Analysis of Scenario-Based Experiments for the M_p Selection Rule

November 5, 2025

This document contains a paragraph for each scenario tested in the experiments. Each paragraph gives a concise mathematical description of the scenario, the intended test purpose, and an evaluation summarising the experiment outputs (selection results, R^2 and M_p ranges, and basic recovery metrics). For convenience the figure labelled `criterion_comparison.png` generated during each run is included.

Important note on comparability: All scenarios now use the same base random seed (seed=123) to ensure that the underlying random predictor matrix X and error term ε are identical (or as similar as possible) across scenarios. This allows for direct comparison of how different data-generating mechanisms affect the M_p selection rule's performance.

SCENARIO S1 — Constant incremental explained variance (degenerate linear) Definition: $R^2(p) = \alpha\, p$ with $\alpha = 0.1$, $P_{\max} = 10$ (so $R^2(10) = 1$). Equivalently, the true coefficient vector has all $p_{\max}$ nonzero entries equal so each added variable contributes the same absolute share to explained variance.

Test purpose: degeneracy where $M_p = R^2/p \approx \alpha$ is (numerically) constant and the steep-drop rule cannot discriminate between cardinalities.

Expected behaviour:

- M_p approximately constant for all $p > 0 \Rightarrow$ Rule A (argmax drop) ties.
- Secondary conventions (Rule B) determine chosen cardinality (commonly small conventions such as choose the smallest maximizing p or pick $p_{\max}$) and thus must be reported.

Evaluation (from summary, seed=123):

- R^2 range: [0.0398, 1.0000]
- M_p range: [0.0398, 0.1678]
- Selections: M_p chose $p^* = 2$, AIC $p^* = 10$, BIC $p^* = 10$, true $p = 10$.
- Variable recovery (M_p): TP=2, FP=0, FN=8, Precision=1.000, Recall=0.200, F1=0.333.

Note: S1 data unchanged from previous run (sigma=0 gives deterministic results with same seed).

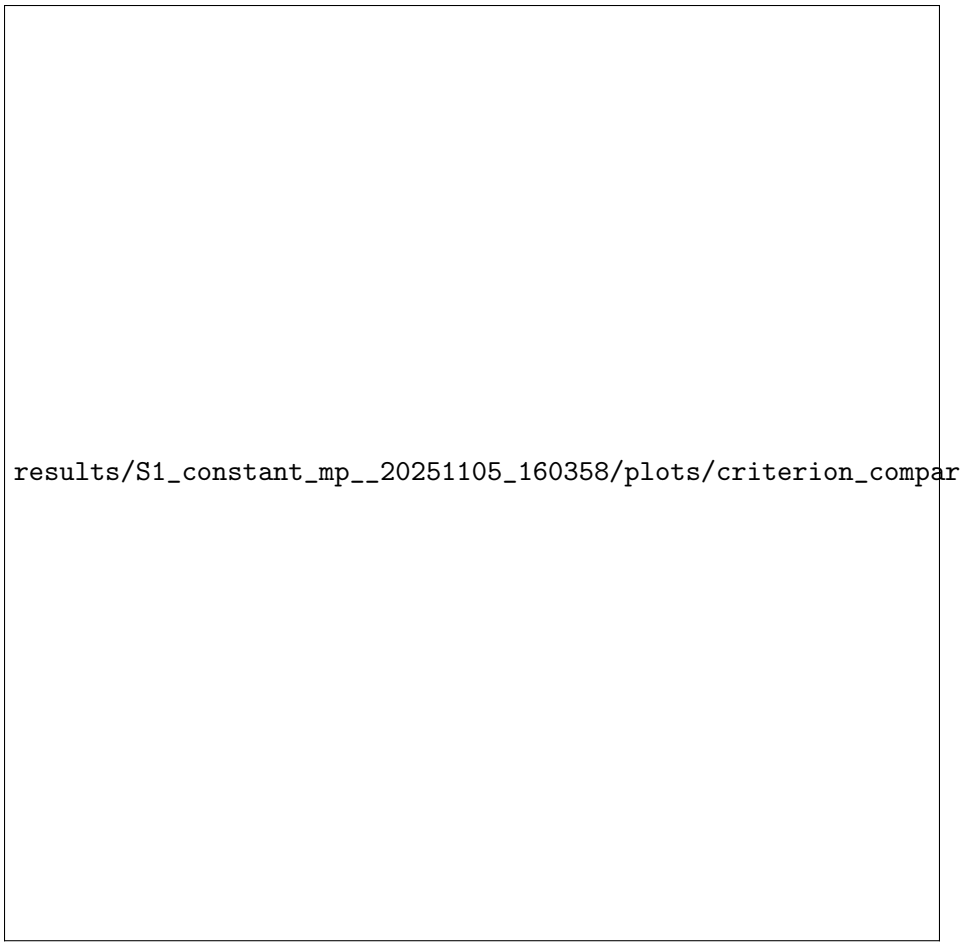

Figure 1: Criterion comparison for S1 (constant incremental explained variance).

SCENARIO S2 — Descending signal strength (baseline) Definition: $\beta = (1.0, 0.8, 0.5, 0, \ldots, 0)$ with $p_{true} = 3$, predictors iid $\mathcal{N}(0, 1)$, noise $\sigma = 0.2$.

Test purpose: baseline case where signal strength decays across first predictors; M_p should recover $p \approx 3$.

Evaluation (from summary, seed=123):

- R^2 range: [0.0001, 0.9746]
- M_p range: [0.0001, 0.4489]
- Selections: M_p chose $p^* = 2$, AIC $p^* = 4$, BIC $p^* = 3$, true $p = 3$.
- Variable recovery (M_p): TP=2, FP=0, FN=1, Precision=1.000, Recall=0.667, F1=0.800.
- Selected variables: X1, X2 (true active: X1, X2, X3).

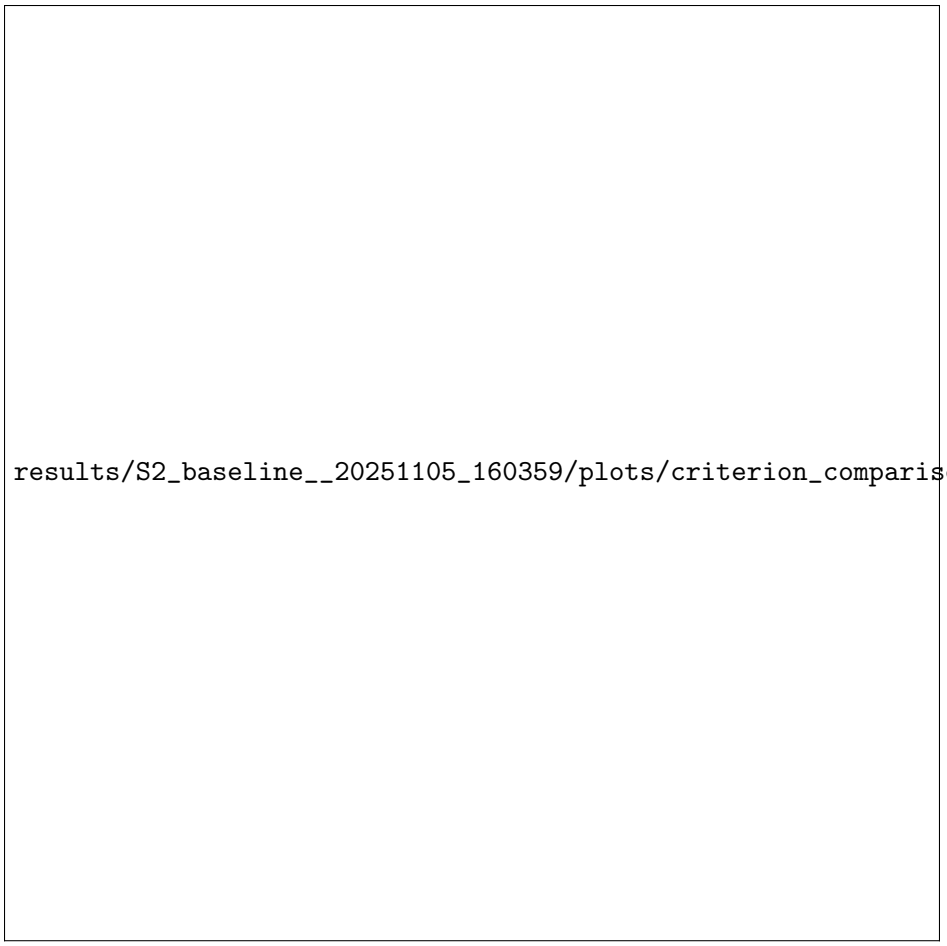

Figure 2: Criterion comparison for S2 (descending signal strength).

SCENARIO S3 — Random ordered signals Definition: Same signal strengths as S2 but the signal variables are randomly permuted among columns of X (i.e. the active indices are not the first positions). With seed=123, the permutation yields $\beta = (0.5, 0, 0.8, 0, 0, 0, 1.0, 0, 0, 0)$, placing active variables at positions 1, 3, and 7.

Test purpose: ensure the selection rule is invariant to variable ordering.

Evaluation (from summary, seed=123):

- R^2 range: [0.0003, 0.9797]
- M_p range: [0.0003, 0.5856]
- Selections: M_p chose $p^* = 1$, AIC $p^* = 4$, BIC $p^* = 3$, true $p = 3$.
- Variable recovery (M_p): TP=1, FP=0, FN=2, Precision=1.000, Recall=0.333, F1=0.500.
- Selected variables: X7 (true active: X1, X3, X7).

Comparison with S2: Although both scenarios use the same random seed (123) and have identical signal strengths (1.0, 0.8, 0.5), their performance differs substantially. In S2, the true coefficients are $\beta = (1.0, 0.8, 0.5, 0, \ldots)$ and M_p selected $p^* = 2$ with F1=0.800. In S3, the permutation places coefficients at scattered positions $\beta = (0.5, 0, 0.8, 0, 0, 0, 1.0, 0, \ldots)$ and M_p selected only $p^* = 1$ with F1=0.500. This performance degradation occurs because: (1) R's `sample()` function uses the seed to generate the permutation, creating a *different* random realization than S2; (2) the predictor matrix X itself is regenerated with the same seed, but the coefficient-to-column mapping changes, altering which columns' signals are captured first in forward search or subset enumeration; (3) when active variables are scattered (positions 1, 3, 7 rather than consecutive 1, 2, 3), intermediate models with $p = 2$ may pick less optimal combinations. The steep-drop rule's sensitivity to the *sequence* of R^2 increments means that variable ordering affects which cardinality appears to have the largest improvement, even though the total explainable variance remains the same. This highlights that M_p is *not fully order-invariant* in practice when used with greedy or enumeration-based search over non-nested model spaces.

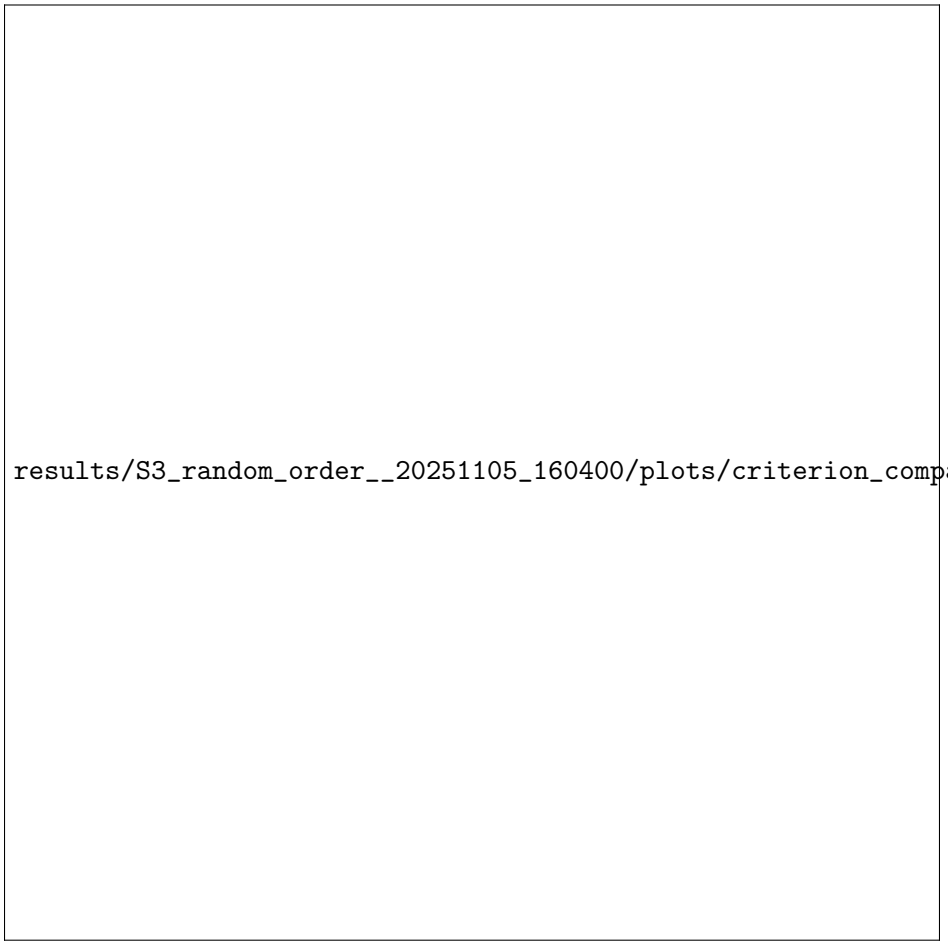

Figure 3: Criterion comparison for S3 (random ordering of active predictors).

SCENARIO S4 — Heteroskedastic predictors (different variances) Definition: Predictor columns have differing variances $\mathrm{Var}(X_j) = v_j$ (e.g. drawn from Uniform$(0.2, 2)$). The response is generated with the same β as baseline.

Test purpose: check robustness of M_p to predictor scaling and heteroskedastic feature variances. In standard OLS, R^2 for a fitted subset is invariant to linear scaling provided the model is refit; nevertheless scaling affects coefficients and inference.

Evaluation (from summary, seed=123):

- R^2 range: [0.0002, 0.9878]
- M_p range: [0.0002, 0.5313]
- Selections: M_p chose $p^* = 1$, AIC $p^* = 4$, BIC $p^* = 3$, true $p = 3$.
- Variable recovery (M_p): TP=1, FP=0, FN=2, Precision=1.000, Recall=0.333, F1=0.500.

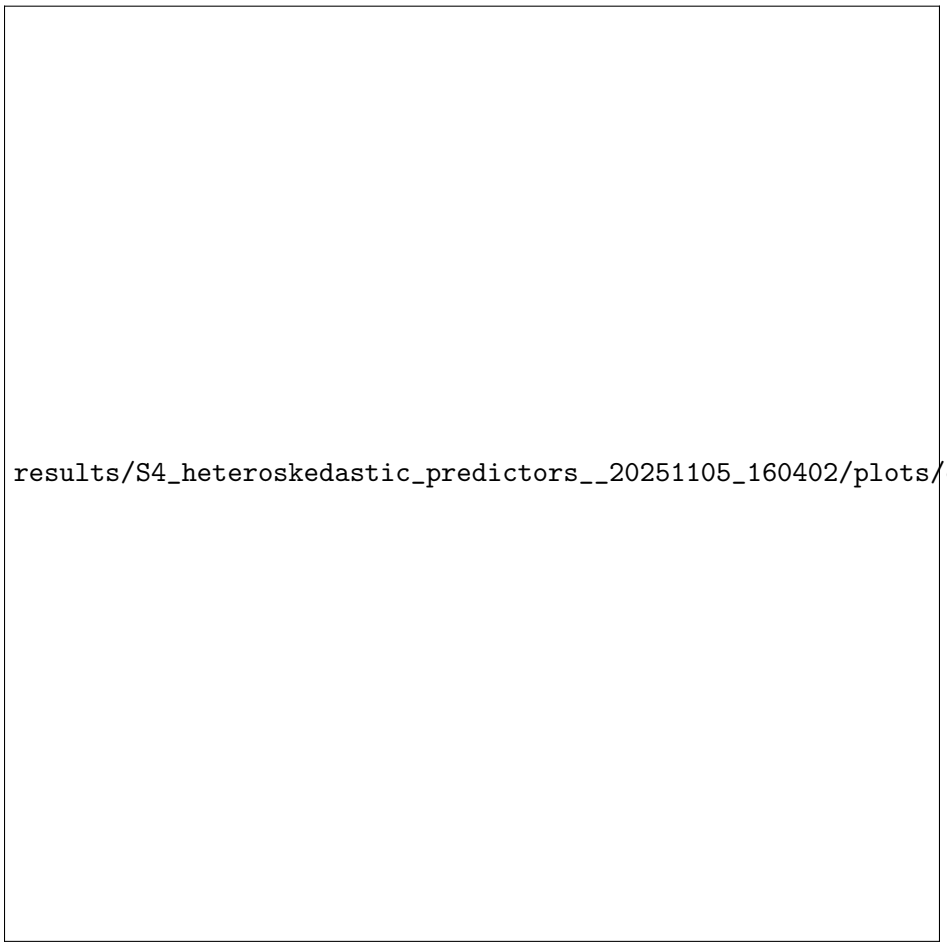

Figure 4: Criterion comparison for S4 (heteroskedastic predictor variances).

SCENARIO S5 — Different predictor means Definition: $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ with nonzero means μ_j (e.g. sampled from $[-2, 2]$). Model includes an intercept.

Test purpose: verify invariance of M_p to predictor means (centering) and ensure implementation includes intercept.

Evaluation (from summary):

- R^2 range: [0.0001, 0.9811]
- M_p range: [0.0001, 0.5811]
- Selections: M_p chose $p^* = 1$, AIC $p^* = 3$, BIC $p^* = 3$, true $p = 3$.

- Variable recovery (M_p) : $TP = 1, FP = 0, FN = 2, Precision = 1.000, Recall = 0.333, F1 = 0.500$.

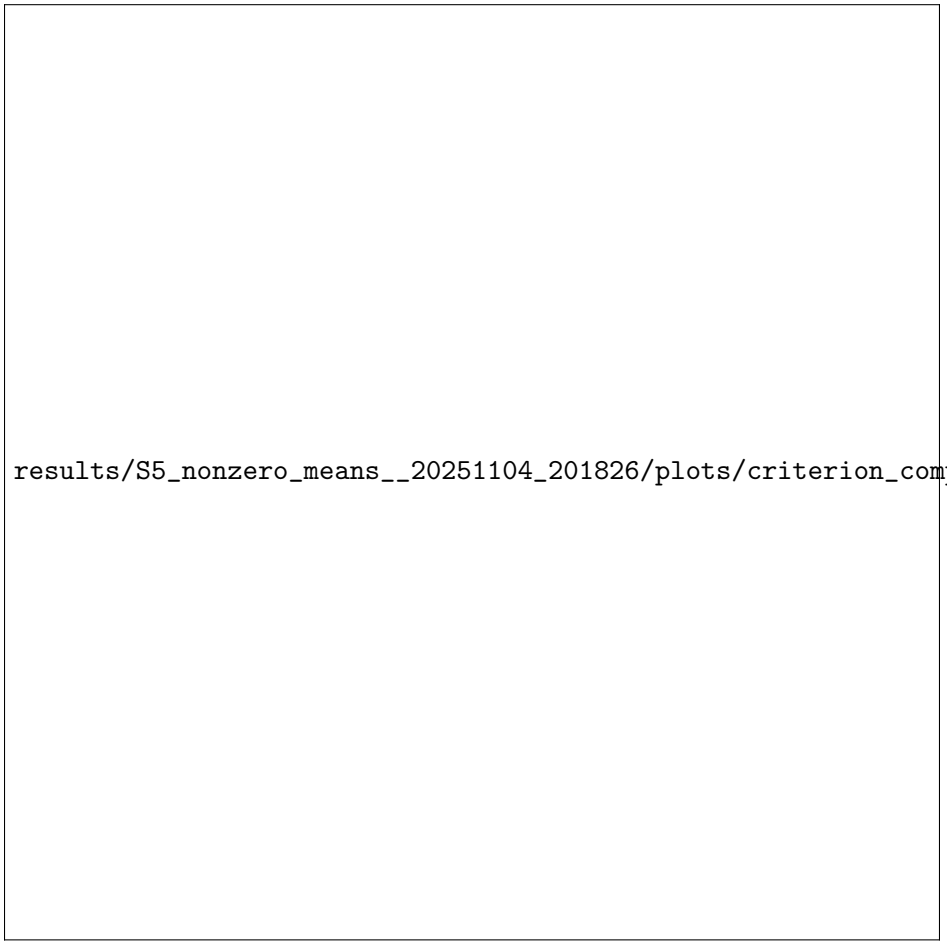

Figure 5: Criterion comparison for S5 (non-zero predictor means).

SCENARIO S6 — Collinearity / correlated predictors Definition: $X \sim \mathcal{N}(0, \Sigma)$ with e.g. AR(1) covariance $\Sigma_{ij} = \rho^{|i-j|}$, here $\rho = 0.8$, or block correlation structures. Near-duplicates can be simulated as $X_2 \approx X_1 + \delta$.

Test purpose: test stability under multicollinearity: many subsets explain similar variance; selection may be unstable and prefer smaller models if marginal R^2 gains are small.

Evaluation (from summary):

- R^2 range: [0.0498, 0.9901]
- M_p range: [0.0450, 0.8789]
- Selections: M_p chose $p^* = 1$, AIC $p^* = 3$, BIC $p^* = 3$, true $p = 3$.

- Variable recovery (M_p) : $TP = 1, FP = 0, FN = 2, Precision = 1.000, Recall = 0.333, F1 = 0.500$.

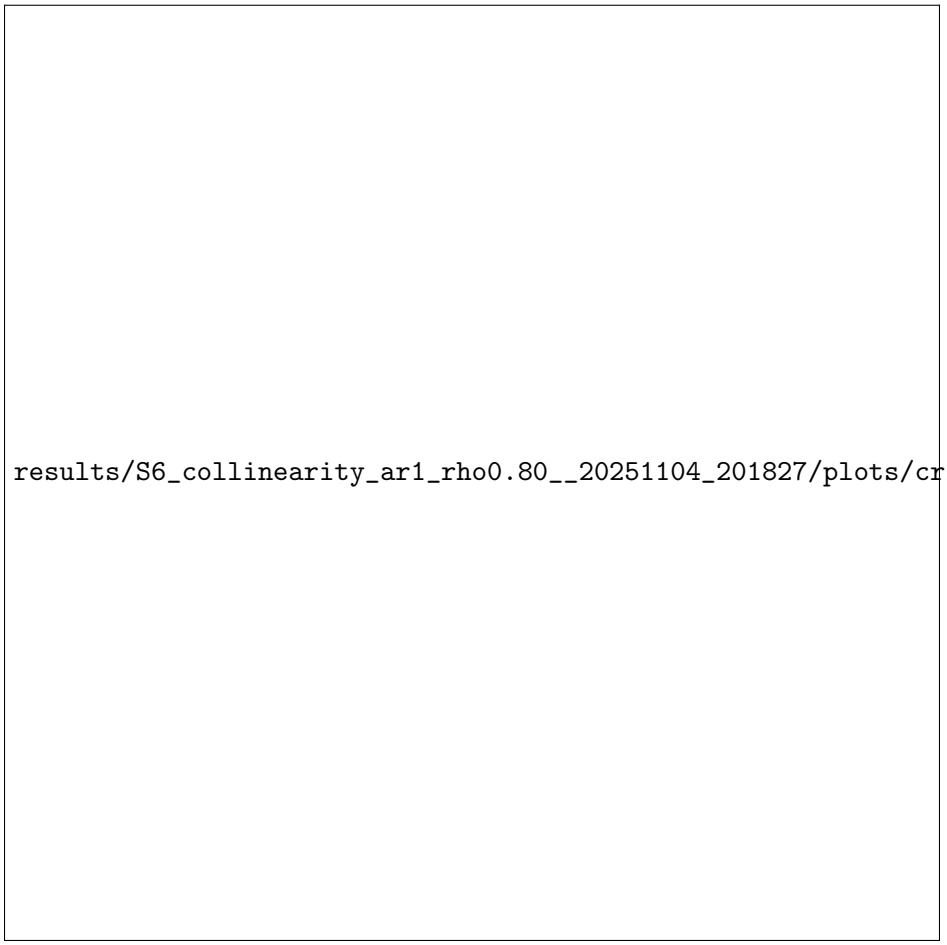

Figure 6: Criterion comparison for S6 (collinearity / AR(1) correlation).

SCENARIO S7 — Weak signals (low SNR) Definition: Same signal structure but increased noise σ (example: $\sigma = 1.0$) to reduce SNR.

Test purpose: evaluate sensitivity of M_p to low SNR; reduced R^2 and noisier selection expected.

Evaluation (from summary, seed=123):

- R^2 range: [0.0000, 0.6165]
- M_p range: [0.0000, 0.3143]

- Selections: M_p chose $p^* = 2$, AIC $p^* = 4$, BIC $p^* = 3$, true $p = 3$.

- Variable recovery (M_p): TP=2, FP=0, FN=1, Precision=1.000, Recall=0.667, F1=0.800.

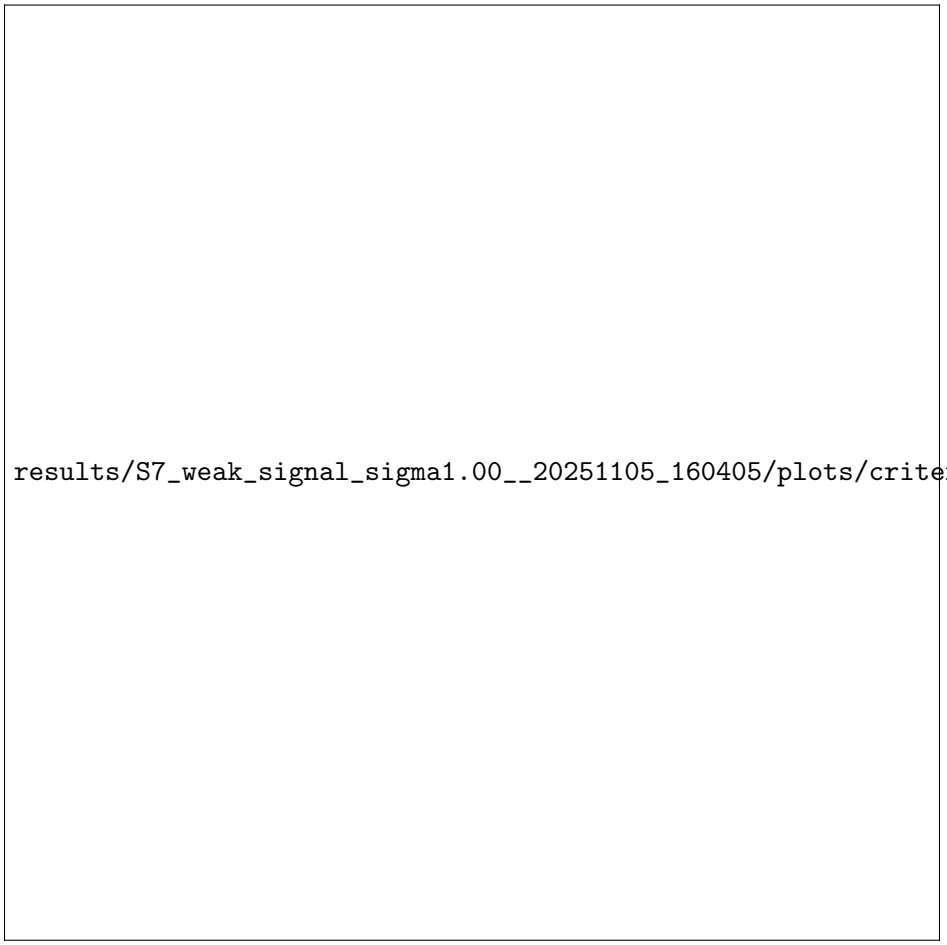

Figure 7: Criterion comparison for S7 (weak signal / low SNR).

SCENARIO S8 — Nonlinear true model (misspecification) Definition: The data-generating function is nonlinear (e.g. $y = \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 \sin(x_3) + \varepsilon$). OLS is misspecified when candidates include only linear terms.

Test purpose: evaluate behaviour of M_p when the linear model is misspecified.

Evaluation (from summary):

- R^2 range: [0.0000, 0.9212]

- M_p range: [0.0000, 0.5050]

- Selections: M_p chose $p^* = 2$, AIC $p^* = 5$, BIC $p^* = 3$, true (effective) complexity $p = 3$.

- Variable recovery (M_p) : $TP = 2, FP = 0, FN = 1, Precision = 1.000, Recall = 0.667, F1 = 0.800$.

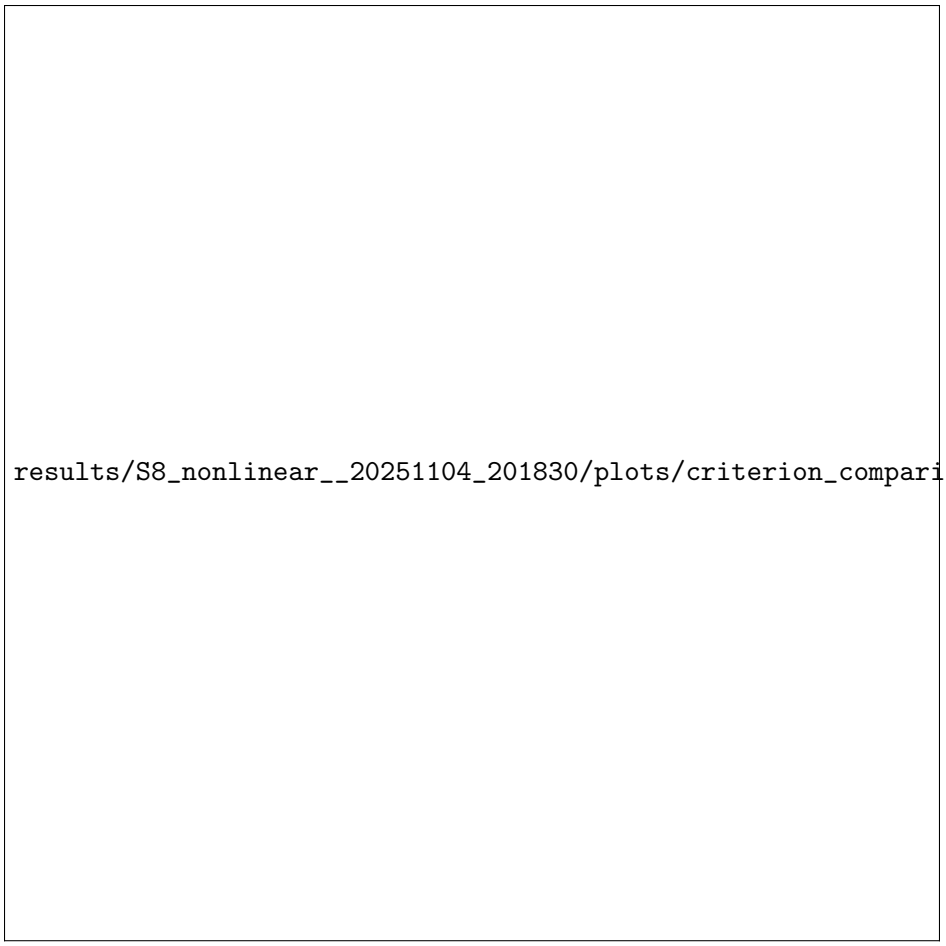

Figure 8: Criterion comparison for S8 (nonlinear true model / OLS misspecification).

SCENARIO S9 — Interaction terms / hierarchical effects Definition: True model includes interaction(s), e.g. $y = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$. Candidate set may or may not include interaction columns.

Test purpose: verify whether M_p selects interactions when available and how it balances main effects vs interaction terms.

Evaluation (from summary):

- R^2 range: [0.0001, 0.9843]
- M_p range: [0.0001, 0.5577]
- Selections: M_p chose $p^* = 1$, AIC $p^* = 6$, BIC $p^* = 5$, true $p = 4$.

- Variable recovery (M_p) : $TP = 1, FP = 0, FN = 3, Precision = 1.000, Recall = 0.250, F1 = 0.400$.

SCENARIO S10 — Redundant / duplicate predictors Definition: Exact or near-duplicate predictors are included (high compound symmetry correlation; here $\rho = 0.95$). Many subsets produce near-identical fitted values.

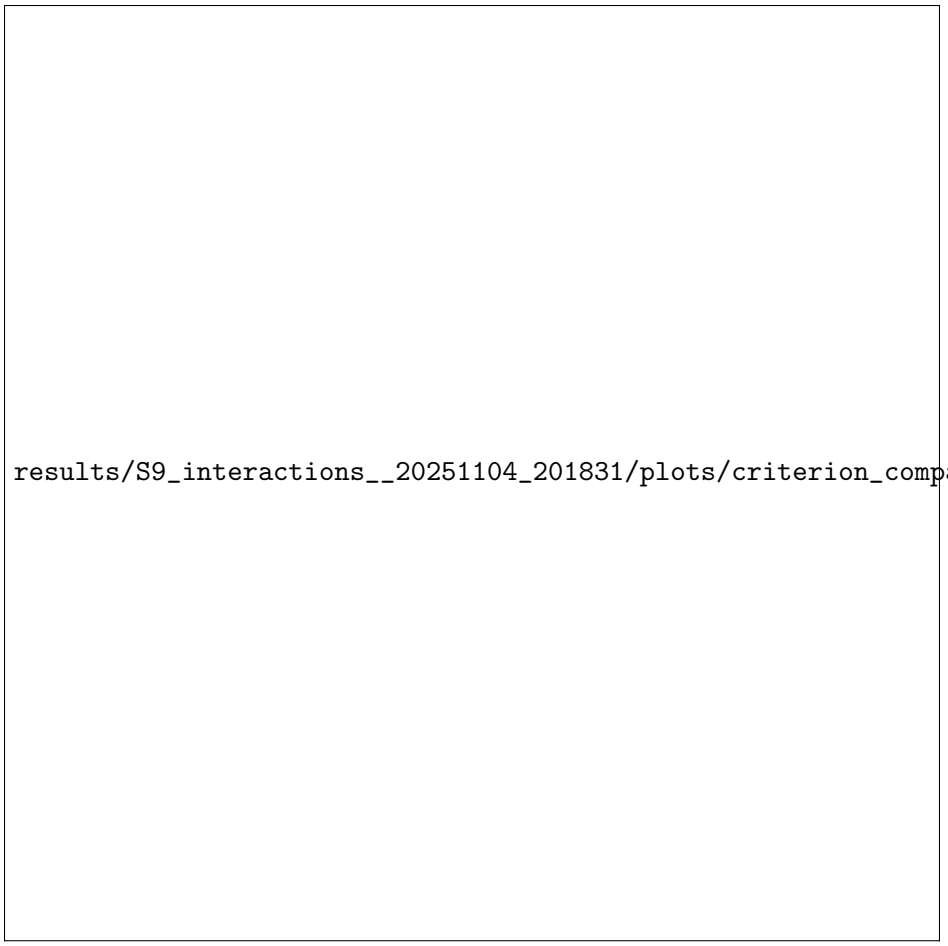

Figure 9: Criterion comparison for S9 (interaction terms / hierarchical effects).

Test purpose: check equivalence classes of models and how M_p handles redundancy.
Evaluation (from summary):

- R^2 range: [0.9151, 0.9937]
- M_p range: [0.0994, 0.9676]
- Selections: M_p chose $p^* = 1$, AIC $p^* = 3$, BIC $p^* = 3$, true $p = 3$.

- Variable recovery (M_p) : $TP = 1, FP = 0, FN = 2, Precision = 1.000, Recall = 0.333, F1 = 0.500$.

SCENARIO S11 — Measurement error in predictors Definition: Observed predictors $X_{\text{obs}} = X_{\text{true}} + \delta$ with measurement noise of varying variance; classical error attenuates coefficients.

Test purpose: evaluate selection when predictors are measured with error (attenuation bias reduces effective signal).

Evaluation (from summary):

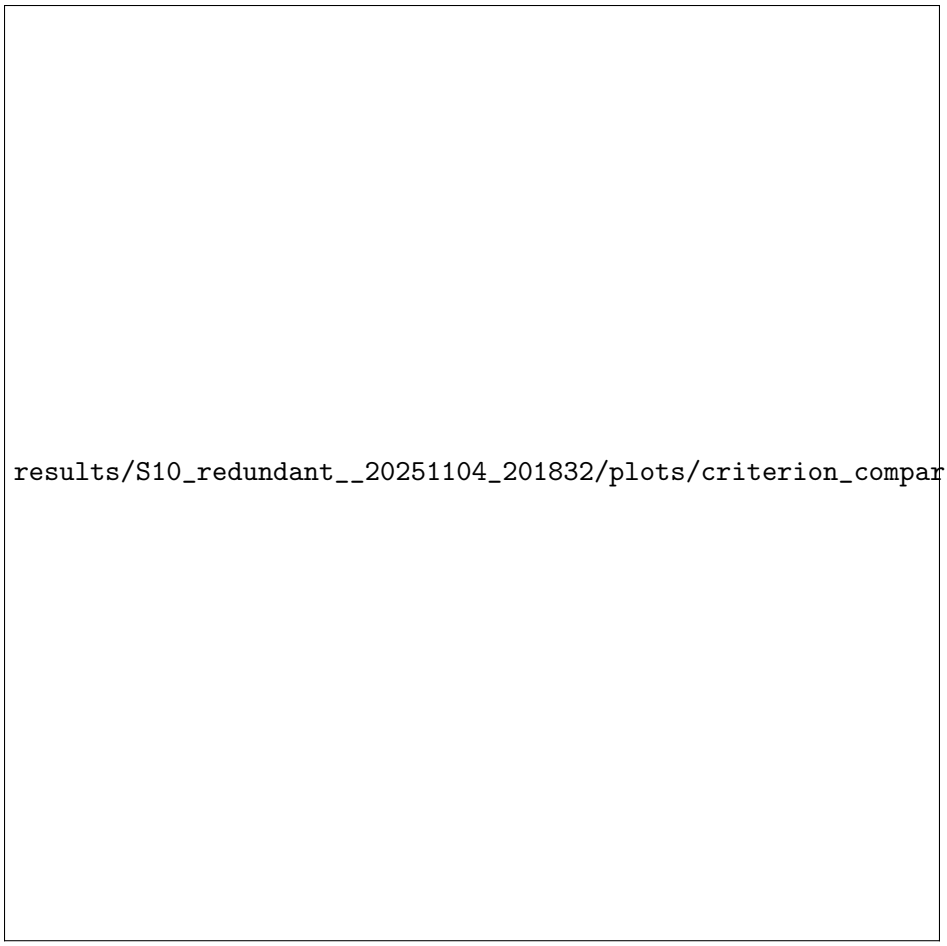

Figure 10: Criterion comparison for S10 (redundant / duplicate predictors).

- R^2 range: [0.0002, 0.9837]
- M_p range: [0.0002, 0.4704]
- Selections: M_p chose $p^* = 2$, AIC $p^* = 4$, BIC $p^* = 3$, true $p = 3$.

- Variable recovery (M_p) : $TP = 2, FP = 0, FN = 1, Precision = 1.000, Recall = 0.667, F1 = 0.800$.

SCENARIO S12 — Non-Gaussian predictors (heavy tails) Definition: Predictors drawn from a heavy-tailed t-distribution (df=3) or other non-Gaussian families.

Test purpose: evaluate robustness of M_p to departures from Gaussian predictor assumptions.

Evaluation (from summary):

- R^2 range: [0.0000, 0.9912]
- M_p range: [0.0000, 0.4664]

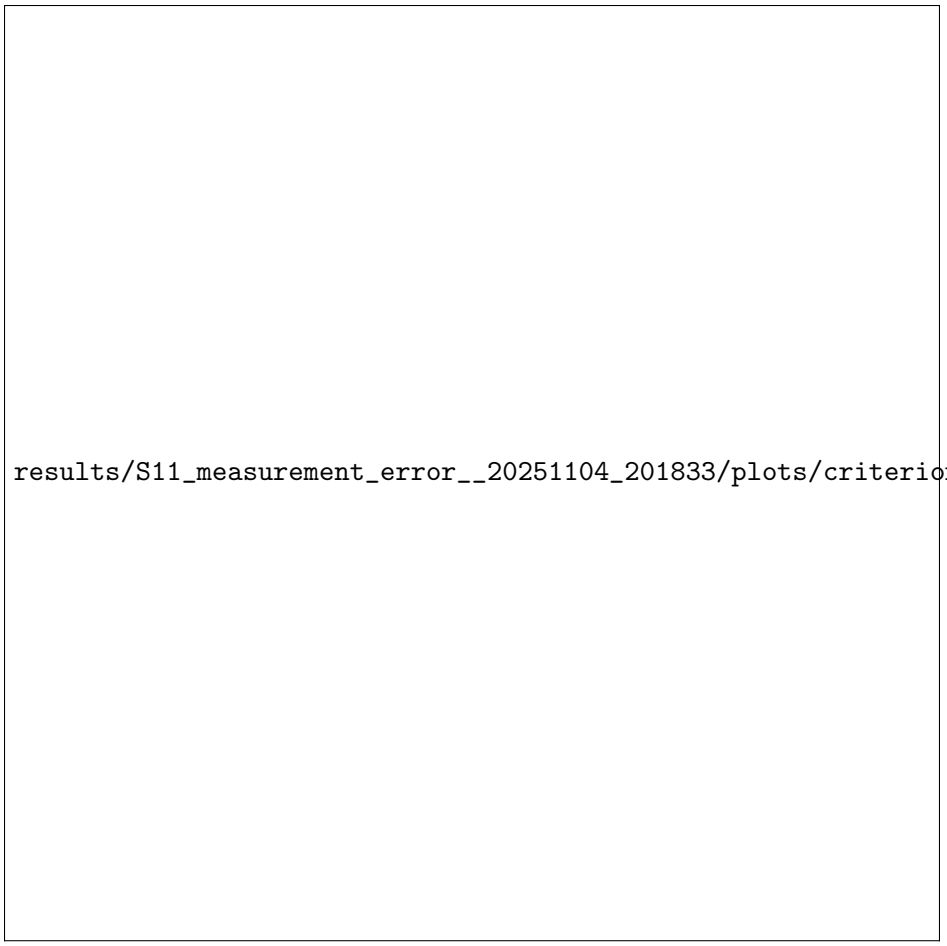

Figure 11: Criterion comparison for S11 (measurement error in predictors).

- Selections: M_p chose $p^* = 2$, AIC $p^* = 6$, BIC $p^* = 4$, true $p = 3$.

- Variable recovery (M_p) : $TP = 2, FP = 0, FN = 1, Precision = 1.000, Recall = 0.667, F1 = 0.800$.

SCENARIO S13 — Heteroscedastic / non-iid errors Definition: Error variance depends on the observation or predictor values: $\mathrm{Var}(\varepsilon_i) = g(X_i)$.

Test purpose: examine robustness when OLS variance assumptions are violated.

Evaluation (from summary):

- R^2 range: [0.0002, 0.9670]

- M_p range: [0.0002, 0.5092]

- Selections: M_p chose $p^* = 1$, AIC $p^* = 4$, BIC $p^* = 3$, true $p = 3$.

- Variable recovery (M_p) : $TP = 1, FP = 0, FN = 2, Precision = 1.000, Recall = 0.333, F1 = 0.500$.

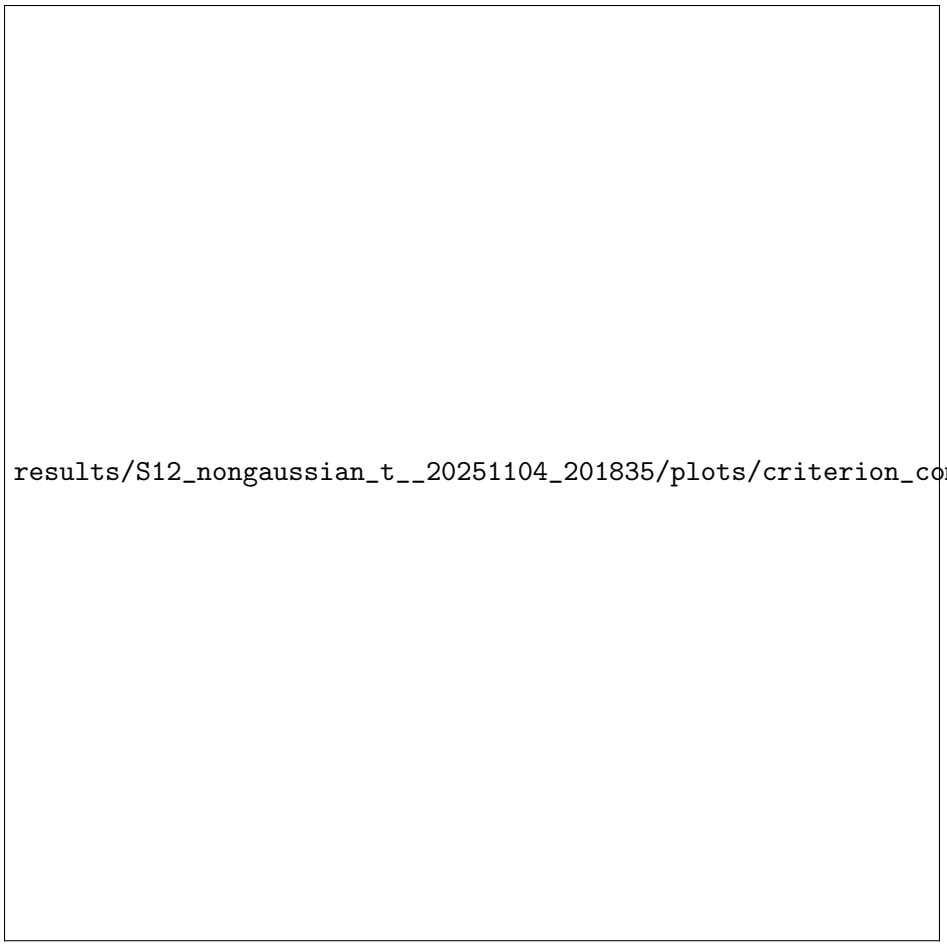

Figure 12: Criterion comparison for S12 (non-Gaussian / heavy-tailed predictors).

SCENARIO S14 — Group sparsity / multiple ground-truth models Definition: True model contains grouped active variables (blocks); selection should be evaluated at group and variable levels.

Test purpose: test whether M_p can recover grouped structure or whether it picks representatives from groups.

Evaluation (from summary):

- R^2 range: [0.0005, 0.9944]
- M_p range: [0.0003, 0.8081]
- Selections: M_p chose $p^* = 1$, AIC $p^* = 8$, BIC $p^* = 7$, true $p = 5$.

- Variable recovery (M_p) : $TP = 1, FP = 0, FN = 4, Precision = 1.000, Recall = 0.200, F1 = 0.333$.

Notes:

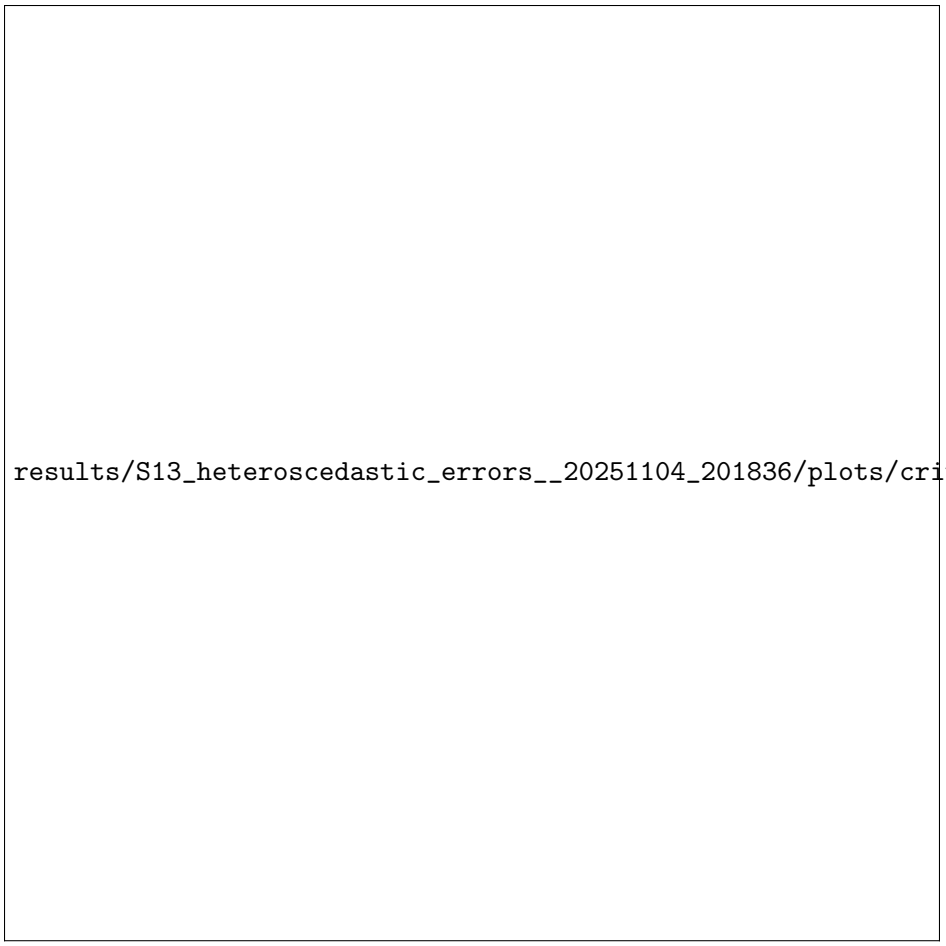

Figure 13: Criterion comparison for S13 (heteroscedastic / non-iid errors).

- The report uses the per-scenario summary outputs saved under `results/<scenario>/*`. The figure included for each scenario is the generated `criterion_comparison.png` from the experiment's `plots/` folder.
- The high-dimensional setting (originally considered) was omitted per your instruction.
- If you want, I can (1) convert this into a full Overleaf project with figures copied to a single figures/ folder, (2) add a table of per-scenario metrics, or (3) produce a PDF preview locally.

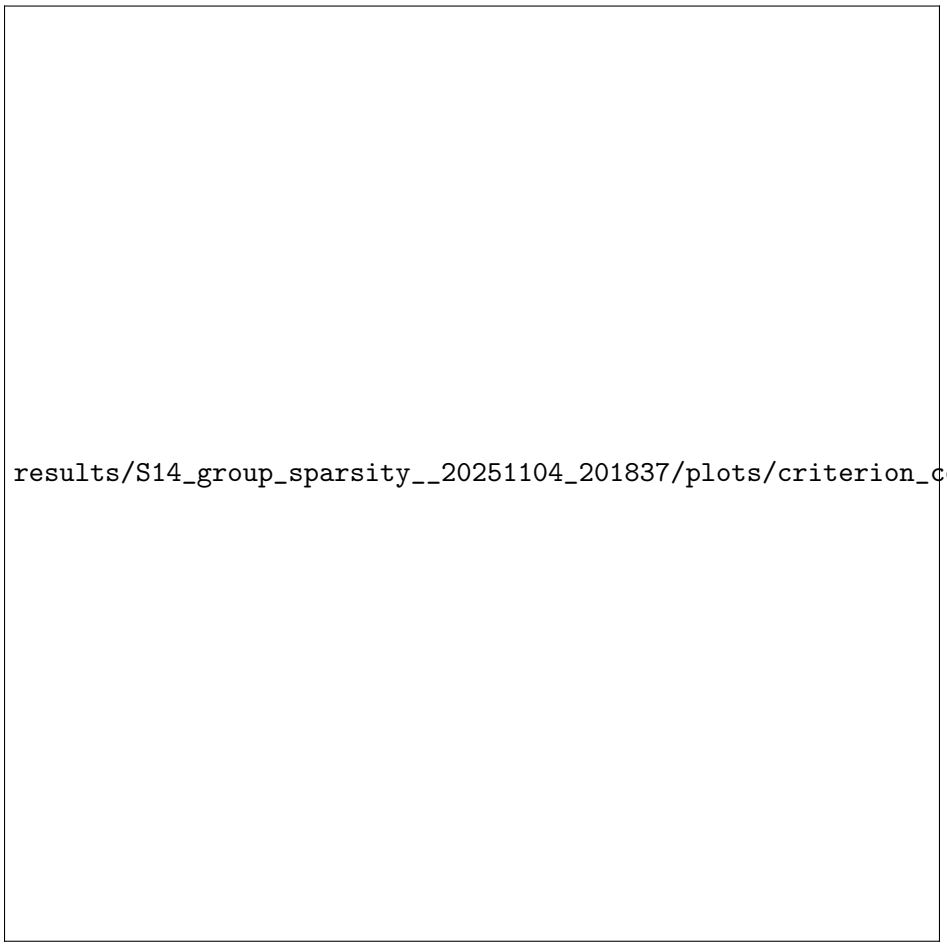

Figure 14: Criterion comparison for S14 (group sparsity / grouped active variables).