	0.000	1.000	0.000	1.000
$\mathrm{NB}(3, 0.7)$	$n = 25$	0.395	0.393	0.339	0.001	0.002	0.407	0.001	0.526
	$n = 100$	0.875	0.875	0.834	0.000	0.000	0.906	0.000	0.945
$\mathrm{BG}(2, 5)$	$n = 5$	0.161	0.170	0.137	0.267	0.269	0.158	0.274	0.025
	$n = 100$	0.565	0.565	0.514	0.676	0.705	0.608	0.706	0.007
$\mathrm{BG}(2, 2)$	$n = 5$	0.122	0.132	0.108	0.205	0.201	0.125	0.207	0.018
	$n = 25$	0.558	0.570	0.504	0.705	0.688	0.611	0.717	0.001
$\mathcal{W}(0.7, 0.8)$	$n = 25$	0.320	0.338	0.282	0.503	0.432	0.353	0.492	0.006
	$n = 100$	0.749	0.759	0.679	0.874	0.792	0.792	0.882	0.000
$\mathcal{W}(0.5, 1.5)$	$n = 25$	0.428	0.431	0.385	0.000	0.001	0.438	0.000	0.567
	$n = 100$	0.948	0.943	0.937	0.000	0.000	0.960	0.000	0.984

Table 5: Conditional power calculations with significance level $\alpha = 0.1$.

As usual for a wide choice of alternative distributions, there is no best test against all alternatives. From standard tests, W^2 has slightly higher powers with small sample sizes. For larger sample sizes, A^2 and W^2 are almost identical. Maximal type test KS has slightly lower powers than the other standard tests. Tests CR, SB, SB_0, $\hat{\theta}$, SWL and SWU are sensitive to the alternative distribution the data comes from. This makes them situational and they lack the versatility of the standard tests. For example, CR, SB, SB_0, $\hat{\theta}$ and SWL outperform the standard tests when the data comes from BG or $\mathcal{W}$ distributions. SWU outperforms other tests for Pois, Bin and NB distributions. $|SW|$ is more versatile and has almost identical powers to the standard quadratic tests. Likelihood based tests need a versatile comparative alternative distribution to perform well. Type I Weibull distribution fits this role, as we can see from $|SW|$ powers. Test SWL is for the case where $\beta < 1$ and SWU for $\beta > 1$. Under those conditions, they outperform $|SW|$.

4.2 Type I Errors

The type I error is defined to be the probability of falsely rejecting the null hypothesis if it is actually true. We simulated this scenario by drawing samples under the null hypothesis, from the geometric distribution with parameter p for various sample sizes n. If the conditional p-value came out lower than the significance level, we had made a type I error.

We used 1000 iterations to calculate each conditional p-value and another 1000 iterations to calculate the type I error.

$\alpha = 0.05$									
Parameter	Sample size	W^2	A^2	KS	CR	SB_0	$\|SW\|$	SWL	SWU
$p = 0.25$	$n = 5$	0.048	0.047	0.035	0.037	0.035	0.044	0.037	0.039
	$n = 25$	0.055	0.054	0.047	0.048	0.052	0.052	0.049	0.050
	$n = 100$	0.056	0.056	0.055	0.054	0.056	0.057	0.056	0.053
$p = 0.5$	$n = 5$	0.027	0.026	0.018	0.018	0.017	0.023	0.017	0.020
	$n = 25$	0.066	0.065	0.056	0.050	0.048	0.064	0.051	0.055
	$n = 100$	0.042	0.045	0.036	0.050	0.049	0.039	0.059	0.050
$p = 0.75$	$n = 5$	0.002	0.002	0.000	0.001	0.001	0.001	0.001	0.001
	$n = 25$	0.032	0.030	0.027	0.032	0.027	0.029	0.031	0.017
	$n = 100$	0.045	0.045	0.040	0.046	0.040	0.046	0.041	0.043
$\alpha = 0.1$									
$p = 0.25$	$n = 5$	0.070	0.066	0.057	0.080	0.081	0.073	0.079	0.073
	$n = 25$	0.105	0.113	0.096	0.111	0.118	0.103	0.112	0.106
	$n = 100$	0.101	0.092	0.087	0.104	0.101	0.103	0.097	0.107
$p = 0.5$	$n = 5$	0.042	0.040	0.038	0.035	0.037	0.039	0.035	0.038
	$n = 25$	0.101	0.108	0.071	0.095	0.089	0.110	0.098	0.107
	$n = 100$	0.123	0.116	0.117	0.105	0.104	0.109	0.105	0.115
$p = 0.75$	$n = 5$	0.006	0.006	0.005	0.002	0.002	0.006	0.002	0.006
	$n = 25$	0.062	0.060	0.060	0.059	0.057	0.070	0.057	0.032
	$n = 100$	0.101	0.102	0.090	0.097	0.095	0.095	0.098	0.078

Table 6: Type I errors for various sample sizes, parameter p-values, test statistics and significance levels α.

Some of the type I errors are slightly above the significance level but this is explained by Monte Carlo errors from calculating the p-value and the error. This is because of the discreteness of the data. If the parameter is $p = 0.75$, the samples consist largely of 0-s and if the sample size is small, we often get only 0-s. In that case $t = 0$ and it is a singular case.

We left out $\hat{\theta}$ and SB from the power study and type I error study because they had identical powers for all alternatives and SB_0 should be preferred over SB.

5 Real Life Data

In this section we use real life data from Bracquemond et al. (2002). The data consist of numbers of inspections between discovery of defects in an industrial process. Conditional samples are used to calculate the distribution of goodness-of-fit test statistics following the recipe from Section 2. Conditional p-values are reported to decide if we should reject or not reject the null hypothesis, that the data comes from the geometric distribution.

In order to have the data on the form considered in this paper, we have subtracted 1 from each observation.

Value	0	1	2	3	4	≥ 5
Observed frequency	6	4	3	3	2	10
Expected frequency, geometric	3.9	3.3	2.9	2.5	2.1	13.3
Expected freqquency, beta-geometric	5.0	3.9	3.1	2.5	2.0	11.5
Expected frequency, discrete Weibull	6.0	3.6	2.7	2.2	1.8	11.7

Table 7: Real life data. Observed and estimated expected frequencies for three different models.

The data is given in Table 7. Note that the data $x_i \geq 5$ are lumped together in the table for illustrative purposes. The observed values for these 10 observations are 6. 8, 10, 12, 13, 16, 17, 25, 28, and are used in the simulations and calculations.

Conditional p-values for the various tests are given in Table 8, calculated with 10000 Monte Carlo samples from the conditional distribution.

Statistic	W^2	A^2	KS	CR	SB	SB_0	$\hat{\theta}$	$\|SW\|$	SWL	SWU
p^{cond}	0.107	0.117	0.315	0.042	0.134	0.134	0.134	0.110	0.047	0.953

Table 8: Conditional p-values obtained by simulating 10000 data sets from the conditional distribution.

The standard tests as well as the tests versus beta-geometric distribution still indicate the possibility of a geometric distribution, having p-values > 0.10, while the hypothesis of geometric distribution is in fact rejected at 5% significance level by the CR test and the one-sided test versus the type I discrete Weibull distribution with $\beta < 1$. This possibility of the Weibull distribution is also indicated by the fitted expected frequencies as shown in Table 7.

The VGLM R-package gives the maximum likelihood estimates for the beta-geometric model given by

$$\begin{aligned} \hat{\pi} &= 0.1772 \\ \hat{\theta} &= 0.0502 \\ \hat{p} &= 0.1378 \text{ (geometric distribution)} \end{aligned}$$

Also, the VGLM R-package gives a p-value for a likelihood ratio test verus the beta-geometric to be 0.3276. This is higher than the values for SB and SB_0, e.g. The reason might be that the asymptotic chi-square distribution of the likelihood ratio is not appropriate for these data.

The DiscreteWeibull R-package estimates a type I Weibull model giving $\hat{q} = 0.784, \hat{\beta} = 0.794$, indicating a decreasing hazard rate, which corresponds well to the above rejection of the one-sided test versus $\beta < 1$.

6 Conclusion and Future Work

In this paper we studied goodness-of-fit tests for discrete distributions obtained by conditioning on the sufficient statistic under the null hypothesis. We developed in particular a method to draw conditional samples from the geometric distribution. These samples are used for calculation of p-values for various goodness-of-fit tests. In addition to considering standard goodness-of-fit tests, we derived new likelihood based test statistics for testing of the geometric distribution versus heterogeneity, as well as versus discrete Weibull distributions with both increasing and decreasing hazard.

A power study was conducted to check how the tests perform against data from different alternative distributions. Our simulations suggested that the two-sided test versus the type I discrete Weibull distributions was able to detect bad fit for data from various alternative distributions. The power results for this test, $|SW|$, were in fact generally similar to the ones obtained for the standard quadratic goodness-of-fit tests.

The tests versus heterogeneous geometric distributions, CR and SB_0, are doing well for alternatives of this kind, as one should expect, and then usually much better than the standard tests. The tests for heterogeneity are, however, mostly inferior versus other alternatives. The reason is presumably that heterogeneity leads to increased variance as compared to the geometric distribution. On the other hand, Weibull distributions with decreasing hazard lead to an increased variance.

Real life data from Bracquemond et al. (2002) was considered and it was tested whether the geometric distribution is a good fit. The calculated p-values suggested that the geometric distribution might not be a good fit according to some of the tests.

For further work, a general method could be described for the case where $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is sufficient for the family of distributions under the null hypothesis. This is the case for the power series distributions where the probability distribution is of the form, see Kyriakoussis et al. (1998) or González-Barrios et al. (2006),

$$\mathbf{P}(X = x) = \frac{a(x)\theta^x}{\eta(\theta)} \text{ for } x = 0, 1, 2, \ldots \tag{11}$$

where $a(x) \geq 0$, $\theta > 0$, $\eta(\theta) = \sum_{y=0}^{\infty} a(y)\theta^y$. The Poisson, binomial, negative binomial and geometric distributions are all of this kind. It can be shown from (11) (González-Barrios et al., 2006) that for samples $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from this distribution, we have

$$\mathbf{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) \propto \prod_{i=1}^{n} a(x_i) \text{ when } \sum_{i=1}^{n} x_i = t.$$

Conditional samples with a given sum t can hence be obtained by the Metropolis-Hastings algorithm using samples from the conditional geometric distribution as proposals.

Appendix

Proof of Lemma 1

Proof: The function is defined correctly, because for any $(k_1, k_2, \ldots, k_{n-1}) \in L_2$

$k_1 - 1 + (k_2 - 2) - (k_1 - 1) + \ldots + (k_{n-1} - (n-1)) - (k_{n-2} - (n-2)) + t - (k_{n-1} - (n-1)) = t$

and $k_1 - 1, (k_2 - 2) - (k_1 - 1), \ldots, t - (k_{n-1} - (n-1)) \in \mathbb{N}_0$, as $0 < k_1 < k_2 < \ldots < k_{n-1} < t + n$. Let us assume that $(k_1, k_2, \ldots, k_{n-1}), (v_1, v_2, \ldots, v_{n-1}) \in L_2$, such that

$$\phi(k_1, k_2, \ldots, k_{n-1}) = \phi(v_1, v_2, \ldots, v_{n-1}).$$

This implies that

$$\begin{aligned} k_1 + 1 = v_1 + 1 &\Rightarrow k_1 = v_1, \\ (k_2 - 2) - (k_1 - 1) = (v_2 - 2) - (v_1 - 1) &\Rightarrow k_2 = v_2, \\ &\ldots \\ (k_{n-1} - (n-1)) - (k_{n-2} - (n-2)) = (v_{n-1} - (n-1)) - (v_{n-2} - (n-2)) &\Rightarrow k_{n-1} = v_{n-1}, \end{aligned}$$

and we can conclude that ϕ is injective.

Let us fix an element $(x_1, x_2, \ldots, x_n) \in L_1$, then

$$\phi\left(x_1 + 1, x_1 + x_2 + 2, \ldots, \sum_{i=1}^{n-1} x_i + (n-1)\right) = (x_1, x_2, \ldots, x_n)$$

and

$$0 < x_1 + 1 < x_1 + x_2 + 2 < \ldots < \sum_{i=1}^{n-1} x_i + n - 1 < t + n,$$

which implies that ϕ is surjective. Injectivity and surjectivity imply that ϕ is bijective. □

Lemma 3 *Let $n, t \in \mathbb{N}$ and $(k_1, \ldots, k_{n-1}) \in L_2$ be an arbitrary sample drawn according to algorithm 1. Then it is drawn uniformly, i.e.*

$$\mathbf{P}(k_1, \ldots, k_{n-1}) = \frac{1}{\binom{n+t-1}{n-1}}$$

for each $(k_1, \ldots, k_{n-1}) \in L_2$. The probability follows from Lemma 2.

Proof: The task is to distribute t stars and $n-1$ bars randomly on the positions $1, 2, \ldots, t+n-1$. We start from the right. Then the probability of placing a bar in position $t+n-1$ is $\frac{n-1}{t+n-1}$. Then we proceed conditionally to the left and multiply probabilities of placing bars or stars in order to calculate probabilities of a given configuration.

More precisely, let $(k_1, k_2, \ldots, k_{n-1})$ be an arbitrary sample drawn according to Algorithm 1. We want to calculate

$$\mathbf{P}(k_1, \ldots, k_{n-1}). \tag{12}$$

We know that the algorithm accepted $n-1$ integers (i.e., placements k_j of the bars) in the process. Also, let V denote the number of integers that were not accepted. In total the algorithm ran $V+n-1$ iterations. The probability (2) is a product of $V+n-1$ probabilities. Let us look at the denominator and nominator separately. In the denominator we have

$$(t+n-1) \cdot (t+n-2) \cdot \ldots \cdot (t-V-1). \tag{13}$$

In the numerator we have

$$(n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1 = (n-1)! \tag{14}$$

from the accepted integers. In the numerator there is also

$$t \cdot (t-1) \cdot \ldots \cdot (t-V-1) \tag{15}$$

from the integers that were not accepted. Combining (13), (14) and (15) we get

$$\begin{aligned} \mathbf{P}(k_1, k_2, \ldots, k_{n-1}) &= \frac{(n-1)! t(t-1)\cdots(t-V-1)}{(t+n-1)\cdots(t-V-1)} \\ &= \frac{(n-1)! t(t-1)\cdots(t-V-1)(t-V-2)\cdots 2\cdot 1}{(t+n-1)\cdots(t-V-1)(t-V-2)\cdots 2\cdot 1} \\ &= \frac{(n-1)!\ t!}{(t+n-1)!} \\ &= \frac{1}{\binom{t+n-1}{n-1}}. \end{aligned}$$

□

```
Data: $t$ and $n$
Result: $k_1, \ldots, k_{n-1}$
initialization;
$N = 0$ ;                                        // number of accepted integers
$V = 0$ ;                                        // number of not accepted integers
$I = t + n - 1$ ;                                // integer to consider
while $N < n - 1$ do
    Draw $p \sim U[0, 1]$;
    if $p < (n - 1 - N)/(t + n - 1 - N - V)$ then
        $k_{n-1-N} = I$ ;                        // integer $I$ was accepted
        $N = N + 1$;
        $I = I - 1$;
        Continue
    end
    if $p \geq (n - 1 - N)/(t + n - 1 - N - V)$ then
        $V = V + 1$ ;                            // integer $I$ was not accepted
        $I = I - 1$;
        Continue
    end
end
```

Algorithm 1: Draw $k_1, \ldots, k_{n-1}$ Uniformly

```
**Data:** Integer $V$, Data set $x$, sufficient statistic $T(x) = t$ and test statistic $D$
**Result:** $p^{\text{cond}}$
initialization;
count = 0;
**for** 1 *to* $V$ **do**
    Draw $y \sim X \mid T(X) = t$ ;
    **if** $D(y) \geq D(x)$ **then**
        count = count + 1;
    **end**
**end**
$p^{\text{cond}} = \text{count}/V$
```

Algorithm 2: Monte Carlo Conditional p-value

Conditional Sampling from the Negative Binomial Distribution

Let $Y_1 \sim \text{NB}(r_1, p), Y_2 \sim \text{NB}(r_2, p), \ldots, Y_n \sim \text{NB}(r_n, p)$ be independent random variables, where the parameters $r_1, \ldots, r_n$ are assumed to be known. Then $T(\mathbf{Y}) = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for p. In this subsection, we will show how the algorithm for the geometric distribution in Section 2.1 can be used to draw samples from the conditional distribution $\mathbf{Y} \mid T(\mathbf{Y}) = t$.

Note first that an argument like the one leading to (1) gives the following expression:

$$\mathbf{P}(Y_1 = y_1, \ldots, Y_n = y_n | \sum_{i=1}^{n} Y_i = t) = \frac{\prod_{i=1}^{n} \binom{y_i - r_i - 1}{y_i}}{\binom{t+R-1}{t}}$$

where $R = \sum_{i=1}^{n} r_i$.

The following shows that we can sample from this conditional distribution by using the algorithm for the geometric distriburion. Note first that we can write for $i = 1, \ldots, n$,

$$Y_i = \sum_{j=1}^{r_i} X_{ij}$$

where $X_{ij},\ i = 1, \ldots, n; j = 1, \ldots, r_i$ are i.i.d. from Geom(p). Then

$$\begin{aligned}
\mathbf{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = t) &= \mathbf{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid T(Y_1, Y_2, \ldots, Y_n) = t) \\
&= \mathbf{P}\left(\sum_{j=1}^{r_1} X_{1j} = y_1, \ldots, \sum_{j=1}^{r_n} X_{nj} = y_n \;\middle|\; \sum_{i=1}^{n}\sum_{j=1}^{r_i} X_{ij} = t\right) \\
&= \sum_{\substack{(x_{ij}) : \sum_j x_{1j} = y_1, \\ ,\ldots, \\ ,\sum_j x_{nj} = y_n}} \mathbf{P}\left(X_{ij} = x_{ij} \colon i = 1, \ldots, n,\ j = 1, \ldots, r_i \;\middle|\; \sum_{i,j} X_{ij} = t\right).
\end{aligned}$$

It follows from this that we can use the method for drawing samples from the conditional geometric distribution to draw condtional samples in the negative binomial case. More precisely, we can first draw a sample $x_1, x_2, \ldots x_R$, where $R = r_1 + r_2 + \ldots + r_n$, from the conditional distribution of $X_1, X_2, \ldots, X_R \mid T(\mathbf{X}) = t$, where $X_i \sim$ Geom(p) for $i = 1, 2, \ldots, R$ and let

$$y_1 = \sum_{i=1}^{r_1} x_i,\ y_2 = \sum_{i=r_1+1}^{r_1+r_2} x_i, \ldots, y_n = \sum_{i=r_1+\ldots+r_{n-1}+1}^{R} x_i.$$

We end up with a sample $\mathbf{y}$ from the desired conditional distribution $\mathbf{Y} \mid T(\mathbf{Y}) = t$.

For simulation in practice, notice that Algorithm 1 with input t and R gives numbers $k_1, k_2, \ldots, k_{R-1}$. Using the transformation ϕ in Lemma 1, it is then seen that we have

$$\begin{aligned}
y_1 &= k_{r_1} - r_1 \\
y_2 &= k_{r_1+r_2} - k_{r_1} - r_2 \\
&\vdots \\
y_i &= k_{r_1+\ldots+r_i} - k_{r_1+\ldots+r_{i-1}} - r_i \\
&\vdots \\
y_{n-1} &= k_{r_1+\ldots+r_{n-1}} - k_{r_1+\ldots+r_{n-2}} - r_{n-1} \\
y_n &= t - k_{r_1+\ldots+r_{n-1}} + R - r_n.
\end{aligned}$$

Some discrete distributions

$X \sim \mathrm{NB}(r,p)$	$\mathbf{P}(X=x)=\binom{x+r-1}{x}(1-p)^x p^r,\ x=0,1,2,\dots$
$X \sim \mathrm{Pois}(\lambda)$	$\mathbf{P}(X=x)=\frac{\lambda^x e^{-\lambda}}{x!},\ x=0,1,2,\dots$
$X \sim \mathrm{Geom}(p)$	$\mathbf{P}(X=x)=p(1-p)^x,\ x=0,1,2,\dots$
$X \sim \mathrm{Bin}(n,p)$	$\mathbf{P}(X=x)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,2,\dots,n$
$X \sim \mathrm{Ber}(p)$	$\mathbf{P}(X=x)=p^x(1-p)^{1-x},\ x=0,1$
$X \sim \mathcal{W}(q,\beta)$	$\mathbf{P}(X=x)=q^{x^\beta}-q^{(x+1)^\beta},\ x=0,1,2,\dots$
$\mathbf{X} \sim \mathrm{Mult}(t,\pi_1,\pi_2,\dots,\pi_n)$	$\mathbf{P}(\mathbf{X}=\mathbf{x})=\frac{n!\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_n^{x_n}}{x_1!x_2!\cdots x_n!},\ \sum_{i=1}^n x_i=t$
$X \sim \mathrm{BG}(\alpha,\beta)$	$\mathbf{P}(X=x)=\frac{B(\alpha+1,x+\beta)}{B(\alpha,\beta)},\ x=0,1,2,\dots$

Table 9: Distributions

References

Barlow, R. E., A. W. Marshall, F. Proschan, et al. (1963). Properties of probability distributions with monotone hazard rate. *The Annals of Mathematical Statistics 34*(2), 375–389.

Beltrán-Beltrán, J. I. and F. J. O'Reilly (2019). On goodness of fit tests for the Poisson, negative binomial and binomial distributions. *Statistical Papers 60*(1), 1–18.

Bracquemond, C., E. Crétois, and O. Gaudoin (2002). A comparative study of goodness-of-fit tests for the geometric distribution and application to discrete time reliability. *Laboratoire Jean Kuntzmann, Applied Mathematics and Computer Science, Technical Report*.

Bracquemond, C. and O. Gaudoin (2003). A survey on discrete lifetime distributions. *International Journal of Reliability, Quality and Safety Engineering 10*(01), 69–98.

D'Agostino, R. B. and M. A. Stephens (Eds.) (1986). *Goodness-of-fit Techniques*. New York, NY, USA: Marcel Dekker, Inc.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Volume 1. New York: Wiley.

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics 6*(1), 17–24.

González-Barrios, J. M., F. O'Reilly, and R. Rueda (2006). Goodness of fit for discrete random variables using the conditional density. *Metrika 64*(1), 77–94.

Heller, B. (1986). A goodness-of-fit test for the negative binomial distribution applicable to large sets of small samples. In *Developments in Water Science*, Volume 27, pp. 215–220. Elsevier.

Kyriakoussis, A., G. Li, and A. Papadopoulos (1998). On characterization and goodness-of-fit test of some discrete distribution families. *Journal of Statistical Planning and Inference 74*(2), 215–228.

Lindqvist, B. H., R. Erlemann, and G. Taraldsen (2020). Conditional Monte Carlo revisited. *Submitted and under revision.*

Lindqvist, B. H. and G. Taraldsen (2005). Monte Carlo conditioning on a sufficient statistic. *Biometrika 92*(2), 451–464.

Lockhart, R. A., F. O'Reilly, and M. Stephens (2009). Exact conditional tests and approximate bootstrap tests for the von mises distribution. *Journal of Statistical Theory and Practice 3*(3), 543–554.

Lockhart, R. A., F. J. O'Reilly, and M. A. Stephens (2007). Use of the gibbs sampler to obtain conditional tests, with applications. *Biometrika 94*(4), 992–998.

Nakagawa, T. and S. Osaki (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability 24*(5), 300–301.

Ozonur, D., E. Gökpinar, F. Gökpinar, and H. Bayrak (2013). Comparisons of the goodness of fit tests for the geometric distribution. *Gazi University Journal of Science 26*(3), 369–375.

Paul, S. R. (2005). Testing goodness of fit of the geometric distribution: an application to human fecundability data. *Journal of Modern Applied Statistical Methods 4*(2), 8.

Puig, P. and C. H. Weiß (2020). Some goodness-of-fit tests for the Poisson distribution with applications in biodosimetry. *Computational Statistics & Data Analysis 144*, 106878.

Rao, C. R. and I. Chakravarti (1956). Some small sample tests of significance for a Poisson distribution. *Biometrics 12*(3), 264–282.

Rueda, R., F. O. Reilly, and V. Perez-Abreu (1991). Goodness of fit for the Poisson distribution based on the probability generating function. *Communications in Statistics - Theory and Methods 20*(10), 3093–3110.

Singh, B., P. Pudir, and S. Maheshwari (2014). Parameter estimation of beta-geometric model with application to human fecundability data. *arXiv preprint arXiv:1405.6392*.

Spinelli, J. J. and M. A. Stephens (1997). Cramér-von Mises tests of fit for the Poisson distribution. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 25*(2), 257–268.

Vila, R., E. Y. Nakano, and H. Saulo (2019). Theoretical results on the discrete Weibull distribution of Nakagawa and Osaki. *Statistics 53*(2), 339–363.

Weinberg, C. R. and B. C. Gladen (1986). The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 547–560.

Wilf, H. S. (1999). East side, west side . . . - an introduction to combinatorial families-with maple programming. ”`https://www.math.upenn.edu/~wilf/lecnotes.html`”.

Cramér-von Mises Tests for Change Points

Rasmus Erlemann, Richard Lockhart and Rihan Yao

Submitted and under revision, available at

`https://arxiv.org/abs/2010.07072`

Cramér-von Mises tests for Change Points

Rasmus Erlemann
NTNU, Department of Mathematical Sciences
Richard Lockhart
SFU, Department of Statistics and Actuarial Science
Rihan Yao
SFU, Department of Statistics and Actuarial Science

September 25, 2020

Abstract

We study two nonparametric tests of the hypothesis that a sequence of independent observations is identically distributed against the alternative that at a single change point the distribution changes. The tests are based on the Cramér-von Mises two-sample test computed at every possible change point. One test uses the largest such test statistic over all possible change points; the other averages over all possible change points. Large sample theory for the average statistic is shown to provide useful p-values much more quickly than bootstrapping, particularly in long sequences. Power is analyzed for contiguous alternatives. The average statistic is shown to have limiting power larger than its level for such alternative sequences. Evidence is presented that this is not true for the maximal statistic. Asymptotic methods and bootstrapping are used for constructing the test distribution. Performance of the tests is checked with a Monte Carlo power study for various alternative distributions.

Keywords: Asymptotic Distribution; Change Point Detection; Cramér-von Mises Two-sample Test; Nonparametric Test Statistics; Monte Carlo Simulation.

1 Introduction

Consider a sequence of independent observations $X_1, \ldots, X_n$. We propose tests of the null hypothesis that the X_i are independent and identically distributed (iid) with unknown continuous distribution H against the change point alternative that there is some (unknown) c with $1 \le c < n$ such that $X_1, \ldots, X_c$ are iid with continuous distribution F and then $X_{c+1}, \ldots, X_n$ are iid with some other continuous distribution G. We will consider tests based on two sample empirical distribution function tests for equality of distribution, focusing on the two-sample Cramér-von Mises test.

If the time c of the potential change point were specified in advance we could test the hypothesis that $F = G = H$ using any two sample test for equality of two distributions. The two-sample Cramér-von Mises test is one well known possibility. Notation may be simpler to read if we used the shorthand $d = n - c$. Let

$$F_c(x) = \frac{1}{c} \sum_{i=1}^{c} 1(X_i \le x)$$

be the empirical distribution function of the first c observations and

$$G_d(x) = \frac{1}{d} \sum_{i=c+1}^{n} 1(X_i \le x)$$

be the empirical distribution function of the remaining d observations. The combined empirical distribution function H_n of the entire sample is

$$H_n(x) = \frac{cF_c(x) + dG_d(x)}{n}.$$

The two-sample Cramér-von Mises test of the hypothesis $F = G$ is based on the statistic

$$W_n(c) = \frac{cd}{n} \int_{-\infty}^{\infty} \{F_c(x) - G_d(x)\}^2 \, dH_n(x).$$

For a thorough discussion of this nonparametric test and a simple computing formula in terms of the ranks of the first c values of X in the whole sample see Anderson (1962). The distribution of the test statistic does not depend on H under the null hypothesis provided H is a continuous function.

A number of authors have suggested adapting this statistic to the change point problem. See, for instance, Picard (1985) and Brodsky and Darkhovsky (1993) where the two natural possible test statistics considered herein are suggested and studied briefly. The first of these tests can be used both to assess the existence of a change point and to estimate the location of the change if it exists. The statistic in question is

$$W_{\max} \equiv \max_{1 \le c \le n-1} W_n(c).$$

We shall also use $W_{\max}$ to define the estimated change point

$$\hat{c}_n = \operatorname*{arg\,max}_{1 \le c \le n-1} W_n(c);$$

thus $\hat{c}_n$ is the value of c achieving the maximum. (We remark that the statistic W_n is discrete and in small samples there is some modest probability that $\hat{c}_n$ will not be unique; this lack of uniqueness plays no role in the hypothesis testing problem.)

We prefer, however, the statistic

$$\overline{W}_n(X_1, \ldots, X_n) = \overline{W}_n \equiv \frac{1}{n-1} \sum_{c=1}^{n-1} W_n(c).$$

We offer several potential rationales for our choice:

- In many goodness-of-fit contexts quadratic statistics like ours outperform maximal statistics. For instance, the Cramér-von Mises goodness-of-fit test is generally more powerful than the Kolmogorov-Smirnov test; see, for instance, Stephens (1986).

- Quadratic statistics such as we propose often have simpler large sample theory than do maximal statistics like the Kolmogorov-Smirnov test. Generally speaking the former have limiting distributions which are linear combination of chi-squares while the latter have limiting laws which are those of the supremum of a Gaussian process. The actual laws of these suprema are known only in special cases (although inequalities can often provide useful upper bounds on p-values).

- The large sample theory in question often provides a more accurate approximation for quadratic statistics than it does for maximal statistics. For example, see Mohd Razali and Yap (2011) and Büning (2002).

In Section 2 we present large sample distribution theory under the null hypothesis, show how to compute p-values based on this large sample theory and demonstrate that the asymptotic approximations are quite accurate for $n \geq 100$, particularly in the important lower tail. Section 3 presents a short power study showing that over a wide range of alternatives the statistic $\bar{W}$ is more powerful than $W_{\max}$. Section 4 presents asymptotic power calculations against contiguous sequences of alternatives; these permit useful approximations to the power of $\bar{W}$ in cases where the null is not obviously false. By contrast, the limit theory for $W_{\max}$ does not lend itself to easy power calculations. We conjecture, however, that in this context of contiguous alternatives the statistic $W_{\max}$ has the defect that, unlike $\bar{W}$, its power converges to its level. In this section we present some further Monte Carlo studies relevant to contiguous sequences of alternatives. Finally we present some discussion in Section 6. We give proofs and evidence for the conjecture in the Appendix.

2 Null limit theory

Suppose that the null hypothesis holds and the $X_1, \ldots, X_n$ are iid with *continuous* cdf H. Then for all c we have

$$W(X_1, \ldots, X_c, X_{c+1}, \ldots, X_n) = W(H(X_1), \ldots, H(X_c); H(X_{c+1}), \ldots, H(X_n)).$$

Thus in computing distribution theory under the null we may, and will, assume that H is the uniform distribution; to emphasize the point we let $U_1, U_2, \cdots$ be an iid sequence of Uniform random variables; the joint law of $(H(X_1), \ldots, H(X_n))$ is the same as that of $(U_1, \ldots, U_n)$.

Large sample theory for the two sample Cramér-von Mises statistic is well known: if c depends on n in such a way that $c/n \to s \in (0,1)$ (or even just $\min\{c, n-c\} \to \infty$) then

$$W_n(c) \Rightarrow \sum_{j=1}^{\infty} \frac{Z_j^2}{\pi^2 j^2}$$

where the Z_i are iid standard normal; see Anderson (1962). (Notice that the limit is free of s.) Our statistic has a related limit given as follows.

Theorem 1 *As $n \to \infty$ we have, under the null hypothesis,*

$$\overline{W}_n \Rightarrow \overline{W}_\infty \equiv \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \frac{Z_{jk}^2}{j(j+1)\pi^2 k^2}$$

where the Z_{jk} are iid standard normal.

The theorem is a consequence, as usual, of a suitable weak convergence result which we now present; the Gaussian process limit we derive is mentioned in Picard (1985); the specific weights in Theorem 1 do not seem to have been previously described.

We begin by defining the partial sum empirical process (van der Vaart and Wellner, 1996, p. 225), for $(s,t) \in [0,1]^2$, by

$$\mathbb{Z}_n(s,t) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le ns} \left\{ 1(U_i \le t) - t \right\}.$$

Our statistic can be described in terms of this process. Notice that

$$F_c(t) = \frac{\sqrt{n}}{c} \mathbb{Z}_n(c/n, t) + t$$

and that

$$G_d(t) = \frac{\sqrt{n}}{d} \left\{ \mathbb{Z}_n(1,t) - \mathbb{Z}_n(c/n,t) \right\} + t.$$

Thus

$$F_c(t) - G_d(t) = \sqrt{n} \left\{ \frac{\mathbb{Z}_n(c/n,t)}{c} - \frac{\mathbb{Z}_n(1,t) - \mathbb{Z}_n(c/n,t)}{d} \right\}.$$

We now define a process $\mathbb{W}_n(s,t)$ for $0 < s < 1$ and $0 \le t \le 1$ by

$$\mathbb{W}_n(s,t) = \sqrt{s(1-s)} \left\{ \frac{\mathbb{Z}_n(s,t)}{s} - \frac{\mathbb{Z}_n(1,t) - \mathbb{Z}_n(s,t)}{1-s} \right\} = \frac{\mathbb{Z}_n(s,t) - s\mathbb{Z}_n(1,t)}{\sqrt{s(1-s)}}.$$

For given c our two sample test statistic is given by

$$W_n(c) = \int_0^1 \left\{ \mathbb{W}_n(c/n,t) \right\}^2 \, dH_n(t).$$

The processes $\mathbb{Z}_n$ and $\mathbb{W}_n$ have well known weak limits given the in following theorem. It will also prove useful to introduce the notation

$$\mathbb{B}_n(s,t) = \mathbb{Z}_n(s,t) - s\mathbb{Z}_n(1,t).$$

Theorem 2 *Under the null hypothesis:*

1. *As $n \to \infty$,*

$$\mathbb{Z}_n(s,t) \rightsquigarrow \mathbb{Z}_\infty$$

a mean 0 Gaussian Process with covariance function

$$\rho_Z(s,t;s',t') = s \wedge s' \psi(t,t')$$

where $\psi(t,t') = t \wedge t' - tt'$;

2. As $n \to \infty$,

$$\mathbb{B}_n(s,t) \rightsquigarrow \mathbb{B}_\infty$$

a mean 0 Gaussian Process with covariance function

$$\rho_B(s,t;s',t') = \psi(s,s')\psi(t,t');$$

3. As $n \to \infty$,

$$\mathbb{W}_n(s,t) \rightsquigarrow \mathbb{W}_\infty$$

a mean 0 Gaussian Process with covariance function

$$\rho_W(s,t;s',t') = \chi(s,s')\psi(t,t')$$

where

$$\chi(s,s') = \frac{\psi(s,s')}{\sqrt{s(1-s)s'(1-s')}}.$$

The process $\mathbb{B}$ is called a Brownian pillow by some writers or a 4 side tied down Brownian motion; see, for instance Zhang (2014) or McKeague and Sun (1996). The process $\mathbb{Z}$ is a Blum-Kiefer-Rosenblatt process ; see Blum et al. (1961).

We now record well known facts about the eigenvalues of the covariance ρ_W. The covariance kernel ψ is that of a Brownian Bridge. It has eigenvalues of the form $1/(\pi^2 k^2)$ for $k = 1, 2, \cdots$ with corresponding orthonormal eigenfunctions $f_{\psi,k}(u) = \sqrt{2}\sin(\pi k u)$. The covariance kernel χ arises in the study of the Anderson-Darling goodness-of-fit test. It has eigenvalues of the form $1/\{j(j+1)\}$ for $j = 1, 2, \cdots$. The corresponding orthonormal eigenfunctions are associated Legendre functions. The j^{th} eigenfunction is

$$f_{\chi,j}(u) = 2\sqrt{\frac{2j+1}{j(j+1)}}\sqrt{s(1-s)}q_j(2s-1)$$

where the q_j are polynomials of degree $j-1$ defined recursively as follows:

$$q_1(u) = 1,$$

$$q_2(u) = 3u$$

and for $j \geq 2$

$$q_{j+1}(u) = \frac{1}{j}\left\{(2j+1)uq_j(u) - (j+1)q_{j-1}(u)\right\}.$$

It follows that the eigenvalues of ρ_W consist of all possible products

$$\lambda_{jk} = \frac{1}{j(j+1)\pi^2 k^2}$$

with corresponding eigenfunctions

$$f_{\chi,j}(s)f_{\psi,k}(t).$$

The expansion in Theorem 1 is then Parseval's identity with

$$Z_{jk} = \int_0^1 \int_0^1 \mathbb{W}(s,t)f_{\chi,j}(s)f_{\psi,k}(t)\,ds\,dt.$$

2.1 Numerical Work

The distribution of $\overline{W}_\infty$ can be computed numerically in order to provide approximate, asymptotically valid, p-values. Our desired approximation to the p-value is

$$P(\overline{W}_n > w_{\text{obs}}) \approx P(\overline{W}_\infty > w_{\text{obs}})$$

where w_{obs} is the value of $\overline{W}_n$ observed in the data. Define

$$\lambda_{jk} = \frac{1}{\pi^2 j(j+1)k^2}.$$

In practice, we truncate the infinite sum defining $\overline{W}_\infty$, retaining the terms with the largest values of λ_{jk}, and replace the neglected terms by their expected value. So we write

$$\begin{aligned}\overline{W}_\infty &= \overline{W}_M + T_M \\ &= \sum_{jk\le M} \lambda_{jk} Z_{jk}^2 + \sum_{jk>M} \lambda_{jk} Z_{jk}^2.\end{aligned}$$

We then approximate T_M by its expected value:

$$\mu_M \equiv \sum_{jk>M} \lambda_{jk} \mathrm{E}\left(Z_{jk}^2\right) = \sum_{jk>M} \lambda_{jk}.$$

Since the mean of $\overline{W}_\infty$ is

$$\sum_{j,k} \lambda_{jk} = \frac{1}{6}$$

the mean of T_M may be computed by

$$\frac{1}{6} - \sum_{jk\le M} \lambda_{jk}.$$

Our approximation becomes

$$P(\overline{W}_n > w_{\text{obs}}) \approx P(\overline{W}_M + \mu_M > w_{\text{obs}}).$$

The latter quantity may now be computed by using numerical Fourier inversion following Imhof (1961). The `R` package `CompQuadForm` (see Duchesne and Lafaye de Micheaux, 2010) implements this computation in the function `imhof`; we use this software in our numerical work below.

We have evaluated the quality of our asymptotic approximation to the null distribution of $\overline{W}$ in a small Monte Carlo study. Since this distribution does not depend on H when the null hypothesis holds we generated $N = 10,000$ samples of size $n = 200, 500, 1000$. Figure 1 shows a Q-Q plot for these 10,000 values for $n = 200$ to check the uniformity of their distribution. Specifically, we plot the order statistics against the uniform plotting points $1/(N+1), \ldots, N/(N+1)$. Figure 2 is an enlargement of the smallest 10% of these values since the quality of the approximation is most important for small p-values. In both cases it is seen that the approximation is excellent. For completeness, however, we note that the hypothesis of exact uniformity of these 10,000 p-values is rejected ($P \approx 0.01$) by the Anderson-Darling test. Applied to the smallest 1,000 p-values, rescaled so that p-value number 1,001 from the bottom becomes 1, the Anderson-Darling p-value is actually 0.99. We conclude the uniform approximation is very good at reasonable sample sizes, particularly in the important lower tail. For p-values over 0.5 we believe that the truncation we must do in order to compute the limit law is slightly off but argue that inaccuracy in the upper tail of p-values is not very consequential.

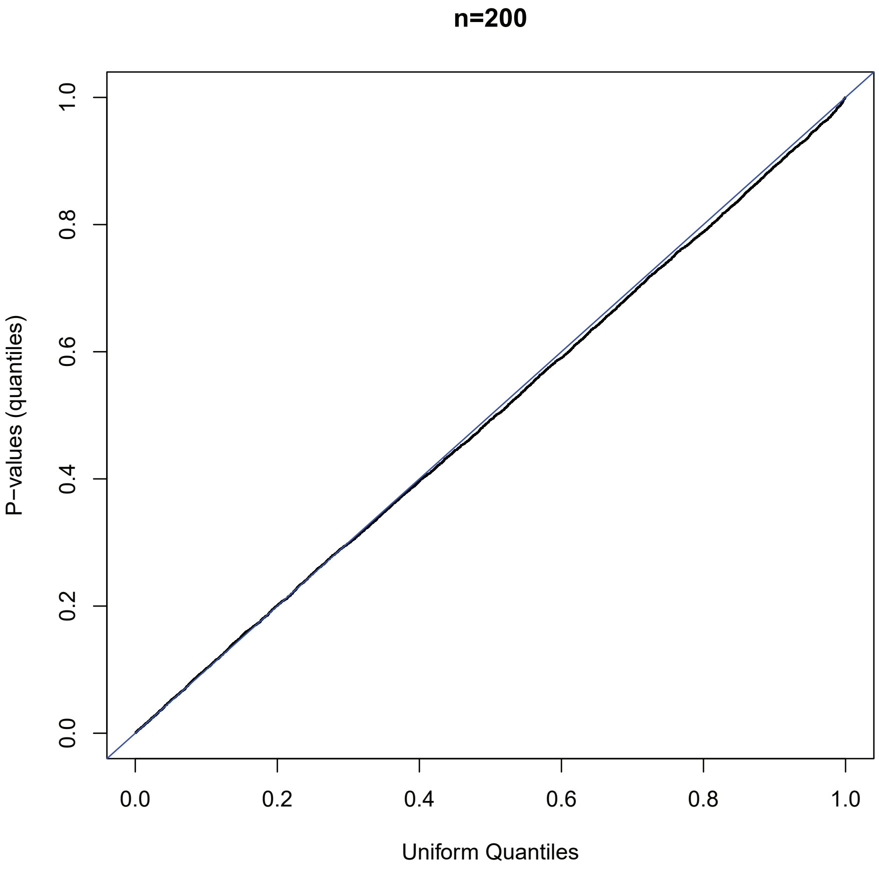

Figure 1: Ordered p-values plotted against uniform quantiles for 10,000 iid Monte Carlo samples from a continuous distribution. The blue line is the uniform cumulative distribution function; exact p-values have a uniform distribution; the graph shows this approximation is good.

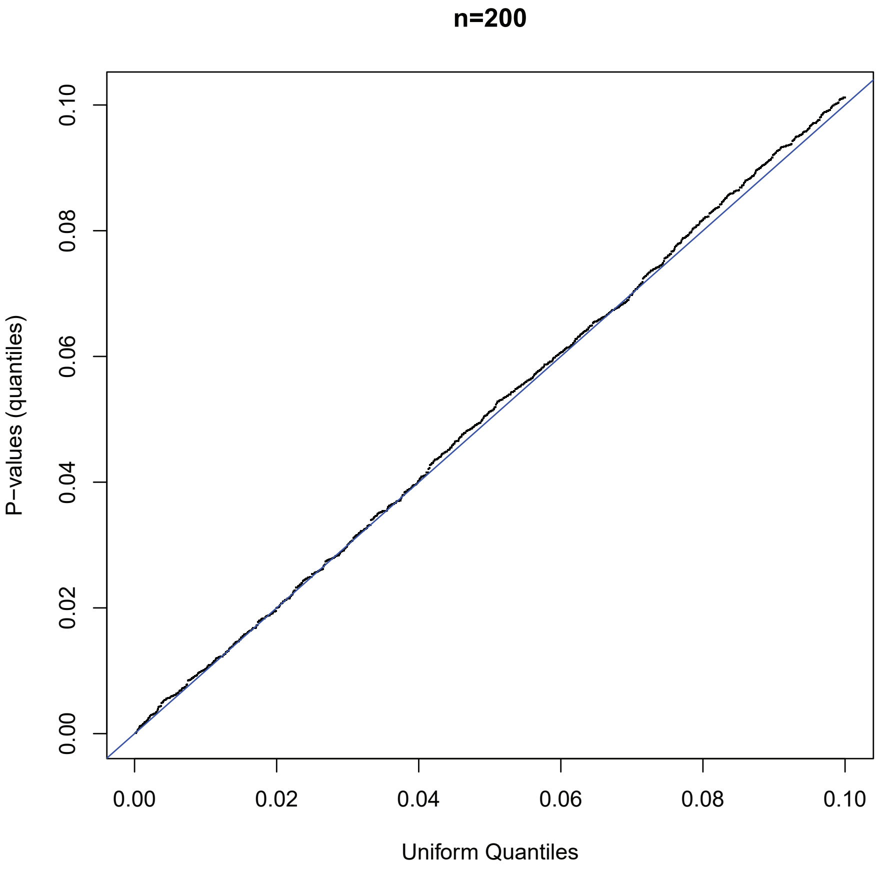

Figure 2: Exploded view of Figure 1 showing the lower 10% of the distribution of the ordered p-values plotted against uniform quantiles for 10,000 iid Monte Carlo samples from a continuous distribution. The blue line is the uniform cumulative distribution function; exact p-values have a uniform distribution; the graph shows this approximation is very good in the important lower tail.

3 Monte Carlo Power approximations

We undertook a variety of Monte Carlo simulation studies to compare the power of $\overline{W}_n$ to $W_{\max}$. In Table 3 we show the percentage of samples rejected in 10,000 trials by the two methods at the levels $\alpha = 0.05$ and $\alpha = 0.1$. We consider samples of size $n \in \{20, 50, 100\}$. In one experiment recorded in the table we generated data from the Gamma distributions where the parameters change at $c = n/2$. In another experiment we change from the Gamma distribution to the Normal distribution at $c = n/2$; in this case neither the mean nor the variance changes. While our tests are designed to detect single change points we have included two trials in which there are three segments which change between various Gamma distributions. One changes from shape 1, scale 2 to shape 2, scale 1 at the 40% point and then to shape 0.5, scale 4 at the 60% point. All three of these have the same mean. The other changes from shape 1, scale 2 to shape 2, scale 3, and back to shape 1, scale 2; the changes happen after 30% and then 70% of the data. Finally we present two experiments with samples from the normal distribution; in one the mean changes at $c = n/2$ and in the other the standard deviation changes at the same point. In all these trials the parameter values in the distributions in a given segment do not change as the sample size changes; this may be compared with the further Monte Carlo results in Section 4.

		$\alpha = 0.1$		$\alpha = 0.05$	
Alternative	Sample size	$W_{\max}$	$\overline{W}_n$	$W_{\max}$	$\overline{W}_n$
$X_1, \ldots, X_{0.5n} \sim \text{Gamma}(1,2)$, $X_{0.5n+1}, \ldots, X_n \sim \text{Gamma}(2,2)$	$n = 20$	47.9	50.7	35.0	37.5
	$n = 50$	82.3	85.7	73.9	77.4
	$n = 100$	98.3	98.9	96.3	96.9
$X_1, \ldots, X_{0.5n} \sim \text{Gamma}(1,2)$, $X_{0.5n+1}, \ldots, X_n \sim \mathcal{N}(2,2)$	$n = 20$	12.9	13.7	6.9	7.2
	$n = 50$	16.1	19.2	9.0	11.2
	$n = 100$	22.1	31.2	13.7	19.0
$X_1, \ldots, X_{0.4n} \sim \text{Gamma}(1,2)$, $X_{0.4n+1}, \ldots, X_{0.6n} \sim \text{Gamma}(2,1)$ $X_{0.6n+1}, \ldots, X_n \sim \text{Gamma}(0.5,4)$	$n = 20$	17.5	16.5	10.0	9.2
	$n = 50$	24.6	25.5	15.5	15.9
	$n = 100$	38.3	42.8	27.3	28.5
$X_1, \ldots, X_{0.3n} \sim \text{Gamma}(1,2)$, $X_{0.3n+1}, \ldots, X_{0.7n} \sim \text{Gamma}(2,3)$ $X_{0.7n+1}, \ldots, X_n \sim \text{Gamma}(1,2)$	$n = 20$	29.0	20.6	15.8	7.9
	$n = 50$	72.3	71.6	54.4	48.1
	$n = 100$	98.3	98.6	94.1	94.6
$X_1, \ldots, X_{0.5n} \sim \mathcal{N}(0,1)$, $X_{0.5n+1}, \ldots, X_n \sim \mathcal{N}(0,3)$	$n = 20$	18.2	22.0	10.8	11.3
	$n = 50$	29.6	56.0	17.0	33.0
	$n = 100$	66.3	93.4	45.0	81.2
$X_1, \ldots, X_{0.5n} \sim \text{Exp}(1)$, $X_{0.5n+1}, \ldots, X_n \sim \text{Exp}(1.5)$	$n = 20$	15.8	16.4	9.1	9.3
	$n = 50$	23.4	26.9	14.9	17.5
	$n = 100$	35.8	42.7	25.0	31.0

Table 1: Powers (percentage) from various alternative distributions and significance levels 0.1 and 0.05. Critical points were calculated with $100,000$ and Powers by $10,000$ Monte Carlo simulations. The notation Gamma(α, β) indicates sampling from a Gamma distribution with shape α and scale β. The parameters in the normal distribution are mean and variance as usual. The parameter in the Exponential distribution is the mean.

It will be seen that, except for very small samples, when there is a single change point the test using $\overline{W}_n$ has better power than $W_{\max}$. Since it is also far faster to compute p-values for $\overline{W}_n$ using the highly accurate asymptotic law we recommend $\overline{W}$ over $W_{\max}$. At the same time we observe that the procedure is specifically designed to choose between 1 change point and no change points and not to estimate and find multiple change points. In particular, for one of the alternatives in Table 3 with 2 change points the statistic $W_{\max}$ is usually more sensitive than $\overline{W}_n$.

The results presented here show how the powers grow with sample size when the two distributions are fixed. Other experiments, not reported here, show that both statistics have better power when the change is near the center of the sequence.

More Monte Carlo power calculations are presented in Section 5 below with a focus on contiguous alternatives.

4 Power approximations: contiguous alternatives

We now compute approximate distribution theory for $\bar{W}_n$ when the null hypothesis is false and the extent of the change at the change point is big enough to be detectable but not obvious; that is, we study situations where the best possible power in large samples stays away from 1. To do so we consider a sequence of alternatives indexed by n and assume that these alternatives are contiguous to a sequence for which the null hypothesis of no change holds. To be specific our null hypothesis sequence will have X_i iid for $1 \le i \le n$ with density h and cdf H. For the alternative we suppose that there is a value c_0 such that for $1 \le i \le c_0$, the X_i are iid with density f and that for $c_0 + 1 \le i \le n$ the X_i are iid with density g. All of f, g, h, and the true change point c_0 may depend on n but the dependence will be hidden in our notation. Under the null hypothesis the joint density of $X_1, \ldots, X_n$ is

$$\mathbf{f}_{0n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} h(x_i).$$

Under the alternative the joint density becomes

$$\mathbf{f}_{1n}(x_1, \ldots, x_n) = \prod_{i=1}^{c_0} f(x_i) \prod_{c_0+1}^{n} g(x_i).$$

The log-likelihood ratio of these two is

$$\begin{aligned}\Lambda_n &= \ln\left\{\mathbf{f}_{1,n}(X_1, \ldots, X_n)/\mathbf{f}_{0n}(X_1, \ldots, X_n)\right\} \\ &= \sum_{i=1}^{c_0} \ln\left\{f(X_i)/h(X_i)\right\} + \sum_{i=c_0+1}^{n} \ln\left\{g(X_i)/h(X_i)\right\}.\end{aligned}$$

The sequence of alternatives $\mathbf{f}_{1n}$ is contiguous to the null sequence $\mathbf{f}_{0n}$ if, computing under the null hypothesis, we have

$$\Lambda_n \rightsquigarrow N(-\tau^2/2, \tau^2) \tag{1}$$

for some $0 \le \tau < \infty$. If we define $U_i = H(X_i)$ then under the null hypothesis the U_i are iid Uniform[0,1]. Under the alternative $U_1, \ldots, U_{c_0}$ are iid with density

$\tilde{f}(u) = f(H^{-1}(u))/h\left(H^{-1}(u)\right)$ while $U_{c_0+1},\ldots,U_n$ are iid with density $\tilde{g}(u) = g(H^{-1}(u))/h\left(H^{-1}(u)\right)$. The likelihood ratio becomes

$$\tilde{\Lambda}_n = \sum_{i=1}^{c_0} \ln\left\{\tilde{f}(U_i)\right\} + \sum_{i=c_0+1}^{n} \ln\left\{\tilde{g}(U_i)\right\}.$$

Since our test statistics are invariant to a monotone transformation applied to each individual data point we will take H to be Uniform[0,1] and then drop the tildes from our notation. The quantity

$$S_n = \sum_{i=1}^{c_0} \phi_f(X_i)/\sqrt{n} + \sum_{i=c_0+1}^{n} \phi_g(X_i)/\sqrt{n}$$

is needed in our theorem.

Theorem 3 *Assume*

A1 *There are two functions ϕ_f and ϕ_g in $L_2[0,1]$ such that*

$$\lim_{n\to\infty} \sqrt{n}(f-1) = \phi_f$$

and

$$\lim_{n\to\infty} \sqrt{n}(g-1) = \phi_g.$$

A2 *There is a $u \in (0,1)$ such that*

$$\lim_{n\to\infty} \frac{c_0}{n} = u.$$

Then as $n \to \infty$ we have, under the sequence of alternative hypotheses specified by f, g, and c,

1. *The log-likelihood ratio satisfies*

$$\Lambda_n = S_n + o_P(1) \rightsquigarrow N(-\tau^2/2, \tau^2)$$

where

$$\tau^2 = u\int_0^1 \phi_f^2(t)\,dt + (1-u)\int_0^1 \phi_g^2(t)\,dt.$$

2. *The process $\mathbb{W}_n$ converges weakly to a Gaussian process with covariance ρ and mean*

$$\mu(s,t) = \mu_\chi(s)\mu_\psi(t)$$

where

$$\mu_\chi(s) = \sqrt{s(1-s)}\left\{\frac{1-u}{1-s}1(s \le u) + \frac{u}{s}1(s > u)\right\}$$

and

$$\mu_\psi(t) = \left[\mathrm{E}\left\{\phi_f(U)1(U \le t)\right\} - \mathrm{E}\left\{\phi_g(U)1(U \le t)\right\}\right].$$

3. *and*

$$\overline{W}_n \rightsquigarrow \overline{W}_\infty \equiv \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{\left(Z_{jk}+\eta_j\tau_k\right)^2}{j(j+1)\pi^2 k^2}$$

where the Z_{jk} are iid standard normal,

$$\eta_j = \int_0^1 \mu_\chi(s) f_{j,\chi}(s)\,ds,$$

and

$$\tau_k = \int_0^1 \mu_\psi(t) f_{j,\psi}(t)\,dt.$$

As with the null distribution, this limiting alternative distribution for $\overline{W}$ can be computed using the `R` package `CompQuadForm`. As an example we take f to be standard normal and g to be normal with mean μ and standard deviation σ. The two parameters are assumed to depend on n in such a way that

$$\sqrt{n}\mu \to \gamma_1 \quad \text{and} \quad \sqrt{n}(\sigma-1) \to \gamma_2.$$

It is convenient to take $h = f$. Under the null the data $X_1, \dots, X_n$ are iid standard normal. The functions $\tilde{f}$ and $\tilde{g}$ are then given by $\tilde{f} \equiv 0$ and

$$\tilde{g}(u) = \frac{\phi\left\{\frac{\Phi^{-1}(u)-\mu}{\sigma}\right)}{\phi\left\{\Phi^{-1}(u)\right\}}.$$

Under these conditions we may check that condition **A1** holds with $\phi_f = 0$ and

$$\phi_g(u) = \gamma_2\left[\left\{\Phi^{-1}(u)\right\}^2 - 1\right] + \gamma_1\Phi^{-1}(u).$$

5 Large sample behaviour of $W_{\max}$

The statistic $W_{\max}$ is more challenging to analyze because the weak convergence result in Theorem 2 asserts convergence in $\ell_\infty^{\mathrm{loc}}((0,1)\times[0,1])$. By $\ell_\infty^{\mathrm{loc}}((0,1)\times[0,1])$ we mean the space of functions on $(0,1)\times[0,1]$ which are bounded on compact subsets of their domain. We give this the topology of uniform convergence on compacts. See van der Vaart and Wellner (1996). Our proof of Theorem 1 shows that our statistic is a continuous function on a subset of $\ell_\infty^{\mathrm{loc}}((0,1)\times[0,1])$ to which sample paths of $\mathbb{W}_\infty$ are almost sure to belong. We are not able to establish the corresponding result for $W_{\max}$. Traditionally this problem has been handled either by fixing a small $\epsilon > 0$ and redefining $W_{\max}$ by maximizing only over $\{c : \epsilon \le c/n \le 1-\epsilon\}$ or by careful analysis of the behaviour of the process and the

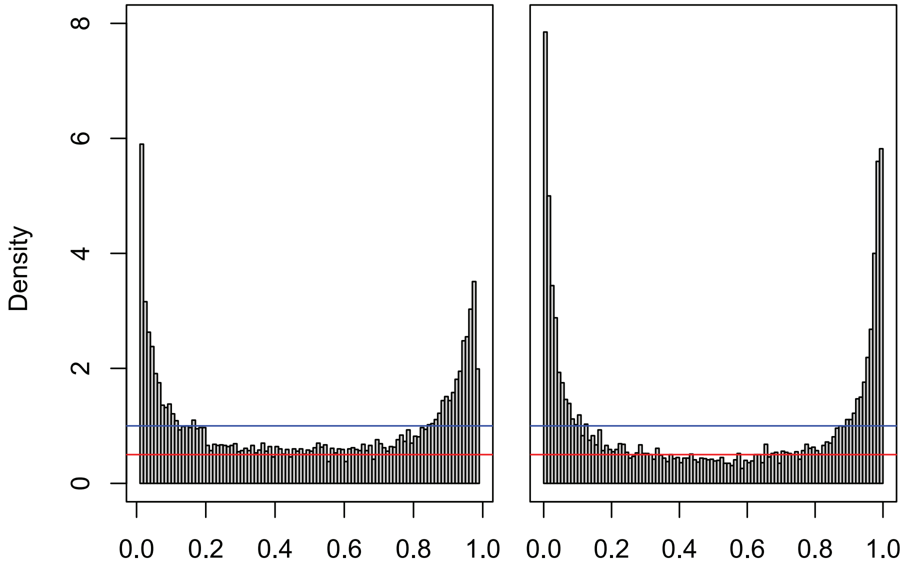

Figure 3: Histograms of values of estimated change points for sample sizes $n = 100$ on the left and $n = 500$ on the right. The null hypothesis is true and 10,000 samples were used for each histogram. The x-axis shows $\hat{c}/n$ and the y-axis is a probability density scale. The two figures have the same scales on each axis. Horizontal lines at height 1 (blue) and 0.5 (red) are provided to help see the extent to which the distribution on the right is more concentrated around 0 and 1 than the distribution on the left.

test statistic for c/n close to 0 or to 1. For instance, Jaeschke (1979) considers a weighted Kolmogorov-Smirnov test for the uniform distribution and shows that the supremum of the weighted empirical process has, after suitable normalization, an extreme value distribution.

We have not pursued either of these ideas but offer here some evidence that this statistic has some important defects. First we look at a small simulation study. We generated 10,000 samples of size 100 and 500 from the null hypothesis. In Figure 3 we plot histograms of the value $\hat{c}$ which maximizes $W_n(c)$ over $1 \leq c \leq n - 1$. Observe that as the sample size grows the histogram concentrates near 0 and 1 (though the convergence is slow). We can prove:

Proposition 1 *Under the null hypothesis and under any sequence of contiguous alternatives*

$$\min\left\{\frac{\hat{c}}{n}, \frac{n-\hat{c}}{n}\right\} \to 0$$

in probability. Under the null hypothesis, the distribution of $\hat{c}/n$ converges to a Bernoulli(0.5) *law.*

This means that, even for data from detectable (but not obvious) alternatives, our test statistic $W_{\max}$ usually compares the distribution of a tiny fraction of the data to that of the vast majority of the data even when the true change point is in the middle of the sequence. We also conjecture:

Conjecture 1 *For any sequence of contiguous alternatives the difference between the power and the level of a test based on $W_{\max}$ goes to 0 as $n \to \infty$.*

Gamma, shape=$1 + b/\sqrt{n}$, break at $n/2$

			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
	$\bar{W}$	MC	11.70	13.96	14.83	14.71	15.91
$b = 2$	$\bar{W}$	Asym	11.79	13.59	14.61	14.67	15.70
	$W_{\max}$	MC	12.13	12.00	12.36	11.41	11.80
	$\bar{W}$	MC	18.50	25.18	26.48	27.74	29.52
$b = 3$	$\bar{W}$	Asym	18.72	24.73	26.12	27.66	29.25
	$W_{\max}$	MC	18.62	22.05	21.84	21.34	21.88
	$\bar{W}$	MC	34.95	52.67	57.39	61.28	65.62
$b = 5$	$\bar{W}$	Asym	35.26	51.97	57.06	61.18	65.35
	$W_{\max}$	MC	35.48	47.60	50.07	52.76	54.46

Gamma, shape=$1 + b/\sqrt{n}$, break at $3n/10$

	$\bar{W}$	MC	9.24	11.29	11.73	11.83	13.21
$b = 2$	$\bar{W}$	Asym	9.42	10.86	11.47	11.79	13.10
	$W_{\max}$	MC	10.00	10.60	10.48	9.86	10.37
	$\bar{W}$	MC	13.41	20.04	20.26	21.80	23.15
$b = 3$	$\bar{W}$	Asym	13.54	19.56	19.92	21.66	22.98
	$W_{\max}$	MC	14.81	18.07	17.38	18.00	17.97
	$\bar{W}$	MC	22.42	41.53	45.54	48.59	53.34
$b = 5$	$\bar{W}$	Asym	22.75	40.87	45.11	48.52	53.07
	$W_{\max}$	MC	26.43	39.44	41.36	43.09	45.72

Table 2: Powers (percentage) for change from Gamma(shape= $1+b/\sqrt{n}$, scale=1) to Gamma(1,1) at the indicated breakpoint, $n/2$ in the top and $3n/10$ in the bottom. Powers are based on 10,000 samples and either use Monte Carlo critical points (based on 100,000 samples) or asymptotic critical points as indicated by 'MC' or 'Asym'. All tests are at the level $\alpha = 0.05$.

Normal, $\sigma = 1 + b/\sqrt{n}$, break at $n/2$

			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
$b = 2$	$\bar{W}$	MC	5.61	5.97	5.65	5.66	5.91
	$\bar{W}$	Asym	5.69	5.77	5.40	5.61	5.83
	$W_{\max}$	MC	6.70	5.72	5.19	4.80	5.25
$b = 3$	$\bar{W}$	MC	6.11	7.04	6.87	6.75	7.40
	$\bar{W}$	Asym	6.20	6.66	6.66	6.73	7.23
	$W_{\max}$	MC	7.67	6.49	5.71	5.36	5.55
$b = 5$	$\bar{W}$	MC	6.76	9.55	11.10	11.32	13.56
	$\bar{W}$	Asym	6.79	9.24	10.79	11.25	13.33
	$W_{\max}$	MC	8.99	7.91	6.99	6.84	6.88

Normal, $\sigma = 1 + b/\sqrt{n}$, break at $0.3n/10$

$b = 2$	$\bar{W}$	MC	6.26	6.49	5.80	5.63	5.76
	$\bar{W}$	Asym	6.37	6.17	5.63	5.63	5.68
	$W_{\max}$	MC	7.12	6.08	5.72	5.22	5.42
$b = 3$	$\bar{W}$	MC	6.91	7.37	6.74	6.41	6.95
	$\bar{W}$	Asym	7.09	7.10	6.51	6.39	6.80
	$W_{\max}$	MC	8.18	7.08	6.29	5.94	5.95
$b = 5$	$\bar{W}$	MC	7.89	9.40	9.92	9.91	11.13
	$\bar{W}$	Asym	8.09	8.99	9.65	9.79	10.98
	$W_{\max}$	MC	9.80	8.96	8.04	7.67	7.19

Table 3: Powers (percentage) for change from Normal(0,$\sigma = 1 + b/\sqrt{n}$) to Normal(0,1) at the indicated breakpoint, namely, $n/2$ in the top and $3n/10$ in the bottom. Powers are based on 10,000 samples and either use Monte Carlo critical points (based on 100,000 samples) or asymptotic critical points as indicated by 'MC' or 'Asym'. All tests are at the level $\alpha = 0.05$.

Here is some Monte Carlo evidence from a simulation study. In Tables 2 and 3 we study four alternatives at sample sizes $n = 10, 50, 100, 200, 500$. For each sample size we draw 10,000 samples of size n. The first c observations in each sample have some parameter of the form $a + b/\sqrt{n}$ and the remaining $n - c$ have parameter a. We used the Gamma distribution and the normal distribution and tried $c = 0.5n$ and $c = 0.3n$ for each distribution. In the Gamma case we tried changing the shape parameter with $a = 1$ while holding the scale parameter at 1. The tables show the expected convergence (although we have not computed the power predicted by our theory in Section 4.

For the statistic $W_{\max}$ the tables show, in the normal case, the power declining towards the level (which is 5% here). For the Gamma cases studied here the power is rising but slowly for distant alternatives (large values of b) and declining very slowly for less distant alternatives (smaller values of b). Our experience in general is that for more distant alternatives it requires larger sample sizes before the power of $W_{\max}$ begins to drop.

Our conjecture is motivated by an analogy with Lockhart (1991) in which it is shown that goodness-of-fit test statistics which depend only on $o(n)$ tail order statistics have the property asserted in the second conjecture. In the Appendix we prove the proposition and provide partial details showing how we would hope to prove our conjecture, if we could.

6 Discussion

It is a general principle that procedures with optimal frequency properties are found by searching among Bayes procedures. It is also generally the case that optimal Bayes procedures involve averaging rather than maximizing. These heuristics motivate considering testing for change points by using test statistics which are averages over possible change points rather than maxima. In this paper we have used this heuristic to motivate an average two sample goodness of fit statistic when we are concerned about general changes in distribution, rather than simple changes in mean, in a sequence of independent data points. We have shown the resulting test statistic has computable large sample theory which can be used to provide very accurate p-values. Moreover we have shown that averaging over possible change points is generally more sensitive to alternatives than maximizing over possible change points.

The basic idea can be used in other contexts. Consider, for instance, testing for a change in mean. We describe first the unrealistic situation in which the standard deviation is known and then how to handle estimation of that SD. Suppose $X_1, \ldots, X_n$ are independent and we wish to test the null hypothesis that they are iid with unknown mean μ and known standard deviation σ (which we take to be 1 for notational convenience) against the alternative that the mean changes after the data point number c. The usual Z statistic is

$$T_c = \left(\frac{X_1 + \cdots + X_c}{c} - \frac{X_{c+1} + \cdots + X_n}{n-c} \right) / \sqrt{\frac{1}{c} + \frac{1}{n-c}}.$$

Our proposal would be to use the two sided test

$$\overline{T^2} = \frac{1}{n-1} \sum_{c=1}^{n} T_c^2.$$

This statistic has mean 1 under the null hypothesis of no change in mean. Arguments similar to those in Section 2 show that this statistic has the same limiting distribution, under the null, as the well known Anderson-Darling goodness-of-fit statistic.

In the more reasonable case where the (assumed common) standard deviation is unknown will use the statistic

$$T_s^2 = \overline{T^2}/s^2$$

where s^2 is some estimate of σ^2 which is consistent under the null hypothesis. The sample standard deviation is one possibility though this can be badly biased under the alternative. An estimate which is rather less precise but still likely to be quite accurate under the alternative hypothesis is

$$s_1^2 = \frac{\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2}{2(n-1)}.$$

Notice that under the alternative hypothesis all but one term in this average is an unbiased estimate of σ^2; the bias in the estimator is $\Delta_\mu^2/(2n)$ where Δ_μ denotes the change in the mean at the true change point. Under the null our estimate is unbiased. The statistic T_s^2 also has the same limiting distribution as the well known Anderson-Darling goodness-of-fit statistic when the null holds.

Other nonparametric goodness of fit tests can be used instead of the Cramér-von Mises test. For example a Bayesian test Labadi et al. (2014), likelihood tests Csörgö et al. (1997) or other two-sample tests Büning (2002). Sample size, the kind of alternative distribution from which we expect the data to come and the expected index of the change point should likely be used to choose the best test. Finding the asymptotic distribution for less well-known tests can be difficult. Bootstrapping can be used instead. This deserves further research.

Appendix

Proof of Theorems 1 and 2.

The weak limit $\mathbb{Z}$ given below is discussed in Picard (1985) but we provide details for completeness.

We prove Theorem 2 first. Define the partial sum empirical process (van der Vaart and Wellner, 1996, p. 225), for $(s,t) \in [0,1]^2$, by

$$\mathbb{Z}_n(s,t) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le ns} \{1(U_i \le t) - t\}.$$

Our statistic can be described in terms of this process. Notice that

$$F_c(t) = \frac{\sqrt{n}}{c}\mathbb{Z}_n(c/n, t) + t$$

and that

$$G_d(t) = \frac{\sqrt{n}}{d}\left\{\mathbb{Z}_n(1, t) - \mathbb{Z}_n(c/n, t)\right\} + t.$$

Thus

$$F_c(t) - G_d(t) = \sqrt{n}\left\{\frac{\mathbb{Z}_n(c/n, t)}{c} - \frac{\mathbb{Z}_n(1, t) - \mathbb{Z}_n(c/n, t)}{d}\right\}.$$

Now define the process $\mathbb{W}_n(s, t)$ for $0 < s < 1$ and $0 \le t \le 1$ by

$$\mathbb{W}_n(s, t) = \sqrt{s(1 - s)}\left\{\frac{\mathbb{Z}_n(s, t)}{s} - \frac{\mathbb{Z}_n(1, t) - \mathbb{Z}_n(s, t)}{1 - s}\right\}.$$

For given c our two sample test statistic is given by

$$W_n(c) = \int_0^1 \left\{\mathbb{W}_n(c/n, t)\right\}^2 \, dH_n(t).$$

Let ν_n be the probability measure on $(0, 1)$ putting mass $1/(n - 1)$ on each point of the form c/n for $1 \le c \le n - 1$. Our statistic is

$$\overline{W}_n = \int_0^1 \int_0^1 \left\{\mathbb{W}_n(s, t)\right\}^2 \, dH_n(t)\, d\nu_n(s).$$

We now break the proof of our two results into steps consisting of a statement followed by a detailed proof. In each case the assertions are intended to hold under the null hypothesis and the assumption that the common distribution H is continuous.

Step 1: The process $\mathbb{Z}_n$ converges weakly in $\ell_\infty([0, 1]^2)$ to a tight, centred, Gaussian process $\mathbb{Z}$ with covariance

$$\rho(s, t; s', t') = (s \wedge s')(t \wedge t' - tt').$$

See van der Vaart and Wellner (1996).

Step 2: Hence the process $\mathbb{W}_n$ converges weakly in $\ell_\infty^{\text{loc}}((0, 1) \times [0, 1])$ to the tight centred Gaussian process

$$\mathbb{W}(s, t) = \sqrt{s(1 - s)}\left\{\frac{\mathbb{Z}(s, t)}{s} - \frac{\mathbb{Z}(1, t) - \mathbb{Z}(s, t)}{1 - s}\right\}.$$

This process has continuous sample paths (on $(0,1)\times[0,1]$) and the covariance given in the statement of the theorem.

Step 3: For any sequence c_n with $\epsilon_n \equiv c_n/n \to 0$ we have

$$\left\{\int_0^{\epsilon_n} + \int_{1-\epsilon_n}^1\right\}\{\mathbb{W}_n(c/n,t)\}^2\; dH_n(t)d\nu_n(s) = \frac{\sum_{i=1}^{c_n} W_n(i) + \sum_{i=n+1-c_n}^{n} W_n(i)}{n-1} \to 0$$

in probability. Under the null hypothesis the mean of $W_n(c)$ is $1/6+1/(6n)$; see Anderson (1962). The expected value of the indicated quantity is thus

$$\frac{2c_n}{n-1}\left(\frac{1}{6}+\frac{1}{6n}\right) \to 0.$$

Step 4: The integral

$$W_\infty = \int_0^1\int_0^1 \mathbb{W}^2(s,t)dt\,ds$$

is almost surely finite. Since all the variates involved are non-negative we may compute

$$\mathrm{E}(W_\infty) = \mathrm{E}\left(\int_0^1\int_0^1 \mathbb{W}^2(s,t)dt\,ds\right) = \int_0^1 \chi(s,s)\,ds\int_0^1 \psi(t,t)\,dt = 1/6 < \infty.$$

Step 5: For any sequence ϵ_n tending to 0 as $n\to\infty$ we have, by taking expectations,

$$\left\{\int_0^{\epsilon_n} + \int_{1-\epsilon_n}^1\right\}\int_0^1 \mathbb{W}^2(s,t)dt\,ds \to 0$$

in probability.

Step 6: The tensor product kernel

$$\rho = \chi\otimes\psi(s,t;s',t') = \chi(s,s')\psi(t,t')$$

is compact and has eigenvalue-eigenfunction pairs

$$\lambda_{jk} = \frac{1}{j(j+1)}\frac{1}{\pi^2k^2}, \quad f_{jk}(s,t) = f_{\chi,j}(s)f_{\psi,k}(t)$$

indexed by j,k each running from 1 to ∞. It follows as usual that the family

$$Z_{jk} = \frac{1}{\sqrt{\lambda_{jk}}}\int_0^1\int_0^1 \mathbb{W}(s,t)f_{jk}(s,t)dt\,ds$$

defines a family of independent standard normal variables. Parseval's identity is then

$$\int_0^1 \int_0^1 \mathbb{W}^2(s,t)dt\,ds = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{Z_{jk}^2}{j(j+1)\pi^2 k^2}.$$

Step 7: For each fixed $\epsilon > 0$ we have

$$\int_{\epsilon}^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t)dH_n(t)\,d\nu_n(s) - \frac{1}{n-1} \sum_{n\epsilon < i < n(1-\epsilon)} W_n^2(i) \to 0$$

in probability. This is an easy consequence of the fact that for $i/n \le s < (i+1)/n$ we have $\int_0^1 \mathbb{W}_n^2(s,t)dF_n(t) = \mathbb{W}_n^2(i)$.

Step 8: For each fixed $\epsilon > 0$ we have

$$\int_{\epsilon}^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t)dH_n(t)\,d\nu_n(s) - \int_{\epsilon}^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t)\,dt\,ds \to 0$$

Under the null hypothesis H_n converges weakly to the uniform law on the unit interval. Moreover ν_n converges weakly to Lebesgue measure on the unit interval. The weak convergence result in Step 2 above uses a topology of uniform convergence on compacts such as the set $[\epsilon, 1-\epsilon] \times [0,1]$ and this implies the desired result.

Step 9: For each fixed $\epsilon > 0$ we have

$$\int_{\epsilon}^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t)dt\,ds \rightsquigarrow \int_{\epsilon}^{1-\epsilon} \int_0^1 \mathbb{W}^2(s,t)\,dt\,ds.$$

This is a direct consequence of weak convergence using the continuous mapping theorem.

Step 10: There is a metric d on the set of probability measures on the real line for which the metric topology is the topology of weak convergence. For each fixed $\epsilon > 0$ we have

$$d\left(\mathcal{L}\left(\int_{\epsilon}^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t)dt\,ds\right), \mathcal{L}\left(\int_{\epsilon}^{1-\epsilon} \int_0^1 \mathbb{W}^2(s,t)\,dt\,ds\right)\right) \to 0.$$

There is then a sequence $\epsilon_n \to 0$ so slowly that this convergence continues to hold with ϵ replaced by ϵ_n and so that the convergences in Steps to 7 and 8 continue to hold. Notice that by Step 5

$$d\left(\mathcal{L}\left(\int_{\epsilon_n}^{1-\epsilon_n} \int_0^1 \mathbb{W}^2(s,t)\,dt\,ds\right), \mathcal{L}\left(\int_0^1 \int_0^1 \mathbb{W}^2(s,t)\,dt\,ds\right)\right) \to 0.$$

for this sequence.

Step 11: For the sequence chosen in Step 10 we therefore have

$$\frac{1}{n-1}\sum_{n\epsilon_n<i<n(1-\epsilon_n)} W_n^2(i) \rightsquigarrow \int_0^1\int_0^1 \mathbb{W}^2(s,t)\,dt\,ds.$$

In view of Step 1 we see

$$W_n \rightsquigarrow \int_0^1\int_0^1 \mathbb{W}^2(s,t)\,dt\,ds$$

The law of the limit is, by Step 6, that of

$$\sum_{j=1}^\infty\sum_{k=1}^\infty \frac{Z_{jk}^2}{j(j+1)\pi^2k^2}.$$

This completes the proofs of Theorems 1 and 2.

Proof of Theorem 3.

This is standard so we present only an outline. Conditions **A1** and **A2** can be used to prove that

$$\Lambda_n - S_n \to 0$$

in probability under the null. The Lindeberg Central limit theorem then establishes the first conclusion of the Theorem. For more detailed arguments in a similar context see Guttorp and Lockhart (1988). Thus, under the conditions of the theorem the sequence of alternatives is contiguous to a sequence for which the null holds.

Contiguity implies that tightness under the null sequence extends to tightness under the alternative sequence. This proves tightness, under the alternative, of the sequence of processes $\mathbb{W}_n$. Thus we need only compute the limiting finite dimensional distributions under the alternative sequence. As usual we apply LeCam's Third Lemma (again similar arguments are in Guttorp and Lockhart (1988)) to reduce the problem to studying the joint law, under the null hypothesis, of Λ_n and the vector $(\mathbb{W}_n(s_1,t_1),\ldots,\mathbb{W}_n(s_k,t_k))$ for an arbitrary sequence of time points $t_1,\ldots,t_k$ all in $[0,1]$.

The null distribution theory presented above (see Step 1 in the proof of Theorem 2) shows that, under the null hypothesis,

$$(\mathbb{W}_n(s_1,t_1),\ldots,\mathbb{W}_n(s_k,t_k)) \rightsquigarrow MVN_k(0,\mathbf{R}_W)$$

where $\mathbf{R}_W$ is the $k\times k$ matrix with i,jth entry

$$R_{Wij} = \rho_W(s_i,t_i;s_j,t_j).$$

The Lindeberg Central Limit Theorem may now be used to show that the vector

$$(S_n, \mathbb{W}_n(s_1,t_1), \ldots, \mathbb{W}_n(s_k,t_k))$$

converges in distribution to multivariate normal with mean vector $(-\tau^2/2, 0, \ldots, 0)$ and variance covariance matrix of the form

$$\left[\begin{array}{cc} \tau^2 & \mathbf{c}^\top \\ \mathbf{c} & \mathbf{R}_W \end{array} \right].$$

Here the vector $\mathbf{c}$ is the limiting covariance which is found, after some algebra, to be

$$c_i = \mu(s_i,t_i) = \mu_\chi(s_i)\mu_\psi(t_i).$$

This completes the proof of the second assertion of the Theorem.

The third step is standard; Guttorp and Lockhart (1988) does similar problems.

Proof of Proposition 1

Fix $0 < \delta < 1/2$ and let A_n denote the event $\{\delta \le \hat{c}/n \le 1-\delta\}$. We will show that

$$\lim_{n\to\infty} P(A_n) = 0.$$

This will prove Proposition 1. To this end fix $0 < \epsilon < \delta$. Define

$$M_n = \sup_{\delta \le s \le 1-\delta} \int_0^1 \frac{\mathbb{B}_n^2(s,t)}{s(1-s)}\, dt$$

and

$$M_n'(\epsilon) = \sup_{\epsilon \le s \le \delta} \int_0^1 \frac{\mathbb{B}_n^2(s,t)}{s(1-s)}\, dt.$$

Then

$$A_n \subset \{M_n'(\epsilon) < M_n\}.$$

Weak convergence of $\mathbb{B}_n$ to $\mathbb{B}$ guarantees that

$$\limsup_{n\to\infty} P(A_n) \le \limsup_{n\to\infty} P\{M_n'(\epsilon) < M_n\} \le P(M'(\epsilon) \le M)$$

where

$$M = \sup_{\delta \le s \le 1-\delta} \int_0^1 \frac{\mathbb{B}^2(s,t)}{s(1-s)}\, dt$$

and

$$M'(\epsilon) = \sup_{\epsilon < s \le \delta} \int_0^1 \frac{\mathbb{B}^2(s,t)}{s(1-s)}\, dt.$$

We claim that

$$\lim_{\epsilon \to 0} P(M'(\epsilon) \le M) = 0 \tag{2}$$

and this will prove

$$\limsup_{n\to\infty} P(A_n) = 0$$

and Proposition 1.

Assertion (2) would follow from a law of the iterated logarithm (as $s \to 0$) for the process

$$s \mapsto \frac{\int_0^1 \mathbb{B}^2(s,t)\,dt}{s(1-s)}.$$

While we expect such a result to hold we have not tried to prove anything along those lines. We will establish instead the lower bound

$$\limsup_{s\to 0} \int_0^1 \frac{\pi^2 \mathbb{B}^2(s,t)}{2\log\{\log(1/s)\}s(1-s)}\,dt \ge 1$$

almost surely which is enough to imply (2). We enumerate the steps needed:

1. Let

   $$I_{\mathbb{B}}(s) = \int_0^1 \frac{\mathbb{B}^2(s,t)}{s}\,dt,$$

   and

   $$I_{\mathbb{Z}}(s) = \int_0^1 \frac{\mathbb{Z}^2(s,t)}{s}\,dt.$$

   Then

   $$\begin{aligned}\int_0^1 \frac{\mathbb{B}^2(s,t)}{s(1-s)}\,dtI &= \int_0^1 \frac{\{\mathbb{Z}(s,t) - s\mathbb{Z}(1,t)\}^2}{s(1-s)}\,dt \\ &\ge \frac{I_{\mathbb{Z}}(s)}{1-s} + \frac{s}{1-s}\int_0^1 \mathbb{Z}^2(1,t)\,dt - \frac{2s}{1-s}\sqrt{I_{\mathbb{Z}}(s)\int_0^1 \mathbb{Z}^2(1,t)\,dt}.\end{aligned}$$

   From this we deduce that it is enough to show that

   $$\limsup_{s\to 0} \int_0^1 \frac{\pi^2 \mathbb{Z}^2(s,t)}{2\log\{\log(1/s)\}s}\,dt \ge 1 \tag{3}$$

   almost surely.

2. For each fixed s the process

   $$t \mapsto \frac{\mathbb{Z}(s,t)}{\sqrt{s}}$$

is a Brownian Bridge. If we put

$$W(s) = \int_0^1 \frac{\mathbb{Z}^2(s,t)}{s}\,dt$$

then each $W(s)$ has the same distribution as the limit law of the usual Cramér-von Mises statistic which is the law of

$$\sum_{j=1}^{\infty} \lambda_j Z_j^2.$$

In this representation the Z_j are iid standard normal and the eigenvalues λ_j are given, for $j = 1, 2, \ldots$, by

$$\lambda_j = \frac{1}{\pi^2 j^2}.$$

3. The process $\mathbb{Z}$ has independent increments in s and for each $0 < s' < s$ the process
$$t \mapsto \frac{\mathbb{Z}(s,t) - \mathbb{Z}(s',t)}{\sqrt{s - s'}}$$
has the same law as
$$t \mapsto \frac{\mathbb{Z}(s,t)}{\sqrt{s}}$$

4. Now fix $s_0 = 1$ and some $r < 1$ to be chosen later. Define $s_n = s_0 r^n$ for $n = 1, 2, \ldots$. Put
$$W_n = \int_0^1 \frac{\mathbb{Z}^2(s_n,t)}{s_n}\,dt$$
and
$$W_n^* = \int_0^1 \frac{\{\mathbb{Z}(s_n,t) - \mathbb{Z}(s_{n+1},t)\}^2}{s_n - s_{n+1}}\,dt.$$
All of these variables have the law of $W(s)$ described above.

5. Fix $\epsilon > 0$. Let A_n be the event $W_n^* > 2(1-\epsilon)\lambda_1 \log(\log(1/s_n))$ and B_n be the event $W_{n+1} \le 2(1+\epsilon)\lambda_1 \log(\log(1/s_n))$. We will show that we can choose r small enough so that

   (a) The event that A_n occurs infinitely often (i.o.) has probability 1.

   (b) The event that B_n occurs for all large n has probability 1.

6. So the event $A_n \cap B_n$ i.o. has probability 1.

7. On the event $A_n \cap B_n$ we have $W_n \geq 2(1-\epsilon)\lambda_1 \log(\log(1/s_n)$ so that this event occurs infinitely often.

8. This proves

$$P\left\{W_n \geq 2(1-\epsilon)\lambda_1 \log(\log(1/s_n)) \,\text{i.o.}\right\} = 1.$$

which establishes (3). The definition of contiguity is that any sequence of events whose probability converges to 0 under the null has probability converging to 0 under the alternative. This finishes the proof of Proposition 1.

Evidence for Conjecture 1

For $\epsilon > 0$ we define

$$I_n(\epsilon) = \{c : 1 \leq c \leq n\epsilon \text{ or } 1 \leq n - c \leq n\epsilon\}.$$

Proposition 1 establishes that there is a sequence $\epsilon_n \searrow 0$ such

$$\lim_{n\to\infty} P(\hat{c}_n \in I_n(\epsilon_n)) = 1.$$

Thus

$$P\left[W_{\max} = \max\{W_n(c) : c \in I_n(\epsilon_n)\}\right] \to 1. \tag{4}$$

We now outline the steps in our strategy for proving the conjecture before giving some evidence for each step.

Step 1 There are constants a_n and b_n and a random variable V such that

$$a_n W_{\max} - b_n \rightsquigarrow V$$

and V has a continuous limit distribution.

Step 2: So

$$a_n \max\{W_n(c) : c \in I_n(\epsilon_n)\} - b_n \rightsquigarrow V.$$

Step 3: There are random variables $\tilde{W}_n(c)$ such that under the null hypothesis

$$a_n \max\{|W_n(c) - \tilde{W}_n(c)| : c \in I_n(\epsilon_n)\} \to 0$$

and such that for each $c \in I_n(\epsilon_n)$ the variable $\tilde{W}_n(c)$ is measurable with respect to the σfield generated by $X_c, c \in I_n(\epsilon_n)$. To be specific we define, for $c < n/2$,

$$\tilde{W}_n(c) = \int_0^1 c\{F_c(u) - u\}^2 \, du$$

and, for $c > n/2$,

$$\tilde{W}_n(c) = \int_0^1 d\{G_d(u) - u\}^2\, du.$$

(Recall the shorthand $d = n - c$.)

Step 4: Define

$$\tilde{\Lambda}_n = \sum_{n\epsilon_n < c \le c_0} \phi_f(X_c)/\sqrt{n} - \sum_{c_0 < c < n - n\epsilon_n} \phi_g(X_c)/\sqrt{n}.$$

The log-likelihood ratio Λ_n satisfies

$$\Lambda_n - \tilde{\Lambda}_n \to 0$$

in probability, under the null hypothesis.

Step 5: Since $\tilde{W}_{\max}$ is independent of $\tilde{\Lambda}_n$ we may apply LeCam's third lemma to show that under the sequence of contiguous alternatives we have

$$a_n W_{\max} - b_n \rightsquigarrow V$$

Step 6: Since this limit law is the same as under the null we must power minus level tends to 0.

For some of these steps we can fill in partial evidence.

For Step 1 we would hope to follow the ideas in Jaeschke (1979) to show that the limit V has an extreme value distribution. In that paper the maximizer of the usual empirical process, standardized by dividing by its standard deviation, is shown to have an extreme value limit with constants analogous to a_n and b_n involving $\sqrt{\log\log n}$ and $\log\log\log n$.

Step 2 is a consequence of Step 1 and (4).

In Step 3 we would hope to use the closeness of H_n to the uniform distribution to convert the $dH_n(u)$ integrals to du integrals. Then we write

$$\frac{cd}{n} \int_0^1 \{F_c(u) - G_d(u)\}\, 2\, du$$

as a sum of three terms

$$T_1 = \frac{d}{n} \int_0^1 c\left\{F_c(u) - u\right\}^2\, du,$$

$$T_2 = \frac{c}{n} \int_0^1 d\left\{G_d(u) - u\right\}^2\, du,$$

and

$$T_3 = -2\frac{\sqrt{cd}}{n}\int_0^1 \sqrt{c}\,\{F_c(u)-u\}\cdot\sqrt{d}\,\{G_d(u)-u\}\,du.$$

The integrals in T_1 and T_2 are both one sample Cramér-von Mises statistics so they are on the order 1. For any sequence $c = c_n$ such that $c_n/n \to 0$ the coefficient in front of T_2 is $o(1)$. So T_2 is negligible relative to T_1. The Cauchy-Schwarz inequality then shows T_3 is negligible relative to T_2. There is a parallel argument when $c_n/m \to 1$.

Step 4 is not conjecture; its proof is straightforward from the assumptions of the Conjecture. Steps 5 and 6 are exactly parallel to the arguments in Lockhart (1991).

References

T. W. Anderson. On the distribution of the two-sample Cramér-von Mises criterion. *Ann. Math. Statist.*, 33:1148–1159, 1962.

J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.*, 32(2):485–498, 06 1961. doi: 10.1214/aoms/1177705055. URL `https://doi.org/10.1214/aoms/1177705055`.

E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change Point Problems.* Springer Netherlands, 1993.

Herbert Büning. Robustness and power of modified Lepage, Kolmogorov-Smirnov and Cramér-von Mises two-sample tests. *Journal of Applied Statistics*, 29(6): 907–924, 2002.

M. Csörgö, , and L. Horváth. *Limit Theorems in Change-Point Analysis.* Wiley Series in Probability and Statistics. Wiley, 1997.

Pierre Duchesne and Pierre Lafaye de Micheaux. Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54:858–862, 2010.

P. Guttorp and R. A. Lockhart. On the asymptotic distribution of quadratic forms in uniform order statistics. *The Annals of Statistics*, 16:433–449, 1988.

J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48:419–426, 1961.

D. Jaeschke. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Ann. Statist.*, 7(1):108–115, 01 1979.

Luai Al Labadi, Emad Masuadi, and Mahmoud Zarepour. Two-sample Bayesian nonparametric goodness-of-fit test, 2014.

R. A. Lockhart. Overweight tails are inefficient. *The Annals of Statistics*, 19(4): 2254–2258, 1991.

Ian W. McKeague and Yanqing Sun. Transformations of gaussian random fields to brownian sheet and nonparametric change-point tests. *Statistics & Probability Letters*, 28(4):311–319, 1996. doi: 10.1016/0167-7152(95)00140-9. URL `https://doi.org/10.1016/0167-7152(95)00140-9`.

Nornadiah Mohd Razali and Bee Yap. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Analytics*, 2, 01 2011.

Dominique Picard. Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4):841–867, 1985.

Michael A Stephens. Tests based on EDF statistics. In Ralph B D'Agostino and Michael A Stephens, editors, *Goodness-of-fit Techniques*, chapter 4, pages 97–193. Marcel Dekker, New York, 1986.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

Tonglin Zhang. A kolmogorov-smirnov type test for independence between marks and points of marked point processes. *Electron. J. Statist.*, 8(2):2557–2584, 2014. doi: 10.1214/14-EJS961. URL `https://doi.org/10.1214/14-EJS961`.

ISBN 978-82-471-9838-4 (printed ver.)
ISBN 978-82-471-9354-9 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology