Simulation4

Mengqi Liu

Jul 21, 2023

Recap

- N: number of samples one time.
- M: number of bins.
- H0: $X \perp\!\!\!\perp Y \mid Z$, H1: $X \not\perp\!\!\!\perp Y \mid Z$
- Methods:
  - "Cor_kernel" is the method put forward in "Local permutation tests for conditional independence" and I revised it a little bit to make it suitable for continuous X, Y, Z according to page 20 of "Minimax Optimal Conditional Independence Testing"(which is also one of cited paper of the former).
  - "Linear_reg_y": regress Y on X, Z and take the *absolute* coefficient of X as the test statistic.
  - "Linear_reg_x": regress X on Y, Z and take the *absolute* coefficient of Y as the test statistic.
  - "Double_reg": regress Y on Z and regress X on Z separately. Take the *absolute* correlation between residuals from two linear regressions as the test statistic.
  - "double_Cor_kernel": double-binning method (each bin will be divided into sub bins).
  - "Linear_reg_y_sub", "Linear_reg_x_sub", "Double_reg_sub": conduct "Linear_reg_y", "Linear_reg_x" and "Double_reg" with permutation in sub-partitions.
- $\alpha = 0.05$
- $X = f_x(Z) + \epsilon, Y = f_y(Z) + \epsilon$
- Noise ϵ:
  - various a
  - H0
    - normal: $N(Z, a)$
    - uni: $\mathrm{Unif}([-a, +a])$
    - poi: Poisson$(1/2)$
    - skewed_normal: $N(Z, a)$
  - H1
    - normal: $N\left([0,0], \begin{pmatrix} a_1 & 0.8 \cdot \sqrt{a_1 a_2} \\ 0.8 \cdot \sqrt{a_1 a_2} & a_2 \end{pmatrix}\right)$
    - skewed_normal: $N\left([0,0], \begin{pmatrix} a_1 & 0.8 \cdot \sqrt{a_1 a_2} \\ 0.8 \cdot \sqrt{a_1 a_2} & a_2 \end{pmatrix}\right)$, skewness = $[5, -5]$
- $N = 100, Z \sim \mathrm{Unif}([0, 10))$, $M \in \{2, 5, 10, 16, 25, 50\}$.

Smoothness

- is depicted by distance between $X \mid Z$ and $X|\tilde{Z}$, where $\tilde{Z}$ is the representative in each bin.

- different distance metrics:

  - Hellinger distance / Rényi Divergence (*Local permutation tests for conditional independence*)
  - TV distance (*Nearest-Neighbor Sampling Based Conditional Independence Testing*)

Questions

1. double binning
2. choice of number of sub-partitions
3. smoothness: yes+no no+no yes+yes
4. choice of xfunc, yfunc

New gains

- The number of sub-partitions in each interval will effect the power and type-1 control of "cor_kernel_sub" method.

- High type-I error without double binning.

  - sensitive to different distributions.
  - "Linear_reg_x" is the worst
  - experiment 5's last figure : special failure
- Weak advantage of power.

  - distributions matter.
  - varying degrees of smoothness matter. (experiment 5)
- "Linear_reg_y_sub", "Linear_reg_x_sub" and "Double_reg_sub" have almost the same performance. (have checked the code)

- Failure in part II: $X|Z$ is smooth and $Y|Z$ is the more non-smooth due to polynomial f_y\$.

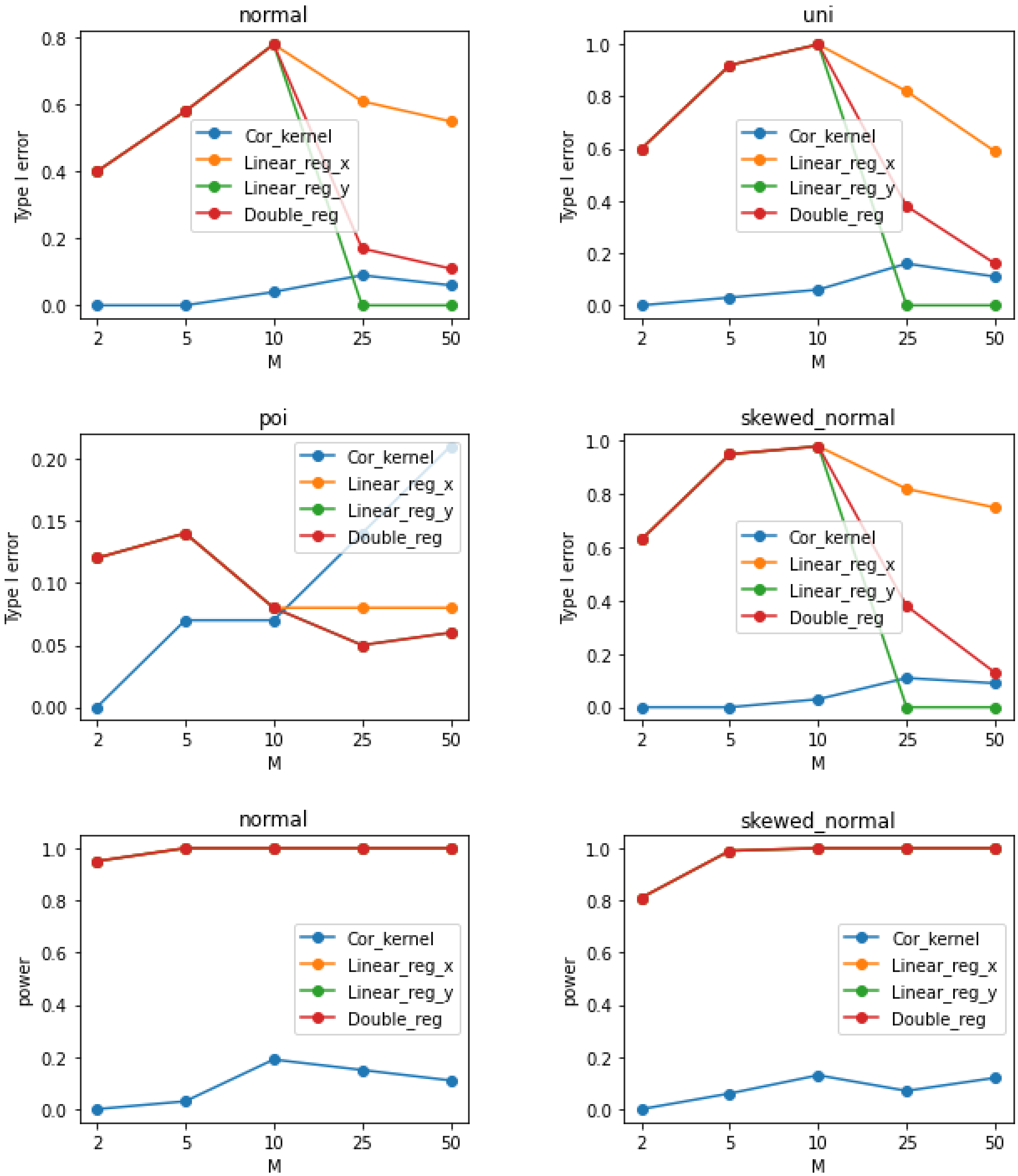
Part I: Linear f_x & f_y in Z

Smooth $X|Z$ and non-smooth $Y|Z$

---

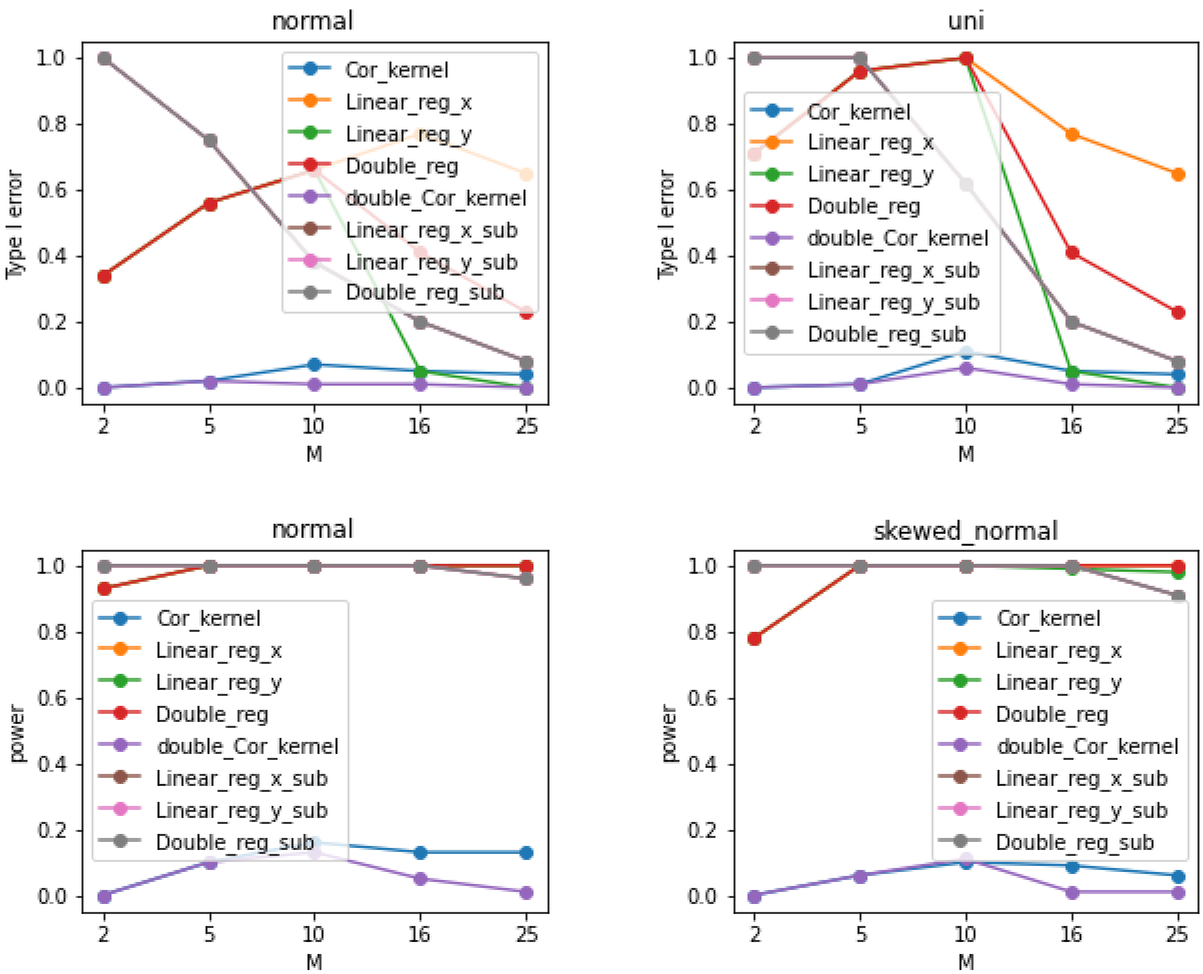
experiment 1

$f_x(Z) = Z$, $f_x(Z) = Z$, sub $= 4$, $\epsilon_x \sim N(\cdot, 1)$, $\epsilon_y \sim N(\cdot, 0.1)$.



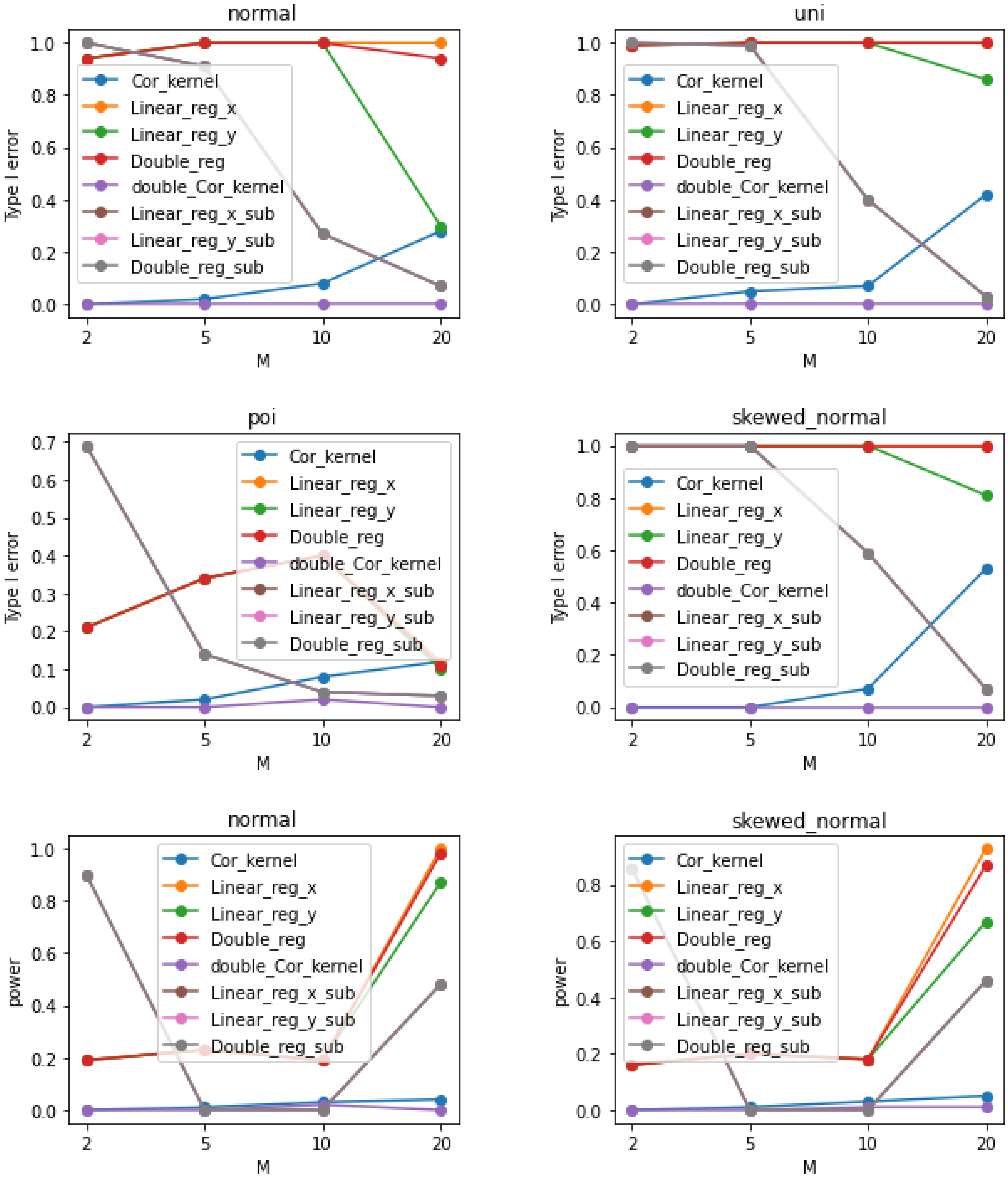
experiment 2

$f_x(Z) = Z$, $f_x(Z) = Z$, sub $= 2$, $\epsilon_x \sim N(\cdot, 1)$, $\epsilon_y \sim N(\cdot, 0.1)$.



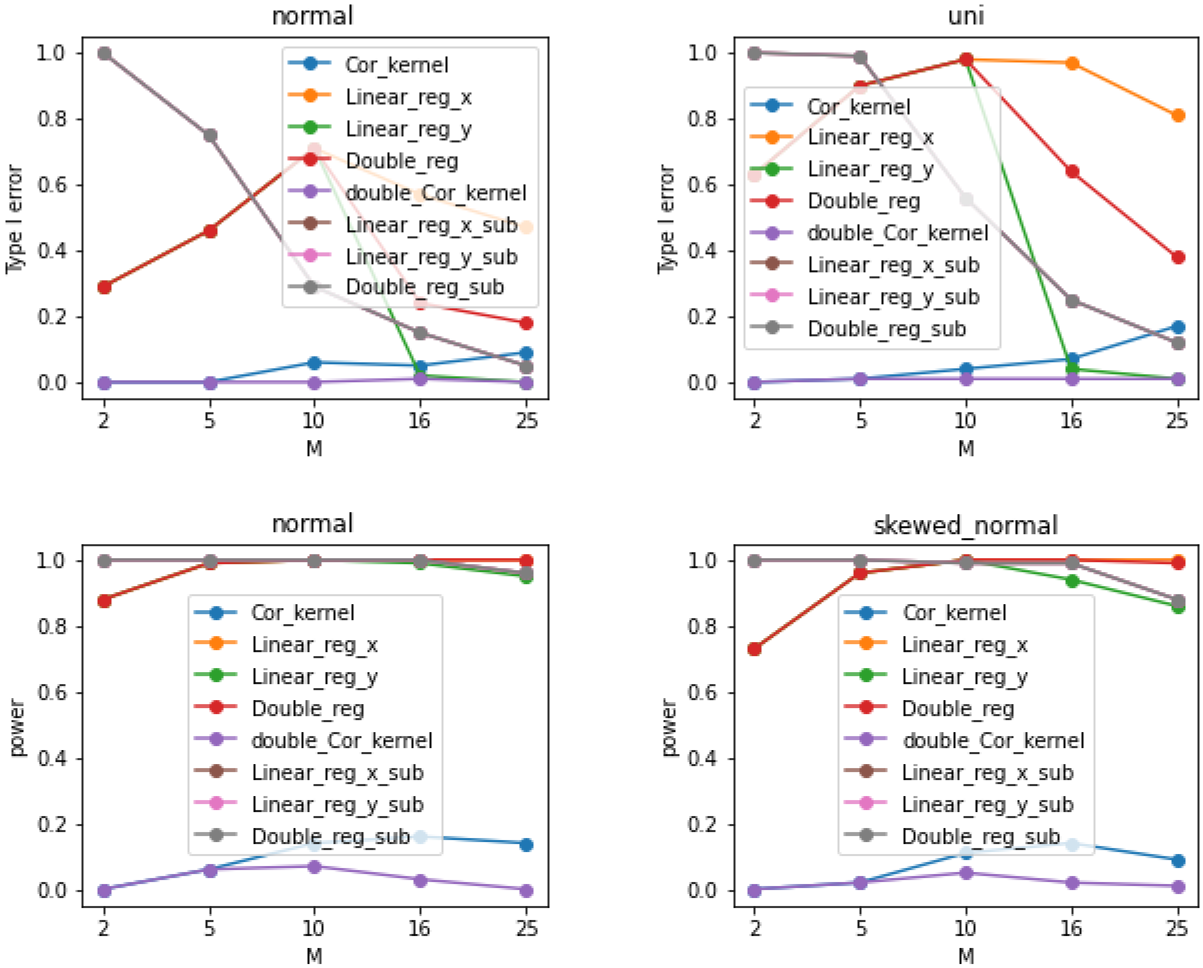
experiment 3

$f_x(Z) = Z$, $f_x(Z) = -Z$, sub $= 4$, $\epsilon_x \sim N(\cdot, 1)$, $\epsilon_y \sim N(\cdot, 0.1)$.



experiment 4

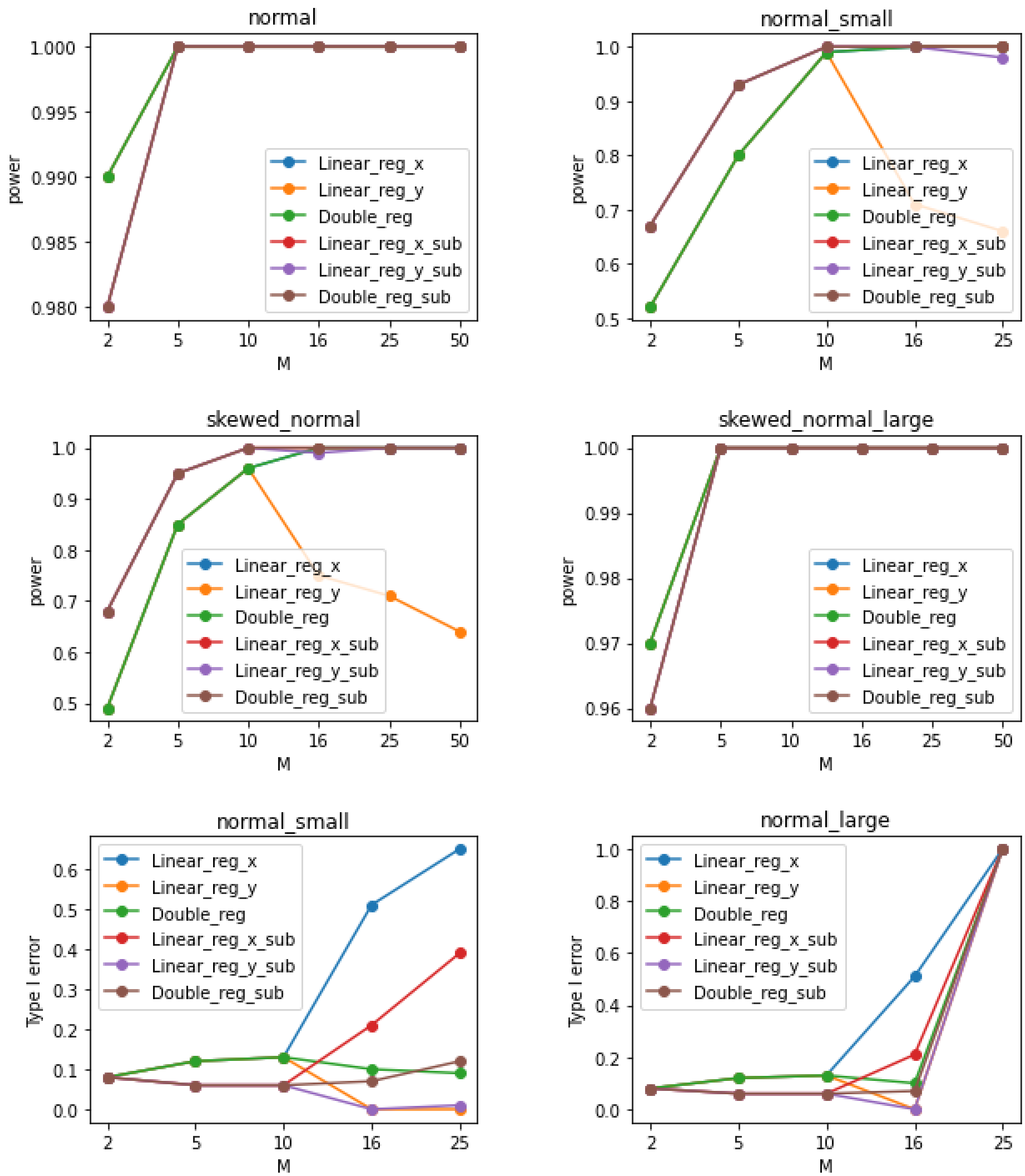
$f_x(Z) = Z$, $f_x(Z) = 5Z$, sub $= 2$, $\epsilon_x \sim N(\cdot, 1)$, $\epsilon_y \sim N(\cdot, 1)$.



experiment 5

$f_x(Z) = Z$, $f_x(Z) = Z$, sub $= 2$.

Distribution	Var(ϵ_x)	Var(ϵ_y)
normal / skewed_normal	1	0.1
normal_small	5	0.01
normal_large / skewed_normal_large	5	0.5

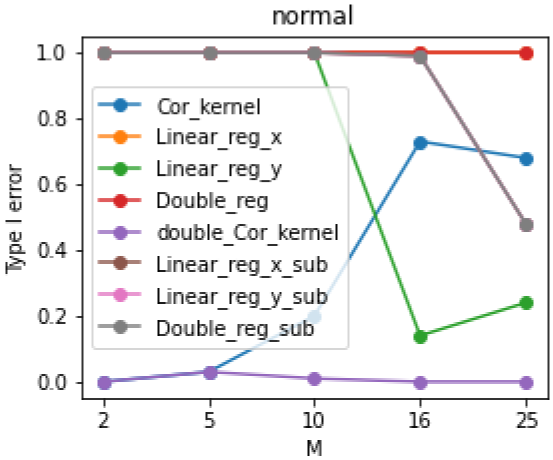


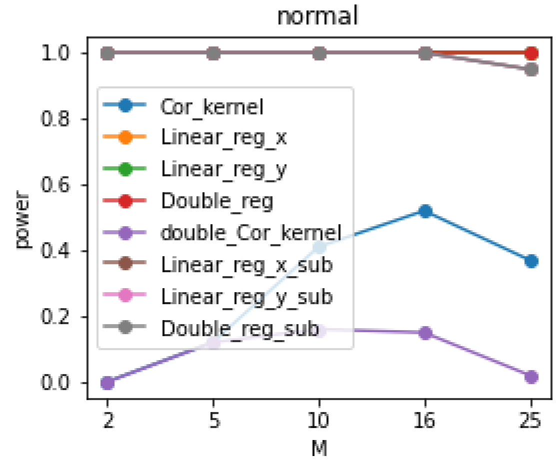
Non-smooth $X|Z$ and non-smooth $Y|Z$

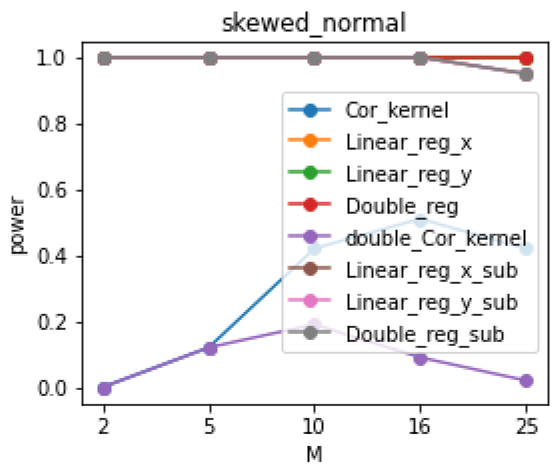
---

experiment 6

$f_x(Z) = Z$, $f_x(Z) = Z$, sub $= 2$, $\epsilon_x \sim N(\cdot, 0.1)$, $\epsilon_y \sim N(\cdot, 0.1)$.





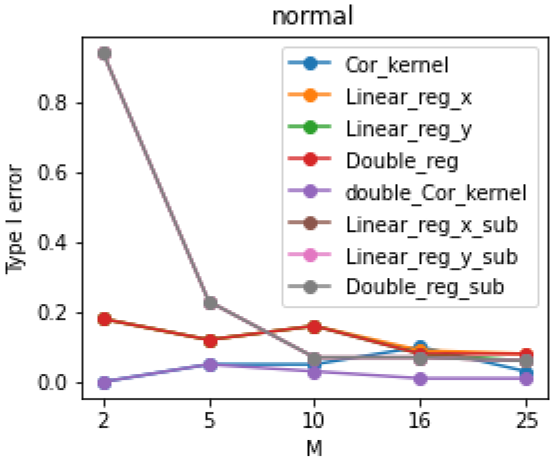


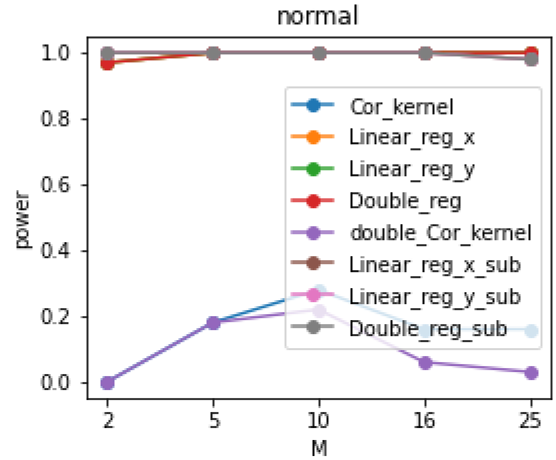
Smooth $X|Z$ and smooth $Y|Z$

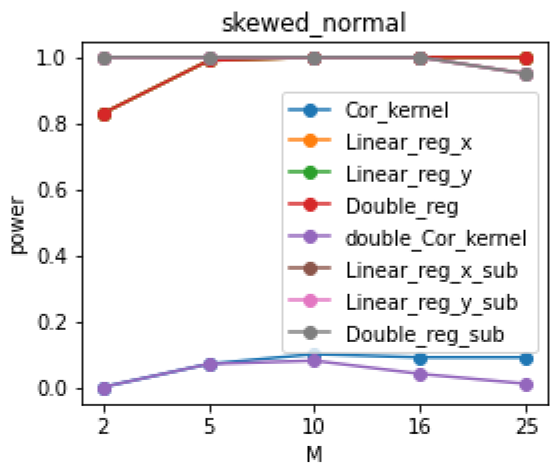
---

experiment 7

$f_x(Z) = Z$, $f_x(Z) = Z$, sub $= 2$, $\epsilon_x \sim N(\cdot, 1)$, $\epsilon_y \sim N(\cdot, 1)$.



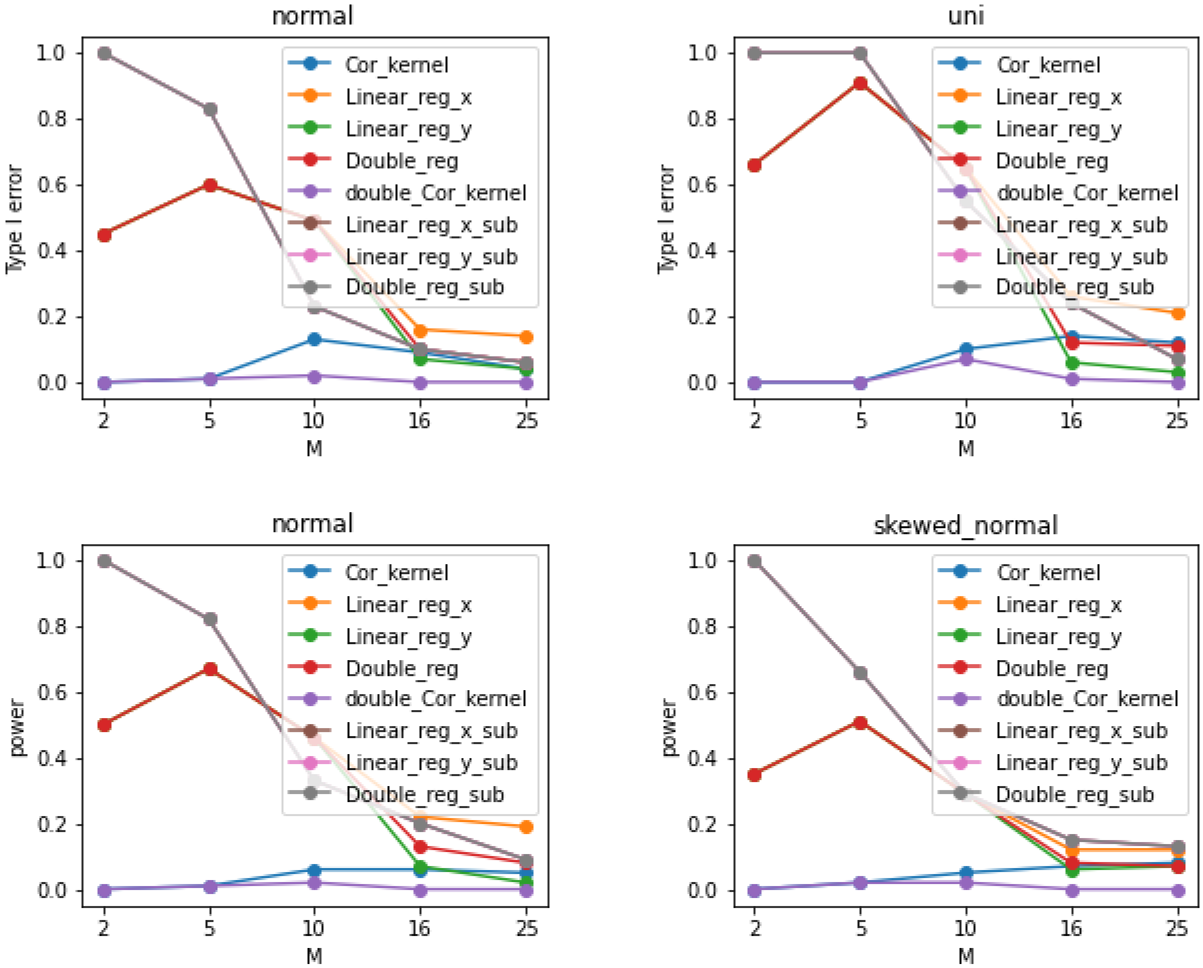




Part II: Linear f_x & non-linear f_y in Z

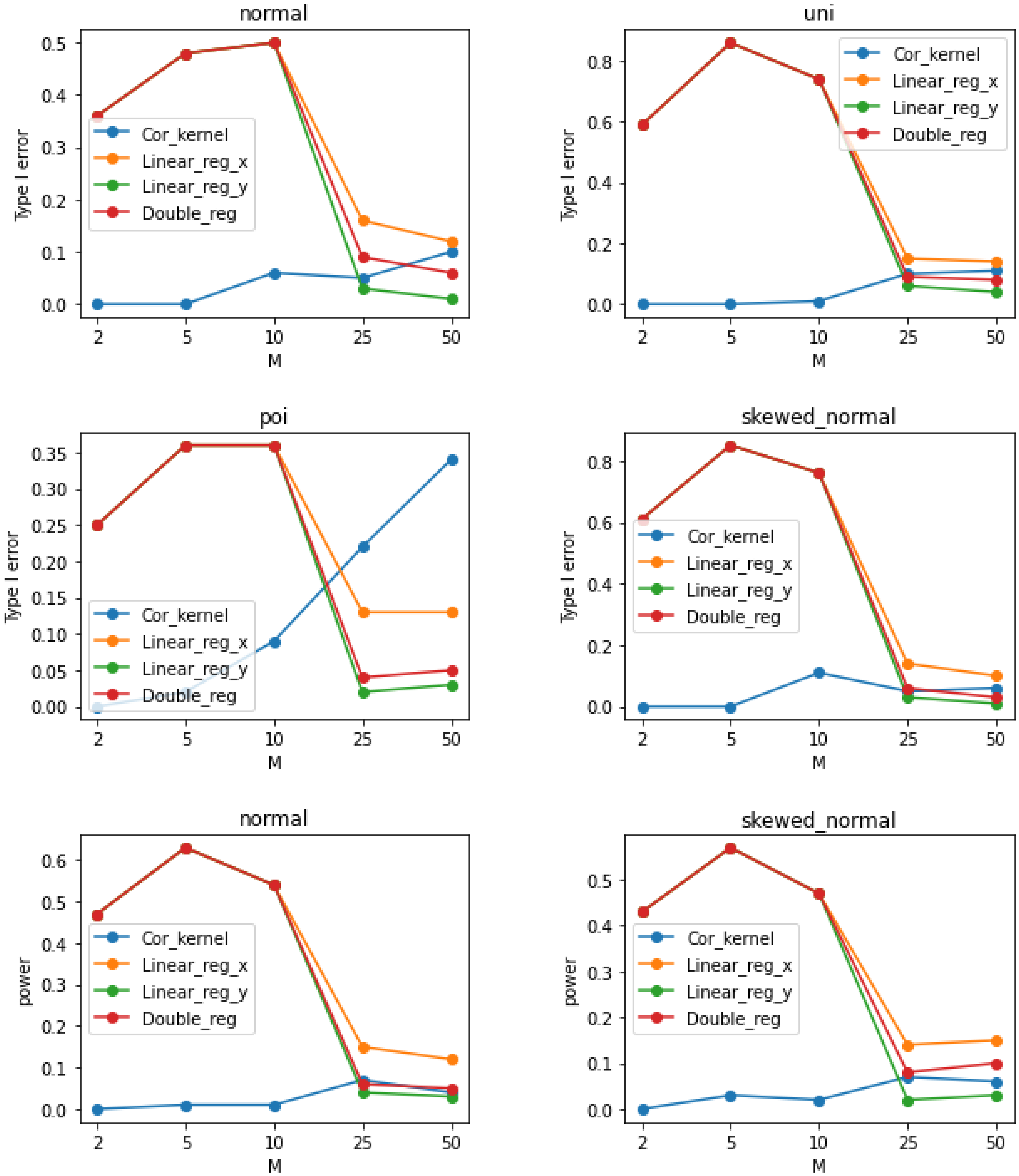
experiment 8

$f_x(Z) = Z$, $f_x(Z) = Z + 3Z^2 + 2$, sub $= 2$, $\epsilon_x \sim N(\cdot, 1)$, $\epsilon_y \sim N(\cdot, 0.1)$.



experiment 9

$f_x(Z) = Z$, $f_x(Z) = Z + 3Z^2 + 2$, sub $= 4$, $\epsilon_x \sim N(\cdot, 1)$, $\epsilon_y \sim N(\cdot, 0.1)$.



normal
uni
poi
skewed_normal
normal
skewed_normal
Type I error
power
M
Cor_kernel
Linear_reg_x
Linear_reg_y
Double_reg
2
5
10
25
50