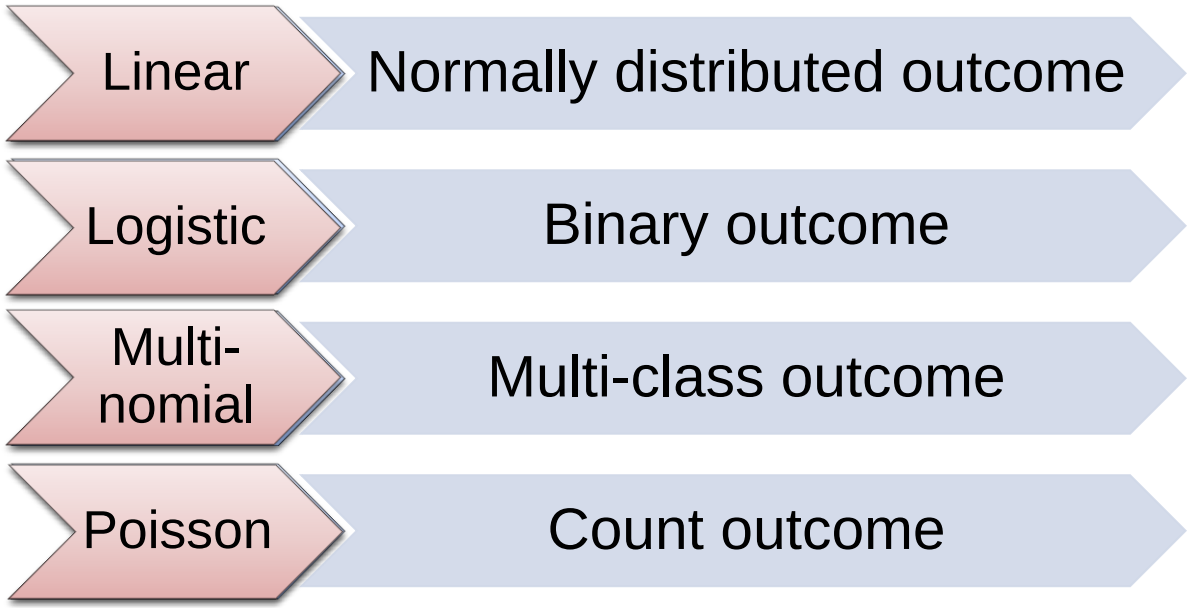
Random generalized linear model: a highly accurate and interpretable ensemble predictor

Song L, Langfelder P, Horvath S. BMC Bioinformatics 2013

Steve Horvath (shorvath@mednet.ucla.edu)
University of California, Los Angeles

Generalized linear model (GLM)

- Flexible generalization of ordinary linear regression.
- Allows for outcomes that have other than a normal distribution.
- R implementation considers all models and link functions implemented in the R function glm



Aside: randomGLM predictor also applies to survival outcomes

Common prediction algorithms

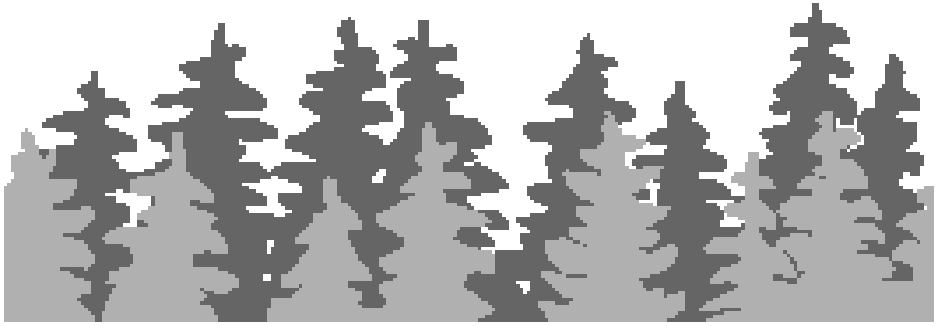
- Generalized linear model (GLM)
- Penalized regression models
  - Ridge regression, elastic net, lasso.
- Recursive partitioning and regression trees (rpart)
- Linear discriminant analysis (LDA)
  - Special case: diagonal linear discriminant analysis (DLDA)
- K nearest neighbor (KNN)
- Support vector machines (SVM)
- Shrunken centroids (SC) (Tibshirani et al 2002, PNAS)
- Ensemble predictors:
  - Combination of a set of individual predictors.
  - Special case: random forest (RF), combination of tree predictors.

Bagging

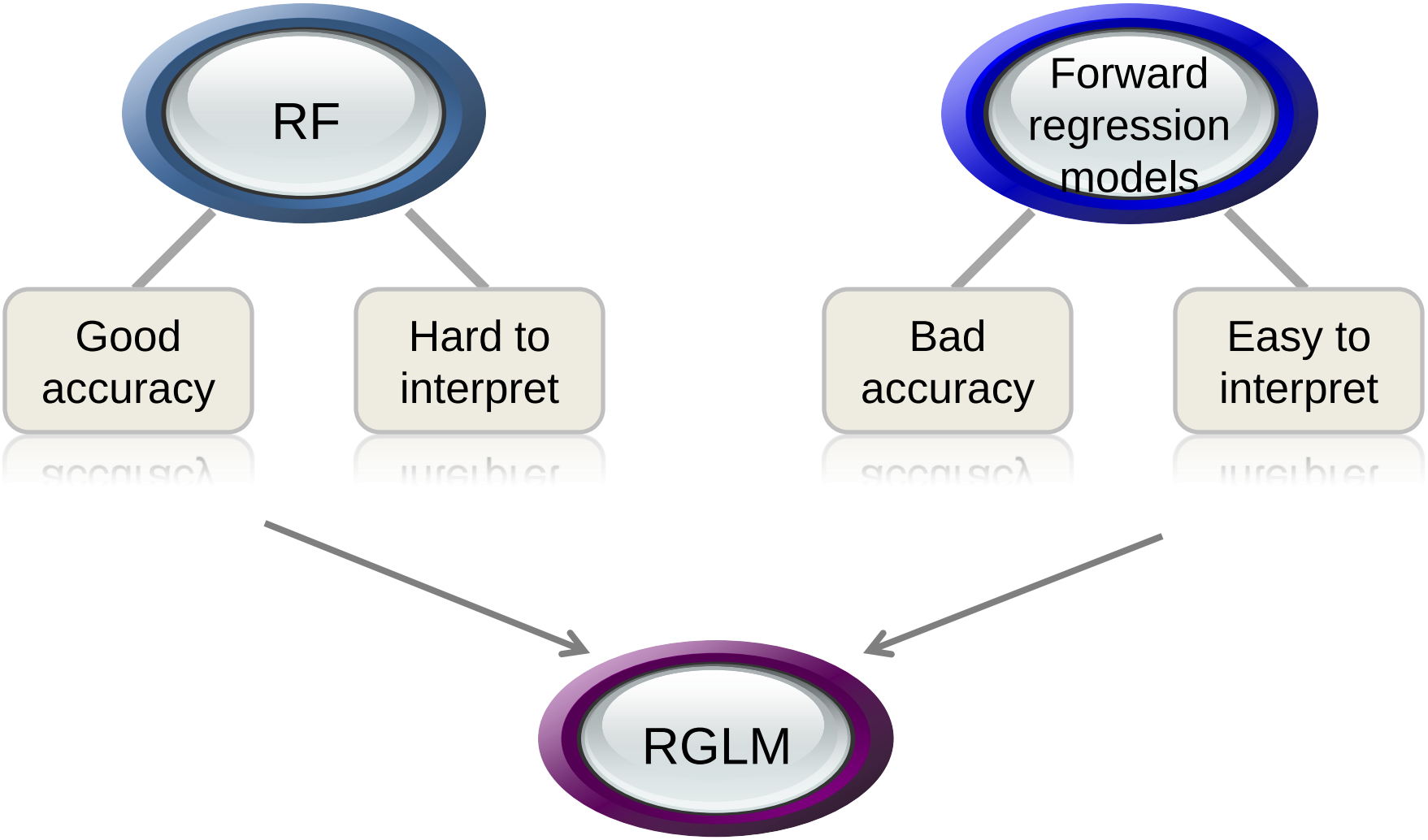
- Bagging = Bootstrap aggregating
- Nonparametric Bootstrap (standard bagging):
- Bag is drawn at random with replacement from the original training data set
- individual predictors (base learners) can be aggregated by plurality voting
- Relevant citation: Breiman (1996)

Random Forest (RF)

- An RF is a collection of tree predictors such that each tree depends on the values of an independently sampled random vector.



Rationale behind RGLM



Breiman L: **Random Forests.** *Machine Learning* 2001, **45:**5–32.
Derksen S, Keselman HJ: **Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables.** *British JMathematical Stat Psychology* 1992, **45**(2):265–282.

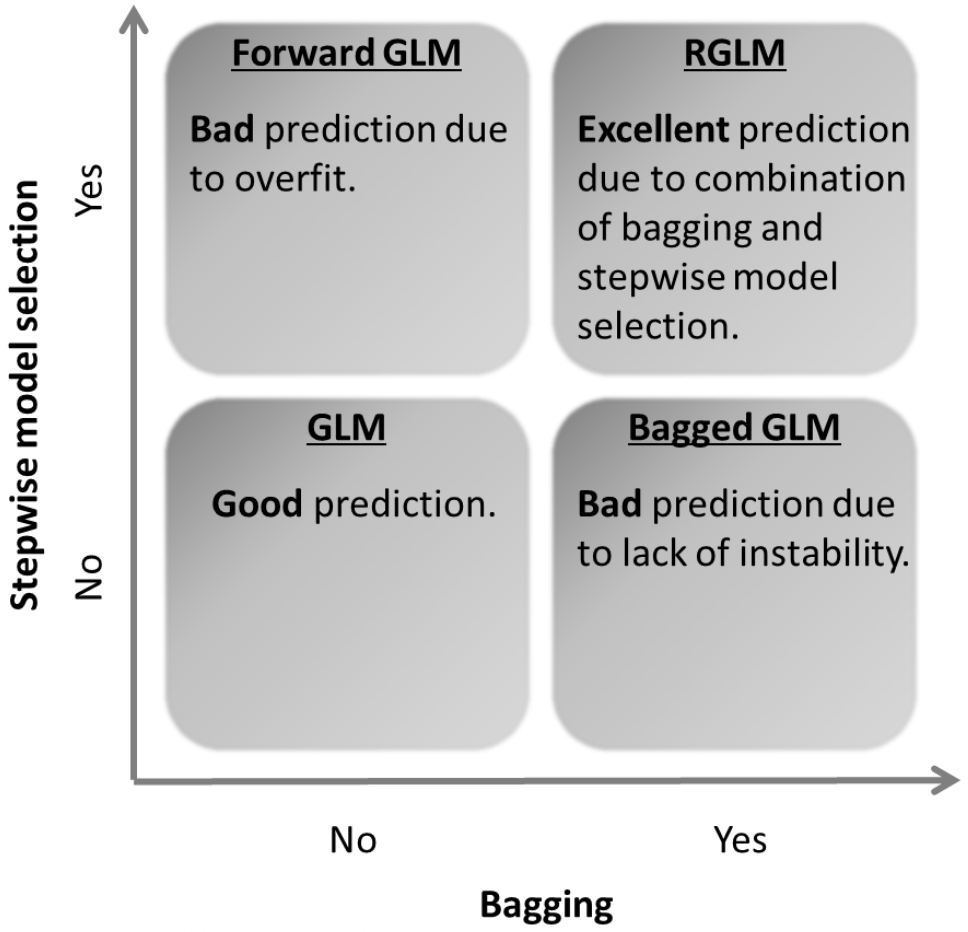
RGLM construction

RGLM construction

- RGLM: an ensemble predictor based on bootstrap aggregation (bagging) of generalized linear models whose covariates are selected using forward regression according to AIC criteria.

RGLM construction combines 2 seemingly wrong choices, forward regression and bagging, for GLMs to arrive at a superior method. Two wrongs make a right.

Not mentioned here: additional elements of randomness.

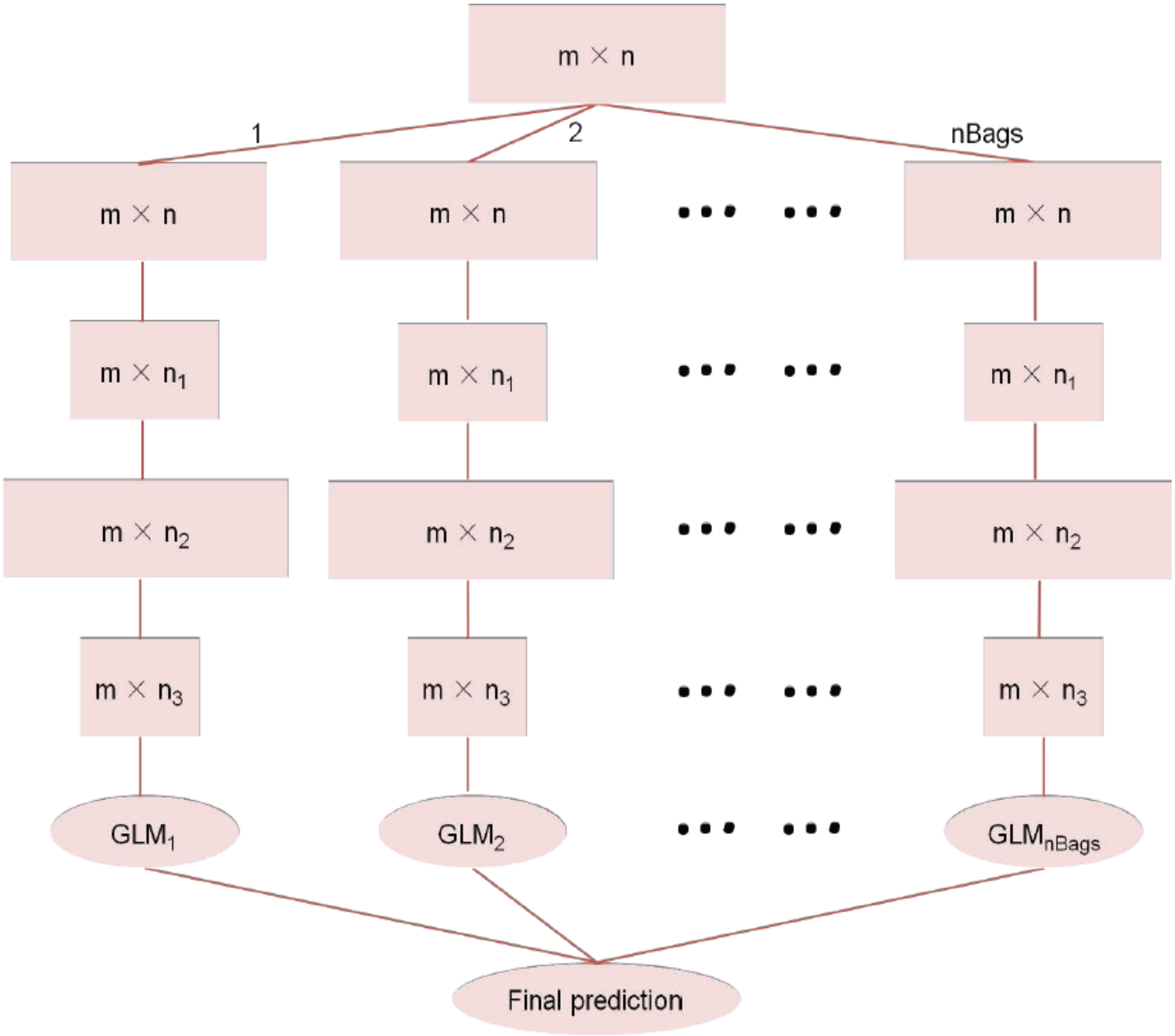


Breiman L: Random Forests. Machine Learning 2001, 45:5–32.

Derksen S, Keselman HJ: Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British JMathematical Stat Psychology 1992, 45(2):265–282.

RGLM construction

1. **Training data.**
   m samples across n features.
2. **Bootstrapping.**
   Select bootstrap samples. Repeat nBags times.
3. **Random subspace.**
   Randomly select n_1=nFeaturesInBag features.
*. **Optional feature interactions.**
   n_2=nFeaturesInBag +no.of.interactions.
4. **Feature selection.**
   Select n_3 = nCandidateCovariates features based on correlation.
5. **Build forward selected GLMs.**
6. **Aggregate predictions over bags.**



RGLM evaluation

RGLM prediction evaluation

- Binary outcome prediction:
  - 20 disease-related expression data sets. — RGLM ties for 1st.
  - 700 comparisons with dichotomized gene traits. — RGLM ranks 1st.
  - 12 UCI benchmark data sets. — RGLM ties for 1st.
  - 180 simulations. — RGLM ties for 1st.

  Accuracy: proportion of observations corrected classified.

- Continuous outcome prediction:
  - Mouse tissue data with 21 clinical traits. — RGLM ranks 1st.
  - 700 comparisons with continuous gene traits. — RGLM ranks 1st.
  - 180 simulations. — RGLM ranks 1st.

  Accuracy: correlation between observed and predicted outcome.

RGLM often outperforms alternative prediction methods like random forest in both binary and continuous outcome predictions.

20 disease-related expression data sets

Data set	Samples	Features
adenocarcinoma	76	9868
brain	42	5597
breast2	77	4869
breast3	95	4869
colon	62	2000
leukemia	38	3051
lymphoma	62	4026
NCI60	61	5244
prostate	102	6033
srbct	63	2308
BrainTumor2	50	10367
DLBCL	77	5469
lung1	58	10000
lung2	46	10000
lung3	71	10000
psoriasis1	180	10000
psoriasis2	82	10000
MSstage	26	10000
MSdiagnosis1	27	10000
MSdiagnosis2	44	10000

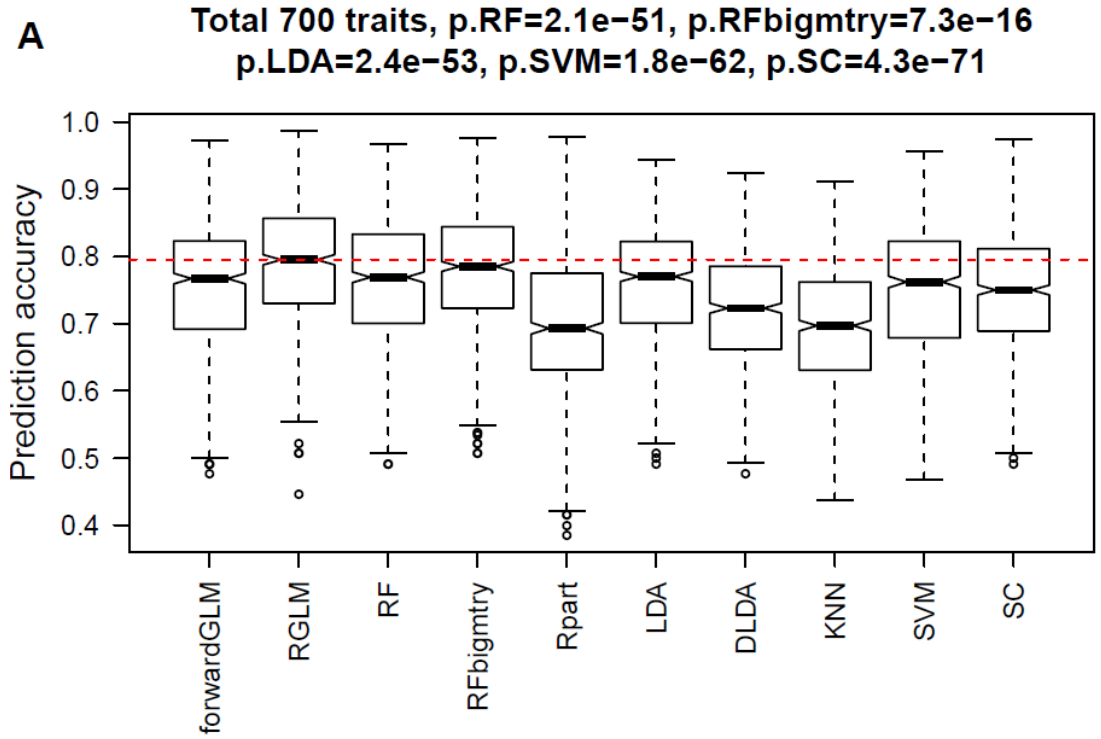
Prediction accuracy in 20 disease-related expression data sets

- RGLM achieves the highest mean accuracy, but not significantly better than RFbigmtry, DLDA and SC.

Data set	RGLM	RF	RFbigmtry	Rpart	LDA	DLDA	KNN	SVM	SC
adenocarcinoma	0.842	0.842	0.842	0.737	0.842	0.744	0.842	0.842	0.803
brain	0.881	0.810	0.833	0.762	0.810	0.929	0.881	0.786	0.929
breast2	0.623	0.610	0.636	0.584	0.610	0.636	0.584	0.558	0.636
breast3	0.705	0.695	0.716	0.611	0.695	0.705	0.669	0.674	0.700
colon	0.855	0.823	0.823	0.726	0.855	0.839	0.774	0.774	0.871
leukemia	0.921	0.895	0.921	0.816	0.868	0.974	0.974	0.763	0.974
lymphoma	0.968	1.000	1.000	0.903	0.960	0.984	0.984	1.000	0.984
NCI60	0.902	0.869	0.869	0.738	0.885	0.902	0.852	0.869	0.918
prostate	0.931	0.892	0.902	0.853	0.873	0.627	0.804	0.853	0.912
srbct	1.000	0.944	0.984	0.921	0.857	0.905	0.952	0.873	1.000
BrainTumor2	0.760	0.750	0.740	0.620	0.760	0.700	0.700	0.660	0.720
DLBCL	0.909	0.851	0.883	0.831	0.922	0.779	0.870	0.792	0.857
lung1	0.931	0.931	0.931	0.828	0.914	0.931	0.931	0.897	0.914
lung2	0.935	0.935	0.935	0.826	0.957	0.978	0.935	0.848	0.978
lung3	0.901	0.901	0.887	0.803	0.873	0.859	0.831	0.859	0.887
psoriasis1	0.989	0.994	0.989	0.978	0.994	0.989	0.989	0.983	0.989
psoriasis2	0.963	0.988	0.976	0.963	0.976	0.963	0.963	0.963	0.963
MSstage1	0.846	0.846	0.846	0.423	0.769	0.769	0.808	0.769	0.769
MSdiagnosis1	0.963	0.926	0.926	0.556	0.889	0.889	0.963	0.926	0.926
MSdiagnosis2	0.591	0.614	0.614	0.568	0.545	0.568	0.568	0.568	0.523
MeanAccuracy	0.871	0.856	0.863	0.752	0.843	0.833	0.844	0.813	0.863
Rank	1	4	2.5	9	6	7	5	8	2.5
Pvalue	NA	0.029	0.079	0.00014	0.0075	0.05	0.014	0.00042	0.37

700 gene expression comparisons with dichotomized gene traits

- 700 = 7*100. Start with 7 human and mouse expression data sets. Randomly choose 100 genes as gene traits for each data set, dichotomize at median.
- RGLM performs significantly better than other methods, although the increase in accuracy is often minor.



12 UCI machine learning benchmark data sets

- 12 famous data sets with binary or dichotomized outcomes.
- Different from many genomic data sets, they have large sample sizes and few features.

Data set	Sample	Features
BreastCancer	699	9
HouseVotes84	435	16
Ionosphere	351	34
diabetes	768	8
Sonar	208	60
ringnorm	300	20
threenorm	300	20
twonorm	300	20
Glass	214	9
Satellite	6435	36
Vehicle	846	18
Vowel	990	10

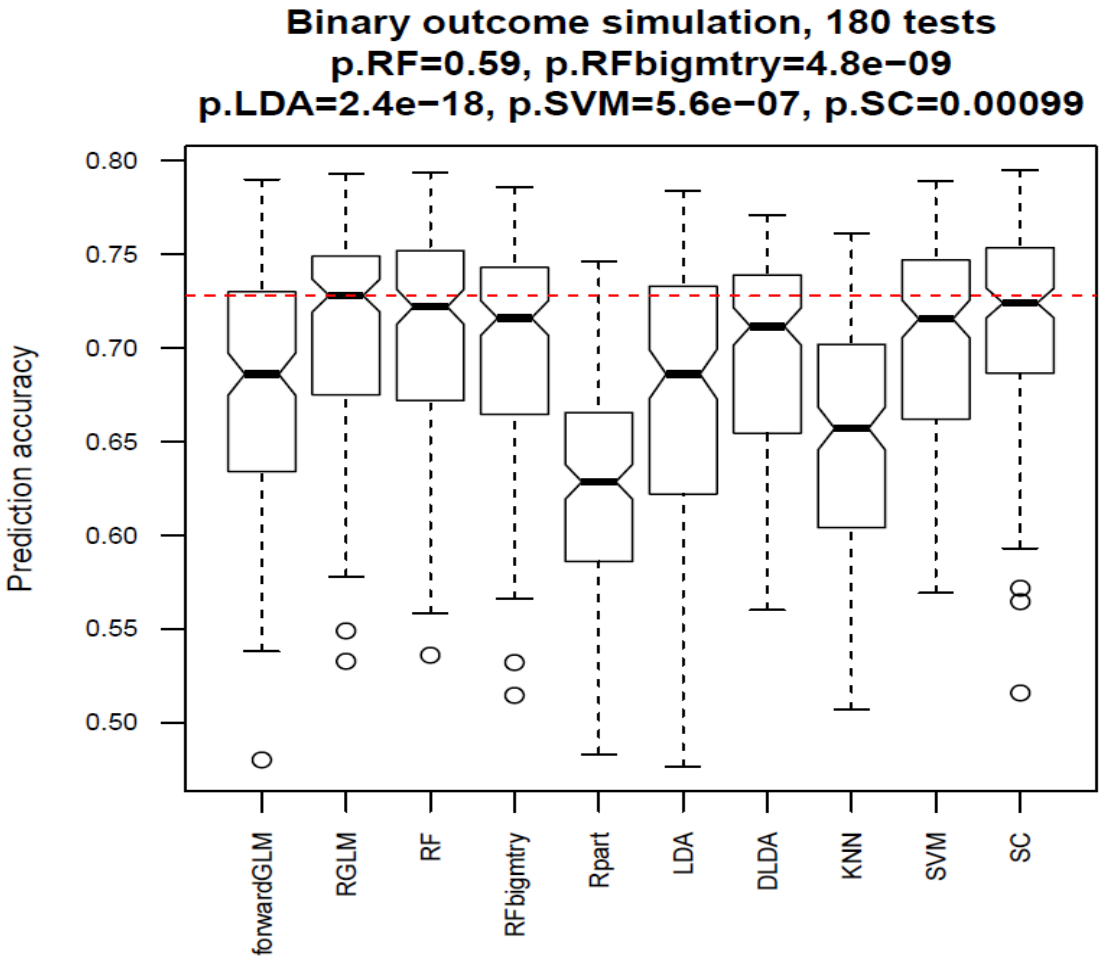
12 UCI machine learning benchmark data sets

- RGLM.inter2 (RGLM considering 2-way interactions between features) ties with RF and SVM.
- RGLM without interaction terms does not work nearly as well.
- Pairwise interaction terms may improve the performance of RGLM in data sets with few features.

Data set	RGLM	RGLM.inter2	RF	RFbigmtry	Rpart	LDA	DLDA	KNN	SVM	SC
BreastCancer	0.964	0.959	0.969	0.961	0.941	0.957	0.959	0.966	0.967	0.956
HouseVotes84	0.961	0.963	0.958	0.954	0.954	0.951	0.914	0.924	0.958	0.938
Ionosphere	0.883	0.946	0.932	0.917	0.875	0.863	0.809	0.849	0.940	0.829
diabetes	0.768	0.759	0.759	0.754	0.741	0.768	0.732	0.740	0.757	0.743
Sonar	0.769	0.837	0.817	0.788	0.707	0.726	0.697	0.812	0.822	0.726
ringnorm	0.577	0.973	0.940	0.910	0.770	0.567	0.570	0.590	0.977	0.535
threenorm	0.803	0.827	0.807	0.777	0.653	0.817	0.825	0.815	0.853	0.817
twonorm	0.937	0.953	0.947	0.920	0.733	0.957	0.960	0.947	0.953	0.960
Glass	0.636	0.743	0.827	0.799	0.729	0.659	0.531	0.808	0.748	0.645
Satellite	0.986	0.987	0.988	0.985	0.961	0.985	0.734	0.990	0.988	0.803
Vehicle	0.965	0.986	0.986	0.973	0.944	0.967	0.729	0.909	0.974	0.752
Vowel	0.936	0.986	0.983	0.976	0.950	0.938	0.853	0.999	0.991	0.909
MeanAccuracy	0.849	0.910	0.909	0.893	0.830	0.846	0.776	0.862	0.911	0.801
Rank	6	2	2	4	8	7	10	5	2	9
Pvalue	0.0093	NA	0.26	0.042	0.00049	0.0093	0.0067	0.11	0.96	0.0015

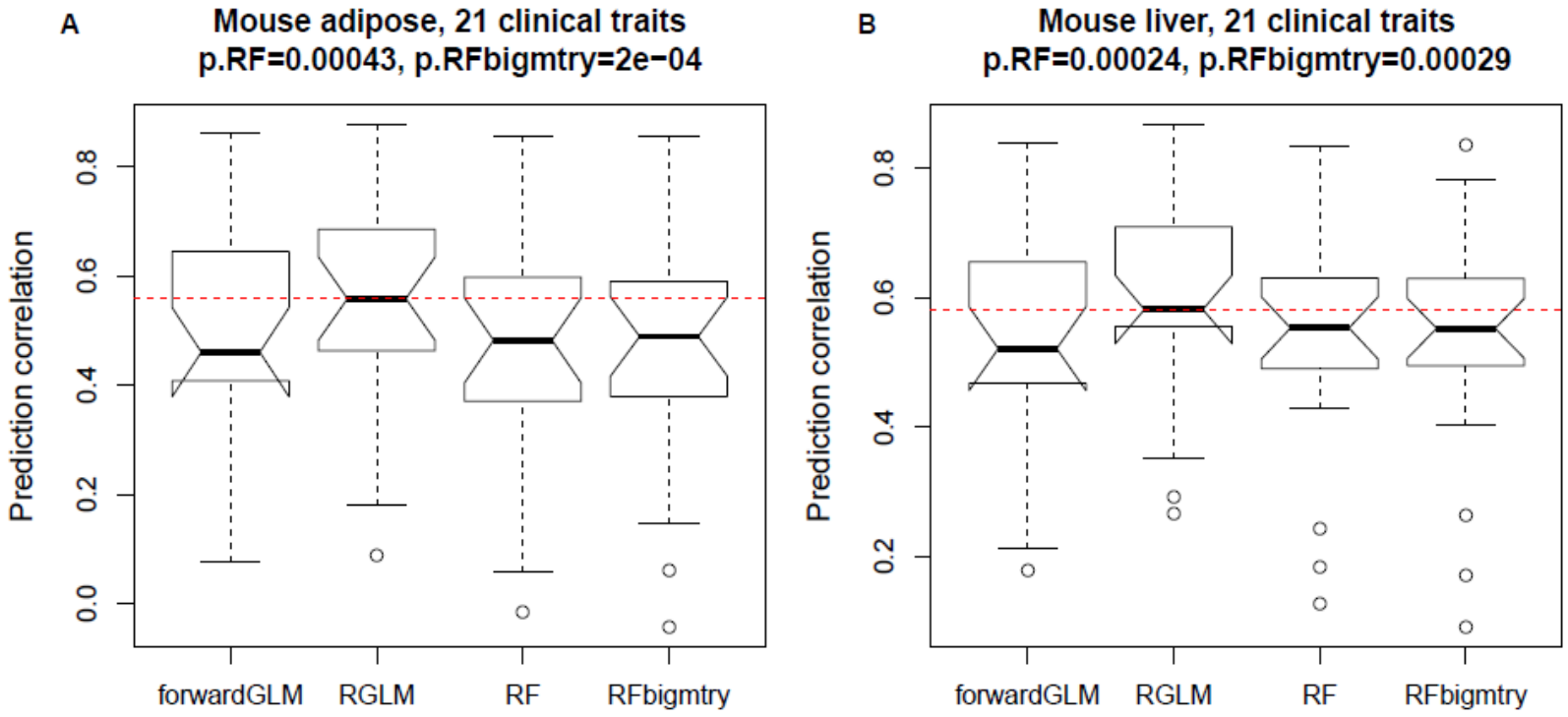
180 simulations

- Number of features varies from 60 to 10000, training set sample size varies from 50 to 2000, test set sample size is fixed to 1000.
- RGLM ties with RF.

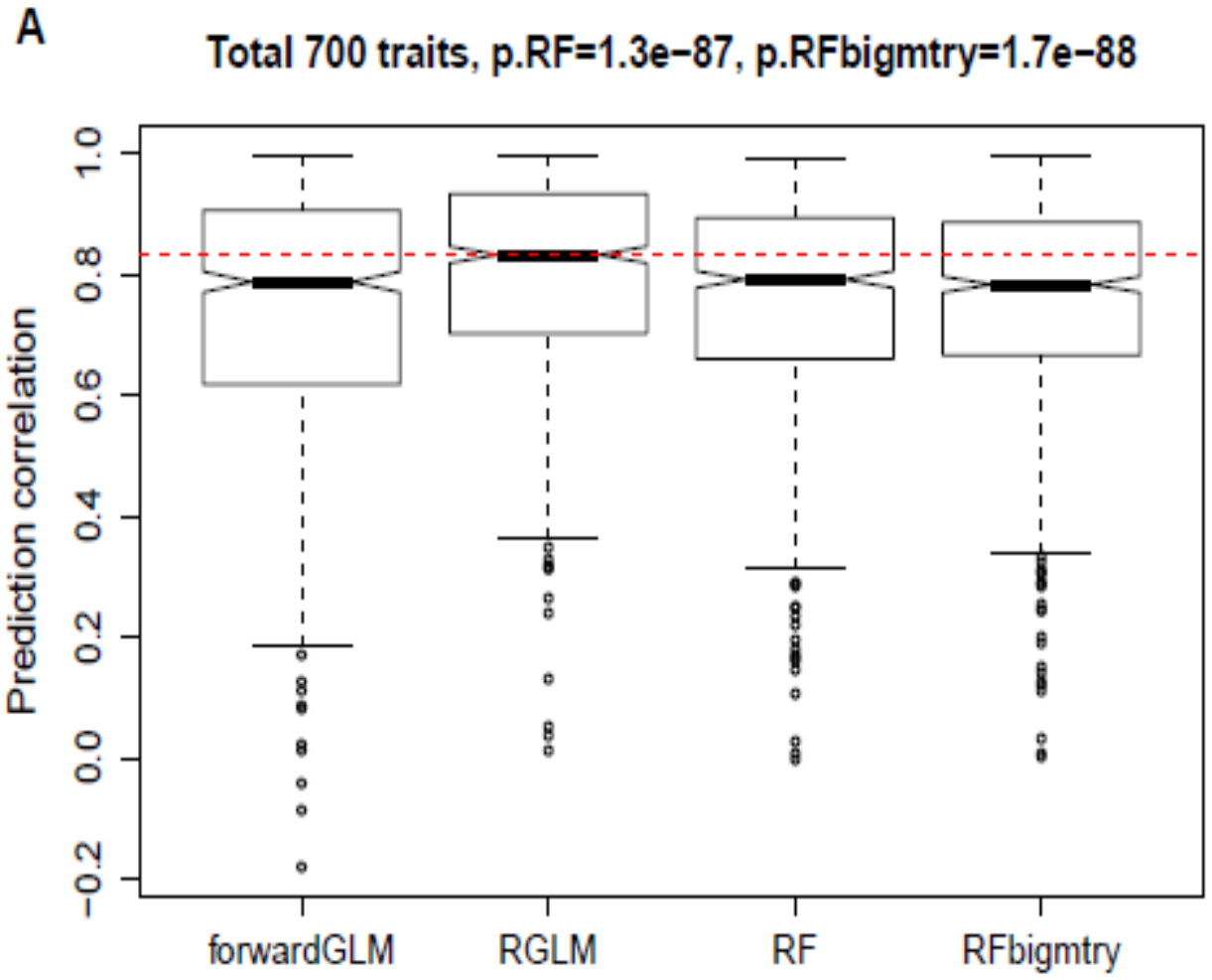


Mouse tissue data with 21 clinical traits

- RGLM performs best when predicting 21 continuous physiological traits based on adipose or liver expression data.
- Data from Jake Lusis

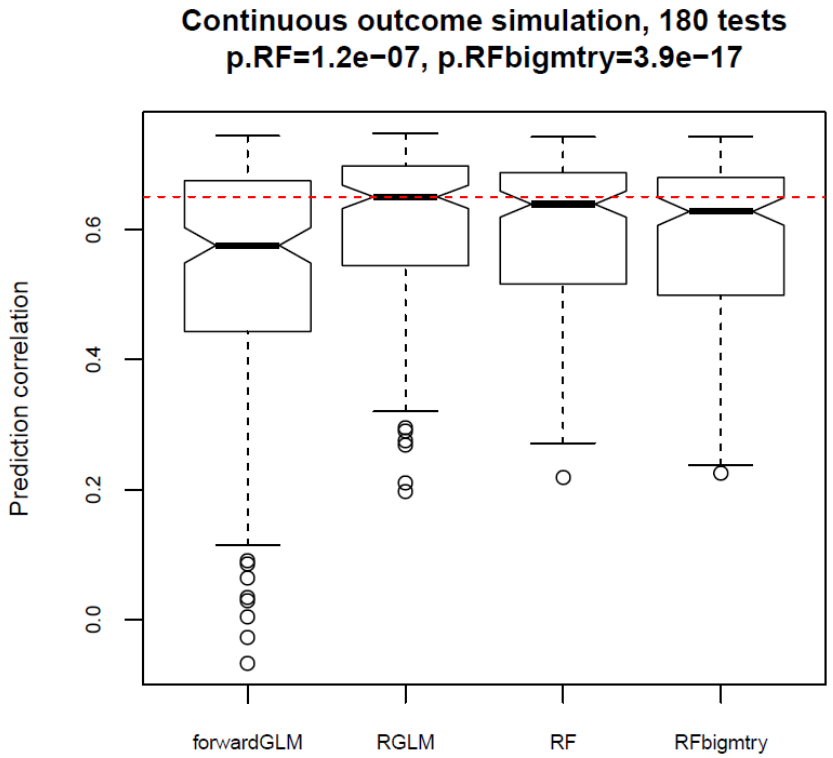


700 gene expression comparisons with continuous gene traits



180 simulations

- Number of features varies from 60 to 10000, training set sample size varies from 50 to 2000, test set sample size is fixed to 1000.
- RGLM performs best.

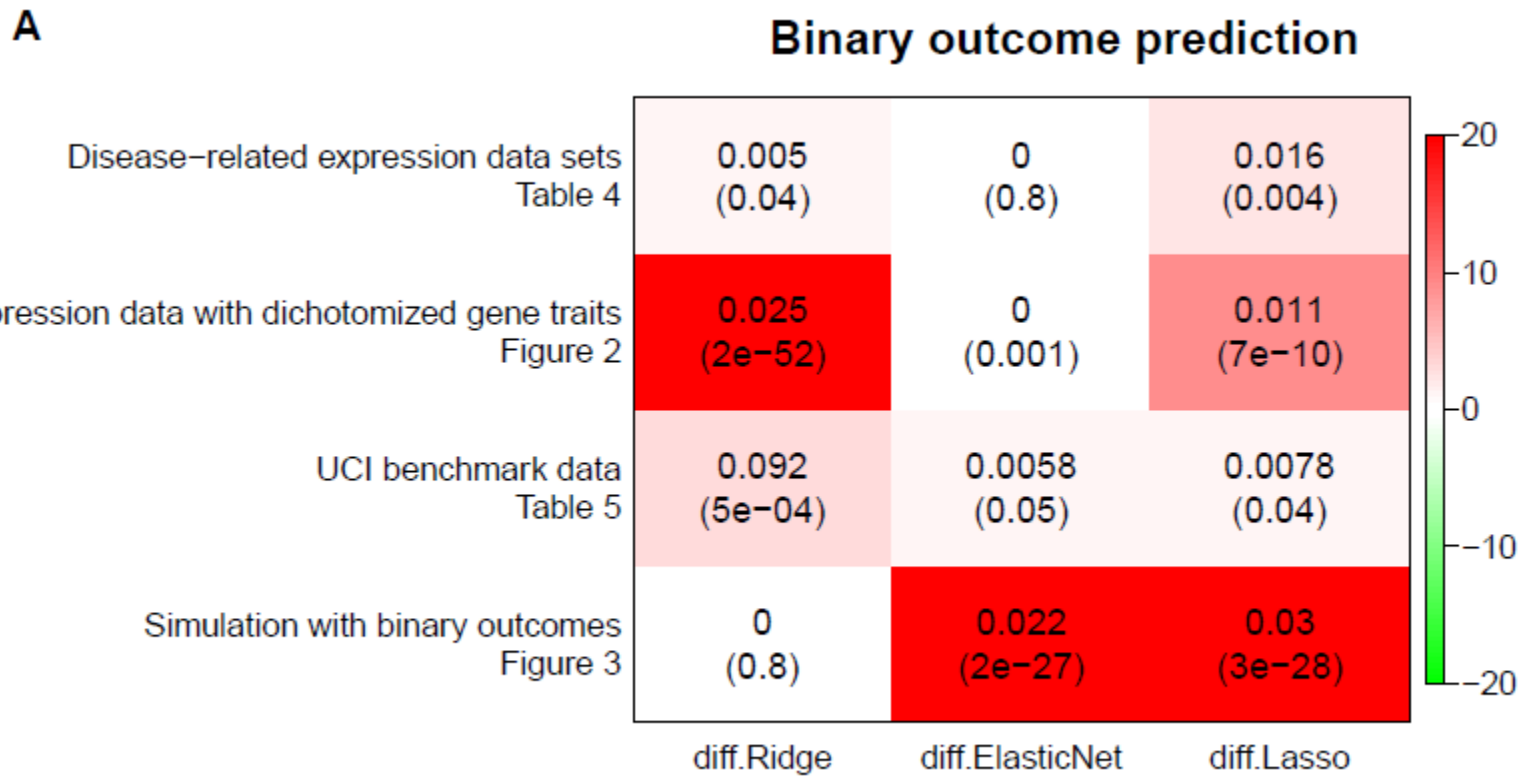


Comparing RGLM with penalized regression models implemented in R package glmnet

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization Paths for Generalized Linear Models via Coordinate Descent,
Journal of Statistical Software, Vol. 33(1), 1-22 Feb 2010

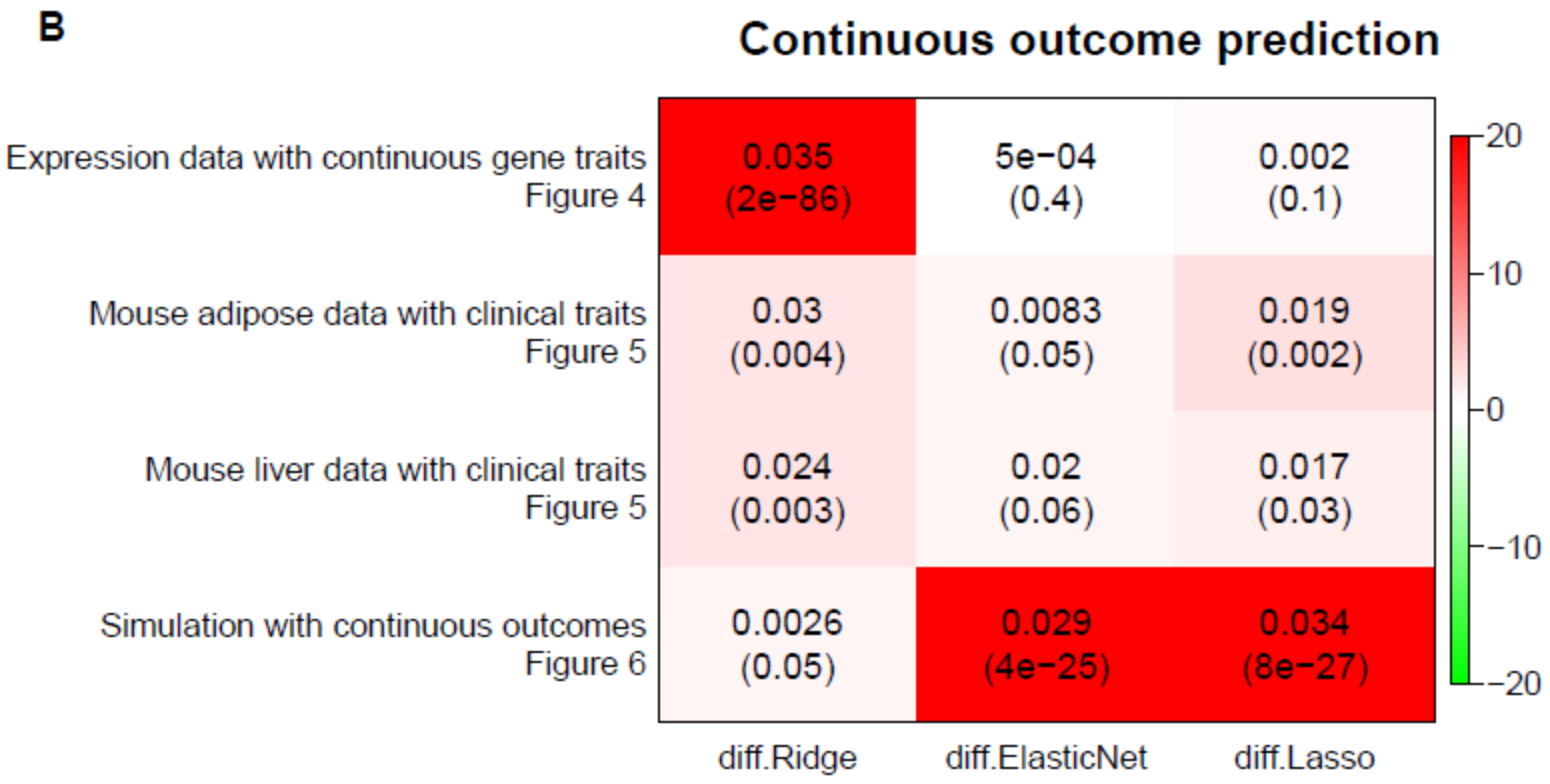
Overall, RGLM is significantly better than ridge regression, elastic net, and lasso for binary outcomes

Table contains differences in accuracy
(and corresponding p-value in brackets)



In general, RGLM is significantly better than ridge regression, elastic net, and lasso for <u>continuous</u> outcomes

Table contains differences in accuracy
(and corresponding p-value in brackets)



Ensemble thinning

Thinned version of RGLM

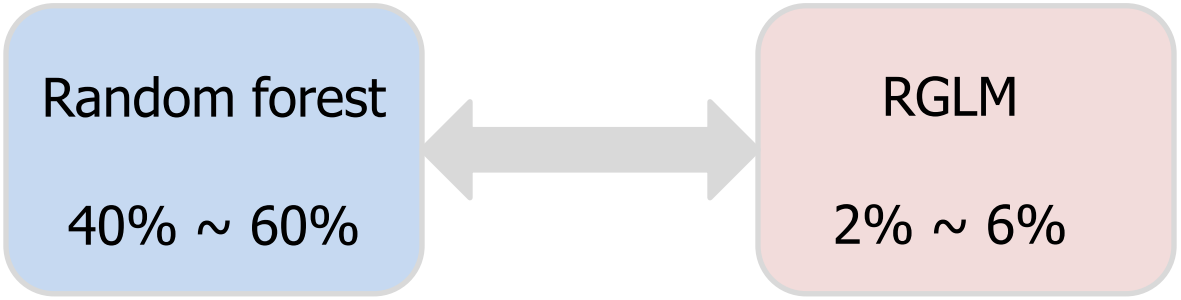
Goal:
Define a sparse predictor that involves few features, i.e. thin the RGLM out by removing rarely occuring features.

Observation:
Since forward variable selection is used for each GLM, some features are rarely selected and contribute little to the ensemble prediction.

Idea:
1) Omit features that are rarely used by the GLMs.
2) Refit each GLM (per bag) without the omitted features.

How many features are being used ?

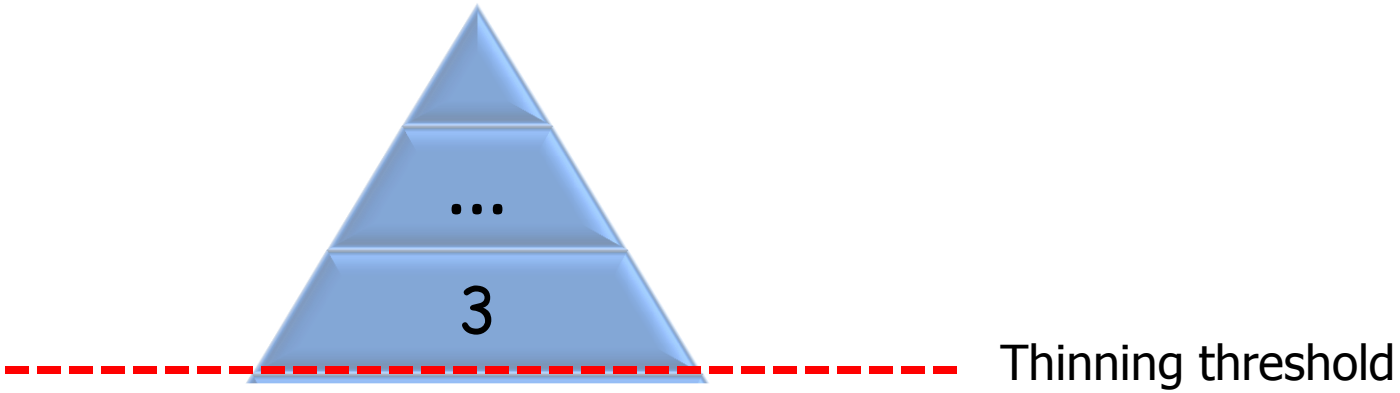
- Example: binary outcome gene expression analysis with 700 comparisons. Total number of features is around 5000 for each comparison.
- We find that RGLM uses far fewer features than the RF

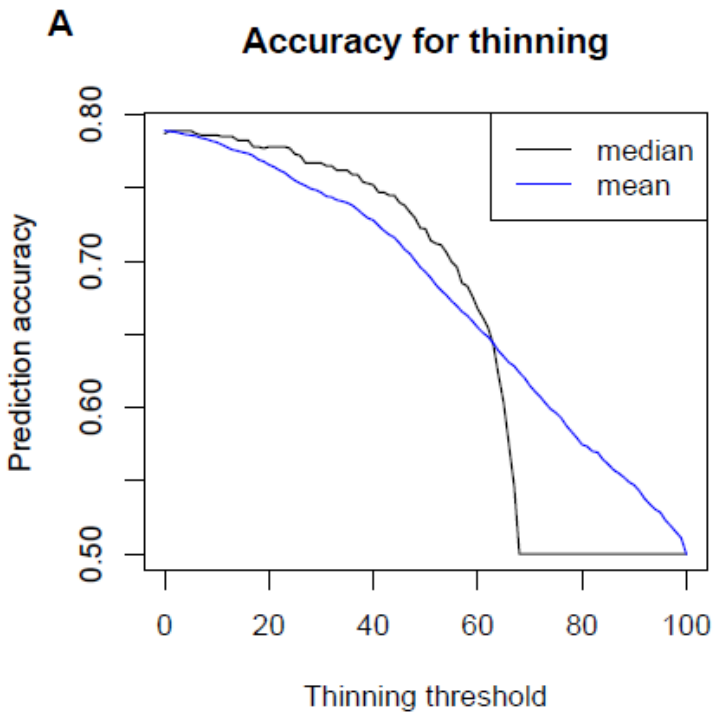


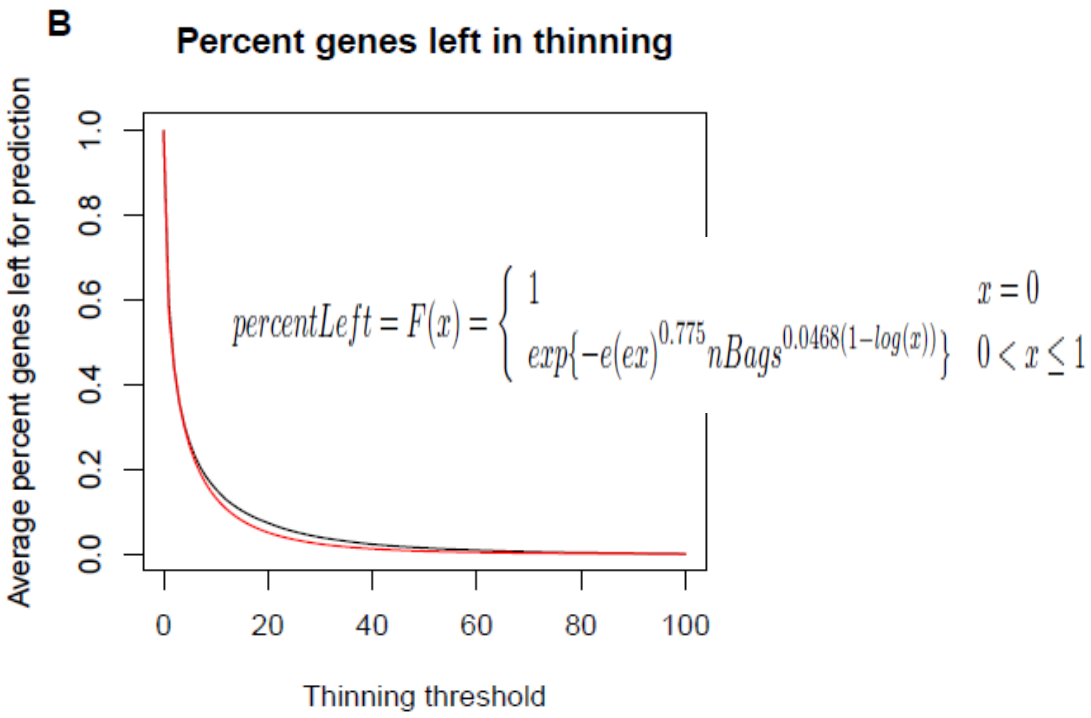
- Reason: RGLM uses forward selection with AIC criterion in each bag
- **Question:** Can we further thin the RGLM predictor out by removing rarely used features?

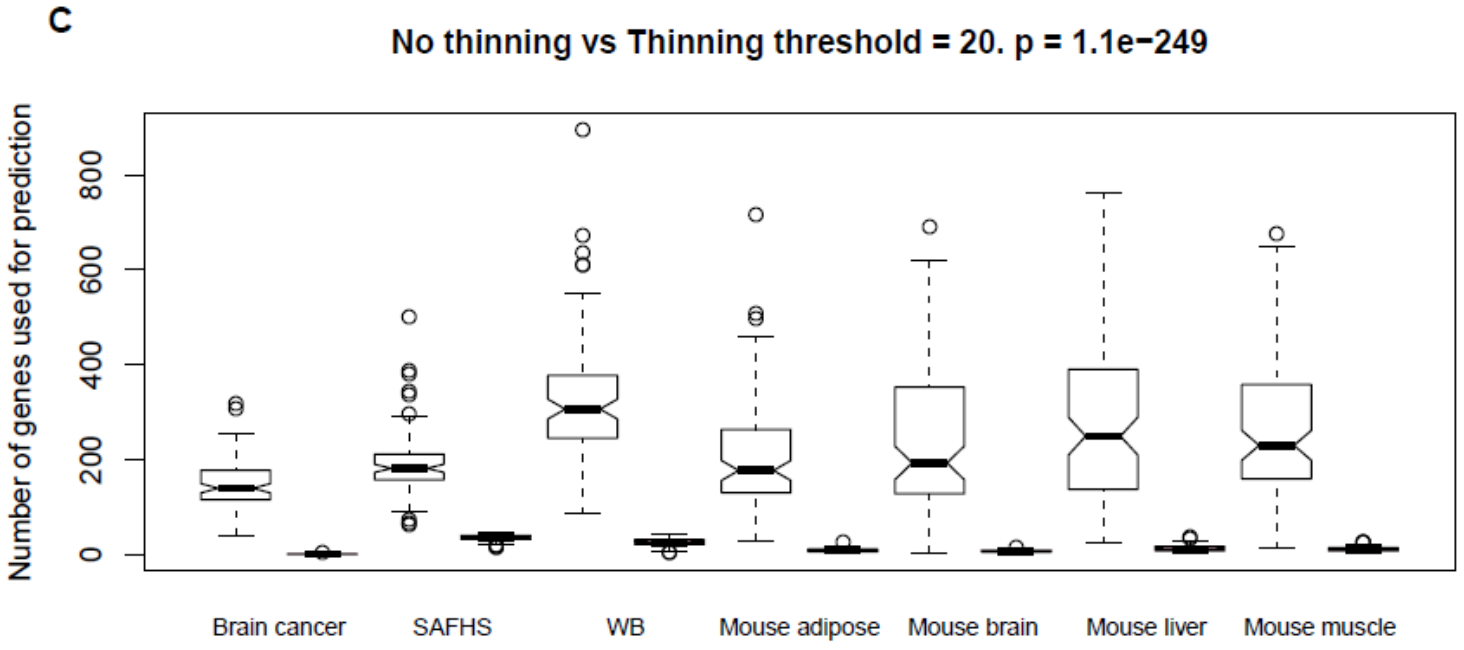
RGLM predictor thinning

- For thinning use the RGLM variable importance measure:
  timesSelectedByForwardRegression that counts the number of times a feature is selected by a GLM (across the number of bags)









- Over 80% features removed
- Median accuracy decreases only 0.009
- Mean accuracy decreases 0.023

Including mandatory covariates

- In many applications, one has a set of mandatory covariates that should be part of each model.
- Example: When it comes to predicting lung disease (COPD) then it makes sense to include smoking status and age in each logistic model
  - and let randomGLM select additional gene expression levels, see
- Straightforward in the randomGLM model:
  - use argument “mandatoryCovariates” in the randomGLM R function, see help(randomGLM)

RGLM pros and cons

- Pros
  - Astonishing accuracy: it often outperforms existing methods.
  - Few features contribute to the prediction especially if RGLM thinning is used.
  - Easy to interpret since it involves relatively few features and uses GLMs.
  - Provides useful by-products as part of its construction including out-of-bag estimates of the prediction accuracy, variable importance measures.
  - GLM formulation allows one to apply the RGLM to different types of outcomes: binary, quantitative, count, multi-class, survival.
  - RGLM allows one to force specific features into regression models in all bags, i.e. mandatory covariates.
- Cons
  - Slower than many common predictors due to the forward selection step (AIC criterion). Work-around: *randomGLM* R implementation allows users to parallelize the calculation.

R software implementation

- The RGLM method is implemented in the freely available R package randomGLM.
- Peter Langfelder contributed and maintains the package.
- Tutorials can be found at the following webpage: http://labs.genetics.ucla.edu/horvath/RGLM
- Can be applied to survival time outcome Surv(time,death)

Random generalized linear model: a highly accurate and interpretable ensemble predictor

Lin Song, Peter Langfelder, Steve Horvath

Human Genetics and Biostatistics, University of California, Los Angeles

SHorvath (at) mednet (dot) ucla (dot) edu
Peter (dot) Langfelder (at) gmail (dot) com

BMC Bioinformatics **14**:5 (2013). DOI: 10.1186/1471-2105-14-5 (link opens in a new tab/window)

R software implementation

- The RGLM method is implemented in the freely available R package *randomGLM*.
- *randomGLM* function outputs training set predictions, out-of-bag predictions, test set predictions, coefficient values, and variable importance measures
- *predict* function for test set predictions
- Tutorials can be found at the following webpage: http://labs.genetics.ucla.edu/horvath/RGLM.

Conclusions

- RGLM shows superior prediction accuracy compared to existing methods, such as random forest, in the majority of studies using simulation, gene expression and machine learning benchmark data sets. Both binary and continuous outcome prediction were considered.
- RGLM is recommended for high-dimensional data, while RGLM.inter2 is recommended for low-dimensional data.
- OOB estimates of the accuracy can be used to inform parameter choices
- RGLM variable importance measure, *timesSelectedByForwardRegression*, allows one to define a "thinned" ensemble predictor with excellent prediction accuracy using only a small fraction of original variables.
- RGLM variable importance measures correlate with other importance measures but are not identical to them. Future evaluations are needed.

Selected references (more can be found in the article)

Song L, Langfelder P, et al (2013) Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC Bioinformatics. PMID: 23323760, PMCID: PMC3645958

[1] Breiman L: Bagging Predictors. Machine Learning 1996, 24:123-140.

[2] Breiman L: Random Forests. Machine Learning 2001, 45:5-32.

[3] Dudoit S, Fridlyand J, Speed TP: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association 2002, 97(457):77-87.

[4] Diaz-Uriarte R, Alvarez de Andres S: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006, 7:3.

[5] Frank A, Asuncion A: UCI Machine Learning Repository 2010, [http://archive.ics.uci.edu/ml].

[6] Meinshausen N, Buhlmann P: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2010, 72(4):417-473.

[7] Perlich C, Provost F, Simono® JS: Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. JOURNAL OF MACHINE LEARNING RESEARCH 2003, 4:211-255.

[8] Buhlmann, Yu B: Analyzing Bagging. Annals of Statistics 2002, 30:927-961.