MRR Project

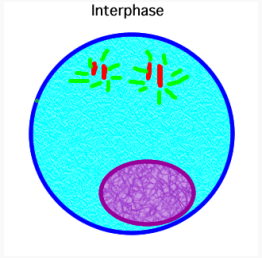
Prediction of Yeast gene expression

Kossir El Mehdi – Bougrine Rayan

TABLE OF CONTENTS

01 Motivation

Eukaryotic cell division is a complex process, with many layers of regulation at the level of gene transcription, protein production, localization, modification, and degradation



Cell cycle consists four phases. It moves through each phase, it also passes through several checkpoints.

Many genes specific to the cell cycle are regulated transcriptionally and are expressed just before they are needed

The precise determination of the moment of maximum expression is, therefore, important for understanding the cell division process.

02 Data presentation

Data set

- The data used in this analysis comes from R package «**spls**» it's the yeast Cell Cycle dataset used in Chun and Keles (2010).
- **542 genes and 106 transcription factors**
- **Each column in the target variable corresponds to mRNA levels.**

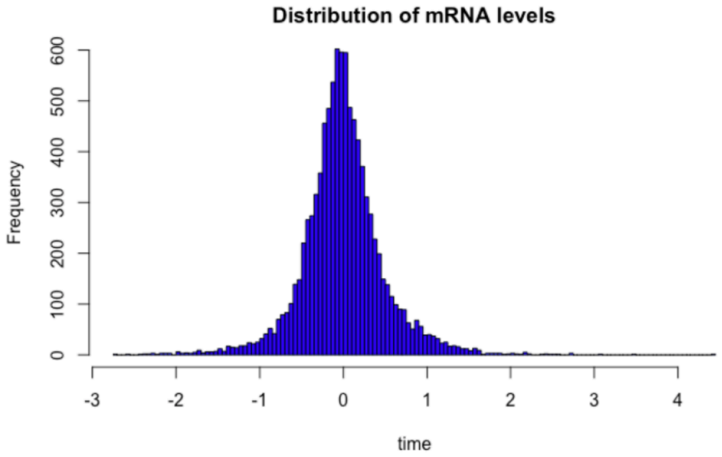
Note :

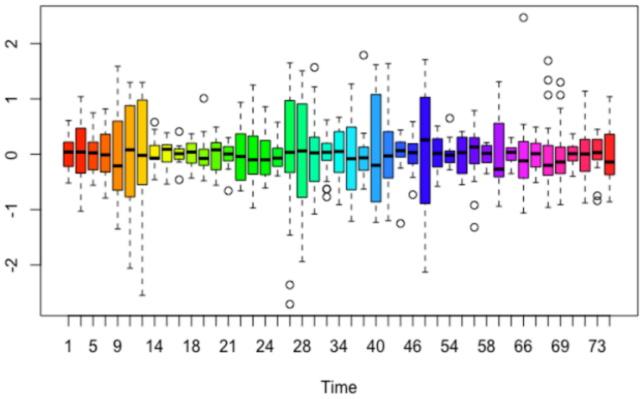
Here we present a unique approach to identify yeast cell cycle-regulated gene expressions. We rely on ChIP-chip binding and putative binding sites of transcription factors.

All our variables are quantitative, and we don't have any missing values.

03 mRNA levels problem

In order to identify the relationships between **transcription factors** and **gene expression**, we summarize our mRNA levels **(total of 18 measurements).**



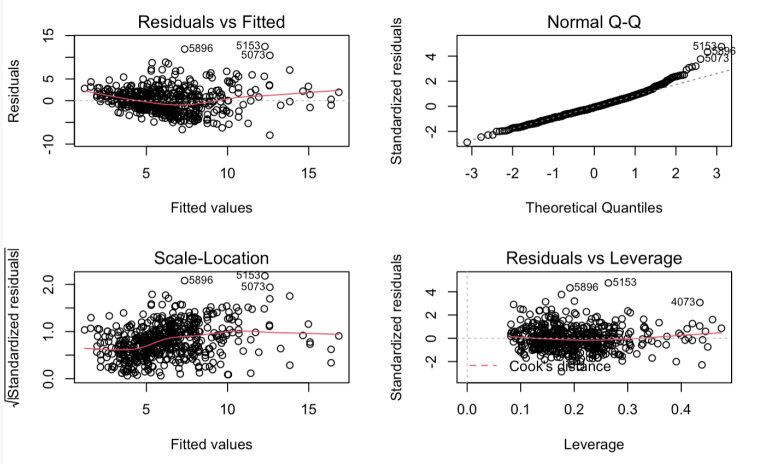


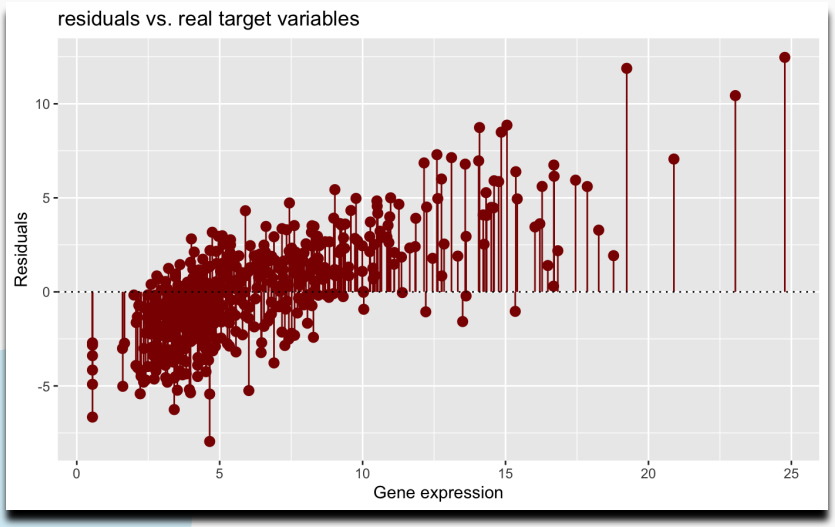
As we can see we assume that we have a **gaussian process**

Our Data are log2 transformed
We use the sum of the absolute values to summarize our target variable

04 Multiple linear regression

We fit a multiple linear regression model to our data .



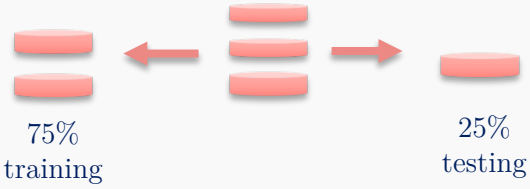


As we can see our model is **moderately efficient**, this is due to the high **collinearity between the variables**.

RMSE = 2.73832

05 Variable Selection

Using stepwise regression methods



Method	Forward	Backward	Stepwise
AIC	2730.15	2729.895	2727.707

We note that we were able to reduce the number of **TFs from 106 to 29** while maintaining approximatively the predictive quality of the model

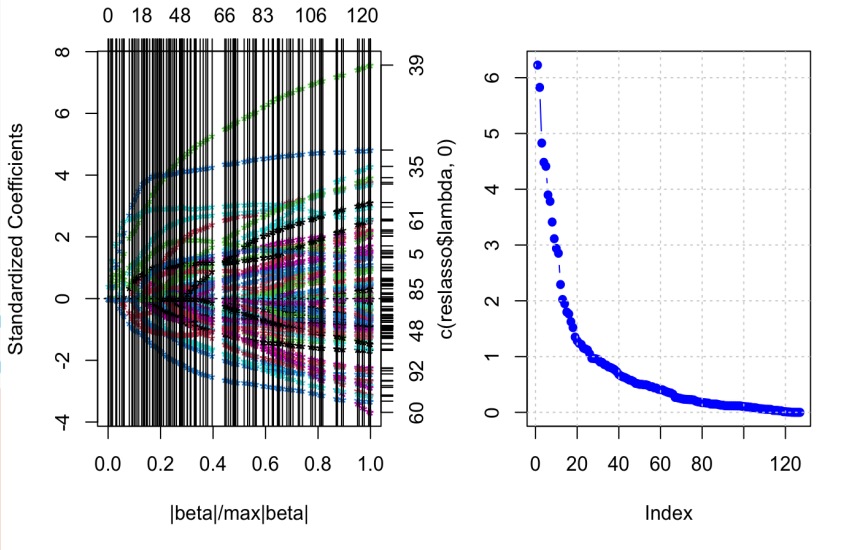
Subsest of tancription factors found using stepwise regression				
ARG81_YPD	ARG81_YPD	ARG81_YPD	ARG81_YPD	ARG81_YPD
FZF1_YPD	GCR1_YPD	HIR2_YPD	IXR1_YPD	MBP1_YPD
MET31_YPD	MSN4_YPD	NDD1_YPD	PUT3_YPD	RAP1_YPD
RFX1_YPD	RIM101_YPD	ROX1_YPD	SFP1_YPD	SIP4_YPD
SOK2_YPD	STE12_YPD	SUM1_YPD	SWI5_YPD	SWI6_YPD
YJL206C_YPD	ZMS1_YPD	STP1_YPD	HAP4_YPD	

RMSE =2.836957

Variable Selection

Using ridge, lasso and elastic net regression

Method	Ridge	Lasso	Elastic
RMSE	0.9368407	0.8172558	0.8992425

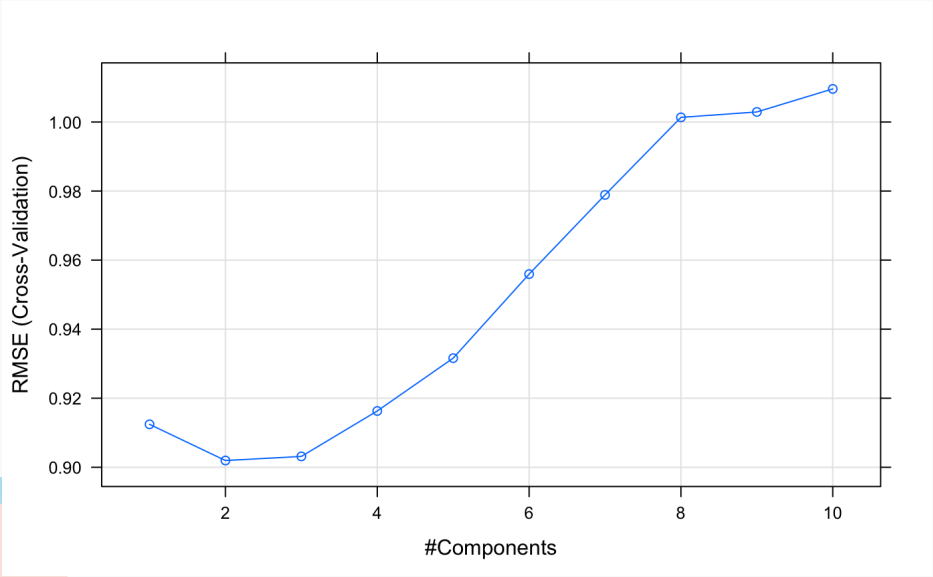


By using penalized regression methods we were able to improve the quality of our model and reduce variables specially using **LASSO**

106->34 variables

06 PCR regression

We preform a PCR to our data. The basic idea behind it is to calculate the principal components and then use some of these components as predictors in a linear regression model fitted using the typical least squares procedure.

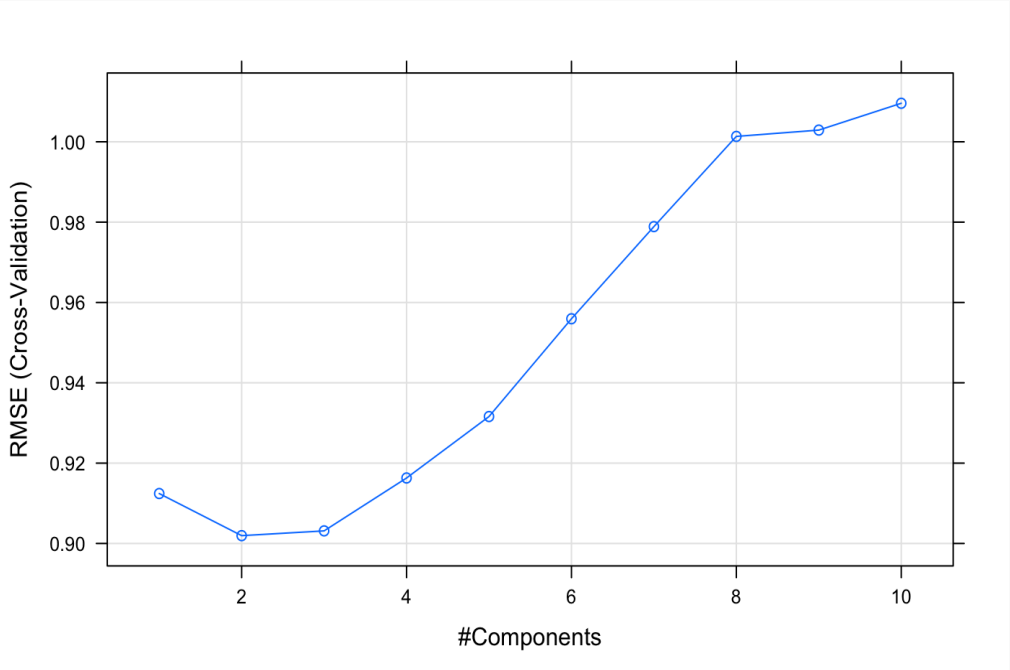


6 components
&
RMSE =0.8992425

A possible drawback of PCR is that we have no guarantee that the selected principal components are associated with the outcome.

06 PLS regression

An alternative to PCR is the Partial Least Squares (PLS) regression, which identifies new principal components that not only summarizes the original predictors, but also that are related to the outcome. These components are then used to fit the regression model.



2 components
&
RMSE =0.8530325

CONCLUSION

As we can see the best model remains the lasso, the subset of the significant transcription factors is in the .Rmd file. For further analysis we can fit a sine curve to our Y variables and transform them into categorical one in order to know through the transcription factors if we will have a strong gene expression or not.

RESOURCES...

- http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/152-principal-component-and-partial-least-squares-regression-essentials/#partial-least-squares-regression
- https://www.nature.com/scitable/topicpage/eukaryotes-and-cell-cycle-14046014/#:~:text=In%20eukaryotes%2C%20the%20cell%20cycle,when%20the%20cell%20actually%20divides.&text=Later%2C%20during%20G2%2C%20the,readiness%20to%20proceed%20to%20mitosis.
- https://www.wikipedia.org/
- https://www.molbiolcell.org/doi/full/10.1091/mbc.9.12.3273
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3734186/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC25624/