Sparse covariance and precision matrix estimation under matched design

Yukun Li, Raji Balasubramanian

University of Massachusetts Amherst

March 2021

Covariance Matrix and Precision Matrix

- Covariance matrix and precision matrix are two essential parts in multivariate analysis since they represent marginal and conditional dependence structures respectively.
- Let $\boldsymbol{X}$ be multivariate normal with covariance matrix $\boldsymbol{\Sigma}$, and the precision matrix $\boldsymbol{\Omega}$ is defined to be the inverse of the covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. For $i \neq j$,

  $\Sigma_{ij} = 0$ if and only if X_i and X_j are marginally independent

  $\Omega_{ij} = 0$ if and only if X_i and X_j are conditionally independent given all other variables

- The estimation of high dimensional covariance/precision matrix based on few sample observations is a difficult problem. Sample covariance matrix is singular and noninvertible.

Matched case-control Study

- Case-control study: One of the observational study designs for performing clinical research. Economical, quick to perform, and easy to implement.
- Matched case-control study: Match the cases and controls for confounding factors.
- Matched data are not randomly collected.
- Problem: Recover the covariance/precision matrix of variables under matched case-control design

Inverse probability weighting

- A well-known technique used in controlling for selection biases in non-experimental studies and in many missing data problems.
- Create a pseudo population by giving a weight to each observation in the case control sample.
- Weight: inverse of sampling probability. Up-weight the observations which have low probability of being in the case control sample, and down-weight those that have high probability.

Method: Low Dimension

- Covariance estimation: Sample covariance with inverse probability weighting

$$\boldsymbol{S}_{ipw} = \frac{1}{\sum_{i=1}^{n} w_i - 1} \sum_{i=1}^{n} w_i (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^*)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^*)^T \tag{1}$$

  where w_i is weight for the i^{th} subject, $\hat{\boldsymbol{\mu}}^* = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \boldsymbol{X}_i$ is the weighted sample mean.

- Precision estimation: Inverse of estimated covariance

$$\hat{\Omega} = \boldsymbol{S}_{ipw}^{-1} \tag{2}$$

Method: High Dimension

- Given a set $\boldsymbol{X}$ of n i.i.d. vectors following $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have the weighted log likelihood function

$$\log f(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{\sum_{i=1}^{n} w_i}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} w_i \left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{X}_i - \boldsymbol{\mu}\right) + c$$

Adopting the idea from Bien & Tibshrirani (2010) and Friedman et al. (2007), we can add lasso penalty to the likelihood function.

- Covaraince matrix:

$$\hat{\boldsymbol{\Sigma}} = argmax_{\boldsymbol{\Sigma}} \{ -\log\det \boldsymbol{\Sigma} - \mathrm{tr}(\boldsymbol{S}^*_{ipw} \boldsymbol{\Sigma}^{-1}) - \lambda \|\boldsymbol{\Sigma}\|_1 \} \tag{3}$$

- Precision matrix:

$$\hat{\boldsymbol{\Omega}} = argmax_{\boldsymbol{\Omega}} \{ \log\det \boldsymbol{\Omega} - \mathrm{tr}(\boldsymbol{S}^*_{ipw} \boldsymbol{\Omega}) - \lambda \|\boldsymbol{\Omega}\|_1 \} \tag{4}$$

where $\boldsymbol{S}^*_{ipw} = \frac{1}{\sum_{i=\mathbf{1}}^{n} w_i} \sum_{i=1}^{n} w_i \left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)^T \left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)$, and λ is the tuning parameter

Method: High Dimensional Covariance Matrix Estimation

- Problem (3)

$$\hat{\Sigma} = argmax_{\Sigma}\{-\log\det\Sigma - \mathrm{tr}(\boldsymbol{S}^{*}_{ipw}\Sigma^{-1}) - \lambda\|\Sigma\|_1\}$$

- This problem is not convex. According to Bien & Tibshrirani (2010), it can be solved by majorize-minimize iteration

$$\hat{\Sigma}^{(t)} = \mathrm{argmin}_{\Sigma}\left[\mathrm{tr}\left\{\left(\hat{\Sigma}^{(t-1)}\right)^{-1}\Sigma\right\} + \mathrm{tr}\left(\Sigma^{-1}\boldsymbol{S}^{*}_{ipw}\right) + \lambda\|\Sigma\|_1\right]$$

- Then generalized gradient descent can be applied to solve this problem

$$\Sigma \leftarrow \mathcal{S}\left\{\Sigma - t\left(\Sigma_0^{-1} - \Sigma^{-1}\boldsymbol{S}^{*}_{ipw}\Sigma^{-1}\right), \lambda t\right\} \tag{5}$$

where $\mathcal{S}$ is the elementwise soft-thresholding operator defined by $\mathcal{S}(A,B)_{ij} = \mathrm{sign}\left(A_{ij}\right)\left(A_{ij} - B_{ij}\right)_{+}$.

Method: High Dimensional Precision Matrix Estimation

- Problem (4)

$$\hat{\mathbf{\Omega}} = argmax_{\mathbf{\Omega}}\{\log\det\mathbf{\Omega} - \mathrm{tr}(\boldsymbol{S}^{*}_{ipw}\mathbf{\Omega}) - \lambda\|\mathbf{\Omega}\|_1\}$$

- Graphical Lasso (Friedman et al.(2007)) can be applied to solve this problem by optimizing over each row and corresponding column of the covariance matrix in a block coordinate descent fashion.

Method: Weight estimation

- Suppose the subjects in the cohort are divided into J strata based on $\boldsymbol{Z}$
- Randomly select n_1 cases from the N_1 cases in the cohort
- Select controls by matching for strata J
- The weight can be estimated by

$$\hat{w}_{i1} = N_1/n_1$$
$$\hat{w}_{i0} = N_{0j}/n_{0j} \text{ for control subject in stratum j}$$

where
N_1: number of case subjects in the cohort
n_1: number of case subjects in the matched data
N_{0j}: number of control subjects in stratum j in the cohort
n_{0j}: number of control subjects in stratum j in the matched sample

Simulation Study: Low Dimension Scenario

- Cohort sample size N = 20000
- Matching variable: $\boldsymbol{Z} = (Z_1, Z_2)$, where $Z_1 \sim N(0, 1)$, $Z_2 \sim Ber(0.5)$
- Variables of interest: $\boldsymbol{X} = (X_1, X_2, ..., X_20)^T \sim N(\boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Z}}, Cov(\boldsymbol{X}|\boldsymbol{Z}))$ where

$$\boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Z}} = \begin{cases} 0.1 + \alpha_Z * Z_1 + 0.1 * Z_2 & \text{for } X_1, ..., X_{10} \\ 0 & \text{Others} \end{cases}$$

  and $Cov(\boldsymbol{X}|\boldsymbol{Z})$ has 20% non-zero elements randomly selected from U(-2, -1) or U(1, 2).
- Case-control status Y

$$logit(Y) = \beta_0 + \sum_{i=1}^{10} 0.5X_i + \beta_Z * Z_1 + 0.2 * Z_2$$

- Set $\beta_Z = 1.2$ and $\alpha_Z = \{0.3, 0.5, 0.7\}$ to investigate the effect of α_Z.
- β_0: chosen to make the proportion of cases in the cohort is 20%

Simulation Study: Low Dimension Scenario

- Matching procedure
  - Cohort subjects are divided into 10 strata based on their propensity scores $P(Y = 1|Z_1, Z_2)$, which can be estimated by logistic regression;
  - Randomly select 500 cases from the cohort;
  - Match the cases with 500 controls according to the strata defined by propensity scores;
  - Bin the adjacency strata if their $\frac{\#\text{ of controls}}{\#\text{ of cases}}$ are greater than 50
- We compare the following 4 covariance estimators to the true covariance matrix of $\boldsymbol{X}$:
  Sample covariance of $\boldsymbol{X}$ from random samples : $\boldsymbol{S}_{random}$
  Sample covariance of $\boldsymbol{X}$ from case-control samples: $\boldsymbol{S}_{CC}$
  Sample covariance of $\boldsymbol{X}$ from matched case-control samples: $\boldsymbol{S}_{MCC}$
  Proposed weighted covariance of $\boldsymbol{X}$ from matched case-control samples: $\boldsymbol{S}_{ipw}$

Simulation Study: Low Dimension Scenario

- Compare the four estimates with the true covariance matrix of $\boldsymbol{X}$

$$\begin{aligned} Cov(\boldsymbol{X}) &= E(Cov(\boldsymbol{X}|\boldsymbol{Z})) + Cov(E(\boldsymbol{X}|\boldsymbol{Z})) \\ &= Cov(\boldsymbol{X}|\boldsymbol{Z}) + Cov(\boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Z}}) \end{aligned}$$

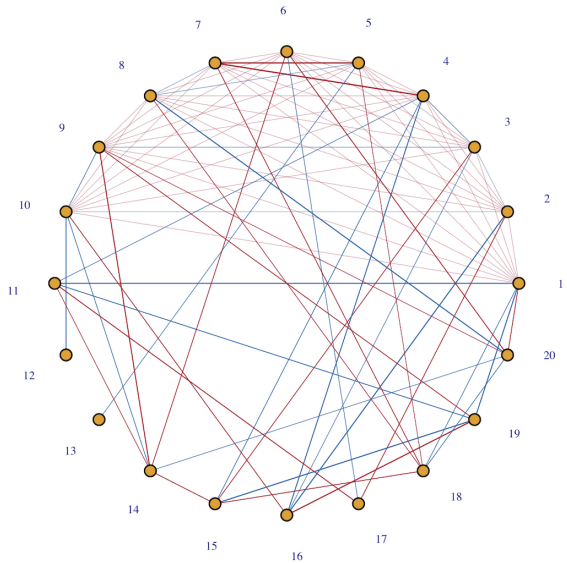


Figure: Visualization of true Cov($\boldsymbol{X}$)

Simulation Study: Low Dimension Scenario

- Error metric:

$$\text{Average Relative Bias} = \frac{1}{p^2} \sum_{i}^{p} \sum_{j}^{p} \frac{\hat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}}{1 + abs(\mathbf{\Sigma}_{ij})} \tag{6}$$

- For low dimensional setting, comparison of precision matrix estimators can be directly obtained by inverting the covariance estimates.

Simulation Result: Low Dimension Scenario

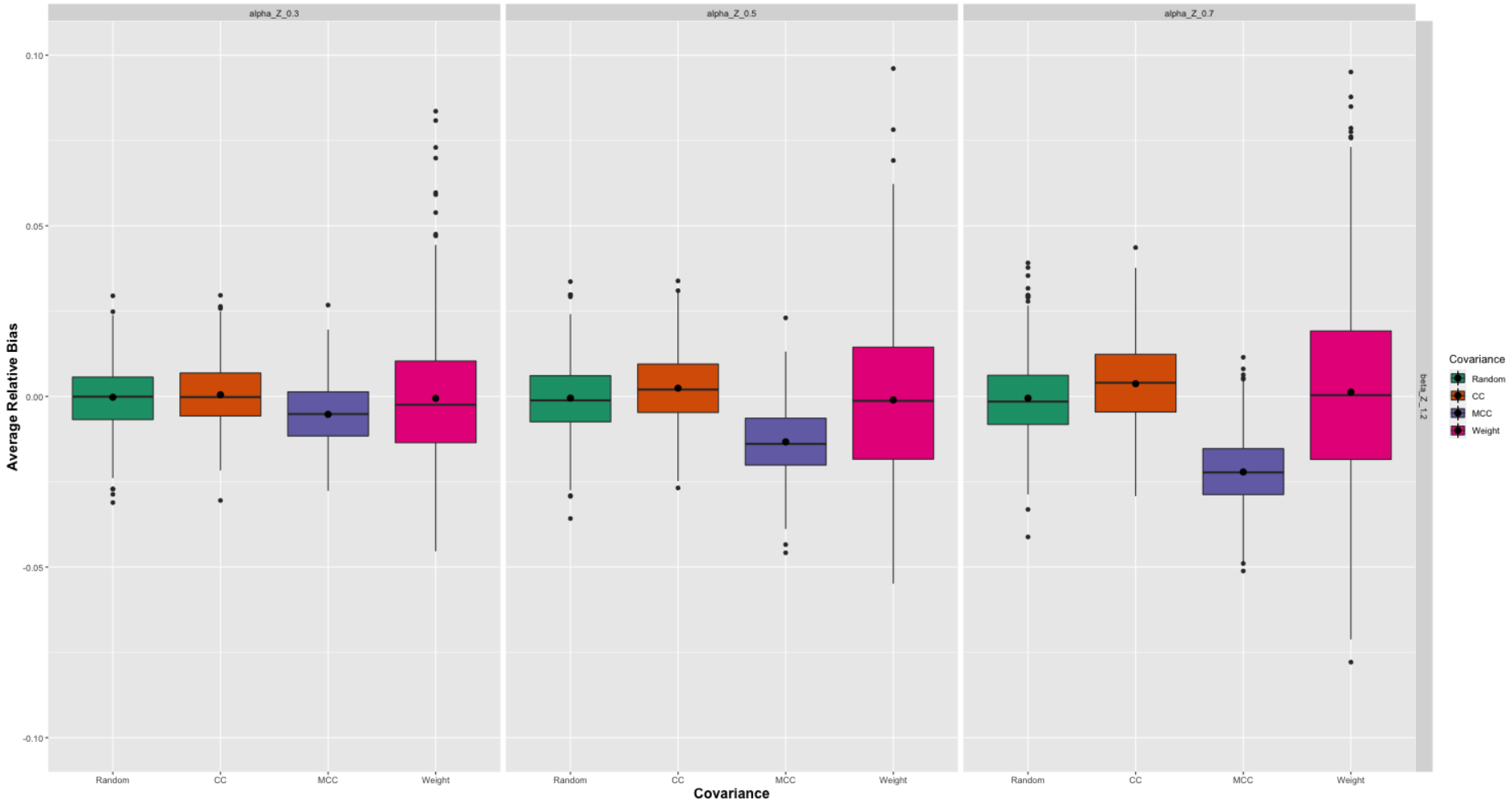


Figure: Covariance matrix estimation for low dimension scenario

Effect of Binning

- Bin the adjacency strata if their $\frac{\#\text{ of controls}}{\#\text{ of cases}}$ are greater than 50, 25 and 12.5

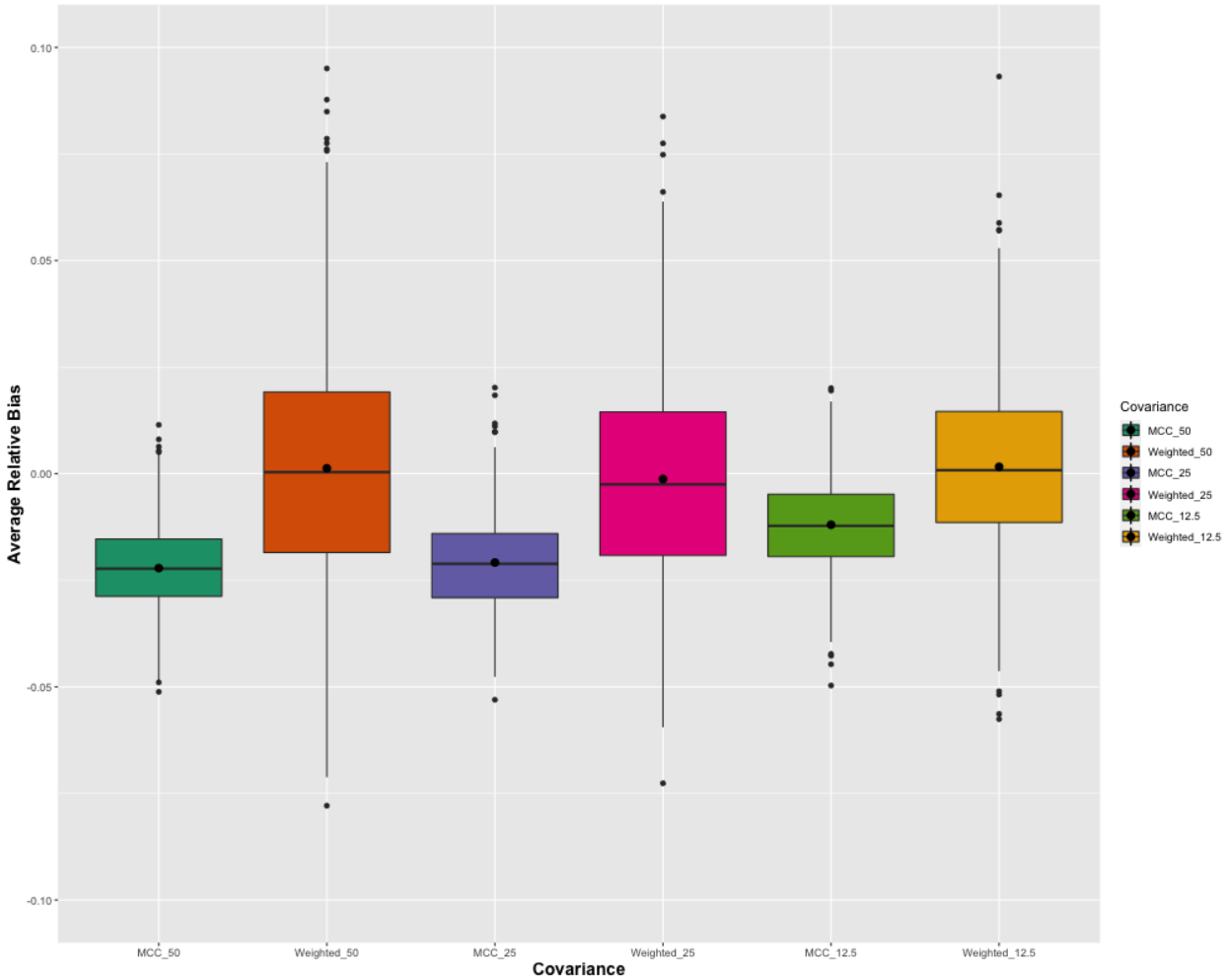
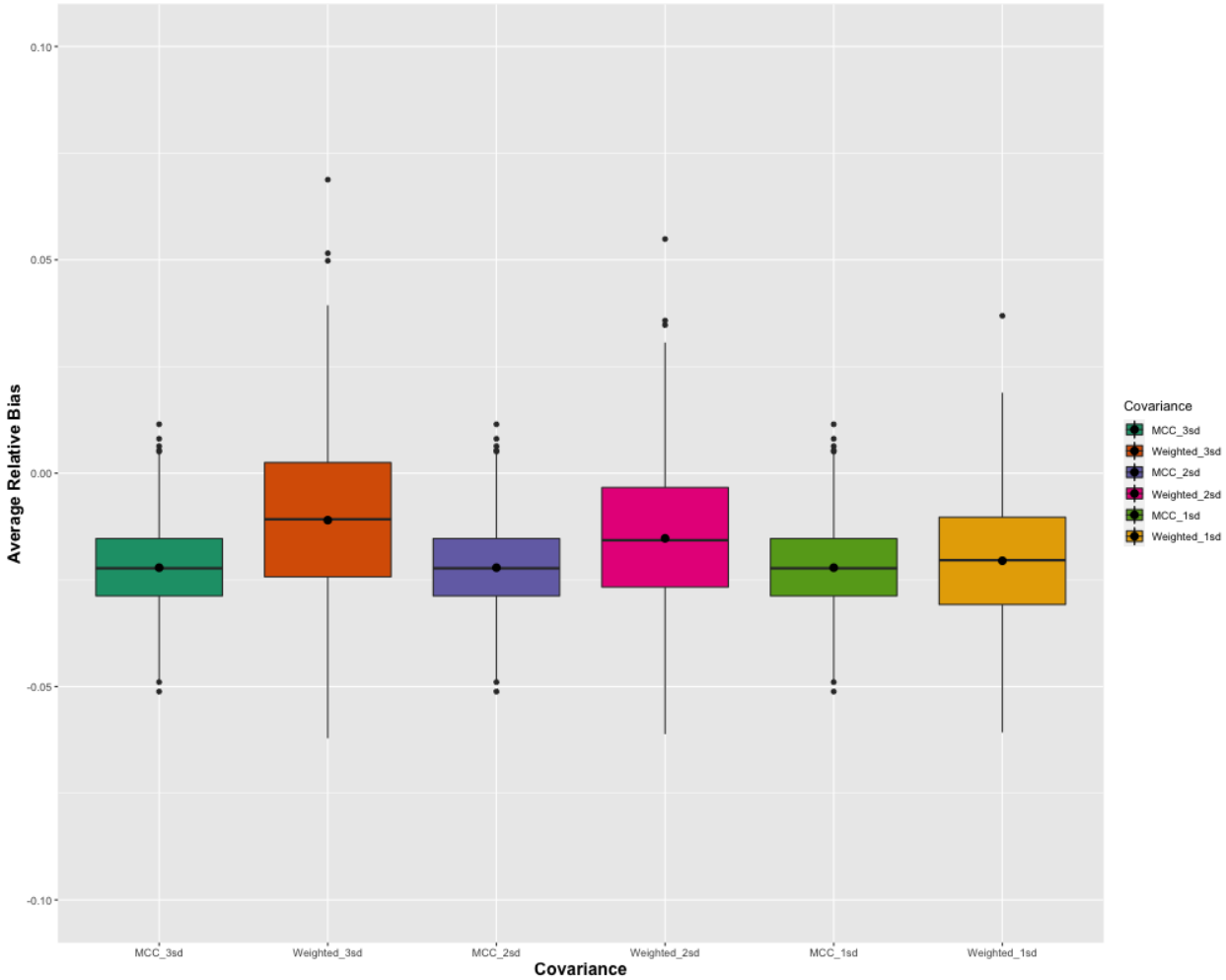


Figure: Effect of binning when $\alpha_Z = 0.7$ and $\beta_Z = 1.2$

Effect of Thresholding

- Thresholding can help to reduce the effect of extreme weights by truncating them at a maximum allowable weight.
- Choice of threshold: mean + 3/2/1*SD



Conclusion

- Proposed weighted estimator in low dimensional setting and a likelihood-based estimation procedure in high dimensional setting
- Simulation for low dimensional setting shows simple covariance estimation from matched case control sample has significant bias, and estimation from our proposed method has almost no bias for all simulation scenarios.
- Variance reduction methods are needed for our proposed method.
- Simulation for high dimensional setting will be conducted.

Reference

- Bien, Jacob, and Robert J. Tibshirani. "Sparse estimation of a covariance matrix." Biometrika 98.4 (2011): 807-820.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." Biostatistics 9.3 (2008): 432-441.