Testing for Equality of Distributions Under Known Additional Measurement Error

Maxime Riscard, Luis F Campos (or vise versa)

May 31, 2018

[**LFC**] This is written extremely loosely for now, just want to get everything down

1   Introduction

Known measurement error in Astronomy and related fields:

In the field of Astronomy, Astrophysics, etc. many, if not most, of the measurements astronomers use to answer questions have known measurement error. This error is known (or approximately known) because extensive testing is done on the ground to understand the behavior of potential observations when the observatory goes live. Typically, the data is analyzed without regard to this error, or is accounted for in rudimentary ways, such as reweighing.

The specific problem we're thinking about:

A problem was brought to us that is commonly found in Astronomy, we have a set of measurements coming from two different sets of galaxies and we want to know

if the distributions of these measurements are the same. What is typically done is the measurements are collected and the a simple Kolmogorov-Smirnov Test (Smirnov [1948]) is conducted for equality of the distributions. The measurement error remains unaccounted for and invalidates the K-S test when the error is heteroskedastic, or dependent. We seek to find ways to analyze this type of data accounting for known measurement error that is generalizable and theoretically-grounded. But we will confine ourselves to the context of Hypothesis testing for equality of distributions for now.

Our Strategy:

Generally, we use deconvolution (cite) in conjunction with parametric bootstrap to account for known measurement error distributions, then we use simple distributional distance metrics to test the equality of distributions. This strategy is helpful in several ways. The user can select different deconvolution techniques depending on the assumptions about the underlying process they are willing to make. The performance of different deconvolution techniques is dictated by the data (sample size, true distribution) and the measurement error distributions (smooth, super-smooth). Using a parametric bootstrap allows the user to define the measurement error distributions for their specific instrument and measurement. In some situations, practitioners happily assume symmetric super-smooth measurement error distributions, like the Normal distribution. But there are situations, e.g. real positive outcomes, where the measurement errors are not symmetric where other models are better fits. This can be accounted for in the parametric bootstrap since we simulate the error process and can implement it with any known distribution. The strategy also allows us to generalize away from any one test, like the K-S test, by enabling the user to define a test statistic that measures differences in distribution. Giving the user the freedom to select the most meaningful measure of difference for their specific problem.

2 Problem Set-Up

In this section we set notation and set of the problem of distributional testing with known measurement error.

We assume the data following data generating process for our data.

Denote the noise-free outcomes X_i and Y_j, these are drawn from distributions F_X and F_Y respectively, $i = 1, ..., n_X$ and $j = 1, ..., n_Y$. We, instead, observe noisy-observations

$$\widetilde{X_i} = X_i + \epsilon_{X,i} \quad \text{and} \quad \widetilde{Y_j} = Y_j + \epsilon_{Y,j}. \tag{1}$$

The distribution of the errors are known, for example they can be assumed to be normal

$$\epsilon_{X,i} \sim \mathcal{N}\left(0, \sigma_{X,i}^2\right) \quad \text{and} \quad \epsilon_{Y,j} \sim \mathcal{N}\left(0, \sigma_{Y,j}^2\right) \tag{2}$$

with known variances $\sigma_{X,i}^2$ and $\sigma_{Y,j}^2$. Let's collect the known variances into vectors as

$$\sigma_X = (\sigma_{X,1}, \sigma_{X,2}, ..., \sigma_{X,n_X}) \quad \text{and} \quad \sigma_Y = (\sigma_{Y,1}, \sigma_{Y,2}, ..., \sigma_{Y,n_Y}).$$

We want to test the equality of the distributions from which the noise-free observations are drawn $H_0 : F_X = F_Y$. Under homoskedastic noise, i.e. $\sigma_{X,i}^2 = \sigma_{Y,i}^2 = \sigma^2$, the distributions of the noisy observations are the same under the null distribution and hence the K-S statistic is valid, though it will lose power as the noise increases. However, under heteroskedastic noise this will not be the case. We use alternative methods for testing this hypothesis.

3 Methods

3.1 Deconvolution

In this section we describe the deconvolution problem, some methods for performing deconvolution, and concerns in using this for Hypothesis testing.

3.2 Two-sample tests for equality of distributions

In this section we explain some of the basic details for tests for equality of distributions, when they fail and set up the next section on different measures of distributional equality.

3.3 Test Statistics for Distributional equality

Here we explore different measures of distributional similarity equality, KS statistic, AD statistic, Earth Movers, KL, etc. We can discuss their benefits and drawbacks in the context of testing for distributional equality.

3.4 Putting the Pieces Together

As described in Section 3.1, deconvolution involves approximating an underlying distribution of a set of noisy observations when the error's distribution and magnitude are known. So, given the set of observations $\widetilde{X}$ and error magnitudes σ_X we can approximate the underlying distribution F_X. We denote this application of the deconvolution procedure as

$$\widehat{F}_X = \texttt{decon}(\widetilde{X}, \sigma_X). \tag{3}$$

Data: $\widetilde{X}, \widetilde{Y}, \sigma_X, \sigma_Y$

Result: p-value

$\phi_{obs} = \Phi(\widetilde{X}, \widetilde{Y})$

$\widehat{F}_0 = \texttt{decon}((\widetilde{X}, \widetilde{Y}), (\sigma_X, \sigma_Y))$

for b *in* $1, ..., B$ **do**

  for i *in* $1, ..., n_X$ **do**

   $X_i^{(b)} \sim \widehat{F}_0 \quad \text{and} \quad \varepsilon_{X,i}^{(b)} \sim N\left(0, \sigma_{X,i}^2\right)$

   $\widetilde{X}_i^{(b)} = X_i^{(b)} + \varepsilon_{X,i}^{(b)}$

  end

  for j *in* $1, ..., n_Y$ **do**

   $Y_j^{(b)} \sim \widehat{F}_0 \quad \text{and} \quad \varepsilon_{Y,j}^{(b)} \sim N\left(0, \sigma_{Y,j}^2\right)$

   $\widetilde{Y}_j^{(b)} = Y_j^{(b)} + \varepsilon_{Y,j}^{(b)}$

  end

  $\phi^{(b)} = \Phi(\widetilde{X}^{(b)}, \widetilde{Y}^{(b)})$

end

Calculate:

$p = \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}\{\phi_{obs} > \phi^{(b)}\}$

Algorithm 2: How to write algorithms

4   Simulation Study

References

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19(2):279–281, 06 1948. doi: 10.1214/aoms/1177730256. URL https://doi.org/10.1214/aoms/1177730256.