Subset Selection for Insignificant subsets

Chenghui Zheng

2024-08-29

(a) Linear Model with Independent Predictors

$$Y_1 \sim 1.5X_1 + 1.5X_2 + 2X_3 + 2X_4 + 2X_5 + 3X_6 + 4X_7 + 5X_8 + \epsilon$$

(b) Linear Model with Correlated Predictors

$$Y_2 \sim 1.5X_1 + 1.5X_2 + 2X_3 + 2X_4 + 2X_5 + 3X_6 + 4X_7 + 5X_8 + \epsilon$$

Where $X_1 \not\perp\!\!\!\perp X_2$ and $\text{cov}(X_1, X_2) = 0, 0.5, 0.75, 0.9$ respectively.

(c) Linear Model with Correlated Predictors and Different SNR

$$Y_3 \sim 1.5X_1 + 1.5X_2 + 2X_3 + 2X_4 + 2X_5 + 3X_6 + 4X_7 + 5X_8 + \epsilon$$

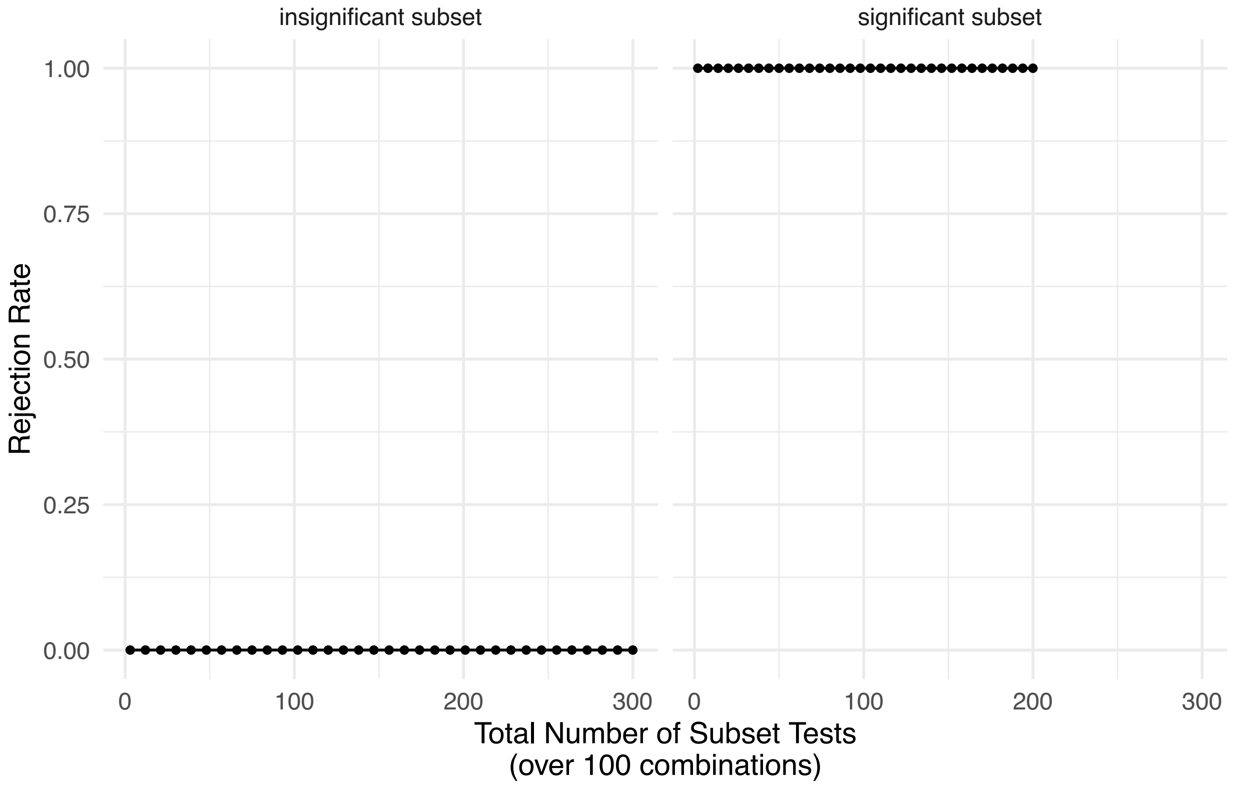
Where $\text{cov}(X_i, X_j) = \rho^{|i-j|}$ and $\epsilon \sim N(0, \sigma^2)$ with $\sigma^2 = 0.1, 0.5, 0.75, 2.1$.

(d) Non-linear Model

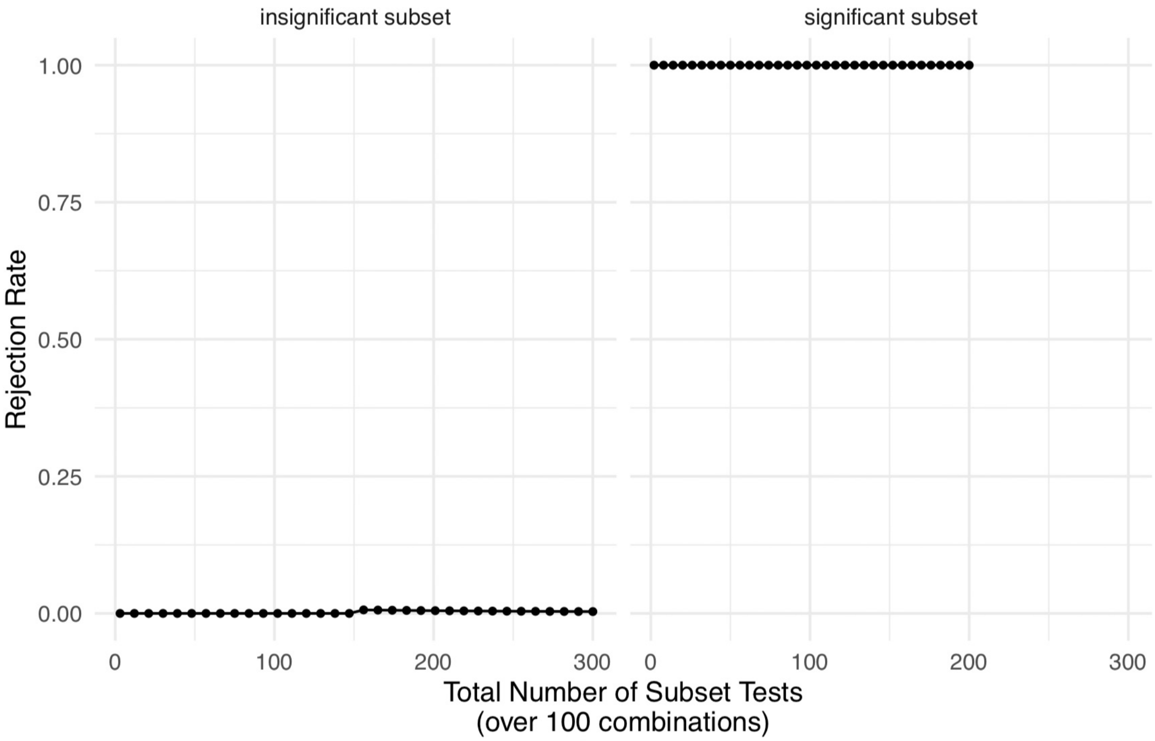
$$Y_4 \sim 2X_1^2 + 2\cos(4X_2) + \sin(X_3) + \exp(X_4/3) + 3X_5 + X_6^3 + 5X_7 + \max(0, X_8)$$

GCM Subsets selection(old T(n),100 simulations, each 500 instances)

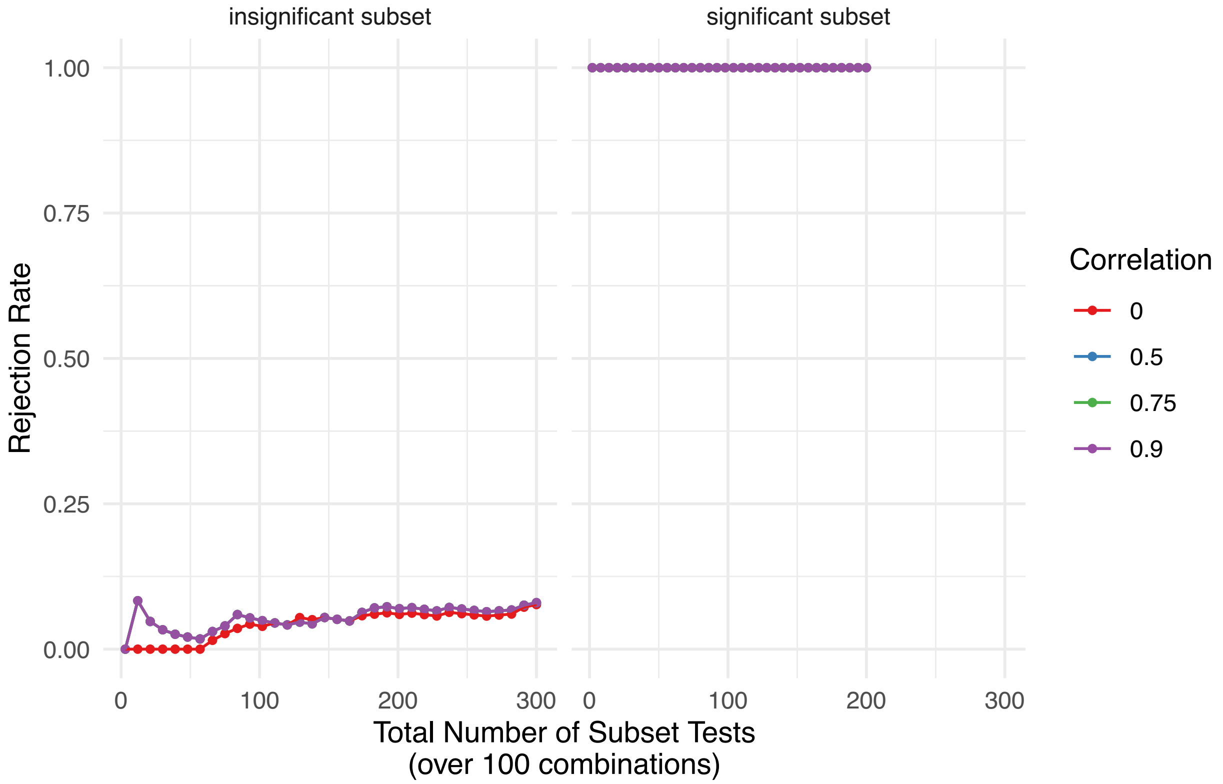
Case a): old T(n)



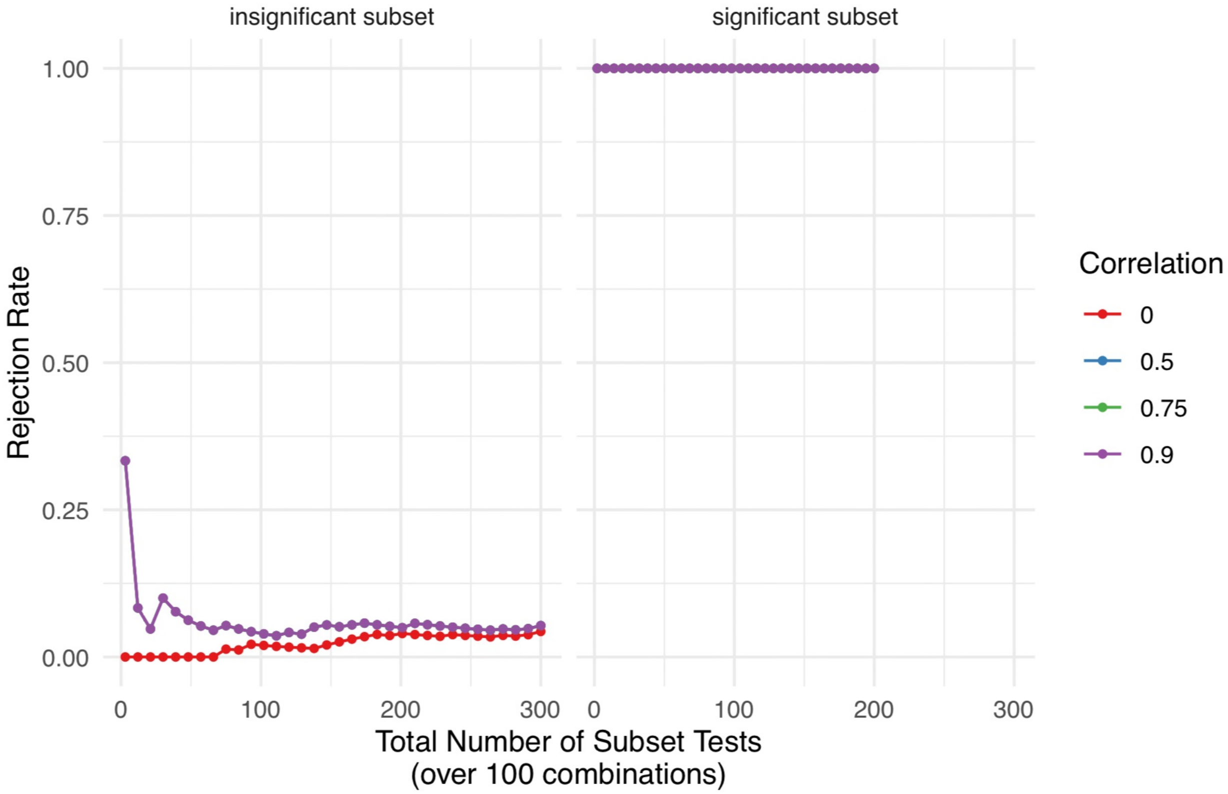
new T(n) Case a):



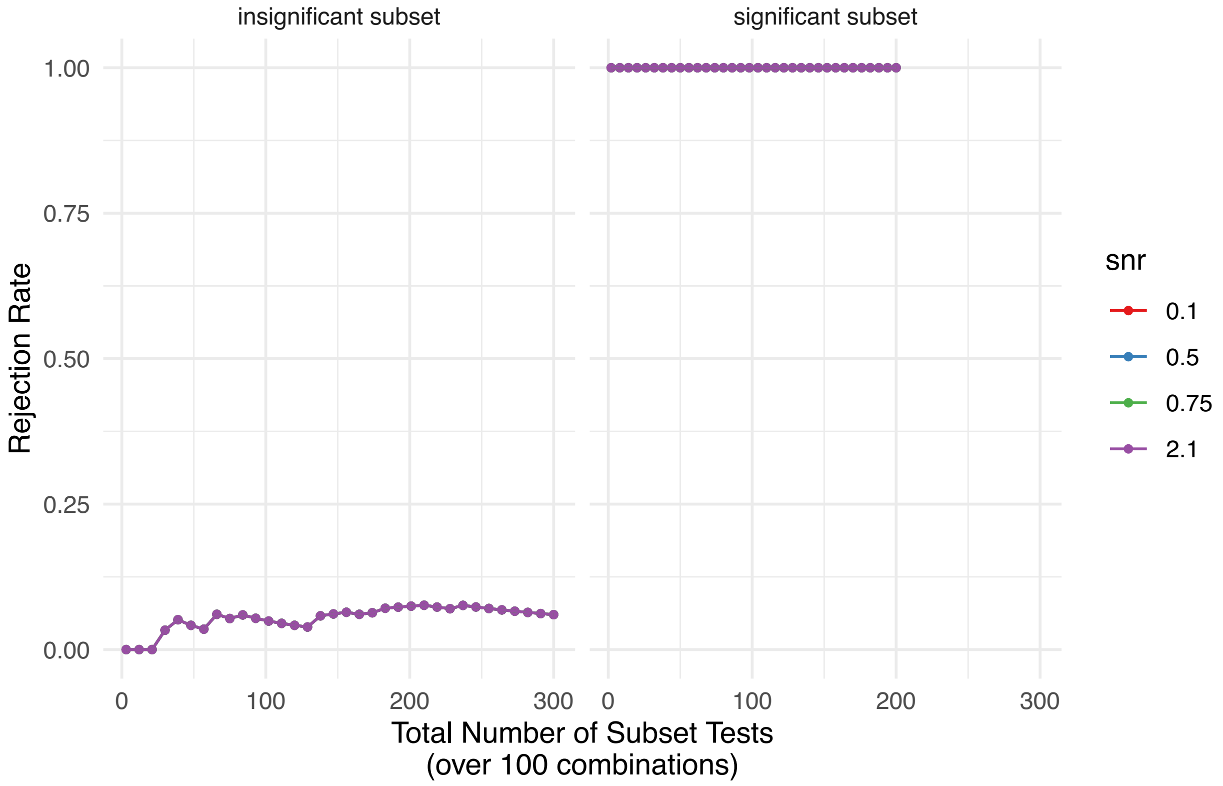
Case b): old $T(n)$



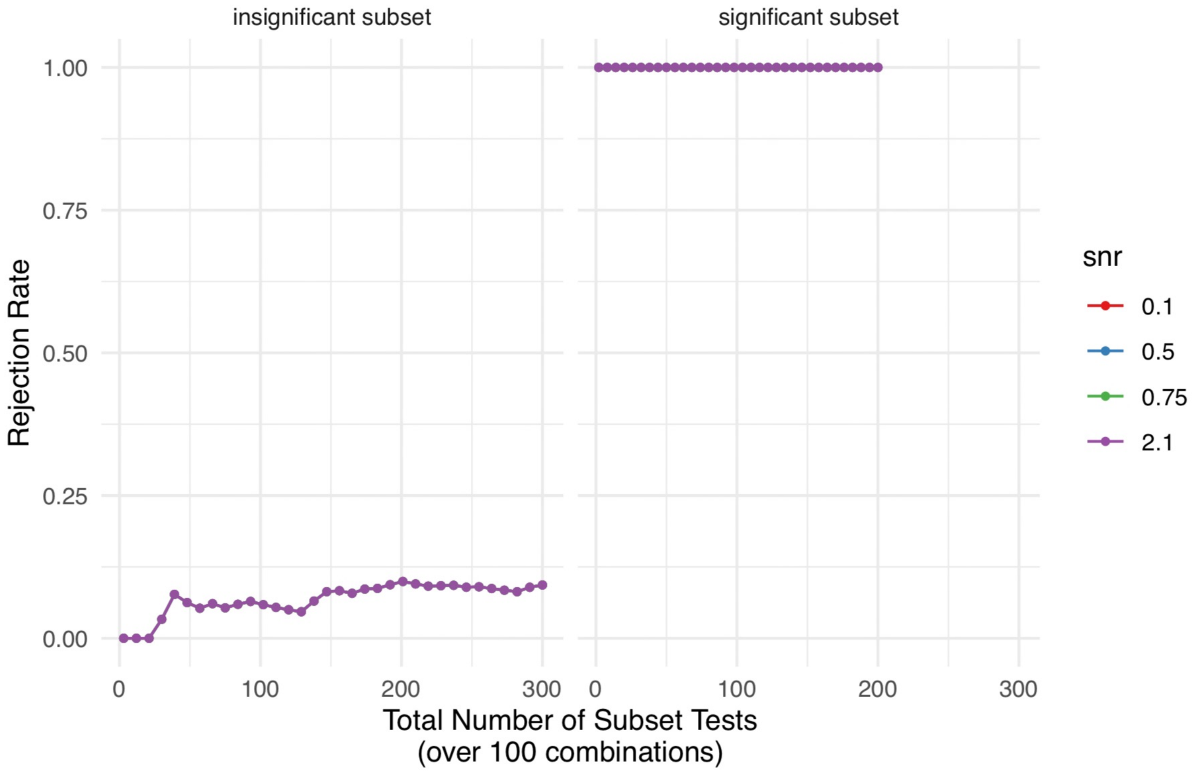
Case b): new $T(n)$



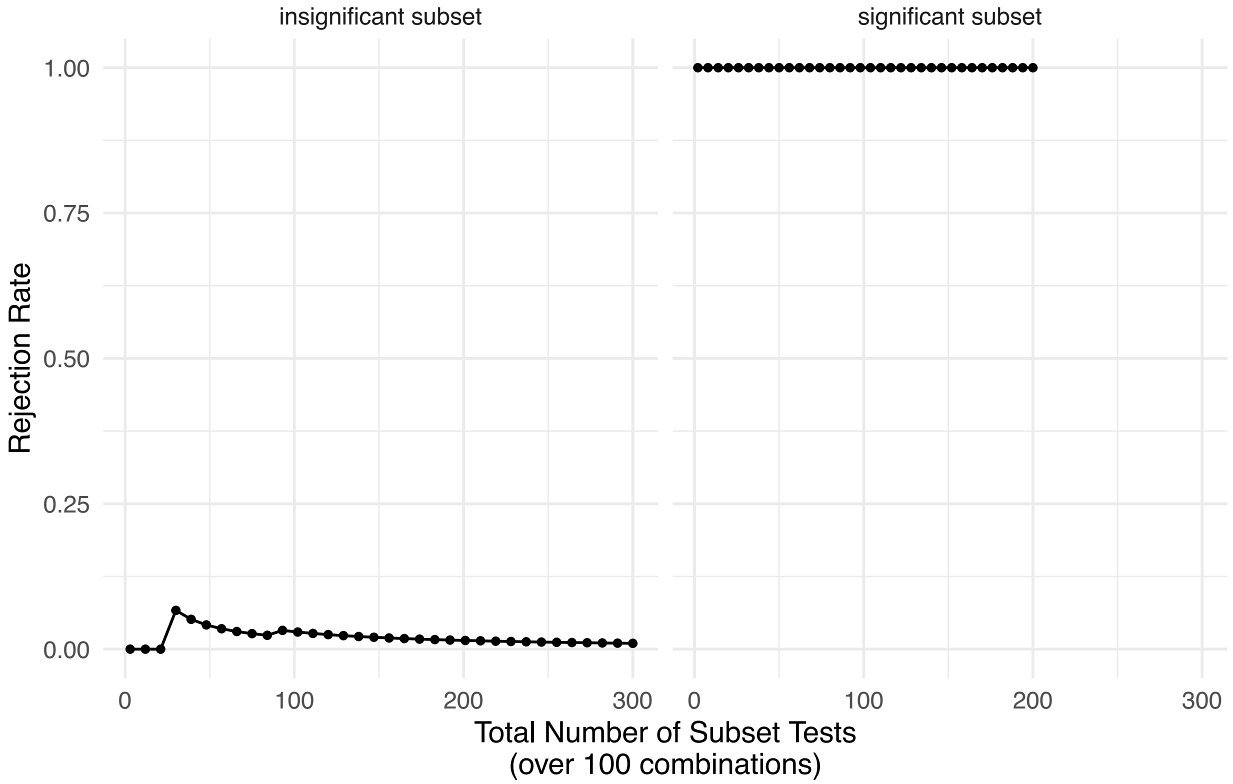
Case c): old T(n)



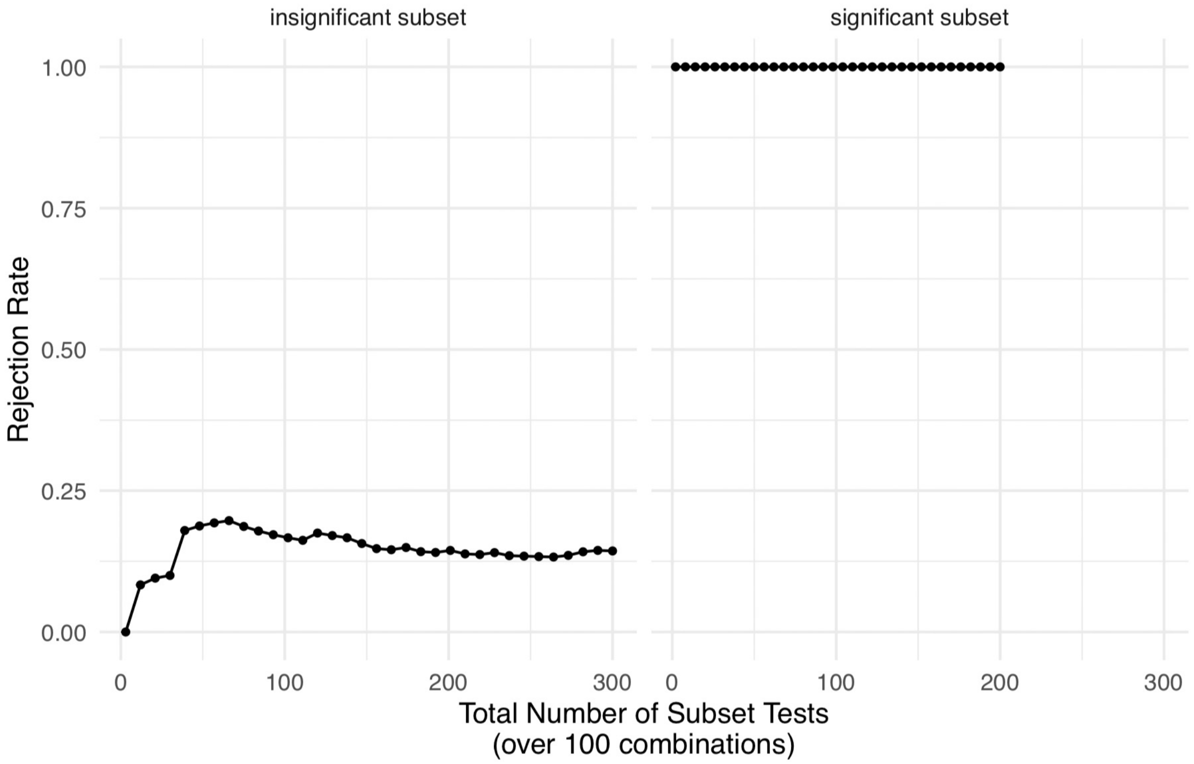
Case c): new T(n)



Case d): old T(n)

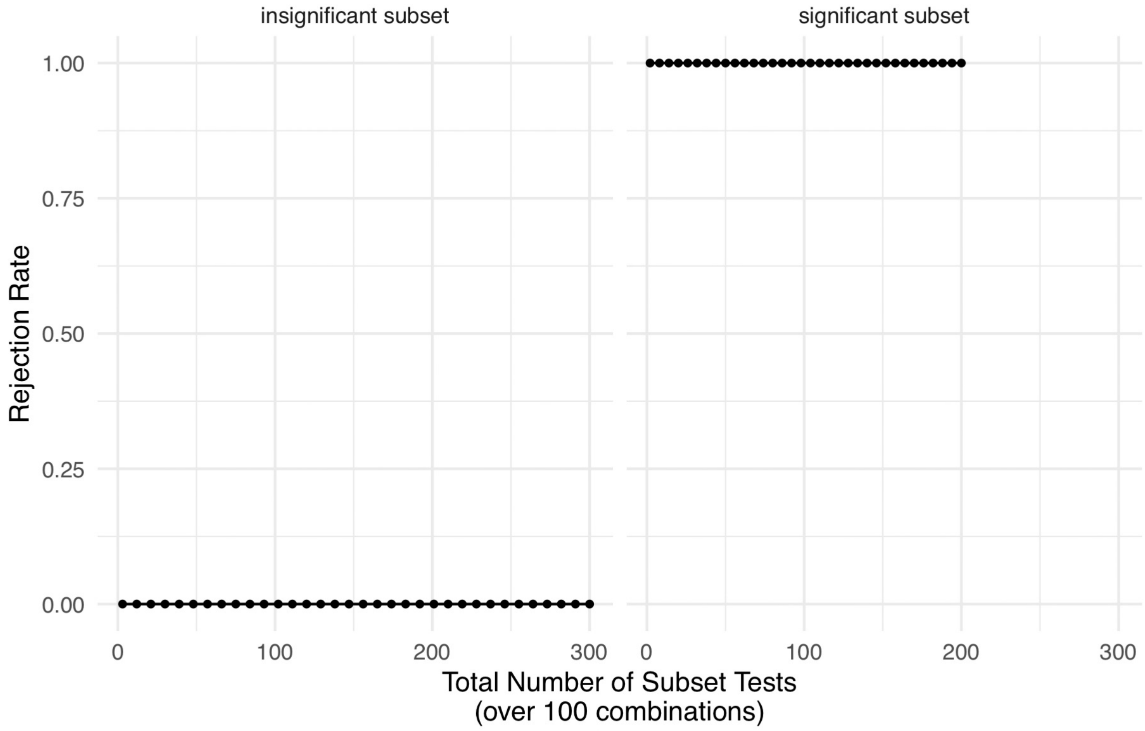


Case d): new T(n)

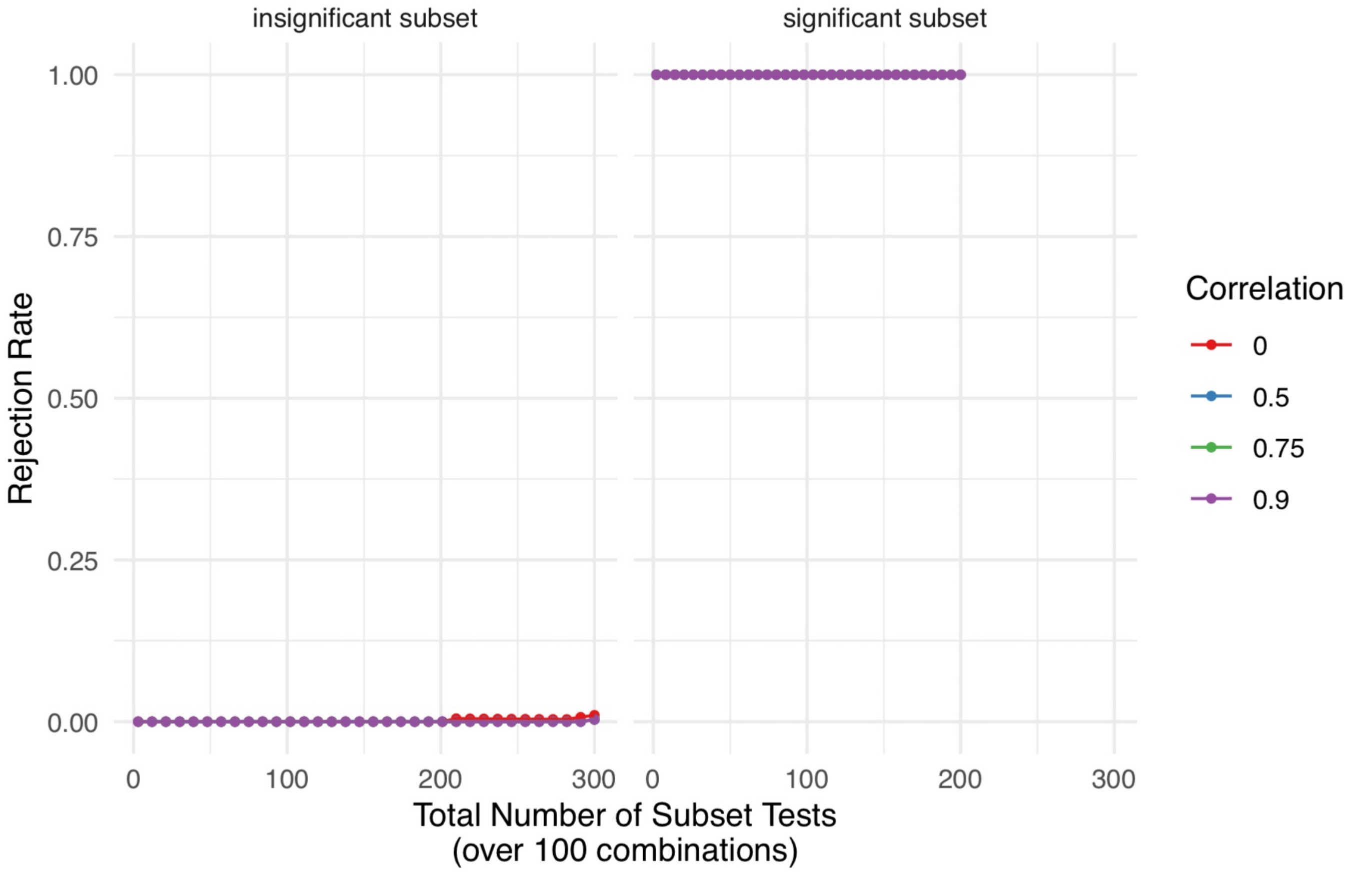


Loco

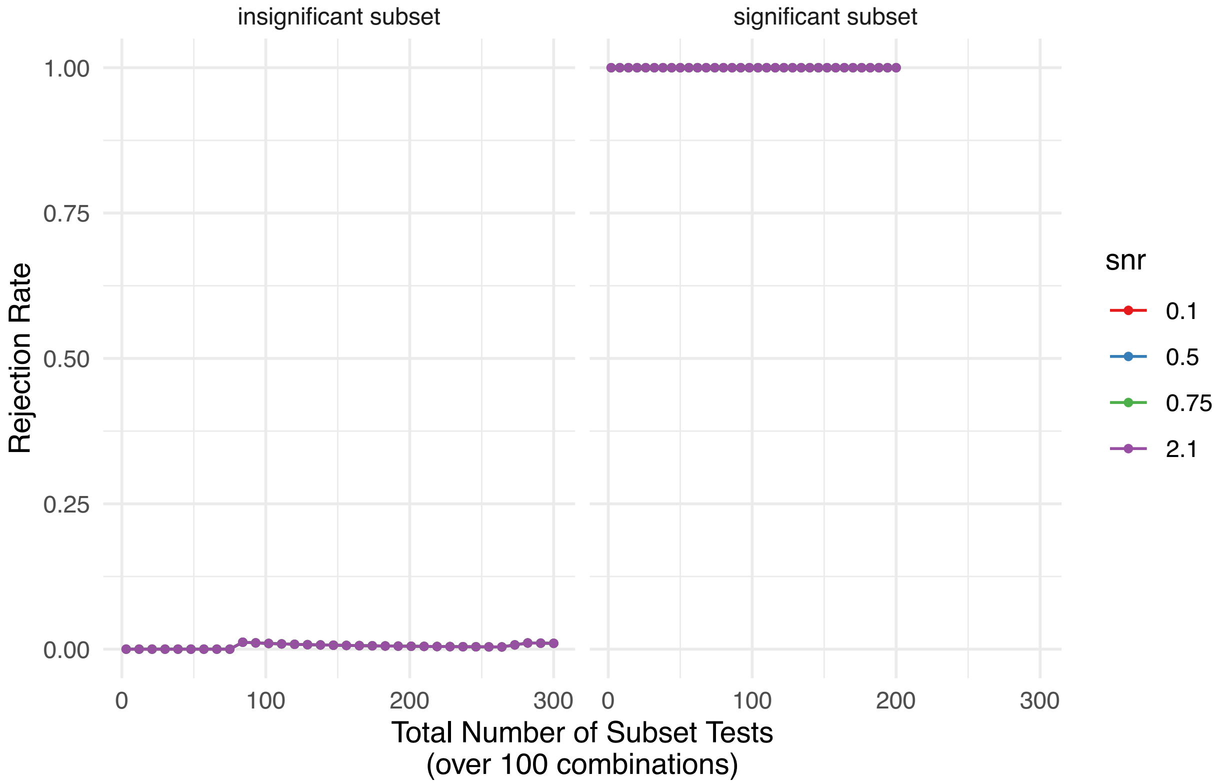
Case a):



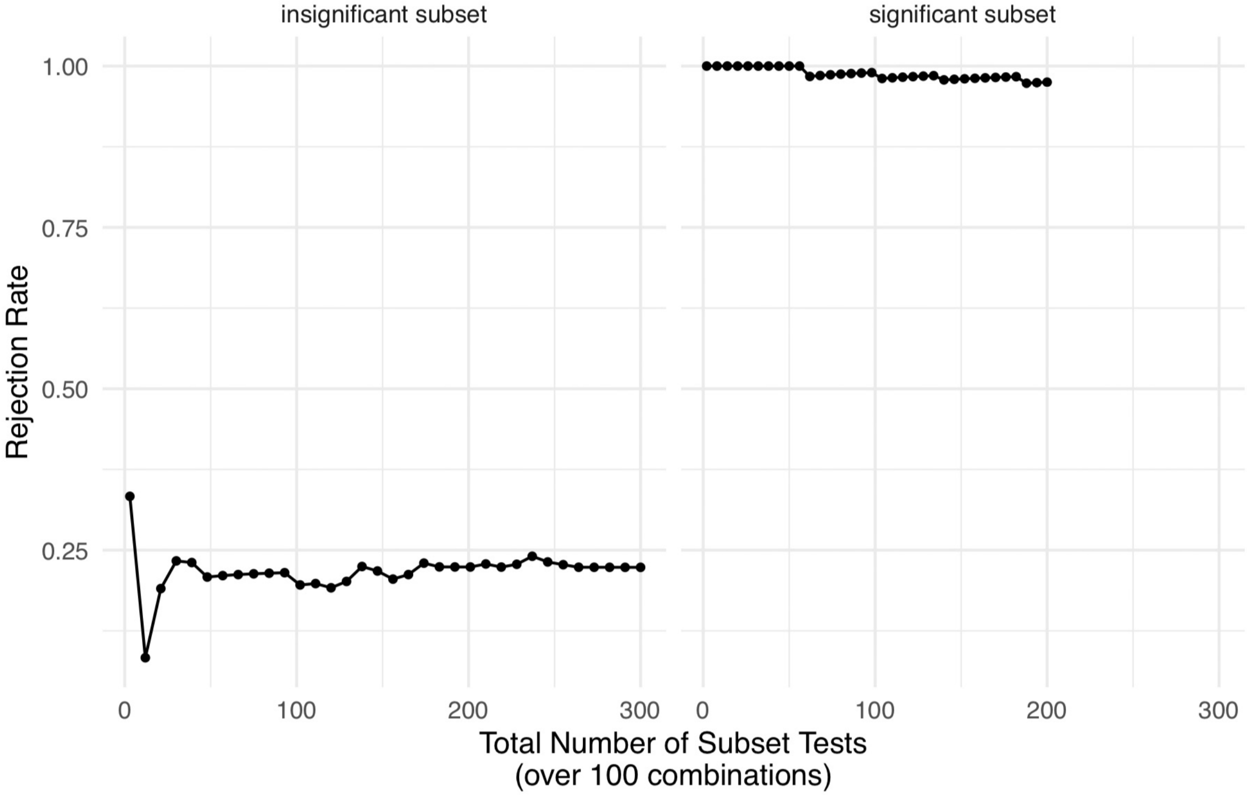
Case b):



Case c):


insignificant subset
significant subset
Rejection Rate
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Total Number of Subset Tests
(over 100 combinations)
snr
0.1
0.5
0.75
2.1

Case d):


insignificant subset
significant subset
Rejection Rate
1.00
0.75
0.50
0.25
0
100
200
300
Total Number of Subset Tests
(over 100 combinations)