Combining observed and simulated data

Suppose we observe $X_{obs} \in \mathbb{R}^{n \times p}$ and $y_{obs} \in \mathbb{R}^n$ with $p \gg n$
We want to estimate β such that

$$y_{obs} = X_{obs}\beta + \epsilon_1, \quad \epsilon_1 \sim N(0, \sigma^2 I)$$

Additionally, suppose we have access to simulated data of the same nature: $X_{sim} \in \mathbb{R}^{m \times p}$ and $y_{sim} \in \mathbb{R}^m$ with $p \asymp m$. We assume

$$y_{sim} = X_{sim}(\beta + \Delta) + \epsilon_2, \quad \epsilon_2 \sim N(0, \sigma^2 I), \quad \Delta \sim N(0, \Sigma_\Delta)$$

where Δ is a bias term. We are interested in the conditions under which including (X_{sim}, y_{sim}) improves the estimation of β.

Ridge regression three ways

Let $\tilde{\Sigma} = X_{sim}\Sigma_\Delta X_{sim}^T + \sigma^2 I_m$. We consider three different linear systems we can solve to recover β.

$$y_{obs} = X_{obs}\beta + \epsilon_1, \quad \epsilon_1 \sim N(0, \sigma^2 I_n) \tag{1}$$

$$\begin{bmatrix} y_{obs} \\ y_{sim} \end{bmatrix} = \begin{bmatrix} X_{obs} \\ X_{sim} \end{bmatrix} \beta + \epsilon_2, \quad \epsilon_2 \sim N\left(0, \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & \tilde{\Sigma} \end{bmatrix}\right) \tag{2}$$

$$\begin{bmatrix} y_{obs} \\ \tilde{\Sigma}^{-\frac{1}{2}} y_{sim} \end{bmatrix} = \begin{bmatrix} X_{obs} \\ \tilde{\Sigma}^{-\frac{1}{2}} X_{sim} \end{bmatrix} \beta + \epsilon_3, \quad \epsilon_3 \sim N\left(0, \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & I_m \end{bmatrix}\right) \tag{3}$$

Because $p \gg n$, we will add a ridge penalty when estimating these models.

Risk of the estimators

Suppose X_{obs} and X_{sim} are rotated such that $X_{obs}^T X_{obs} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $X_{sim}^T X_{sim} = \text{diag}(\delta_1, \ldots, \delta_p)$. Also assume $\Sigma_\Delta = \text{diag}(\alpha_1, \ldots, \alpha_p)$. Then, for a given ridge regularization parameter λ, the risk $\mathbb{E}\|\hat{\beta} - \beta\|_2^2$ for each of these estimators is:

$$\text{risk}(\hat{\beta}_1) = \frac{\sigma^2}{n} \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + \lambda)^2} + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda}{\lambda_j + \lambda} \right)^2$$

$$\text{risk}(\hat{\beta}_2) = \frac{\sigma^2}{n+m} \sum_{j=1}^{p} \frac{\lambda_j + \delta_j + \frac{1}{\sigma^2}\delta_j^2 \alpha_j}{(\lambda_j + \delta_j + \lambda)^2} + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda}{\lambda_j + \delta_j + \lambda} \right)^2$$

$$\text{risk}(\hat{\beta}_3) = \frac{\sigma^2}{n+m} \sum_{j=1}^{p} \frac{\lambda_j + \frac{1}{\sigma^2}\xi_j}{(\lambda_j + \xi_j + \lambda)^2} + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda}{\lambda_j + \xi_j + \lambda} \right)^2$$

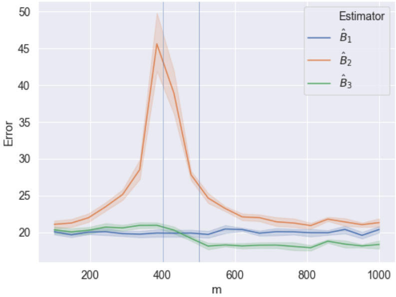
$$\xi_j = \frac{1}{\sigma^2}\delta_j \left(1 - \frac{\delta_j \alpha_j}{\sigma^2 + \delta_j \alpha_j} \right)$$

Simulations: varying m

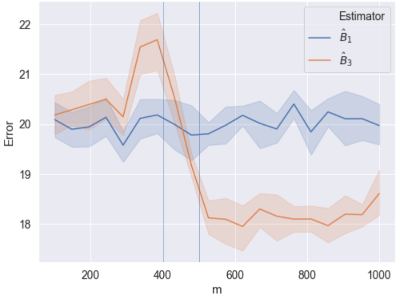
First we consider how the number of samples in X_{sim} impacts the error of the estimator relative to a fixed n and p. We generate data as follows:

$$X_{obs} \sim N(0, I_n), X_{sim} \sim N(0, I_m), \beta_j \sim N(0, 1), \Delta \sim N(0, \eta^2 I_p)$$

where $n = 100, p = 500, \eta^2 = 1, \lambda = 0.1$



Here we see that estimator (2) performs extremely poorly when m = p-n



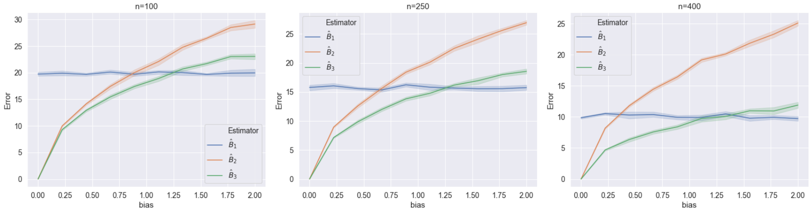
Zooming in on estimators (1) and (3), we see that (3) overtakes the performance of (1) once m >= p.

Simulations: varying η^2

We now investigate the influence of the size of the variance introduced by the bias term, η^2. We do this for a few values of n. Our setup:

$$X_{obs} \sim N(0, I_n), X_{sim} \sim N(0, I_m), \beta_j \sim N(0, 1), \Delta \sim N(0, \eta^2 I_p)$$

$p = 500, m = 1000, \lambda = 0.1$



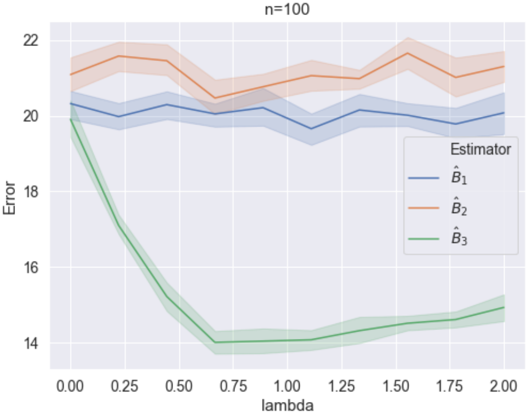
For all values of n, estimator (3) outperforms (1) until the bias term is around 1.25. As n approaches p, the performance of (2) becomes worse than (1) as the bias decreases.

Simulations: varying λ

We now investigate the influence of the regularization parameter λ. We do this for a few values of n. Our setup:

$$X_{obs} \sim N(0, I_n), X_{sim} \sim N(0, I_m), \beta_j \sim N(0, 1), \Delta \sim N(0, \eta^2 I_p)$$

$n = 100, p = 500, m = 1000, \eta^2 = 1$



The regularization parameter seems to not have much of an influence on estimators (1) and (2) but we see a significant improvement in (3) as lambda approaches 1.