Supplementary Material: Sequential Feature Selection and Inference using Multivariate Random Forests

Joshua Mayer, Raziur Rahman, Souparno Ghosh and Ranadip Pal

Simulation

Both MLASSO and MEnet produce weights that reflect importance/relevance of the associated covariates. However, reliability of these weights completely depend upon correct model specification. If the model is misspecified the weights are not reliable indicators of variable importance. We offer a simple simulation study below to illustrate our point.

We generate the feature matrix, $\mathbf{X}$, independently from $\mathcal{N}(0,1)$ and use 3 *signals* and 997 *spurious* features ($m = 0.003M$). Marginally, both responses share the same set of features as follows:

$$Y_j = 3X_1 + 2X_1^2 + 4X_2 + 3X_2^2 + 8X_3 + \epsilon,\ j = 1, 2$$

where $\epsilon \sim \mathcal{N}(0,1)$. Dependence between Y_1 and Y_2 are induced by a Gumbel copula with $\nu = 2$.

We allow $n_{train} = 400$ and $n_{test} = 100$ for each of five folds and for SMuRFS, we fix $q = 5$ and $\alpha = 0.05$. The training set is again evenly splitted into a *secondary training* set and a *secondary test* set. We select variables on the *secondary training* set and build the predictive random forest on the *secondary test* set. We then compute the prediction error on the *primary test*. Table 1 shows the selection accuracy of competing algorithms.

All the algorithms correctly identify the *signals*. Next, we check whether estimates of regression coefficients obtained from MLASSO and MEnet correctly identified the variable importance in Table 2. Note that, the regularization algorithms need a pre-specified feature matrix. If one suspects polynomial functions of features are important, one needs to include those polynomials manually. Consider the simulation example we offer in the main manuscript where the right hand side of the data generating model has a linear component and a logistic component. How many polynomial terms need to be included in that case? In general, when dealing with huge feature set, expanding the feature matrix to accommodate polynomial terms increases computational burden considerably. Such polynomial terms are not customarily included in standard regularization methods. We also have not included the quadratic terms when fitting MLASSO and MEnet in this simulation. Admittedly, we are fitting a misspecified model, but then again, in real data it is almost impossible to rule out model misspecification. Regardless of the misspecification both MLASSO and MEnet correctly identified the *signals*!

Observe that, in some folds, both MLASSO and MEnet correctly identify variable importance only for the linear terms (X_3 dominates both X_1 and X_2 only in the linear terms). The presence of the quadratic terms of X_1 and X_2 leads to misidentification in couple of folds. Furthermore, the spurious features appearing for MLASSO and MEnet in Table 1 imply the existence of several non-zero β associate with spurious features. Ordering of these spurious features is meaningless. Although, we do observe

Table 1: Table showing the selection accuracy of competing algorithms

Method	Fold	Number of true signal	Number of signals identified	Number of spurious features selected
SMuRFS	Fold 1	3	3	0
	Fold 2	3	3	2
	Fold 3	3	3	0
	Fold 4	3	3	1
	Fold 5	3	3	0
MLASSO	Fold 1	3	3	26
	Fold 2	3	3	7
	Fold 3	3	3	21
	Fold 4	3	3	12
	Fold 5	3	3	33
MEnet	Fold 1	3	3	216
	Fold 2	3	3	139
	Fold 3	3	3	191
	Fold 4	3	3	112
	Fold 5	3	3	335

that, across all the folds, weights of the spurious features are smaller than the weights associated with X_1, X_2 and X_3 (last column of Table 2). Regardless, it seems that weights estimated from MLASSO and MEnet may give a distorted picture of relative importance and relevance of features under model misspecification. This demonstration suggests that both MLASSO and MEnet are more robust in terms of identifying important features as compared to ranking features.

Instead, we contend that it is safer to label the features as statistically significant or not.

Finally, we use the secondary test set to generate the predictive multivariate conditional RF. The predictive performances in the primary test sets (across the folds) are shown in Table 3.

Table 2: Table showing the Rank Ordering ability of MLASSO and MEnet for the synthetic data. In each case, β_i is the estimated regression coefficient associated with X_i.

Method	**Fold**	**Response**	β_1	β_2	β_3	$\max \lvert\beta_j\rvert_{j=4}^{1000}$
MLASSO	Fold 1	Y_1	2.25	4.11	6.48	0.56
		Y_2	2.25	4.08	6.45	0.57
	Fold 2	Y_1	3.03	2.26	6.06	0.42
		Y_2	3.01	2.25	6.09	0.43
	Fold 3	Y_1	2.59	1.50	5.61	1.20
		Y_2	2.58	1.53	5.67	1.21
	Fold 4	Y_1	3.65	3.60	6.44	0.41
		Y_2	3.60	3.66	6.42	0.42
	Fold 5	Y_1	3.75	3.56	6.16	1.11
		Y_2	3.73	3.55	6.19	1.08
Enet	Fold 1	Y_1	1.30	1.86	2.94	0.72
		Y_2	1.31	1.84	2.93	0.70
	Fold 2	Y_1	1.60	1.03	2.74	0.61
		Y_2	1.59	1.02	2.76	0.62
	Fold 3	Y_1	1.35	0.75	2.35	0.57
		Y_2	1.34	0.77	2.38	0.57
	Fold 4	Y_1	1.70	1.54	2.62	0.48
		Y_2	1.67	1.55	2.62	0.48
	Fold 5	Y_1	2.18	1.64	3.32	1.13
		Y_2	2.14	1.65	3.34	1.11

Table 3: Prediction performance on the test set for training full

Method	**Variable**	**NMSPE**	**NMAPE**
SMuRFS	Y_1	0.2464	0.3180
	Y_2	0.3070	0.3476
MLASSO	Y_1	0.4300	0.4684
	Y_2	0.4730	0.4793
MEnet	Y_1	0.7419	0.6480
	Y_2	0.7642	0.6471

Results on drug pairs S_{C1} and S_{C2}

Table 4: Enrichment analysis for SMuRFS, *strong*-SMuRFS strong, MLASSO and MEnet methods for whole genome statistical background with 0.4 confidence interval for drug pairs S_{C1} and S_{C2} obtained from GDSC dataset

Method	SMuRFS	*strong*-SMuRFS	MLASSO	MEnet
		AZD-0530 & Erlotinib		
Feature size	791	235	171	172
Number of Nodes	607	180	151	153
Number of edges	2111	220	108	113
Average node degree	6.96	2.44	1.43	1.48
Avg Local clustering coeff	0.438	0.363	0.264	0.269
Expected Number of Edges	1368	110	66	70
PPI enrichment p-value	0	0	1.19e-6	1.32e-6
Ratio of Observed to expected edges	1.54	**2**	1.63	1.61
Pathway Gene Count	6	3	0	0
		AZD6244 & PD-0325901		
Feature size	1825	214	222	227
Number of Nodes	1301	181	202	207
Number of edges	10832	238	155	163
Average node degree	16.7	2.63	1.53	1.57
Avg Local clustering coeff	0.333	0.361	0.317	0.324
Expected Number of Edges	7022	127	116	122
PPI enrichment p-value	0	0	2.73e-4	2.38e-4
Ratio of Observed to expected edges	1.54	**1.87**	1.33	1.33
Pathway Gene Count	30	7	4	1
		Nutlin-3a & PD-0332991		
Feature size	837	222	431	439
Number of Nodes	657	176	374	381
Number of edges	2287	265	512	539
Average node degree	6.96	3.01	2.74	2.83
Avg Local clustering coeff	0.35	0.362	0.337	0.332
Expected Number of Edges	1733	160	426	451
PPI enrichment p-value	0	2.29e-14	3.8e-5	2.9e-5
Ratio of Observed to expected edges	1.32	**1.65**	1.2	1.2
Pathway Gene Count	14	11	8	8

Table 5: Prediction performances of competing methods for drug set S_{C1}

Drug Name	Fold	Feature Selection Algorithm	Number of Features	NMSPE	NMAPE
AZD-0530	1	*strong*-SMuRFS	8	**1.0239**	**0.8274**
		SMuRFS	279	1.0652	0.8568
		MLASSO	25	1.3220	0.8793
		MEnet	26	1.1345	0.8769
	2	*strong*-SMuRFS	82	**1.0070**	**0.6666**
		SMuRFS	240	1.0145	0.6677
		MLASSO	108	1.0421	0.6761
		MEnet	108	1.0332	0.6725
	3	*strong*-SMuRFS	118	1.0197	0.5597
		SMuRFS	349	**1.0011**	0.5579
		MLASSO	32	1.0084	0.5539
		MEnet	32	1.0115	**0.5528**
	4	*strong*-SMuRFS	78	0.9984	0.6516
		SMuRFS	317	0.9929	**0.6512**
		MLASSO	15	**0.9204**	0.6658
		MEnet	15	0.9217	0.6628
	5	*strong*-SMuRFS	78	0.9031	0.5894
		SMuRFS	279	**0.8906**	**0.5850**
		MLASSO	9	0.9375	0.6123
		MEnet	9	0.9308	0.6123
Erlotinib	1	*strong*-SMuRFS	8	0.8293	**0.6472**
		SMuRFS	17	**0.8124**	0.6737
		MLASSO	25	0.8892	0.7705
		MEnet	25	0.8986	0.7773
	2	*strong*-SMuRFS	82	**0.7769**	**0.5658**
		SMuRFS	240	0.7927	0.5764
		MLASSO	108	0.8831	0.5863
		MEnet	108	0.8811	0.5927
	3	*strong*-SMuRFS	118	0.8935	0.6567
		SMuRFS	349	0.8643	**0.6497**
		MLASSO	32	**0.8507**	0.6644
		MEnet	32	0.8685	0.6672
	4	*strong*-SMuRFS	78	**0.8181**	**0.6365**
		SMuRFS	317	0.8479	0.6409
		MLASSO	15	0.8728	0.6614
		MEnet	15	0.8785	0.6648
	5	*strong*-SMuRFS	78	0.8434	0.5193
		SMuRFS	279	**0.8423**	**0.5181**
		MLASSO	9	0.8718	0.5264
		MEnet	9	0.8640	0.5285

Table 6: Prediction performances of competing methods for drug set S_{C2}

Drug Name	Fold	Feature Selection Algorithm	Number of Features	NMSPE	NMAPE
AZD-6244	1	*strong*-SMuRFS	92	**0.8298**	**0.6162**
		SMuRFS	1308	0.8429	6188
		MLASSO	75	0.8390	0.6250
		MEnet	76	0.8390	0.6250
	2	*strong*-SMuRFS	55	**0.8444**	**0.6242**
		SMuRFS	279	0.8983	0.7121
		MLASSO	34	0.8813	0.7039
		MEnet	32	0.8835	0.7071
	3	*strong*-SMuRFS	64	**0.7822**	**0.7148**
		SMuRFS	529	0.7953	0.7241
		MLASSO	41	0.8196	0.7473
		MEnet	43	0.8223	0.7486
	4	*strong*-SMuRFS	27	0.8200	0.6688
		SMuRFS	92	**0.8061**	**0.6672**
		MLASSO	71	0.8574	0.7078
		MEnet	71	0.8504	0.7046
	5	*strong*-SMuRFS	81	**0.8665**	**0.6798**
		SMuRFS	669	0.8711	0.6832
		MLASSO	48	0.8909	0.6870
		MEnet	50	0.8928	0.6848
PD-305901	1	*strong*-SMuRFS	92	**0.7668**	**0.6623**
		SMuRFS	1308	0.7849	0.6747
		MLASSO	75	0.8104	0.6803
		MEnet	76	0.8174	0.6840
	2	*strong*-SMuRFS	55	0.7261	**0.6410**
		SMuRFS	279	**0.7227**	0.6485
		MLASSO	34	0.7501	0.6612
		MEnet	32	0.7534	0.6651
	3	*strong*-SMuRFS	64	0.8530	**0.5916**
		SMuRFS	529	0.8524	0.5960
		MLASSO	41	**0.8438**	0.6209
		MEnet	43	0.8469	0.6205
	4	*strong*-SMuRFS	27	0.6788	0.6344
		SMuRFS	92	**0.6751**	**0.6322**
		MLASSO	71	0.6804	0.6406
		MEnet	71	0.6834	0.6415
	5	*strong*-SMuRFS	81	**0.7867**	**0.6699**
		SMuRFS	669	0.7919	0.6741
		MLASSO	48	0.8208	0.6905
		MEnet	50	0.8210	0.6906

Pseudocode for SMuRFS algorithm

```
Inputs: ntree, mtry, alpha, prop.test, data
For i = 1:ntree
{
  select mtry covariates without replacement from remaining covariates
  select a bootstrp sample of size n_data from the data
  grow a conditional inference tree using mtry covariates
  find the minimum p-value among the covariates across all nodes
  select the covariates with Bonferroni corrected p-values > alpha
  from training data obtain a sample of size prop.test * n_data without replacement
  conduct a permuation test for each of the selected covariate
  delete the covariates with  Bonferroni corrected p-value > alpha

}
return(remaining covariates)
```