1 Introduction

Collinearity, the non-independence of predictor variables, may cause adverse effects on multiple linear regression, such as the inaccurate estimates of regression parameters due to high variance, the misleading t-test statistics, imprecise prediction on test data set, and etc [1, 2]. In this paper, we attempt to clarify the conditions under which collinearity is problematic, and quantitatively analyze how pronounced the deleterious effects of collinearity are. To be specific, we design a Monte Carlo simulation experiment to study the roles of four important factors (i.e., the collinearity structure, the values of true regression coefficients $\boldsymbol{\beta}$, the sample size N and the strength of errors R^2) in influencing collinearity's effects. The implementation code can be found in this link.

2 Simulation Methodology and Experimental Results

In this section, we first discuss the interactions of collinearity structure with the other three factors (i.e., N, $\boldsymbol{\beta}$, and R^2) (Section 2.1) and then delve deeper into the effects of collinearity structures. We focus on three frequently-seen collinearity structures. The first one is that regression predictors come from a multivariate normal distribution with covariance matrix specified by researchers. In this case, we analyze how the sign and magnitude of covariance between two predictors affects the variance of coefficient estimates (Section 2.2). Then, a more complicated collinearity structure is studied, where the collinearity of predictors is constructed through functions (i.e., one predictor is expressed as a function of other predictors) (Section 2.3). Finally, we consider the case when collinearity is changing over time (Section 2.4). The detailed simulation methodology and experimental results are elaborated in each sub-section.

2.1 The Interactions of Collinearity Structure with Other Factors

We briefly introduce the experiment design for analyzing the joint effect of collinearity structure, N, $\boldsymbol{\beta}$ and R^2 on multiple linear regression, and showcase experiment outcomes and findings. Following the practice of [2], we generate the design matrix $\boldsymbol{X}$ from a multivariate normal distribution consistent with the specified covariance matrix $\boldsymbol{C}$ (in this section, $\boldsymbol{C}$ is a 4×4 symmetric matrix), and calculate the response $\boldsymbol{Y}$ according to formula $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $Var(\boldsymbol{\epsilon}) = f^2 * s^2$, $s^2 = \boldsymbol{\beta}^T cov(\boldsymbol{X})\boldsymbol{\beta}$ and $f = (\frac{1}{R^2} - 1)^{\frac{1}{2}}$. This data generating framework enables us to configure the level of collinearity, $\boldsymbol{\beta}$, N and R^2 in a fine-grained manner. (The detailed configuration can be found in Appendix A. There are 192 different combinations of the design factors in total.) A Monte Carlo simulation is then performed to compute the estimated standard deviation (SD) [1] [2] of every coefficient for each combination of collinearity level, $\boldsymbol{\beta}$, N and R^2.

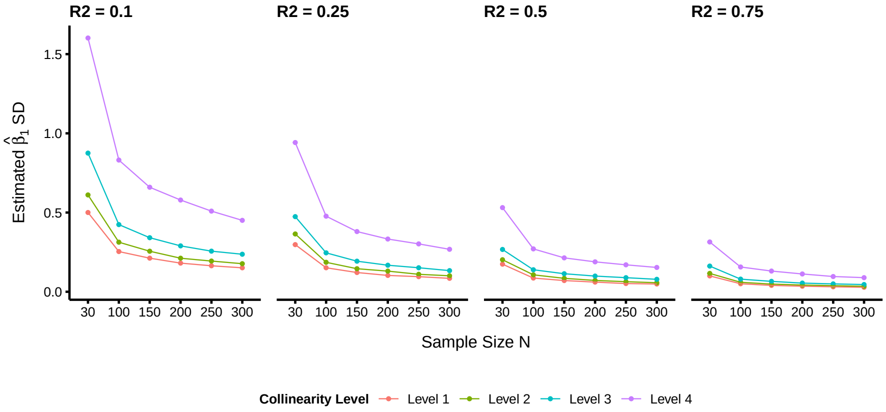

Figure 1: Joint Effect of Collinearity Structure, N and R^2 on Multiple Linear Regression: The higher the $\hat{\beta}_1$ SD value, the less accurate the coefficient estimate is. Level 1-4 reflect increasing levels of collinearity.

Figure 1 reports the estimated $\hat{\beta}_1$ SD for different levels of collinearity and sample size for each of the four levels of R^2. From

[1] The estimated SD of coefficient $\hat{\beta}_1$ is $[\sum_{i=1}^{n}(\hat{\beta}_{1i} - \bar{\hat{\beta}}_1)^2/(n-1)]^{\frac{1}{2}}$, where n is the number of samples generated by Monte Carlo simulation for each factor combination, and $\bar{\hat{\beta}}_1$ is the mean of estimated $\hat{\beta}_{1i}$.

[2] The other statistics like the accuracy of coefficient/coefficient standard error, defined in [2], can also be used to measure how much the variance of regression coefficients is inflated, but they yield similar analysis results. So, for the brevity of presentation, we only display the SD of coefficients.

this figure, we get the following three observations: (1) The $\hat{\beta}_1$ SD is related inversely to R^2 and N, and proportional to the level of collinearity. As the R^2 and N increase, the overall $\hat{\beta}_1$ SD decreases regardless of collinearity level. The inverse trend can be found when collinearity level increases. (2) The collinearity will be problematic only if the simulation is at some extreme cases. From the figure, we can find that only when R^2 is 0.25 and/or sample size is at the lowest level (e.g., 30), the substantial inaccuracy in $\hat{\beta}_1$ estimation can be observed. For normal cases (e.g., $R^2 > 0.5$ and $N > 150$), the problem of collinearity may not be significant. (3) The curve of level 4 collinearity differs significantly from the curves of three lower level collinearity. As shown in the graph, across all R^2 and N, the level 4 collinearity can always bring about significant increase in $\hat{\beta}_1$ SD. But for level 1-3, at most cases, their curves are similar and even overlapped.

It's also meaningful to clarify the reason why we don't show and discuss factor $\boldsymbol{\beta}$ in Figure 1 and how the collinearity levels, N and R^2 influence the estimates of other coefficients (e.g., $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$). We conduct an ANOVA study to verify that excluding factor $\boldsymbol{\beta}$ from the discussion in this section is reasonable. The ANOVA table in Appendix B provides the analysis of how the estimation accuracy of $\hat{\beta}_1$ is affected by the four design factors and their interactions. As shown in that table, even though the factor $\boldsymbol{\beta}$ (i.e., X_model) is statistically significant, the variance it can explain is not substantial and excluding it from discussion will not alter the basic relationship. As for the experiment results of $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$, the Figure 3 gives a panorama. The curves of $\hat{\beta}_1$ SD and $\hat{\beta}_2$ SD are nearly the same. This is because $\hat{\beta}_1$ and $\hat{\beta}_2$ are interchangeable in our simulation settings. The collinearity level has relatively limited effect on $\hat{\beta}_3$ SD, since the correlation between the third predictor and the other predictors is weak. For the independent predictor, the forth predictor, the influence of collinearity level is negligible.

2.2 Collinearity Structure: Sign and Magnitude

Unlike previous section where covariance between predictors is constrained to be positive numbers, we conduct a thorough analysis in this section with covariance varying from negative to positive, and discuss the experiment results with regard to bivariate collinearity cases and multivariate collinearity cases. We leverage the same data generating framework described in Section 2.1, but only treat covariance matrix $\boldsymbol{C}$ and the values of true regression coefficients $\boldsymbol{\beta}$ as variables. For bivariate collinearity cases, we set the $\boldsymbol{C}$ to be a 2×2 symmetric matrix with $\rho_{12} \in [-0.8, 0.8]$. (Please note that since the predictor variances are always set to 1, the correlation ρ_{12} is equal to the corresponding covariance.) The true model coefficients (β_1, β_2) is a vector of length 1, which is at an angle $\theta \in [0, \frac{\pi}{2}]$ to x-axis (i.e., $(\beta_1, \beta_2) = (\cos\theta, \sin\theta)$). For multivariate collinearity cases, $\boldsymbol{C}$ (a 3×3 matrix) contains $\rho_{13}, \rho_{23} \in [-0.45, 0.45]$ and $\rho_{12} = 0$. $(\beta_1, \beta_2, \beta_3) = (\sin\psi * \cos\theta, \sin\psi * \sin\theta, \cos\psi)$, where $\psi, \theta \in [0, \frac{\pi}{2}]$. We use Monte Carlo simulation to explore these factors' effects on $\hat{\beta}$ SD.

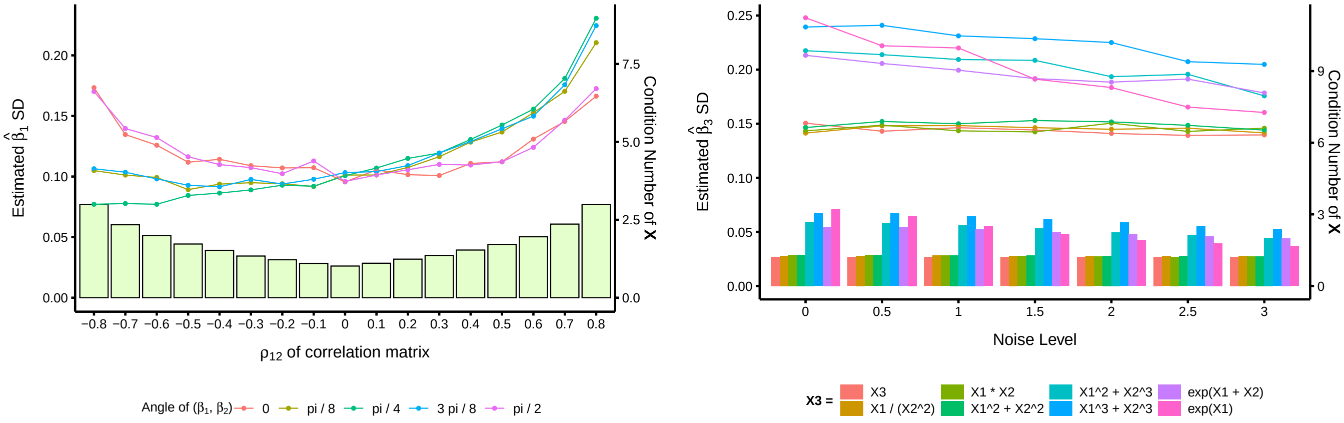

Figure 2: The Experimental Results of Section 2.2 and Section 2.3: The LHS figure illustrates the estimated $\hat{\beta}_1$ SD under different ρ_{12} and $\boldsymbol{\beta}$ settings using line plot. The condition numbers of design matrix $\boldsymbol{X}$ is shown in bar plot. The RHS figure depicts the changes of estimated $\hat{\beta}_3$ SD (line plot) and condition numbers (bar plot) over different noise levels and X_3 settings. Noise level controls the variance of ϵ in $X_3 = f(X_1, X_2) + \epsilon$. High noise level will weaken the correlation between X_3 and X_1, X_2. $X_3 = X_3$ indicates that X_3 is independent.

For bivariate collinearity cases, the experimental results are shown in Figure 2's LHS, and we can get the following three findings: (1) The effects of negative and positive correlations are asymmetric. We find that, for two ρ_{12}s with the same absolute value (e.g. -0.6 and 0.6), their corresponding $\hat{\beta}_1$ SDs are not necessarily identical. For instance, when $\theta = \frac{\pi}{4}$, the $\hat{\beta}_1$ SD at $\rho_{12} = -0.8$ is significantly different from the value at $\rho_{12} = 0.8$. (2) It's possible to get a more accurate estimate of coefficient by increasing the collinearity level. As shown in the figure, when $\theta = \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}$, increasing negative collinearity can decrease the variance of coefficient estimates. (3) The condition number can't always accurately identify the collinearity problems. The condition numbers of different ρ_{12} settings form a U-shape bar plot, which implies that the coefficient variance will inflate as the correlation becomes negative, but, in fact, the variance can contract in some cases. (The experimental results of this part match the prediction in [3]. For a more detailed analysis, we refer readers to that paper.)

For multivariate collinearity cases, two interesting findings are found: (1) When $\psi = 0$ or $\psi = \frac{\pi}{2}$, the estimated SD of $\hat{\beta}_1$ is only related to ρ_{13}. Similar finding can be observed for $\hat{\beta}_2$ SD. This can be verified by viewing the first and forth rows of the Figure 4 and Figure 5. (2) In some special cases (e.g., $\theta = \frac{\pi}{4}, \psi = \frac{\pi}{4}$), $\hat{\beta}$ SD gets its minimum/maximum when $\rho_{13}, \rho_{23} = -0.45$ / $\rho_{13}, \rho_{23} = 0.45$. This is similar to the finding of bivariate collinearity cases.

2.3 Collinearity Structure: Formed by Functions

We consider the situation where the third predictor X_3 is a function of the first two mutually independent predictors X_1 and X_2, and report the findings of Monte Carlo simulation. We generate the samples for X_1 and X_2 from a multivariate normal distribution whose $\rho_{12} = 0$, and then calculate the samples of X_3 using the function $X_3 = f_i(X_1, X_2) + \epsilon$, where ϵ is the random normal noise with standard deviation set to be "$Noise\ Level$". When noise level increases, the correlation between X_3 and X_1, X_2 decreases. As for functions $f_i(\cdot)$, we test eight representative non-linear functions: (1) $X_3 = X_3$ returns an independent X_3; (2) $X_3 = X_1/X_2^2$; (3) $X_3 = X_1 * X_2$; (4) $X_3 = X_1^2 + X_2^2$; (5) $X_1^2 + X_2^3$; (6) $X_3 = X_1^3 + X_2^3$; (7) $X_3 = \exp(X_1 + X_2)$; (8) $X_3 = \exp(X_1)$.

The Monte Carlo simulation results are shown in Figure 2's RHS, and we summarize our findings as three-fold: (1) The quadratic and ratio functions act just like independent variables, while cubic and exponential functions increase collinearity. As shown in the figure, functions (1)-(4) don't inflate $\hat{\beta}_3$ SD or condition number. But for functions (5)-(8), significantly high $\hat{\beta}_3$ SD can be observed. (2) The effects of $Noise\ Level$ on different non-linear functions are inconsistent. For exponential function (8), $\hat{\beta}_3$ SD decreases much more rapidly as the $Noise\ Level$ increases, compared to functions (5)-(7). (3) The condition number is an effective diagnostic for detecting collinearity in this case. A high condition number corresponds to high $\hat{\beta}_3$ SD, regardless of the non-linear functions we use to generate X_3.

2.4 Collinearity Structure: Temporal Patterns

We contrive a Monte Carlo simulation to examine whether the changes in collinearity structure over time will degrade the prediction accuracy of multiple linear regression model. As discussed in [1], when the collinearity between predictors varies with the time, a multiple linear regression model pre-trained on old dataset may produce inaccurate prediction on new test dataset. We follow their practice to design our experiment. For train set, we craft a set of collinear predictors using the equation: $X_{i+1} = X_i + \epsilon$, where X_1 is drawn from a uniform distribution between 0 and 1, and noise ϵ follows $N(0, decay^2)$. The "*decay*" denotes how fast the collinearity decreases for the predictors in a set, and high "*decay*" results in low collinearity. We select the decay levels ranging from 0.002 to 1. As for testing set, we generate five types of testing sets for each training set to mimic the changes in collinearity structures. The first type is "same". It generates data with the same collinearity structure as training data's. The second and third types are "low decay" and "high decay", where the decay is half and twice that of training data. The fourth type "non-linear" only applies "*decay*" to high values. The final one "all independent" produces mutually independent predictors.

In Figure 7, we report the testing Root Mean Squared Error (RMSE) of the models trained and tested on different datasets. We get two non-trivial findings: (1) On "same", "low decay" and "high decay" testing sets, obvious prediction inaccuracy can't be found. For these testing, the models fitted on training set with any level of collinearity give nearly identical testing RMSE. (2) Models trained on datasets with high collinearity perform bad on "non-linear" and "all independent" testing sets. For these two datasets, when the decay level goes towards 0, the testing RMSE increases exponentially.

3 Conclusion

In this paper, we use Monte Carlo simulation to study the effects of four collinearity-related factors on multiple linear regression, and get some interesting findings. For example, under some conditions, the increase of collinearity level will decrease the $\hat{\beta}$ SD. The change of collinearity structure may compromise model prediction. We try our best to provide a comprehensive survey and we hope this paper can be a useful guidance for researchers who want to deal with collinearity.

References

[1] Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography 36*, 1 (2013), 27–46.

[2] Mason, C. H., and Perreault, W. D. Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research 28* (1991), 268–280.

[3] Mela, C. F., and Kopalle, P. K. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics 34*, 6 (2002), 667–677.

Appendices

A Simulation Configuration of Section 2.1

Collinearity Levels	**1**				**2**				**3**				**4**			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
X_1	1.0				1.0				1.0				1.0			
X_2	0.5	1.0			0.65	1.0			0.8	1.0			0.95	1.0		
X_3	0.2	0.2	1.0		0.4	0.4	1.0		0.6	0.6	1.0		0.8	0.8	1.0	
X_4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0

Table 1: Four Collinearity Levels: This table includes four covariance matrices which correspond to four collinearity levels.

$\boldsymbol{\beta}$	β_1	β_2	β_3	β_4	Intercept
Model 1	0.5	0.265	0.0	0.25	2
Model 2	0.4	0.4	0.0	0.25	2

Table 2: The True Coefficients $\boldsymbol{\beta}$

Sample Size N: 30, 100, 150, 200, 250, 300.

R^2: 0.1, 0.25, 0.5, 0.75.

B Extra Experimental Results of Section 2.1

```
Analysis of Variance Table

Response: acc_coef_beta1
                                             Df Sum Sq Mean Sq    F value    Pr(>F)
X_col_level                                   3 1428.5  476.18 11761.7127 < 2.2e-16 ***
X_R2                                          3 2085.9  695.30 17173.9899 < 2.2e-16 ***
X_sample_size                                 5 1550.3  310.07  7658.6934 < 2.2e-16 ***
X_model                                       1    1.3    1.33    32.7549 1.047e-08 ***
X_col_level:X_R2                              9  480.1   53.35  1317.7058 < 2.2e-16 ***
X_col_level:X_sample_size                    15  352.5   23.50   580.4504 < 2.2e-16 ***
X_R2:X_sample_size                           15  532.6   35.50   876.9632 < 2.2e-16 ***
X_col_level:X_model                           3    0.5    0.16     4.0668  0.006728 **
X_R2:X_model                                  3    0.2    0.07     1.7742  0.149651
X_sample_size:X_model                         5    0.6    0.13     3.2060  0.006761 **
X_col_level:X_R2:X_sample_size               45  119.6    2.66    65.6320 < 2.2e-16 ***
X_col_level:X_R2:X_model                      9    0.2    0.02     0.4520  0.906886
X_col_level:X_sample_size:X_model            15    0.3    0.02     0.5116  0.936197
X_R2:X_sample_size:X_model                   15    0.2    0.01     0.3318  0.992318
X_col_level:X_R2:X_sample_size:X_model       45    0.8    0.02     0.4420  0.999575
Residuals                                191808 7765.5    0.04
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

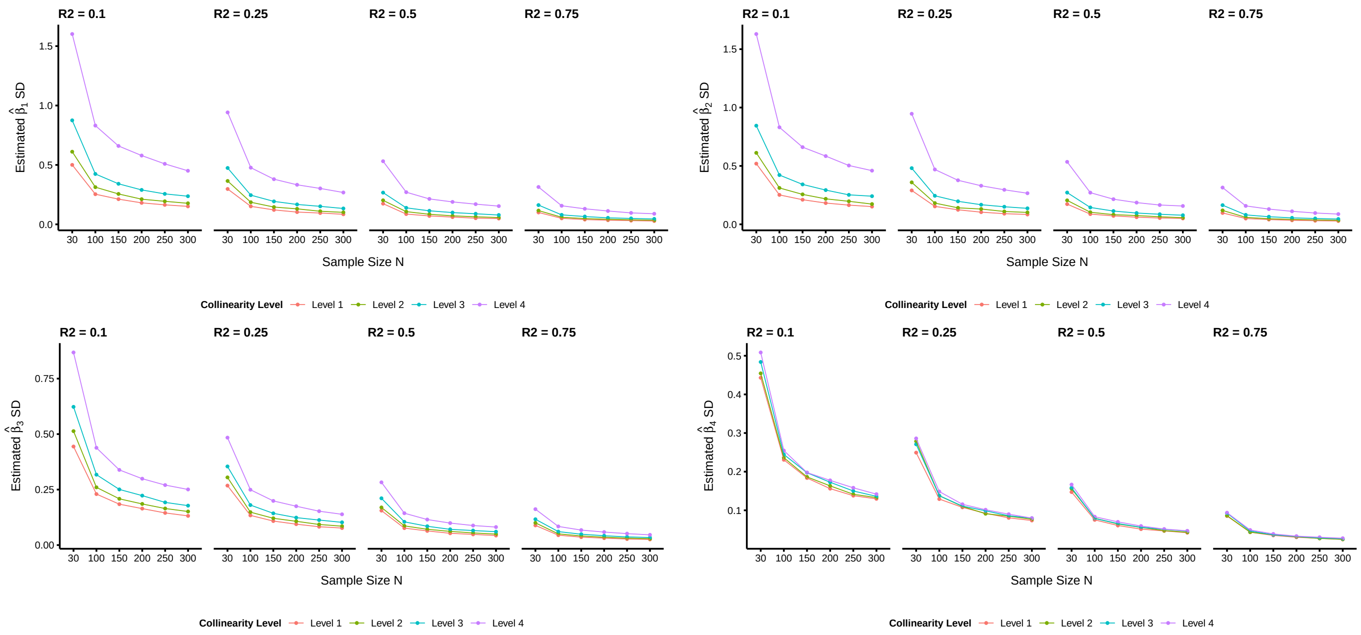

Figure 3: Joint Effect of Collinearity Structure, N and R^2 on Multiple Linear Regression: The higher the $\hat{\beta}$ SD value, the less accurate the coefficient estimate is. Level 1-4 reflect increasing levels of collinearity. This figure depicts the $\hat{\beta}$ SD of four coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$. Please note that the scales of their Y-axes are different.

C Multivariate Collinearity Cases of Section 2.2

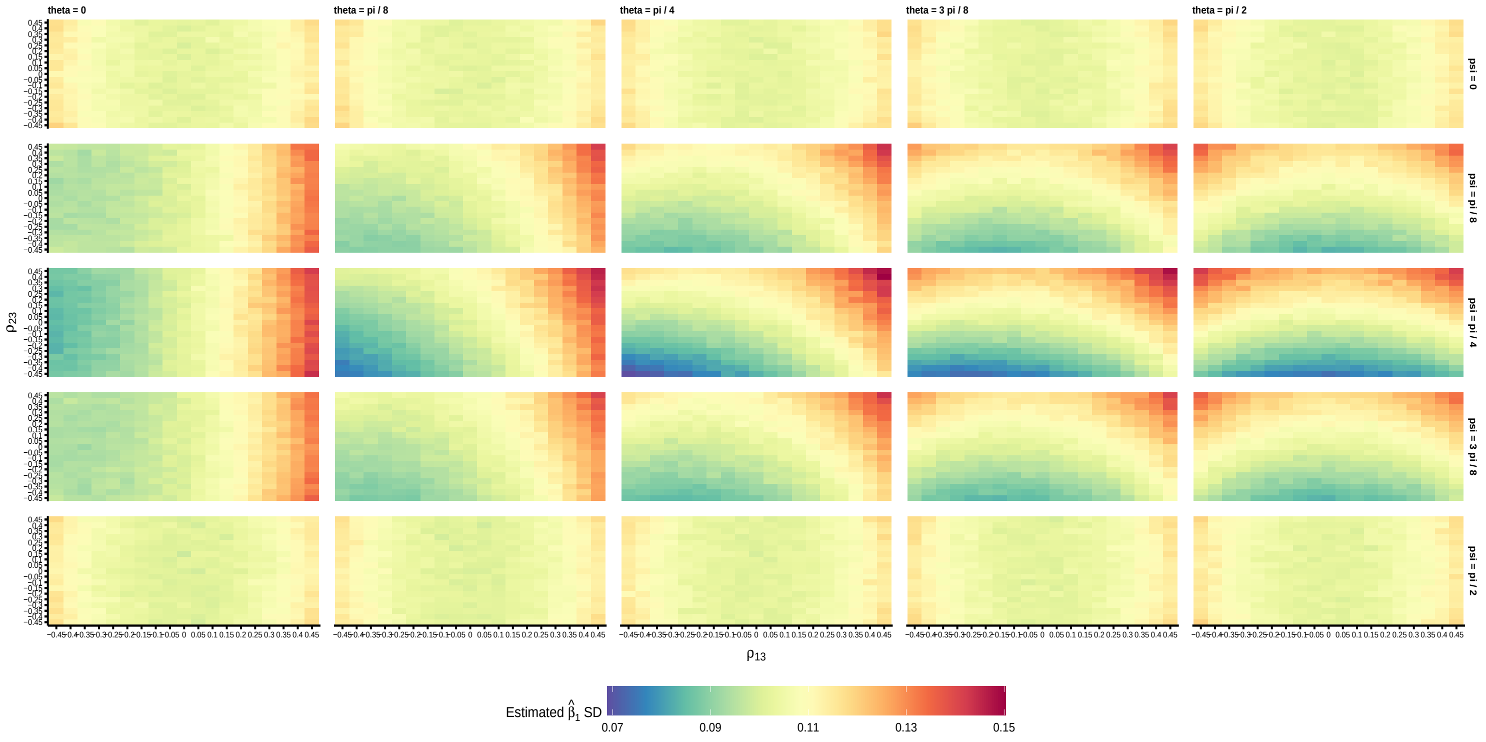

Figure 4: The Estimated $\hat{\beta}_1$ SD across different ρ_{13}, ρ_{23}, ψ and θ settings: Red represents high SD, and blue denotes low SD.

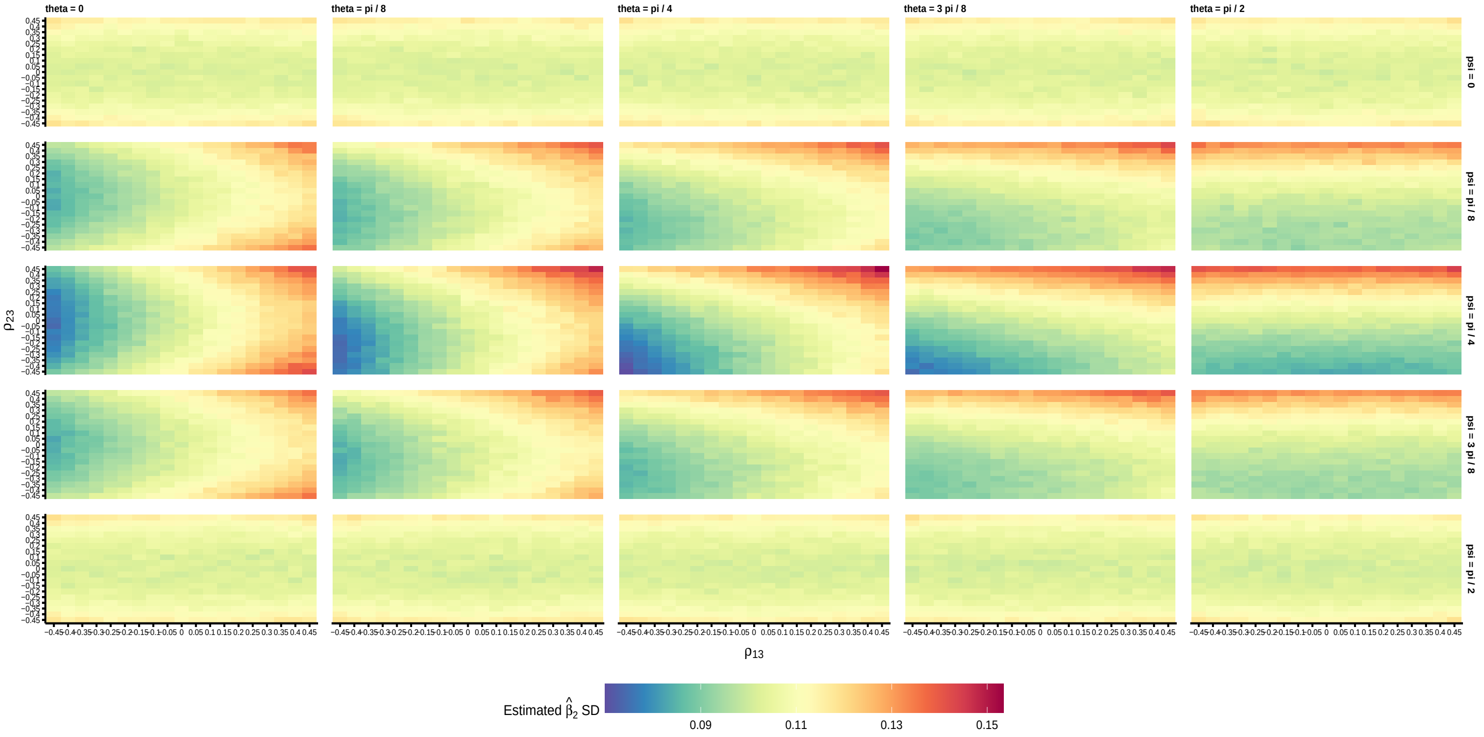

Figure 5: The Estimated $\hat{\beta}_2$ SD across different ρ_{13}, ρ_{23}, ψ and θ settings: Red represents high SD, and blue denotes low SD.

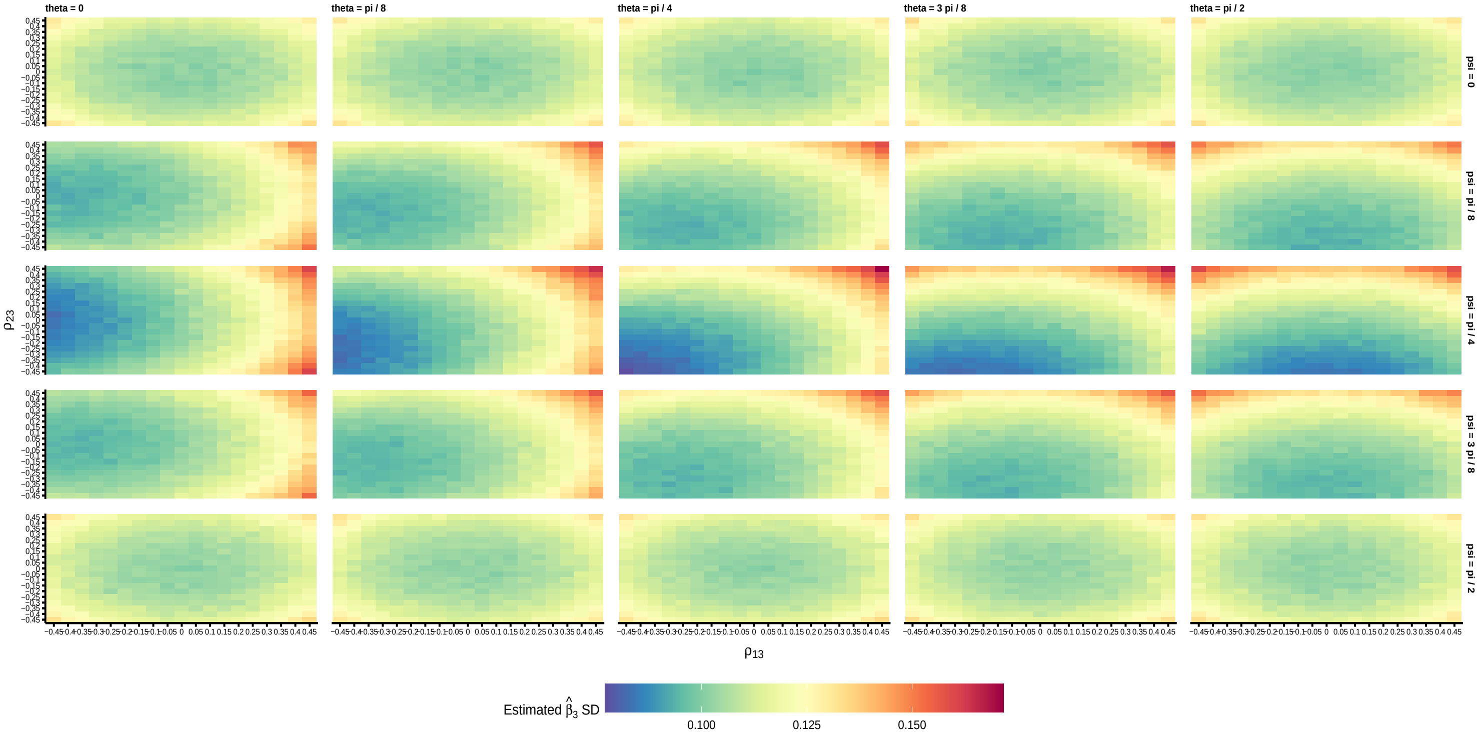

Figure 6: The Estimated $\hat{\beta}_3$ SD across different ρ_{13}, ρ_{23}, ψ and θ settings: Red represents high SD, and blue denotes low SD.

D Experimental Results of Section 2.4

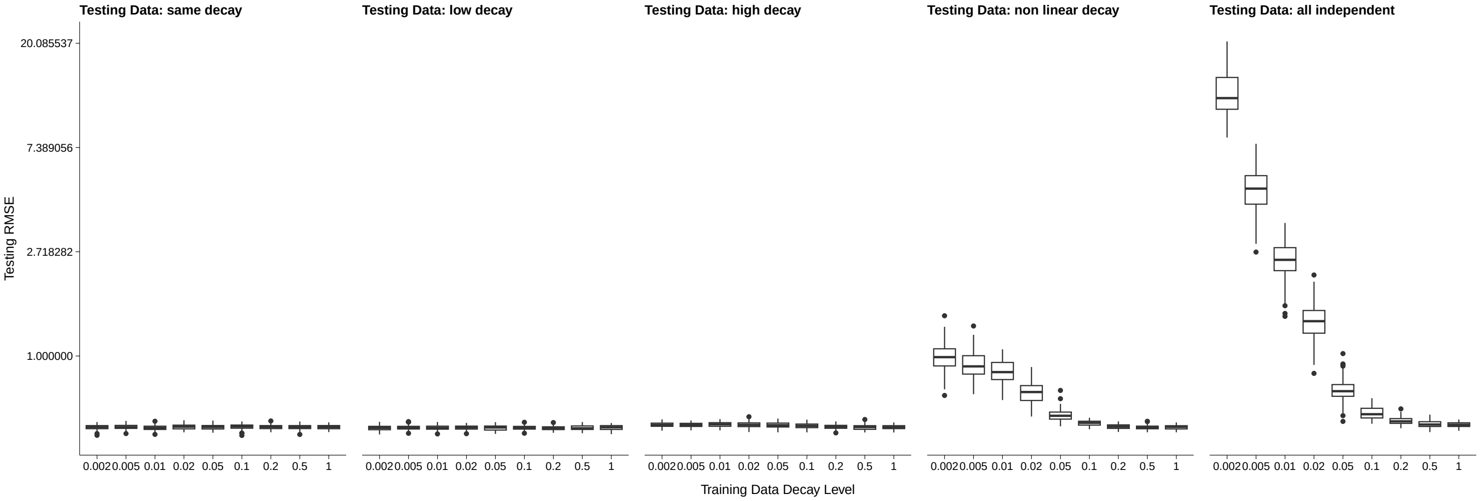

Figure 7: The Testing RMSE of models trained and tested on different datasets: A low decay level corresponds to high collinearity. Testing RMSE is $[\sum(y-\hat{y})^2/n]^{\frac{1}{2}}$, where y is the true value, $\hat{y}$ is the predicted value and n is the sample size.