Supplemental Appendix: Sparse Estimation and Uncertainty with Application to Subgroup Analysis

A Proof of Relative Efficiency of Oracle Estimator and OLS.

Proof: Denote as X_S the submatrix of X for which $\beta_k \neq 0$ and the Gram matrix for X as

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} X_i^\top X_i = \Sigma_X \tag{1}$$

and in block-partition form

$$\Sigma_X = \begin{pmatrix} \Sigma_{SS} & \Sigma_{SS^\complement} \\ \Sigma_{SS^\complement}^\top & \Sigma_{S^\complement S^\complement} \end{pmatrix} \tag{2}$$

Σ_X is invertible, since the least squares estimate exists and is unique. Since Σ_X is invertible, every square submatrix of Σ_X is also invertible.

The asymptotic relative efficiency of the least squares estimate and Oracle estimate is then

$$\lim_{N\to\infty} \frac{\frac{\sigma^2}{N}\mathrm{Tr}\left\{\Sigma^{-1}\right\}}{\frac{\sigma^2}{N}\mathrm{Tr}\left\{\Sigma_{SS}^{-1}\right\}} = \frac{\mathrm{Tr}\left\{\Sigma^{-1}\right\}}{\mathrm{Tr}\left\{\Sigma_{SS}^{-1}\right\}} \tag{3}$$

By the block inverse partition formula,

$$\mathrm{Tr}\left(\Sigma_X^{-1}\right) = \mathrm{Tr}\left\{ \begin{pmatrix} \Sigma_{SS} & \Sigma_{SS^\complement} \\ \Sigma_{SS^\complement}^\top & \Sigma_{S^\complement S^\complement} \end{pmatrix}^{-1} \right\} \tag{4}$$

$$= \mathrm{Tr}\left\{ \left(\Sigma_{SS} - \Sigma_{SS^\complement}\Sigma_{S^\complement S^\complement}^{-1}\Sigma_{SS^\complement}^\top\right)^{-1} \right\} + \mathrm{Tr}\left\{ \left(\Sigma_{S^\complement S^\complement} - \Sigma_{SS^\complement}^\top\Sigma_{S^\complement S^\complement}^{-1}\Sigma_{SS^\complement}\right)^{-1} \right\} \tag{5}$$

Consider the first summand inside the parentheses on the r.h.s. and apply Morrison-Woodbury-Sherman

$$\left(\Sigma_{SS} - \Sigma_{SS^\complement}\Sigma_{S^\complement S^\complement}^{-1}\Sigma_{SS^\complement}^\top\right)^{-1} = \Sigma_{SS}^{-1} + \Sigma_{SS}^{-1}\Sigma_{SS^\complement}\left(\Sigma_{S^\complement S^\complement} - \Sigma_{SS^\complement}^\top\Sigma_{SS}^{-1}\Sigma_{SS^\complement}\right)^{-1}\Sigma_{SS^\complement}^\top\Sigma_{SS}^{-1} \tag{6}$$

By Cauchy-Schwarz, the term $\Sigma_{S^\complement S^\complement} - \Sigma_{SS^\complement}^\top\Sigma_{SS}^{-1}$ is positive semi-definite, see e.g. Tripathi (1999, esp. the last line of the proof of Theorem 1.1.). By symmetry, we get an analogous result for the second summand in side the trace operator,

This gives

$$\begin{aligned}\operatorname{Tr}\left(\Sigma_X^{-1}\right) = \operatorname{Tr}\left(\Sigma_{SS}^{-1}\right) + \operatorname{Tr}\left(\Sigma_{S^{\complement}S^{\complement}}^{-1}\right) + & \qquad (7)\\ \operatorname{Tr}\left(\Sigma_{SS}^{-1}\Sigma_{SS^{\complement}}\left(\Sigma_{S^{\complement}S^{\complement}} - \Sigma_{SS^{\complement}}^{\top}\Sigma_{SS}^{-1}\Sigma_{SS^{\complement}}\right)^{-1}\Sigma_{SS^{\complement}}^{\top}\Sigma_{SS}^{-1}\right) + & \\ \operatorname{Tr}\left(\Sigma_{S^{\complement}S^{\complement}}^{-1}\Sigma_{SS^{\complement}}^{\top}\left(\Sigma_{SS} - \Sigma_{SS^{\complement}}\Sigma_{S^{\complement}S^{\complement}}^{-1}\Sigma_{SS^{\complement}}^{\top}\right)^{-1}\Sigma_{SS^{\complement}}\Sigma_{S^{\complement}S^{\complement}}^{-1}\right) & \\ \geq \operatorname{Tr}\left(\Sigma_{SS}^{-1}\right) & \qquad (8)\end{aligned}$$

and therefore an estimator with the Oracle Property is asymptotically more efficient than least squares.

To establish when equality holds, if $X = X_S$, then clearly the asymptotic relative efficiency is 1. For only if, the inequality above is an equality only when $\operatorname{Tr}\left(\Sigma_{S^{\complement}S^{\complement}}^{-1}\right) = 0$, which is not possible unless $X = X_S$.

B Preliminaries

We offer three sets of preliminary results . First, we show that the weights, $\widehat{w}_k$, and magnitude of $|\widehat{\beta}_k|$ are inversely related. Second, we formally differentiate between "large" and "small" estimates. This will help us derive bounds on $\widehat{w}_k$. Third, we provide a bound on $\widehat{\lambda}$. Note that we refer to the k^{th} order statistic of vector a as $a_{(k)}$, where $a_{(1)}$ is the smallest element of a.

B.1 Inverse relationship between weights and effect size.

Proposition 1

$$\frac{\partial \widehat{w}_k}{\partial |\widehat{\beta}_k|} = -\widehat{\lambda}\sqrt{\frac{\widehat{1}}{\sigma^2}}\operatorname{Var}(w_k|\cdot) < 0. \qquad (9)$$

Derivation: *The weights are calculated as*

$$\widehat{w}_k = \mathbb{E}(w_k|\cdot) = \frac{\int_{w=0}^{\infty} w e^{-w^{\widehat{\gamma}} - \widehat{\lambda} w\sqrt{\frac{\widehat{1}}{\sigma^2}}|\widehat{\beta}_k|}\, dw}{\int_{w=0}^{\infty} e^{-w^{\widehat{\gamma}} - \widehat{\lambda} w\sqrt{\frac{\widehat{1}}{\sigma^2}}|\widehat{\beta}_k|}\, dw}. \qquad (10)$$

Denote as $A = e^{-w\widehat{\gamma} - \widehat{\lambda} w \sqrt{\frac{\widehat{1}}{\sigma^2}} |\widehat{\beta}_k|}$. *Then,*

$$\frac{\partial \widehat{w}_k}{\partial |\widehat{\beta}_k|} = \frac{-\int_{w=0}^{\infty} A dw \times \int_{w=0}^{\infty} w^2 \widehat{\lambda} \sqrt{\frac{\widehat{1}}{\sigma^2}} A dw + \int_{w=0}^{\infty} w A dw \int_{w=0}^{\infty} w \widehat{\lambda} \sqrt{\frac{\widehat{1}}{\sigma^2}} A dw}{\left(\int_{w=0}^{\infty} A dw\right)^2} \tag{11}$$

$$= -\widehat{\lambda} \sqrt{\frac{\widehat{1}}{\sigma^2}} \left\{ \frac{\int_{w=0}^{\infty} w^2 A dw}{\int_{w=0}^{\infty} A dw} - \left(\frac{\int_{w=0}^{\infty} w A dw}{\int_{w=0}^{\infty} A dw} \right)^2 \right\} \tag{12}$$

$$= -\widehat{\lambda} \sqrt{\frac{\widehat{1}}{\sigma^2}} \mathrm{Var}(w_k|\cdot) \tag{13}$$

where moving the derivative under the integral in the first line is allowed by the monotone convergence theorem.

This result allows us to associate the largest weight, $\widehat{w}_{(K)}$ with the smallest estimate, $\widehat{\beta}_{(1)}$, the second largest weight with the second smallest estimate, and so on. In general, weight $\widehat{w}_{(k)}$ is associated with $|\widehat{\beta}|_{(K-k+1)}$

B.2 Separating large and small weights and effect estimates.

We next distinguish between weights near zero from weights close to the maximal value $\widehat{\overline{\gamma}}$. This is our equivalent of either assuming the estimates are "well-separated" (Belloni and Chernozhukov, 2013), or separating "relevant" from "irrelevant" effects (Buhlmann and van de Geer, 2013). The key difference is that these authors separate large and small "true" effects, whereas we separate large and small estimated effects. As is common in the literature, our bounds will be more informative the better we can distinguish between zero- and non-zero effect estimates.

We separate the weights into two groups. In the kernel for $\Pr(w_k|\cdot)$, the numerator in Equation 10, is approximately exponential for large $|\beta_k|$, small w_k, and is approximately constant for $|\beta_k| \approx 0$, w_k large. Define as

$$p_k(C_1, C_2) = \max \left\{ \Pr\left(\widehat{w}_k > \frac{C_1 \log(\widehat{S})}{\lambda \widehat{\sigma} |\widehat{\beta}_k|} \right), \Pr\left(\widehat{w}_k < C_2 \widehat{\overline{\gamma}} \right) \right\};\ C_1 > 0, 0 < C_2 < 1 \tag{14}$$

where the first inequality allows us to bound with some high probability small weights from above and the second, larger weights from below. We use this distinction to differentiate between weights tending to zero (the lefthand set) and those tending to the maximum (the righthand set).

$$\widehat{S} = \left\{ k : \Pr\left(\widehat{w}_k > \frac{C_1 \log\left(\left|\widehat{S}\right|\right)}{\lambda \widehat{\sigma} |\widehat{\beta}_k|} \right) < \Pr\left(\widehat{w}_k < C_2 \widehat{\overline{\gamma}} \right) \right\}. \tag{15}$$

The $\log(|\widehat{S}|)$ term on the left comes from using the union bound applied to $\{p_k\}_{k=1}^K$ and a subexponential (rather than subgaussian) bound applied to each value p_k, as the kernel is approximately exponential in this range. Define

$$\Pr(\max(p_k) > C_3 \log(K)) = p_w(C_1, C_2, C_3). \tag{16}$$

such that, with probablity at least $p_w(C_1, C_2, C_3)$, the weights can be bounded by one of the bounds above, i.e. is either "small" or "large."

Lastly, denote as $\underline{C}_1$ the value that satisfies

$$\Pr\left(\widehat{w}_k > \frac{C_1 \log(|\widehat{S}|)}{\widehat{\lambda}\widehat{\sigma}|\widehat{\beta}_k|}\right) = \Pr\left(\widehat{w}_k \leq \frac{\underline{C}_1}{\widehat{\lambda}\widehat{\sigma}|\widehat{\beta}_k| \log(|\widehat{S}|)}\right) \tag{17}$$

which will give us a lower bound on all $\widehat{w}_k$ with probability at least $p_w(C_1, C_2, C_3)$.

B.3 Bounding the tuning parameter $\widehat{\lambda}$.

Given the results above, we can bound $\widehat{\lambda}$. For the Oracle results below, we need to bound $\widehat{\lambda}$ from below, though we note that a similar bound of the same order of N, K can be found using the strategy below.

As $\lambda^2|\cdot \sim \Gamma(\sqrt{N}K, \frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2 + \rho)$, this gives

$$\widehat{\lambda^2} = \frac{\sqrt{N}K}{\frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2 + \rho}. \tag{18}$$

Change of variables gives $\lambda|\cdot \sim$ generalizedGamma $\left(2 \times (\frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2 + \rho)^{-1/2}, 2\sqrt{N}K, 2\right)$, which gives the estimate

$$\widehat{\lambda} = \frac{\widetilde{\Gamma}(\sqrt{N}K + 1/2)/\widetilde{\Gamma}(\sqrt{N}K)}{\sqrt{\frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2 + \rho}} \tag{19}$$

with $\widetilde{\Gamma}()$ the Gamma function. Note $\widehat{\lambda^2} \geq (\widehat{\lambda})^2$ and if $\sqrt{N}K > 1$, then $\Gamma(3/2)^2(\widehat{\lambda})^2 = \frac{4}{\pi}(\widehat{\lambda})^2 > \widehat{\lambda^2}$. Lastly,

$$1/\tau_k^2|\cdot \sim \text{InvGaussian}\left(\lambda w_k \sigma/|\beta_k|, w_k^2\lambda^2\right) \Rightarrow \tag{20}$$

$$\sum_{k=1}^K \widehat{\tau}_k^2 = \sum_{k=1}^K \frac{|\widehat{\beta}_k|}{\widehat{\lambda}\widehat{w}_k\widehat{\sigma}} + \frac{1}{\widehat{\lambda^2}\widehat{w}_k^2} \tag{21}$$

and we use the bound

$$\sum_{k=1}^{K}\widehat{\tau}_k^2 \leq \frac{|\widehat{S}|\times|\widehat{\beta}_{(K)}|}{\widehat{\lambda}\widehat{w}_{(1)}\widehat{\sigma}} + \frac{|\widehat{S}|}{(\widehat{\lambda})^2\widehat{w}_{(1)}^2} + \frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{\left(K-|\widehat{S}|-1\right)}}{\widehat{\lambda}\widehat{w}_{\left(K-|\widehat{S}|-1\right)}\widehat{\sigma}} + \frac{\left(K-|\widehat{S}|\right)}{(\widehat{\lambda})^2\widehat{w}^2_{\left(K-|\widehat{S}|-1\right)}} \tag{22}$$

$$\leq \frac{|\widehat{S}|\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(\widehat{S})} + \frac{|\widehat{S}|\widehat{\sigma}^2\widehat{\beta}_k^2}{\underline{C}_1^2\log(\widehat{S})^2} + \frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{\left(K-|\widehat{S}|-1\right)}}{\widehat{\lambda}C_2\widehat{\overline{\gamma}}\widehat{\sigma}} + \frac{\left(K-|\widehat{S}|\right)}{(\widehat{\lambda})^2C_2^2\widehat{\overline{\gamma}}^2} \tag{23}$$

$$= \frac{|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1\log(\widehat{S})+\widehat{\sigma}^2\right)}{\underline{C}_1^2\log(\widehat{S})^2} + \frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{\left(K-|\widehat{S}|-1\right)}}{\widehat{\lambda}C_2\widehat{\overline{\gamma}}\widehat{\sigma}} + \frac{\left(K-|\widehat{S}|\right)}{(\widehat{\lambda})^2C_2^2\widehat{\overline{\gamma}}^2} \tag{24}$$

The first line follows from the inverse relationship between $|\widehat{\beta}_k|$ and $\widehat{w}_k$; the second comes from the lower bounds on $\widehat{w}_k$ in $\widehat{S}$ and $\widehat{S}^{\complement}$. The third line is just simplifying.

Combining inequalities gives

$$\frac{4}{\pi}(\widehat{\lambda})^2 \geq \widehat{\lambda}^2 = \frac{\sqrt{N}K}{\frac{1}{2}\sum_{k=1}^{K}\widehat{\tau}_k^2+\rho} \tag{25}$$

$$\Rightarrow (\widehat{\lambda})^2 \geq \frac{\pi}{4}\times\frac{\sqrt{N}K}{\frac{|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1\log(\widehat{S})+\widehat{\sigma}^2\right)}{2\underline{C}_1^2\log(\widehat{S})^2} + \frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{(K-|\widehat{S}|-1)}}{2\widehat{\lambda}C_2\widehat{\overline{\gamma}}\widehat{\sigma}} + \frac{\left(K-|\widehat{S}|\right)}{2(\widehat{\lambda})^2C_2^2\widehat{\overline{\gamma}}^2}+\rho} \tag{26}$$

$$\Rightarrow \widehat{\lambda} \geq \frac{\pi}{4}\times\frac{\sqrt{N}K}{\frac{\widehat{\lambda}|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1\log(\widehat{S})+\widehat{\sigma}^2\right)}{2\underline{C}_1^2\log(\widehat{S})^2} + \frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{(K-|\widehat{S}|-1)}}{2C_2\widehat{\overline{\gamma}}\widehat{\sigma}} + \frac{\left(K-|\widehat{S}|\right)}{2\widehat{\lambda}C_2^2\widehat{\overline{\gamma}}^2}+\rho\widehat{\lambda}} \tag{27}$$

where the second line comes from substituting from Inequality 24 and the third from multiplying both sides by $1/\widehat{\lambda}$. Cross-multiplying gives a quadratic equation in $\widehat{\lambda}$ of the form $\widetilde{a}(\widehat{\lambda})^2+\widetilde{b}\widehat{\lambda}+\widetilde{c}>0$ where[1]

$$\widetilde{a} = \frac{|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1\log(\widehat{S})+\widehat{\sigma}^2\right)}{2\underline{C}_1^2\log(\widehat{S})^2}+\rho \tag{28}$$

$$\widetilde{b} = \frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{\left(K-|\widehat{S}|-1\right)}}{2C_2\widehat{\overline{\gamma}}\widehat{\sigma}} \tag{29}$$

$$\widetilde{c} = -\left(\frac{\pi}{4}\sqrt{N}K - \frac{\left(K-|\widehat{S}|\right)}{2C_2^2\widehat{\overline{\gamma}}^2}\right). \tag{30}$$

[1]We use the convention $0\log 0 = 0$

The quadratic equation gives

$$\widehat{\lambda} \geq \frac{-\frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{(K-|\widehat{S}|-1)}}{2C_2\widehat{\widehat{\gamma}}\widehat{\sigma}} + \sqrt{\left\{\frac{\left(K-|\widehat{S}|\right)|\widehat{\beta}|_{(K-|\widehat{S}|-1)}}{2C_2\widehat{\widehat{\gamma}}\widehat{\sigma}}\right\}^2 + 4\left\{\frac{\widehat{\lambda}|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1\log(\widehat{S})+\widehat{\sigma}^2\right)}{2\underline{C}_1^2\log(\widehat{S})^2} + \rho\right\} \times \left\{\frac{\pi}{4}\sqrt{N}K - \frac{\left(K-|\widehat{S}|\right)}{2C_2^2\widehat{\widehat{\gamma}}^2}\right\}}}{2\frac{|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1\log(\widehat{S})+\widehat{\sigma}^2\right)}{2\underline{C}_1^2\log(\widehat{S})^2} + 2\rho} \tag{31}$$

which, for growing N and K, is of order $N^{1/4}K^{1/2}$ by the bound in 31.

C Variance Estimation

We sample from the approximate sampling distribution of the the LASSOplus estimator at each Gibbs update:

$$\beta_k \mathbf{1}\left(\left|\widehat{\beta}_k^{sp}\right| \geq \frac{\lambda w_k \sigma_{sp}}{N-1}\right) \tag{32}$$

$$\approx \beta_k \Phi\left\{\left|\left|\widehat{\beta}_k^{sp}/\widehat{\sigma}_{ls}\right| - \frac{\lambda w_k \sigma_{sp}}{\widehat{\sigma}_{ls}\times(N-1)}\right|\right\} \tag{33}$$

$$= \beta_k \Phi\left\{\sqrt{N-1}\left|\left|\widehat{\beta}_k^{sp}/\sigma\right| - \frac{\lambda w_k \sigma_{sp}}{\sigma\times(N-1)}\right|\right\} \tag{34}$$

$$= g\left(\beta_k, \widehat{\beta}_k^{sp}, \sigma, \sigma_{sp}, \lambda, w_k\right) \tag{35}$$

where $\Phi(a)$ is the cumulative distribution for a standard normal random variable and we approximate the standard error of the least squares coefficient as $\widehat{\sigma}_{ls} \approx \sigma/\sqrt{N-1}$. Define

$$z_k = \sqrt{N-1}\left|\left|\widehat{\beta}_k^{sp}/\sigma\right| - \frac{\lambda w_k \sigma_{sp}}{\sigma\times(N-1)}\right| \tag{36}$$

$$p_k = \Phi\left\{z_k\right\} \tag{37}$$

Define the vector of partial derivatives

$$\nabla g\left(\beta_k, \widehat{\beta}_k^{sp}, \sigma, \sigma_{sp}, \lambda, w_k\right) = \left[\frac{\partial g(\cdot)}{\partial \beta_k}, \frac{\partial g(\cdot)}{\partial \widehat{\beta}_k^{sp}}, \frac{\partial g(\cdot)}{\partial \sigma}, \frac{\partial g(\cdot)}{\partial \sigma_{sp}}, \frac{\partial g(\cdot)}{\partial \lambda}, \frac{\partial g(\cdot)}{\partial w_k}\right]^\top \tag{38}$$

$$= \begin{bmatrix} p_k \\ \beta_k \times \phi(z_k) \times \sqrt{N-1}/\sigma\, \mathrm{sgn}(\widehat{\beta}_k^{sp}) \\ \beta_k \times \phi(z_k) \times \sqrt{N-1}\left(-\frac{\left|\widehat{\beta}_k^{sp}\right|}{\sigma^2} + \frac{\lambda w_k \sigma_{sp}}{\sigma^2\times(N-1)}\right) \\ \beta_k \times \phi(z_k) \times \sqrt{N-1} \times \frac{\lambda w_k}{\sigma\times(N-1)} \\ \beta_k \times \phi(z_k) \times \sqrt{N-1} \times \frac{w_k \sigma_{sp}}{\sigma\times(N-1)} \\ \beta_k \times \phi(z_k) \times \sqrt{N-1} \times \frac{\lambda \sigma_{sp}}{\sigma\times(N-1)} \end{bmatrix} \tag{39}$$

and the 6×6 matrix

$$V = \text{diag}\left[\text{Var}(\beta_k), \text{Var}(\widehat{\beta}_k^{sp}), \text{Var}(\sigma), \text{Var}(\sigma_{sp}), \text{Var}(\lambda), \text{Var}(w_k)\right] \tag{40}$$

where we are assuming zero covariance between elements. All elements of V are calculated analytically from the variance of the conditional pseudoposterior densities except for $\text{Var}(w_k)$ which is calculated from the approximate density used in the griddy Gibbs sampler. Our approximate variance is then

$$\widehat{\sigma}_j^2 = \nabla g^\top(\cdot) V \nabla g(\cdot) \tag{41}$$

.

D EM Updates for LASSOplus-EM

For our EM implementation, we treat in $\beta^{plus-EM}$ and σ^2 as parameters and the remaining parameters as "missing," i.e. to be estimated. As we have already calculated the conditional posterior densities for all parameters, the EM updates is straightforward.

Standardize Y and all columns of X to be mean-zero, sample variance one. Initialize $\forall k : \widehat{\beta}_k \leftarrow u_k$ with $u_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.01)$; $\widehat{\sigma}^2 \leftarrow ||Y - X\widehat{\beta}||_2^2/N$; $\widehat{\lambda} \leftarrow 1$; $\widehat{w}_k \leftarrow 1$.

At each given step, the most current updates from the previous steps are used. To convergence,

- E-steps
  1. $\forall k : \widehat{(1/\tau_k^2)} \leftarrow \widehat{\lambda}\widehat{w}_k\widehat{\sigma}/|\widehat{\beta}_k|$; $\widehat{\tau_k^2} \leftarrow |\widehat{\beta}_k|/(\widehat{\lambda}\widehat{w}_k\widehat{\sigma}) + 1/(\widehat{\lambda}^2\widehat{w}_k^2)$
  2. $\widehat{\lambda} \leftarrow \frac{\widetilde{\Gamma}(\sqrt{N}K+1/2)/\widetilde{\Gamma}(\sqrt{N}K)}{\sqrt{\frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2 + \rho}}$; $\widehat{\lambda^2} \leftarrow \frac{\sqrt{N}K}{\frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2 + \rho}$ with $\widetilde{\Gamma}()$ the gamma function.
  3. $\forall k$: update $\widehat{w}_k$ via numerical integration using kernel $\Pr(w_k|\cdot) \propto e^{-w^{\widehat{\gamma}} - \widehat{\lambda} w \sqrt{\widehat{\frac{1}{\sigma^2}}}|\widehat{\beta}_k|}$
  4. Update $\widehat{\gamma}$ via numerical integration using kernel $\Pr(\gamma|\cdot) \propto \gamma e^{-\sum_{k=1}^K \widehat{w}_k^{\gamma} - \gamma}$
- M-Steps
- $\widehat{\sigma^2} \leftarrow \frac{\sum_{i=1}^N (Y_i - X_i^\top \widehat{\beta})^2 + \sum_{k=1}^K (\widehat{\beta}_k)^2 \times \widehat{\frac{1}{\tau_k^2}}}{N+K}$; $\widehat{\frac{1}{\sigma^2}} \leftarrow \frac{N+K-2}{\sum_{i=1}^N (Y_i - X_i^\top \widehat{\beta})^2 + \sum_{k=1}^K (\widehat{\beta}_k)^2 \times \widehat{\frac{1}{\tau_k^2}}}$
- Conditional M-steps: $\forall k : \widehat{\beta^k} \leftarrow \frac{\sum_{i=1}^N X_{ik}(Y_i - \sum_{k' \neq k} X_{ik'}\widehat{\beta}_{k'})}{(N-1) + \widehat{\frac{1}{\tau_k^2}}}$ where it is understood that at update $\widetilde{k}$, updated estimates of $\left\{\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_{\widetilde{k}-1}\right\}$ are used.

LASSOplus updates:

- $\widehat{\sigma}^2_{sp} \leftarrow\leftarrow \frac{\sum_{i=1}^N (Y_i - X_i^\top \widehat{\beta})^2 + \sum_{k=1}^K (\widehat{\beta}_k)^2 \times \widehat{\frac{1}{\tau_k^2}}}{\sqrt{N}+K}$
- $\widehat{\beta}_k^{plus} \leftarrow \widehat{\beta}_k \mathbf{1}\left(\left| \sum_{i=1}^N X_{ik}(Y_i - \sum_{k' \neq k} X_{ik'} \widehat{\beta}_{k'}) \right| > \widehat{\lambda} \widehat{w}_k \sqrt{\widehat{\sigma}^2_{sp}} \right)$

E Independence between Adjusted Higher-Order Terms and Lower-Order Terms

We prove first that, under the residualized construction, the least squares coefficient on the a higher-order interaction term is uncorrelated with the coefficients on lower-order terms. By this means, the effect of the higher-order term does not vary with its lower-order components, and hence can be interpreted on its own. We then extend the result to the conditional pseudoposterior density of the estimates.

Denote the $N \times 1$ vector of outcomes Y, $N \times L$ matrix of lower-order terms $\mathbf{X}_{lower}$ and vector of mean-zero, equivariant errors ϵ. Define as $X_{inter} = [X_{inter}]_i = \prod_{1 \leq l' \leq L} x_{il'}$, the elementwise product of the lower-order terms. Assume $[\mathbf{X}_{lower} | X_{inter}]$ is full rank. Using parameters $\{\beta_0, \vec{\beta}_l, \beta_{inter}\}$, define the model, with $\vec{\beta}_l$ an $L \times 1$ vector and the others scalars, as

$$Y = \mathbf{X}_{lower} \vec{\beta}_l + X_{inter} \beta_{inter} + \epsilon. \tag{42}$$

Define the matrices

$$\mathbf{X} = [\mathbf{X}_{lower} | X_{inter}] \tag{43}$$

$$\mathbf{M}_{lower} = \mathbf{I}_L - (\mathbf{X}_{lower})(\mathbf{X}_{lower}^\top \mathbf{X}_{lower})^{-1} \mathbf{X}_{lower}^\top \tag{44}$$

$$\widetilde{X}_{inter} = \mathbf{M}_{lower} X_{inter} \tag{45}$$

$$\mathbf{X}_{adjust} = [\mathbf{X}_{lower} | \widetilde{X}_{inter}] \tag{46}$$

The vector $\widetilde{X}_{inter}$ is the residualized interaction term described in the text, giving parameterization

$$Y = \mathbf{X}_{lower} \vec{\widetilde{\beta}}_l + \widetilde{X}_{inter} \widetilde{\beta}_{inter} + \epsilon \tag{47}$$

where the error vector ϵ, stays unchanged since the two parameterizations differ only by a linear transformation of the covariates.

The covariance of the least squares estimates in the first parameterization is proportional to the inverse of the cross product of the design matrix. Using the block-partition matrix formula gives

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix} (\mathbf{X}_{lower}^\top\mathbf{X}_{lower}) & \mathbf{X}_{lower}^\top X_{inter} \\ X_{inter}^\top\mathbf{X}_{lower} & X_{inter}^\top X_{inter} \end{bmatrix}^{-1} \tag{48}$$

$$= \begin{bmatrix} \left(\mathbf{X}_{lower}^\top\mathbf{X}_{lower} - \frac{1}{c_0}\mathbf{X}_{lower}^\top X_{inter}X_{inter}^\top\mathbf{X}_{lower}\right)^{-1} & -\frac{1}{c_0}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1}X_{lower}^\top X_{inter} \\ -\frac{1}{c_0}X_{inter}^\top\mathbf{X}_{lower}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1} & \frac{1}{c_0} \end{bmatrix} \tag{49}$$

with the constant $c_0 = X_{inter}^\top X_{inter} - X_{inter}^\top\mathbf{X}_{lower}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top X_{inter}$. This implies

$$\text{Cov}(\widehat{\beta}_{inter}, \widehat{\beta}_k) \propto -\left[\frac{1}{c_0}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top X_{inter}\right]_j \text{ for } j \in \{1, 2, \ldots, L\} \tag{50}$$

In general, this covariance will not be zero, suggesting that under the normal parameterization the effect of the interaction term varies with movements in its lower order terms. Repeating the same exercise with a model parameterized in terms of $\widetilde{X}_{inter}$ gives

$$\text{Cov}\left(\widehat{\widetilde{\beta}}_{inter}, \widehat{\widetilde{\beta}}_k\right) \propto -\left[\frac{1}{c_0}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top\widetilde{X}_{inter}\right]_j \tag{51}$$

$$= -\left[\frac{1}{c_0}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top\mathbf{M}_{lower}X_{inter}\right]_j \tag{52}$$

$$= 0 \text{ for } j \in \{1, 2, \ldots, L\} \tag{53}$$

Therefore, under the parameterization with residualized interaction terms, the marginal effect of each interaction term is uncorrelated with that of its lower order terms. To extend to the multivariate case, assume the full design matrix of all effects is full-rank, and all other effects have been partialed out. The case of $K > N$ requires an assumption similar to the restricted eigenvalue assumption (Bickel, Ritov and Tsybakov, 2009), that all submatrices of size $L+2$ are full rank and all components of the submatrices not in $\mathbf{X}$ are linearly independent of $\mathbf{X}$. Partialing out out with respect to the other covariates in either case leaves the results unchanged.

Next, we show the result holds for the conditional pseudoposterior density under a conditional independent normal prior, as with the augmented LASSOplus. Assume $[\vec{\beta_l}^\top, \beta_{inter}]^\top \sim \mathcal{N}(0_{L+1}, D)$ with D an $(L+1)\times(L+1)$ diagonal matrix with positive entries along the diagonal. In this case, the conditional posterior of $[\vec{\beta_l}^\top, \beta_{inter}]^\top$ under a normal likelihood takes the form

$$\Pr([\vec{\beta_l}^\top, \beta_{inter}]^\top|\cdot) \sim \mathcal{N}(A^{-1}\mathbf{X}^\top Y, \sigma^2 A^{-1}) \tag{54}$$

with $A = \mathbf{X}^\top \mathbf{X} + D$. Carrying through the same derivation as above gives the posterior covariance between β_{L+1}, the parameter on the interaction term, and β_k, $1 \leq j \leq L$, as

$$A^{-1}_{j,L+1} \propto -\left[\frac{1}{c_0'}(\mathbf{X}^\top_{lower}\mathbf{X}_{lower} + D_{1:L,1:L})^{-1}\mathbf{X}^\top_{lower}X_{inter}\right]_j \text{ for } j \in \{1, 2, \ldots, L\} \tag{55}$$

which will not be 0, in general. In this case, the constant $c_0' = (X_{inter} + D_{L+1,L+1})^\top (X_{inter} + D_{L+1,L+1}) - X^\top_{inter}\mathbf{X}_{lower}(\mathbf{X}^\top_{lower}\mathbf{X}_{lower} + D_{1:L,1:L})^{-1}\mathbf{X}^\top_{lower}X_{inter}$.

Considering the residualized interaction term instead of the standard term gives

$$A^{-1}_{j,L+1} \propto -\left[\frac{1}{c_0'}(\mathbf{X}^\top_{lower}\mathbf{X}_{lower} + D_{1:L,1:L})^{-1}\mathbf{X}^\top_{lower}\widetilde{X}_{inter}\right]_j = 0 \text{ for } j \in \{1, 2, \ldots, L\} \tag{56}$$

F Alternative Estimators

For the LASSO and adaptive LASSO, we found the BIC statistic of Wang and Leng (2007) performed poorly when $K > N$, sometimes including dozens of false positives. We instead use a standard BIC statistic where we take the degrees of freedom as the number of non-zero coefficients Zou, Hastie and Tibshirani (2007).

In terms of uncertainty estimates, we implement the approximate confidence intervals for the LASSOplus. We use the posterior intervals for the horseshoe model. For the frequentist LASSO and adaptive LASSO, we implement the perturbation method of Minnier, Tian and Cai (2011). For $p \in \{1, 2, \ldots, P\}$ for some large P, the method requires fitting

$$\widehat{\beta}^{alasso,p}(\lambda|w_\cdot, g_\cdot) = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} g_i^p (Y_i - X_i^\top \beta)^2 + \lambda \sum_{k=1}^{K} w_k |\beta_k|; \tag{57}$$

$$w_k = 1/|\beta_k^0| \tag{58}$$

where the weights are $g_i^p \overset{\text{i.i.d.}}{\sim} \exp(1)$. For the LASSO, we simply take $w_k = 1$ for all k. Minnier, Tian and Cai (2011) prove that the set $\{\widehat{\beta}^{alasso,p}(\lambda|w_\cdot, g_\cdot)\}_{p=1}^P$ will achieve nominal coverage asymptotically, though the result does not hold for the LASSO. We fit the perturbation method to both for comparison. We found the perturbation method performs better than the parametric bootstrap suggested by Chatterjee and Lahiri (2011, 2013), so we do not present the results.

We next move on to the LASSO+OLS method of Belloni and Chernozhukov (2013), hereafter BC.The empirical process approach selects the tuning parameter in order to bound $2\max\left(\epsilon^\top X_{\cdot k}\right)$ with some high probability. BC note that, up to a scale parameter σ, the tuning parameter value can

be simulated quite easily, and they define $\Lambda(1-\alpha_{sig}|X)$ as the $1-\alpha_{sig}$ quantile of $2\max\left(\epsilon^\top X_{\cdot k}/\sigma_b\right)$ for $\mathbb{E}(\epsilon_i|X_i)=0$; $\text{Var}(\epsilon_i|X_i)=\sigma_b^2$ as approximated through a simulation.

Second stage variable selection. Tuning λ in order to satisfy the Oracle Inequality will generally over-select effects. The reason is that the LASSO induces bias in the coefficient estimaates, and that bias leaves a gap for irrelevant effects that are correlated with the relevant effects to be drawn in and selected. Several methods in the empirical process framework have used the Oracle Inequality-tuned LASSO to over-select covariates and then, in a second stage, select a subset of these.

One way to do so is simply thresholding the LASSO estimates, so

$$\widehat{\beta}^{thresh} = \widehat{\beta}^{L} \odot \mathbf{1}\left(\left|\widehat{\beta}^{L}\right| > \tau\right) \tag{59}$$

where the inequality and multiplication $\odot$ are taken elementwise. A second option is to take then rerun OLS on variables that survive the threshold. Define X_{thresh} as the submatrix of X corresponding with elements of $\widehat{\beta}^{thresh}(\tau)$ that are not zero. Then,

$$\widehat{\beta}^{thresh+OLS}(\tau) = (X_{thresh}^\top X_{thresh})^{-1} X_{thresh}^\top Y. \tag{60}$$

In the case X_{thresh} is rank-deficient, either ridge regression or partial least squares can be used (Liu and Yu, 2013). The post LASSO OLS estimator is then $\widehat{\beta}^{thresh+OLS}(0)$, which is simply OLS used on all selected LASSO covariates.

Belloni and Chernozhukov (2013) propose a different means of selecting a subset of relevant effects and eliminating the first-stage LASSO bias. Denote $Q(\theta) = ||Y - X\theta||_2^2$. The select τ such that

$$t_\gamma = \max_{t\geq 0} Q\left(\widehat{\beta}^{thresh+OLS}(\tau)\right) - Q\left(\widehat{\beta}^{L}\right) \leq \gamma \tag{61}$$

for $\gamma \leq 0$. Taking $\gamma = 0$ returns the sparsest OLS-reflated model that generates a lower residual sum of squares than the LASSO estimator. We follow the suggestion of Belloni and Chernozhukov (2013, expr 2.14) and take $\gamma = \left\{Q\left(\widehat{\beta}^{thresh+OLS}(0)\right) - Q\left(\widehat{\beta}^{L}\right)\right\}/2$ in the simulations.

References

Belloni, Alexandre and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19(2):521–547.

Bickel, Peter, Ya'acov Ritov and Alexandre Tsybakov. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics* 37(4):1705–1732.

Buhlmann, Peter and Sara van de Geer. 2013. *Statistics for High-Dimensional Data*. Springer.

Chatterjee, A and SN Lahiri. 2011. "Bootstrapping Lasso Estimators." *Journal of the American Statistical Association* 106(494):608–625.

Chatterjee, A and SN Lahiri. 2013. "Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap." *The Annals of Statistics* 41(3):1232–1259.

Liu, H. and B. Yu. 2013. "Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression." *Electronic Journal of Statistics* 7(3124–3169).

Minnier, Jessica, Lu Tian and Tianxi Cai. 2011. "A perturbation method for inference on regularized regression estimates." *Journal of the American Statistical Association* 106(496).

Tripathi, Gautam. 1999. "A matrix extension of the Cauchy-Schwarz inequality." *Economics Letters* 63:1–3.

Wang, Hansheng and Chenlei Leng. 2007. "Unified LASSO Estimation by Least Squares Approximation." *Journal of the American Statistical Association* 102(479):1039–1048.

Zou, Hui, Trevor Hastie and Robert Tibshirani. 2007. "On the degrees of freedom of the lasso." *The Annals of Statistics* 35(5):2173–2192.