Subset Selection Ensembles

Stefan Van Aelst
Anthony Christidis
Ruben Zamar

KU Leuven, Department of Mathematics
University of British Columbia, Department of Statistics

SCRI 2023, Academia Sinica, Taiwan

Linear regression

$$\boxed{y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma\varepsilon_i \qquad i = 1, \ldots, n}$$

- Response y_i
- Predictors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$
- Independent and identically distributed errors ε_i
- Vector of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$

Centering and scaling

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0 \qquad j = 1, \dots, p$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \qquad j = 1, \dots, p$$

Notation:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Least squares

The classical estimator is the least squares estimator (Gauss, 1795) which solves

$$\boxed{\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}$$

- Optimal when the errors are i.i.d. normal
- Easy to compute

High dimensional data

- Data with $p > n$ are common nowadays in fields like chemometrics, genomics, and many others.
- Bias-variance trade-off
    - Larger models have less bias but more variance.
    - Unless n is very large ($n/p > 20$, say) trading-off some bias for a decrease in variance may be reasonable.
- **Sparsity**: many of the candidate variables included in the model are not very useful.

$\hookrightarrow$ A possible approach: fit LS to a reduced subset of predictors, but which one?

Best subset selection (Garside, 1965)

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq t$$

- $t \leq \min(n-1, p)$ is the number of nonzero coefficients in $\boldsymbol{\beta}$.
- t is often chosen via cross-validation. (Beale et al., 1967)
- Trade a small bias for a large reduction in variance.
- Highly interpretable.
- Nonconvex optimization problem, exact solution is not feasible. (Welch, 1982)
- Modern algorithms for high quality approximate solutions. (Hazimeh and Mazumder, 2020)

Ensemble methods

$$\boxed{\hat{f}(\mathbf{x}) = \bar{f}(\mathbf{x}) = \sum_{g=1}^{G} \hat{f}_g(\mathbf{x})/G}$$

- High prediction accuracy.
- Mean squared prediction error (Ueda and Nakano, 1996):

$$\text{MSPE}\left[\hat{f}\right] = \text{Bias}\left[\bar{f}\right]^2 + \text{Var}\left[\bar{f}\right] + \sigma^2$$

with

$$\text{Bias}\left[\bar{f}\right] = \overline{\text{Bias}} \quad \text{and} \quad \text{Var}\left[\bar{f}\right] = \frac{1}{G}\,\overline{\text{Var}} + \frac{G-1}{G}\,\overline{\text{Cov}}$$

- Aggregate G diverse models.
- Lack interpretability.

Data driven ensembles

- Ensemble a relatively small number of sparse models.
- Each model provides a good fit to the data.
- The models are learned simultaneously from the data.
- Diversity between models is induced by restricting the sharing of predictors between different models.

Best split selection

Best split selection aims to find G models $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^g;\ 1 \leq g \leq G$ such that

$$\min_{\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^G \in \mathbb{R}^p} \sum_{g=1}^{G} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}^g\|_2^2 \quad \text{subject to} \quad \begin{cases} \|\boldsymbol{\beta}^g\|_0 \leq t, & 1 \leq g \leq G, \\ \|\boldsymbol{\beta}_j^{\cdot}\|_0 \leq u, & 1 \leq j \leq p. \end{cases}$$

with $\boldsymbol{\beta}_j^{\cdot} = (\beta_j^1, \beta_j^2, \ldots, \beta_j^G)^T \in \mathbb{R}^G$

- For $t \leq \min(n-1, p)$ the penalty $\|\boldsymbol{\beta}^g\|_0 \leq t$ imposes sparsity on the individual models.
- For $u \leq G$ the penalty $\|\boldsymbol{\beta}_j^{\cdot}\|_0 \leq u$ induces diversity among the models.
- Both t and u are selected in a data-driven manner.

Best split selection

The ensemble model is obtained by

$$\boxed{\hat{\beta} = \bar{\beta} = \frac{1}{G} \sum_{g=1}^{G} \hat{\beta}^{g}.}$$

- The ensemble is an interpretable, sparse linear model!
- Finding the exact best split selection solution is a huge combinatorial problem.

⟶ We need a good approximate algorithm.

Algorithm for fixed t and u

- Initial solutions $\tilde{\boldsymbol{\beta}}^1, \ldots, \tilde{\boldsymbol{\beta}}^G$.
- Apply projected subset gradient descent to the G models cyclically until convergence.
  For each model g an upper bound for the loss function

  $$\mathcal{L}_n\left(\boldsymbol{\beta}^g | \mathbf{y}, \boldsymbol{X}\right) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}^g\|^2$$

  is given by its quadratic approximation

  $$\mathcal{L}_n^Q\left(\boldsymbol{\beta}^g | \mathbf{y}, \boldsymbol{X}, \tilde{\boldsymbol{\beta}}^g\right) = \mathcal{L}_n\left(\tilde{\boldsymbol{\beta}}^g | \mathbf{y}, \boldsymbol{X}\right) + \nabla_{\boldsymbol{\beta}}\mathcal{L}_n\left(\tilde{\boldsymbol{\beta}}^g | \mathbf{y}, \boldsymbol{X}\right)^T\left(\boldsymbol{\beta}^g - \tilde{\boldsymbol{\beta}}^g\right) + \frac{1}{2}C\left\|\boldsymbol{\beta}^g - \tilde{\boldsymbol{\beta}}^g\right\|_2^2$$

  with $C = 2\|\boldsymbol{X}^T\boldsymbol{X}\|_2$.

Projected subset gradient descent

For each model g we iteratively solve

$$\min_{\boldsymbol{\beta}^g} \mathcal{L}_n^Q\left(\boldsymbol{\beta}^g|\mathbf{y}, \boldsymbol{X}, \tilde{\boldsymbol{\beta}}^g\right) = \min_{\boldsymbol{\beta}^g} \left\|\boldsymbol{\beta}^g - \left(\tilde{\boldsymbol{\beta}}^g - \frac{1}{C}\nabla_{\boldsymbol{\beta}}\mathcal{L}_n\left(\tilde{\boldsymbol{\beta}}^g|\mathbf{y}, \boldsymbol{X}\right)\right)\right\|_2^2$$

which needs to be minimized under the constraints $||\boldsymbol{\beta}^g||_0 \leq t$ and $||\boldsymbol{\beta}_{j\cdot}||_0 \leq u$ for $1 \leq j \leq p$.
Let S^g contain all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ whose components only differ from zero for feasible predictors which are not yet included in u other models, then

$$\underset{||\boldsymbol{\beta}^g||_0 \leq t, \boldsymbol{\beta}^g \in S^g}{\operatorname{argmin}} \mathcal{L}_n^Q\left(\boldsymbol{\beta}^g|\mathbf{y}, \boldsymbol{X}, \tilde{\boldsymbol{\beta}}^g\right) = \mathcal{P}\left(\tilde{\boldsymbol{\beta}}^g - \frac{1}{C}\nabla_{\boldsymbol{\beta}}\mathcal{L}_n\left(\tilde{\boldsymbol{\beta}}^g|\mathbf{y}, \boldsymbol{X}\right);\ S^g, t\right)$$

$\mathcal{P}(v;\ S, t)$ is the projected subset operator which retains the t largest elements in absolute value of the vector v that belong to the set S.

KU LEUVEN

Initial solutions

- We run the algorithm for $u = 1, \ldots, G$ (for a fixed t).
- For $u > 1$ we use the solution at $u - 1$ as initial solution. (warm starts)
- We repeat this procedure on a grid of t values (a subset of $\{1, \ldots, n - 1\}$).
- The optimal values of t and u are selected by CV.

$\hookrightarrow$ We need to generate an initial solution for $u = 1$.

Stepwise split selection

- For $u = 1$, the G models cannot share predictors.
- We generalize the stepwise forward selection procedure to construct multiple models:

1. Set $\tilde{\boldsymbol{\beta}}^1, \ldots, \tilde{\boldsymbol{\beta}}^G = \mathbf{0}$, i.e. all models are empty and take all available predictors as initial set of candidate predictors.
2. Repeat until all models are saturated ($||\tilde{\boldsymbol{\beta}}^1||_0 = \cdots = ||\tilde{\boldsymbol{\beta}}^G||_0 = n - 1$) or no predictor yields a sufficient improvement anymore
    a. For each unsaturated model find the candidate predictor that yields the largest improvement for this model and calculate the p-value for this candidate predictor.
    b. If the smallest p-value is below a threshold γ, then add the candidate predictor to the corresponding model and remove it from the set of candidate predictors.
3. Apply the lasso to each of the G models.

KU LEUVEN

The number of models

What is the effect of G on the performance of best split selection?

MSPE evaluated on a test set of size 2 000 (relative to σ^2).

	$\zeta = \mathbf{0.1}$			$\zeta = \mathbf{0.2}$			$\zeta = \mathbf{0.4}$		
G	**MSPE**	$\overline{\textbf{MSPE}}$	$\overline{\textbf{Cor}}$	**MSPE**	$\overline{\textbf{MSPE}}$	$\overline{\textbf{Cor}}$	**MSPE**	$\overline{\textbf{MSPE}}$	$\overline{\textbf{Cor}}$
1	1.39	–	–	1.30	–	–	1.24	–	–
2	1.29	1.56	0.85	1.31	1.55	0.87	1.28	1.56	0.84
3	1.21	1.65	0.82	1.23	1.62	0.85	1.21	1.55	0.85
4	1.23	1.77	0.80	1.20	1.70	0.83	1.19	1.65	0.83
5	1.19	1.80	0.79	1.16	1.72	0.82	1.15	1.63	0.83

The number of models

What is the effect of G on the recall and precision of best split selection?

- Recall: $\mathrm{RC} = \dfrac{\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0, \hat{\beta}_j \neq 0)}{\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)}$
- Precision: $\mathrm{PR} = \dfrac{\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0, \hat{\beta}_j \neq 0)}{\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0)}$

The number of models

What is the effect of G on the recall and precision of best split selection?

	$\zeta = \mathbf{0.1}$		$\zeta = \mathbf{0.2}$		$\zeta = \mathbf{0.4}$	
G	**RC**	**PR**	**RC**	**PR**	**RC**	**PR**
1	0.45	0.54	0.31	0.61	0.19	0.69
2	0.56	1.00	0.28	1.00	0.16	1.00
3	0.79	0.98	0.42	1.00	0.21	1.00
4	0.81	0.90	0.56	1.00	0.30	1.00
5	0.84	0.85	0.67	0.99	0.34	1.00

Performance comparison

We compare the following methods in R

1. **Stepwise** forward regression (lars).
2. **Lasso** (glmnet).
3. **EN**: Elastic Net with $\alpha = 3/4$ (glmnet).
4. **Fast-BSS**: Best subset selection (L0Learn).
5. **Step-SplitReg** (stepSplitReg).
6. **SplitReg-Lasso** (SplitReg).
7. **SplitReg-EN** with $\alpha = 3/4$ (SplitReg).
8. **Fast-BSpS**: Best split selection with $G = 5$ (PSGD).
9. **RGLM**: Random GLM (RGLM).
10. **RF**: Random Forest (randomForest).
11. **XGBoost**: Extreme Gradient Boosting (xgboost).

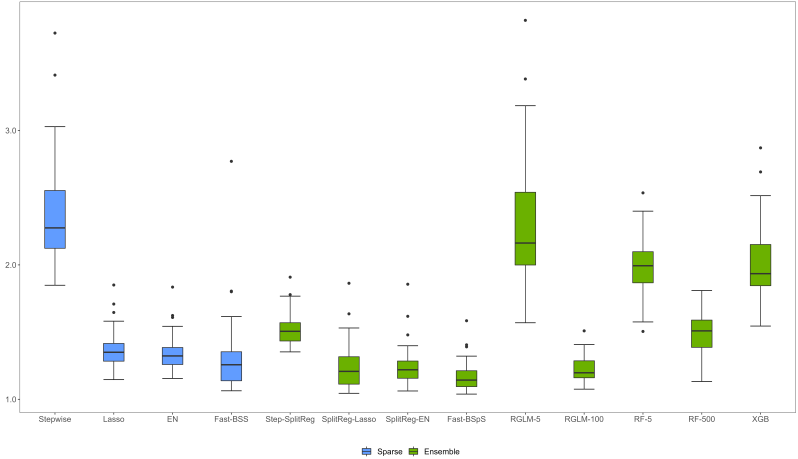
Simulation design

Model: $y_i = \mathbf{x}_i'\boldsymbol{\beta}_0 + \sigma\epsilon_i, \quad 1 \leq i \leq n.$

- $n = 50$ and p is 150 or 500.
- The number of active (i.e. nonzero) variables is $p_0 = [p\zeta]$ with $\zeta \in \{0.1, 0.2, 0.4\}$.
- The errors ϵ_i are standard normal distributed.
- The $\mathbf{x}_i \in \mathbb{R}^p$ are multivariate normal with zero mean and covariance matrix $\boldsymbol{\Sigma}$ with 1 on the diagonal and
  Scenario 1: All variables have correlation ρ with each other.
  Scenario 2: Only active variables have correlation ρ with each other.
- $\rho \in \{0.2, 0.5, 0.8\}$
- σ is chosen such that the signal to noise ratio SNR $= \boldsymbol{\beta}_0'\boldsymbol{\Sigma}\boldsymbol{\beta}_0/\sigma^2$ equals 1, 3 or 5.
- Performance is measured by averaging over $N = 50$ replicates.

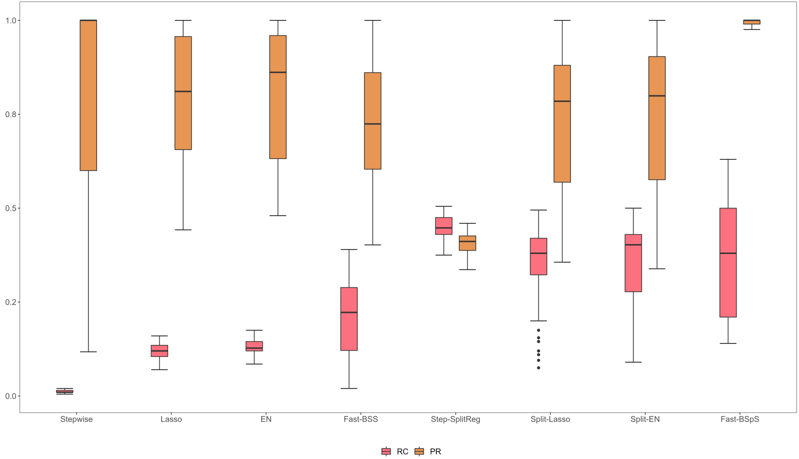
KU LEUVEN

MSPE



MSPEs for Scenario 2 with $\rho = 0.5$, $p = 500$, $n = 50$, SNR $= 5$ and $\zeta = 0.4$.

Recall and precision



MSPEs for Scenario 2 with $\rho = 0.5$, $p = 500$, $n = 50$, SNR $= 5$ and $\zeta = 0.4$.

Average rank of methods over all settings

	$p = 500$			$p = 150$			Overall Rank		
Method	**MSPE**	**RC**	**PR**	**MSPE**	**RC**	**PR**	**MSPE**	**RC**	**PR**
Stepwise	12.06	11.00	**3.87**	11.17	11.00	**3.09**	11.62	11.00	**3.48**
Lasso	7.20	9.81	4.17	6.50	9.78	**3.67**	6.85	9.80	**3.92**
EN	6.17	8.81	4.13	5.93	8.72	4.30	6.05	8.77	4.21
Fast-BSS	4.81	6.89	6.02	5.52	6.75	4.93	5.16	6.82	5.47
Step-SplitReg	9.07	**1.85**	10.26	6.96	5.21	8.96	8.02	3.53	9.61
SplitReg-Lasso	3.57	5.06	6.09	3.33	5.55	5.00	3.45	5.30	5.54
SplitReg-EN	**2.85**	3.89	5.57	**2.74**	4.60	5.41	**2.80**	4.25	5.49
Fast-BSpS	**2.56**	3.56	**3.20**	**2.09**	**2.28**	5.78	**2.33**	**2.92**	4.49
RGLM-5	12.24	**3.24**	8.46	12.69	**1.46**	9.50	12.46	**2.35**	8.98
RGLM-100	3.63	–	–	6.50	–	–	5.06	–	–
RF-5	10.02	7.65	10.15	10.30	6.13	10.67	10.16	6.89	10.41
RF-500	5.69	–	–	5.83	–	–	5.76	–	–
XGB	11.13	4.24	4.07	11.44	4.52	4.70	11.29	4.38	4.38

Application

Bardet-Biedl syndrome (BBS) gene expression dataset (Li et al., 2020)

- Data of 120 mammalian-eye tissue samples.
- Response: expression level of TRIM32 (tripartite motif-containing protein 32).
- Predictors: expression levels of $p = 200$ relevant genes from mammalian-eye tissue samples (Scheetz et al., 2006).
- We randomly split the full dataset $N = 50$ times into a training set of size $n = 30$ and a test set of size $m = 90$.
- For Fast-BSpS we used $G = 5$ and the grids $u \in \{1, 2, 3, 4, 5\}$ and $t \in \{0.3n, 0.4n, 0.5n\} = \{9, 12, 15\}$.

Application: MSPE

Method	**MSPE**	$\overline{\textbf{MSPE}}$
Stepwise	0.84 (0.30)	–
Lasso	0.65 (0.25)	–
EN	0.63 (0.24)	–
Fast-BSS	0.59 (0.18)	–
Step-SplitReg	0.57 (0.19)	0.92 (0.22)
SplitReg-Lasso	0.63 (0.24)	0.65 (0.23)
SplitReg-EN	0.62 (0.23)	**0.63 (0.23)**
Fast-BSpS	**0.45 (0.08)**	**0.60 (0.10)**
RGLM	**0.45 (0.10)**	1.67 (0.35)
RF	0.67 (0.17)	1.03 (0.19)
XGB	0.84 (0.25)	1.04 (0.23)

Application: important genes

Genes can be ranked in order of importance according to the number of individual models they appear in. Let A_k denote the set of genes that appears in at least k models, then we have

$$|A_4| = 0, \ |A_3| = 20, \ |A_2| = 27, \ |A_1| = 28.$$

- Fast-BSpS thus uses only 28 genes.
- 20 genes appear in 3 different models.
- 7 appear in two different models.
- 1 gene is used in only 1 model.

Conclusion

- Best split selection yields a highly interpretable ensemble model with excellent prediction accuracy.
- We developed an efficient approximate algorithm.
- `R` packages `stepSplitReg` and `PSGD` are available on CRAN.
- The framework can be extended to many settings.

Conclusion

- Best split selection yields a highly interpretable ensemble model with excellent prediction accuracy.
- We developed an efficient approximate algorithm.
- `R` packages `stepSplitReg` and `PSGD` are available on CRAN.
- The framework can be extended to many settings.

Thank you for your attention!

Christidis, A.-A., Van Aelst, S., and Zamar, R. (2023). "Multi-Model Subset Selection," https://arxiv.org/abs/2204.08100.

KU LEUVEN