Angela Montanari

Variable screening in high dimensional regression via random projection ensembles

Angela Montanari[1*]; Laura Anderlucci[1]; Matteo Farnè[1]; Giuliano Galimberti[1]

[1] University of Bologna, Dept. of Statistical Sciences, via Belle Arti 41, Bologna, Italy – angela.montanari@unibo.it; laura.anderlucci@unibo.it; matteo.farne@unibo.it; giuliano.galimberti@unibo.it

Abstract:
In this paper we propose a variable selection method for multiple linear regression which is based on ensembles of axis-aligned random projections and accounts for partial correlation between each predictor and the response. Performances of the proposed method are evaluated on simulated and real data.

Keywords:
Multiple regression; axis-aligned random projections; sparsity; variable ranking; prediction accuracy

1. Introduction:

It is well known that, when dealing with high dimensional data, most of the classical multivariate methods cannot be applied or give unreliable results and it is known as well that when the number of observed variables p is large the relevant information may be contained in an s-dimensional subset of the observed variables.

In the context of multiple linear regression this means that the vector of regression coefficients for the model involving all the p variables is sparse. The ordinary approach for variable selection based on stepwise methods has turned out to produce very unstable results and new alternative solutions have recently appeared in the literature. The problem, for instance, has been addressed by either directly applying L_1 norm regularization to the original data (Tibshirani,1996) or by screening the variables to identify the most relevant ones and then applying an L_1 penalty to the selected subset (Fan, J. & Lv, J.,2008). The reasons for this two-step approach lie in the high computational load inherent in the penalized approach.

In this paper we propose a new method for variable selection in multiple linear regression which is based on random projections of the covariates. The use of random projections to reduce the dimensionality of a data set is becoming increasingly popular in the multivariate statistical literature. The common trait of the most effective solutions consists in randomly combining the p columns of the data matrix $\boldsymbol{X}$, thus mapping the data onto a random d-dimensional (with $d \ll p$) subspace on which classical analyses can be performed. The results obtained on different random projections are then summarized by ensemble methods in order to obtain the final estimates. Successful applications include supervised classification (Cannings, T.I. & Samworth, R.J., 2017), large covariance estimation (Marzetta *et al.*, 2011), large-scale regression (Thanei *et al.*, 2017) and sparse principal components (Gataric *et al.*, 2020).

2. Methodology - Predictor selection via Random Projections:

In our proposal we exploit the special feature of axis-aligned random projections, which represent a fast and analytically tractable way to perform random variable selection.
Given a data matrix $\boldsymbol{X}$ we consider $\boldsymbol{XA}$ where $\boldsymbol{A}$ is a $p \times d$ axis aligned random matrix.
The least squares problem is than rephrased in terms of $\boldsymbol{XA}$ as

$$\boldsymbol{b_A} = arg \min_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{XAb}\|$$

Where the vector $\boldsymbol{y}$ includes the observed values on the response variable.
Many different $\boldsymbol{A}$ matrices are considered. In particular we consider B_1 sets composed by B_2 random projections each and within each block of B_2 projections we chose the one for which the fitted regression model shows the largest R^2.
As the matrix $\boldsymbol{A}$ is axis aligned, only a few variables will contribute to $\boldsymbol{b_A}$ in each selected projection but, combining the models fitted in all the B_1 top projections, we can obtain a ranking of the p variables; after cutting the ranking at the assumed sparsity level s we identify the most relevant predictors for y. The pseudo code is displayed in Figure 1. Theoretical aspects related to the properties of the proposed method are also analysed.

3. Results:

To study and to evaluate the performance of the proposed method, we partially reproduce the numerical study of Fan, J. & Lv, J. (2008). In particular, Fan and Lv consider two main scenarios to validate their Sure Independence Screening (SIS) method: independent and correlated features. In addition, the prediction accuracy of our method is computed on a real dataset.

Simulation I: independent features. The first scenario considers a linear model with IID standard Gaussian predictors and Gaussian noise with standard deviation σ=1.5. Two settings with (n, p)=(200,1000) and (n, p)=(500,2000) are considered. The number s of relevant predictors is 8 and 18, and the corresponding non-zero coefficients are randomly chosen as follows.
Let us set $a = 4 \log(n)/n^{1/2}$ and $5 \log(n)/n^{1/2}$, respectively; the non-zero coefficients are of the form $(-1)^u a|z|$ for each model, where u is drawn from a Bernoulli distribution with parameter 0.4 and z is drawn from the standard Gaussian distribution. In particular, the L_2-norms of $\boldsymbol{\beta}$ in the two simulated models are 6.695 and 9.582. For each model 100 data sets are simulated; the size of the projected space d is set to 10 and 500 blocks of 50 axis-aligned projections each are considered.
In order to facilitate the comparison with the results of Fan and Lv, Figure 2 reports the distribution of the minimum number of variables to be selected in order to include the true model. More than the 70% of the datasets ranked the relevant variables as first. Such results clearly outperform those of SIS reported in Figure 5 (a), page 862 of Fan, J. & Lv, J. (2008).

Simulation II: dependent features. The scenario with dependent features considers three settings with (n, p, s) equal to (200,1000,5), (200,1000,8) and (800,2000,14), s denoting the number of non-zero coefficients.
The three p-vectors $\boldsymbol{\beta}$ are generated in the same way as in simulation I. Let's set $(\sigma,a) = (1,\ 2 \log(n)/n^{1/2})$, $(1.5,\ 4 \log(n)/n^{1/2})$, $(2,\ 4 \log(n)/n^{1/2})$.
In particular, the L_2-norms of $\boldsymbol{\beta}$ in the three simulated models are 3.618, 6.696 and 6.788.
To introduce correlation between predictors, an $s \times s$ symmetric positive definite matrix $\boldsymbol{C}$ was generated with condition number about $n^{1/2}/\log(n)$; samples of s predictors $X_1, \ldots, X_s$ are then generated from $N(0,\boldsymbol{C})$. The remaining predictors are taken as $X_i = Z_i + (1 - r)\, X_1$, $I = 2s + 1, \ldots, p$, with $r = 1 - 4 \log(n)/p$, $1 - 5 \log(n)/p$ and $1 - 5 \log(n)/p$, being $Z_{s+1},\ldots,Z_p \sim N(0,\boldsymbol{I}_{p-s})$.
For each model 100 data sets are simulated; the size of the projected space d is set to 10, B_1=500, B_2=50.

Figure 3 includes the distribution of the minimum number of selected variables that is required to include the true model: compared with the independent case, the algorithm requires a larger model size; however, such number is still very limited, particularly if compared with that of SIS (see Figure 6 (a)-(b), page 863 of Fan, J. & Lv, J. (2008)).

Data: standardized data matrix $\mathbf{X}_{n\times p}$ and response vector $\mathbf{y}$
d = dimension of each random model;
B_1= dimension of the ensemble;
B_2= no. of random models within each block;
for i *in* $1,\ldots,B_1$ **do**
  for j *in* $1,\ldots,B_2$ **do**
    generate axis-aligned random matrix $\mathbf{A}_{p\times d}$;
    compute $\tilde{\mathbf{X}}_{n\times d} = \mathbf{X}\mathbf{A}$;
    compute R^2 of the regression model $\hat{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{b}$;
  end
  retain the regression coefficients $\mathbf{b}_i^*$ corresponding to the model with the largest R^2 and set the other $(p-d)$ to zero;
end
average the $|\mathbf{b}_i^*|$s and retain the s predictors with maximum values.

Figure 1. Pseudo-code of the proposed algorithm for variable screening.

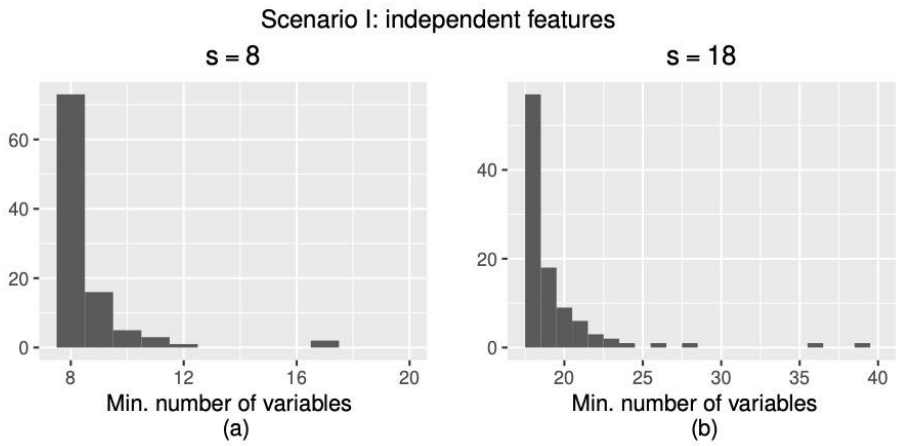

Figure 2. Scenario 1: Distribution of the minimum number of selected variables that is required to include the true model when (a) n=200 and p=1000 and (b) n=800 and p=2000.

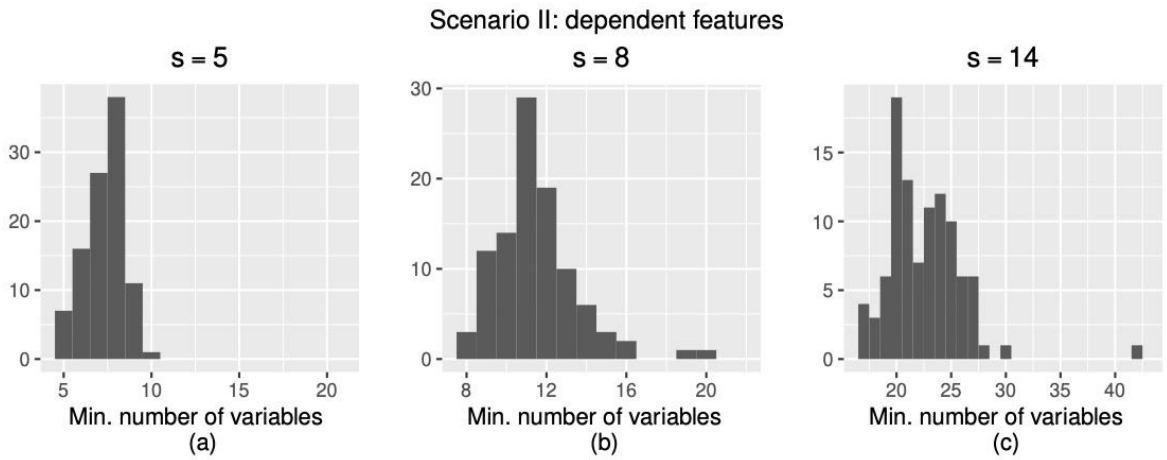

Figure 3. Scenario 2: Distribution of the minimum number of selected variables that is required to include the true model when (a)-(b) n=200 and p=1000 and (c) n=800 and p=2000.

Real data: **Twitter social media buzz.** The dataset (from UCI Machine Learning Repository) includes 8000 observations; the goal is to predict the popularity of topics on Twitter as quantified by its mean number of active discussions given 1378 predictor variables (e.g. number of authors contributing to the topic over time, average discussion lengths, number of interactions between authors etc). The aim is find a subset of predictors that is relevant for the prediction of the response.
The variable screening of the Twitter dataset was carried out by both sparse linear regression via random projections ensembles and Sure Independence Screening, in 4-fold Cross-Validation. The former run with B_1 = 1000, B_2 = 50 and d = 20.
At the end of the procedure, the first 150 predictors were retained for both methods and the mean square errors (MSE) computed. The top 150 predictors of both methods have then been further filtered out by the Lasso. The results are reported in Table 1.

Table 1. Real data example. Prediction accuracy of the regression models that retain the first 150 predictors (MSE_{150}) and after employing the lasso (MSE – Lasso). Last column reports the cross-validation size of each model (Avg size (sd) – Lasso).

	MSE_{150}	MSE - Lasso	Avg size (sd) - Lasso
lm-RPE	27409.19	31824.75	70 (36.31)
SIS	42232.19	31860.85	78.75 (21.82)

4. Discussion and Conclusion:

This paper presents a novel approach to sparse linear regression via Random Projections that accounts for partial correlation between predictors; as the simulation studies and the analisys of real data highlight, the proposed method improves upon SIS which only considers marginal correlations. The optimal choice of the tuning parameters, B_1, B_2, d, and the estimation of s are object of ongoing research.

References:

1. Cannings, T.I. and Samworth, R.J. (2017), Random-projection ensemble classification. J. R. Stat. Soc. B, 79: 959-1035. https://doi.org/10.1111/rssb.12228
2. Fan, J. and Lv, J. (2008), Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70: 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

3. Gataric M., Wang T. & Samworth R.J. (2020), Sparse principal component analysis via axis-aligned random projections. J. R. Stat. Soc. B, 82: 329-359. https://doi.org/10.1111/rssb.12360
4. Marzetta T. L., Tucci G. H. & Simon S. H. (2011), A Random Matrix-Theoretic Approach to Handling Singular Covariance Estimates, IEEE Transactions on Information Theory, vol. 57, no. 9, pp. 6256-6271, doi: 10.1109/TIT.2011.2162175.
5. Thanei G.A., Heinze C. & Meinshausen N. (2017) Random Projections for Large-Scale Regression. In: Ahmed S. (eds) Big and Complex Data Analysis. Contributions to Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-41573-4_3
6. Tibshirani, R. (1996), Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58: 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x