Predicting Age from sperm methylation data

Peter Sarvari

Introduction:

Our main aim is to find a model (feature set used, their weights and regularization coefficient) to predict age from sperm methylation levels measured using the Illumina 450K array or 850K array. This model should be robust and give reliable results regardless of experimental settings, time and location. We have two independent datasets at hand. One is a 16GB file consisting of 1033 samples and 850K features provided by the Reproductive Medicine Associates, New Jersey (RMA). We will refer to this data as IVF data. The other is a 2GB file consisting of 371 samples and 450K features provided by Prof. Timothy Jenkins (BYU). We will refer to this data as Aging data. Age predicting from sperm methylation is a novel topic and we could only find one paper examining this, Jenkins et al, 2018 [1]. First, we will show that with our feature selection method, it is possible to achieve a higher R-squared value for both datasets than with the features suggested by the paper. Then we show that whereas the features discussed in the paper only work well for the Aging dataset, our features work well for both the Aging and the IVF dataset. Hence our features seem to generalize well on independent datasets. Finally, we will show that partial least squares regression outperforms Lasso in terms of R-squared on an independent dataset using our selected features.

Methods:

Here we describe our primary feature selection method: We average the 450K / 850K methylation levels within genomic intervals defined as follows:

a. Human sperm hypo-methylated regions (HMR)
b. Human sperm hypo-methylated regions overlapping promoters (HMR + prom)
c. Human sperm hypo-methylated regions not overlapping promoters (HMR no prom)

The precise genomic regions that these intervals consider can be found on the UCSC Genome Browser. The averaging in these 3 intervals yield 36506, 16285 and 20221 features, respectively. This corresponds to a more than 10-fold reduction in predictor size.

In this analysis we apply three main machine learning methods:

1. Ridge regression
2. Lasso regression
3. Partial Least-squares Regression (PLS)

We always apply a 10-fold cross-validation technique to choose regularization parameter in case of Ridge and Lasso regression and to choose the number of components in case of PLS.

Whenever we evaluate on the same dataset that we train a model on, we again apply 10-fold cross-validation and report the mean R-squared (R2) value (variance explained) and the standard deviation. When we evaluate the model on an independent dataset from what it was trained on, we report the R-squared value. For the Ridge and Lasso regressions, we use the R package called 'glmnet'. For PLS, we use the 'fit.simpls' function from the R package called 'pls'.

1. Ridge regression

We use ridge regression to reproduce the results of Jenkins et al[1] and to show that when we train the model on a dataset containing the features selected by our feature selection method, the resulting mean R2 is higher than the one that can be obtained by training on the data with the 50 features suggested by Jenkins et al[1].

We also use Ridge regression to show the baseline R2 result using all the features available in the big data matrix.

2. Lasso regression

We use Lasso to further select features (secondary feature selection) to be used on an independent dataset. Selected features are the ones that

   a. Have non-zero coefficient at least once if Lasso is run in a 10-fold cross-validation fashion
   b. Have non-zero coefficient if Lasso is run on the whole dataset (no 10-fold cross-validation)

In case a. we always retrain the data on the independent dataset using the features selected by Lasso. We also utilize 10-fold cross-validation to get the mean R2 on test set chosen from the independent dataset. We use this method to validate the chosen features on independent data.

In case b. we do try both retraining on the independent dataset and transferring the model (weights and regularization coefficient) directly from the original training (IVF) dataset that was used for Lasso feature selection to the independent (Aging) dataset. We use this method to achieve main aim of this project (see Introduction). We try feature centering around zero during training and well as outcome centering, although we do the latter only to give us diagnostic understanding of the model and we acknowledge that outcome centering is not possible when the outcome is unknown.

3. Partial Least-squares Regression (PLS)

PLS is another method we use, that, in this case, happens to give the best result for model transfer between the two datasets getting us the closest to the main aim of this project. Since the 'pslr' function that is used in R to fit Partial Least-squares Regression does not execute with our big data, we had to custom write the 'plsr' function using the 'simpls.fit' function directly as suggested in the original paper[2] as a method to reduce computational overhead. To choose the

number of components, we use the root mean-squared error metric (RMSE) and 10-fold cross-validation.

Results and discussion:

1. Using all the features included in the data matrix, we get the following results:

Data	Features	Mean R2 in test set	Mean R2 in training set
IVF	850K	0.752	0.999
Aging	450K	0.758	0.999

The mean training R2 is huge suggesting the possibility of overfitting. Since we're already applying regularization, another way to reduce overfitting is to reduce the number of features.

2. Reproducing the Jenkins et al[1] results

We requested the data (Aging data) from the authors and received slightly more samples (371) than what was originally used in the paper (329). The paper defined R2 as the correlation coefficient squared between the actual outcomes and the predicted outcomes. This does not equal the variance explained by the regularized regression model and can yield high R2 even if the prediction has a high mean-squared error. This definition gives an R2 of 0.819 +/- 0.054. Defining R2 as the variance explained, we obtain a slightly lower value: 0.801 +/- 0.054.

We tried the 50 features suggested by the paper on the IVF dataset. 10-fold cross-validation gave an R2 of 0.549 +/- 0.070. This is much worse than we expected based on the results on the Aging dataset. Hence, we conclude that these 50 features do not generalize well.

3. Validating primary feature reduction method. The results (mean R2 +/- standard deviation) of Ridge regression are summarized in the table below.

Data	HMR	HMR + prom	HMR no prom	Jenkins et al.
IVF	0.811 +/- 0.043	0.730 +/- 0.046	0.805 +/- 0.047	0.549 +/- 0.070
Aging	0.824 +/- 0.075	0.802 +/- 0.061	0.816 +/- 0.083	0.801 +/- 0.056

All of our generated predictor sets perform better than the 50 features suggested by Jenkins et al.

4. Validating secondary feature selection, 10-fold cross-validation method. The results (mean R2 +/- standard deviation) of Lasso are summarized in the table below.

Data	HMR	HMR + prom	HMR no prom
IVF	0.798 +/- 0.055	0.744 +/- 0.040	0.804 +/- 0.053
Aging	0.840 +/- 0.069	0.785 +/- 0.073	0.822 +/- 0.097

We are focusing on the results on the Aging data here, since the regression on Aging data only used the features that were selected by the Lasso on the IVF dataset. Note that the HMR features yielded the highest mean R2, 0.840 so far. Hence, we conclude that the secondary feature selection method is also useful.

Moreover, we show that similar results can be obtained without using 10-fold cross-validation:

Data	HMR	HMR + prom	HMR no prom
Aging	0.832 +/- 0.052	0.766 +/- 0.161	0.829 +/- 0.089

5. Results of transferring the weights and regularization coefficient from IVF to Aging data

Without feature centering around zero (subtracting the mean of each predictor from its value), the R2 values were very low: the best R2 value was 0.329 for the HMR + prom feature set, the R2 was negative for the Aging dataset with the HMR and HMR no prom predictor sets.
With feature centering, the R2 values significantly improve for all predictor sets. It is clear that if we center features, we will never get a good prediction on a new dataset if the outcome (age) averages are not similar. This is because centering sets the average value of the features, corresponding the average outcome, to be the same, so if this is not true, we introduce bias. In our case, the average age in the IVF data is 4 years higher. Hence, the R2 further improves once we center the outcome of both datasets around zero. Since feature centering improved our results, we must conclude that the original bias was bigger than the one introduced by feature centering. This bias could have arisen because of the difference between the Illumina 450K and 850K arrays, the difference between the light measuring device used in the experiments, or simply due to the geographical difference between the two cohorts in the two datasets.

The following three figures illustrate how feature centering and then outcome centering improves the result:

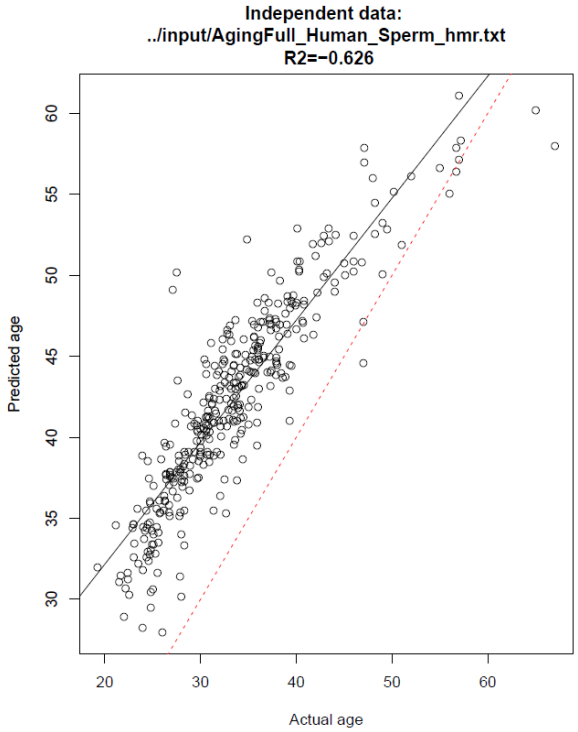

Figure 1 - no centering

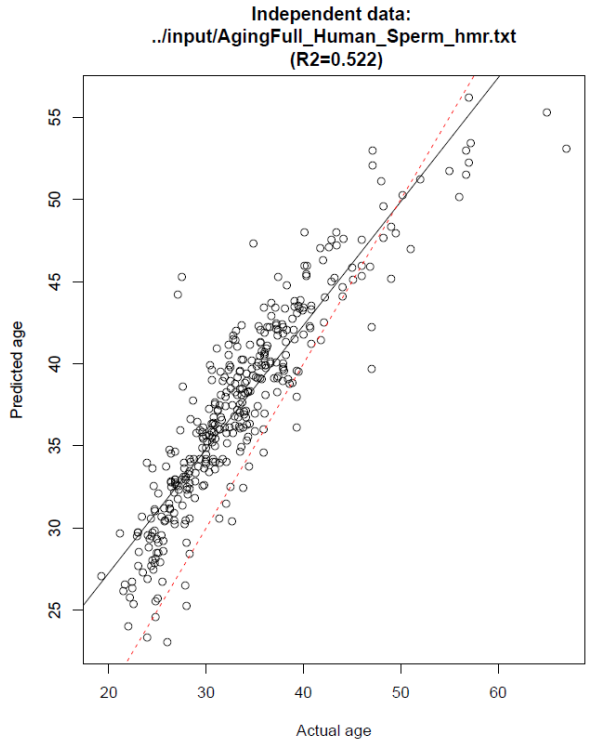

Figure 2 - feature centering

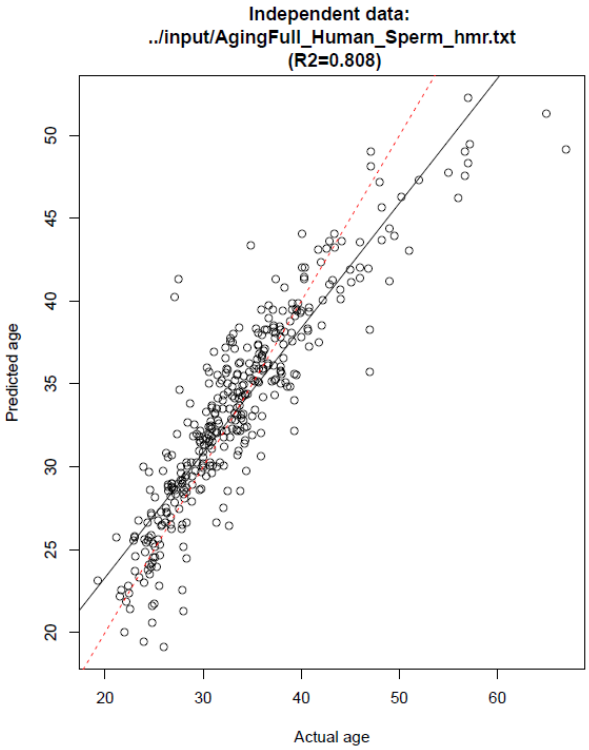

Figure 3 - feature and age centering

We then tried PLS to see whether it can improve these results. Figure 4 shows that PLS gives an R2 value in the acceptable range (>0.7) when trained on the IVF dataset and evaluated on the Aging dataset using the HMR + prom feature set.

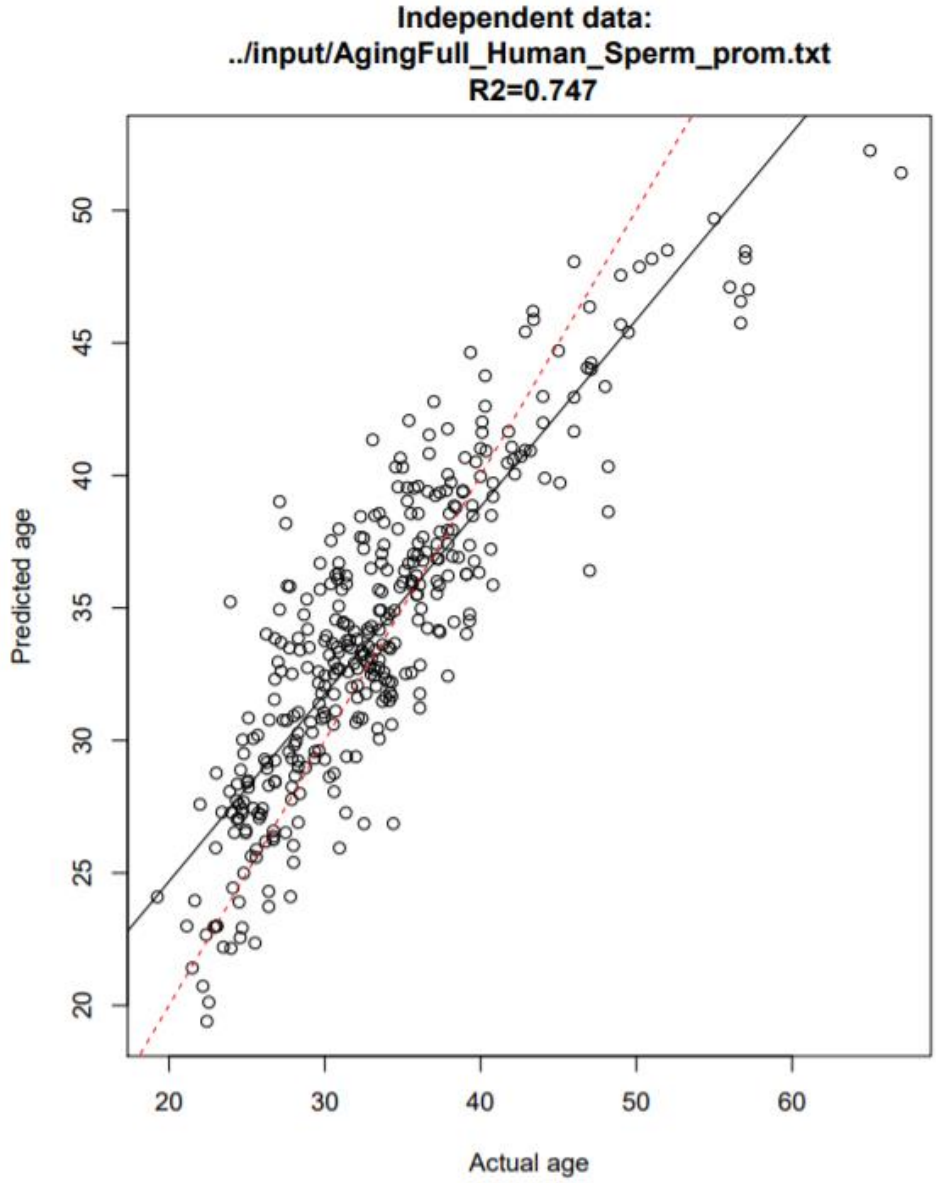

Figure 4 - Result of PLS trained on the IVF data and evaluated on the Aging data

The comparison between Lasso, Lasso with feature centering and PLS is given in the table below.

Data	HMR	HMR + prom	HMR no prom
Lasso	-0.626	0.329	-0.704
Lasso with centering	0.522	0.470	0.512
PLS	0.644	0.748	0.542

From the above table, we can see that PLS outperforms Lasso. Hence, we recommended using PLS to build a fixed machine learning model (with weights and regularization coefficient) that can be used to predict age from new independent sperm methylation datasets. PLS might perform this well, because it only uses 20 components, whereas Lasso used around 550 features. Hence PLS is less prone to overfitting. Another reason why PLS performs this well could be that while it makes the components uncorrelated, it happens to get rid of the bias in methylation levels between these two datasets. If the latter is true, then we might have just gotten lucky and hence we cannot guarantee superior performance of PLS without confirming the result on a new independent dataset.

Conclusion:

Our main aim was to find a model (feature set used, their weights and regularization coefficient) to predict age from sperm methylation levels measured using the Illumina 450K array or 850K array. First, we showed that with our feature selection method, it is possible to achieve a higher R-squared value for both the Aging and IVF datasets than with the 50 features suggested by the Jenkins et al[1] and with the full predictor set. We then showed that whereas the features discussed in the paper only work well for the Aging dataset, our features work well for both the Aging and the IVF dataset. Hence our features seem to generalize well on independent datasets. Finally, we showed that partial least squares regression (PLS) outperforms Lasso in terms of R-squared on an independent dataset using our selected features. In fact, PLS gave an R2 value of 0.748 when trained on the IVF dataset and evaluated on the Aging dataset using the HMR + prom predictor set.

1. Jenkins, T. G., Aston, K. I., Cairns, B., Smith, A. & Carrell, D. T. Paternal germ line aging: DNA methylation age prediction from human sperm. *BMC Genomics* **19**, 1–10 (2018).
2. Wehrens, R. The pls Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Softw.* **18**, (2007).