What is the dimensionality of salt tolerance in barely?

Dennis Perrett
Matrikelnummer 5578599
dennis.perrett@student.uni-tuebingen.de

Philipp Hölscher
Matrikelnummer 6301374
philipp.hoelscher@student.uni-tuebingen.de

Abstract

To understand the latent structure of salt tolerance in Barely, we used Saade et al.'s (2016) data set of 1336 Barely genotypes grown with and without salt treatment under field conditions. Using regression and latent variable modelling methods we find that salt tolerance is not a purely uni-dimensional construct and not able to be fully measured by traditional yield-based indices.

1 Introduction

Climate change and exponential population growth (Roser et al. [2013]) coupled with linear increases (ourworldindata.com [2023]) in crop yield is increasing pressure on our agricultural system. Rising sea levels and droughts result in higher soil salinity hindering crop growth, threatening food production. Understanding how plants respond to salt stress allows for targeted plant development for more resistant crops. In this work we highlight difficulties faced in traditional analysis due to the nature of the data and complexity of the plant system.

2 Methods

2.1 Data and Methods Introduction

To better understand salt tolerance in Barley, Saade et al. (2016) curated and published a novel data set of 5360 individual plants, spanning 1336 genotypes across two growth cycles. We use this dataset and, inspired by Agarwal et al. (2019), attempt to model salt tolerance directly by means of OLS and quantile regression. We then extend upon these models and apply a structural equation model (SEM) to highlight and investigate the complexity of the plant system. We transform the dataset with the function $f(x) = x/\sqrt{x}$ for consistency and comparability with Saade et al.'s (2016) proposed salt *Stress Weighted Performance* (SWP) measure; a yield based measure that considers magnitude as well as difference. We use Statsmodels (Seabold and Perktold [2010]) (Python), Semopy (Igolkina and Meshcheryakov [2020]) (Python) and Lavaan (Rosseel [2012]) (R) for our analysis.

2.2 Linear Regression

Simple models provide a strong scientific benefit. A simple model is easy to interpret and is a powerful tool for scientific conclusions. We start by fitting a simple OLS regression model in the form of:

$$SWP = \beta_0 + \beta_1 HEA + \beta_2 MAT + \beta_3 RIP + \beta_4 HEI + \beta_5 TGW + \beta_6 EAR + \beta_7 GPE + \beta_8 DRY + \beta_9 HI$$

Project Report for *Data Literacy* 2022/23 where SWP is the stress-weighted performance index proposed by Saade et al. (2016).

Although well understood, easy to implement and interpret, and inexpensive to compute, OLS demands that certain assumptions hold in order to retain its properties as a Best Unbiased Linear Estimator (BLUE). These assumptions include i.i.d, normally distributed residuals with equal variance. Below we investigate these assumptions, highlight invalid assumptions and provide some potential solutions.

Due to the nature of the dataset we first address the multicolinearity of our data. This impacts interpretation of the model's coefficients, which is the main purpose of the analysis in this form. A typical way of handling this is dimensionality reduction through PCA or removing independent variables from the model, however, this does not solve the problem of interpretability with respect to the plant features.

In accordance with Agarwal et al. (2019), we drop Thousand grain mass (TGW) and dry weight per m^2 (DRY) from our model. This removes our ability to interpret the effect of these parameters, however we gain some interpretability for remaining estimated parameters. Nevertheless, this method still fails to address the cause of the multicolinearity and begs the question whether a coefficient relates directly to the corresponding plant feature, or to a latent variable underlying the plant features.

Secondly we investigate the residuals for violations of the OLS assumptions.

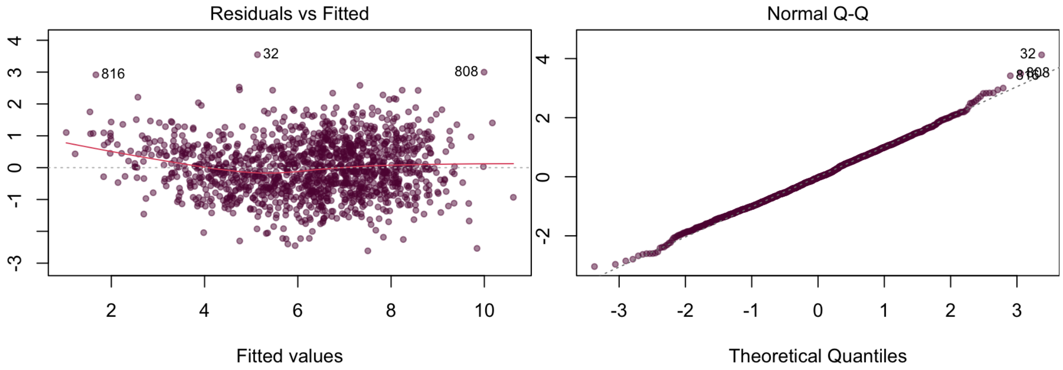


Figure 1: The Residual vs Fitted plot displays clear Heteroscedasticity. The Normal Q-Q plot shows normality of the residuals.

While the residuals are normally distributed, it is clear from the plots that we are faced with heteroscedasticity. The result is that the F and t-statistics are not appropriately distributed under the null(s), which leads to incorrect standard errors and p-values. These issues make the application of OLS regression difficult, as we are forced to make changes that would severely hinder the usefulness of the model.

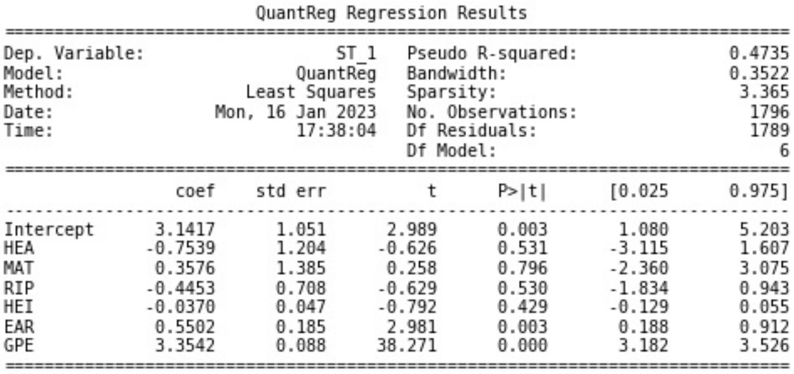
3 Quantile Regression

Quantile regression is an extension of OLS in which quantile regression models a chosen quantile, rather than the conditional mean. Moreover, quantile regression does not assume homoscedasticity Koenker and Bassett [1978].

While circumventing the technical problem of heteroscedasticity, the output below (Figure 2) highlights a distinct shortcoming of quantile regression: The model does not describe the data as a whole, but rather only a single quantile.

We can see from the Pseudo R-squared that the model is a less than impressive fit for the 75th quantile (0.47). Most estimates are insignificant which leads to hard to justify interpretations.

While technically a relatively sound approach, after resolving these issues we still face the problem mentioned above, namely, do the coefficients correspond to plant features directly, or to an underlying latent variable?



```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                   ST_1   Pseudo R-squared:               0.4735
Model:                       QuantReg   Bandwidth:                      0.3522
Method:                 Least Squares   Sparsity:                        3.365
Date:                Mon, 16 Jan 2023   No. Observations:                 1796
Time:                        17:38:04   Df Residuals:                     1789
                                        Df Model:                            6
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.1417      1.051      2.989      0.003       1.080       5.203
HEA           -0.7539      1.204     -0.626      0.531      -3.115       1.607
MAT            0.3576      1.385      0.258      0.796      -2.360       3.075
RIP           -0.4453      0.708     -0.629      0.530      -1.834       0.943
HEI           -0.0370      0.047     -0.792      0.429      -0.129       0.055
EAR            0.5502      0.185      2.981      0.003       0.188       0.912
GPE            3.3542      0.088     38.271      0.000       3.182       3.526
==============================================================================
```

Figure 2: Output of the 0.75th quantile regression fit. Note the low Pseudo R^2 and number of insignificant parameter estimates.

4 Structural Equation Model

Structural equation modeling (SEM), an extension of confirmatory factor analysis (CFA) and path analysis, is a method used to analyse complex relationships between variables, both observed and unobserved (latent). SEM aims to fit a user-specified model (e.g.: user defines which observed variables load on which latent factors) that can accurately reconstruct the variance-covariance matrix of the data. The process of SEM starts with investigating the dimensionality of the data, in our case, eigenvalue decomposition, and attempting to identify meaningful latent structure. We hypothesize two latent traits: a growth-related *vegetative* and a yield-focused *reproduction* trait (see Figure 3). A *unidimensional*, *higher order*, and *bidimensional* model (among many others) are fit and compared to investigate the inter-relation of the data.

Eigenvalue decomposition suggests that the data is mostly one-dimensional, with 67% explained by dimension 1 and only 13% explained by the second dimension. This is likely due to salt tolerance being generally unidimensional (enough salt will simply kill the plant) and the dataset being curated with highly correlated variables.

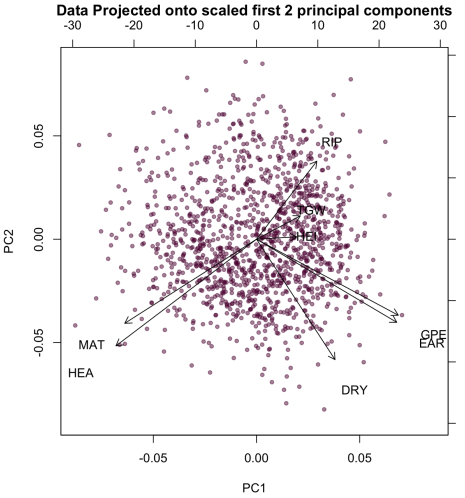


Figure 3: Biplot of projected data. Note the two distinct directions of the factor loadings.

4.1 Model fits

We conduct a likelihood ratio test (lavTestLRT) and analyse various fit measures to determine the model which describes the data the best. The test compares models with a significant result suggesting the more complex model has a better fit.

Table 1: Fit indices of 3 Latent Trait models.

Model	df	χ^2	χ^2/df	CFI	TLI	RMSEA	LavTest P-value
Unidimensional	20	62.51	3.13*	0.76	0.66	0.040	-
Higher order model	18	36.73	2.04*	0.89	0.83	0.03**	0.00**
Bidimensional	17	24.27*	1.43**	0.98**	0.96**	0.02**	0.00**

Based on table 1, we see that the *bidimensional* model has the best fit. The χ^2 test and associated p-value tend to prefer complex models. Accordingly, we consider a range of fit measures (Stone [2021]).

We sidestep analysis of the 22 parameter estimates (due to scope restrictions) to focus on the question of dimensionality. To briefly evaluate the performance and interpret the latent factors we plot the 2 latent traits against the Saade et al.'s (2016) SWP measure for salt tolerance and see that the reproduction factor correlates (positively) strongly, while the vegetative factor correlations weakly negatively. We test both a model with correlated, and uncorrelated latent factors. Analysis shows no significant correlation. As such, the factors are uncorrelated in the plots by design. Interestingly, the, albeit very low, correlation between SWP and the vegetative factor shows that SWP is indeed capturing slightly more than just yield based information.

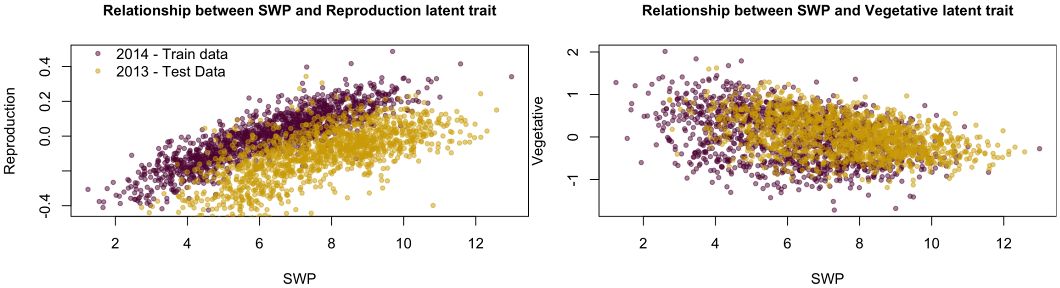


Figure 4: Relationships between the latent traits and the Saade et al.'s (2016) SWP measure for salt tolerance. Note the near identical performance on the unseen test data. SWP considers the magnitude of raw yield, which was generally higher in 2013.

4.2 Discussion

A simple, brief analysis suggests that the addition of salt results in a, for the most part, increase in ripening time (RIP), maturity time (MAT) and flowering time (HEA) (*https://github.com/Philipp-42/salt-tolerance-in-barely-.git*). These observed variables, along with total grain weight (TGW) and plant height (HEI), load on the vegetative factor. The second latent trait, in turn, is strongly correlated with grain per ear (GPE), dry weight (DRY) and ear number per plant (EAR), along with TGW and HEI.

We conclude the multicolinearity present in the dataset can be attributed to two interpretable latent traits, namely vegetative and reproduction factors. A few technical, and theoretical aspects of the model, (not discussed) hint that there are further latent traits not fully captured by the data, again highlighting the issues of a non-comprehensive measure of plant performance / salt tolerance. While SWP nicely captures the reproductive aspect of salt tolerance, it almost fully neglects the vegetative salt tolerance. Research neglecting the full dimensionality of salt tolerance runs the risk of biased, if not misguided conclusions.

References

Stephanie Saade, Andreas Maurer, Mohammed Shahid, Helena Oakey, Sandra M. Schmöckel, Sónia Negrão, Klaus Pillen, and Mark Tester. Yield-related salinity tolerance traits identified in a nested association mapping (nam) population of wild barley. *Scientific Reports*, 6(1):32586, 2016. doi: 10.1038/srep32586. URL `https://doi.org/10.1038/srep32586`.

Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Lucas Rodés-Guirao. World population growth. *Our World in Data*, 2013. https://ourworldindata.org/world-population-growth.

ourworldindata.com. Crop Yield, tonnes per hectare, 2023. URL `https://ourworldindata.org/explorers/crop-yields?tab=chart&facet=none&country=NZL~OWID_WRL&hideControls=false&Crop=Wheat&Metric=Actual+yield`.

Agarwal G, Saade S, Shahid M, Tester M, and Sun Y. Quantile function modeling with application to salinity tolerance analysis of plant data. *BMC Plant Biol.*, 2019.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Anna A. Igolkina and Georgy Meshcheryakov. semopy: A python package for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 0(0):1–12, 2020. doi: 10.1080/10705511.2019.1704289. URL https://doi.org/10.1080/10705511.2019.1704289.

Yves Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012. doi: 10.18637/jss.v048.i02.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913643.

BM Stone. The ethical use of fit indices in structural equation modeling: Recommendations for psychologists. *Front. Psychol.*, 2021.