Resampling-Based Inference for High-Dimensional Regression

Inferenza Tramite Ricampionamento per Regressioni ad Alta Dimensionalità

Anna Vesely and Jelle J. Goeman and Angela Andreella and Livio Finos

Abstract We propose a novel procedure for resampling-based multiple testing in high-dimensional regression. First, we construct permutation test statistics for each individual hypothesis by means of repeated random splits of the data. In each split, half of the observations is used to perform variable selection, and half to build test statistics for the selected variables. Then we define an asymptotically exact test for any subset of hypotheses by aggregating the individual statistics through a suitable function, e.g., maximum or weighted sums. The procedure is flexible, allowing different selection techniques and combining functions. It can be embedded into closed testing methods to make simultaneous confidence statements on the proportion of true discoveries (TDP) of all subsets, valid even under post-hoc selection.

Abstract *Si propone un metodo basato sul ricampionamento per test multipli in regressioni ad alta dimensionalità. Si definiscono statistiche test per ogni singola ipotesi tramite partizioni casuali dei dati, in cui metà delle osservazioni sono usate per selezionare le variabili, e metà per costruire statistiche per le variabili selezionate. Aggregando tali statistiche, ad es. con il massimo o somme pesate, si ottiene un test asintoticamente esatto per ogni sottoinsieme di ipotesi. La procedura è flessibile, poiché ammette diverse tecniche di selezione e diverse funzioni di combinazione. Può essere utilizzata all'interno di metodi di closed testing per ottenere*

Anna Vesely
Department of Developmental Psychology and Socialization, University of Padua, Italy, e-mail: anna.vesely@unipd.it

Jelle J. Goeman
Biomedical Data Sciences, Leiden University Medical Center, The Netherlands, e-mail: j.j.geoman@lumc.nl

Angela Andreella
Department of Economics, University of Venice, Italy, e-mail: angela.andreella@unive.it

Livio Finos
Department of Developmental Psychology and Socialization, University of Padua, Italy, e-mail: livio.finos@unipd.it

limiti di confidenza per la proporzione di variabili attive (TDP), simultaneamente su tutti i sottoinsiemi di ipotesi.

Key words: high-dimensional regression, multiple testing, Multisplit, resampling-based test, true discovery proportion

1 Introduction

In linear regression, interest usually lies in discovering relevant predictor variables and assessing statistical significance; however, many challenges arise in high-dimensional settings. Researchers are often interested in studying subsets of variables with an exploratory approach, quantifying activation inside. Moreover, when they do not know a priori which subsets they are interested in, they may want to study many and make the selection post hoc.

We propose a multiple testing method for high-dimensional linear regression that defines a test for any subset of variables and ensures asymptotic error control. We use the permutation framework, which is often more powerful than the parametric approach, especially when considering multiple hypotheses [2]. The method relies on two building blocks: the Multisplit proposed by [4], that computes adjusted p-values for all variables using variable selection techniques and repeated splits of the data, and the sign-flipping test given in [3].

First, we construct permutation test statistics for all variables. Then we aggregate these individual statistics to define an asymptotically exact test for any subset of variables. Different combining functions are possible, including the maximum and weighted sums. As we can test any subset, the procedure can be embedded into closed testing methods that give simultaneous confidence sets for the true discovery proportion (TDP), such as [1] and [6]. This way we are able to provide confidence statements on the TDP of all subsets, valid even under post-hoc selection.

The structure of the paper is the following. We introduce the model and its assumptions in Sect. 2, then we define the method in Sect. 3. Finally, in Sect. 4 we compare the proposed method and the Multisplit through simulations.

2 High-Dimensional Linear Regression

We consider a linear regression framework with n observations and m variables, potentially high-dimensional ($n < m$). The model is

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I) \tag{1}$$

where $\mathcal{N}_n$ denotes the multivariate normal distribution, and $I \in \mathbb{R}^{n \times n}$ is the identity matrix. Here $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times m}$ is a fixed design matrix,

$\beta \in \mathbb{R}^m$ is the vector of coefficients and $\varepsilon \in \mathbb{R}^n$ is a random error vector. We assume that X has rank m, and $X^\top X/n$ converges to a finite positive semi-definite matrix as $n \to \infty$.

We are interested in exploring which variables in X are active, meaning that they have non-null coefficients. Let $M = \{1,\dots,m\}$ be the set of variable indices, and $N = \{j \in M : \beta_j = 0\}$ the unknown subset corresponding to inactive variables. For any $j \in M$, we may define the null hypothesis $H_j : \beta_j = 0$, that is true when $j \in N$, regardless of the value of other variables' coefficients. We want to study more variables taken together, i.e., test intersection hypotheses of the form

$$H_S = \bigcap_{j \in S} H_j : \beta_j = 0 \text{ for all } j \in S \qquad (S \subseteq M,\ S \neq \emptyset)$$

with significance level $\alpha \in [0,1)$. H_S is true if all variables in S are inactive ($S \subseteq N$).

To test any H_S, we will rely on a variable selection procedure that estimates the set of active variables with $A \subseteq M$. As in [4], we assume the following properties.

Sparsity $\quad |A| \leq n/2$.
Screening property $\quad \lim_{n\to\infty} P(M \setminus N \subseteq A) = 1$.

The ideal selection procedure, for which the screening property always holds, is an oracle method that selects all truly active variables, plus eventually some others. Even though this is not available in practice, it can be used in simulations to show the performance of the proposed method when the assumptions are ensured. When studying real data, we suggest using the Lasso [5] with a suitable calibration of the λ parameter, so that it selects enough variables for the screening property to be likely. If m_1 is an estimate of the expected number of active variables, we recommend choosing λ so that the Lasso selects $\min(2m_1, n/2)$ variables.

3 Resampling-Based Multisplit

We propose an asymptotically exact test for any intersection hypothesis H_S corresponding to a non-empty set $S = \{j_1,\dots,j_s\} \subseteq M$. The method will efficiently construct permutation test statistics for H_S by combining statistics for the individual hypotheses H_j with $j \in S$. We take as combining function any $g : \mathbb{R}^s \longrightarrow \mathbb{R}$ which is increasing in each argument, such as the maximum or (weighted) sums.

To define permutation test statistics for all individual hypotheses H_j, we use Q random splits of the data and B random sign-flipping transformations. The values of Q and B do not need to grow with m; larger values of B tend to give more power, but to have non-zero power we only need $B \geq 1/\alpha$ [2]. Hence fix B diagonal sign-flipping matrices $F_1,\dots,F_B \in \mathbb{R}^{n\times n}$, where $F_1 = I$ is the identity, while the diagonal elements of the other matrices are independently and uniformly drawn from $\{-1,1\}$. As in [4], for each split q we randomly partition observations into two equally-sized subsets $\mathscr{D}_0^q$ and $\mathscr{D}^q$, and we use $\mathscr{D}_0^q$ to estimate the set of active variables with $A^q \subseteq M$. Then we use $\mathscr{D}^q$ and A^q to compute test statistics similarly to [3], as follows.

For each split q, we restrict the design matrix X to observations in $\mathscr{D}^q$ and variables in A^q, obtaining $X^q = X_{\mathscr{D}^q,A^q}$. For each selected variable $j \in A^q$, we define X^q_{-j} as X^q without the j-th column, then we construct the split's residual maker matrix

$$R^q_{-j} = 0 \in \mathbb{R}^{n\times n} \qquad \text{except} \qquad R^q_{-j;\mathscr{D}^q,\mathscr{D}^q} = I - X^q_{-j}(X^{q\top}_{-j}X_{-j})^{-1}X^{q\top}_{-j}$$

where all elements are zero except those corresponding to observations in $\mathscr{D}^q$.

We give statistics for all variables by aggregating information over the Q splits:

$$C_{j,b} = \sum_{q\,:\,j\in A^q} R^q_{-j}F_bR^q_{-j} \in \mathbb{R}^{n\times n}$$

$$T^b_j = \begin{cases} 0 \text{ if } \|X_j^\top C_{j,b}\| = 0 \\ \|X_j^\top C_{j,b}\|^{-1}|X_j^\top C_{j,b}Y| \text{ otherwise} \end{cases} \qquad (j \in M,\ b \in \{1,\dots,B\})$$

where X_j is the j-th column of X.

To test H_S, it is sufficient to combine the individual statistics T^b_j as

$$T^b_S = g\left(T^b_{j_1},\dots,T^b_{j_s}\right) \qquad (b \in \{1,\dots,B\}).$$

As a critical value we take $T_S^{(\lceil(1-\alpha)B\rceil)}$, where $T_S^{(1)} \leq \ldots \leq T_S^{(B)}$ are the sorted statistics, and $\lceil\cdot\rceil$ denotes the ceiling function.

Theorem 1. *The test that rejects H_S when $T^1_S > T_S^{(\lceil(1-\alpha)B\rceil)}$ is asymptotically an α-level test.*

Proof. Assume that H_S is true, and consider any couple of variables $j,h \in S$, any transformation b, and any split q. Suppose that the variable selection procedure selects all active variables; by the screening property, this is true at least asymptotically, so this assumption does not affect asymptotic results. As H_j is true and A^q contains all active variables, we can write

$$Y_{\mathscr{D}^q} = X_{-j;\mathscr{D}^q,A^q}\beta_{-j;A^q} + \varepsilon_{\mathscr{D}^q}, \qquad \varepsilon_{\mathscr{D}^q} \sim \mathscr{N}_{n/2}(0,\sigma^2 I).$$

For any selected variable $j \in A^q$, the matrix R^q_{-j} has non-null elements only corresponding to observations in $\mathscr{D}^q$, and so the effective score for this model is

$$V^{qb}_j = \frac{1}{\sqrt{n}}X_j^\top R^q_{-j}F_bR^q_{-j}Y = V^{*qb}_j + o_P(1), \qquad V^{*qb}_j = \frac{1}{\sqrt{n}}X_j^\top R^q_{-j}F_b\varepsilon$$

(see Theorem 2 in [3]). Hence the sB-dimensional vectors

$$\mathbf{V}_S = (V^1_{j_1},\dots,V^B_{j_1},\dots,V^1_{j_s},\dots,V^B_{j_s})^\top, \qquad V^b_j = \sum_{q\,:\,j\in A^q} V^{qb}_j$$

$$\mathbf{V}^*_S = (V^{*1}_{j_1},\dots,V^{*B}_{j_1},\dots,V^{*1}_{j_s},\dots,V^{*B}_{j_s})^\top, \qquad V^{*b}_j = \sum_{q\,:\,j\in A^q} V^{*qb}_j$$

are asymptotically equivalent. Notice that T_j^b is the standardization of $|V_j^b|$. Since

$$\mathbf{V}_S^* \xrightarrow[n\to\infty]{\mathrm{d}} Z \sim \mathscr{N}_{sB}(0, \Xi \otimes I)$$

$$I \in \mathrm{IR}^{B\times B}, \qquad \Xi = (\xi_{k\ell}) \in \mathrm{IR}^{s\times s}, \qquad \xi_{k\ell} = \sigma^2 \lim_{n\to\infty} \frac{1}{n} X_{j_k}^\top C_{j_k,1} C_{j_\ell,1} X_{j_\ell}$$

the B vectors $(V_{j_1}^1, \dots, V_{j_s}^1), \dots, (V_{j_1}^B, \dots, V_{j_s}^B)$ converge to i.i.d. random vectors, and so do the vectors $(T_{j_1}^1, \dots, T_{j_s}^1), \dots, (T_{j_1}^B, \dots, T_{j_s}^B)$. Therefore the combinations of their elements $T_S^1, \dots, T_S^B$ converge to i.i.d. random variables. Moreover, high values of T_S^1 correspond to evidence against H_S. From Lemma 1 in [3],

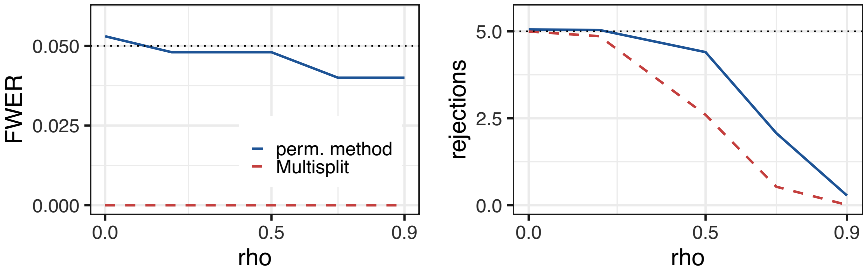
$$\lim_{n\to\infty} P\left(T_S^1 > T_S^{(\lceil(1-\alpha)B\rceil)}\right) = \frac{\lfloor \alpha B \rfloor}{B} \leq \alpha.$$

□

To summarize, we have constructed permutation test statistics for all variables in high-dimensional regression, that are sufficient to define an asymptotically exact test for any intersection hypothesis H_S. The test is obtained combining the statistics for variables in S through any function g that is increasing in each argument.

4 Simulations

We compare the proposed method with the Multisplit, using simulation settings similar to those in [4], with $n = m = 100$. We simulate X from a centered multivariate normal distribution with $\mathrm{cov}(X_j, X_h) = \rho^{|j-h|}$ for $j, h \in M$, taking $\rho \in \{0, 0.2, 0.5, 0.7, 0.9\}$. We compute Y as in (1), where $\beta = (1, \dots, 1, 0, \dots, 0)$ has 5 non-null elements, and σ is such that the signal-to-noise ratio is 4. We take $\alpha = 0.05$,



(a) FWER. The dotted line corresponds to the significance level α.

(b) Number of rejections. The dotted line denotes the number of active variables.

Fig. 1: Results by covariance parameter ρ.

$Q = 50$, $B = 200$, and an oracle selection that returns 10 variables. We simulate data 1000 times and study the set of all variables. For the proposed method, we correct for multiplicity with the maxT-method [8], corresponding to $g = \max$.

Results are shown in Fig. 1. Both methods control the FWER, computed as the proportion of simulations where at least one true null hypothesis is rejected. In terms of rejections, the proposed method is equivalent to the Multisplit when the covariance parameter ρ is small, and more powerful in all the other scenarios.

5 Discussion

We have considered the problem of testing multiple hypotheses in high-dimensional linear regression. Our proposed approach provides asymptotically valid resampling-based tests for any subset of hypotheses, which can be employed within closed testing procedures to make confidence statements on the number of active predictor variables (TDP) within any set. These confidence statements are valid even when the subsets of interest are chosen post hoc, after seeing the data.

First, we have provided a procedure that repeatedly splits the data into two random subsets, using the first to select variables and the second to build permutation test statistics for each variable. Then we have shown that statistics for any intersection hypothesis can be defined by aggregating individual statistics with any function which is increasing in each argument, including the maximum and weighted sums. Our method is extremely flexible, allowing different selection procedures and combining functions. Preliminary simulations show a considerable increase in power over the Multisplit [4]. An implementation of the procedure is available in [7].

References

1. Goeman, J.J., Solari, A.: Multiple testing for exploratory research. Stat. Sci. **26**(4), 584–597 (2011)
2. Hemerik, J., Goeman, J.J.: False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. J. R. Stat. Soc. Series B Stat. Metodol. **80**(1), 137–155 (2018)
3. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. R. Stat. Soc. Series B Stat. Metodol. **82**(3), 841–864 (2020)
4. Meinshausen, N., Meier, L., Bühlmann, P.: p-values for high-dimensional regression. J. Am. Stat. Assoc. **104**(488), 1671–1681 (2009)
5. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Series B Stat. Metodol. **58**(1), 267–288 (1996)
6. Vesely, A., Finos, L., Goeman, J.J.: Permutation-based true discovery guarantee by sum tests. Pre-print arXiv:2102.11759 (2021)
7. Vesely, A.: splitFlip: Resampling-based Multisplit. https://github.com/annavesely/splitFlip. R package version 1.1.0 (2021)
8. Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York (1993)