Predicting Early Bulking in Potatoes

Fetsje Bijma Alessandro Di Bucchianico* Eric Cator
Henk Don Patrick Hafkenscheid Jakub Nowotarski
Bijan Ranjbar-Sahraei

Abstract

Early bulking of potatoes is important for potato breeders for several reasons, including flexibility in scheduling and less influence of weather conditions. In this paper we use statistical models to model tuber growth in order to identify which existing varieties allow for early bulking. We also investigate which genetic properties (SNP's) may be important for early bulking.

KEYWORDS: early bulking, SNP, variance stabilizing transformation, linear regression, sparse data, elastic net

1 Introduction

In this section we provide the necessary background for the problem, and state the main research questions.

1.1 Company background

HZPC (`www.hzpc.nl`) is the world leading developer and seller of high quality seed potatoes. It is an internationally operating Dutch company with head quarters in Joure. HZPC has 320 employees, 800 growers and 55 breeders on 19 locations. To serve its customers better, HZPC has an R&D department in Metslawier. The main goal is to develop new varieties of potatoes that meet the needs of consumers and industrial partners by advanced data-driven breeding techniques.

1.2 Problem description

Tuber bulking is the 4th growth stage in the development of a potato (see Figure 1). Tuber cells expand with the accumulation of water, nutrients and carbohydrates. Tubers become the dominant site for deposition of carbohydrates and mobile inorganic nutrients.

*Corresponding author.

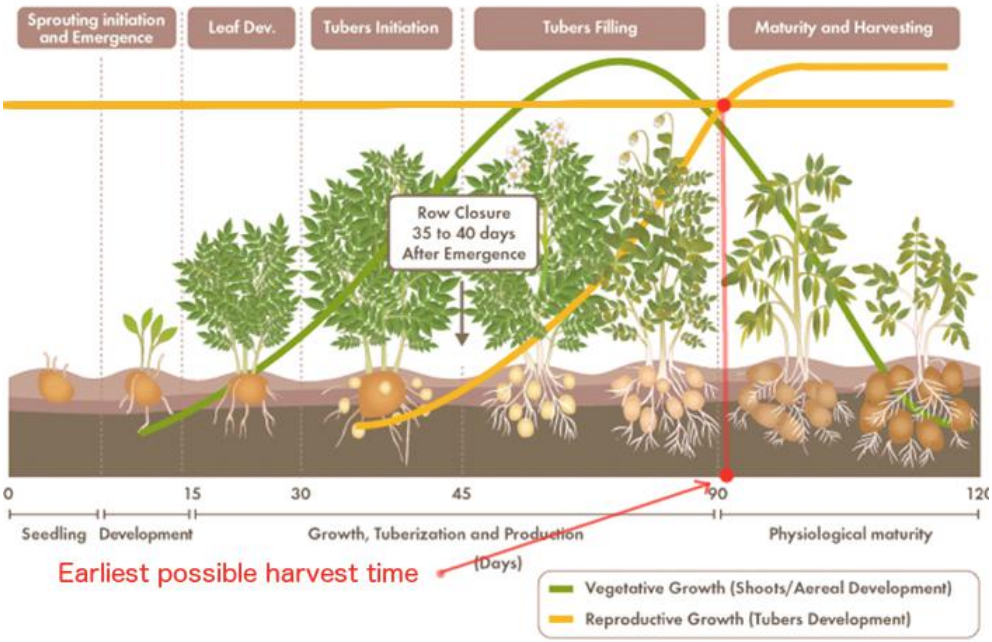

Figure 1: Early bulking (source: `www.sqm.com`).

HZPC wishes to breed early bulking varieties in order to be able to harvest as early as possible high quantities of tubers with desirable sizes . The benefits of early bulking are the opportunities to have new harvests as early as possible, more flexibility with scheduling (since it takes less until harvest), and less influence of climate factors such as rain and humidity.

In order to search for early bulking varieties in an efficient way, there is a need for a statistical model that predicts the tuber filling in length and volume in time per variety and to find the genetic parameters that have a significant effect on early bulking performance. Furthermore, a simple and efficient strategy should be designed for selection of early bulking varieties. More concretely, we will address the following research questions:

Research questions:

Question 1 How to model tuber growth and predict which varieties are more likely to bulk early?

Question 2 How to identify important genetic properties for early bulking?

With respect to Question 1, HZPC is interested in the mass of harvested tubers with

tuber size 45 mm or more as well as subtraits of varieties like tuber filling (length and diameter of tubers) and the number of tubers per plant. Tuber size is commonly defined by potato breeders in terms of "square size", i.e. the length of the side of the smallest square in which the tuber fits. A complication for the development of models for Question 1 is that the number of tubers and sizes of the tubers are correlated (if there are more tubers, then they are likely to be smaller).

For Question 2, the goal is to find models with causal explanations in terms of DNA differences so that one effectively and efficiently measure early bulking in breeding programs. We note that a complication here is that important traits are usually determined by several genes simultaneously.

2 Available data

In this section we describe the data that we could use to address the research questions. In Subsection 2.1 we describe the field data for Question 1, while in Subsection 2.2 we describe the genetic data for Question 2.

2.1 Tuber data

Data of trial fields of the the years 2011-2015 were made available to us by HZPC. These trial fields were laid out using the following experimental design (see also Figure 2):

- 100 varieties of tubers.
- 4 different harvest times.
- 2 replicates per harvest time.

The data set of the year 2015 is very detailed and contains for every individual tuber length, width, height, square size (as defined in Subsection 1.2, weight and volume. For the previous years (2011–2014) only summarized data were available through the number of tubers and total weight for each field plot, square size category and harvest time.

2.2 Genetic data

The genetic data set made available by HZPC is in the form of frequency counts of SNP's (single nucleotide polymorphisms, pronounced "snips"). A SNP is a genetic variation at a specific position in the genome in the form of the replacement of a single nucleotide at a specific base location. The SNP's in the data set only allow two different alleles (i.e., two different nucleotides), so a 0 indicates no variation (the most frequent nucleotide) and a 1 indicates the genetic variation (the alternative nucleotide, which must occur in at least 1% of the population. The values in the data set are integers from 0 to 4, since the SNP's are determined for the 4 chromosomes of a tuber

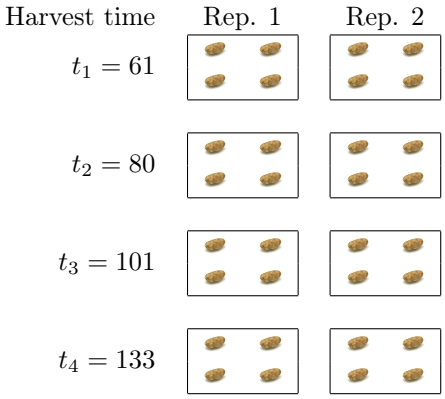

Figure 2: Experimental design

(2 from the father and 2 from the mother). SNP data are available for 113 varieties for the years 2011-2014, and for only 12 of the varieties of the 2015 field trial.

3 Tuber growth modelling

Before we try to make a statistical model for tuber growth, we performed a small exploratory data analysis to check for data quality issues, variation between individual tubers as well as get an idea of the time evolution of tuber growth. Figure 3 shows that there is considerable variation between the individual tubers within varieties. There is no clear difference between the two replicates (indicated by different colours). The main interest of HZPC is the weight of tubers with a square size of at least 45 mm. After examining various plots, we found that a log-log relationship seems to be a suitable model for tuber size and tuber weight since the data points in the plots lie reasonably well on a straight line and the deviations from the straight line are less than for the other standard relationships that we tried out (see Figure 4). So the log-log transformation is also a variance stabilizing transformation. Other plots showed a moderate plot effect, i.e., there is some variation between the weights of tubers of the same variety but planted on different parts (plots) of the experimental field. Since the main interest of HZPC is the total weight of tubers with square size at least 45 mm, we decided to fit a joint model for log-weight and log-square size as function of time and number of tubers instead of model for weight and time. This model allows us to predict yield as function of time. In view of the considerable variation between tubers, we decided to model every tuber individually. Based on Figure 3 we assume a quadratic function as a simple form for the time evolution. To be more precise, we

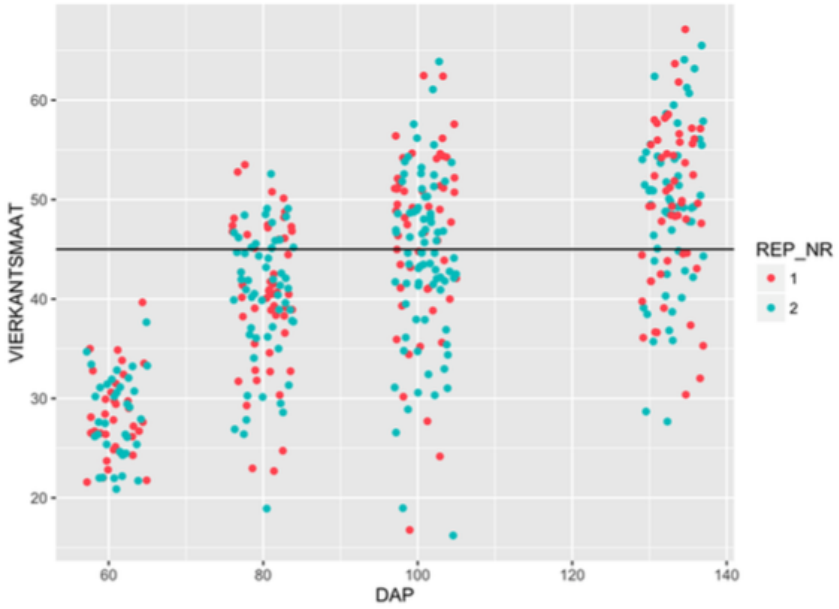

Figure 3: Scatter plot of tuber size ("square size") as function of days after planting

fitted the following linear regression model[1]. Define for each variety v the following quantities:

- $Y_1^v(t)$ = log of **square size** of the tubers at time t
- $Y_2^v(t)$ = log of **weight** of the tubers at time t
- $N^v(t)$ number of tubers belonging to the same potato plant at time t.

Then our model is

$$(Y_1^v(t), Y_2^v(t)) = \begin{pmatrix} 1 & t & t^2 & N^v(t) \end{pmatrix} \begin{pmatrix} \beta_{11}^v & \beta_{12}^v \\ \beta_{21}^v & \beta_{22}^v \\ \beta_{31}^v & \beta_{32}^v \\ \lambda_1^v & \lambda_2^v \end{pmatrix} + (\varepsilon_1^v(t), \varepsilon_2^v(t))\,., \qquad (1)$$

where $(\varepsilon_1^v(t), \varepsilon_2^v(t)) \sim \mathcal{N}((0,0), \Sigma^v)$. We obtained estimates for the parameters β, λ and Σ by using maximum likelihood.

In order to predict the total weight of a potato plant, we multiplied the estimates $N^v(t)$ into the model and compute for each t the expected total weight of big tubers (e.g., tubers with square size at least 45 mm). A graphical representation of our results is presented in Figure 5.

[1]Note that although the time evolution is described as a quadratic function, the parameters appear in a linear way in the regression function.

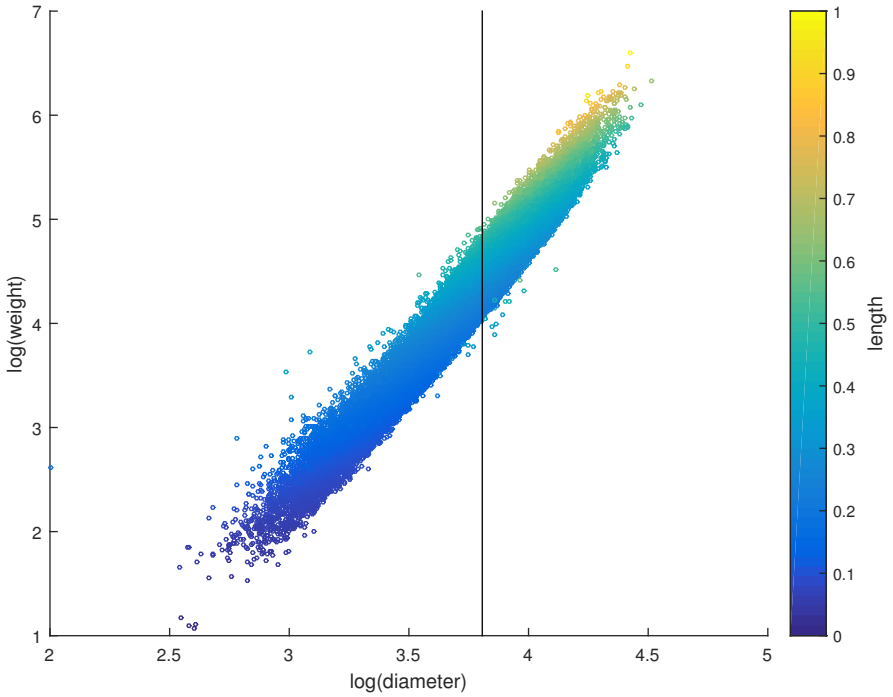

Figure 4: Log-log plot of tuber size ("square size") and weight as function of days after planting

4 Genetic properties and early bulking

In the previous section we made models to predict the early bulking properties of existing varieties. In order to develop new varieties with favourable early bulking performance, it is important to study the genetic properties of early bulking varieties. Therefore we now turn to the genetic data described in Subsection 2.2. Our approach consists in trying to build a linear regression model with the SNP's as independent (explanatory) variables and the total weights per variety of the tubers with square size at least 45 mm. Since the data set contains 113 varieties and 11763 SNP's, we have many more parameters than observations. Thus we cannot perform an ordinary linear regression. However, we may safely assume that only a few SNP's may influence the early bulking performance of a variety. In other words, a sparse model may be appropriate. Sparse models may be fitted using special variants of linear regression, in which the least squares criterion is replaced by another criterion that puts an extra penalty on the number of selected explanatory variables. These variants are sophisticated counterparts of the traditional backward and forward model selection methods. The first example of such a method is the lasso introduced by Tibshirani (see Tibshirani (1996), which makes use of an ℓ_1-criterion rather than the standard

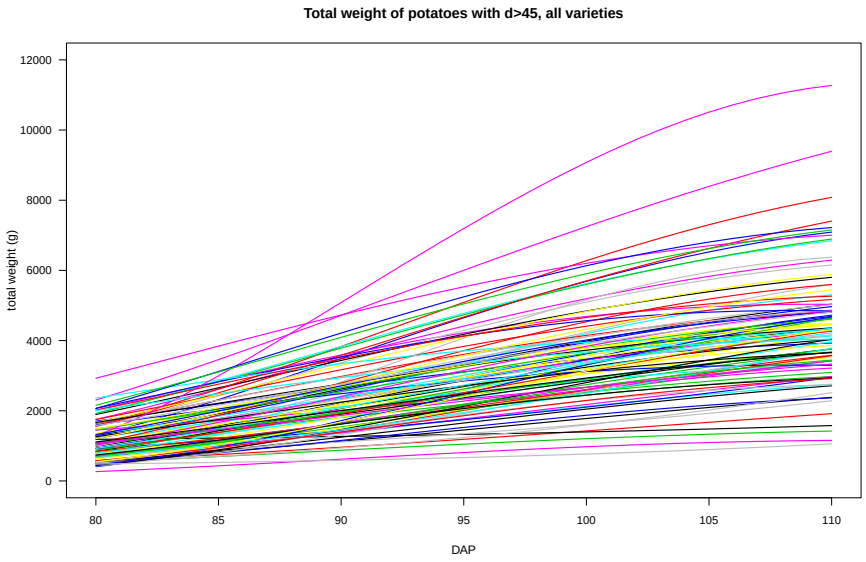

Figure 5: Time profile of total weight of big tubers

ℓ_2-criterion used in the least squares approach. Further refinements are the elastic net in which the criterion involves both an ℓ_1-term and an ℓ_2-term, with an automatic choice of the relative weights of these terms (see Zou and Hastie (2005)) and least angle regression which features a continuous way of including explanatory variables (see Efron et al. (2004)). We refer to Hesterberg et al. (2008) for a gentle and lucid introduction to these advanced regression methods and to Hastie et al. (2015) for an accessible monograph on methods for sparse data like in our case (i.e., we expect that only a few SNP's will influence early bulking performance).

For our analysis we used the data of the $2011-2014$ field trials since they contain SNP data for 113 varieties. It should be noted however, that there are several missing values. Certain SNP's may be difficult to obtain since the maximum number of missing values per SNP equals 51 and there are 266 SNP's with more than 10% missing values.

We followed a two-step approach:

1. apply multiple imputation to fill in missing values

2. apply elastic net to preselect important SNP's

The elastic net regression method requires complete cases. One could leave out the SNP's with missing values, but that would lead to an underestimation of the standard error. Therefore we decided to apply imputation. One should choose a suitable imputation method by considering the possible mechanism causing the missing data. In our case the SNP observations were obtained per variety using a complicated

procedure to extract the relevant genetic data. Due to the complicated nature of the extraction procedure, determination of SNP's may fail at certain locations in the genome. Since there is no indication that this depends on the variety, the missing data mechanism that is appropriate is "missing completely at random" as introduced in Rubin (1976). We chose "predictive mean matching" as imputation method, since it is likely that varieties with similar SNP values for non-missing data entries will have similar SNP entries for missing data (see Van Buuren (2012) for a comprehensive overview of both theoretical and practical issues related to imputation). The analysis was performed using the statistical software `R` with the following packages:

1. the mice package for Step 1
2. the glmnet package for Step 2

In our analysis we used the following linear regression model:

$$\begin{pmatrix} W_1 \\ \vdots \\ W_{113} \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,11673} \\ \vdots & \vdots & & \vdots \\ 1 & x_{113,1} & \dots & x_{113,11673} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_{11673} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{113} \end{pmatrix}, \qquad (2)$$

where

- W_i is total weight of all tubers with square size 45+ of variety i ($i = 1, \dots, 113$).
- $x(i, j)$ is the value for SNP-j and variety i.

We selected elastic net as regression method instead of the lasso, since the lasso can only select as many variables as there are observations and it does not behave well in case of correlated independent variables (cf. (Hastie et al., 2015, Section 4.2)) so that is hard to make valid statements about which SNP's are important indicators for early bulking performance. Although there is SNP data for 113 varieties, we could only use the 69 varieties because of lacking early bulking data. We used elastic net with $\alpha = 0.5$ in order to have an equal weight of the ℓ_1- and ℓ_2-penalties, since this gave the best result. Cross-validation was used to obtain an optimal value of the λ parameter in the elastic net. After performing the elastic net analysis, we first removed all parameters (SNP's) that had a zero estimate which yielded a list of 140 SNP's worth investigating further. A further look at the results revealed some spurious effects caused by the effect that some SNP's had only 1 observation for a certain value and the remaining observations for one other value or for which there was only value of the SNP (in other words, such SNP's were constant for all varieties and thus no inference could be made for the effect of these SNP's). We removed these SNP's after doing the imputation and the elastic net analysis because it was much easier to remove this SNP's in the relatively small list of SNP's with complete data that remained. Note that we decided not to remove several SNP's with only 2 possible values, one value of which has only 2, 3 or 4 observations or SNP's with several values, but one which has only 1 observation (these SNP's could also lead to spurious effects, see e.g. Figure 6 for an illustration of the possible leverage effect in the form of box plots).

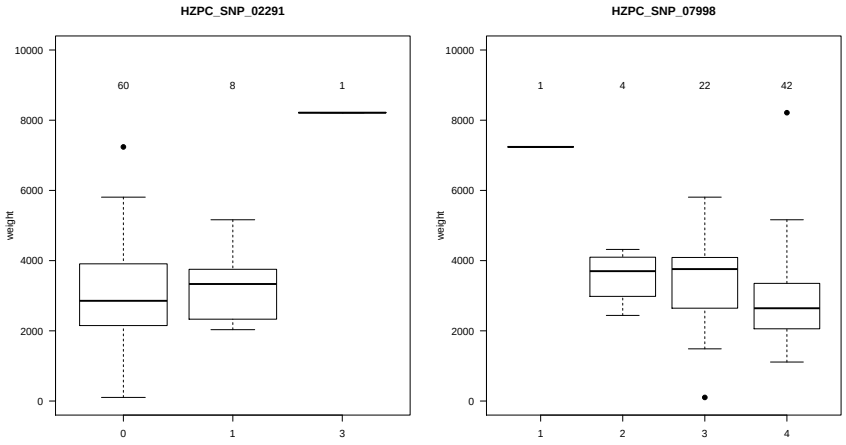

Figure 6: Selected SNP's, possibly spurious effects

After these steps, the list of potentially interesting SNP's reduced to 35 SNP's, only 1 of which has a positive effect (higher number of chromosomes with a modification give a higher weight) and the remaining 34 have a negative effect. So the data is indeed sparse, since this means that at most 1% of the SNP's seem to influence the early bulking performance. In view of possible correlations between the SNP's, one should be careful in identifying which SNP's influence early bulking performance.

5 Discussion

In this section we summarize our main conclusions and results. Based on these conclusions and results, we also indicate we also give some recommendations for future research.

5.1 Key insights

We list our key insights for the questions separately.

Question 1 How to model tuber growth and predict which varieties are more likely to bulk early?

1. There is a linear relation between log-weight and log-square size
2. A $\log - \log$ transformation has a variance stabilizing effect (this is important as constant variance is one of the assumptions of standard linear regression models)
3. There is a moderate plot effect

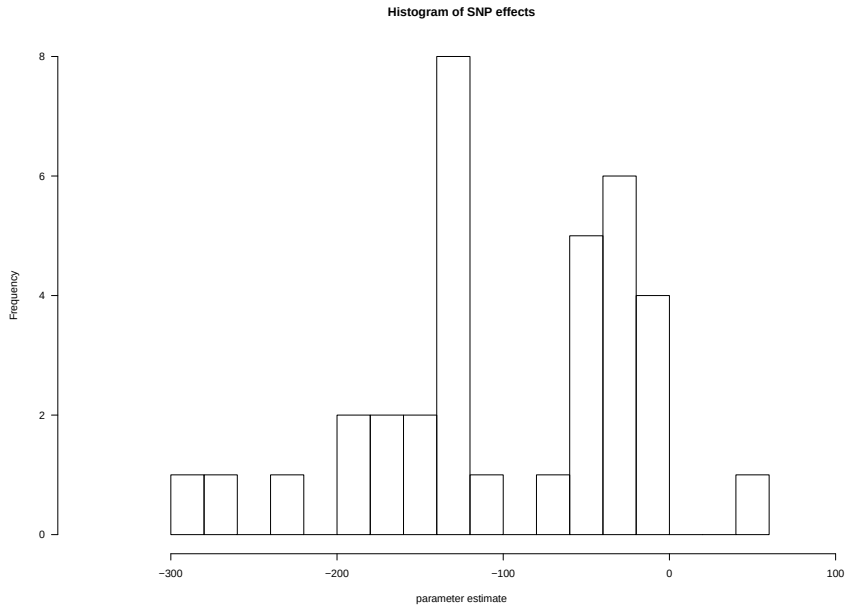

Figure 7: Selected SNPs, positive and negative effects

4. The number of tubers stabilizes after the second harvest time.

Question 2 How to identify important genetic properties (SNP's) for early bulking?

1. Do not include SNP's that are almost constant for all varieties since they may lead to spurious results.
2. A regression analysis with SNP's as predictor variables is possible in spite of the fact that there are many more SNP's than varieties using the elastic net approach
3. At most 1% of the SNP's show a significant effect.
4. Both positive and negative effects occur.

5.2 Future research

There are several ways in which research on the two main questions of this paper could be pursued.

For the growth modeling question, a further investigation of model accuracy should be undertaken and a sensitivity analysis should be performed with respect to harvest times. The growth model should also be enhanced with a plot effect in view of

the observed moderate plot effect. One should explore different shapes for the time profiles, e.g., $\sqrt{t}$.

For the genetic properties question, one should further explore the elastic net model. First of all one should perform modeling diagnostics in particular the normality assumption. In case of normality problems, one could try Box-Cox transformation or model the joint distribution. A further analysis of the relative importance of the significant SNP's is also important. There are several ways to do this, ranging from applying recently developed post-selection inference methods (see e.g., Section 6.3 of Hastie et al. (2015) and Chapter 11 of Bühlmann and van de Geer (2011) to variants of the lasso and elastic net that allow for group effects (i.e., methods that single out groups of highly correlated parameters, see e.g., Bach et al. (2012) for an overview of relevant methods). Apart from these statistical approaches, we also recommend to use more refined genetic data than SNP's.

Acknowledgement We would like to thank Jacqueline Verdijck-Lamers, Hans van Doorn, Rob Klooster, and Pieter-Jelte Lindenbergh of HZPC for introducing us to this interesting problem and for their extensive support to us during the SWI week.

References

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin, 2011.

S. van Buuren. *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, Florida, 2012.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity*. CRC Press, Boca Raton, Florida, 2015.

T. Hesterberg, N. Choi, L. Meier, and C. Fraley. Least angle and ℓ_1 penalized regression: A review. *Statistical Surveys*, 2:61–93, 2008.

D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.