Prediction of multiple traits in potato plants using gene expression profiles

Renny Doig
Supervisor: Dr. Hong Gu
Data courtesy of: Dr. Helen Tai

Abstract

Gene expression profiles of potato plants will be analyzed using multiple techniques designed for the ultra-high dimensional data which characterizes these profiles. Several phenotypic traits will be considered as response variables of interest; the goal being to determine which gene markers are important in predicting each response. Two approaches will be taken: feature selection and dimension reduction both of which are designed for dealing with high-dimensional data.

Introduction

Gene expression profiles are counts of how often a large number of genes, often in the thousands or tens of thousands, are transcribed by an individual and are commonly used to predict phenotypic expression in an individual. These types of data have many more predictors than observations therefore many standard statistical procedures do not work well. There are two approaches for dealing with high-dimensional data: feature selection and dimension reduction. Feature selection chooses a subset of the original predictors which are most useful in predicting the response while dimension reduction transforms the feature space so that the features can be projected onto a lower dimensional subspace that retains most of the predictive accuracy of the entire feature space. This analysis explores two different approaches to feature selection and one approach for dimension reduction. The purpose of these three techniques will be to reduce the problem of predicting phenotypes using gene expression to a lower-dimensional problem which is easier to analyze and interpret.

Methods

Data

Serial analysis of gene expression was used to collect gene expression data from potato plants seventy days after planting. Counts of 6248 gene transcriptions will be used as the initial set of predictors for all analyses in this thesis. There are five response variables which are all considered separately: emergence, flowering time, maturity, chipping, and glucose. Emergence is the number of days that it takes for a plant to emerge from the soil and reach a height of ten centimeters. Flowering time is the number of days before a plant starts to flower; plants that did not flower were not considered for analyses involving this response. Maturity is score of healthyness of a potato plant; the score is assigned at one hundred days and rated on a scale from one (dead) to nine (healthy). This was treated as a continuous quantity for the purposes of this analysis. The amount of glucose present in a plant was measured using strip assays. To measure chipping a chip slice was taken off of the potato plant, fried, then given a score between 10 (dark) and 100 (light).

Serial analysis of gene expression

Serial analysis of gene expression (SAGE) is a technique used to measure the frequency with which certain genes are expressed in an individual. The process described by Hu and Polyak (2006) first obtains short tag sequences for each gene. These tags correspond to mRNA which is created every time a gene is expressed by an individual. Then, using the sequencing platform Illumina, the tags are sequenced and counts of gene expression are collected. Prior to use the tags were filtered to retain only those genes which were expressed at least three times in at least one sample. Then a negative binomial test was performed to determine which tags showed significant variation.

Tools

Cross Validation

Cross validation is a method for estimating the test prediction error of a model by training a model on a training set then calculating the prediction error based on a separate validation set. This technique is often used as a means to determine the optimal values for hyper-parameters in a model by choosing the parameter value that minimizes the cross-validated error. The particular form of cross validation used in this analysis is K-fold cross validation. This splits the training set into K parts; $K-1$ of these parts are treated as the new training

set and the K^{th} partition is treated as the validation set. The cross-validation process is performed K times, each time corresponding to a different partition being treated as the validation set. The cross-validated error is the average of these K prediction errors. More details of this method can be found in Hastie et al. (2009).

Subsampling Ranking Forward selection

Subsampling Ranking Forward selection (SuRF) is a feature selection method developed by Liu et al. (2018) which uses subsampling, LASSO, and forward selection. The method takes stratified subsamples of the data and fits a generalized LASSO to each subsample. Each of the variables is ranked corresponding to the number of subsamples in which the variable is 'selected' by LASSO. A permutation-based approach is taken to estimate the empirical null distribution. Then forward selection is performed by computing the log-likelihood ratio statistic for candidate variables and comparing to the empirical null distribution. Significance is determined by computing the p-value based on the null distribution; if multiple variables are significant the variable with the highest rank is selected. This is implemented in the `SuRF` package developed by Liu and Kenney (2018).

Gradient boosted trees

Boosted trees are an aggregration of sequentially grown trees where each tree individually is a weak predictor. The gradient boosting approach fits each subsequent tree on the negative gradient of the loss function evaluated over the training data by least squares estimation. Under squared error loss this results in each tree being fit on the residuals from the previous tree in the sequence. There are three hyperparameters in the gradient boosted tree model: shrinkage, interaction depth, and the number of trees. The shrinkage paramter, ν, regulates the effect of adding the next tree at each step of the sequence; small values correspond to very little weight being placed on each individual tree. Interaction depth, J, determines the level of interaction allowed between predictors (ie. how many splits in each tree). The number of trees grown is denoted by M. The relative importance of each predictor used in training the boosted trees can be assessed by the reduction in squared error caused by selecting the predictor at a split in a tree averaged over all M trees. The implementation of gradient boosted trees used here is `xgboost` developed by Chen et al. (2019).

Sure Independence Screening

Sure Independence Screening (SIS) is a feature selection method that was introduced by Fan and Lv (2008) specifically for use on ultra-high dimensional data which gene expression profiles are a common example of. It is an iterative method where regression with a smoothly

clipped absolute deviation (SCAD) penalty is used to evaluate the correlation between the response and all of the (unselected) predictors; at each step the predictor most correlated with the response is chosen. To avoid selecting colinear predictors at each step of the process the residuals from the previous step are used as ther response in the current step. This process is repeated until a prespecified number of predictors are chosen. This is implemented in the `SIS` package developed by Saldana and Feng (2018).

Poisson log-Normal Principal Components Analysis

Principal Component Analysis (PCA) is a dimension reduction technique which identifies the orthogonal axes corresponding to the directions of greatest variance. Poisson-lognormal PCA (PLN-PCA) is an extension of PCA which models the data conditional on some latent variable as following a Poisson family distribution. This framework provides a means of separating the Poisson measurement error from the covariance between predictors. The latent variable follows a q-dimensional Gaussian distribution; this q defines the rank of the reduced feature space. Using gradient-based optimization the two components of the variance are estimated. It is important to note that unlike classical PCA, in PLN-PCA a solution of rank q is not nested in the solution of rank $q+1$. Following the framework for generalized linear models this technique allows for the offsets to be considered when modelling the Poisson error. PLN-PCA provides a unique definition of R^2 that will be used in assessing the accuracy of lower rank approximations:

$$R_q^2 = (\ell_q - \ell_{min})/(\ell_{max} - \ell_{min})$$

where ℓ_q is the log-likelihood of the q-dimensional reduction, ℓ_{max} is the log-likelihood of the saturated model, and ℓ_{min} is the log-likelihood of the null model. This methodology was implemented in `R` in the `PLNmodels` package which was developed by Chiquet et al. (2018a). In this package only the 473 predictors with the highest relative importance are given importance ranks.

Principal Component Regression

Principal Component Regression (PCR) is a modification of the typical multiple regression framework which uses derived inputs in place of the original predictors. The process is identical to the multiple regression framework except that the first M principal components are used in place of the original predictors.

Analysis

This section describes how the tools described previously will be used to: a) rank and select a subset of the gene tags and b) define a new feature space with reduced dimension. The analyses are carried out with each of five traits as responses in separate models.

Feature Selection

Both SuRF and gradient boosted trees were used as feature selection tools to assess which gene tags were most strongly related to each of the physical traits. SuRF was used to select tags by choosing all predictors with a p-value below a pre-specified cut-off. To account for the sequencing depth of each observation the data were normalized by the sequencing depth for each observation. All traits were assumed to follow Gaussian distribution. Since SuRF is conservative in its estimate of the p-values a cut-off of 0.2 will be used to select features.

Gradient boosted trees were used to fit each of the traits to the gene profiles. Ten-fold cross-validation was used to obtain the optimal values for the hyperparameters. Typical values for the shrinkage parameter are small (< 0.1) so the values of ν that were tested were 0.001, 0.01, 0.1, and 1. Hastie et al. (2009) suggest that if the interaction depth is greater than three then a value of six is often sufficient so the range of values tested for interaction depth is 1, 2, 3, and 6. The number of trees grown is often a large number so the values checked for M are 200, 400, 600, and 800. The relative importance measure of the boosted trees was used to rank the predictors.

Dimension Reduction

PLN-PCA was used to find a reduced feature space which would be accurate in predicting the glucose in a potato. It was found that implementing PLN-PCA using the entire set of predictors did not produce a sufficiently concise feature space. Figure A.1 shows that incorporating a latent space with dimension as large as 25 still did not yield an R^2 which leveled off; this is not was expected given the results quoted by Chiquet et al. (2018b). To fix this issue SIS was first used to reduce the number of features to 100 using glucose as a response. The reduced set of predictors was then analyzed using PLN-PCA to obtain a reduced rank feature space. To assess the adequacy of this reduced feature space in predicting glucose a PCR was fit using the scores from the component analysis. Unlike the feature selection component of this analysis glucose is the only trait considered as a response for assessing the dimension reduction techniques.

Results

Feature Selection

Cross-Validation

Cross-validation was performed to obtain optimal hyperparameters for the gradient-boosted trees. This was done on five separate models each corresponding to one of the phylogenic traits. Table 1 summarizes the optimal values for the three hyperparameters. These values are the values that will be used to train a gradient-boosted tree and obtain the importance rankings. The relationship between the cross-validated error and the hyperparameters is illustrated in Figure 1.

Trait	Shrinkage	Interaction Depth	No. of Trees
Chipping Score	0.01	3	400
Emergence	0.01	1	400
Glucose	0.01	1	400
Flowering Time	0.01	1	200
Maturity Score	0.1	1	200

Table 1: Table of optimal hyperparameter values for gradient-boosted trees for each phylogenic trait.

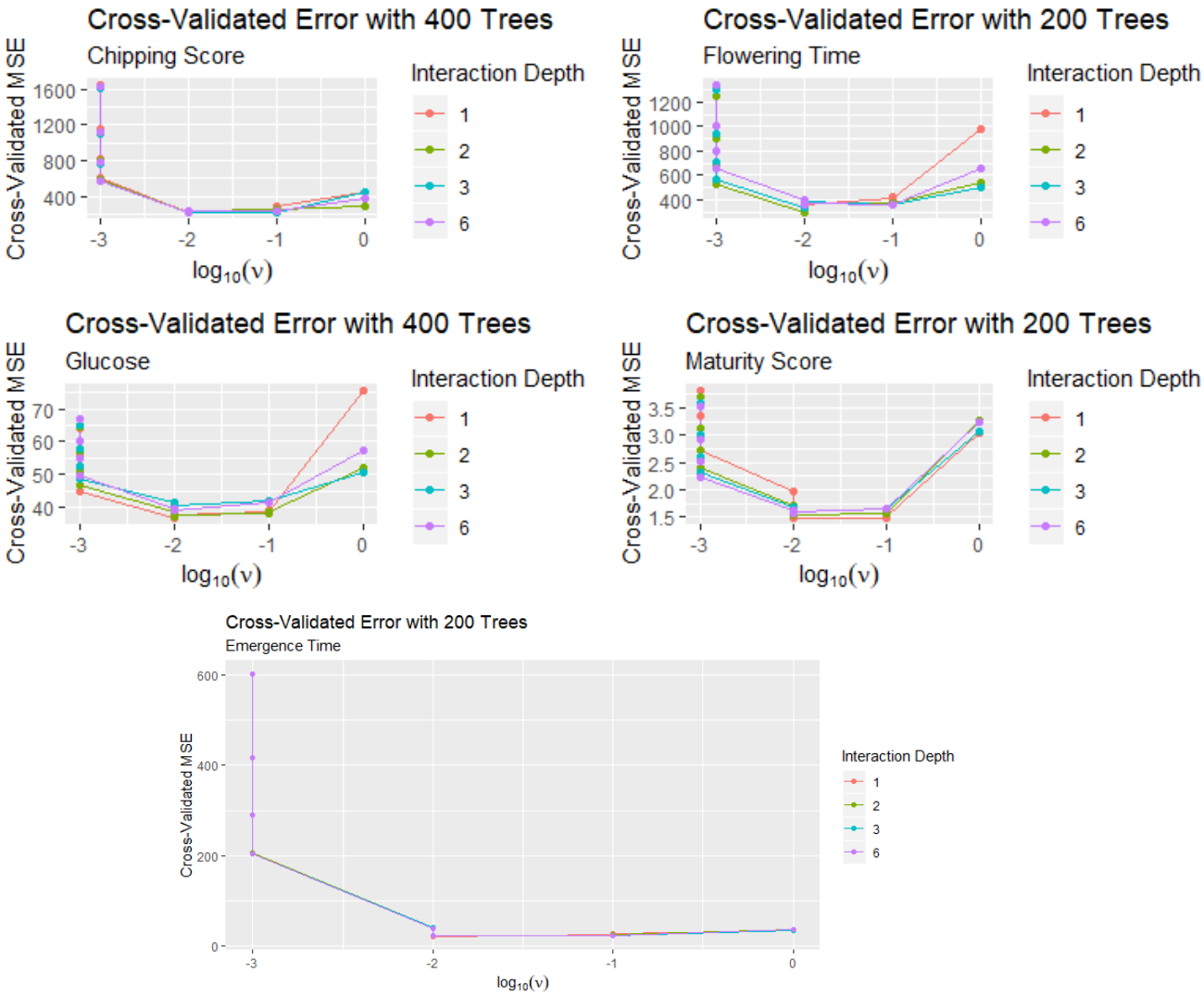

Figure 1: Cross-validated errors for all five phenotypic traits plotted against $\log_{10}$ of the shrinkage parameter and for each interaction depth. Only plots for the optimal number of trees is shown.

Selected Features

Before comparing the variables selected by SuRF and the gradient-boosted trees the table below, Table 2, compares the cross-validated error for each methods performance on all five of the responses.

Trait	SuRF	Boosted Trees	SuRF/Trees (%)
Chipping Score	173.98	207.87	83.7
Emergence	10.94	19.66	55.6
Glucose	33.95	36.77	92.3
Flowering Time	217.25	289.20	75.1
Maturity Score	1.30	1.49	87.2

Table 2: 10-fold cross-validated test error for SuRF and gradient-boosted trees for each of the five traits as well as SuRF CV error as a percent of the boosted tree CV error.

The feature selection results for both SuRF and the gradient-boosted trees are summarized in Table 3. For each trait the table lists the gene markers selected by SuRF with a 0.2 p-value cut-off. Next to each of these tags is the rank corresponding to its relative importance as estimated by the gradient-boosted trees. Any tags that were selected by SuRF but were not in the top 473 predictors ranked by the gradient-boosted trees have a '-' in that column.

Chipping Score		Emergence		Glucose		Flowering Time		Maturity	
SuRF	Imp.	SuRF	Imp.	SuRF	Imp.	SuRF	Imp.	SuRF	Imp.
TAG4566	1	TAG3136	13	TAG2112	8	TAG4173	14	TAG4774	4
TAG55	3	TAG1230	1			TAG2818	-	TAG4395	7
TAG636	137	TAG4634	-			TAG4029	8	TAG666	-
						TAG4725	-	TAG5212	-
						TAG2101	-	TAG2686	-
								TAG806	-
								TAG2914	84
								TAG180	2

Table 3: The tags selected by SuRF and their corresponding importance rank provided by the gradient-boosted trees.

Dimension Reduction

SIS was used to select the one hundred genes best correlated with glucose. The model accuracy statistics corresponding to latent variables of ranks 1 through 15 as estimated by PLN-PCA are shown in Figure 2 below.

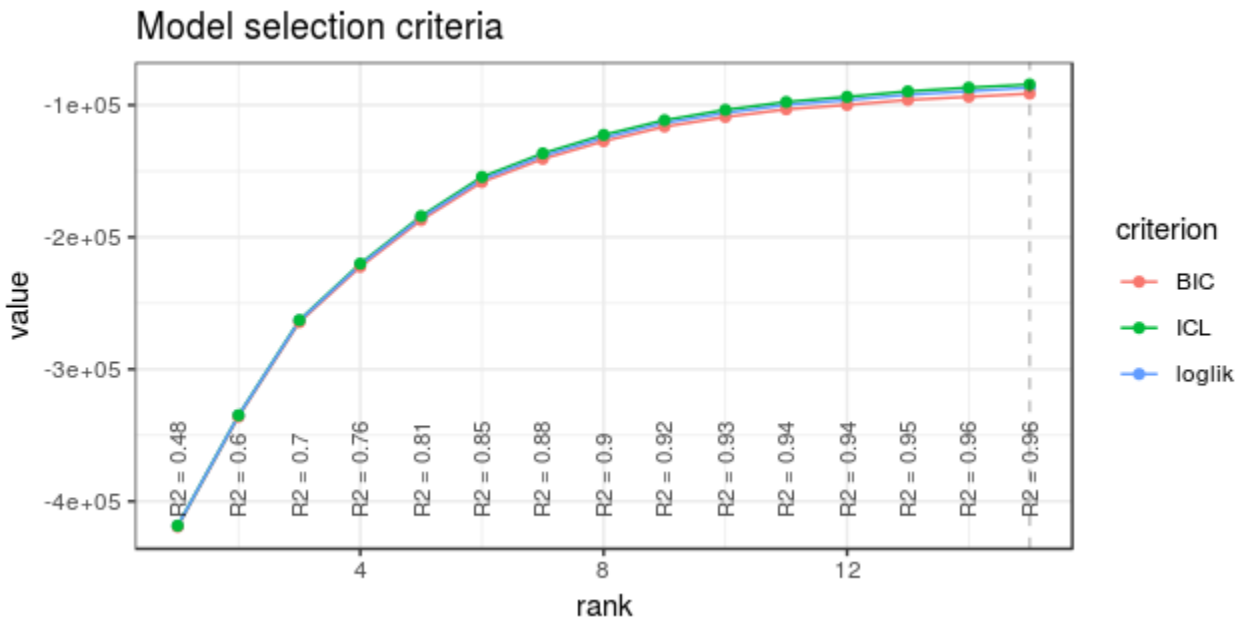

Figure 2: R^2 and model selection criteria for the first 15 reduced feature spaces using $nsis = 100$.

The model which had a latent variable of rank 8 for the principal component analysis was selected as it is the lowest rank which accounts for at least 90% of the variance in the response as quantified by the R^2 measure. Using the scores estimated from this model we fit a PCR which yielded the following coeffiencts and corresponding p-values. Coefficients which have a p-value sufficiently low to suggest significant departure from zero are in bold (excluding the intercept).

Term	Estimate	p-value
Intercept	6.903	$< 2 \times 10^{-16}$
1st Component	**0.079**	**0.001**
2nd Component	0.034	0.245
3rd Component	0.085	0.158
4th Component	-0.036	0.613
5th Component	**0.167**	**0.029**
6th Component	0.093	0.371
7th Component	0.002	0.987
8th Component	**-0.588**	**0.001**

Table 4: Coefficient estimates and corresponding p-values for PCR fit using scores from 8-dimensional feature space.

The p-value for the regression itself was 0.0002 and the multiple R^2 and adjusted R^2 were 0.1049 and 0.0788 respectively. In Figure 3 below we have four plots showing the relationships between the three significant terms and glucose as well as how glucose changes with respect to both the 5th and 8th scores.

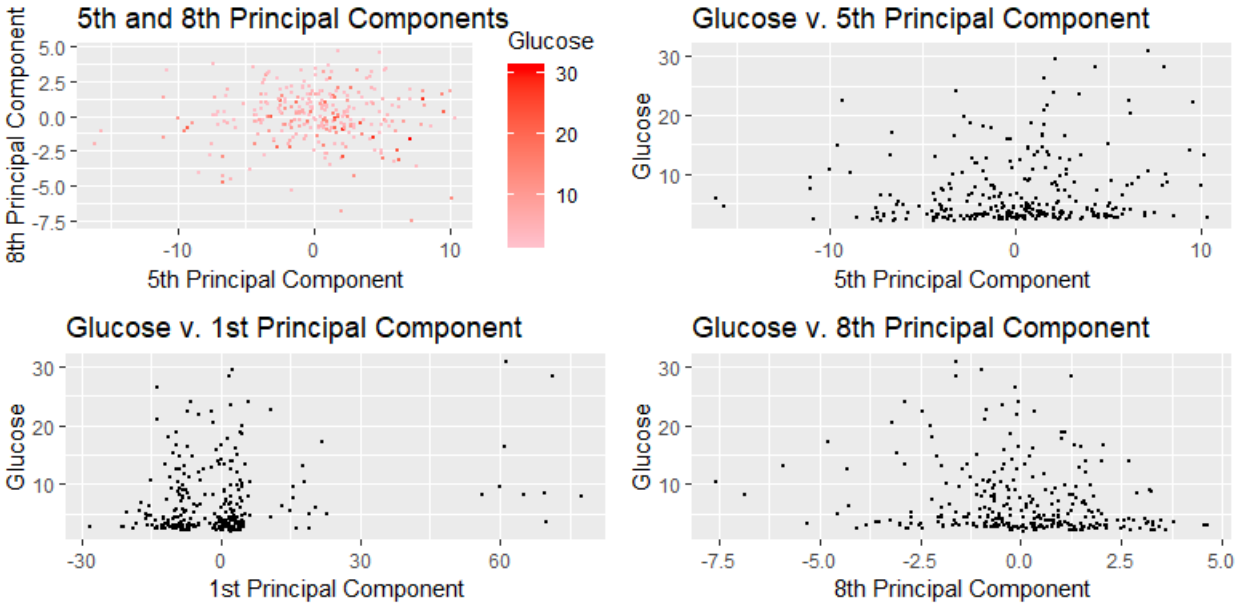

Figure 3: The top left plot shows the relationship between the 5th and 8th scores and how glucose changes as the two change. The remaining three plots show the relationship between the three significant terms (1st, 5th, and 8th scores) and glucose.

Conclusions

Feature Selection

Considering the hyperparameters tuned for the boosted trees it is interesting to note that four of the five traits considered had an optimal interaction depth of one. This means that the trees did not need to be split at all; that the predictor space was sufficiently homogenous in the response that splitting at any predictor reduced the predictive ability of the model. This is the decision tree equivalent of fitting a regression line using only an intercept term and suggests that the predictors do not have a strong relationship with the response, linearly or non-linearly.

Comparing the cross-validated prediction errors between SuRF and the boosted trees for each of the traits shows that for all five SuRF had greater predictive accuracy. For most of

the traits the cross-validated error did not differ greatly between the two methods; glucose and maturity score were within 15% and chipping score and flowering time were within 25%. When emergence was considered however, SuRF produced a test error roughly half of that produced by the gradient-boosted trees. As trees do not perform well when the underlying relationship is stricly linear this suggests that the relationship between the gene expressions and emergence time can be mostly explained by linear terms.

Even at a p-value cut-off of 0.2 SuRF did not select very many predictors for most of the response, the notable exception being maturity. This suggests that only a small number of genes are more useful in predicting the responses than others. When the response 'chipping score' is considered the top two gene tags selected by SuRF are also ranked very high by the gradient-boosted trees; this agreement between the two methods suggests that tags 4566 and 55 are closely related to the chipping of a potato chip.

For the emergence time response the tag selected second by SuRF was ranked first by the gradient-boosted trees while the first tag selected by SuRF was not ranked very high and the third wasn't ranked at all. As the performance of the boosted trees was much worse than that of SuRF, as noted above, the importance rankings should only be considered lightly. Overall it is reasonable to conclude that tags 3136 and 1230 are reasonable variable selections for this trait.

The glucose measured in the potato plants only had one predictor selected by SuRF even with a cut-off of 0.2. This suggests that tag 2112 is the best-suited gene tag for predicting glucose level out of the other tags measured by a substantial margin; it was also ranked fairly well by the gradient-boosted trees.

There were five predictors selected by SuRF for flowering time, however three of them were not in the top 473 predictors as ranked by the gradient-boosted trees. The other two tags were ranked fairly high by the gradient-boosted trees. Recall that SuRF considers only linear relationships and trees consider only non-linear relationships. The discrepancy in the rankings of the two methods suggests that there are some predictors with a strong linear relationship with flowering time and other with a strong non-linear relationship, but there is little overlap between the two.

SuRF selected the most predictors when considering the relationship with maturity score; of the eight selected three were ranked in the top ten by the gradient-boosted trees. The remaining five did not have high importance so this suggests that tags 4774, 4395, and 180 are the most useful in predicting the maturity of a potato plant out of the markers considered. Overall these results provided at least one gene marker that, relative to the rest of the predictors, stood out as a suitable predictor for each response.

Dimension Reduction

The results from the PCR show that the reduced dimension space estimated by PLN-PCA was able to pick up some signal from the glucose. The scatterplots in Figure 3 do confirm that there is a weak relationship between the 5th and 8th principal components and glucose in the potato plants. Additionally, the plot in the topright of Figure 3 confirms this relationship.

There are several points that suggest that there is either not a strong relationship between the genes and the phenotypic expressions or at the very least such a relationship is not well captured using the PCA framework. The first such point is that the number of predictors considered in the component analysis needed to be reduced a priori as using the entire dataset did not produce accurate approximations. This is further supported by the very small R^2 values for the principal component regression as well as the scatterplots in Figure 3. This supports the assertion in the previous section that only a small number of gene markers are important predictors of the responses. If only a very small number of predictors are useful then finding a subspace which is a transformation of all of the predictors is unlikely to perform well, even when considering a latent variable of small rank.

A Additional Plots

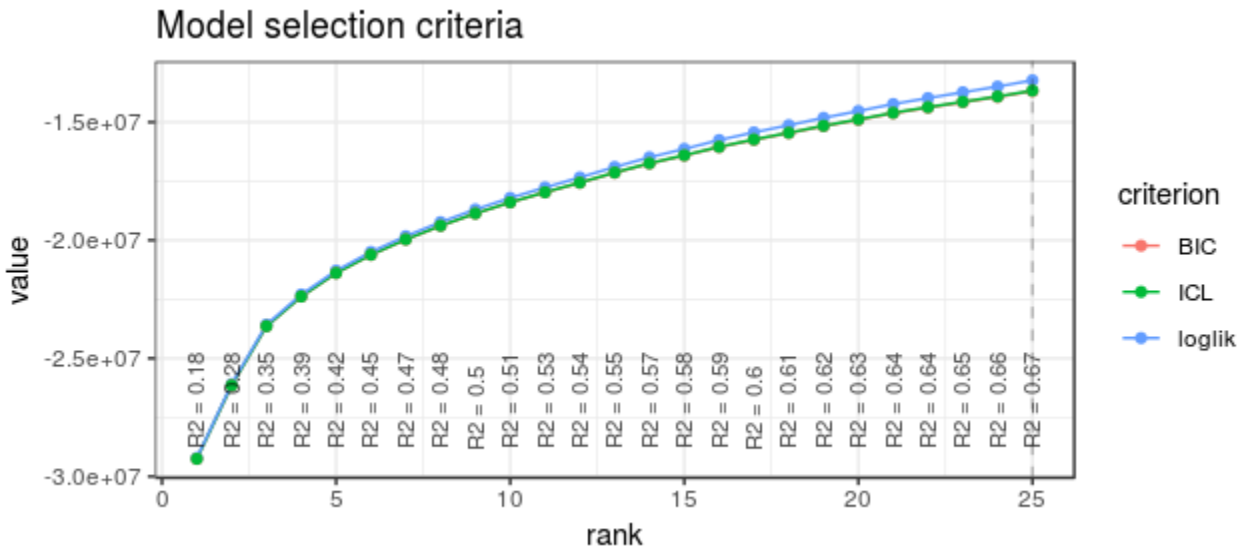

Figure A.1: Model accuracy measures for the first 25 rank approximations produced by PLN-PCA. Point of interest is that even at rank 25 the R^2 value is only 0.67.

References

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2019). *xgboost: Extreme Gradient Boosting*. R package version 0.82.1.

Chiquet, J., Mariadassou, M., and Robin, S. (2018a). *PLNmodels: Poisson Lognormal Models*. https://github.com/jchiquet/PLNmodels (dev version).

Chiquet, J., Mariadassou, M., and Robin, S. (2018b). Variational inference for probabilistic poisson pca. *Annals of Applied Statistics*.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Royal Statistical Society*, 70.

Hastie, T., Tibshirani, R., and Freidman, J. (2009). *Elements of Statistical Learning*. Springer, second edition.

Hu, M. and Polyak, K. (2006). Serial analysis of gene expression. *Nature Protocols*, 1:1743–1760.

Liu, L., Gu, H., Limbergen, J. V., and Kenney, T. (2018). Surf: a new method for sparse variable selection, with application in microbiome data analysis. *Biometrics*, 63:1–25.

Liu, L. and Kenney, T. (2018). *SuRF: Subsampling Ranking Forward-selection*. R package version 1.0.0.

Saldana, D. F. and Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, 83(2):1–25.