# (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 $\beta$.

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 $\beta$ 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 $\beta$ as $\hat{\beta} = 0.2995$ (runtime: 47s). Figure 1 shows the screeplots and latent postions with and without covariates.

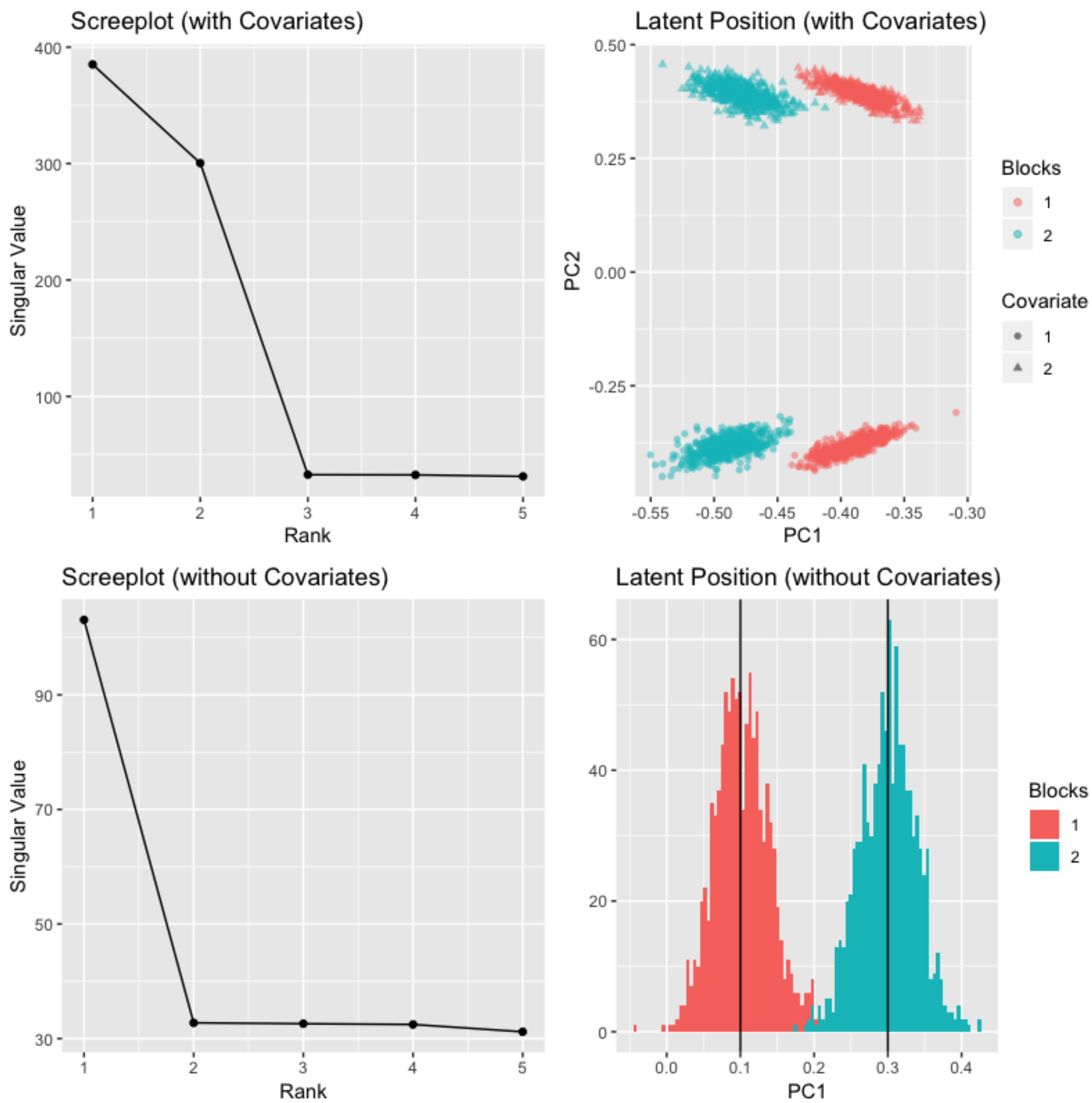


Figure 1: $K = 2,\ d = 1,\ n = 2000,\ \beta = 0.3$, Balanced

### 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 $\beta$ as $\hat{\beta} = 0.1999$ (runtime: 3m28s). Figure 2 shows the screeplots and latent postions with and without covariates.

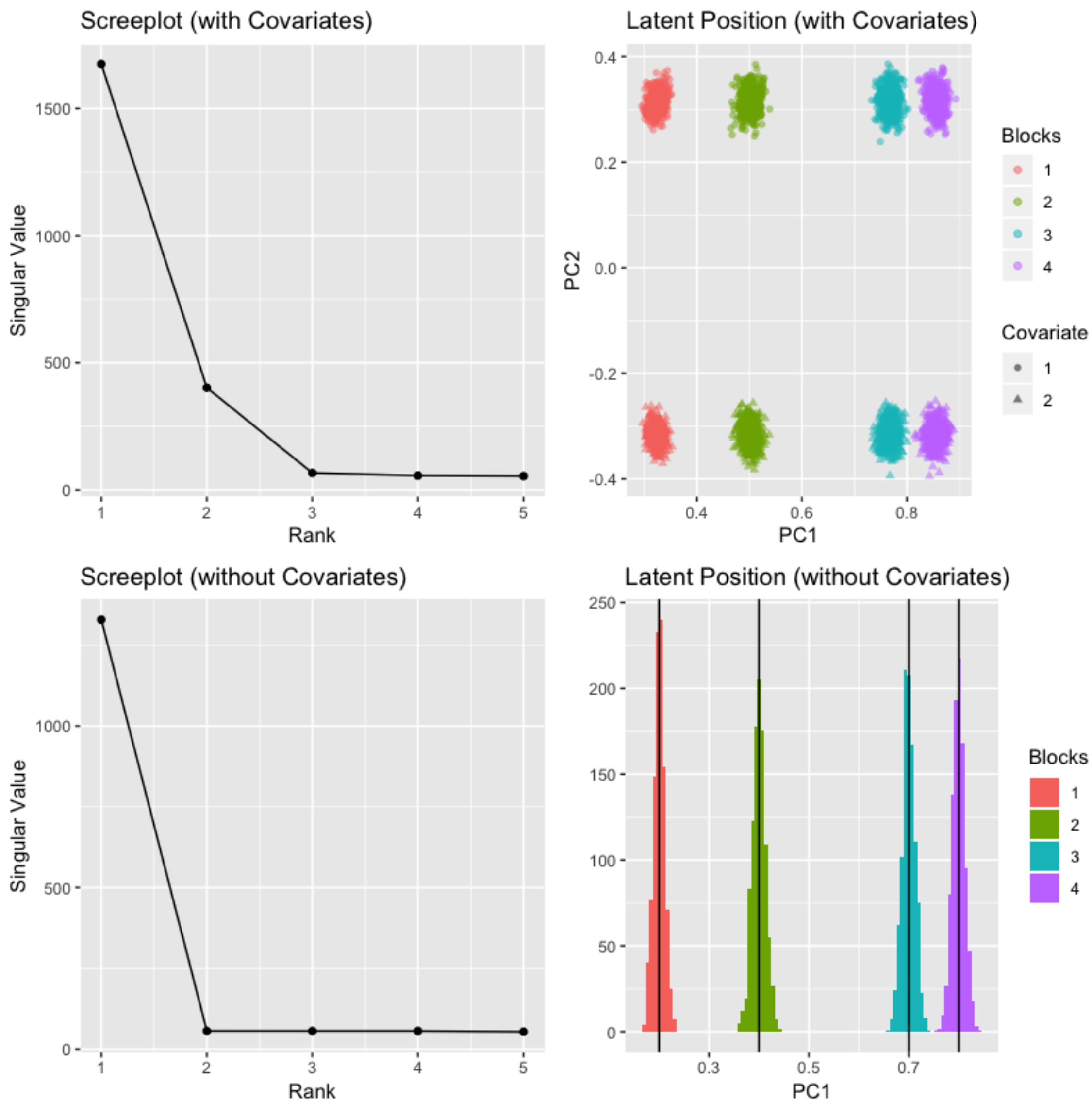


Figure 2: $K = 4,\ d = 1,\ n = 4000,\ \beta = 0.2$, Balanced

### 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 $\beta$ as $\hat{\beta} = 0.1383$ (runtime: 35m9s). Figure 3 shows the screeplots and latent postions with and without covariates.

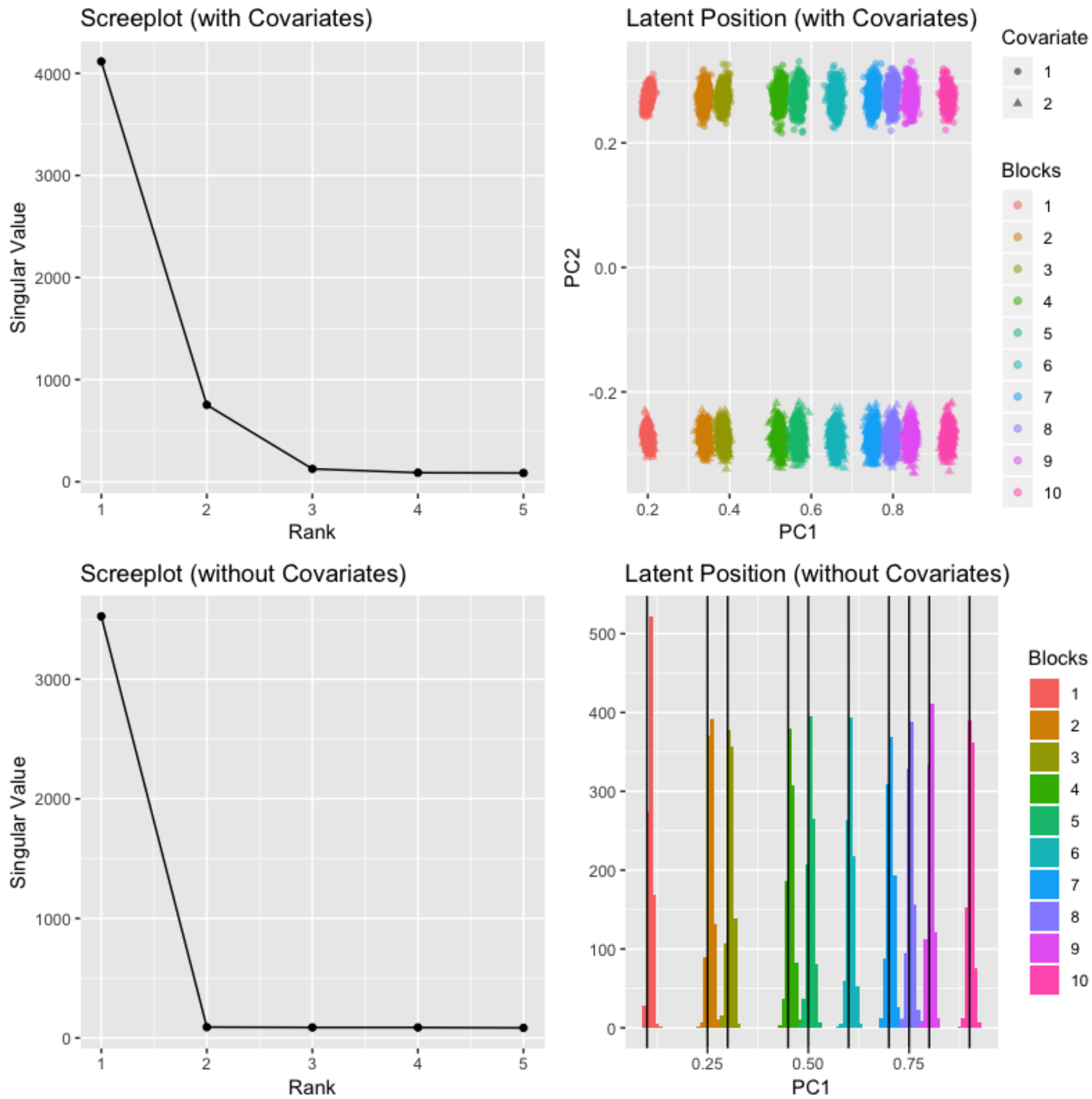


Figure 3: $K = 10,\ d = 1,\ n = 10000,\ \beta = 0.15$, Balanced

## 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 $\beta$ as $\hat{\beta} = -0.2000$ (runtime: 48s). Figure 4 shows the screeplots and latent postions with and without covariates.

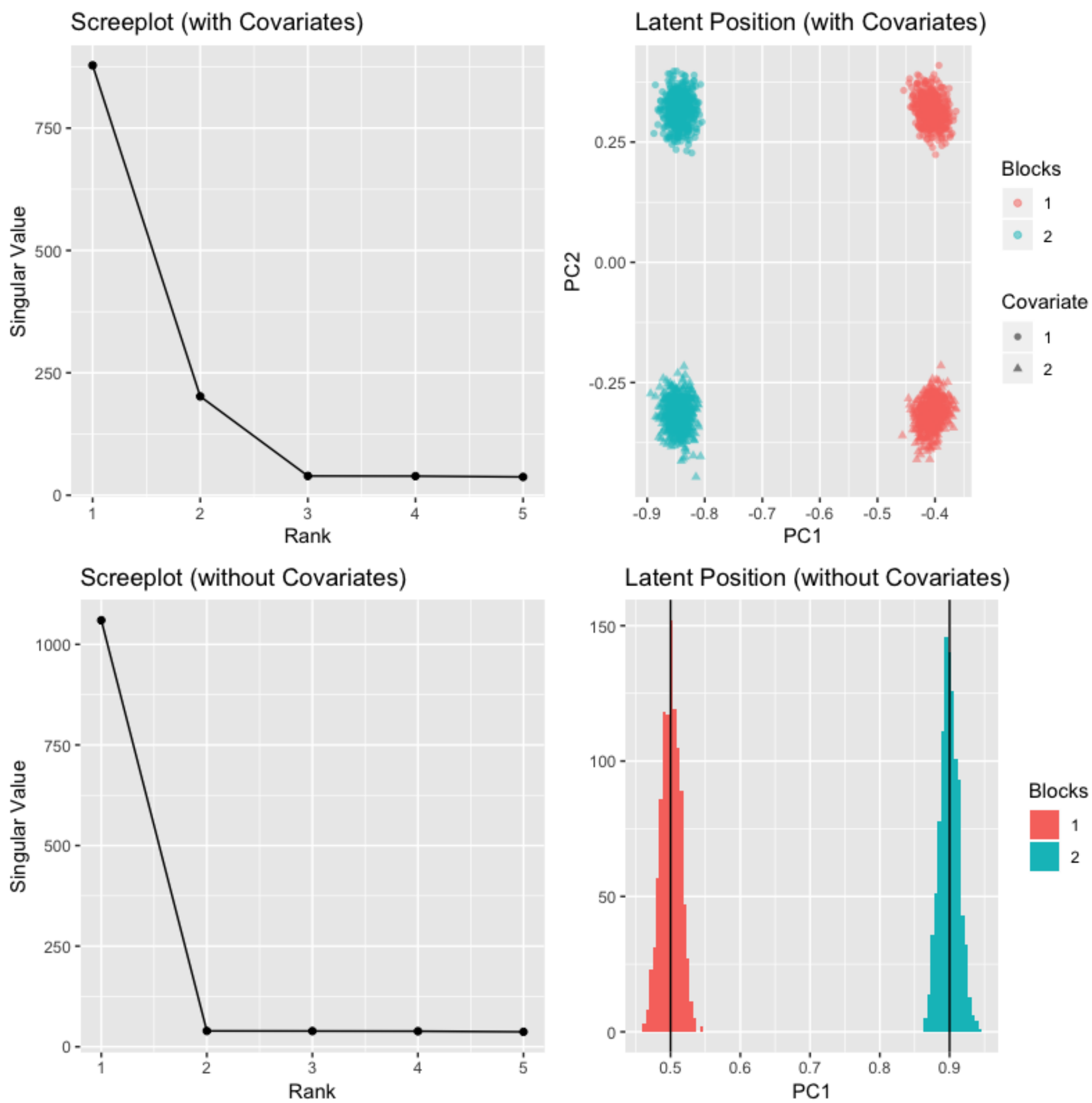


Figure 4: $K = 2,\ d = 1,\ n = 2000,\ \beta = -0.2$, Balanced

### 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 $\beta$ as $\hat{\beta} = -0.2999$ (runtime: 3m30s). Figure 5 shows the screeplots and latent postions with and without covariates.

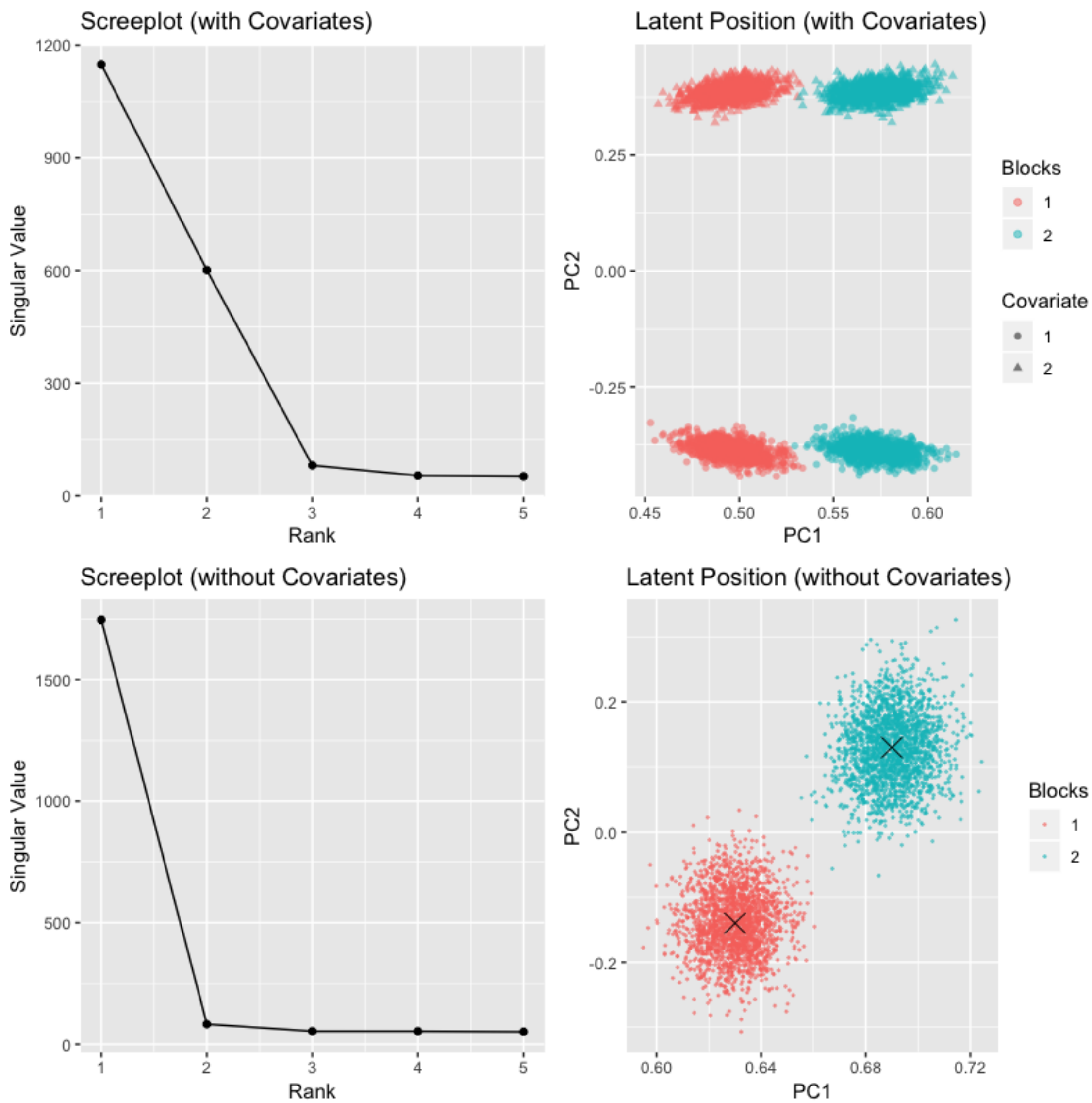


Figure 5: $K = 2,\ d = 2,\ n = 4000,\ \beta = -0.3$, Balanced

### 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 $\beta$ as $\hat{\beta} = -0.3492$ (runtime: 46s). Figure 6 shows the screeplots and latent postions with and without covariates.

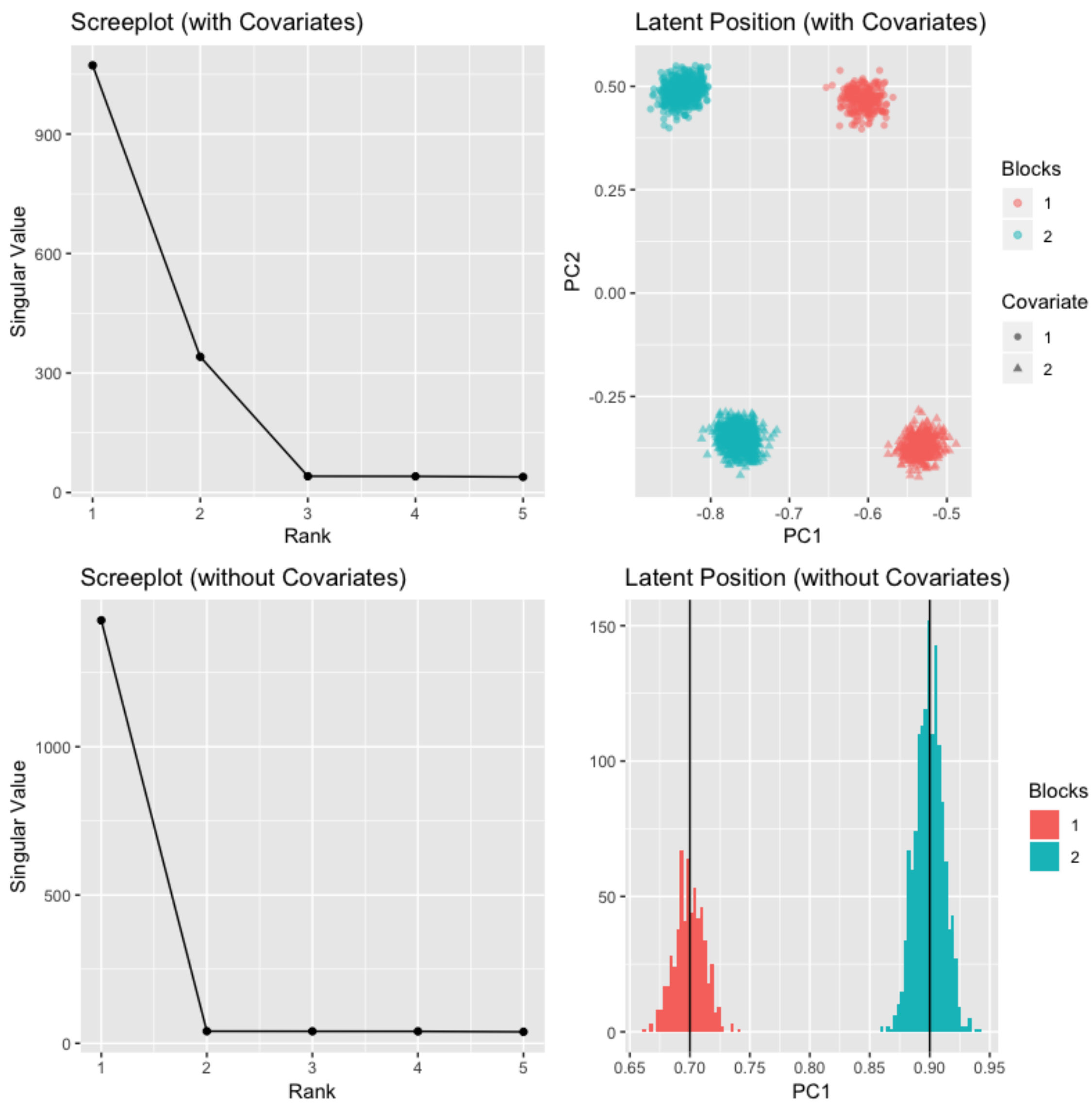


Figure 6: $K = 2,\ d = 1,\ n = 2000,\ \beta = -0.35$, Size of Block = (0.3, 0.7), Size of Gender = (0.4, 0.6)

### 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 $\beta$ as $\hat{\beta} = 0.3992$ (runtime: 3m33s). Figure 7 shows the screeplots and latent postions with and without covariates.

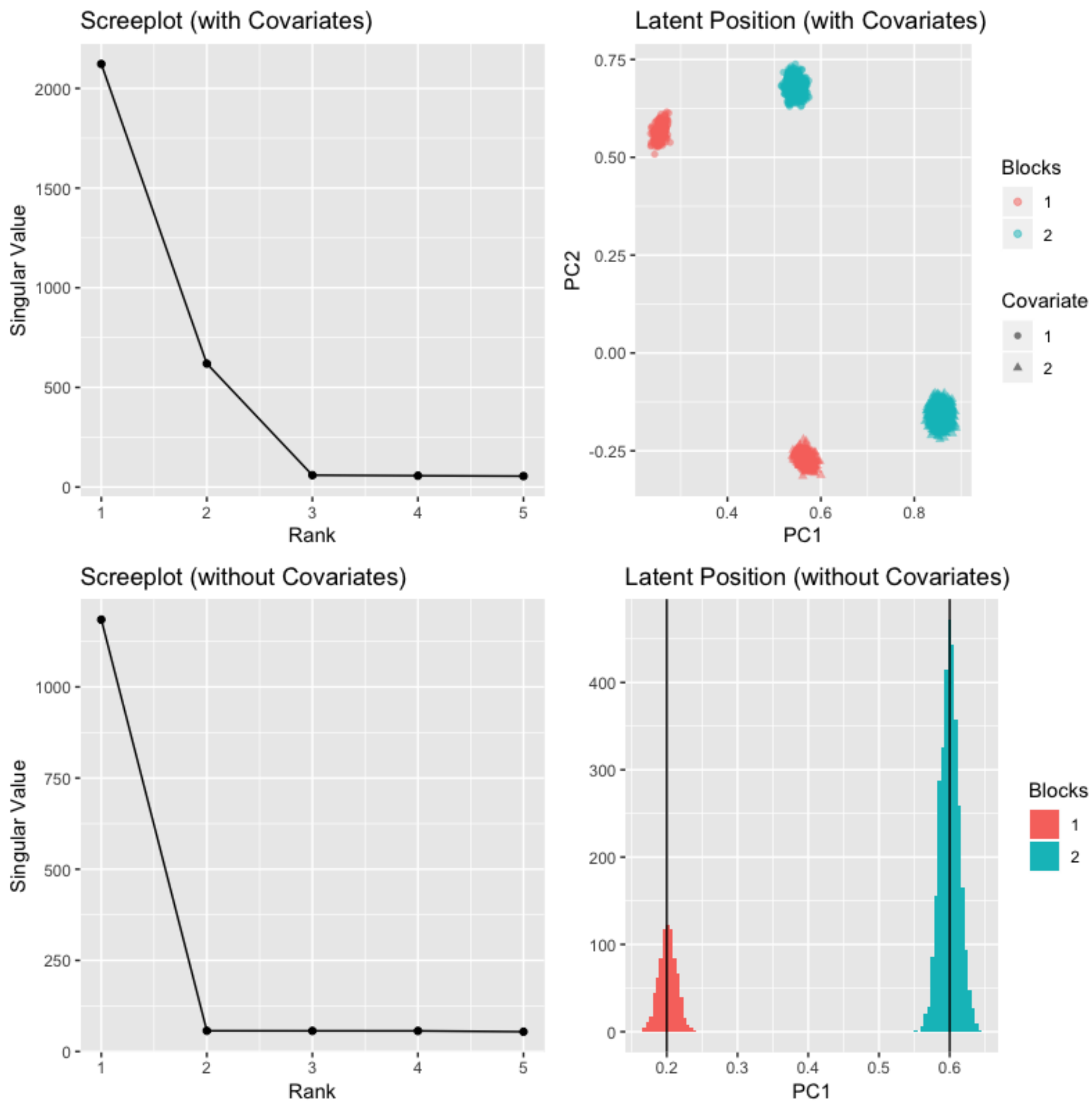


Figure 7: $K = 2,\ d = 1,\ n = 4000,\ \beta = 0.4$, Size of Block = (0.2, 0.8), Size of Gender = (0.3, 0.7)

### 2.4 Summary

With respect to the estimation of $\beta$, we summarize our simulations as follows.

| $n$ | $K$ | $d$ | **Size of Blocks** | **Size of Gender** | $\beta$ | $\hat{\beta}$ | **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 $\beta$) 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 $\hat{P} = \hat{X}\mathbf{I}_{d_1,d_2}\hat{X}^\top$.

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.

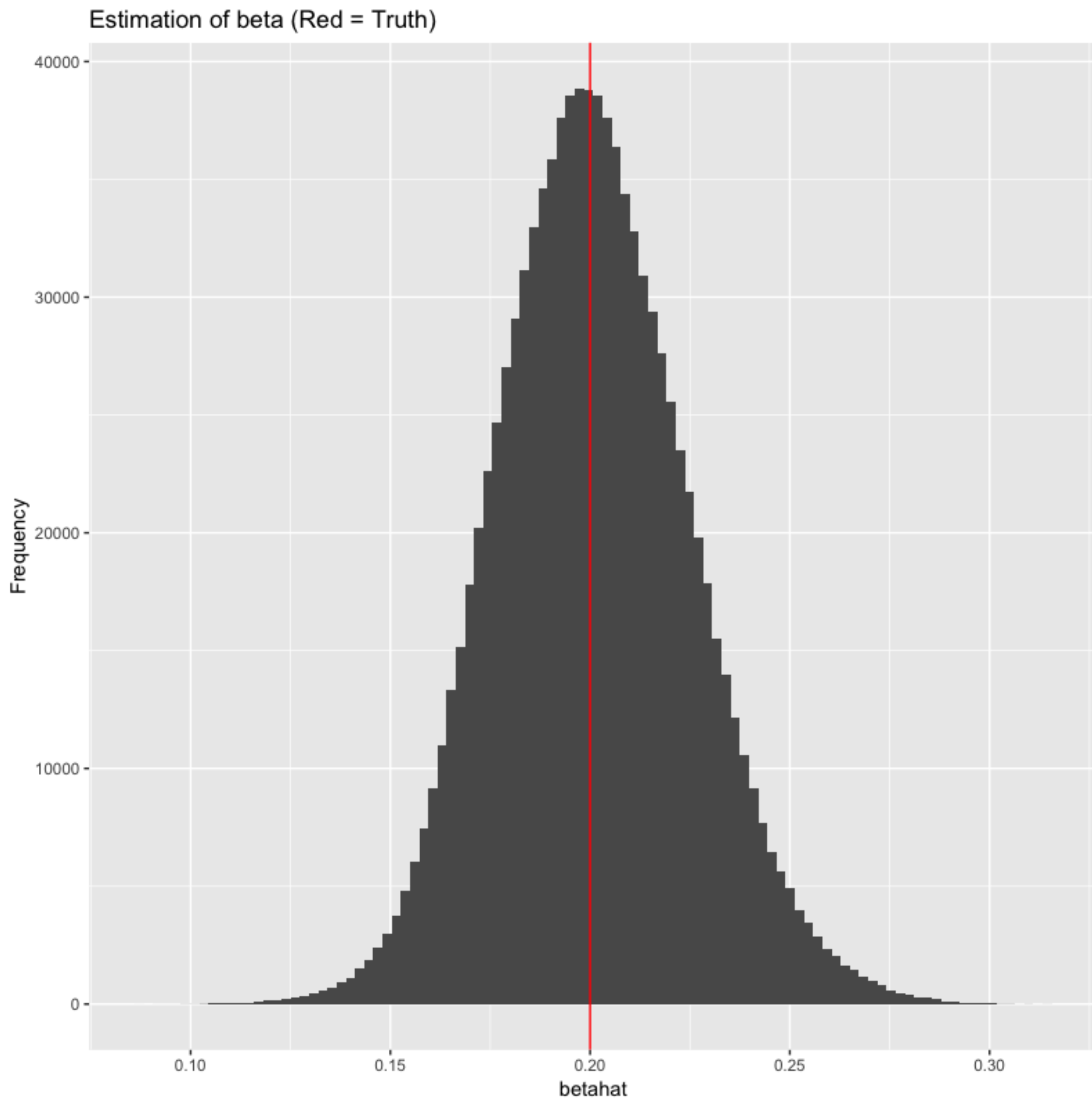


Figure 8: $K = 4,\ d = 1,\ n = 4000,\ \beta = 0.2$, Balanced

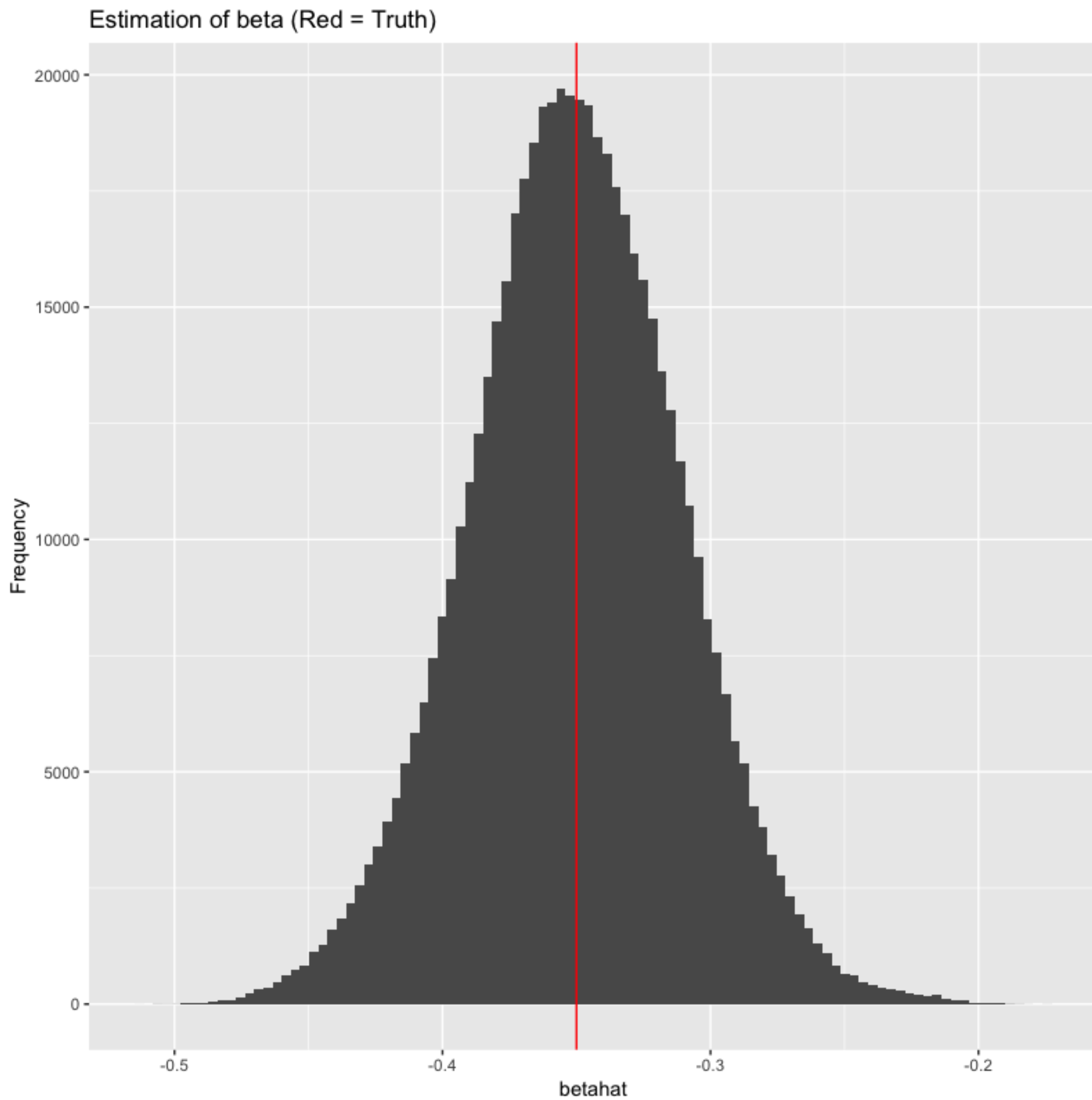


Figure 9: $K = 2,\ d = 1,\ n = 2000,\ \beta = -0.35$, Size of Block = (0.3, 0.7), Size of Gender = (0.4, 0.6)

# 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.