(G)RDPG with Covariates

Cong Mu

1 Notes

Consider (G)RDPG with covariates as

$$P_{ij} = X_i^\top X_j + \beta 1_{\{Z_i = Z_j\}}. \tag{1}$$

Given an adjacency matrix A with observed covariates Z, we use the following procedure to estimate β.

1. Estimate $\hat{X}$ using Adjacency Spectral Embedding (ASE).
2. Cluster $\hat{X}$ using Gaussian Mixture Model (GMM) and compute the means of clusters $\hat{\mu}$.
3. Construct matrix $\mathbf{I}_{d_1,d_2} = \text{diag}(1, \cdots, 1, -1, \cdots, -1)$ with d_1 ones followed by d_2 minus ones on its diagonal, $d_1 \geq 1$ and $d_2 \geq 0$ are two integers satisfying $d_1 + d_2 = \hat{d}$ where $\hat{d}$ is the embeded dimension. (c.f. [1])
4. Compute matrix $B_{\hat{\mu}} = \hat{\mu}\mathbf{I}_{d_1,d_2}\hat{\mu}^\top$ and cluster the diagonal of $B_{\hat{\mu}}$.
5. Estimate β by subtracting all corresponding (based on clusters in last step) terms in $B_{\hat{\mu}}$ and taking the mean.
6. Post analysis such as removing the effects of covariates.

We do some simulations to test whether this procedure could work well if we change

- Number of Blocks
- Dimension of Latent Position
- Size of Each Block and Each Gender (Binary Covariate)

2 Simulation

2.1 Number of Blocks

Here we fix dimension of latent position $d = 1$, the size of each block and each gender (binary covariate) to be balanced, and consider number of blocks $K = 2, 4, 10$.

2.1.1 K = 2, n = 2000

We consider latent position to be $[0.1, 0.3]$, i.e. $p = 0.1$, $q = 0.3$ and $\beta = 0.3$. With our procedure, we estimate β as $\hat{\beta} = 0.2995$ (runtime: 47s). Figure 1 shows the screeplots and latent postions with and without covariates.

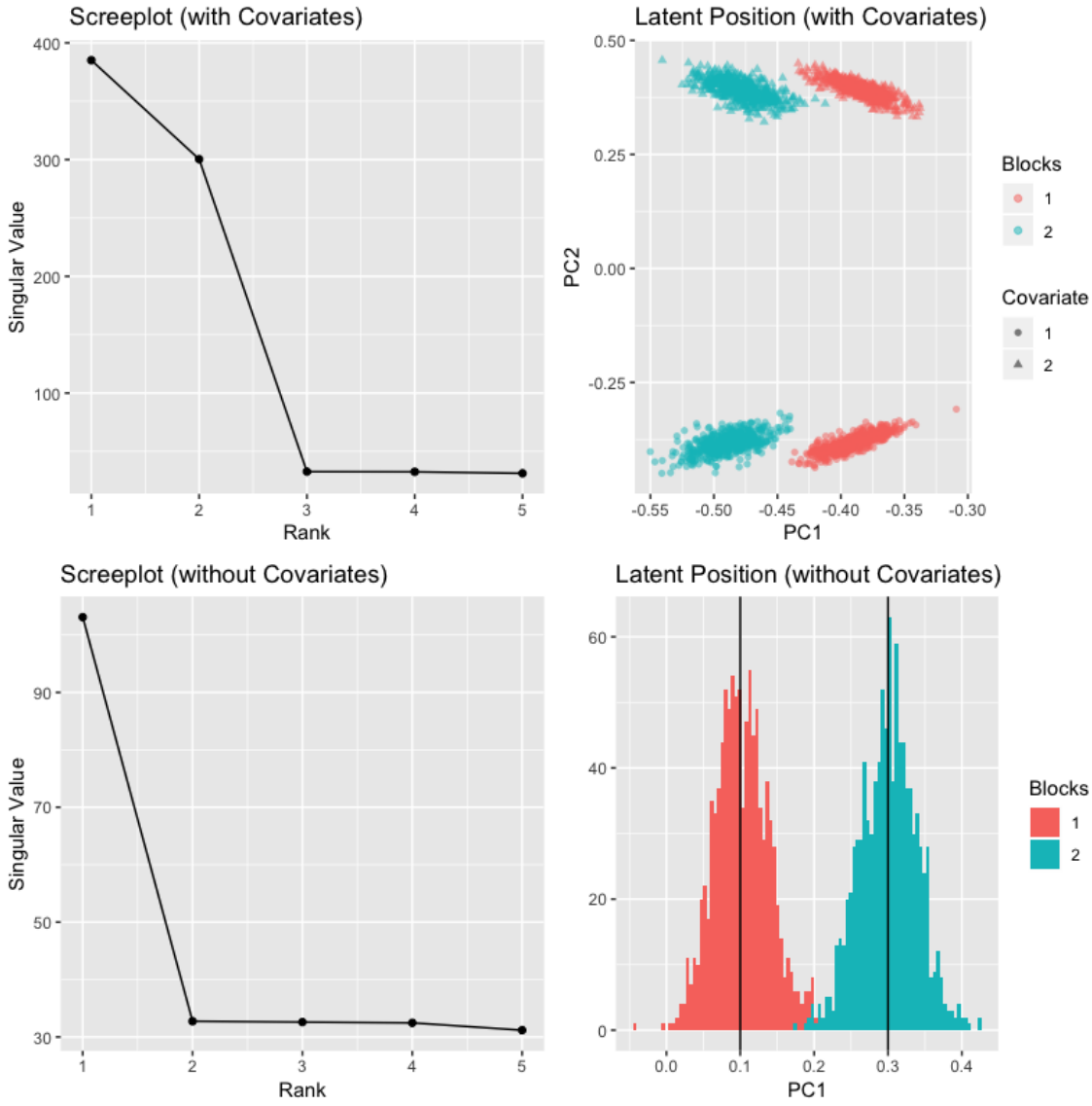

Figure 1: $K = 2$, $d = 1$, $n = 2000$, $\beta = 0.3$, Balanced

2.1.2 K = 4, n = 4000

We consider latent position to be $[0.2, 0.4, 0.7, 0.8]$, and $\beta = 0.2$. With our procedure, we estimate β as $\hat{\beta} = 0.1999$ (runtime: 3m28s). Figure 2 shows the screeplots and latent postions with and without covariates.

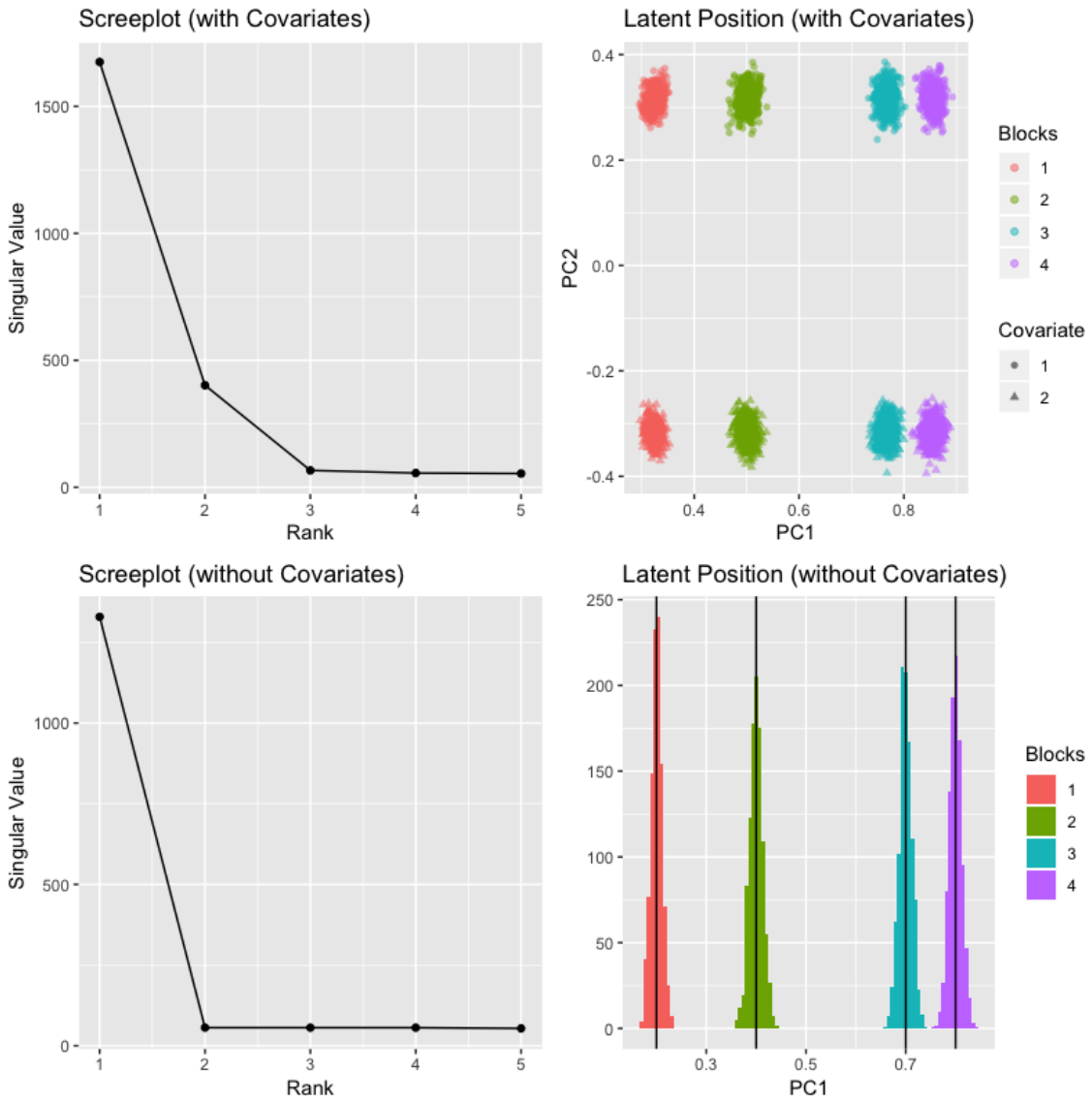

Figure 2: $K = 4,\ d = 1,\ n = 4000,\ \beta = 0.2$, Balanced

2.1.3 K = 10, n = 10000

We consider latent position to be $[0.1, 0.25, 0.3, 0.45, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9]$, and $\beta = 0.15$. With our procedure, we estimate β as $\hat{\beta} = 0.1383$ (runtime: 35m9s). Figure 3 shows the screeplots and latent postions with and without covariates.

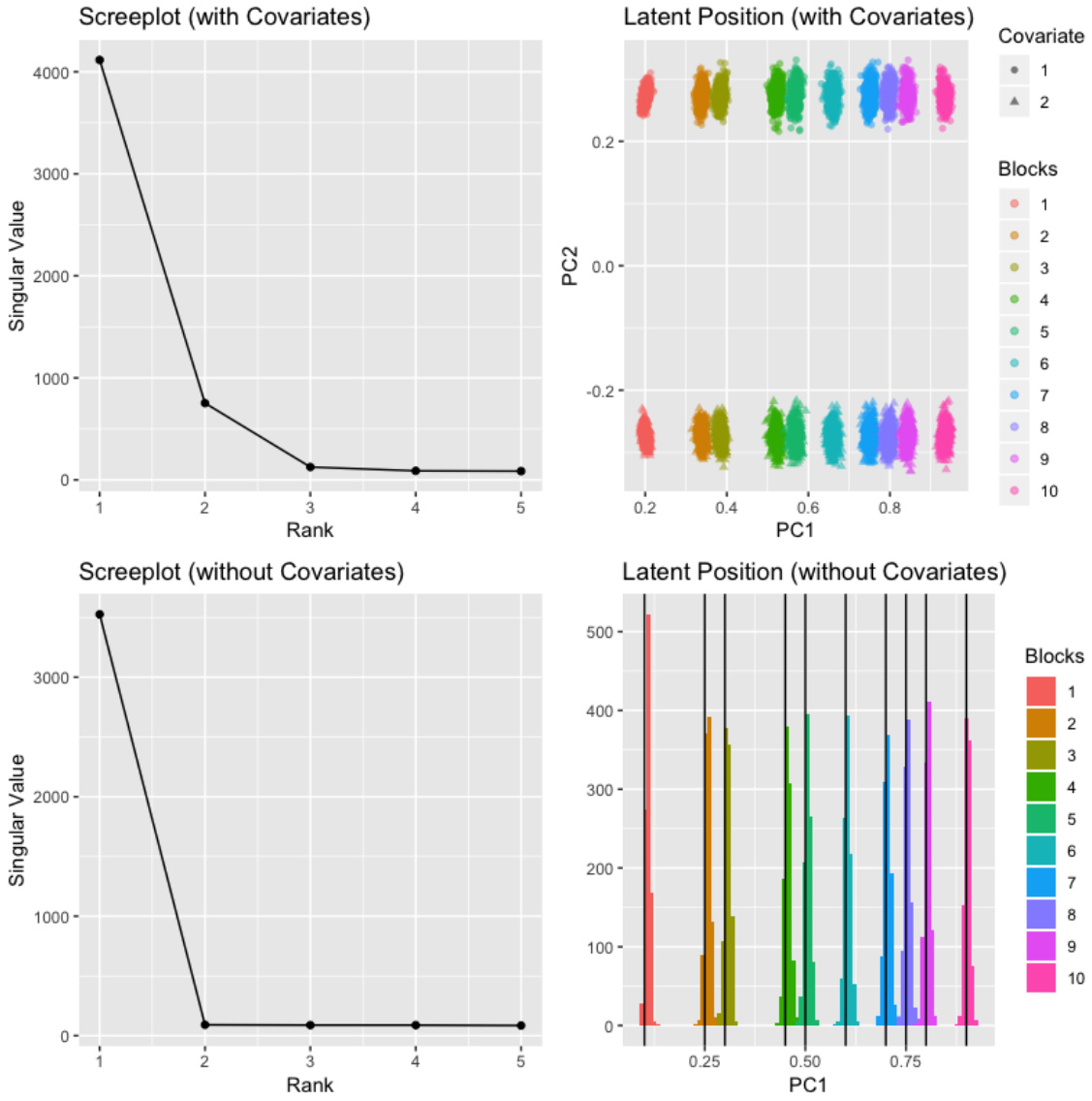

Figure 3: $K = 10$, $d = 1$, $n = 10000$, $\beta = 0.15$, Balanced

2.2 Dimension of Latent Posistion

Here we fix number of blocks $K = 2$, the size of each block and each gender (binary covariate) to be balanced, and consider dimension of latent position $d = 1, 2$.

2.2.1 d = 1, n = 2000

We consider latent position to be $[0.5, 0.9]$, i.e. $p = 0.5$, $q = 0.9$ and $\beta = -0.2$. With our procedure, we estimate β as $\hat{\beta} = -0.2000$ (runtime: 48s). Figure 4 shows the screeplots and latent postions with and without covariates.

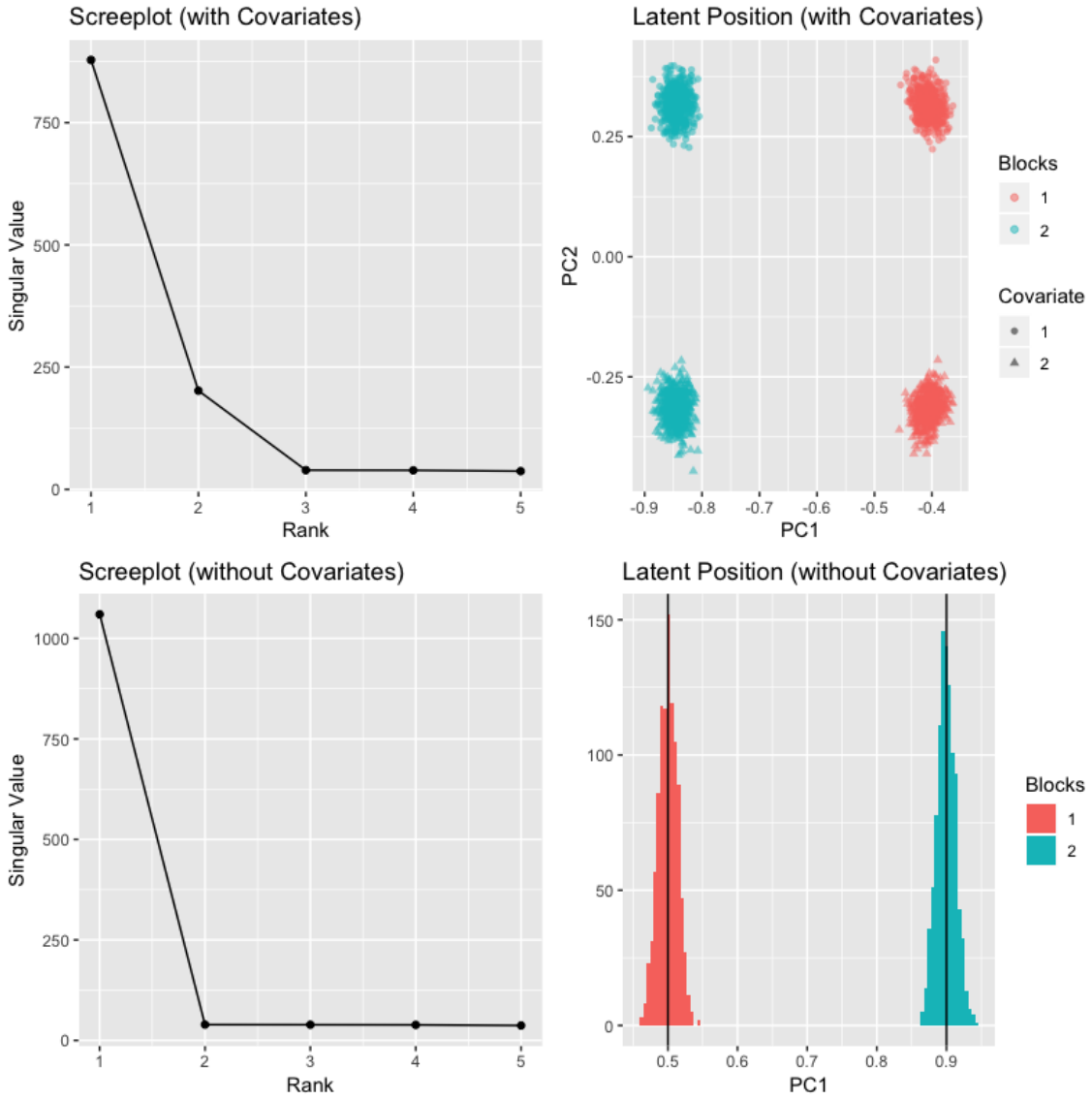

Figure 4: $K = 2$, $d = 1$, $n = 2000$, $\beta = -0.2$, Balanced

2.2.2 d = 2, n = 4000

We consider latent position to be $x_1 = [0.63, -0.14]$, $x_2 = [0.69, 0.13]$, and $\beta = -0.3$. With our procedure, we estimate β as $\hat{\beta} = -0.2999$ (runtime: 3m30s). Figure 5 shows the screeplots and latent postions with and without covariates.

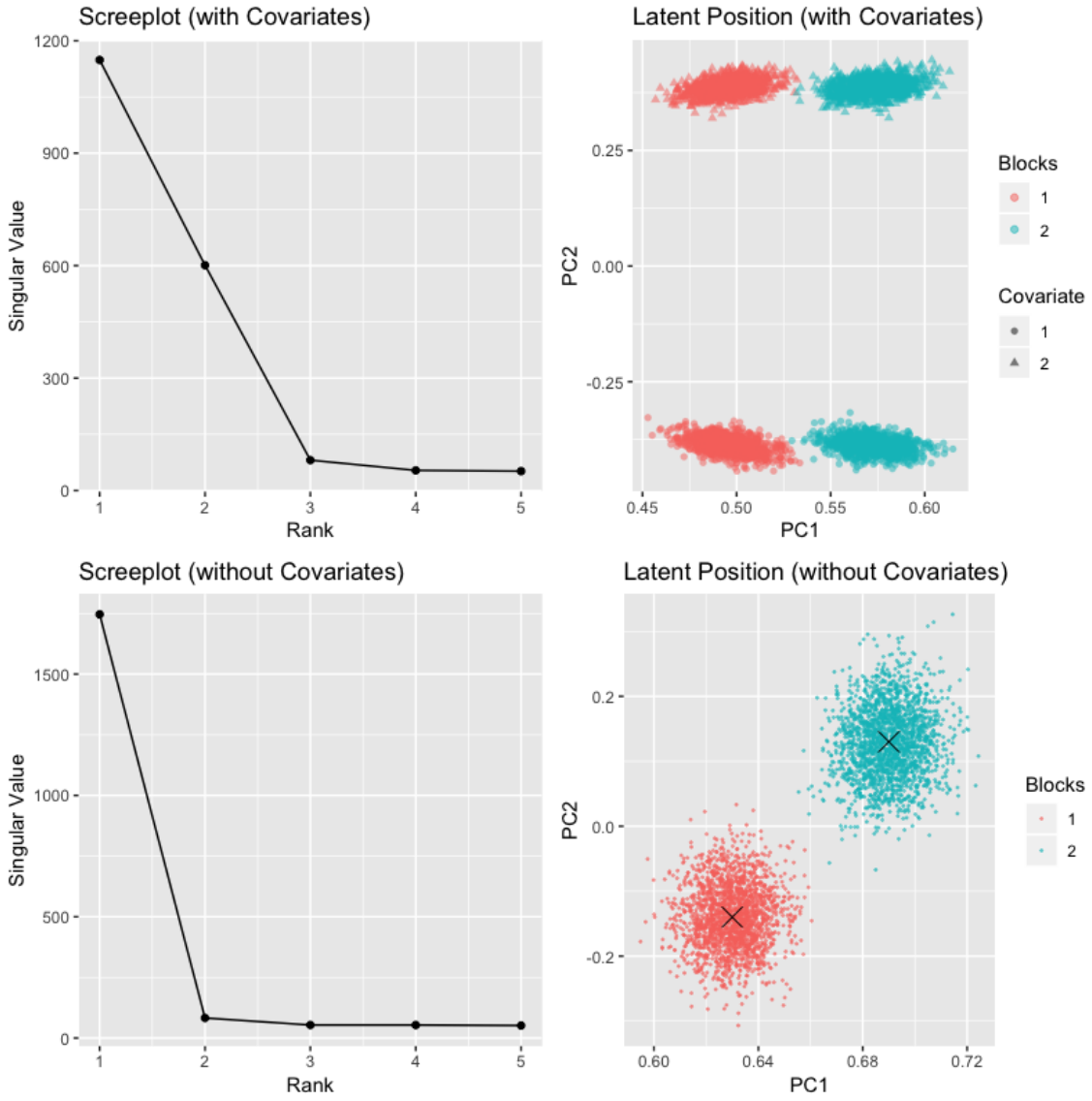

Figure 5: $K = 2$, $d = 2$, $n = 4000$, $\beta = -0.3$, Balanced

2.3 Size of Each Block and Each Gender (Binary Covariate)

Here we fix number of blocks $K = 2$, dimension of latent position $d = 1$, and consider the size of each block and the size of each gender to be unbalanced.

2.3.1 Size of Block = (0.3, 0.7), Size of Gender = (0.4, 0.6), n = 2000

We consider latent position to be $[0.7, 0.9]$, i.e. $p = 0.7$, $q = 0.9$ and $\beta = -0.35$. With our procedure, we estimate β as $\hat{\beta} = -0.3492$ (runtime: 46s). Figure 6 shows the screeplots and latent postions with and without covariates.

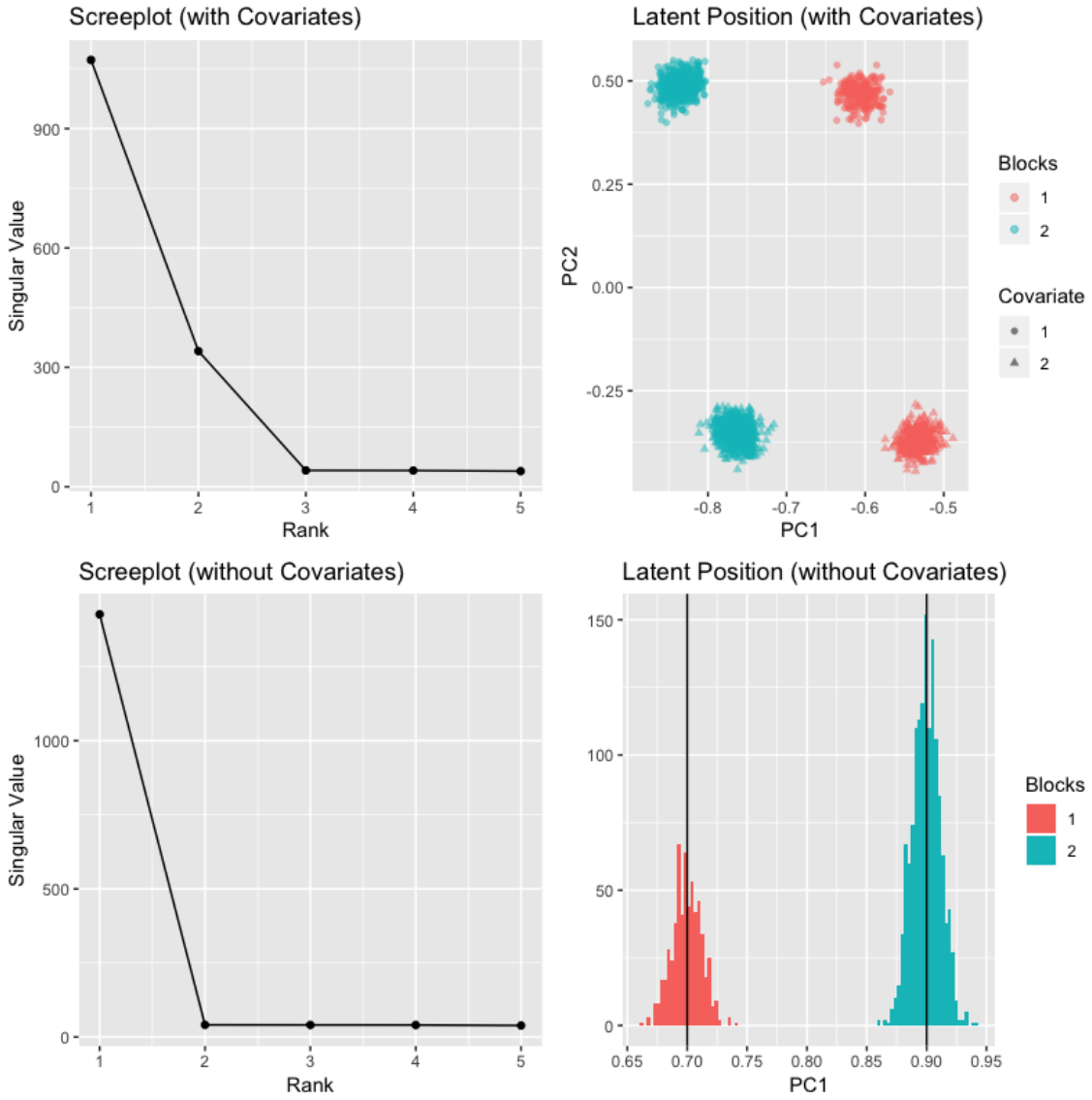

Figure 6: $K = 2$, $d = 1$, $n = 2000$, $\beta = -0.35$, Size of Block = (0.3, 0.7), Size of Gender = (0.4, 0.6)

2.3.2 Size of Block = (0.2, 0.8), Size of Gender = (0.3, 0.7), n = 4000

We consider latent position to be $[0.2, 0.6]$, i.e. $p = 0.2$, $q = 0.6$ and $\beta = 0.4$. With our procedure, we estimate β as $\hat{\beta} = 0.3992$ (runtime: 3m33s). Figure 7 shows the screeplots and latent postions with and without covariates.

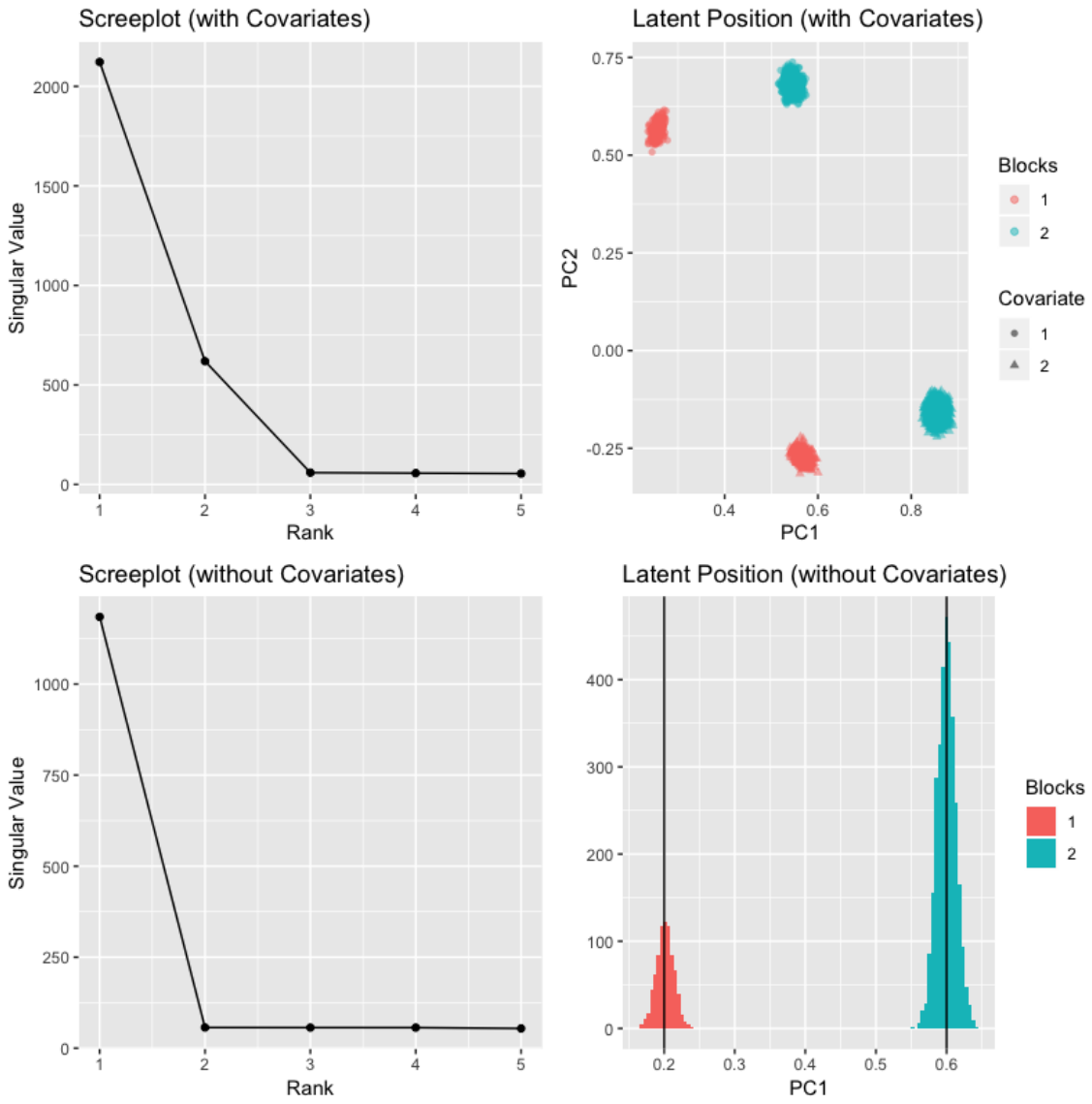

Figure 7: $K = 2$, $d = 1$, $n = 4000$, $\beta = 0.4$, Size of Block = (0.2, 0.8), Size of Gender = (0.3, 0.7)

2.4 Summary

With respect to the estimation of β, we summarize our simulations as follows.

n	K	d	**Size of Blocks**	**Size of Gender**	β	$\hat{\beta}$	**Runtime**
2000	2	1	Balanced	Balanced	0.3	0.2995	47s
4000	4	1	Balanced	Balanced	0.2	0.1999	3m28s
10000	10	1	Balanced	Balanced	0.15	0.1383	35m9s
2000	2	1	Balanced	Balanced	-0.2	-0.2000	48s
4000	2	2	Balanced	Balanced	-0.3	-0.2999	3m30s
2000	2	1	(0.3, 0.7)	(0.4, 0.6)	-0.35	-0.3492	46s
4000	2	1	(0.2, 0.8)	(0.3, 0.7)	0.4	0.3992	3m33s

Table 1: Summary of Simulation

In general, the procedure seems to work well under different settings (in terms of estimating β) and has a tractable computational complexity.

3 Distribution of $\hat{\beta}$

To get the uncertanity of the $\hat{\beta}$, we consider the following procedure.

1. Estimate $\hat{X}$ using Adjacency Spectral Embedding (ASE).
2. Cluster $\hat{X}$ using Gaussian Mixture Model (GMM) and compute the means of clusters $\hat{\mu}$.
3. Construct matrix $\mathbf{I}_{d_1,d_2} = \text{diag}(1, \cdots, 1, -1, \cdots, -1)$ with d_1 ones followed by d_2 minus ones on its diagonal, $d_1 \geq 1$ and $d_2 \geq 0$ are two integers satisfying $d_1 + d_2 = \hat{d}$ where $\hat{d}$ is the embeded dimension. (c.f. [1])
4. Compute matrix $B_{\hat{\mu}} = \hat{\mu}\mathbf{I}_{d_1,d_2}\hat{\mu}^\top$ and cluster the diagonal of $B_{\hat{\mu}}$.
5. Compute $\hat{P} = \hat{X}\mathbf{I}_{d_1,d_2}\hat{X}^\top$ (or explode $B_{\hat{\mu}}$).
6. Estimate the distribution of $\hat{\beta}$ by subtracting all paired (based on clusters in **Step 2** and **Step 4**) terms in $\hat{P}$.

Figure 8 and Figure 9 are two examples using our procedure.

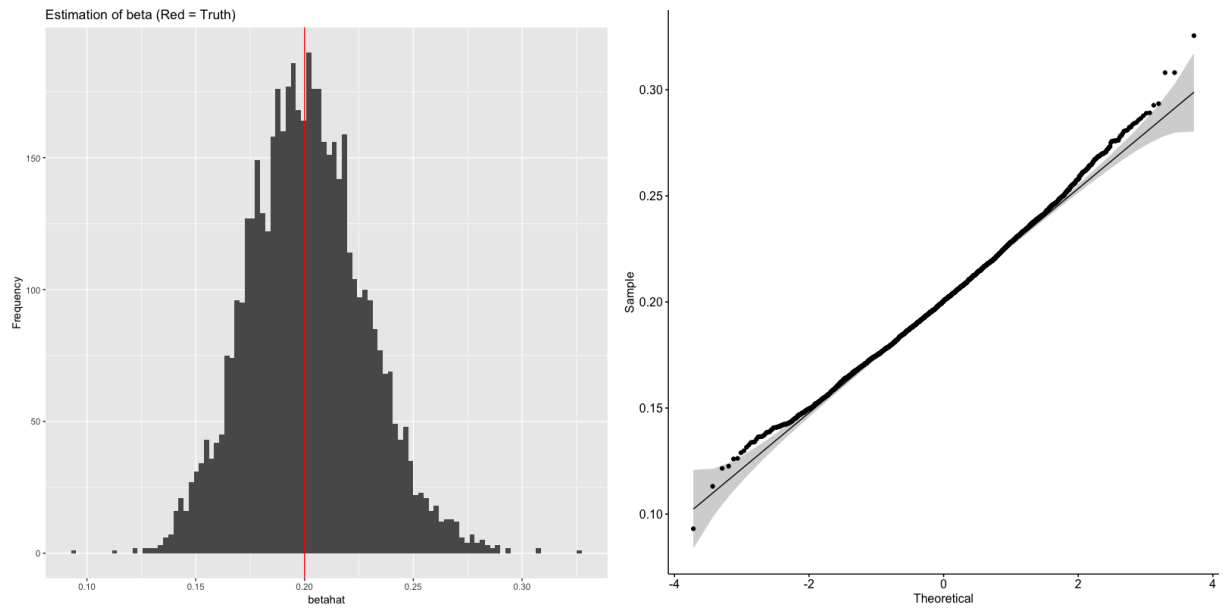

Figure 8: $K = 4,\ d = 1,\ n = 4000,\ \beta = 0.2$, Size of Block = (0.3, 0.7), Size of Gender = (0.4, 0.6), Runtime=1h23m54s

The Shapiro-Wilk Test (H_0: normal) gave `p-value < 0.01`.

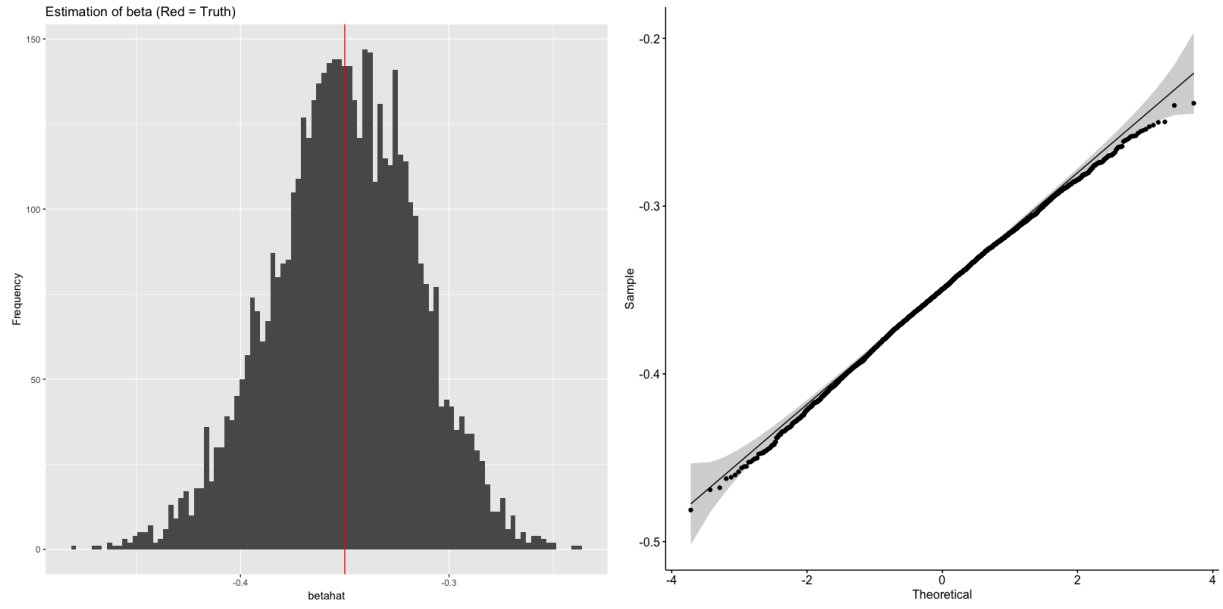

Figure 9: $K = 2,\ d = 1,\ n = 2000,\ \beta = -0.35$, Balanced, Runtime=13m33s

The Shapiro-Wilk Test (H_0: normal) gave `p-value < 0.01`.

4 Extensions

Bipartite

Here we fix dimension of latent position $d = 1$, number of blocks for worker $K_X = 4$, number of blocks for firm $K_Y = 2$, latent position for worker to be $[0.1, 0.3, 0.4, 0.6]$, latent position for firm to be $[0.7, 0.9]$, size of each block to be balanced. We do some simple simluation (try different number of workers n and number of firms m) to check whether the procedure would still work. Note that here we generate adjacency matrix **WITHOUT** condition that one worker could only work for one firm.

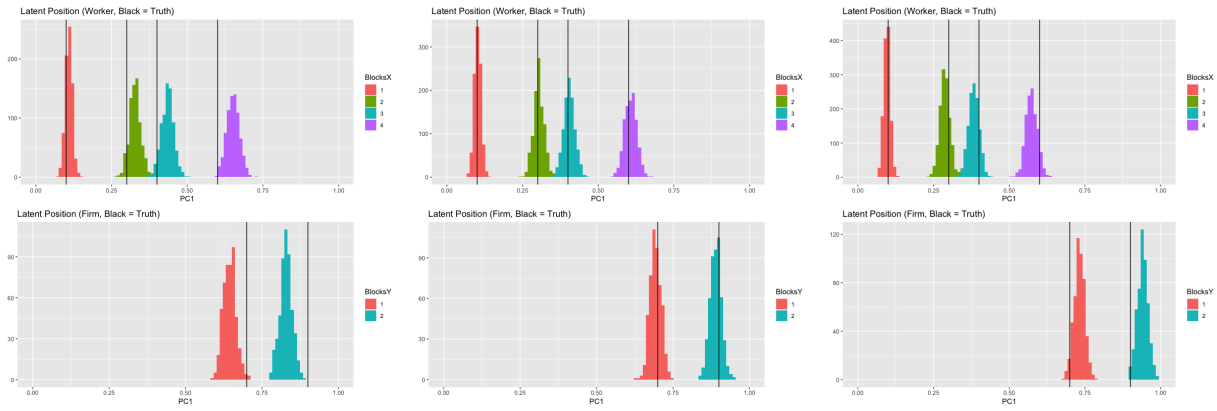

Figure 10: $n = [3000, 4000, 5000],\ m = 1000$

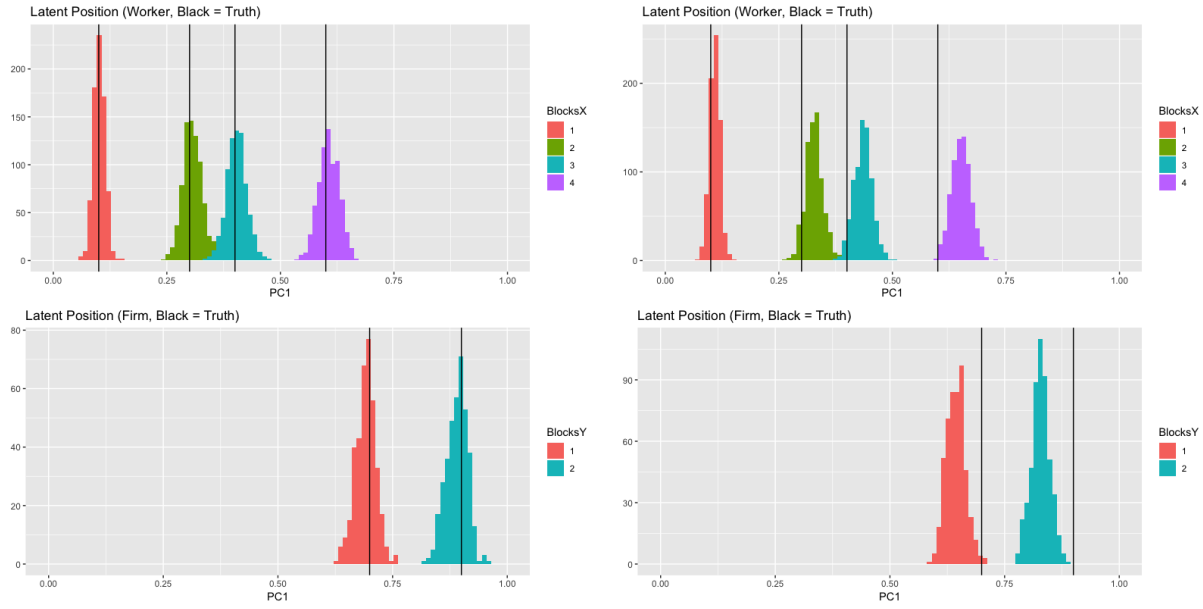

Figure 11: $n = 3000,\ m = [750, 1000]$

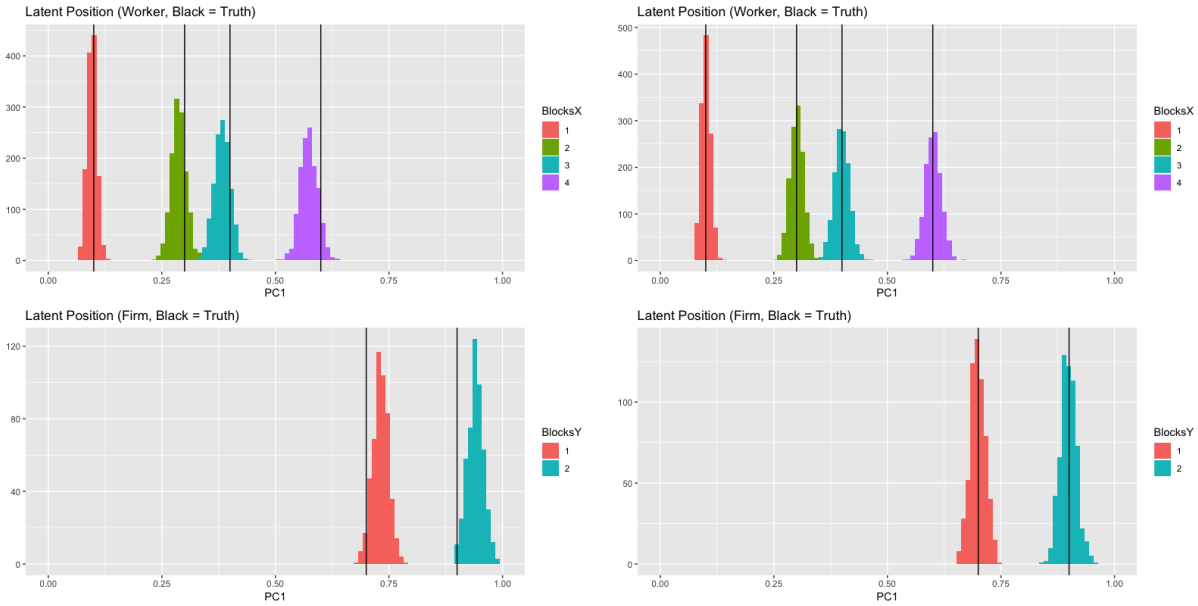

Figure 12: $n = 5000,\ m = [1000, 1200]$

Poisson

Case 1: Generate Data using 'Real' Data

Given the 51×51 states transitions data, we use the following steps to simulate data and test our procedure

1. Decompose $A_{transitions} = USV^\top$.
2. Compute $\hat{X} = U_{\hat{d}} S_{\hat{d}}^{1/2}$ and $\hat{Y} = V_{\hat{d}} S_{\hat{d}}^{1/2}$ where $\hat{d}$ is the embeded dimension.
3. Cluster $\hat{X}$ and $\hat{Y}$ using Gaussian Mixture Model (GMM) and compute the means of clusters $\hat{\mu}_X$ and $\hat{\mu}_Y$. Consequently, we could also get the number of blocks K_X and K_Y.
4. Given the number of nodes (firms) n, use $\hat{\mu}_X$ and $\hat{\mu}_Y$ as the latent position of each block to generate balanced X and Y, i.e. the size of each block for X is n/K_X and the size of each block for Y is n/K_Y.
5. Calculate $P = XY^\top$ and simulate A as $A_{ij} \sim Poisson(P_{ij})$ where $i, j = 1, 2, \cdots, n$.
6. Follow the **Step 1 - Step 3** for A instead of $A_{transitions}$.

Using data from `d01.csv`. Let $\hat{d} = 1$ and $n = 3000$, we have $K_X = 5,\ K_Y = 6,\ \hat{\mu}_X = [0.5863, 1.6801, 2.9824, 9.6007, 926.3],\ \hat{\mu}_Y = [0.5199, 0.7677, 1.6405, 3.0695, 9.8996, 926.9891]$ and

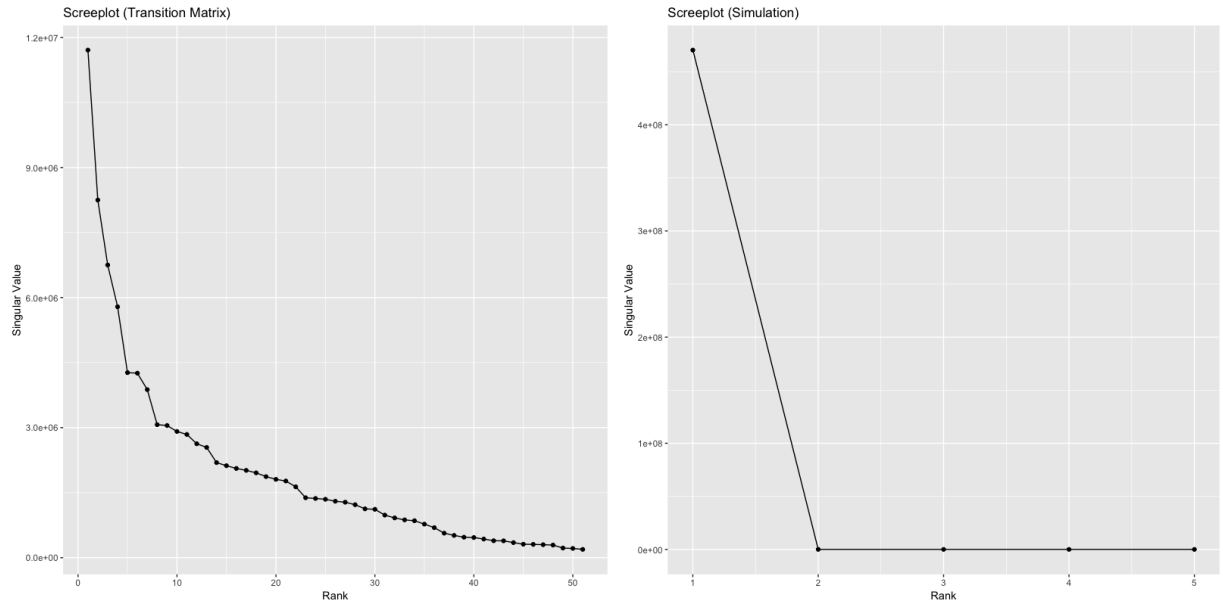

Figure 13: $n = 3000,\ \hat{d} = 1$

We summarize the true and estimated latent position as follows.

Latent Position for X						
Truth	0.5863	1.6801	2.9824	9.6007	926.3	
Estimated (Mean)	0.5604	1.6058	2.8505	9.1765	885.3571	
Latent Position for Y						
Truth	0.5199	0.7677	1.6405	3.0695	9.8996	926.9891
Estimated (Mean)	0.5439	0.8033	1.7164	3.2115	10.3576	969.8564

Table 2: Summary of true and estimated latent position

Case 2: Generate Data Directly

Instead of using $\hat{\mu}_X$ and $\hat{\mu}_Y$ as the latent position, we directly initialize the latent position and follow the **Step 5 - Step 6** in **Case 1**. Let $\hat{d} = 1,\ n = 2000,\ K_X = 4,\ K_Y = 4$, latent position for X to be $[1, 2, 10, 20]$ and latent position for Y to be $[4, 5, 13, 18]$. Follow our procedure, we could have

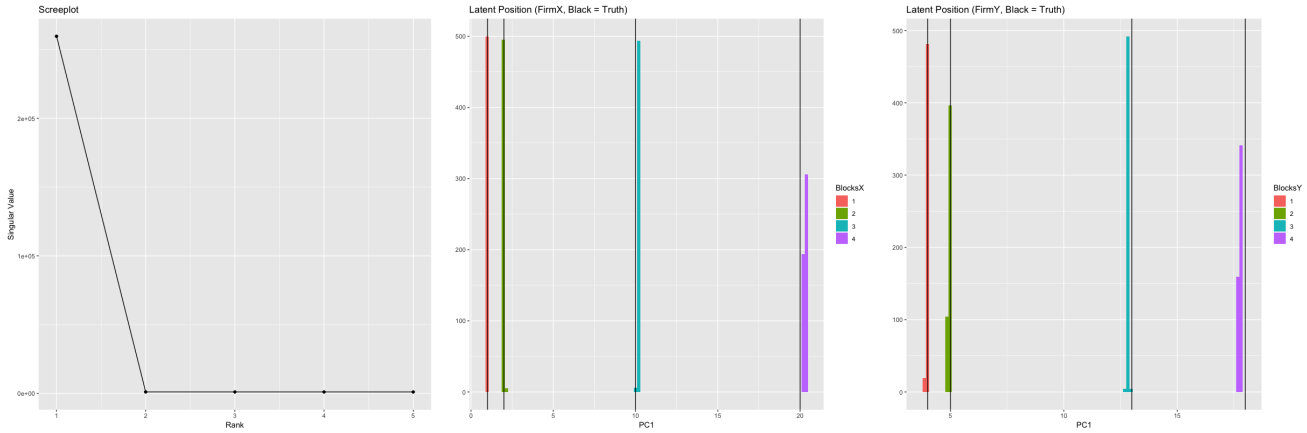

Figure 14: $n = 2000,\ \hat{d} = 1,\ K_X = K_Y = 4,\ latent_X = [1, 2, 10, 20],\ latent_Y = [4, 5, 13, 18]$, Balanced

We summarize the true and estimated latent position as follows.

Latent Position for X				
Truth	1	2	10	20
Estimated (Mean)	1.0144	2.0287	10.1395	20.2804
Latent Position for Y				
Truth	4	5	13	18
Estimated (Mean)	3.9444	4.9299	12.8200	17.7491

Table 3: Summary of true and estimated latent position

We also tried one unbalanced example. Let $\hat{d} = 1,\ n = 4000,\ K_X = 4,\ K_Y = 4$, latent position for X to be

$[1, 2, 10, 20]$, latent position for Y to be $[4, 5, 13, 18]$ and the size of each block to be $[0.25, 0.25, 0.15, 0.35]$. Follow our procedure, we could have

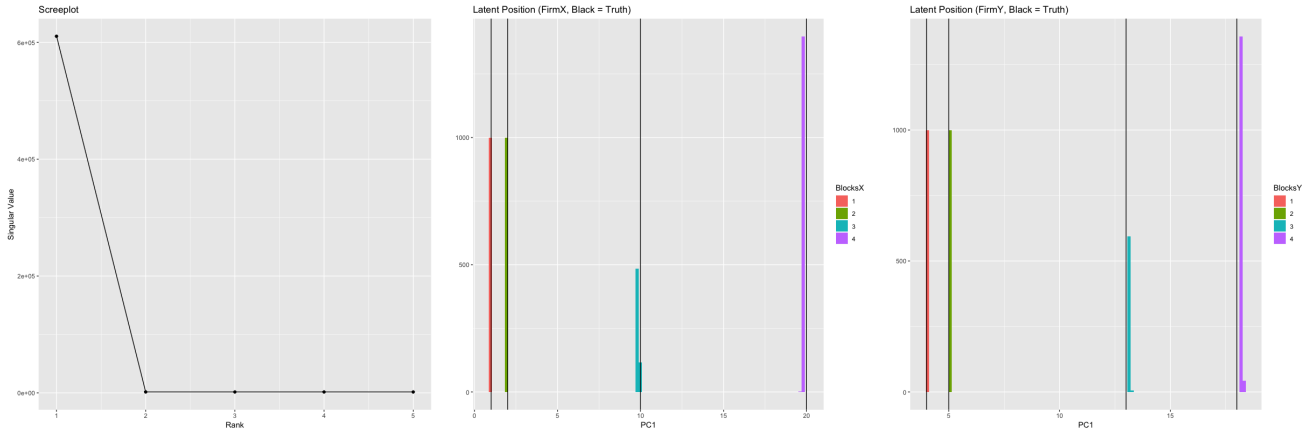

Figure 15: $n = 4000,\ \hat{d} = 1,\ K_X = K_Y = 4,\ latent_X = [1, 2, 10, 20],\ latent_Y = [4, 5, 13, 18]$, Unbalanced

We summarize the true and estimated latent position as follows.

Latent Position for X				
Truth	1	2	10	20
Estimated (Mean)	0.9881	1.9760	9.8834	19.7638
Latent Position for Y				
Truth	4	5	13	18
Estimated (Mean)	4.0473	5.0601	13.1563	18.2151

Table 4: Summary of true and estimated latent position

Multiple Covariates

We start with two binary covariates, gender (male, female) and race (white and non-white). Similar as (1), we consider (G)RDPG with covariates as

$$P_{ij} = X_i^\top X_j + \beta_1 1_{\{Z_i^{(1)}=Z_j^{(1)}\}} + \beta_2 1_{\{Z_i^{(2)}=Z_j^{(2)}\}}. \tag{2}$$

For number of block $K = 2$, dimension of latent position $d = 1$, latent position to be $[p, q]$, we have the block probability matrix as

$$B_{cov} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \tag{3}$$

where

$$B_{11} = \begin{array}{c} \\ (male, white)_1 \\ (male, non\text{-}white)_1 \\ (female, white)_1 \\ (female, non\text{-}white)_1 \end{array} \begin{array}{c} \begin{array}{cccc} (male, white)_1 & (male, non\text{-}white)_1 & (female, white)_1 & (female, non\text{-}white)_1 \end{array} \\ \begin{pmatrix} p^2+\beta_1+\beta_2 & p^2+\beta_1 & p^2+\beta_2 & p^2 \\ p^2+\beta_1 & p^2+\beta_1+\beta_2 & p^2 & p^2+\beta_2 \\ p^2+\beta_2 & p^2 & p^2+\beta_1+\beta_2 & p^2+\beta_1 \\ p^2 & p^2+\beta_2 & p^2+\beta_1 & p^2+\beta_1+\beta_2 \end{pmatrix} \end{array} \tag{4}$$

$$B_{12} = \begin{array}{c} \\ (male, white)_1 \\ (male, non\text{-}white)_1 \\ (female, white)_1 \\ (female, non\text{-}white)_1 \end{array} \begin{array}{c} \begin{array}{cccc} (male, white)_2 & (male, non\text{-}white)_2 & (female, white)_2 & (female, non\text{-}white)_2 \end{array} \\ \begin{pmatrix} pq+\beta_1+\beta_2 & pq+\beta_1 & pq+\beta_2 & pq \\ pq+\beta_1 & pq+\beta_1+\beta_2 & pq & pq+\beta_2 \\ pq+\beta_2 & pq & pq+\beta_1+\beta_2 & pq+\beta_1 \\ pq & pq+\beta_2 & pq+\beta_1 & pq+\beta_1+\beta_2 \end{pmatrix} \end{array} \tag{5}$$

$$B_{21} = \begin{array}{c} \\ (male, white)_2 \\ (male, non\text{-}white)_2 \\ (female, white)_2 \\ (female, non\text{-}white)_2 \end{array} \begin{array}{c} \begin{array}{cccc} (male, white)_1 & (male, non\text{-}white)_1 & (female, white)_1 & (female, non\text{-}white)_1 \end{array} \\ \begin{pmatrix} pq+\beta_1+\beta_2 & pq+\beta_1 & pq+\beta_2 & pq \\ pq+\beta_1 & pq+\beta_1+\beta_2 & pq & pq+\beta_2 \\ pq+\beta_2 & pq & pq+\beta_1+\beta_2 & pq+\beta_1 \\ pq & pq+\beta_2 & pq+\beta_1 & pq+\beta_1+\beta_2 \end{pmatrix} \end{array} \tag{6}$$

$$B_{22} = \begin{array}{c} \\ (male, white)_2 \\ (male, non\text{-}white)_2 \\ (female, white)_2 \\ (female, non\text{-}white)_2 \end{array} \begin{array}{c} \begin{array}{cccc} (male, white)_2 & (male, non\text{-}white)_2 & (female, white)_2 & (female, non\text{-}white)_2 \end{array} \\ \begin{pmatrix} q^2+\beta_1+\beta_2 & q^2+\beta_1 & q^2+\beta_2 & q^2 \\ q^2+\beta_1 & q^2+\beta_1+\beta_2 & q^2 & q^2+\beta_2 \\ q^2+\beta_2 & q^2 & q^2+\beta_1+\beta_2 & q^2+\beta_1 \\ q^2 & q^2+\beta_2 & q^2+\beta_1 & q^2+\beta_1+\beta_2 \end{pmatrix} \end{array} \tag{7}$$

Let $p = 0.1, q = 0.6, \beta_1 = 0.1, \beta_2 = 0.3, n = 4000$. Start with the balanced case, we have

$$B_{cov} = \begin{pmatrix} 0.41 & 0.11 & 0.31 & 0.01 & 0.46 & 0.16 & 0.36 & 0.06 \\ 0.11 & 0.41 & 0.01 & 0.31 & 0.16 & 0.46 & 0.06 & 0.36 \\ 0.31 & 0.01 & 0.41 & 0.11 & 0.36 & 0.06 & 0.46 & 0.16 \\ 0.01 & 0.31 & 0.11 & 0.41 & 0.06 & 0.36 & 0.16 & 0.46 \\ 0.46 & 0.16 & 0.36 & 0.06 & 0.76 & 0.46 & 0.66 & 0.36 \\ 0.16 & 0.46 & 0.06 & 0.36 & 0.46 & 0.76 & 0.36 & 0.66 \\ 0.36 & 0.06 & 0.46 & 0.16 & 0.66 & 0.36 & 0.76 & 0.46 \\ 0.06 & 0.36 & 0.16 & 0.46 & 0.36 & 0.66 & 0.46 & 0.76 \end{pmatrix} \tag{8}$$

With similar procedure in **Section 1**, we compute the following matrix

$$B_{\hat{\mu}} = \begin{pmatrix} 0.76 & 0.36 & 0.06 & 0.46 & 0.66 & 0.36 & 0.46 & 0.16 \\ 0.36 & 0.76 & 0.46 & 0.06 & 0.46 & 0.16 & 0.66 & 0.36 \\ 0.06 & 0.46 & 0.41 & 0.01 & 0.16 & 0.11 & 0.36 & 0.31 \\ 0.46 & 0.06 & 0.01 & 0.41 & 0.36 & 0.31 & 0.16 & 0.11 \\ 0.66 & 0.46 & 0.16 & 0.36 & 0.76 & 0.46 & 0.36 & 0.06 \\ 0.36 & 0.16 & 0.11 & 0.31 & 0.46 & 0.41 & 0.06 & 0.01 \\ 0.46 & 0.66 & 0.36 & 0.16 & 0.36 & 0.06 & 0.76 & 0.46 \\ 0.16 & 0.36 & 0.31 & 0.11 & 0.06 & 0.01 & 0.46 & 0.41 \end{pmatrix} \tag{9}$$

Then we could get $\hat{\beta}_1 = 0.1000$ and $\hat{\beta}_2 = 0.2996$.

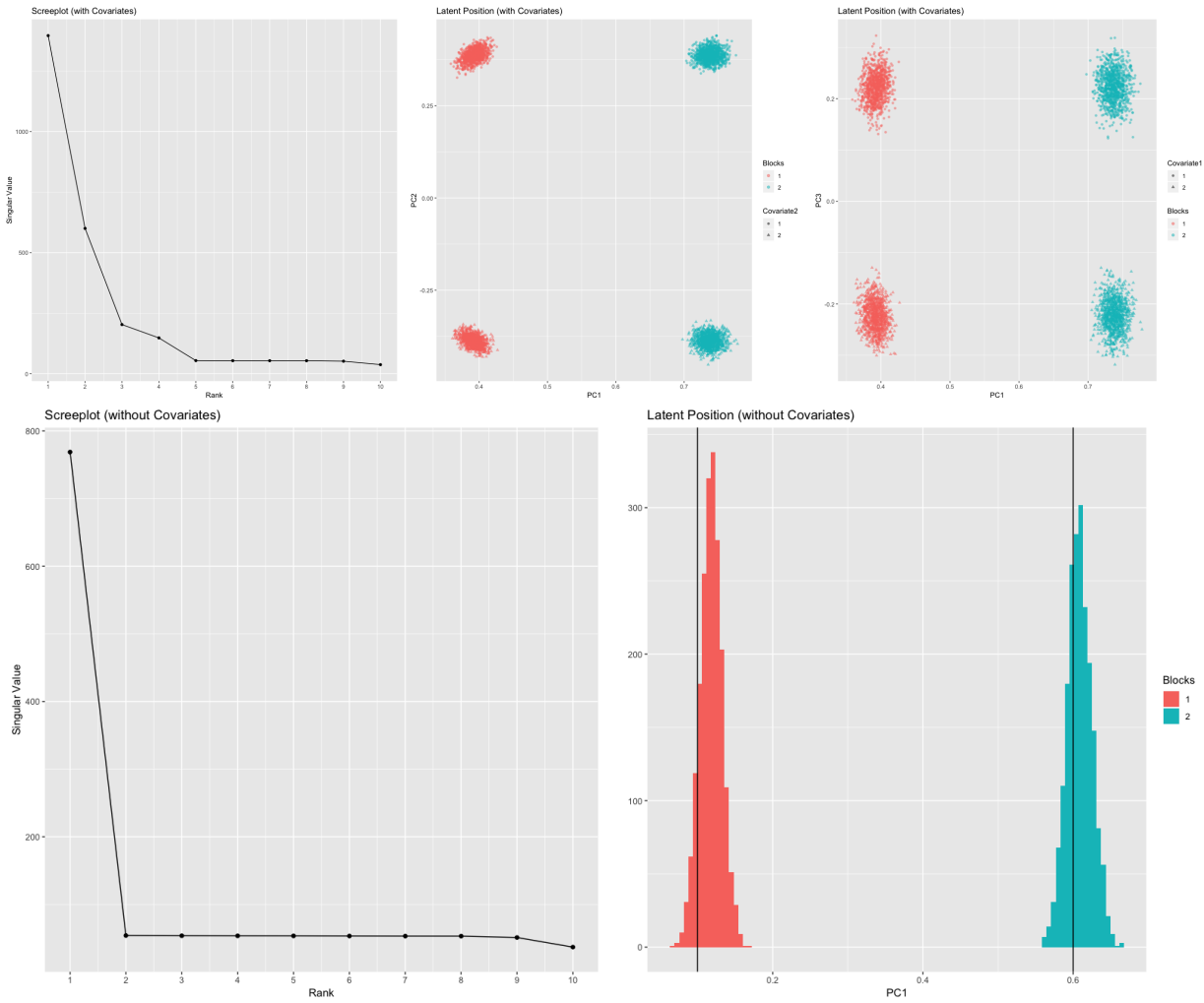

Figure 16: $K = 2,\ d = 1,\ n = 4000,\ \beta = [0.1, 0.3]$, Balanced

And follow the similar procedure in **Section 3**, we have

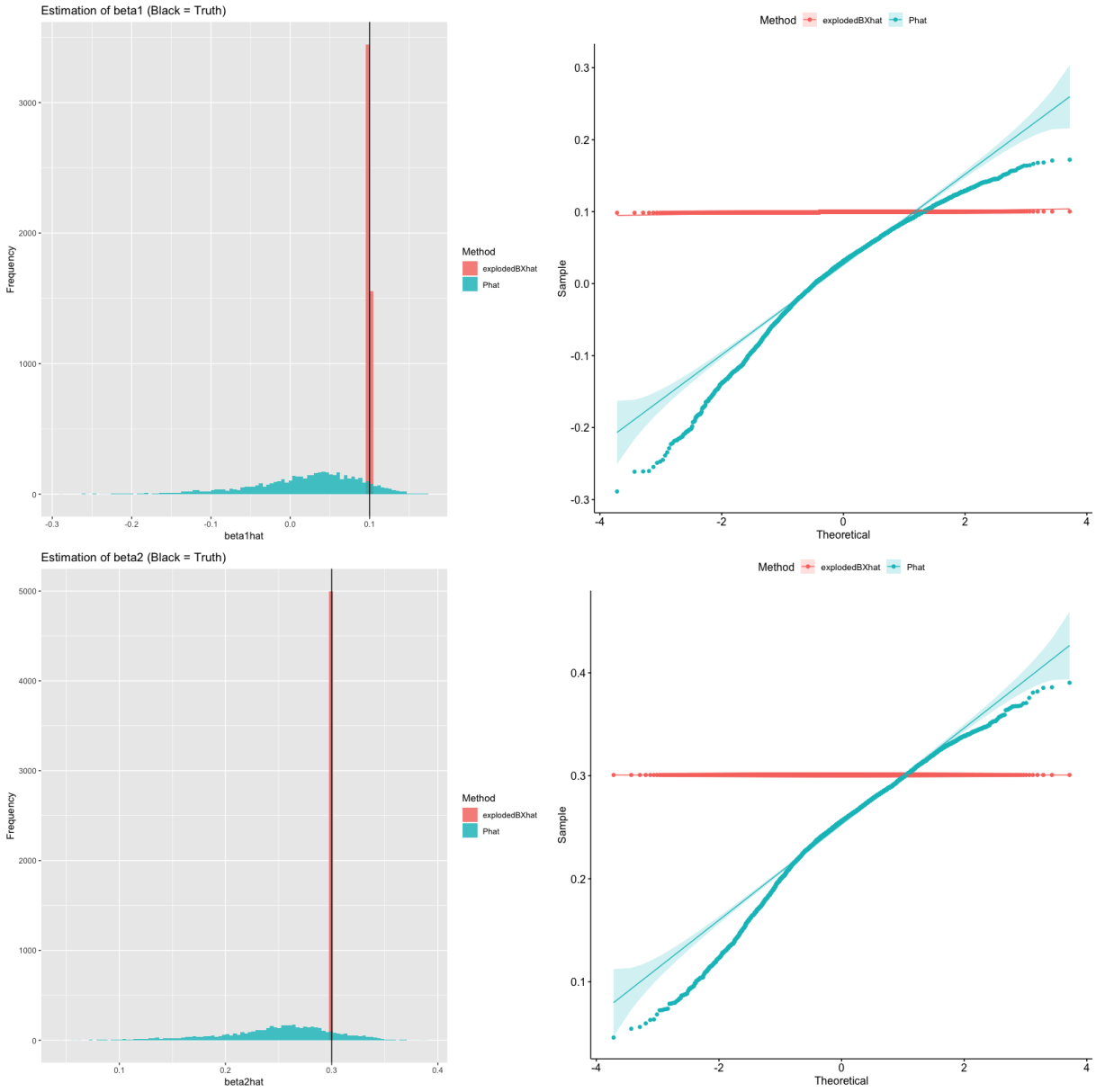

Figure 17: $K = 2,\ d = 1,\ n = 4000,\ \beta = [0.1, 0.3]$, Balanced

The Shapiro-Wilk Test (H_0: normal) gave `p-value < 0.01` for both $\hat{\beta}_1$ and $\hat{\beta}_2$.

Continuous Covariates

We start with one covariate that could take 4 discrete values. Let number of blocks $K = 2$, dimension of latent position $d = 1$. Consider latent position to be $[0.1, 0.7]$, $\beta = 0.3$ and $n = 4000$. Start with the balanced case, we have

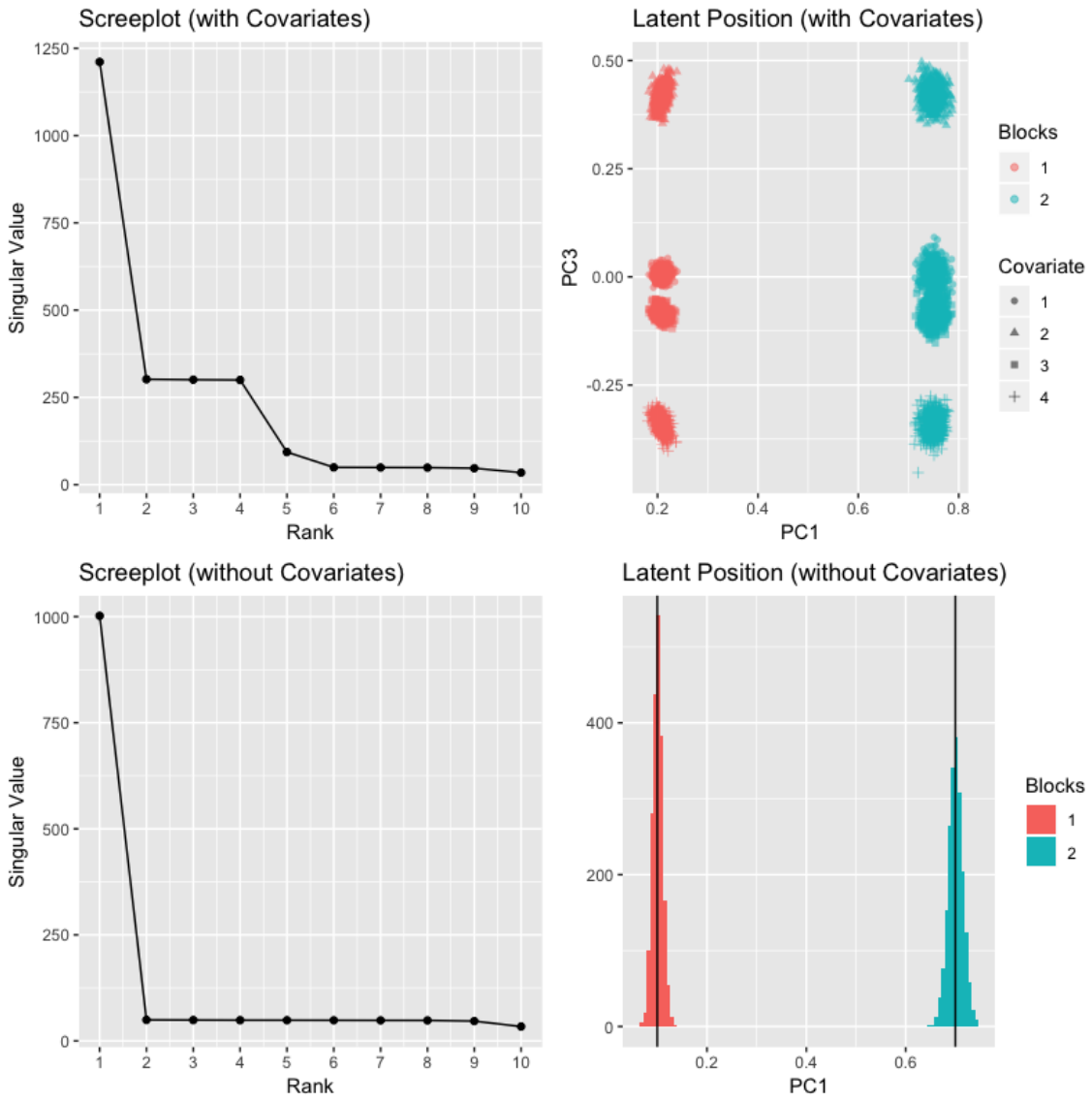

Figure 18: $K = 2,\ d = 1,\ n = 4000,\ \beta = 0.3$, Balanced

Follow the similar procedure in **Section 1**, we have $\hat{\beta} = 0.2992$. And follow the similar procedure in **Section 3**, we have

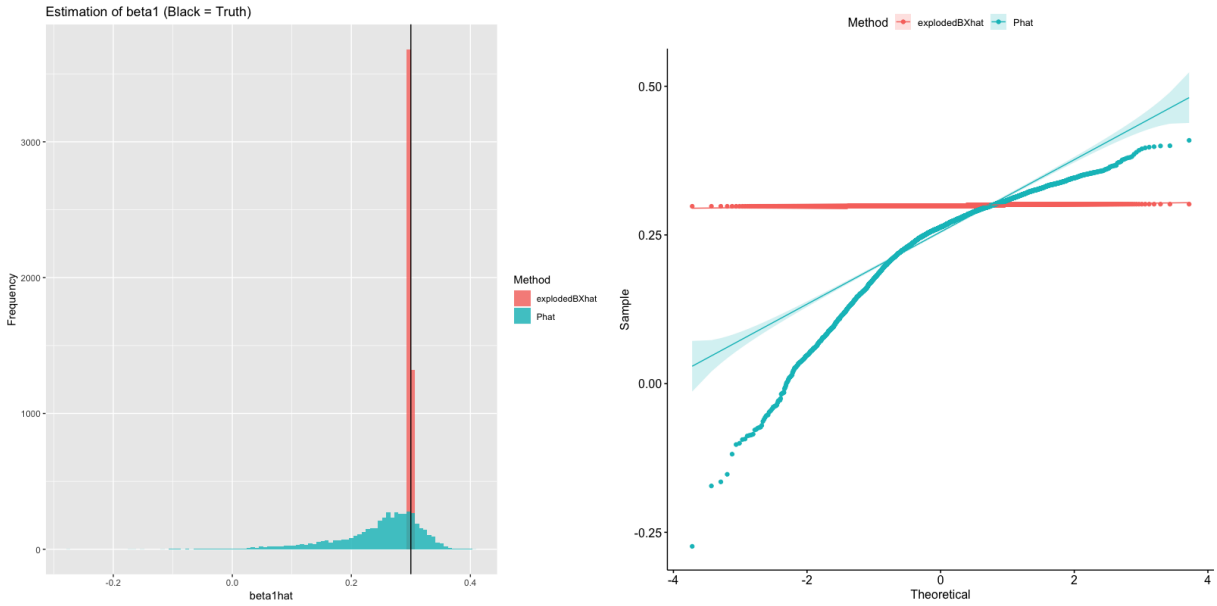

Figure 19: $K = 2,\ d = 1,\ n = 4000,\ \beta = 0.3$, Balanced

The Shapiro-Wilk Test (H_0: normal) gave `p-value < 0.01`.

References

[1] Rubin-Delanchy, P., Priebe, C. E., Tang, M., & Cape, J. (2017). A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint*, arXiv:1709.05506.