A Distribution-Free Goodness-of-Fit Test

Using K-Nearest Neighbor Coincidences

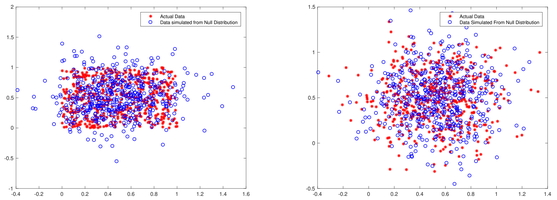
Dong Xu & Leif Ellingson

Texas Tech University Department of Mathematics and Statistics

Goals

- Develop a procedure to test whether data comes from a hypothesized probability distribution regardless of the dimensionality of the data and whether or not it is in a Euclidean space.
- Compare the results of our proposed procedure to leading methods for univariate data.

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 have the same color.
- Figure on the right gives data from two different distribution. 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: $X_1, \ldots, X_{n_1}$. The distribution of X_i has unknown pdf $f(x)$. We assume that f is continuous a.e. with respect to the Lebesgue measure.

We defined a test for testing the following hypotheses:

$H_0 : f = f_0 \quad a.e. \qquad VS. \qquad H_A : \quad f$ and f_0 differ on a set of positive measure

- Some famous goodness-of-fit tests can do this, such as Kolmogorov-Smirnov, Cramer-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.

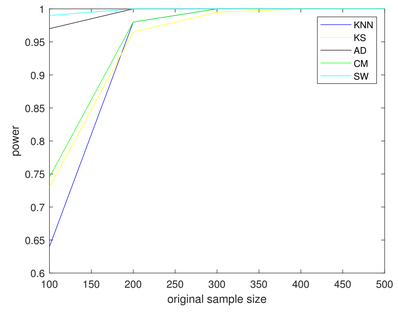
3 An Initial K-Nearest Neighbor Goodness of-Fit-Test

- 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 NN is defined as the observation closest to a given point, the second NN is the next closest point, and so on.
- Our purpose is to test this $f(x)$ is significantly approach the distribution $f_0(x)$ in our assumption. To do this, we simulate $Y_1, \ldots, Y_{n_2}$ from $f_0(x)$.
- To do this, we first defined the variable : $Z_i = X_i$ when $1 \le i \le n_1$, $Z_i = Y_i$ when $1 \le i \le n_2$.
- Define the rth (r=1,...,k) nearest neighbor to Z_i [denoted by $N_r(Z_i)$] as the point Z_j satisfying $|Z_v - Z_i| \le |Z_j - Z_i|$ for exactly r-1 value of v, $1 \le v \le n$, for $v \ne i, j$, and write

$$I_i(r) = \begin{cases} 1, & \text{if } Z_i \text{ and } N_r(Z_i) \text{ belong to the same sample} \\ 0, & \text{otherwise} \end{cases}$$

  Thus, the test statistic $T_{n,k} = \sum_{i=1}^{n} \sum_{r=1}^{k} I_i(r)$ represents the number of all k nearest neighbor coincidences.
- H_0 is rejected for large values of $T_{n,k}$.
- By restating $T_{n,k}$ in a different manner, Henze showed that the permutation distribution of $T_{n,k}$ is asymptotically normal with a specific mean and variance, so computations of the permutations are not needed in practice. Hence large value of $T_{n,k}$ provide strong evidence against null hypothesis.

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



- In this figure, Shapiro-Wilk test beats everyone else from the beginning to the end.
- Following is the Anderson-Darling test.
- In the univariate case, our KNN test(blue curve) does not behave well when n_1 is small, this procedure is only comparable to other tests when the sample size is relatively large.
- When n_1 approaches 400, all the curves reach the upper bound which is 1.00.

4 An Improved Test Using Repeated Simulation

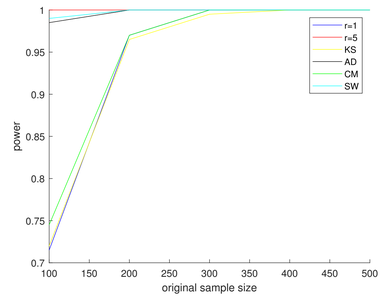
- A good way to improve the reliability of a test is to collect more information from the sample.
- However, the original sample size of observed data is usually fixed.
- So we create a method by using multiple repeated simulated samples from $f_0(x)$.
- Following is the algorithm of repeated simulation.

1. Generate $Y_i^{(s)}$(s = 1,...,r) from f_0 instead of just Y_i
2. Let $Z_i = X_i$ when $1 \le i \le n_1$, $Z_i = Y_i$ when $1 \le i \le n_2$.
3. Calculate Henze's statistic T_n and tentatively reject H_0 at level γ if

$$\sqrt{n}\left(T_{n,k}^{(s)} - E(T_{n,k}^{(s)})\right) / \sqrt{Var(T_{n,k}^{(s)})} > z_\gamma, \text{where} n = n_1 + n_2$$

4. Repeat steps 1 and 3 r times.
5. Reject H_0 at level α if you tentatively reject H_0 greater than $.5r$ times.

Note: Choose γ so that $\alpha = \sum_{j=.5r}^{r} \binom{r}{j} \gamma^j (1-\gamma)^{r-j}$



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

- Norbert Henze (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. Ann.Statist.Volume 16, Number 2(1988),772-783
- Szekely, G. J. and Rizzo, M. L. (2004) Testing for Equal Distributions in High Dimension, InterStat, Nov. (5).