Penalized Mixed Models to Adjust for Batch Effects and Unobserved Confounding in High Dimensional Regression

Yujing Lu Patrick Breheny

IOWA

Summary

Confounding can lead to spurious associations. In high-dimensional studies, recent research has shown that even when confounders are unobserved, they can still leave traces upon multiple features, which makes it possible to adjust for them.

In this study, we assess how unobserved confounding introduces bias and variability into the data. We quantify the magnitude and structure of these effects by examining the ratios between bias, signal, and noise. We specifically investigate the impact of the amount and complexity of unobserved confounding on the performance of LASSO, principal components LASSO (PC-LASSO), and **penalized linear mixed models** (PLMMs). We find that:

- **Both methods outperform regular LASSO** as the amount of confounding increases.
- **PLMMs are more robust in handling complex confounding structures** than PC-LASSO.
- In terms of preventing spurious associations, **PLMMs select signals more precisely than PC-LASSO.**
- **PLMMs outperform PC-LASSO with semi-synthetic data** as well.

Linear Confounding Model

We consider the following setting:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},\\ \mathbf{X} &= \mathbf{D} + \mathbf{Z}\mathbf{A}^\top,\\ \boldsymbol{\varepsilon} &\sim \mathrm{N}(\mathbf{0}, \sigma_e^2\mathbf{I}_n),\end{aligned}$$

where

- $\mathbf{y}$ is an $n \times 1$ vector of outcomes
- $\mathbf{X}$ is an $n \times p$ matrix of observed features
- $\mathbf{Z}$ is an $n \times q$ matrix of unobserved confounders
- $\mathbf{D}$ is an $n \times p$ matrix that is independent of $\mathbf{Z}$
- $\mathbf{A}$ is an $p \times q$ matrix that controls the structure and strength of confounding
- $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\gamma} \in \mathbb{R}^q$ are the effects of features and confounders, respectively

References

- Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K. and Zhao, L. (2019). Models as approximations i: Consequences illustrated with linear regression. Statistical Science, 34 523–544.
- Ćevid, D., Bühlmann, P. and Meinshausen, N. (2020). Spectral deconfounding via perturbed sparse linear models. Journal of Machine Learning Research, 21 1–41.
- Chernozhukov, V., Hansen, C. and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. The Annals of Statistics, 45 39 – 76.
- Jia, J. and Rohe, K. (2015). Preconditioning the Lasso for sign consistency. Electronic Journal of Statistics, 9 1150 – 1172.

Methods

When the confounding effect $\mathbf{Z}\boldsymbol{\gamma}$ is unobserved, the two methods below adjust for it from different perspectives.

PC-LASSO: PC-LASSO makes adjustments in the mean structure by including principal components (PCs) derived from the observed features $\mathbf{X}$, in the hope that the unobserved confounding may be captured by the leading PCs. $\widehat{\boldsymbol{\beta}}$ is obtained by minimizing

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{C}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\mathbf{C}$ is an $n \times k$ matrix containing the first k PCs. $\boldsymbol{\alpha} \in \mathbb{R}^k$ is not penalized.

PLMM: Instead of adjusting the mean structure, PLMM carries out its adjustment through the variance, treating the unobserved confounding as a random effect with mean zero and $\mathbb{V}(\mathbf{Z}\boldsymbol{\gamma}) = \sigma_s^2\mathbf{K}$. Then $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma_s^2\mathbf{K} + \sigma_e^2\mathbf{I}_n$. Pre-multiplying $\mathbf{X}$ and $\mathbf{y}$ by $\boldsymbol{\Sigma}^{-1/2}$ so that $\boldsymbol{\beta}$ can be estimated by minimizing

$$\frac{1}{2n}\|\boldsymbol{\Sigma}^{-1/2}\mathbf{y} - \boldsymbol{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

$\mathbf{K}$ can be estimated using $\frac{1}{p}\mathbf{X}\mathbf{X}^\top$ after standardizing $\mathbf{X}$.

Decomposition of Confounding

When $\mathbf{Z}$ is unobserved, not including $\mathbf{Z}$ in the linear model results in model misspecification. The best population linear approximation to y is obtained by only projecting onto $\mathbf{X}$, so that

$$\tilde{\boldsymbol{\beta}} = \mathbb{E}(\mathbf{x}\mathbf{x}^\top)^{-1}\mathbb{E}\left(\mathbf{x}(\mathbf{x}^\top\boldsymbol{\beta} + \mathbf{z}^\top\boldsymbol{\gamma})\right) = \boldsymbol{\beta} + \boldsymbol{\tau},$$

where $\boldsymbol{\tau} = \mathbb{E}(\mathbf{x}\mathbf{x}^\top)^{-1}\mathbb{E}(\mathbf{x}\mathbf{z}^\top)\boldsymbol{\gamma}$.

The least squares estimator $\widehat{\boldsymbol{\beta}}$ will converge to $\boldsymbol{\beta} + \boldsymbol{\tau}$, not $\boldsymbol{\beta}$. $\boldsymbol{\tau}$ is the **bias** that is introduced by hidden confounding. It can also be interpreted as the extent of confounding effect $\mathbf{Z}\boldsymbol{\gamma}$ that can be projected onto $\mathbf{X}$. The part that cannot be projected onto $\mathbf{X}$ enters the model as **noise**: $\psi = \mathbf{z}^\top\boldsymbol{\gamma} - \mathbf{x}^\top\boldsymbol{\tau}$.

Partitioning the unobserved confounding effects into **bias and noise** allows us to compute

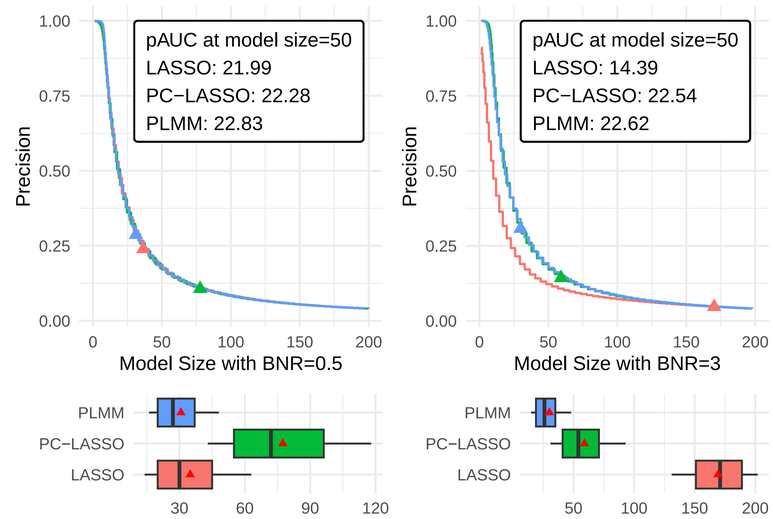
$$\text{Bias-to-Noise Ratio (BNR)} = \frac{\boldsymbol{\tau}^\top\mathbb{V}(\mathbf{x})\boldsymbol{\tau}}{\mathbb{V}(\psi \mid \boldsymbol{\tau}) + 1},$$

$$\text{Signal-to-Noise Ratio (SNR)} = \frac{\boldsymbol{\beta}^\top\mathbb{V}(\mathbf{x})\boldsymbol{\beta}}{\mathbb{V}(\psi \mid \boldsymbol{\tau}) + 1}.$$

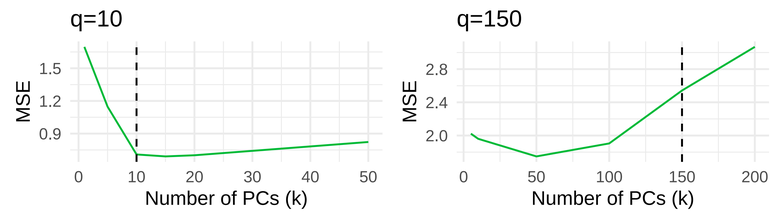
- Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11 733–739.
- Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. Nature Methods, 8 833–835.
- Wang, Y. and Blei, D. M. (2019). The blessings of multiple causes. Journal of the American Statistical Association, 114 1574–1596.

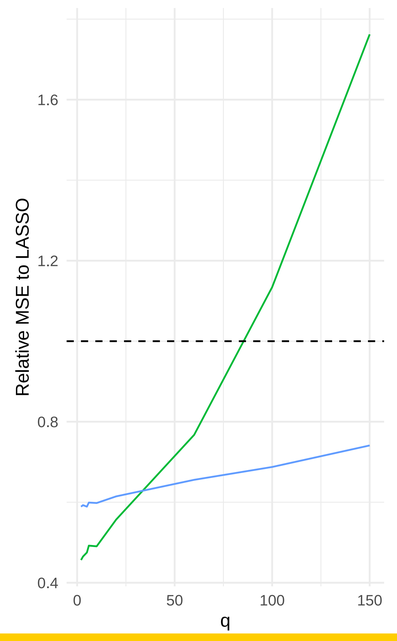
Results

Keep SNR $= 1.5, \mathbb{V}(\psi \mid \boldsymbol{\tau}) = 1$, changing BNR.



BNR $= 1.5$, SNR $= 1.5, \mathbb{V}(\psi \mid \boldsymbol{\tau}) = 1$.





Semi-synthetic data with real $\mathbf{X}$, $\mathbf{Z}$

g	0	1	2
Relative MSE to LASSO			
PC-LASSO	1.99	1.32	1.16
PLMM	1.38	0.92	0.84
pAUC at Model Size 50			
LASSO	12.78	10.46	6.95
PC-LASSO	27.67	26.20	23.32
PLMM	35.56	34.12	29.58
Average Model Size			
LASSO	23.70	49.50	60.12
PC-LASSO	68.36	84.80	88.36
PLMM	25.80	23.80	23.90