Testing the Equality of the Distribution of two Datasets or How good is your MC?

Dr. Wolfgang Rolke

Dr. Angel Lopez

Promise:

- Everyone in the room will understand everything in this talk
- Some might even find it useful

$X_i,..,X_n,Y_1,..,Y_m$ in R^D

$X_1,..,X_n \sim F$, $Y_1,..,Y_m \sim G$

One of these might be MC data

H_o: $F=G$ vs H_a: $F \neq G$

Many methods in $D=1$, especially if F=Normal (Kolmogorov-Smirnoff etc.)

Little if $D>1$

- Bickel, P. J. (1969). A distribution free version of the Smirnov two-sample test in the multivariate case. Annals of Mathematical Statistics 40 1-23.
- Friedman, J. H. and Rafsky, L. C.(1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Annals of Statistics 7(4) 697-717.
- Zech, G. and Aslan, B. A Multivariate Two-Sample Test Based on the Concept of Minimum Energy, Proceedings of Phystat2003, SLAC, Stanford.
- Bickel, P. J. and Breiman, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Annals of Probability 11 185-214.
- Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor coincidences. Annals of Statistics16 772-783.

Idea of Method

Consider X_j and ask: is its nearest neighbor from the X or the Y dataset?

Under H_o both possibilities are equally likely (relative to n and m)

Let Z_j be 1 if nearest neighbor is from X's, 0 otherwise

Then $Z_j \sim Ber((n-1)/(n+m-1))$

Then $\Sigma Z_j \sim Bin(n, (n-1)/(n+m-1))$

Well, almost

Extension

Let Z_{ji} be 1 if i^{th} nearest neighbor is from X's, 0 otherwise, i=1,..,k

Then $Z=\Sigma_j\Sigma_i Z_{ij} \sim \mathrm{Bin}(nk, (n-1)/(n+m-1))$

Again, almost

p-value=$P(V \geq Z)$ where $V \sim \mathrm{Bin}(nk, (n-1)/(n+m-1))$

Often needs standardizing of data

What about the “almost”?

If n and m are “small”, find null distribution using a permutation type test:

Idea: randomly reorder X’s and Y’s, divide them again in n X’s and m Y’s

Null hypothesis now true by definition

Find Z

Repeat many times (say 1000), get distribution of Z’s

Practical Questions:

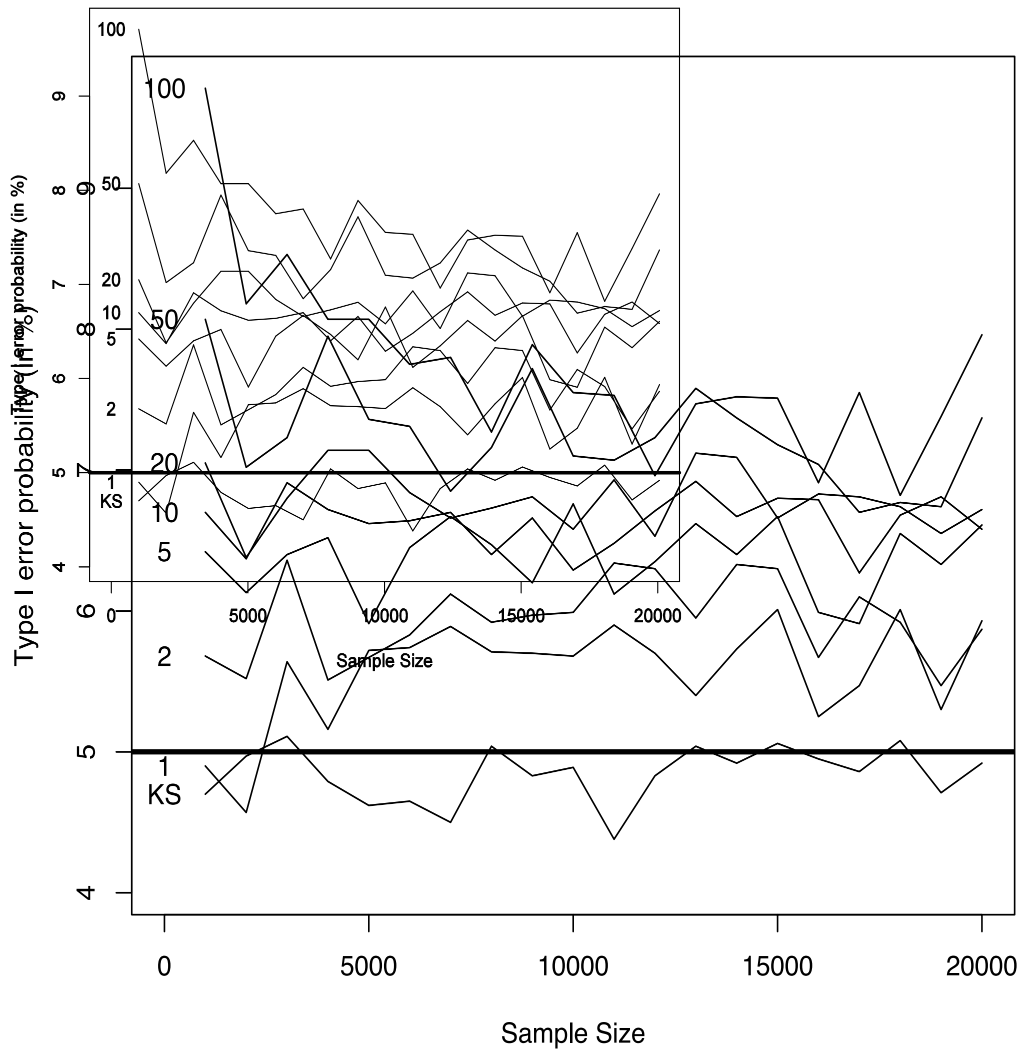
If one dataset is MC, how to choose sample size?

What k?

Binomial Approximation or Permutation test?

Answers from some mini MC

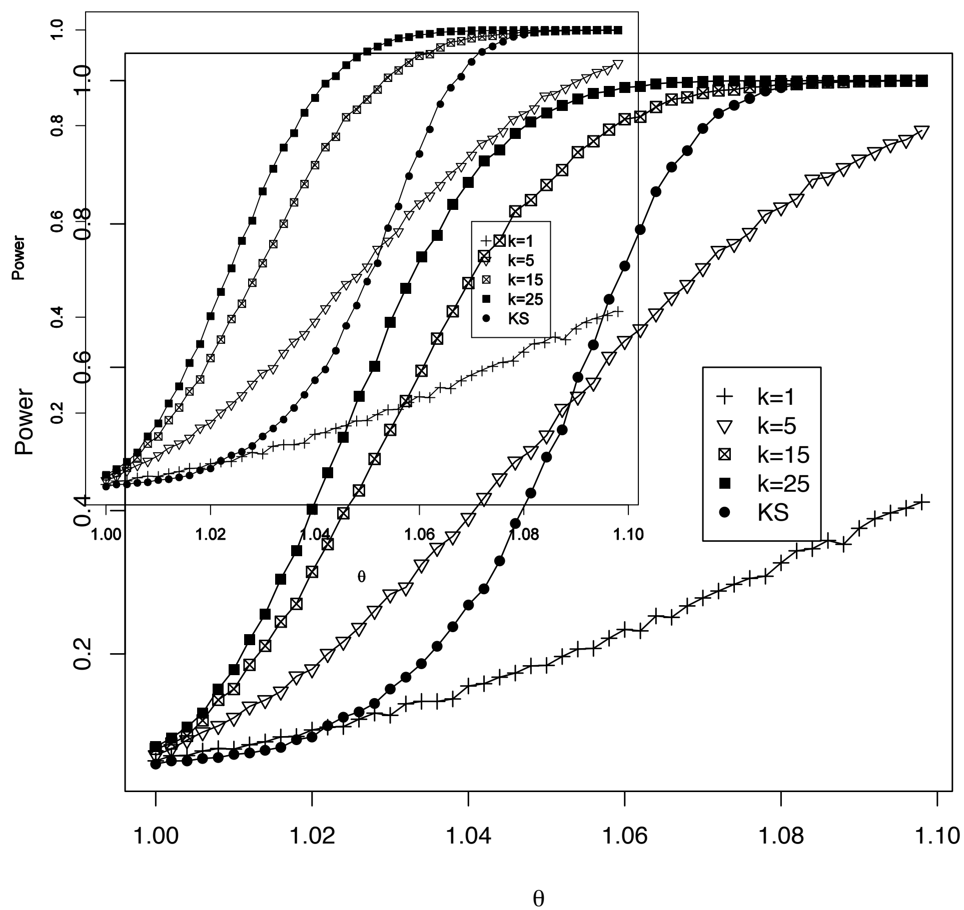
- D=1 and compare with Kolmogorov-Smirnov (KS) test.
- n=m from 1000-20000
- F=G=U[0,1]
- k=1, 2, 5, 10, 20, 50, 100
- repeat 10000 times
- nominal type I error probability 5%



→ true α goes up with k

- n=m=1000
- F=U[0,1]
- G=U[0,θ], θ from 1 to 1.1
- k=1, 5, 15, 25
- repeat 10000 times
- nominal α=5%

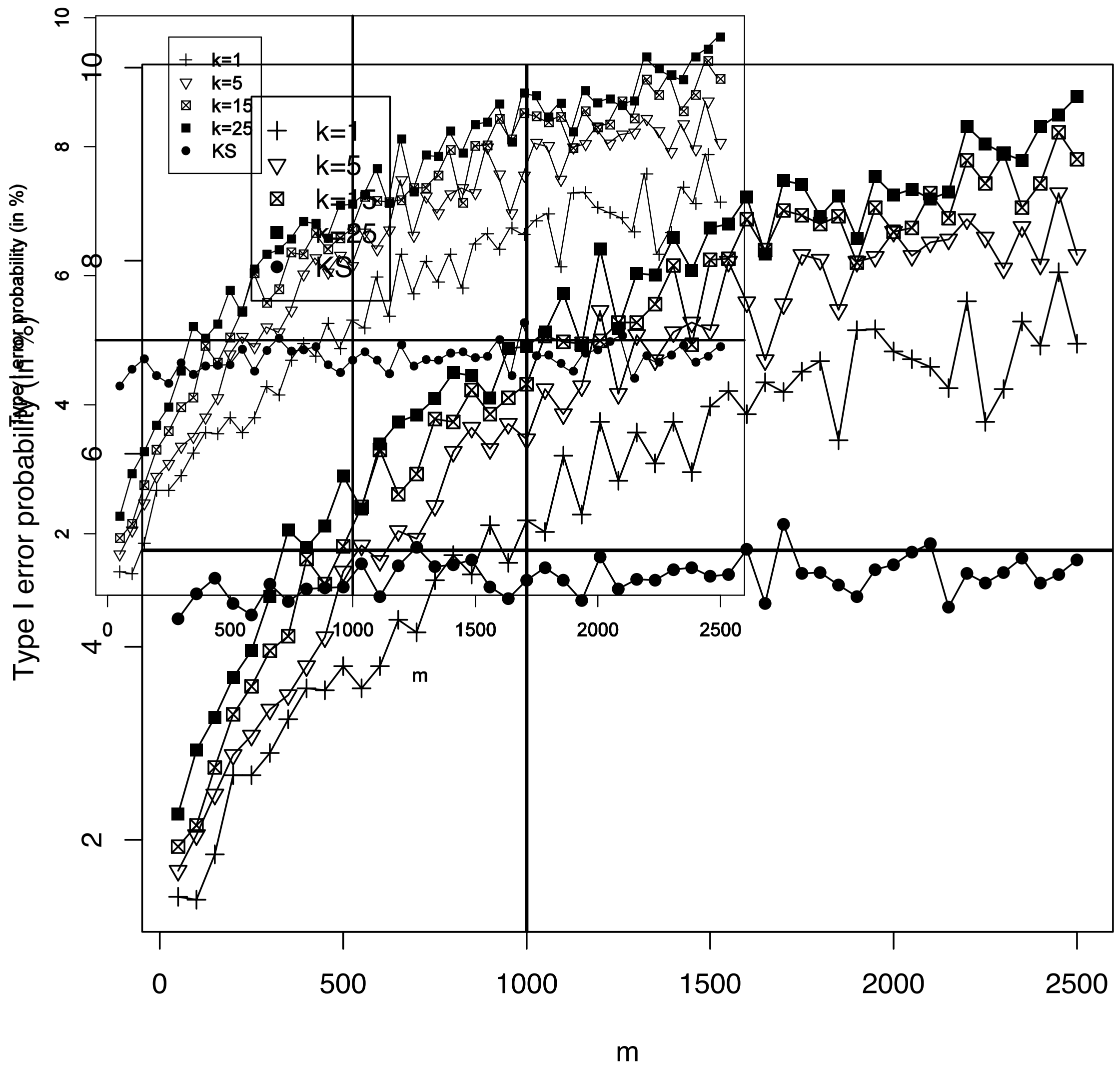
→ Recommend k=10 if n,m>1000, otherwise use permutation test



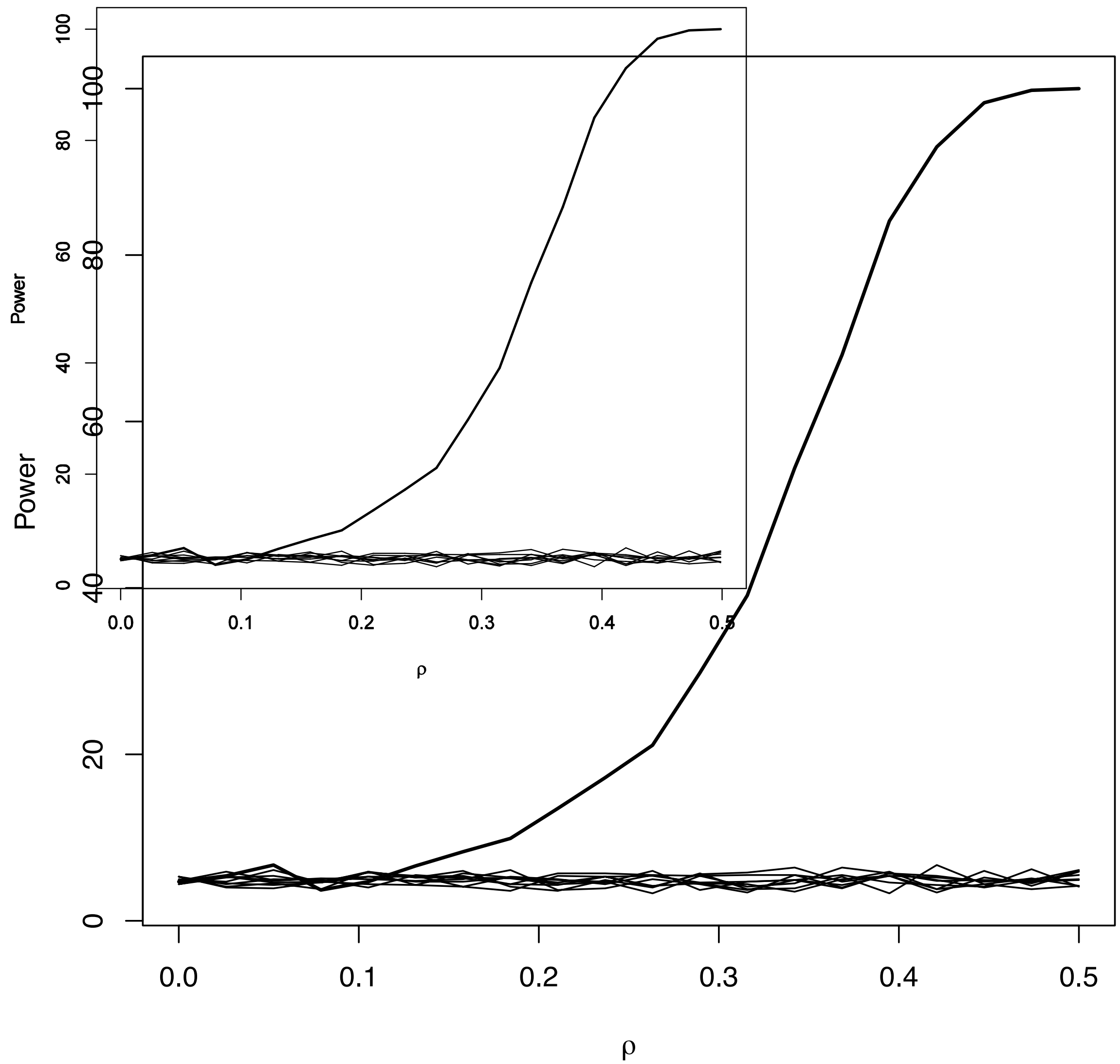
- F=G=U[0,1]
- n=1000
- m goes from 50 to 2500

→ Recommend

n=m



- n=m=1000
- k=10
- F=N(0,I) in d=9
- G=N(0,Σ) with
  cor(X_i,X_j)=ρ if |i-j|=1
  cor(X_i,X_j)=0 if |i-j|>1



Implementation

- C++ routine is available
- uses Binomial approximation or permutation method
- uses a simple search for the k-nearest neighbors
- more sophisticated and faster routines exist and could also be used in combination with our code, see for example Friedman, Baskett and Shustek (1975) An Algorithm for Finding Nearest Neighbors. IEEE Transactions on Computers 24, 1000-1006

Conclusion

- Method tests for equality of distributions
- Compare data to data or data to MC
- Easy to understand and implement

The End