

Supplementary Material

EPR-C3: A Deterministic Constraint-Aware Heuristic for High-Dimensional Subset Selection in Multiple Linear Regression

Jackson J. Alcázar^a

^aCentro de Química Médica, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, 7780272, Chile

1. Algorithmic summaries of comparative methods

This section summarizes the additional subset-search strategies used as comparative baselines in the main manuscript. For consistency with the comparative benchmark, all candidate subsets generated by these methods were ultimately refitted by ordinary least squares (OLS), evaluated using the same admissibility criteria, screened to remove duplicate predictor combinations, and ranked using the same target performance criterion.

Algorithm S1. Forward stepwise regression

Require: Candidate predictors $\mathcal{X}$, response y , maximum model size w , target metric f , significance level α , VIF threshold τ_{VIF} .

Ensure: Ranked set of admissible predictor subsets $\mathcal{M}$.

- 1: $\mathcal{M} \leftarrow \emptyset$ and $S \leftarrow \emptyset$.
- 2: **while** $|S| < w$ **do**
- 3: Generate one-variable additions $S \cup \{x\}$ for all $x \in \mathcal{X} \setminus S$.
- 4: Refit each candidate subset by OLS.
- 5: Retain candidates satisfying the admissibility criteria.
- 6: **if** no admissible candidate remains **then**
- 7: stop the trajectory.
- 8: **end if**
- 9: Select the admissible candidate with the best value of f and update S .
- 10: Store the selected admissible subset in $\mathcal{M}$.
- 11: **end while**
- 12: Remove duplicate predictor sets from $\mathcal{M}$.
- 13: Rank $\mathcal{M}$ by the target metric.
- 14: **return** $\mathcal{M}$.

Algorithm S2. Backward stepwise regression

Require: Candidate predictors $\mathcal{X}$, response y , maximum model size w , initial candidate-pool size k , target metric f , significance level α , VIF threshold τ_{VIF} .

Ensure: Ranked set of admissible predictor subsets $\mathcal{M}$.

- 1: $\mathcal{M} \leftarrow \emptyset$.
- 2: Define an initial candidate pool $\mathcal{X}_k \subseteq \mathcal{X}$ from the ranked predictor set.
- 3: Enumerate all w -predictor subsets from $\mathcal{X}_k$.
- 4: Refit each subset by OLS and retain subsets satisfying the admissibility criteria.
- 5: **if** no admissible w -predictor subset is available **then**
- 6: stop the trajectory.
- 7: **end if**
- 8: Select the admissible w -predictor subset with the best value of f and set it as S .
- 9: Store S in $\mathcal{M}$.
- 10: **while** $|S| > 1$ **do**

- 11: Generate one-variable removals $S \setminus \{x\}$ for all $x \in S$.
- 12: Refit each candidate subset by OLS.
- 13: Retain candidates satisfying the admissibility criteria.
- 14: **if** no admissible candidate remains **then**
- 15: stop the trajectory.
- 16: **end if**
- 17: Select the admissible candidate with the best value of f and update S .
- 18: Store the selected admissible subset in $\mathcal{M}$.
- 19: **end while**
- 20: Remove duplicate predictor sets from $\mathcal{M}$.
- 21: Rank $\mathcal{M}$ by the target metric.
- 22: **return** $\mathcal{M}$.

Algorithm S3. Lasso screening followed by exhaustive OLS subset refinement

Require: Candidate predictors $\mathcal{X}$, response y , maximum model size w , preselection size $k = 20$, target metric f , significance level α , VIF threshold τ_{VIF} .

Ensure: Ranked set of admissible predictor subsets $\mathcal{M}$.

- 1: Standardize the predictor matrix.
- 2: Fit LassoCV using 5-fold cross-validation.
- 3: Rank predictors by the absolute magnitude of their fitted coefficients.
- 4: Select the top $k = 20$ predictors to form the reduced pool $\mathcal{X}_k$.
- 5: Enumerate all subsets of $\mathcal{X}_k$ with size from 1 to w .
- 6: Refit each subset by OLS and compute the target metric f .
- 7: Retain subsets satisfying the admissibility criteria.
- 8: Remove duplicate predictor sets from the retained subsets.
- 9: Rank retained subsets by the target metric.
- 10: **return** the ranked admissible subsets $\mathcal{M}$.

Algorithm S4. Ridge screening followed by exhaustive OLS subset refinement

Require: Candidate predictors $\mathcal{X}$, response y , maximum model size w , preselection size $k = 20$, target metric f , significance level α , VIF threshold τ_{VIF} .

Ensure: Ranked set of admissible predictor subsets $\mathcal{M}$.

- 1: Standardize the predictor matrix.
- 2: Fit RidgeCV.
- 3: Rank predictors by the absolute magnitude of their fitted coefficients.
- 4: Select the top $k = 20$ predictors to form the reduced pool $\mathcal{X}_k$.
- 5: Enumerate all subsets of $\mathcal{X}_k$ with size from 1 to w .
- 6: Refit each subset by OLS and compute the target metric f .
- 7: Retain subsets satisfying the admissibility criteria.
- 8: Remove duplicate predictor sets from the retained subsets.
- 9: Rank retained subsets by the target metric.
- 10: **return** the ranked admissible subsets $\mathcal{M}$.

Algorithm S5. Elastic Net screening followed by exhaustive OLS subset refinement

Require: Candidate predictors $\mathcal{X}$, response y , maximum model size w , preselection size $k = 20$, l_1 -ratio grid $\{0.2, 0.5, 0.8\}$, target metric f , significance level α , VIF threshold τ_{VIF} .

Ensure: Ranked set of admissible predictor subsets $\mathcal{M}$.

- 1: Standardize the predictor matrix.
- 2: Fit ElasticNetCV using 5-fold cross-validation over l_1 ratios of 0.2, 0.5, and 0.8.
- 3: Rank predictors by the absolute magnitude of their fitted coefficients.
- 4: Select the top $k = 20$ predictors to form the reduced pool $\mathcal{X}_k$.
- 5: Enumerate all subsets of $\mathcal{X}_k$ with size from 1 to w .
- 6: Refit each subset by OLS and compute the target metric f .
- 7: Retain subsets satisfying the admissibility criteria.
- 8: Remove duplicate predictor sets from the retained subsets.
- 9: Rank retained subsets by the target metric.
- 10: **return** the ranked admissible subsets $\mathcal{M}$.

Algorithm S6. Correlation Screening and Redundancy Filtering before All-Subsets Search

Require: Candidate predictors $\mathcal{X}$, response y , maximum model size w , preselection size $k = 20$, redundancy threshold τ_r , target metric f , significance level α , VIF threshold τ_{VIF} .

Ensure: Ranked set of admissible predictor subsets $\mathcal{M}$.

- 1: Compute the absolute Pearson correlation between each predictor and y .
- 2: Rank predictors in decreasing order of absolute correlation with the response.
- 3: Initialize the reduced nonredundant pool $\mathcal{X}_k \leftarrow \emptyset$.
- 4: **for** each predictor in the ranked list **do**
- 5: **if** its absolute Pearson correlation with every predictor already in $\mathcal{X}_k$ does not exceed τ_r **then**
- 6: Add the predictor to $\mathcal{X}_k$.
- 7: **end if**
- 8: **if** $|\mathcal{X}_k| = k$ **then**
- 9: stop screening.
- 10: **end if**
- 11: **end for**
- 12: Enumerate all subsets of $\mathcal{X}_k$ with size from 1 to w .
- 13: Refit each subset by OLS and compute the target metric f .
- 14: Retain subsets satisfying the admissibility criteria.
- 15: Remove duplicate predictor sets from the retained subsets.
- 16: Rank retained subsets by the target metric.
- 17: **return** the ranked admissible subsets $\mathcal{M}$.

Algorithm S7. Genetic algorithm for subset search

Require: Candidate predictors $\mathcal{X}$, response y , maximum model size w , population size $N = 1000$, number of generations $G = 20000$, mutation probability $p_m = 0.08$, crossover probability $p_c = 0.8$, target metric f , significance level α , VIF threshold τ_{VIF} .

Ensure: Ranked set of admissible predictor subsets $\mathcal{M}$.

- 1: Encode each candidate subset as a binary chromosome over $\mathcal{X}$.
- 2: Generate an initial random population of $N = 1000$ valid subsets with sizes between 1 and w .
- 3: $\mathcal{M} \leftarrow \emptyset$.
- 4: **for** $g = 1, \dots, G$ **do**
- 5: Decode each chromosome into a predictor subset.
- 6: Refit each subset by OLS and compute the target metric f .
- 7: Retain admissible subsets satisfying the admissibility criteria in $\mathcal{M}$.

- 8: Select elite individuals according to their target-metric scores.
- 9: Generate offspring by single-point crossover with probability $p_c = 0.8$.
- 10: Apply bit-flip mutation with probability $p_m = 0.08$.
- 11: Repair offspring violating the model-size constraint.
- 12: **if** an offspring subset is larger than w **then**
- 13: Randomly remove excess active predictors until $|S| \leq w$.
- 14: **end if**
- 15: **if** an offspring subset is empty **then**
- 16: Randomly activate one predictor to enforce $|S| \geq 1$.
- 17: **end if**
- 18: Form the next population from elite individuals and repaired offspring.
- 19: **end for**
- 20: Remove duplicate predictor sets from $\mathcal{M}$.
- 21: Rank retained subsets by the target metric.
- 22: **return** the ranked admissible subsets $\mathcal{M}$.

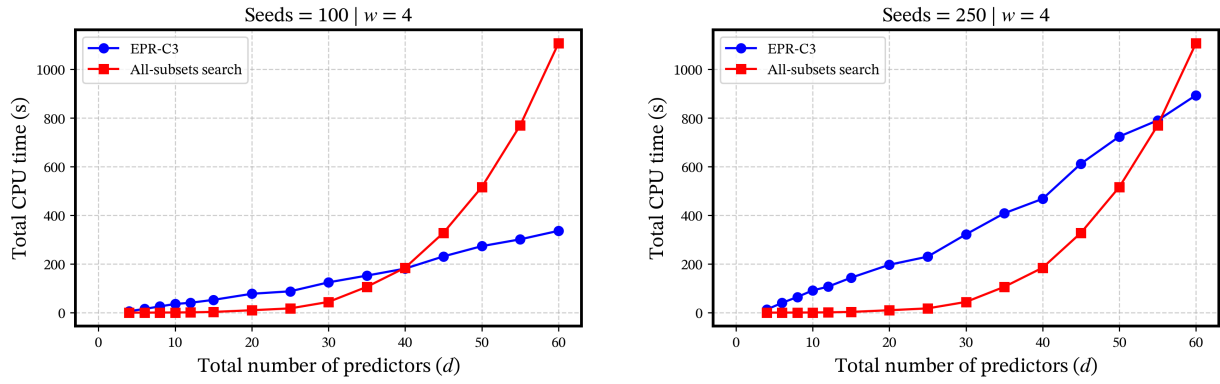


Figure S1: CPU-time scaling of EPR-C3 and all-subsets regression as a function of the total number of predictors (d) under reduced seed counts: (a) 100 seeds and (b) 250 seeds, with $w = 4$.

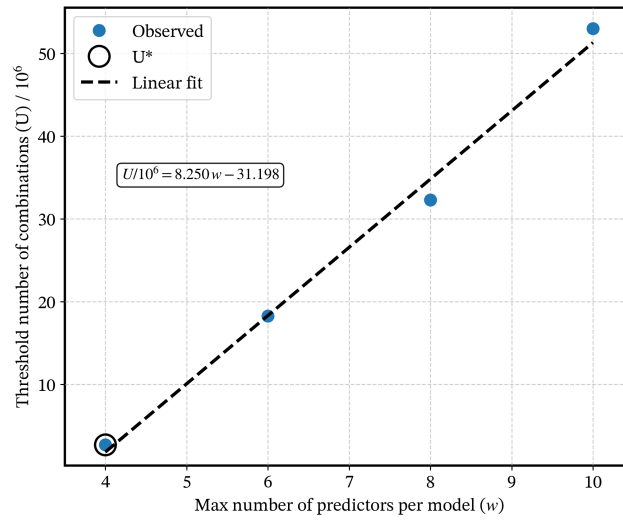


Figure S2: Threshold number of combinations U as a function of the maximum predictors per model (w) for the 1000-seed configuration.