High-Dimensional Linear Model Inference with Conditional Resampling for Gene Discovery

Daniela Corbetta[1,(2)], Paolo Della Penna[1],
Pierre Neuvial[2], Livio Finos[1]

[1]Department of Statistical Sciences, University of Padova
[2]Institut de Mathématiques de Toulouse

30 Jan, 2026

Single-cell RNA-seq data: a high-dimensional inference problem

- scRNA-seq measures gene expression at single-cell resolution across thousands of genes.
- Data are high-dimensional and typically exhibit strong correlation among genes.
- Goal: identify genes (or sets of genes) active in modulating an outcome (disease/biomarker).

Identification of active genes

Aim: Find genes that are active in modulating an outcome of interest

Standard approach: For each gene $i \in \{1, \ldots, p\}$, fit a model using the gene as response and the outcome as covariate and test the null hypothesis of nullity of the coefficient associated with it.

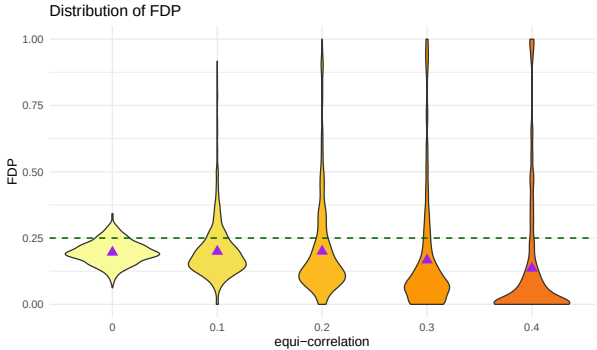
- This approach looks for marginal effects $\rightarrow$ switch to conditional inference.
- high-dimensional problem $\rightarrow$ multiple testing correction performed via FDR controlling methods

Limitations of FDR

- Sets obtained via FDR control methods cannot be manipulated

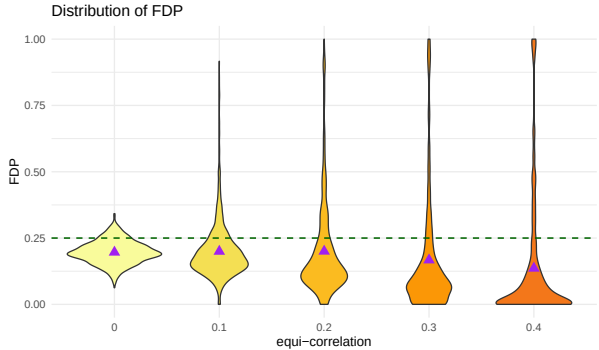
Limitations of FDR

- Sets obtained via FDR control methods cannot be manipulated
- FDR control is not FDP control



Limitations of FDR

- Sets obtained via FDR control methods cannot be manipulated
- FDR control is not FDP control



- Switch to the post-hoc FDP control framework: find a bound on the FDP, $\overline{\mathrm{FDP}} : R \subset \{1, \ldots, p\} \to [0, 1]$, such that for every possible set R,

$$\mathbb{P}\left(\mathrm{FDP}(R) \leq \overline{\mathrm{FDP}}(R), \forall R \subseteq \{1, ..., p\}\right) \geq 1 - \alpha$$

Inferential target: active predictors under dependence

High-dimensional regression model:

$$Y = X\beta + \varepsilon, \qquad p \gg n,$$

with $\beta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$, and $\varepsilon \sim N(0, \sigma^2 I)$.

Aim: provide confidence statements on the set of active predictors

$$\mathcal{A} = \{j : \beta_j \neq 0\},$$

and enable post-hoc guarantees on $\mathrm{FDP}(R)$ for data-dependent choices of R.

Existing approaches

- Debiased lasso (Van de Geer et al., 2014; Zhang & Zhang, 2014)
- Ridge projection (Bühlmann, 2013)
- Multisplit (Vesely et al., 2022)
- Permutation tests (Hemerik et al., 2021)

See Dezeure et al. (2015) for a review

Flipscores: a robust resampling-based test (intuition)

- Consider testing a coefficient via a score-type statistic.
- Approximate the null distribution by random sign-flipping of individual score contributions.
- Attractive in genomics:
  - robust to variance misspecification,
  - naturally accommodates dependence among statistics,
  - resampling structure is well-suited to post-hoc inference.

Flipscores for linear models

Model:

$$Y = x\beta + Z\gamma + \varepsilon$$

Interest in $H_0 : \beta = 0$, $\gamma \in \mathbb{R}^p$ vector of nuisance parameters.
The standardized flipscores (De Santis et al., 2025; Hemerik et al., 2020) test statistic is

$$T(F) = x^\top (I - H)F(I - H)Y / \sqrt{V(T(F))},$$

with $H = Z(Z^\top Z)^{-1}Z^\top$ and $F = \mathrm{diag_n}\{-1, +1\}$.

Flipscores for linear models

Model:

$$Y = x\beta + Z\gamma + \varepsilon$$

Interest in $H_0 : \beta = 0$, $\gamma \in \mathbb{R}^p$ vector of nuisance parameters.
The standardized flipscores (De Santis et al., 2025; Hemerik et al., 2020) test statistic is

$$T(F) = x^\top (I - H)F(I - H)Y / \sqrt{V(T(F))},$$

with $H = Z(Z^\top Z)^{-1}Z^\top$ and $F = \mathrm{diag_n}\{-1, +1\}$.

→ Formulated for a single parameter of interest, but we are interested on doing inference on the whole vector $\{\beta, \gamma\}$

→ Not directly applicable when $p > n$

Flipscores for linear models

Model:

$$Y = x\beta + Z\gamma + \varepsilon$$

Interest in $H_0 : \beta = 0$, $\gamma \in \mathbb{R}^p$ vector of nuisance parameters.
The standardized flipscores (De Santis et al., 2025; Hemerik et al., 2020) test statistic is

$$T(F) = x^\top (I - H)F(I - H)Y / \sqrt{V(T(F))},$$

with $H = Z(Z^\top Z)^{-1}Z^\top$ and $F = \mathrm{diag_n}\{-1, +1\}$.

- $\rightarrow$ Formulated for a single parameter of interest, but we are interested on doing inference on the whole vector $\{\beta, \gamma\}$
- $\rightarrow$ Not directly applicable when $p > n$
  - Perform variable selection on the Z matrix $\rightarrow$ need to deal with the double dipping issue

Inspiration

Zhao et al. (2021) show that, under some conditions, to test $H_0 : \beta = 0$

1. Performing a lasso regression of Y on $[Z, x]$ for an approriate λ value

Inspiration

Zhao et al. (2021) show that, under some conditions, to test $H_0 : \beta = 0$

1. Performing a lasso regression of Y on $[Z, x]$ for an approriate λ value
2. Computing the naive score test statistics, i.e. the parametric score using only the covariates kept by the lasso, $Z_{\hat{\mathcal{A}}_\lambda}$, to compute the projection matrix:

$$T = x^\top (I - H_{\hat{\mathcal{A}}_\lambda}) Y, \quad \text{with } H_{\hat{\mathcal{A}}_\lambda} = Z_{\hat{\mathcal{A}}_\lambda} (Z_{\hat{\mathcal{A}}_\lambda}^\top Z_{\hat{\mathcal{A}}_\lambda})^{-1} Z_{\hat{\mathcal{A}}_\lambda}^\top$$

Inspiration

Zhao et al. (2021) show that, under some conditions, to test $H_0 : \beta = 0$

1. Performing a lasso regression of Y on $[Z, x]$ for an approriate λ value
2. Computing the naive score test statistics, i.e. the parametric score using only the covariates kept by the lasso, $Z_{\hat{\mathcal{A}}_\lambda}$, to compute the projection matrix:

$$T = x^\top (I - H_{\hat{\mathcal{A}}_\lambda}) Y, \quad \text{with } H_{\hat{\mathcal{A}}_\lambda} = Z_{\hat{\mathcal{A}}_\lambda} (Z_{\hat{\mathcal{A}}_\lambda}^\top Z_{\hat{\mathcal{A}}_\lambda})^{-1} Z_{\hat{\mathcal{A}}_\lambda}^\top$$

3. Doing inference based on the asymptotic normal distribution of the naive score test statistics

Inspiration

Zhao et al. (2021) show that, under some conditions, to test $H_0 : \beta = 0$

1. Performing a lasso regression of Y on $[Z, x]$ for an approriate λ value
2. Computing the naive score test statistics, i.e. the parametric score using only the covariates kept by the lasso, $Z_{\hat{\mathcal{A}}_\lambda}$, to compute the projection matrix:

$$T = x^\top (I - H_{\hat{\mathcal{A}}_\lambda}) Y, \quad \text{with } H_{\hat{\mathcal{A}}_\lambda} = Z_{\hat{\mathcal{A}}_\lambda} (Z_{\hat{\mathcal{A}}_\lambda}^\top Z_{\hat{\mathcal{A}}_\lambda})^{-1} Z_{\hat{\mathcal{A}}_\lambda}^\top$$

3. Doing inference based on the asymptotic normal distribution of the naive score test statistics

is asymptotically valid: $\mathbb{P}_{H_0}(\text{reject } H_0) \to \alpha$

Idea

To test $H_0 : \beta = 0$

1. Perform a lasso regression of Y on Z for an appropriate value of λ
2. keep $\hat{\mathcal{A}}_\lambda = \{j : \hat{\gamma}_{j,\lambda} \neq 0\}$
3. Do the flipscores test considering $Z_{\hat{\mathcal{A}}_\lambda}$

Idea

To test $H_0 : \beta = 0$

1. Perform a lasso regression of Y on Z for an appropriate value of λ
2. keep $\hat{\mathcal{A}}_\lambda = \{j : \hat{\gamma}_{j,\lambda} \neq 0\}$
3. Do the flipscores test considering $Z_{\hat{\mathcal{A}}_\lambda}$

This is asymptotically valid

Idea

To test $H_0 : \beta = 0$

1. Perform a lasso regression of Y on Z for an appropriate value of λ
2. keep $\hat{\mathcal{A}}_\lambda = \{j : \hat{\gamma}_{j,\lambda} \neq 0\}$
3. Do the flipscores test considering $Z_{\hat{\mathcal{A}}_\lambda}$

This is asymptotically valid

Problem: we are counting on the screening property

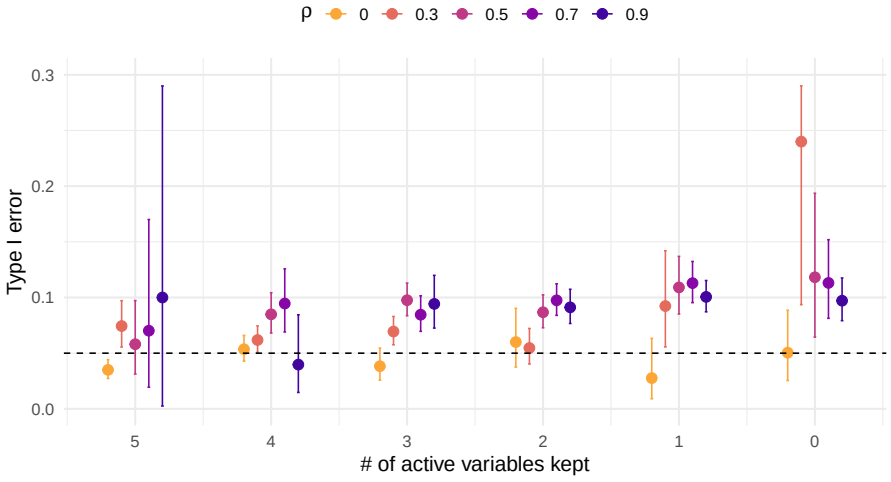
Screening property

Let $\mathcal{A} = \{j : \beta_j \neq 0\}$ be the active set. Then

$$\mathbb{P}(\mathcal{A} \subseteq \hat{\mathcal{A}}_\lambda) \to 1, \quad \text{as } n \to \infty$$

Lack of screening property

Simulation set-up: $n = 50$, $p = 100$, $SNR = 4$, $|\mathcal{A}| = 5$, variables in $[x, Z]$ equicorrelated with correlation $\rho \in \{0, 0.3, 0.5, 0.7, 0.9\}$



Alternative approach with stepwise selection

- Variable selection with the lasso is known to be unstable when the predictors are highly correlated $\rightarrow$ the screening property is often non-satisfied
- Alternative approach: instead of the lasso regression do a stepwise selection on the original variables
- It provides a valid test when the active variables enter the model first

Inference on the full vector of coefficients

The aim is to perform inference on the full vector of coefficients $\{\beta_1, \ldots, \beta_p\}$. The considered model is

$$Y = X^{full}\beta + \varepsilon$$

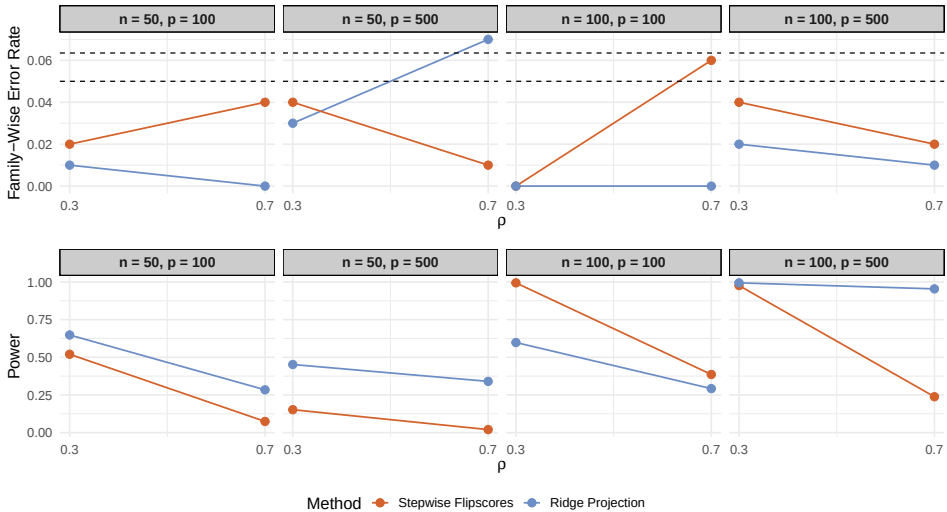
Our proposed strategy is:

- for $j = 1, \ldots, p$,
  - consider as variable of interest x_j and as confounders matrix $Z = X_{-j}^{full}$
  - perform B sign-flipping transformations and get the vector of observed and standardized flipscores test statistics $\{S_j^{obs}, S_j^2, \ldots, S_j^B\}$
- perform a multiple testing correction (for example maxT) and get adjusted p-values

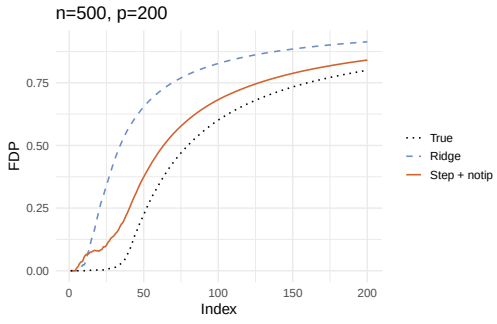
This strategy has a high computational load

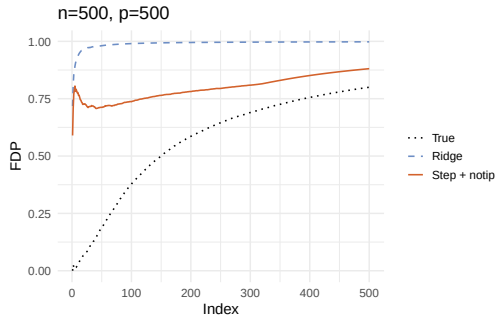
Simulation results - FWER and power

Simulation results for varying levels of n, p and ρ (correlation among x and Z)



Simulation results - FDP





References

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), 1212–1242.

De Santis, R., Goeman, J. J., Hemerik, J., Davenport, S., & Finos, L. (2025). Inference in generalized linear models with robustness to misspecified variances. *Journal of the American Statistical Association*, (pp. 1–10).

Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, (pp. 533–558).

Hemerik, J., Goeman, J. J., & Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3), 841–864.

Hemerik, J., Thoresen, M., & Finos, L. (2021). Permutation testing in high-dimensional linear models: an empirical investigation. *Journal of Statistical Computation and Simulation*, 91(5), 897–914.

Van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202.

Idea of proof

Let $Y = x\beta + Z\gamma + \epsilon = x\beta + U\xi + \epsilon$, with $\xi = DV^\top \gamma$, and let $H_{\mathcal{A}_\lambda} = U_{\mathcal{A}_\lambda} U_{\mathcal{A}_\lambda}^\top$.
We can rewrite the numerator of the standardized flipped score statistics as

$$
\begin{aligned}
N &= x^\top (I - H_{\mathcal{A}_\lambda}) F (I - H_{\mathcal{A}_\lambda}) Y \\
&= x^\top (I - H_{\mathcal{A}_\lambda}) F (I - H_{\mathcal{A}_\lambda}) (Z_{\mathcal{A}_\lambda} \gamma_{\mathcal{A}_\lambda} + Z_{\mathcal{A}_\lambda^c} \gamma_{\mathcal{A}_\lambda^c} + \varepsilon) \\
&= x^\top (I - H_{\mathcal{A}_\lambda}) F (I - H_{\mathcal{A}_\lambda}) (Z_{\mathcal{A}_\lambda^c} \gamma_{\mathcal{A}_\lambda^c} + \varepsilon)
\end{aligned}
$$

Idea of proof

Let $Y = x\beta + Z\gamma + \epsilon = x\beta + U\xi + \epsilon$, with $\xi = DV^\top\gamma$, and let $H_{\mathcal{A}_\lambda} = U_{\mathcal{A}_\lambda}U_{\mathcal{A}_\lambda}^\top$.
We can rewrite the numerator of the standardized flipped score statistics as

$$\begin{aligned} N &= x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})Y \\ &= x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})(Z_{\mathcal{A}_\lambda}\gamma_{\mathcal{A}_\lambda} + Z_{\mathcal{A}_\lambda^c}\gamma_{\mathcal{A}_\lambda^c} + \varepsilon) \\ &= x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})(Z_{\mathcal{A}_\lambda^c}\gamma_{\mathcal{A}_\lambda^c} + \varepsilon) \end{aligned}$$

Hence, we can rewrite the standardized flipped score statistic as

$$\begin{aligned} T(F) &= \frac{x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})(Z_{\mathcal{A}_\lambda^c}\gamma_{\mathcal{A}_\lambda^c} + \varepsilon)}{\sigma\sqrt{x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})x}} \\ &= \frac{x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})\varepsilon}{\sigma\sqrt{x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})x}} + \\ &\quad \frac{x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})Z_{\mathcal{A}_\lambda^c}\gamma_{\mathcal{A}_\lambda^c}}{\sigma\sqrt{x^\top(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})x}} = T_1(F) + T_2(F) \end{aligned}$$

Idea of proof II

- Under the assumption that $\lim_{n\to\infty} \|(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})X\|_\infty / \|(I - H_{\mathcal{A}_\lambda})F(I - H_{\mathcal{A}_\lambda})X\|_2 = 0$, by the Lindeberg-Feller CLT and the dominating convergence theorem, $T_1(F) \xrightarrow{d} N(0,1)$
- Under the assumption $\|Z_{\mathcal{A}_\lambda^c}\gamma_{\mathcal{A}_\lambda^c}\|_2 = o(1)$, $T_2(F) = o(1)$,

so $T(F) \xrightarrow{d} N(0,1)$.

Idea of proof III

- Since the sign-flipped matrices are all independent, the resulting score statistics are independent, i.e., for $i \neq j$ $Cov(T(F_i), T(F_j)) = 0$.
- This implies that the vector $(T(I), \ldots, T(F_w))$ converges to a vector of iid random variables $N(0, 1)$
- For lemma 1 in Hemerik et al. (2020) the test that rejects H_0 when $T_{obs} > T_{\lceil (1-\alpha)w \rceil}$ is an asymptotically valid α level test.