An ensemble learning method for variable selection

Vincent Audigier, Avner Bar-Hen

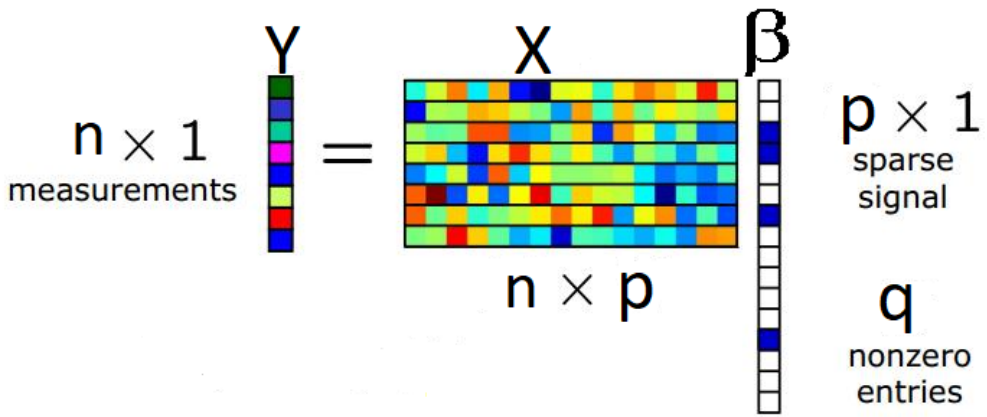
CNAM, MSDMA team, Paris

Journées de la statistique 2018

Context

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I})$$

- β **sparse** with q non-zeros



- **Goal: select a set of features X_j that are likely to be relevant to the response Y**
- Many selection methods with different properties (Stepwise, Lasso, Knockoff, etc)

Challenges

- **Stability** of the selected subset

  → Ensemble method (Genuer et al., 2010; Meinshausen and Bühlmann, 2010)

- **High dimensional data**

  → Shrinkage or preliminary sreening step (Tibshirani, 1996; Wasserman and Roeder, 2009; Foygel Barber and Candes, 2016)

- **Missing data**

  → Multiple imputation (Zhao and Long, 2017)

Issues frequently occur simultaneously!

Proposed algorithm: ensemble regression

Create regression instances

- Sample k variables among the p variables
- Handle missing values (e.g. by stochastic imputation)
- For a given threshold, apply selection procedure (e.g. knockoff) to decide which of the k variables are significantly related to Y.
- Iterate the process B times ($\Rightarrow$ B regression instances).

Aggregate the regression instances

- $r_i = \dfrac{\#\text{ times the variable } X_i \text{ is selected}}{\#\text{ time } X_i \text{ is present in the instances}}$
- Conclude that X_i is significantly related to Y if $r_i > r$.

Properties

- The procedure inherits from the properties of the selection method used (consistency, error rates, ...)
- Improves stability repeating the selection procedure
- Overcomes the high dimensional setting by sampling a subset of variables
- Handles missing value by standard methods (by single stochastic imputation or complete case for small subsets)
- Several parameters to tune

  k number of sampled variables
  B number of regression instances
  r threshold

Choice of k: complete data

k: number of sampled variables

- if the selection method does not handle high dimensional data, choose $k < n$
- choose k **not too small**...

With all explanatory variables

$$\mathbb{V}(Y \mid X) = \sigma^2$$

After a draw of k variables $(X_{j_1}, \ldots, X_{j_k})$

$$\mathbb{V}(Y \mid X_{j_1}, \ldots, X_{j_k}) = \mathbb{V}(\beta_{j_{k+1}} X_{j_{k+1}} + \ldots + \beta_{j_p} X_{j_p}) + \sigma^2$$

... **since missing significant variables increase the noise** in the regression scheme

Choice of k: incomplete data

Aim: cardinal of complete observations $>$ number of variables

- required for complete-case
- useful for sequential imputation

Controlling the probability to face a high dimensional setting

- given a subset of k variables $j_1, ..., j_k$, it is easy to show the **number of complete individuals** $N_{j_1,j_2,...,j_k}$ follows a $\mathcal{B}(n, (1 - \delta_{j_1,j_2,...,j_k}))$
- with $\delta_{j_1,j_2,...,j_k}$ the probability to observe, for each individual, at least one missing value on the subset
- Chernoff inequality:

$$\mathbb{P}(N_{j_1,j_2,...,j_k} < k) \leq \exp\left(-\left(1 - \frac{k}{n(1-\delta_{j_1,j_2,...,j_k})^k}\right)^2 (1-\delta_{j_1,j_2,...,j_k})^k \, n/2\right)$$

Choice of B

B : number of iterations = number of regression instances

- We want all variables are selected, with high probability, in at least $\tilde{B}$ regression instances
- Z_i number of regression instances that contains X_i.
- $Z_i \sim \mathcal{B}(B, k/p)$
- $\mathbb{P}(\min_{i=1,\dots,p} Z_i < \tilde{B}) < p \exp\left(-\left(1 - \frac{p\tilde{B}}{Bk}\right)^2 Bk/(2p)\right)$

$\Rightarrow$ B and p strongly related

Choice of r

r : threshold for selection

Aim : few false positive and few false negative

- r_i : relative frequency of selection for the variable X_i
- r_i can be seen as an estimate of the power (under the hypothesis $\beta_i \neq 0$) or of the α risk (under the hypothesis $\beta_i = 0$)
- Two relevant choices:
  - $r_i > r \Rightarrow \beta_i \neq 0$
  - $r_i < r \Rightarrow \beta_i = 0$

Simulations

- $n = 200$ observations
- $p = 100$ or $p = 300$ variables
- Number of non-zero: $q = 8$
- SNR = 2 ou 4 SNR : $\left(\frac{\text{Var}(Y)-\text{Var}(\varepsilon)}{\text{Var}(\varepsilon)}\right)$
- $\text{Cor}(X_i, X_j) = 0$ or $\text{Cor}(X_i, X_j) = 0.4$
- Missing values on covariates:
  - No
  - MCAR : $\mathcal{B}(0.2)$
  - MAR : Probit $\phi(\beta_0 + Y)$, with β_0 chosen so that 20% of values are missing (in expectation)

Results : $n > p$ complete case

			Algorithm[1]			Full		
rho	snr	methode	ok	-	+	ok	-	+
0	2	Knockoff	7.27	0.73	1.80	7.43	0.57	0.96
0	4	Knockoff	7.68	0.32	1.61	7.78	0.22	1.10
0.4	2	Knockoff	5.39	2.61	3.71	3.92	4.08	0.69
0.4	4	Knockoff	6.45	1.55	2.76	6.24	1.76	0.91
0	2	Lasso	7.54	0.46	1.23	8.00	0.00	18.75
0	4	Lasso	7.74	0.26	1.11	8.00	0.00	17.89
0.4	2	Lasso	5.49	2.51	3.10	7.64	0.36	17.61
0.4	4	Lasso	7.19	0.81	3.56	8.00	0.00	17.64
0	2	Stepwise	6.66	1.34	0.12	8.00	0.00	29.87
0	4	Stepwise	7.36	0.64	0.15	8.00	0.00	28.35
0.4	2	Stepwise	3.82	4.18	0.96	7.20	0.80	29.90
0.4	4	Stepwise	5.77	2.23	1.26	7.90	0.10	30.70

[1] $(k = 6, B = 3000)$

Results : $n > p$ complete case

			Algorithm[1]			Full		
rho	snr	methode	ok	-	+	ok	-	+
0	2	Knockoff	7.27	0.73	1.80	7.43	0.57	0.96
0	4	Knockoff	7.68	0.32	1.61	7.78	0.22	1.10
0.4	2	Knockoff	5.39	2.61	3.71	3.92	4.08	0.69
0.4	4	Knockoff	6.45	1.55	2.76	6.24	1.76	0.91
0	2	Lasso	7.54	0.46	1.23	8.00	0.00	18.75
0	4	Lasso	7.74	0.26	1.11	8.00	0.00	17.89
0.4	2	Lasso	5.49	2.51	3.10	7.64	0.36	17.61
0.4	4	Lasso	7.19	0.81	3.56	8.00	0.00	17.64
0	2	Stepwise	6.66	1.34	0.12	8.00	0.00	29.87
0	4	Stepwise	7.36	0.64	0.15	8.00	0.00	28.35
0.4	2	Stepwise	3.82	4.18	0.96	7.20	0.80	29.90
0.4	4	Stepwise	5.77	2.23	1.26	7.90	0.10	30.70

[1] $(k = 6, B = 3000)$

Results : $n > p$ complete case

			Algorithm[1]			Full		
rho	snr	methode	ok	-	+	ok	-	+
0	2	Knockoff	7.27	0.73	1.80	7.43	0.57	0.96
0	4	Knockoff	7.68	0.32	1.61	7.78	0.22	1.10
0.4	2	Knockoff	5.39	2.61	3.71	3.92	4.08	0.69
0.4	4	Knockoff	6.45	1.55	2.76	6.24	1.76	0.91
0	2	Lasso	7.54	0.46	1.23	8.00	0.00	18.75
0	4	Lasso	7.74	0.26	1.11	8.00	0.00	17.89
0.4	2	Lasso	5.49	2.51	3.10	7.64	0.36	17.61
0.4	4	Lasso	7.19	0.81	3.56	8.00	0.00	17.64
0	2	Stepwise	6.66	1.34	0.12	8.00	0.00	29.87
0	4	Stepwise	7.36	0.64	0.15	8.00	0.00	28.35
0.4	2	Stepwise	3.82	4.18	0.96	7.20	0.80	29.90
0.4	4	Stepwise	5.77	2.23	1.26	7.90	0.10	30.70

[1] $(k = 6, B = 3000)$

Results : $n > p$ incomplete case

			Algorithm[2]			Full		
rho	snr	mech	ok	-	+	ok	-	+
0	2	MCAR	6.25	1.75	0.51	7.43	0.57	0.96
0	2	MAR	6.35	1.65	0.88	7.43	0.57	0.96
0	4	MCAR	7.00	1.00	0.47	7.78	0.22	1.10
0	4	MAR	6.67	1.33	0.75	7.78	0.22	1.10
0.4	2	MCAR	3.52	4.48	1.12	3.92	4.08	0.69
0.4	2	MAR	3.60	4.40	1.51	3.92	4.08	0.69
0.4	4	MCAR	5.29	2.71	1.09	6.24	1.76	0.91
0.4	4	MAR	5.40	2.60	1.57	6.24	1.76	0.91

Table: Resultats pour Knockoff ($n > p$ incomplet)

[2] $k = 10,\ B = 3000$

Results : $n < p$

			Algorithm[3]			Full		
rho	snr	methode	ok	-	+	ok	-	+
0	2	Knockoff	7.19	0.81	2.51	7.94	0.06	1.15
0	4	Knockoff	7.61	0.39	2.63	8.00	0.00	0.91
0.4	2	Knockoff	4.39	3.61	2.92	4.48	3.52	1.00
0.4	4	Knockoff	6.35	1.65	4.31	6.96	1.04	1.08
0	2	Lasso	7.33	0.67	1.83	8.00	0.00	26.44
0	4	Lasso	7.74	0.26	1.99	8.00	0.00	27.98
0.4	2	Lasso	5.21	2.79	2.92	7.40	0.60	20.37
0.4	4	Lasso	6.54	1.46	3.92	7.99	0.01	26.71
0	2	Stepwise	4.57	3.43	0.01	8.00	0.00	35.00
0	4	Stepwise	5.69	2.31	0.02	8.00	0.00	33.45
0.4	2	Stepwise	1.37	6.63	0.09	6.73	1.27	38.11
0.4	4	Stepwise	3.26	4.74	0.25	7.82	0.18	35.39

[3] $(k = 10, B = 10000)$

Conclusions

A new variable selection procedure

- Very simple for applying any variable selection method in complex setting
- Relevant even in a low dimensional setting without missing data

Some limits

- Several tuning parameters
- MAR assumption less plausible with less variables

Perspective

- Estimation of non-zero coefficients

References I

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A): 2178, 2009.

R. Foygel Barber and E. J. Candes. A knockoff filter for high-dimensional selective inference. *ArXiv e-prints*, 2016.

Yize Zhao and Qi Long. Variable selection in the presence of missing data: imputation-based methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5):e1402–n/a, 2017. ISSN 1939-0068. doi: 10.1002/wics.1402. URL http://dx.doi.org/10.1002/wics.1402. e1402.

Results : $n > p$ incomplete case

			Algorithm			Full		
rho	snr	mech	ok	-	+	ok	-	+
0	2	MCAR	5.20	2.80	0.08	8.00	0.00	29.87
0	2	MAR	6.16	1.84	0.92	8.00	0.00	29.87
0	4	MCAR	6.05	1.95	0.11	8.00	0.00	28.35
0	4	MAR	6.43	1.57	0.75	8.00	0.00	28.35
0.4	2	MCAR	2.93	5.07	0.85	7.20	0.80	29.90
0.4	2	MAR	3.53	4.47	1.98	7.20	0.80	29.90
0.4	4	MCAR	5.10	2.90	1.30	7.90	0.10	30.70
0.4	4	MAR	5.34	2.66	2.05	7.90	0.10	30.70

Table: Resultats pour Stepwise ($n > p$ incomplet)

Results : $n > p$ incomplete case

			Algorithm			Full		
rho	snr	mech	ok	-	+	ok	-	+
0	2	MCAR	7.13	0.87	1.48	8.00	0.00	18.75
0	2	MAR	7.33	0.67	3.40	8.00	0.00	18.75
0	4	MCAR	7.50	0.50	1.51	8.00	0.00	17.89
0	4	MAR	7.35	0.65	3.46	8.00	0.00	17.89
0.4	2	MCAR	5.19	2.81	3.58	7.64	0.36	17.61
0.4	2	MAR	5.44	2.56	5.32	7.64	0.36	17.61
0.4	4	MCAR	6.82	1.18	4.11	8.00	0.00	17.64
0.4	4	MAR	6.84	1.16	5.42	8.00	0.00	17.64

Table: Resultats pour Lasso ($n > p$ incomplet)