Sparsity Oriented Importance Learning for High-dimensional Linear Regression

Presentation

By:

Jiahui Dong & Qingyang Fu

INTRO

- Variable importance measure has been an interesting research topic that helps to identify which variables are most important for **understanding**, **interpretation**, **estimation** or **prediction purposes**.
- In this paper, we propose a new variable importance measure, **sparsity oriented importance learning (SOIL)**, for high-dimensional regression from a sparse linear modeling perspective by taking into account the **variable selection uncertainty** via the use of a sensible **model weighting**.

OUTLINE

What:

- What scientific problem is addressed in the paper?
- What is the corresponding mathematical or statistical problem?

Why:

- Why is this problem important?
- Why is this problem challenging?
- Why existing methods can not solve this problem?

How:

- How do the authors tackle this problem?
- How do they justify the proposed method?

What scientific problem is addressed in the paper?

In high-dimensional regression, how to improve **reliability** and **reproducibility** in **model choice**.

What is the corresponding mathematical or statistical problem?

How to make good use of
a **variable importance measure**
to select **variables.**

Why is this problem important?

- Data analysts is **unsatisfied** with **single final model.**
- **Reduce** the list of variables with importance values below a **thresholds**.
- Offering a ranking of variables
- Saving **time** and **cost** in data analysis.
- Helping decision makers to understand the **underlying data process** than trust any single model, and to gain ability to **change or replace** variables.

Why is this problem challenging?

- Variable importance depends on the goal of the analysis and application.

- based on parametric model or nonparametric model.

- Should the importance measure be purely relative to **compare different variables** or should their values have some **meaning on their own**.

Why existing methods can not solve this problem?

- 3 existing method：
  - Simple measure based on final model;
  - LMG:R^2 decomposition;
  - Random Forest
- Simple measure based on Final model: “**Winner takes all**”; Variable selection **uncertainty ignored**; Non-selected variable have **zero importance**.
- LMG(R^2 decomposition): Only deal up to 20 variables
- Random Forest: some **noise variables** receive relatively **large importance values**, even **higher** than almost half of the **true variables**.

How do the authors tackle this problem?

- New variable importance measure: Sparsity Oriented Importance measure (**SOIL**).

- **SOIL**:
  - Two **ingredients**: a **manageable set of models** and a reliable **weighting method** on the models.
  - Measure of importance of the predictors in an absolute scale in [0,1].

SOIL: Sparsity Oriented Importance Learning

Features/advantages:

- Involves **multiple** high-dimensional variable selection methods and **combines** all the **solution path models**.
- Uses external **weighting**(independent of the model selection methods) to avoid bias.
- When the weighting is sensible, the importance of the variables will tend to **0 or 1** as the sample size grows.
- Has excellent performance in the numerical study with satisfying behaviors including **exclusion, inclusion, order preserving, robustness**, etc.

Use of Informative Importance Measures Can Improve the **Reliability** of Data Analysis in Many Ways:

More objective: immediately inspect if "true" variables are missing in the set or unimportant variables are involved.

Finding best model: the most suitable variables for sparse modeling receive higher importance values.

Getting a sense on model selection uncertainty: data analyst will be informed on possible alternative models/covariates.

Helping on the **choice between model selection and model averaging.**

SOIL: General Methodology

- Candidate models:
  - A: $\{A_k\}_{k=1}^{K}$
  - **Full list of all-subset models when p is small**
  - **Group of models when p is large**
- weighting vector:
  - $w = (w_1, w_2, \ldots, w_K)^T$
- SOIL importance measure for the j-th variable, $j \in \{1, \ldots, p\}$:
  - $S_j \equiv S(j; w; A) = \sum_{k=1}^{K} w_k I(j \in A^k)$

- Candidate models:
  - A: $\{A_k\}_{k=1}^{K}$
  - **Full list of all-subset models when p is small**
  - **Group of models when p is large**
- weighting vector:
  - $w = (w_1, w_2, \ldots, w_K)^T$
- SOIL importance measure for the j-th variable, $j \in \{1, \ldots, p\}$:
  - $S_j \equiv S(j; w; A) = \sum_{k=1}^{K} w_k I(j \in A^k)$

SOIL: Theoretical properties

- Consistency of the SOIL
  - Weak consistency
  - Consistency
- Ensure the weighting is concentrated enough around the true model.

Definition 1 (Weak Consistency and Consistency)**.** The weighting vector $\mathbf{w}$ is *weakly consistent* if

$$\frac{\sum_{k=1}^{K} w_k|\mathcal{A}^k \nabla \mathcal{A}^*|}{r^*} \xrightarrow{p} 0, \qquad \text{as } n \to \infty, \tag{1}$$

and $\mathbf{w}$ is *consistent* if

$$\sum_{k=1}^{K} w_k|\mathcal{A}^k \nabla \mathcal{A}^*| \xrightarrow{p} 0, \qquad \text{as } n \to \infty,$$

where ∇ denotes the symmetric difference of two sets and $|\cdot|$ denotes number counting.

Theorem 1. *(a) Under the assumption that the weighting* $\mathbf{w}$ *is weakly consistent, we have:*

$$\frac{\sum_{j\in\mathcal{A}^*} S_j}{r^*} \xrightarrow{p} 1, \qquad \frac{\sum_{j\notin\mathcal{A}^*} S_j}{r^*} \xrightarrow{p} 0, \qquad \text{as } n \to \infty;$$

(b) When the weighting $\mathbf{w}$ *is consistent, we have:*

$$\min_{j\in\mathcal{A}^*} S_j \xrightarrow{p} 1, \qquad \max_{j\notin\mathcal{A}^*} S_j \xrightarrow{p} 0, \qquad \text{as } n \to \infty.$$

Candidate models

- In the high-dimensional setting (p>>n), it is computationally infeasible to use the candidate models with all subsets.
- Using tools for high-dimensional penalized regression

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(\beta_j),$$

- Penalty functions: Lasso, SCAD, MCP
- Apply a method to compute solution paths, e.g. SCAD

$$\boldsymbol{A}_{\mathrm{SCAD}} = \{\mathcal{A}^{\lambda_1}, \mathcal{A}^{\lambda_2}, \ldots, \mathcal{A}^{\lambda_L}\}$$

- Put together the models to form a larger set of candidate models

$$\boldsymbol{A} = \{\boldsymbol{A}_{\mathrm{Lasso}}, \boldsymbol{A}_{\mathrm{AdaptiveLasso}}, \boldsymbol{A}_{\mathrm{SCAD}}, \boldsymbol{A}_{\mathrm{MCP}}\}$$

Weighting

Two weighting methods:

- **ARM** weighting
- **BIC-p** weighting.

- Need to consider the model complexity
  - Prior probability for the model
  - Non-uniform prior

Algorithm 1 The procedure of the ARM weighting for the regression case.

- Randomly split $\mathbf{D}$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of equal size.
- For each $\mathcal{A}^k \in \boldsymbol{A}$, fit a standard linear regression of y on $\mathbf{x}_s^{(k)}$ using the training set $\mathbf{D}_1$ and get the estimated $\widehat{\boldsymbol{\beta}}_s^{(k)}$ and $\widehat{\boldsymbol{\sigma}}_s^{(k)}$.
- For each $\mathcal{A}^k$, compute the prediction $\mathbf{x}_s^{(k)\top}\widehat{\boldsymbol{\beta}}_s^{(k)}$ on the test set $\mathbf{D}_2$.
- Compute the weight w_k for each candidate model:

$$w_k = \frac{e^{-\psi C_k}(\widehat{\boldsymbol{\sigma}}_s^{(k)})^{-n/2}\prod_{i\in\mathbf{D}_2}\exp(-(\widehat{\boldsymbol{\sigma}}_s^{(k)})^{-2}(y_i-\mathbf{x}_{s,i}^{(k)\top}\widehat{\boldsymbol{\beta}}_s^{(k)})^2/2)}{\sum_{l=1}^{K} e^{-\psi C_l}(\widehat{\boldsymbol{\sigma}}_s^{(l)})^{-n/2}\prod_{i\in\mathbf{D}_2}\exp(-(\widehat{\boldsymbol{\sigma}}_s^{(l)})^{-2}(y_i-\mathbf{x}_{s,i}^{(l)\top}\widehat{\boldsymbol{\beta}}_s^{(k)})^2/2)},$$

  for $k=1,\dots,K$, where $C_k = s_k\log\frac{e\cdot p}{s_k} + 2\log(s_k+2)$.
- Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$ for $l=1,\dots,L$, and get $w_k = \frac{1}{L}\sum_{l=1}^{L} w_k^{(l)}$.

- BIC-p: Define $I_k^{BIC} = -2(\log\ell_k + s_k logn)$

- $w_k = \dfrac{\exp(-I\frac{I_k}{2}-\psi C_k)}{\sum_{l=1}^{K}\exp(-\frac{I_l}{2}-\psi C_l)}$

How do they justify the proposed method?

- By comparing SOIL(BIC-p) and SOIL(ARM) with RFI1, RFI2, LMG.
- LMG: relative importance measure by averaging over all possible orderings for R^2 decomposition.
- RFI1: computed from a normalized difference between the prediction error on OOB portion of the data and that on the permuted OOB data for each predictor variable.
- RFI2: the total decrease in node impurities from splitting on a particular variable, averaged over all trees.

Relative Performances of Importance Measures

Example	n	p	Model Settings
			Gaussian Case
1	100	200	$\boldsymbol{\beta}^* = (4,4,4,-6\sqrt{2},\frac{3}{4},0,...,0)^\intercal$
2	150	14+1	$\boldsymbol{\beta}^* = (4,4,4,-6\sqrt{2},\frac{3}{4},0,...,0)^\intercal$. Add $X_{15} = 0.5X_1+2X_4+e$ and $\beta^*_{15} = 0$, where $e \sim N(0, 0.01)$.
3	150	8	$\boldsymbol{\beta}^* = (0,\ldots,0)^\intercal$
4	150	8	$\boldsymbol{\beta}^* = (1,\ldots,1)^\intercal$
S1	150	20	$\boldsymbol{\beta}^* = (4,4,4,-6\sqrt{2},\frac{3}{4},0,...,0)^\intercal$
S2	150	6+6	$\boldsymbol{\beta}^* = (4,4,-6\sqrt{2},\frac{3}{4},0,0)^\intercal$. Add $(X_1^2,X_2^2,X_3^2,X_4^2,X_5^2,X_6^2)$ and corresponding coefficients $(\beta^*_7,\beta^*_8,\ldots,\beta^*_{12})^\intercal = (4,0,1,0,0,0)^\intercal$.
S3	150	6+6	$\boldsymbol{\beta}^* = (4,4,-6\sqrt{2},\frac{3}{4},0,0)^\intercal$. Add $(X_1X_2,X_1X_3,X_1X_4,X_2X_3,X_2X_4,X_3X_4)$ and corresponding coefficients $(\beta^*_7,\beta^*_8,\ldots,\beta^*_{12})^\intercal = (4,2,2,0,0,0)^\intercal$.
			Binomial Case
5	80	6	$\boldsymbol{\beta}^* = \left(1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5},\frac{1}{6},0\right)^\intercal$
6	5000	6	$\boldsymbol{\beta}^* = \left(1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5},\frac{1}{6},0\right)^\intercal$
S4	150	20	$\boldsymbol{\beta}^* = (4,4,4,-6\sqrt{2},\frac{3}{4},0,...,0)^\intercal$
S5	100	200	$\boldsymbol{\beta}^* = (4,4,4,-6\sqrt{2},\frac{3}{4},0,...,0)^\intercal$

Table 1: Simulation settings

Relative Performances of Importance Measures

	SOIL-ARM	SOIL-BIC-p	LMG	RFI1	RFI2
Inclusion/Exclusion	✓	✓			
Tuning in to information	✓	✓			
Robustness to feature correlation	✓	✓			
Robustness against confuser	✓	✓			
Sensitivity to high-order terms	✓	✓			
Pure relativeness			✓	✓	✓
Order preserving	✓	✓			
High-dimensionality	✓	✓		✓	✓
Non-parametricness				✓	✓
Non-negativity	✓	✓	✓		✓

Table 2: Comparison of the characteristics for the importance measures. A "✓" indicates that a specified method has the given property. A blank space indicates the absence of a property.

Relative performances of importance measures

Inclusion/exclusion: address the issue if an importance measure can give a proper sense if a predictor is likely to be needed in the best model to describe the data.

- SOIL-BIC-p and SOIL-ARM have inclusion/exclusion properties in all the examples.

Tuning in to information: the importance obtained should change due to the enrichment of information.

- Only SOIL-BIC-p and SOIL-ARM react to the much-increased info due to sample size increases.

Robustness to feature correlation: SOIL importance show robustness against noise increase and higher feature correlation.

Robustness against confusers: an importance measure oriented towards sparse modeling should assign near zero importance on the confusers.

- SOIL importance measures are much more robust to confusers.

Relative performances of importance measures

Sensitivity to higher-order terms: SOIL importance measures are more sensitive to inclusion of higher-order terms in the model.

Order preserving: refers to the property that the importance reflects the "order" of the variables or not.

- SOIL importance measures exhibit the order preserving property in all the cases.

High-dimensionality: LMG does not support high-dimensional data

Non-negativity: RFI1 does not yield non-negative importance value.

Real Data Example

- **Two** type of dataset:
  - BGS(Berkeley Guidance Study) with small p
  - Bardet data with large p
- SOIL-ARM and SOIL-BIC-p perform reasonably better than the other importance measure.
- SOIL certainly can **miss subtle variables** in the **true** model when the **sample size is small**. But it does not recommend an **unimportant variable as important**.

Conclusion

- In summary, the SOIL method is helpful in different stages of model building. It can be used to narrow down the set of **covariates** for further consideration and for reaching a **final model** with sound considerations.
- More importantly, it provides an **objective view** on reliability of the model and the **model selection uncertainty**.
- Therefore, it can help much improve reproducibility of data analysis that involves variable selection.

谢谢
MERCI
THANK YOU