Testing for Equality of Distributions Under Known Additional Measurement Error

Luis F Campos, Maxime Rischard

March 22, 2019

[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 the problem of distributional testing with known measurement error. We assume the following data generating process for our data.

Denote the noise-free outcomes X_i, $i = 1, \ldots, n_x$ and Y_j, $j = 1, \ldots, n_Y$:

$$\begin{aligned} X_i &\overset{\text{i. i. d.}}{\sim} F_X, \quad F_X \in \mathcal{F}\,, \\ Y_j &\overset{\text{i. i. d.}}{\sim} F_y, \quad F_Y \in \mathcal{F}\,, \end{aligned} \tag{1}$$

where $\mathcal{F}$ is the set of all continuous and univariate probability measures.

We do not directly observe the noise-free outcomes, instead we measure the noisy observations

$$\tilde{X}_i = X_i + \epsilon_{X,i} \quad \text{and} \quad \tilde{Y}_j = Y_j + \epsilon_{Y,j}. \tag{2}$$

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

$$\epsilon_{X,i} \overset{\text{i. i. d.}}{\sim} \mathcal{N}\left(0, \sigma^2_{X,i}\right) \quad \text{and} \quad \epsilon_{Y,j} \overset{\text{i. i. d.}}{\sim} \mathcal{N}\left(0, \sigma^2_{Y,j}\right) \tag{3}$$

with known variances $\sigma^2_{X,i}$ and $\sigma^2_{Y,j}$. 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 $H_0 : F_X = F_Y$. A more formal statement of the hypotheses is:

$$\begin{aligned} &(F_X, F_Y) \in \mathcal{F} \times \mathcal{F} \equiv \Omega \\ H_0 :&(F_X, F_Y) \in \Omega_0, \quad \Omega_0 = \{(F_A, F_B) \in \mathcal{F} \times \mathcal{F},\ s.t.\ F_A = F_B\} \\ H_1 :&(F_X, F_Y) \in \Omega_1, \quad \Omega_1 = \{(F_A, F_B) \in \mathcal{F} \times \mathcal{F},\ s.t.\ F_A \neq F_B\} \end{aligned} \tag{4}$$

which makes it clearer that the hypotheses partition the joint space $\mathcal{F} \times \mathcal{F}$ with $\Omega_0 \cap \Omega_1 = \emptyset$ and $\Omega_0 \cup \Omega_1 = \mathcal{F} \times \mathcal{F}$.

Our goal is to devise a hypothesis test which can distinguish H_0 and H_1. We encode the hypothesis test with a test function $\varphi : \mathbb{R}^{n_X} \times \mathbb{R}^{n_Y} \to \{0, 1\}$ which takes in the noisy observations and returns either 0 (the null hypothesis is chosen) or 1 (the alternative is chosen). The test function can also make use of the known noise variances σ_X^2 and σ_Y^2, but we suppress this in the notation for concision. We further define the power of the test:

$$\beta(F_X, F_Y) = \mathbb{E}_{F_X,F_Y}\left[\varphi(\tilde{X}, \tilde{Y})\right] \tag{5}$$

where $\mathbb{E}_{F_X,F_Y}$ denotes expectations with respect to X, Y, ϵ_X and ϵ_Y. We also define the significance

$$\alpha = \sup_{(F_X,F_Y)\in\Omega_0} \left(\beta(F_X, F_Y)\right) = \sup_{F\in\mathcal{F}} \left(\beta(F, F)\right) , \tag{6}$$

the supremum of the power under the null hypothesis. Under the classical hypothesis testing framework, we choose a nominal significance α^*, and then design a test function φ which maximizes the power while satisfying $\alpha < \alpha^*$.

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, KS statistic, AD statistic, Earth Movers, KL, etc. We can discuss their benefits and drawbacks in the context of testing.

For example, we could calculate the KS distance with the observed noisy data, i.e.

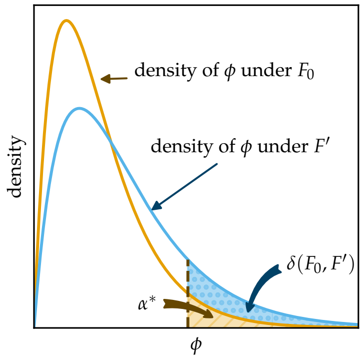
$$\Phi(\tilde{X}, \tilde{Y}) = KS(\hat{F}_{\tilde{X}}, \hat{F}_{\tilde{Y}})$$

This would compare the empirical distributions of the noisy data and return small values for cases where $\hat{F}_{\tilde{X}}$ is close to $\hat{F}_{\tilde{Y}}$ in the KS sense.

We can also discuss their relationship to particular form of deconvolution and discuss how to include a deconvolution step in implementing Φ. For example

$$\Phi(\tilde{X}, \tilde{Y}) = KS(\texttt{decon}(\tilde{X}; \sigma_X), \texttt{decon}(\tilde{Y}; \sigma_Y))$$

Here, we would be accounting for calculating the distances between the "correct" distributions because $\texttt{decon}(\tilde{X}; \sigma_X)$ is approximating F_X and this comparison will be similar to the ideal noiseless compatison $\Phi(X, Y) = KS(\hat{F}_X, \hat{F}_Y)$.



(a) Bootstrap test.

Figure 1: Validity of the bootstrap test.

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 $\tilde{X}$ and error magnitudes σ_X we can approximate the underlying distribution F_X. We denote this application of the deconvolution procedure as

$$\hat{F}_X = \texttt{decon}(\tilde{X}; \sigma_X). \tag{7}$$

3.5 Validity

We turn to the question of the validity of the tests defined by Algorithm 1 and Algorithm 2, where validity means that the condition $\alpha \leq \alpha^*$ is satisfied. To be explicit, the test function under consideration is $\varphi(\tilde{X}, \tilde{Y}) = \mathbb{I}\{p \leq \alpha^*\}$, where p is the output of each algorithm.

In Algorithm 1, the boostrap is used to obtain the distribution of the test statistic

Algorithm 1: Bootstrap test for $F_X = F_Y = F_0$, where F_0 is a pre-specified distribution that can be sampled from. The test statistic $\Phi(\cdot, \cdot)$ is also pre-specified.

```
Data: $\tilde{X}, \tilde{Y}, \sigma_X, \sigma_Y$
Result: p-value: $p$
$\phi_{obs} = \Phi(\tilde{X}, \tilde{Y})$
for $b$ in $1, ..., B$ do
    for $i$ in $1, ..., n_X$ do
        $X_i^{(b)} \sim F_0$ and $\varepsilon_{X,i}^{(b)} \sim N\left(0, \sigma_{X,i}^2\right)$
        $\tilde{X}_i^{(b)} = X_i^{(b)} + \varepsilon_{X,i}^{(b)}$
    end
    for $j$ in $1, ..., n_Y$ do
        $Y_j^{(b)} \sim F_0$ and $\varepsilon_{Y,j}^{(b)} \sim N\left(0, \sigma_{Y,j}^2\right)$
        $\tilde{Y}_j^{(b)} = Y_j^{(b)} + \varepsilon_{Y,j}^{(b)}$
    end
    $\phi^{(b)} = \Phi(\tilde{X}^{(b)}, \tilde{Y}^{(b)})$
end
$p = \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}\{\phi_{obs} > \phi^{(b)}\}$
```

under $F_X = F_Y = F_0$, where F_0 is pre-specified. It follows that if it is indeed the case that $F_X = F_Y = F_0$, then the p-value is approximately uniformly distributed (exactly as $B \to \infty$), and therefore $\beta(F_0, F_0) = \alpha^*$. However, it does not follow that $\alpha = \alpha^*$, as there may very well exist a different $F' \in \mathcal{F}$ for which $\beta(F', F') > \alpha^*$. We can decompose the significance (6) into

$$\alpha = \beta(F_0, F_0) + \sup_{F' \in \mathcal{F}} \big(\underbrace{\beta(F', F') - \beta(F_0, F_0)}_{\equiv \delta(F_0, F')} \big)\,, \tag{8}$$

Algorithm 2: Parametric Bootstrap for Testing Equality of Distributions with known and normal noise distributions. It uses a pre-specified deconvolution function $\texttt{decon}(\cdot;\cdot)$ and statistical distance metric $\Phi(\cdot,\cdot)$

```
Data: $\tilde{X}, \tilde{Y}$, $\sigma_X, \sigma_Y$
Result: p-value: $p$
$\hat{F}_0 = \texttt{decon}((\tilde{X};\tilde{Y}), (\sigma_X, \sigma_Y))$
Return the $p$-value from Algorithm 1 with $F_0 = \hat{F}_0$.
```

and the difference in power $\delta(F_0, F')$ can be seen to be the probability under F' of the test statistic exceeding its $(1-\alpha^*)$ quantile under F_0 minus α^* (illustrated in Figure 1a). The implication is that it is desirable for the distribution of the test statistic under the null to be as insensitive as possible to choice of F_0. Asymptotically, this property is perfectly achieved by the Kolmogorov-Smirnov test statistic **[MR]** is this true? what's the exact statement?.

So far, we have assumed that F_0 is pre-chosen completely arbitrarily. If the true null distribution $F_0^\star$ was known, we would simply choose $F_0 = F_0^\star$ which would guarantee $\alpha = \alpha^*$. With $F_0^\star$ unknown, it remains desirable to choose an $\hat{F}_0$ that is close to $F_0^\star$, so that $\delta(\hat{F}_0, F_0^\star)$ is small. This is the aim of the deconvolution step in Algorithm 2.

4 Simulation Study

4.1 Homoskedastic Error

I claim this above: Under homoskedastic noise, i.e. $\sigma^2_{X,i} = \sigma^2_{Y,i} = \sigma^2$, the distributions of the noisy observations are the same under the null distribution and hence the K-S statistic is valid.

Claim:

Is this true? Let's compare the relative power of using KS directly to using our proposed method.

4.2 Deconvolve and Test?

A separate method has been proposed for this [Paul Green paper]. We could apply deconvolution to the two sets of data separately

$$\hat{F}_X = \texttt{decon}(\tilde{X}; \sigma_X)$$
$$\hat{F}_Y = \texttt{decon}(\tilde{Y}; \sigma_Y)$$

and compare these two distributions directly. Our intuition is that this will be conservative and underpowered. It is unclear as to how this will perform under large samples – let's investigate. Even if this performs well under large sample, hell even if this has the correct size when $n_x, n_Y \to \infty$, it is still a large sample argument.

4.3 Heteroskedastic but equal

Heteroskedastic errors with equal distributions. Set it so that $\sigma_{X_i} \sim \sigma_{Y_j}$ with increasing variability of variances, i.e. start at homoskedatic and create increasingly variable variances.

5 Future Work

The choice of deconvolution and distribution comparison methods discussed in Section 3.1 and Section 3.3 will absolutely have an impact on the power of the resulting test. For example, some deconvolution methods work under additional assumptions and hence, if those assumptions are correct, would have faster convergence rates. Under H_0 this would imply... F0 close to F0hat. This could be negated by ... tests that are insentitive to defficiencies in deconvolution methods. For example, the kernel-based methods have poor tail approximation behavior, but comparing some distance metrics, like earth movers(?) may be less sensitive to the tails and hence the resulting null distribution will not be affected. We shouold investigate this interaction further.

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. 2