A Distribution-Free Goodness-of-Fit Test
Using K-Nearest Neighbor Coincidences

Dong Xu & Leif Ellingson
Texas Tech University Department of Mathematics and Statistics

1 Idea of Nearest Neighbor Coincidence

• In this figure, Red dots represent actual data we collect. Blue empty circles represent data simulated from hypothesis distribution.
• Left figure gives data from two different distributions. Red dots gather together in the middle of the figure and form a rectangular while part of Blue empty circles spread outside it.
• If you locate a specific point by random, the nearest point of it has a quite large probability to belong to the same sample.
• Figure on the right gives data from two different distributions. They are distributed in a pretty similar way that the nearest neighbor of each point has the relatively same probability to be red or blue.

2 Goodness-of-Fit Tests

Suppose we have an independent sample of iid $\mathbb{R}^d$-valued random vectors: $X_1, \ldots, X_n$. The distribution of $X_i$ has unknown pdf $f$. We assume that $f$ is continuous a.e. with respect to the Lebesgue measure.

We aim to test a null hypothesis of the following type:

$H_0: f = f_0 \ a.e.$

$V_n: H_1: f$ and $f_0$ differ on a set of positive measure.

Some famous goodness-of-fit tests can do this, such as Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling procedure.

• However, they are all based on empirical CDFs which are not generalized well to more complicated data.
• Our destination is to develop a method that will be applicable to data on any metric space.

3 An Initial K-Nearest Neighbor Goodness-of-Fit-Test

• Henze (1987) introduced a two-sample test for equality of distributions.
• We develop our methodology based on his idea and modify our test so that it can check the distribution assumptions about the observed data.

The comparison between KNN and other mentioned goodness-of-fit tests is shown in the following figure.

4 An Improved Test Using Repeated Simulation

• A good way to improve the reliability of a test is to collect more information from the sample.
• So we create a method by using multiple repeated simulated samples.

Following is the algorithm of repeated simulation.

1. Generate $Y_1, \ldots, Y_l$ from $f_0$ instead of just $Y_1$.
2. Let $Z_i = X_i$ when $i \leq n_1$, $Z_i = Y_j$ when $i \leq n_2$.
3. Calculate Henze’s statistic $T_n$ and tentatively reject $H_0$ at level $\gamma$ if

$$\sqrt{n} T(n, r) = \sqrt{n} \left( \sum_{i=1}^{n-1} I(r) \right) > z_{1-\gamma}$$

4. Repeat steps 1 and 3 $v$ times.

5. Reject $H_0$ at level $\alpha$ if you tentatively reject $H_0$ greater than $5\alpha$ times. Note: Choose $\gamma$ so that $\gamma = v \cdot \sum_{i=1}^{n-1} I(r)$.

5 Conclusions and Ongoing Work

• We have developed a goodness-of-fit test that does not require the use of CDFs and can be used for more complicated data.
• Our improved test has comparable statistical performance to the classical tests for normality in the univariate case.
• We continue to run simulations for multivariate data and data on manifolds.

6 References