Notes for edits in Disseration

Chapter 01

Chapter 02

Add additional paragraph to motivation?

Chapter 03

\*\*Shoot-Root Manuscript\*\*

Following Sun et al.’s \cite{sun2014model} developmental model, we calculated and chose four key heterochronic parameters, asymptotic growth (a), relative growth rate (r), the timing of inflection point (TI), and the duration of linear growth (L), as phenotypic values to perform QTL mapping. A great variability was observed for growth curve parameters of both phenotypic traits (Table 1). Compared with taproot length, shoot length has a greater rate of growth and reaches the maximum growth rate at an earlier

\*\*Add a paragraph to motivation section\*\* (to end of motivation section)

Continue on about the rest of the chapter as an overview

Chapter 04

\*\*Start of my writing\*\*

A general form of a regression problem is to approach it as a linear combination of basis functions. Framing the problem as such allows for more flexibility to be placed in the model and intersting relationhips the can be explored by relating to functions that are meaningful to the given situation. The most general form of this is

\begin{equation}

f(x) = \sum\_{i=0}^P \theta\_i \phi\_i(x)

(\#eq:gen-form)

\end{equation}

where the $\phi$ are basis functions of the researchers choosing. Under this format you can chooise any function that would fit the need of the given problem and is relevant to the application. The general form for solving for the regression parameters, in our case the genetic effects of the markers and the epistatic interactions between the markers, is as follows:

\*\*Regression by linear combination of basis functions. by Risi Kondor\*\*

Solving for the $\theta’s$

To find the optimal value for $\theta\_0, \theta\_1, . . . , \theta\_P$ we

1. define a loss function L;

2. using the loss function define the empirical risk $Remp(\theta)$ quantifying the loss over all the training data for particular values of $\theta\_0, \theta\_1, . . . , \theta\_P$ ;

3. solve for the particular setting of the parameters (denoted $\theta\_0, \theta\_1, . . . , \theta\_P$)

that minimizes the empirical risk.

We shall use the squared error loss function

$$L(y, f(x)) =\frac{1}{2}(y − f(x))^2$$

This is the simplest possible loss function, and it just says that the loss is proportional to the square of the difference between the predicted value and the true value. The empirical risk is then

To simplify the development, we now introduce the vectors



and the matrix

$$\mathbf{Q} = \left(\begin{array}

{rrr}

\phi\_0(x\_1) & \phi\_2(x\_1) & ... & \phi\_P(x\_1) \\

\phi\_0(x\_2) & \phi\_2(x\_2) & ... & \phi\_P(x\_2) \\

... & ... & ... & ...\\

\phi\_0(x\_2) & \phi\_1(x\_N) & ... & \phi\_P(x\_N)

\end{array}\right)

$$

On the slides X is used for Q, but in the general case where $\phi\_i$ are not linear

functions that might be misleading. The empirical risk can then be written in

the much shorter form

To find $\theta^\star$, we can just set the derivatives of the empirical risk with respect

to each $\theta\_i$ equal to zero

and solve for theta. In short hand, this is written as the single equation

We can then solve for the optimal theta by

In summary, we can use the same formula (3) no matter whether we do linear,

polynomial or RBF regression, \_\_the only thing that changes is the definition of the matrix Q\_\_

This type of formulation has some really nice advantages and reduces some of the intial complexity of working with functional data. By fitting unique functions and applying these as basis functions we can still use OLS calcuations throughout the selection process and to fit the final model. As we have see with high dimensional model selection scenarios, correlation among the predictors being considered can be an issue. This may be made worse if depending upon the selection of the basis functions chosen. With this consideration in mind, one set of basis functions that would help reduce the correlation between predictors are orthogonal polynomials. By definition orthogonal polynomials have an inner product of 0 and hence would not have any correlation between them. This property is desireable when

fitting highdimensional data. Not only does it help reduce the correlation allowing for the model assumptiopns to be more closely met there are also other desriable properties. The set of orthogonal polynomials used for our purposes are the Legendre Polynomials

assessed

Chapter 04 simulation studies

As statistical issues become more complex they are going to be more analytically intractable and computational methods will need to close that gap to show the effectiveness of new models and procedures.Simulation studies were performed to ascertain the validity of the model. Computational verified by cross validation, bootstrapping. Having a training and a testing set to help provide some insight on false positive rates, selecting the correct fit of the polynomial for the genetic and epsitatic effects.

Advantages

* Computationally efficient algorithm for static phenotypes
* Includes epistatic interactions in a meaningful/biologically relevant manner
* Incorporates functional component
  + Parsimonious able to fit growth parameters during the fitting of

Areas of concern

* Many areas of estimation error
* Overfitting because of flexibility
* Need for lab verification
  + Used as screening tool

complex and many moving parts to the selection procedure. Very flexible but this could have it be prone to over fitting at times if not well controlled.

With the complexity and expense that comes with genetic mapping, espically with a functional trait that needs repeated measures

Used as a screening tool for initial findings and exploratory data analysis to aid and guide future research. Needs to be then be lab validated. Especially with something as intricate as epistatic effects between gene markers.

Further investigations are needed to confirm or modify our findings by QTL mapping in natural populations.

Application: Worked Example

if it doesn't take too much time (maybe by the time defense comes around) cross validate the selection proceudre by leaving out samples and running model again. How many times do you get the same results? LOOCV tables

Also consider with holding predictors from the dataset to see if the selection is changed at all. Start with random ones and then maybe strategically select important predictors to remove to see how it affects the fit