DECODE PHENOME-GENOME INTERACTIONS: A DATA SCIENCE APPROACH

By

Abhijnan Chattopdhyay

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2022

ABSTRACT

DECODE PHENOME-GENOME INTERACTIONS: A DATA SCIENCE APPROACH

By

Abhijnan Chattopdhyay

The responses of plants to their environments are determined by multiple interacting genetic factors that themselves may operate through numerous biological mechanisms. Disentangling these complex genome-by-environment interactions is a significant challenge to understanding the underlying biology and developing more robust crops. This dissertation integrates high throughput phenotyping and genome sequencing and aims to harness these multidimensional interactions to test whether different genetic components affect biological processes through similar or distinct mechanisms. First, we present a comparison of different methods that can be practically used for genome-enabled prediction and selection purposes with the help of synthetic datasets with varying levels of difficulty and variability. Using such tools, we have found multiple traits are modulated by similar genomic regions, termed "co-localization". But, the question remains, how can one test for co-segregation, or co-linkages, of multiple phenotypes to specific genetic polymorphisms? From domain knowledge, we can argue that there exist various physical modes of interactions among photosynthetic processes, which result in distinct patterns of interactions between measured parameters. We propose a Bayesian latent variable (LV) approach that tries to imitate such physical modes of interaction among photosynthetic processes by projecting the multivariate phenotypes into lower-dimensional latent factors. Estimation of the entries of the loading matrix (the connection between multidimensional phenotypes to LVs) is through the Automatic Relevance Determination (ARD) prior, which can automatically remove the irrelevant latent factors and add immediate interpretability. This means for a single genotype, the observed latent factors will likely reflect the effects of environmental or developmental effects on mechanistic interconnections. Also, these low-dimensional structure/ latent factors can be genetically mapped using quantitative trait loci (QTL) mapping and can be validated with the linkages from colocalized traits obtained from

univariate QTL analysis. The added advantage of our approach is we can describe specific classes of relationships among multiple phenotypes governed by specific genetic regions that can be shared or specific to environments which can be further used to distinguish functional and genetic linkages among a range of photosynthetic regulatory processes. We extended our setup to integrate multiple environments and showed that the latent variables, either specific to one treatment or shared by various treatments, can be mapped to distinct genetic loci, revealing specific genetic polymorphisms altering the co-regulatory network among phenotypes in $Genotype \times Phenotype \times Environmental$ space. The final piece of my work is to model the association/correlation between phenotypes as a function of genetic and environmental explanatory variables to pin down distinct mechanisms. We develop an efficient estimation methodology called Correlation Modeling under Pairwise Likelihood Estimation (CMPLE), aided by a novel Minorize-Maximize (MM) algorithm, and provide statistical inference techniques. Simulation studies mimicking biological data show that the method is beneficial for recovering pertinent information, including different regulatory pathways, and is computationally efficient in handling many parameters. Our approach is also illustrated by analyzing a motivating dataset from recombinant inbred cowpea lines. Using CMPLE, we can identify the specific genetic variations affecting distinct biological mechanisms, namely "Photoprotection" and "Photoinhibition," under various environmental conditions.

Dedicated in the memory of Atindra Mohon Sarkar (Dadu), Gita Sarkar (Didun),
and Binapani Chattopadhyay (Thakuma).

ACKNOWLEDGEMENTS

First and foremost, I am incredibly grateful to my supervisors, Dr. Tapabrata Maiti and Dr. David Mark Kramer, for their invaluable advice, continuous support, and patience during my Ph.D. study. Both have seen my ups and downs as a researcher and kept pushing me for the greater good. It has been an enormous honor for me to have you both as my academic advisors at the Michigan State University. You have inspired me with your immense knowledge and experience and helped shape me as a better researcher and person. This dissertation would not have been possible without your dedication, advice, continuous encouragement, invaluable guidance, and persistent help.

Next, I thank Dr. Samiran Sinha, whose comprehensive support influenced my statistical methods and critical thinking. I have significantly benefited from your wealth of knowledge and meticulous editing. Also, thanks to my committee members, Dr. Chih-Li Sung and Dr. Shrijita Bhattacharya, who offered guidance and support.

I am also indebted to several members of our lab. Special thanks to Isaac, Oliver, and Donghee, with whom I spent countless hours running different statistical models and debugging codes. I gratefully recognize the help of Sebastian and Atsuko for lending biological explanations for all the questions I had regarding Photosynthesis. Also, Thanks to the lovely family of STT for your overwhelming love and support.

Thanks to my roommates, Anurag and Dipti, for making East Lansing a home away from home. Thank you, Atri, Rejada, and Tathagata, for countless hours of playing FIFA and letting me win. Thank you, Anushree, Alex, Shreya, and Sneha, for constantly listening to me rant and talk things out when things became too severe.

Lastly, my family deserves endless gratitude: my father for teaching me to appreciate persistence and inculcate passion for research, my mother for teaching me the act of being selfless, and my brother for teaching me that it is not over until it is over. To my family, I give everything, including this.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

PHENOME-BY-GENOME-BY-ENVIRONMENT INTERACTIONS AND THE SCOPE OF DATA SCIENCE

Portions of this chapter appeared in the following publication:

A. Kanazawa, A. Chattopadhyay, S. Kuhlgert, H. Tuitupou, T. Maiti, and D. M. Kramer, "Light potentials of photosynthetic energy storage in the field: what limits the ability to use or dissipate rapidly increased light energy?," Royal Society of Open Science, vol. 8, p. 211102, 2021

1.1 Background

Generation and testing of models (or hypotheses) are essential components of the scientific method. Development of Artificial Intelligence (AI) and Machine Learning(ML) promises algorithms, tools and techniques that can identify previously unseen connections between phenomena. Though these AI methods can reveal new correlations and connections, they do not provide mechanistic insights. Indeed, it is often unclear how the unseen networks of ML operate, or if the algorithms they develop have any relationship to the true mechanisms that govern the phenomena. Indeed, different ML approaches may lead to similar predictions, but with mechanistic algorithms,inference might be unrelated to the true physical processes.

This issue also affects the robustness of AI/ML for making universally-applicable predictions. The fact that multiple, non-mechanistic algorithms can fit limited sets of data gives rise to the phenomenon of "overfitting" in which model outputs can provide excellent fits to subsets of data that are not universally applicable. More physically realistic models may overcome these issues by constraining algorithms to those that are physically realistic (based on universally-applicable models). Lack of tethering of AI/ML to physical models motivates our proposed work to "bridge the gap" between scientifically feasible phenomenons and AI driven models using experimental data. The aims include the development of tools that allow AI algorithms to be compared to or constrained by hypothetical (physically-relevant) models, thereby enabling "classical" scientific

hypotheses testing as well as the generation of more universally applicable models, reducing the occurrence of overfitting. We chose as a use case the understanding how solar energy transduction enables and limits the energy productivity of crops, is critical for improving the productivity and resilience of crops in a rapidly changing world. Recent development of large scale genotyping and phenotyping technologies demonstrates the opportunity to harness natural and induced variation in photosynthetic processes across various environmental conditions. The major scientific challenge is to understand the complex interactions among the genomics, environment and performance (phenotypes) of plant photosynthesis–a hyper-dimensional problem that is difficult for unaided human understanding.

Our ultimate aim is to enable global analyses of the flood of data from these technologies to generate and test models relevant to meaningful biological functions. These models will represent hypotheses that can be directly tested using more reductionist approaches in the lab. These tools and models will then be used to identify genetic components that can account for the observed diversity of genotype and phenotype variation, and can be used as targets for advanced breeding and engineering efforts .

New high-throughput phenotyping platforms that can rapidly measure multiple phenotypes, allow us to compare the genomic associations of multiple traits. Such platforms generate data "hyper-cubes" that can relate a wide range of parameters (reflecting potentially linked traits), metadata (e.g. environmental conditions) and genomic content. Here, we explore the possibility of using such "co-association" (or co-segregation) maps of hyper-cubic data sets to test models that predict functional and genetic linkages among a range of photosynthetic regulatory processes. We propose to develop a new class of generative models based on dimensioanlity reduction methods for high dimensional phenomics network. These networks provide biologically significant clusters of interrelated photosynthesis traits which can be regarded as fundamental mechanisms across different genotypes and different environmental stress. Representation of such dynamics across "Genotype $\times$ Phenotype $\times$ Environmental" space is crucial to understand the mechanistic bases of the "true" phenomenon and motivates the researchers to use such structural models with different

crops in different climates.

1.2 Photosynthetic model and Electron transfer chain

Consider the case of light capture by photosynthesis [1]. In chloroplasts, photosynthesis can be initiated when light energy is absorbed by pigments (chlorophylls and specific carotenoids). Using high throughput plant phenotyping tools, it is possible to rapidly measure a range of parameters that reflect distinct, mechanistically-related processes related to photosynthetic efficiency on different genotypes under various environmental conditions (light intensity, temperature, humidity, CO_2 levels, time and location). Interpretation of these parameters is based on the literature [2, 3, 4]. Also, through affordable sequencing processes, the gene expressions of any population can be easily obtained, and one can identify SNP markers that can be significantly associated with any phenotype of interest.

Under environmental stresses, e.g., high light intensities, high or low temperatures, lack of water, light input can exceed the capacity to perform photochemistry. This leads to the buildup of photochemical intermediates that can initiate the formation of reactive oxygen species and subsequent photodamage to the photosynthetic machinery, while decreasing the efficiency of photochemistry [5, 6]. Chloroplasts can protect themselves from photodamage by activating various "nonphotochemical quenching" (NPQ) processes that dissipate absorbed light energy, decreasing the accumulation of reactive intermediates. While NPQ can alleviate photodamage, it also decreases photochemical efficiency, and thus the regulation of NPQ is finely adjusted by the chloroplast to balance these tradeoffs.

There are multiple forms of NPQ (rapidly formed "energy-dependent", qE and slowly activated photo-inhibitory quenching, qI), which are activated under different environmental conditions and modified by genetic variations [7, 8]. These altered NPQ responses ("total" NPQ, designated $NPQt$ [9]) can contribute to the canonical qE mechanism where the prediction is that the extent of qE will be positively associated with increased lumen acidification, which will be reflected in a positive correlation between $NPQt$ and the thylakoid pmf, in our case estimated by the $ECSt$

parameter [10]. This predicted association should be modified or broken down under certain conditions or in mutants that lack key components required for activation of the qE response. In some cases, the breakdown in normal photoprotective mechanisms can lead to the buildup of a large fraction of photodamaged PSII centers, as reflected in increased qI (slowly reversible NPQ associated with photodamage). The associated loss of photochemical activity can lead to decreased electron and proton transfer, resulting in decreased pmf, which will be reflected in a negative correlation between $NPQt$ and $ECSt$.

As an empirical evidence, Figure 1.1 shows a positive correlation between $NPQt$ and $ECSt$ in a cowpea RIL population under the control temperature and for a genetic combination of QTL markers in chromosomes 4 and 9 (genotypes with AA allele for both markers) [11] . On the other hand, $NPQt$ and $ECSt$ are negatively correlated under the low temperature (Chilling stress) and for a different genetic combinations of QTL markers in chromosomes 4 and 9 (genotypes with AA allele the first marker and BB allele for the second).

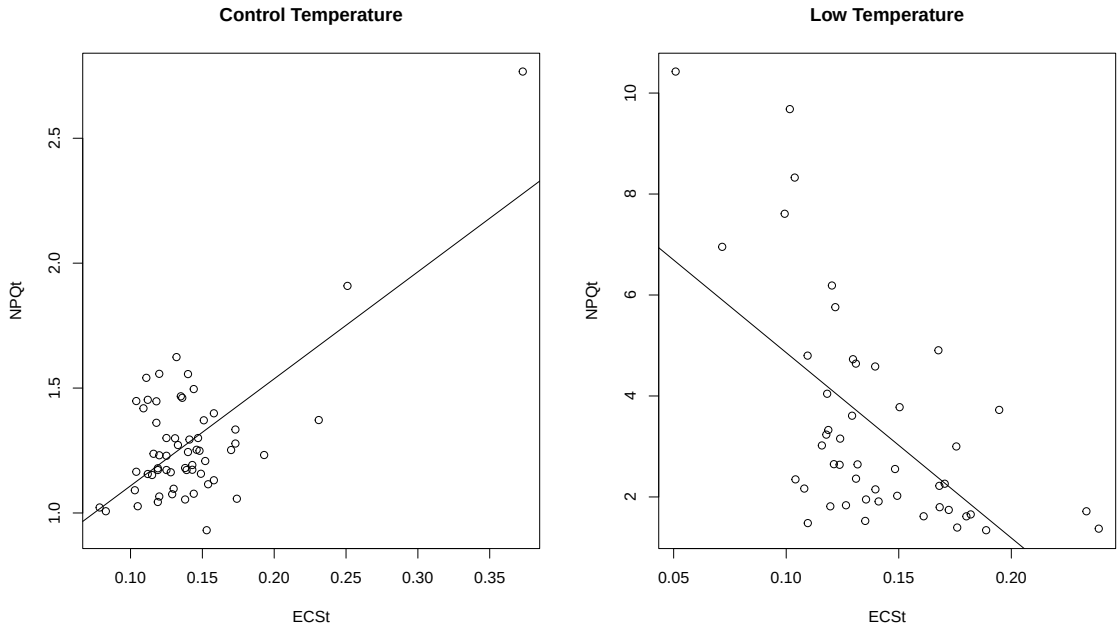

Figure 1.1: Real data example where we identify example of photoprotection and photodamage been regulated by different genetic variations at different environmental conditions

Under ideal conditions, a large fraction of solar energy is used to drive photochemical reactions.

This fraction is usually termed the quantum yield of photochemistry. Productive photochemistry induces a series of electron and proton transfer reactions, resulting in the formation of biochemical energy-storing products, ATP and NADPH, which in turn are used to drive the fixation of CO_2 and other cellular processes. These electron transfers involve two chlorophyll-containing complexes, Photosystem I (PS I) and Photosystem II (PS II), which are essentially connected by the cytochrome b6f complex and mobile electron carriers plastoquinone/plastoquinol (PQ/PQH2) and plastocyanin (PC). In "non-cyclic photophosphorylation", PS II oxidizes the water and releases protons into the lumen, which travels down an electron transport chain to PSI while forming an electrochemical proton gradient (pmf, proton motive force) and passes to NADP+ to make NADPH (Figure 1.2).

Under different abiotic stresses, plants regulate their photosynthetic machinary by triggering various nonphotochemical quenching processes (NPQ). Process (A) (Energy-dependent NPQ (qE)) activated by acidification of the thylakoid lumen resulting in quenching excitation energy through the qE mechanism. This is reflected by the positive correlation between $NPQt$ and $ECSt$. On the other hand, formation of reactive oxygen species can damage PS II, resulting in Process (B) (long-lived photoinhibition-related NPQ (qI)) and decrease the number of active PS II centers which can be observed by the negative correlation between $NPQt$ and $ECSt$. These two processes are illustrated in the Figure 1.3. It is noteworthy that these two forms of NPQ are both induced under conditions where light input exceeds capacity and have similar effects on photochemical efficiency.

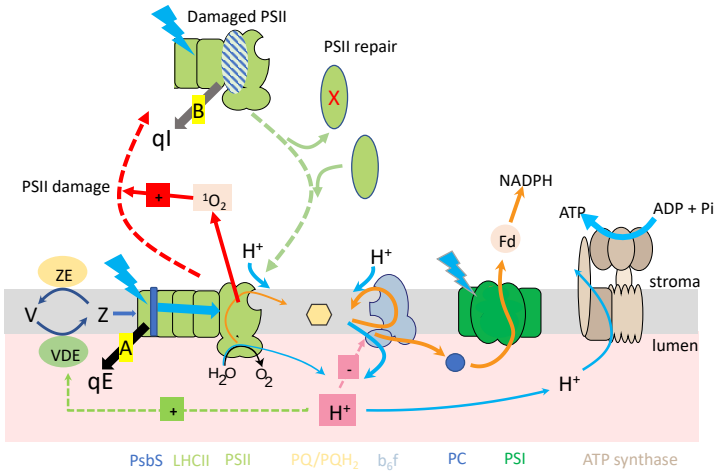

Figure 1.2: Simplified schematics for regulating light energy capture and storage by plant photosynthesis

However, the qE form is typically considered to act as a primary photoprotective mechanism and is readily reversed. In contrast, the qI form involves protein damage, the repair of which requires degradation and resynthesis of the PS II D1 protein, and is thus considered to reflect more severe responses [12]. This patterns are highly influenced by a number of other factors as well. [13] showed that at under lower CO_2 and increasing light, there is a rapid drop in the yield of PS II (ϕ_{II}) and a corresponding rapid rise in the yield of NPQ, together with a decrease in qL. But, under high CO_2 there is a slower drop in the yield of PS II and qL with increasing light, and slower rise in the yield of NPQ. This shows that multiple parts of the photosynthetic machinery indulges in co-regulating the quenching behaviours.

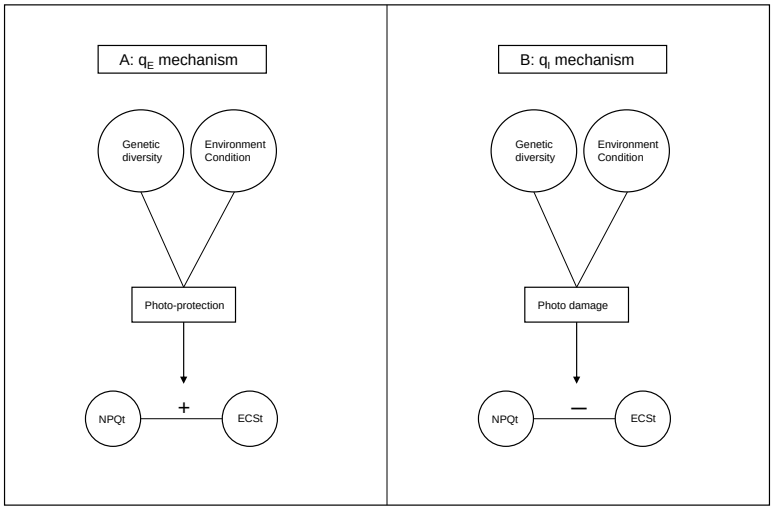

Figure 1.3: Flowchart of photo-protection and photo-damage through purely correlative scheme

This apparent connections between $NPQt$ and $ECSt$ is further impacted by other photosynthetic responses, e.g., ϕ_{II}, qL, etc. as illustrated below. Thus, we have a complex interactions among numerous phenotypic responses which can impact the "beneficial" photoprotective and "harmful" photoinhibitary (photodamage) mechanisms. In real world, the fine balance of such phenotype associations breaks down with certain genetic and environment predictors. Here we aim to associate genetic markers with the corresponding combination of environmental conditions modulating the contributions from these two forms of regulatory mechanism. In fact, by identifying this co-segregation, we can gain better insights about the genetic and environmental determinants of variations in biological system.

1.3 Biological Questions and Big Data Platform

1.3.1 Facilitating science to generate "benchmark" modles

Imagine yourself as a manufacturer of a new model of a car. You have built a world class engineering laboratory to design and produce a car which might be the future of automobile industry. The only problem is you are yet to get a licence to test your models of car outside the lab. So, you would

not be able tell how it might behave on freeways or winding roads or how the tires might fare in snowy conditions. Even if you have carefully designed the new car, you are quite uncertain to assess its performance in real road conditions. A similar problem has been faced by the scientists studying plants, and in particular, photosynthesis, the process by which plants converts light energy into chemical energy generating all our food. This natural process involves net movement of electrons through a series of electron carriers performing a series of chemical reactions known as the light-dependent reactions. In short, Light energy is absorbed by pigment molecules,which passes excited electrons to an electron transport chain activating energetically "downhill" flow of electrons,and thus leading to synthesis of ATP and NADPH. Photosynthesis experts around the globe have been studying photosynthesis in their lab with sophisticated instruments and under specific controlled conditions. From such "reductionist" experiments, researchers are able to dissect complex processes of electron transfers into different component parts of the photosystems. Results from such researches provide a detailed framework of the wonderful biological machine which has powered life for over a billion years.

Such versatile platform enables researchers to demonstrate novel processes under dynamic environmental conditions. Not only that, Science have enlightened us with the genomic artifacts of photosynthetic systems which uses genomic sequencing to identify combinations of genetic loci associated with specific traits and generate elite lines with combinations of those traits through marker assisted breeding. Co-assessing genomic information identifies potentially important genetic loci, helping plants to cope with environmental changes and perils. Coupling sophisticated Phenotyping with Gene-sequencing explore the possibility of test models to predict functional and genetic linkages among a range of Photosynthetic regulatory processes. With the reproducible characteristics at the core of the data generating mechanism, such platform and data generated has the potential to serve as the "benchmark" models for Phenomics applications.

1.3.2 Generating Hypothesis and regulatory Pathways

Biological schematics do not always behave as expected outside the lab. Photosynthesis is highly sensitive to rapid changes in environmental conditions such as light, temperature, humidity, and the availability of water and other nutrients. Understanding different photosynthetic parameters in rapid fluctuations in environmental conditions are critical for plant productivity and the avoidance of photodamage. With this goal of bringing "Nature to the lab", We developed an experimental approach using the open science PhotosynQ platform to probe the "Light Potentials" of photosynthetic processes to rapid increases or decreases in ambient light In this work, we describe an approach to studying the extents and mechanisms or the diversity of such dynamic responses in the field. In a selected set of data on Mentha, we show that the capacity to increase LEF and NPQ upon rapid increases in light are strongly suppressed in leaves previously exposed to low ambient PAR or low leaf temperature.

A simple linear effects model applied over the entire data set indicated strong correlations between LEFamb, PARamb, and Tleaf, suggesting that both environmental factors controlled LEFamb. However, such correlations may be coincidental since PAR and Tleaf are both expected to be dependent on weather or time of day, as it is clear from the solid statistical correlations between PAR and Tleaf. Also, the effects are likely to be co-dependent. For example, at low PARamb, LEFamb should be light-limited and thus have minimal dependence on Tleaf. Still, at higher PARamb, it may be more strongly controlled by temperature-dependent processes.

One approach to disentangling these effects would be to slice the data into segments, e.g., at different ranges of PARamb, and test for correlations with Tleaf within each piece. However, arbitrary-chosen ranges for the details can add bias or fail to detect more complex interactions. We thus applied a Gaussian Mixture Model (GMM) clustering approach based on those presented earlier. Because GMM is an unsupervised machine learning method, it can reduce bias in selecting clusters representing regions of distinct interactions among environmental and photosynthetic parameters. GMM assumes that the data points from the population of interest are drawn from a combination (or mixture) of Gaussian distributions with specific parameters and performs an

optimization scheme to a sum of several Gaussian distributions, allowing for an utterly unsupervised process, avoiding potential user bias. An expectation-maximization (EM) algorithm was used to fit the GMM to the dataset, generating a series of Gaussian components (clusters) with distributions characterized by specific means and covariance matrices. The optimal number of groups was determined using the Bayesian Information Criterion (BIC), the value of the maximized log-likelihood, with a penalty on the number of parameters in the model. This approach also allows the comparison of models with differing parameterizations and differing numbers of clusters because the volumes, shapes, and orientations of the covariances can be constrained to those described by defined models.

Clusters obtained through GMM are within the cluster (intracluster) and between cluster (intercluster) variations. Intracluster variations can be analyzed to determine variations in the interactions between parameters and variations in environmental conditions, e.g., to assess if a relationship is modulated in different ways under different ranges of conditions. Also, as will be seen in the Discussion, intercluster variations (differences in the mean and covariances between clusters) can be used to differentiate distinct patterns of behavior, or mechanistic interactions, between conditions.

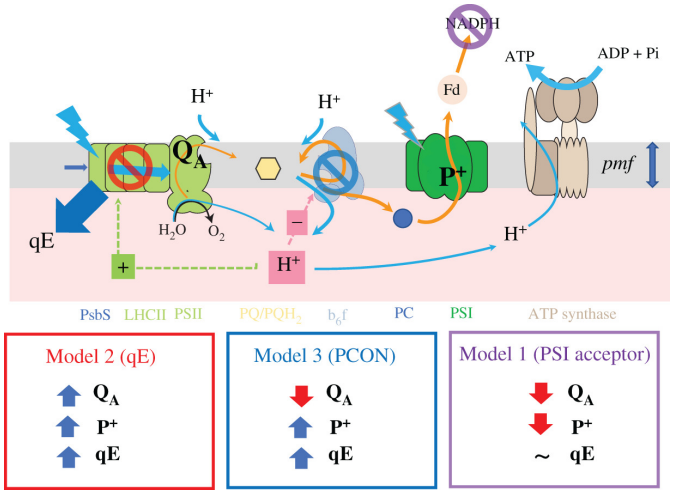

Figure 1.4: Three basic mechanistic models describing proposed processes that can limit the LPs of photosynthetic and photoprotective mechanisms

Using an unsupervised statistical clustering approach, we showed that these effects could be independent of each other under some environmental conditions while likely interacting under

others. This enables to compare the responses of multiple photosynthetic processes, and we were able to test for contributions from several mechanistic Models (Figure : 1.4) for limitations to LEF and NPQ potentials: 1) Limitations in photosystem I (PSI) electron acceptors; 2) increased thylakoid proton motive force (pmf) leading to rapid increases in NPQ in the form of qE, and 3) increased pmf leading to robust photosynthetic control of plastoquinol oxidation at the cytochrome b6f complex (PCON).

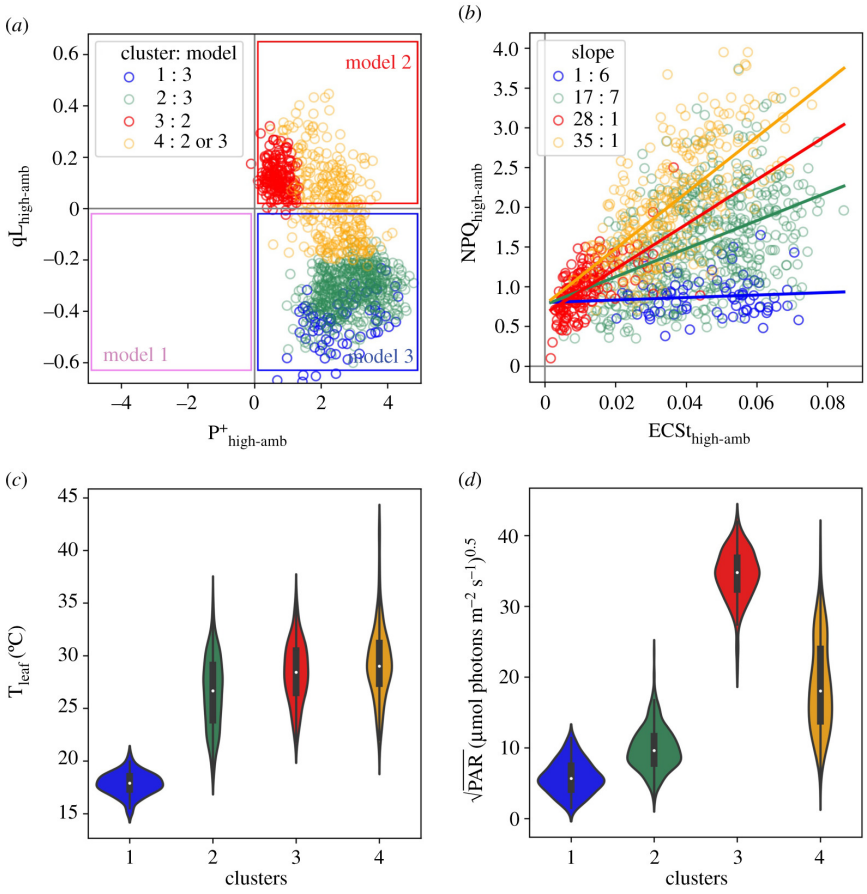

Figure 1.5: Relationships among measured parameters, predicted model behaviours and clustering

Figure: 1.5b plots the dependence of $NPQ_{high-amb}$, which can be attributed to light-induced qE changes, on light-induced pmf changes ($ECSt_{high-amb}$). A generally positive correlation was observed between $NPQ_{high-amb}$and ($ECSt_{high-amb}$), but with high variability, especially at higher values. Applying the clustering obtained for Figure: 1.5a on top of the data in Figure: 1.5b, we see that this variability can be explained by the environmental conditions and the modes of behaviours.

Specifically, we see clear evidence for condition-dependent suppression of rapid activation of qE in response to increases in pmf. Particularly, the sensitivities of $NPQ_{high-amb}$ to $ECSt_{high-amb}$, as indicated by the slopes in figure 8b, were smallest in clusters 1 (slope ~ 1.6) and 2 (slope ~ 17.7), which comprise those with Model 3-like behaviour and occurred at low Tleaf and PARamb values. Higher sensitivities of $NPQ_{high-amb}$ to ($ECSt_{high-amb}$) were seen for clusters 3 (slope ~ 28.1) and 4 (slope ~ 35.1), which comprised those associated with Models 2 and intermediate, and occurred at higher Tleaf and PARamb values.

To assess what controlled the switch between Models 2 and 3, we performed GMM (using $qL_{high-amb}$, $P^{+}_{high-amb}$, Tleaf as inputs). Four distinct clusters were observed (see symbol colours, 1.5a). Intercluster comparisons show that points in clusters 1 and 2 fell exclusively in the region predicted for Model 3. Cluster 3 fell entirely within the region predicted for Model 2. Cluster 4 extended between these regions, possibly indicating contributions from both mechanisms. The clusters falling in the Model 3 region were associated with relatively low Tleaf (1.5c) and PARamb (1.5d), compared with those associated with Model 2 or intermediate behaviours, suggesting that Model 2 prevailed at higher Tleaf and/or PARamb, while Model 3 prevailed at lower values. Within the GMM clusters , $qL_{high-amb}$ was dependent predominantly on Tleaf (cluster 3), PARamb (cluster 1), or both (clusters 2 and 4). This dependence suggests that Tleaf and PARamb acted either independently or cooperatively, depending on conditions, affecting the propensity for photosynthesis to adopt Model 2 or 3 behaviours. As a first-order test of the robustness of these clusters by re-analysing randomly selected subpopulations of the data. Over there, we obtained comparable results, i.e. that we would interpret in similar ways, with as few subpopulations as small as 25% of the full dataset, suggesting that the clustering approach was reasonably robust.

In summary, We found no evidence for Model 1 under any of our conditions, indicating that in Mentha, under our conditions, PSI was maintained in oxidized forms. At higher leaf temperatures, Model 2 prevailed, meaning robust control of induced LEF by NPQ. Strikingly, at lower leaf temperatures, we saw evidence for Model 3, where high light-induced increases in pmf but not in NPQ, resulting in a net reduction of QA and oxidation of P700. Thus, the results reveal considerable

temperature-dependent limitations to NPQ, independent of the formation of pmf, that result in the shape of states likely to produce reactive oxygen species. This low-temperature limitation may thus represent a new target for improving the efficiency and robustness of photosynthesis.

1.4 Scope of Data Science

As the concept of co-association maps gain popularity among the Photosynthesis community, both practitioners and scientists aim to harness the possibility of simultaneously gather data and generate models to support hypothesis invoked from it. In particular, with statistical guarantees researchers were able to draw insights from such data and test hypothesis in the context of biological discovery. With the support of both experiments performed in lab and data from field, we want to accomplish a number of biological goals with the theme of this dissertation. In particular, we aim to answer the following questions:

Biological Query 1: Genome-wide regression and prediction performance

Genetic studies are highly complex in nature as it involves the analysis of high dimensional data, with phenotype(s) being regressed upon a large amount of predictor variables. Even under linear setup, the inherent association among the SNPs and the "heritability" factor make the estimation even complex. One such example would be of a complex phenotype where genetic markers from a specific cluster is attributable to a given phenotype. Since different statistical models addresses the regression problem differently and the nature of "true" association between the response and the predictor variables are unknown, it is near impossible to make robust inference both in terms of prediction and variable selection. Here, we explore various methods like Bayesian linear models, G-BLUP (incorporates genome information) and a shallow neural network approach, called LASSONET to make comparisons for prediction accuracy under varied complexity. Based on synthetic datasets encompassing various situations, we discuss the prediction performance and point out several advantages and disadvantages for different methods.

Biological Query 2: Functional and Genetic linkages for Photosynthetic regulatory Process under different conditions

Plants behave differently with changes in different environmental conditions, for example with respect to excess heat, or excess light or combinations of both(we call this as stress). These behaviours of plants observed under different stress conditions are reflected by phenotypic variations in measured photosynthetic parameters causing natural variation in photosynthetic processes under diverse environmental stresses. The differences in the interconnections between the photosynthetic parameters are evident from the following correlation matrices measured under different stresses:

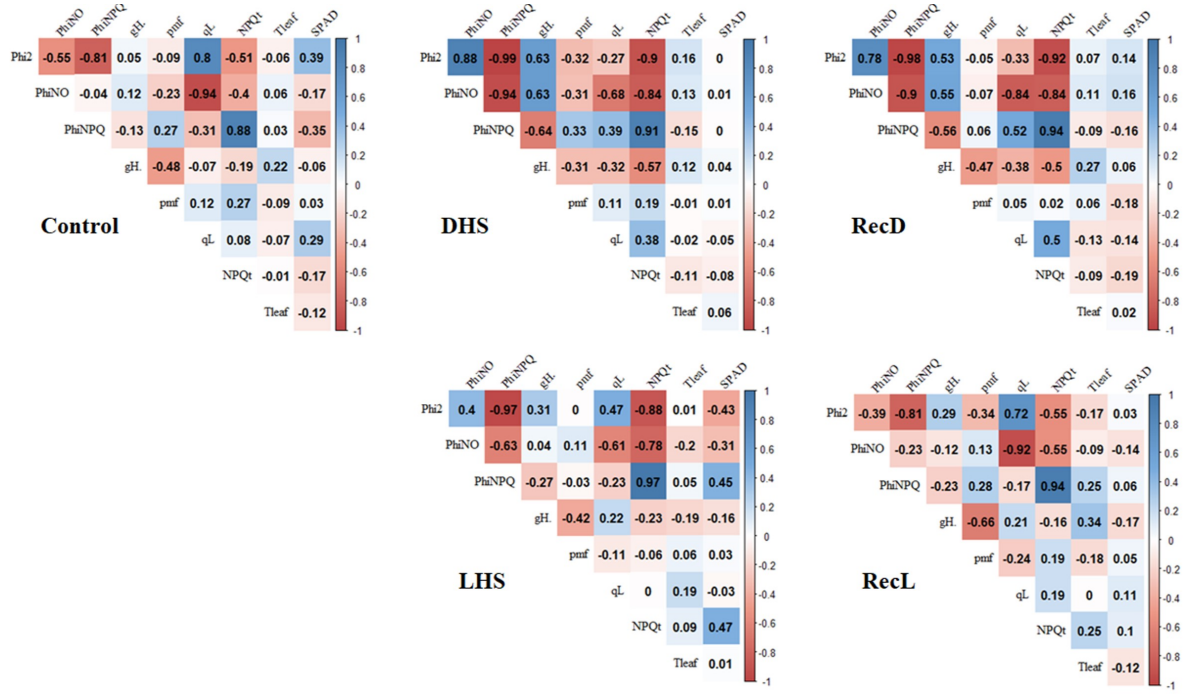

Figure 1.6: Correlations among different phenotypes under different conditions: A) Control/Pre-stress, B) DHS, C) recovery after DHS (RecD), D) LHS, and E) recovery after LHS (RecL).

Studies confirm these interconnections among different phenotypes are responsible for mechanistic bases for adaptations to specific processes of photosynthesis that allow for greater fitness in specific environments. Also, we know about natural variations within a species has come from studies that associate measured phenotypes with specific genomic components, resulting in the familiar quantitative trait loci (QTL) maps. Our objective is to identify specific genomic regions which can potentially different photosynthetic process across different conditions. Biologically speaking,

we explore the possibility of using such “co-association” (or co-segregation) maps of hypercubic data sets (across "Genotype × Phenotype × Environmental" space) to test models that predict functional and genetic linkages among a range of photosynthetic regulatory processes. We explore statistical methods for assessing potential multidimensional linkages and discuss the types of scientific questions that can be asked using the approaches, as well as potential pitfalls.

Biological Query 3: Correlation modeling of multivariate phenotypes in terms of genetic and environmental variables

Quantitative genomics experiments aim to reveal underlying mechanisms that link genotypic variations with multiple biological responses (phenotypes). Interactions/correlations among various phenotypes give new insights into how genetic diversity may have tuned biological processes to enhance fitness under diverse conditions. Dealing with multivariate phenotypes along with the genetic and environmental interactions is a challenging task. One advantage of multivariate GWAS over univariate GWAS is that multivariate GWAS can handle across-trait correlation. But, multivariate GWAS does not explicitly model across-trait correlation in terms of relevant predictors. Also, there exist several statistical methods that deal with variance covariance estimation under generalised linear model. But, in our knowledge, there are no work that can model the interactions between the multiple responses in terms of predictor variables. We therefore provide an innovative framework to dissect genetic configurations behind photosynthetic mechanisms by modeling standard deviations and pairwise correlations among a set of multivariate phenotypes through genetic and environmental predictors. Specifically, our framework, "Correlation Modeling under Pairwise Likelihood Estimation", abbreviated as CMPLE is capable to recover pertinent biological models arising from multi-omics platforms, such as high-throughput phenotyping and genome sequencing. Besides, this procedure has the aspect of many desirable qualities, such as efficient computation, interpretability, and statistical foundation. Our main contributions in this regard are as follows

- Note that, conventional maximum likelihood estimation of parameters fails when one performs standard deviation modeling and pairwise correlation modeling (in terms of predictors).

We have developed a pairwise-composite likelihood approach to estimate the model parameters from the functional forms of standard deviation and correlations among multivariate responses.

- We have proposed a gradient Minorize-Maximize (MM) algorithm to efficiently estimate the model parameters. This guarantees optimization convergence and computational advantages under the given setup.

- Implementing CMPLE on a motivating dataset from a population of cowpea (*Vigna unguiculata. (L.) Walp.)*, we have identified specific genetic variations affecting distinct biological mechanisms, namely "Photoprotection" and "Photoinhibition" under various environmental conditions.

In summary, this research advances statistical approaches for the inference of predictors associated with pairwise correlations in a multivariate setup. Instead of estimating the covariance/precision matrix, the pairwise correlation modeling is relevant in Phenotype $\times$ Genotype $\times$ Environment association studies and helps discover novel bio-physiological pathways in Photosynthesis.

CHAPTER 2

FINDING THE BEST TOOL FOR GENOME-ENABLED-PREDICTION: A COMPARISON STUDY

2.1 Background

Plummeting cost of gene sequencing and genetic marker assays has effectively made it possible to apply them to thousands of individuals in genetic studies. By integrating high-throughput phenotyping data with genomic information, scientists have led statistical genomics into a new era of revolution. The data quality and volume have helped researchers develop tools and techniques that can be efficiently applied for advanced breeding and cultivar improvement. Genome-wide association studies (GWAS) explore an individual's genetic potential and estimate phenotypes using single nucleotide polymorphism (SNP) markers. This process is known as "genome-enabled prediction" (genomic prediction) and can be used to determine breeding program selections (genomic selection). Commonly, genomic prediction is applied early in a breeding program to increase the overall selection procedure and thus helps increase the rate of genetic gain in multiple applications.

Whole-genome-regression (WGR) technique was initially proposed by Meuwissen et al. [14] and has been extensively applied for the genomic analysis of complex traits in plants [15], animals [16] and humans [17, 18]. In WGR, the response phenotypes are regressed over a large number of genetic markers concurrently, invoking a statistical challenge of the "curse of dimensionality" [19] in which the number of predictor variables (e.g., SNPs) is often much larger than the number of observations. This *large-p-with-small-n* regression (where n represents sample size and p denotes the number of predictors) has been extensively studied in statistical and machine learning (ML) literature. Some statistical methods developed for the cause include Bayesian Ridge Regression (BRR), G-BLUP, BayesA, BayesB, BayesC, Bayesian Lasso, etc. Machine learning techniques, such as Reproducing Kernel Hilbert Space, Gradient Boosting Trees, Random Forest, and Artificial Neural Network have also been implemented to cope with the challenge of high dimensional

regressions and have proven to be effective in genetics. However, there are concerns regarding the best tool which can be used for simultaneous feature (marker) selection and complex trait (phenotype) prediction.

Most statistical methods for phenotype prediction in WGR have linear models as their backbone. Nevertheless, in real datasets, the association's nature might also be nonlinear. Machine learning methods such as Shallow neural networks (e.g., single layer NNs) have implemented nonparametric high dimensional regression using nonlinear models and been utilized in multiple applications of the plant, animal, or human genetics. Some studies reflect that NNs can be efficiently trained to obtain high prediction accuracy, but no consistent evidence exists that NNs can outperform linear models. It has been documented that the results obtained from NNs are highly dependent on the genetic architecture, marker density, sample size, span of linkage disequilibrium, and the traits of interest of the species [18]. Hence, empirical evidence suggests that no single approach has uniform superiority across data sets and traits.

Another aspect of WGR deals with choosing optimal SNP markers with high predictive accuracy for the phenotype of interest. Though there have been a very large number of SNPs that are genotyped for a study, most of the methods deal with one SNP at a time for genomic selection. There are several reasons to consider all the SNPs for analysis. The marginal effects of each SNP might have a very different effect from their joint effects. One example of this behavior would be an SNP which is not related to a disease, but correlated with a causal SNP, and will have a marginal association with the disease. Another example might be to think of a situation where several SNPs may have weak marginal effects but strong joint effects. Conditional on causal SNPs which are already included in the model, one would expect that false-positive signals will tend to be weakened while marginally uncorrelated causal SNPs will have a better chance of being selected. Also, the predictive power of a single SNP is assumed to be pretty low. When utilizing a large number of relevant SNPs, one can also improve the prediction power by several folds. While working with a large amount of SNPs, one usually faces the difficulty that the number and extent of spurious associations between the response phenotype and the predictor SNPs increase rapidly

while including many predictors into the model. Also, additional challenges are embodied due to the weak effects of causal variants and strong linkage disequilibrium (LD) among SNPs.

There is a large number of studies based on variable selection methods as discussed earlier, but most of those methods are statistically inaccurate and computationally infeasible for ultra high dimensional p. One of the techniques to resolve this problem is through the sure independence screening (SIS) method [20] by first reducing the dimension to a moderate scale (below sample size) by univariate correlation learning, and then selecting important predictors by a popular variable selection method, such as the LASSO. Similarly, Wu et al. [21] reduced the dimension of predictors to a relatively smaller size using a simple score criterion and subsequently applied the LASSO. One of the shortcomings of this approach is that important features which are marginally uncorrelated with response are more likely to be missed. This is because the univariate screening step is carried out using marginal correlations. One can also modify the SIS implementation by iterative sure independence screening (ISIS) procedure. In this procedure, one can iterate the SIS procedure conditional on the previously selected features which helps in capturing meaningful features that are marginally uncorrelated with the response. In a recent approach, LASSONET tackles this problem within a Neural Network framework, where feature sparsity is attained by incorporating a skip layer [22]. It has proven to perform significantly better with simultaneous feature selection and prediction problems.

In this study, we present a comparison of different methods that can be practically used for genome enabled prediction and selection problems. With the help of synthetic datasets with different level of difficulty and variability, our goal is to find out the best tool that can perform simultaneous variable selection and achieve higher prediction accuracy.

2.2 Methods and Materials

2.2.1 Bayesian Linear Models

Data is collected on a single continuous response, $y_i, i = 1, \ldots, n$, for n individual genotypes. The data equation is given as $y_i = \theta_i + e_i$, where θ_i is the linear predictor that models the expected value

of y_i given the predictors, and e_i are independently and normally distributed random variables with mean zero and variance σ_e^2. The linear predictor, θ_i is further expressed as $\theta_i = \mu + \sum_{j=1}^{p} x_{i,j}\beta_j$, where μ is the overall intercept, $x_{i,j}$ denotes the marker information for the j'th predictor for i'th individual, and β_j denotes the effect of the j'th predictor on the response. Let, η represents the collection of unknown parameters: the intercept, regression coefficients, and the residual variance, expressed as $\eta = \{\mu, \beta_1, \ldots, \beta_p, \sigma_e^2\}$. Since, we are performing our analysis within Bayesian paradigm, we assume prior density in the following way:

$$p(\eta) = p(\mu)p(\sigma_e^2)\prod_{j=1}^{p} p(\beta_j)$$

The joint likelihood is written by,

$$p(\eta|y_1, \ldots, y_n) = \prod_{i=1}^{n} \mathcal{N}(y_i - \mu - \sum_{j=1}^{p} x_{i,j}\beta_j, \sigma_e^2)p(\eta) \tag{2.1}$$

We assign a flat prior to the intercept term, μ and a scaled-inverse χ^2 density to the residual variance, σ_e^2: $p(\sigma_e^2) = \chi^{-2}(\sigma_e^2|\mathcal{S}_e, df_e)$, where the degrees of freedom is $df_e(> 0)$ and the scale parameter is $\mathcal{S}_e(> 0)$. For the regression coefficients, βj, we have assigned either flat or informative priors. The choice of informative priors plays a significant role to attain the different choice of shrinkage. Here, we provide different choices for the priors, for example Gaussian prior for Bayesian ridge regression [23], scaled-t density for BayesA [14], Double Exponential or Laplace prior for Bayesian Lasso [24], mixture of point mass at zero and scaled-t slab for BayesB [14] and mixture of point mass at zero and Gaussian slab for BayesC [25]. We describe the different priors and the choice of hyper-parameters more elaborately in the following paragraph.

Bayesian Ridge Regression

In Bayesian Ridge Regression (BRR), the regression coefficients are assigned IID normal distributions with mean zero and variance σ_β^2. In the second level of hierarchy, we assign a scaled-inverse Chi-squared density, with parameters df_β and $\mathcal{S}_\beta$ for the variance parameter. The joint distribution of the priors along with the hyper-parameters is written as,

$$p(\beta_1, \ldots, \beta_p, \sigma_\beta^2) = \prod_{j=1}^{p} \mathcal{N}(\beta_j|0, \sigma_\beta^2)\chi^{-2}(\sigma_\beta^2|df_\beta, \mathcal{S}_\beta) \tag{2.2}$$

Here, the density is parameterized in a manner, so that the prior expectation and the mode of the variance parameter are $E(\sigma^2_\beta) = \frac{\mathcal{S}_\beta}{df_\beta - 2}$, and $Mode(\sigma^2_\beta) = \frac{\mathcal{S}_\beta}{df_\beta + 2}$, respectively. The values of df_β and $\mathcal{S}_\beta$ are not known. For our analysis, we set the df_β as 5 and solve for the scale parameter for matching the expected R-squared of the model. In genomic studies, this is commonly known as the Best Linear Unbiased predictor (BLUP).

BayesA

In BayesA, the regression coefficients are modeled as a scaled-t density, with parameters df_β and $\mathcal{S}_\beta$. In our setup, this density is constructed as an infinite mixture of scaled-normal densities for computational convenience. Similar to BRR, in the first level of hierarchy, marker regression coefficients are assigned IID normal densities with mean zero and marker specific variance σ^2_β and in the second level of hierarchy, a scaled-inverse Chi-squared density, with parameters df_β and $\mathcal{S}_\beta$ is assigned for the variance parameter. The difference between BRR and BayesA is through the treatment of the scale parameter, $\mathcal{S}_\beta$. Here, the scale parameter is modeled through a gamma density with rate and shape parameters r and s respectively. We have set df_β as 5, s as 1.1 and solve for the rate parameter to match the expected R-squared of the model. In BayesA, the joint distribution of the priors along with the hyper-parameters is written as,

$$p(\beta_1, \ldots, \beta_p, \sigma^2_\beta) = \prod_{j=1}^{p} \mathcal{N}(\beta_j | 0, \sigma^2_\beta) \chi^{-2}(\sigma^2_\beta | df_\beta, \mathcal{S}_\beta) \mathcal{G}(\mathcal{S}_\beta | r, s), \tag{2.3}$$

where $\mathcal{G}(.|.,.)$ denotes a Gamma density.

Bayesian LASSO

The marginal distribution of marker effects in Bayesian LASSO (BL) is double-exponential. Following the work of Park and Casella, we represent the double exponential density as a mixture of scaled normal densities. First level of hierarchy introduces independent normal densities with zero mean and marker specific variance $\tau^2_j \times \sigma^2_e$ on the marker effects. The residual variance, σ^2_e is modeled as a scaled-inverse Chi-square density and the predictor specific scale parameters, τ^2_j are modeled as IID exponentially distributed with rate parameter $\lambda^2/2$. Lastly, λ^2 is assigned a gamma

prior, $\lambda^2 \sim \mathcal{G}(r, s)$. For our setup, we have set s as 1.1 and solved for r to match the expected R-squared of the model. In BL, the joint distribution of the priors along with the hyper-parameters is written as,

$$p(\beta_1, \ldots, \beta_p, \tau_1, \ldots, \tau_p, \lambda^2 | \sigma_e^2) = \prod_{j=1}^{p} \mathcal{N}(\beta_j | 0, \tau_j^2 \times \sigma_e^2) Exp(\tau_j^2 | \frac{\lambda^2}{2}) \mathcal{G}(\lambda^2 | r, s), \tag{2.4}$$

where $Exp(.|.)$ denotes an Exponential density.

BayesB and BayesC

In these cases, the regression coefficients are modeled as IID priors which are expressed as mixtures of point mass at zero and a slab. The slab part is structured with either scaled-t density for BayesB or normal for BayesC. These mixture priors are extensions for BayesA and BRR in the respective cases by incorporating an additional parameter π which represents the prior proportion of non zero predictors. We assign a Beta prior, $\pi \sim Beta(p_o, \pi_0)$ to the mixting parameter. The beta prior is parameterized to achive $E(\pi) = \pi_0$. Also, $p_0 > 0$ is interpreted as the number of prior counts and $\pi_0 \in [0, 1]$. If one chooses $p_0 = 2$ and $\pi_0 = 0.5$, one can obtain a uniform prior in $[0, 1]$, on the other hand a large value of p_0 collapses the prior with point of mass at π_0. The joint distribution of the priors along with the hyper-parameters in BayesB is written as,

$$\begin{aligned} p(\beta_1, \ldots, \beta_p, \sigma_\beta^2, \pi) &= \{\prod_{j=1}^{p} [\pi \mathcal{N}(\beta_j | 0, \sigma_\beta^2) + (1 - \pi) 1(\beta_j = 0)] \chi^{-2}(\sigma_\beta^2 | df_\beta, \mathcal{S}_\beta)\} \\ &\quad \times \mathcal{G}(\mathcal{S}_\beta | r, s) Beta(\pi | p_0, \pi_0), \end{aligned} \tag{2.5}$$

and the joint distribution of the priors along with the hyper-parameters in BayesC is written as,

$$\begin{aligned} p(\beta_1, \ldots, \beta_p, \sigma_\beta^2, \pi) &= \{\prod_{j=1}^{p} [\pi \mathcal{N}(\beta_j | 0, \sigma_\beta^2) + (1 - \pi) 1(\beta_j = 0)] \chi^{-2}(\sigma_\beta^2 | df_\beta, \mathcal{S}_\beta)\} \\ &\quad \times Beta(\pi | p_0, \pi_0), \end{aligned} \tag{2.6}$$

In both the cases, we have set $\pi_0 = 0.5$ and $p_0 = 10$. This signifies a weakly informative beta prior for the mixing parameter with the prior mode as 0.5.

2.2.2 Genomic-BLUP

Instead of the linear models, one can also incorporate different random effects while structuring the conditional expectation function, θ_i. Assuming we have l many random effects $(u_1, \ldots, u_l)$, we can write the conditional expectation for the i'th individual as $\theta_i = \mu + \sum_{j=1}^{p} x_{i,j}\beta_j + \sum_{k=1}^{l} u_{i,k}$. Extending from our previous section, we can write the collection of unknown parameters as $\eta = \{\mu, \beta_1, \ldots, \beta_p, u_i, \ldots, u_l, \sigma_e^2\}$ and express the prior density in the following way:

$$p(\eta) = p(\mu)p(\sigma_e^2)\prod_{j=1}^{p} p(\beta_j)\prod_{k=1}^{l} p(u_k)$$

One common choice is to incorporate Gaussian random effects with some specified covariance structure. In Bayesian settings, people have extensively studied this form in terms of Reproducing Kernel Hilbert Space Regression (RKHS). Gianola et al. [26] have proposed this approach for prediction purpose in genomic studies. The general idea of RKHS is as follows: First, one need to specify the Reproducing Kernel (RK), which is a positive definite functional mapping from the pairs of individuals into the real line. For example, given two genotypes, x_i and $x_{i'}$, we can construct the reproducing kernel as a real valued function, $k(x_i, x_{i'})$ that maps the genotype pair $\{x_i, x_{i'}\}$ into a real line satisfying the condition $\sum_i \sum_{i'} \alpha_i \alpha_{i'} k(x_i, x_{i'}) > 0$, for any non zero coefficients α_i and $\alpha_{i'}$. Next, we represent the regression function as a linear combination of basis functions determined through the reproducing kernel. In Bayesian settings, the RKHS can be expressed as

$$y_i = \mu + u_i + e_i, \quad p(\mu, u, e) \propto \mathcal{N}(0, K\sigma_u^2)\mathcal{N}(0, I\sigma_e^2),$$

where $K = \{k(x_i, x_{i'})\}$ is a $n \times n$ kernel matrix. In Genomic-BLUP (G-BLUP), one incorporates only one random effects which represents the linear regression on the marker densities, $g \sim \mathcal{N}(0, G\sigma_g^2)$, where G stands for the marker information matrix. For practical problems and ease of interpretations, we have standardized the G matrix to have an average diagonal value of approximately one. Janss et al.[27] argued the equivalence between the RKHS regression through Gaussian process and random regressions on principal components. In our implementation, we have used the eigen value decomposition of the genomic matrix, G to make use of this equivalence.

2.2.3 LASSONET

In linear models, LASSO is a very popular tool that assigns zero weights to the most redundant features through l_1 regularization which results in feature sparsity/feature selection. With the backdrop of neural networks, Lemhadri et al. [22] developed LASSONET which can perform global feature selection by adding a residual (skip) layer and allowing a predictor to participate in any hidden layer if the residual layer is active. This method integrates feature selection with the learning of parameters directly, which helps in delivering an entire regularization path with a range of feature sparsity. The objective function being implemented in LASSONET is as follows,

$$\begin{aligned} \text{minimize}_{\theta,W} \quad & L(\theta, W) + \lambda||\theta||_1 \\ \text{subject to} \quad & ||W_j^{(1)}||_\infty \leq M|\theta_j|, j = 1, \ldots, d. \end{aligned} \tag{2.7}$$

The advantages of using this tool is that it uses only a subset of the features and the linear and non linear components are optimized jointly, allowing the flexibility to capture non-linearity. The key idea of this procedure is the constraint

$$|W_{j,k}^{(1)}| \leq M|\theta_j|$$

, which budgets the total amount of non linearity involving the predictor j with respect to to the relative effct importance of X_j as the main effect. Training the LASSONET deals with two operations: at first, a vanilla gradient step is applied on all model parameters followed by a hierarchical proximal operator being applied on the input layer pair $(\theta, W^{(1)})$. Also, this helps in gaining huge computational efficiency. Authors argued that the LASSONET regularization path has an equivalent training cost of training a *single* model. They have also suggested to use a default value of M=10, for the hierarchy coefficient.

2.3 Experiments

In this section, we discuss the performance of different methods in terms of prediction and selection accuracy based on synthetic datasets. In the synthetic dataset, we have simulated data on a single response generated through marker genotypes from a real dataset from CIMMYT global wheat

breeding program. This dataset comprises phenotypic, genotypic, and pedigree information of 599 wheat lines was made publicly available by Crossa et al. [15]. Each line was genotyped for 1279 diversity array technology (DArT) markers. Similar to the RIL population for cowpea, at each marker, there were two homozygous genotypes possible, and they were coded as 0 or 1. For our analysis, we have taken three different cases: (1) significant equispaced markers, (2) significant markers in one cluster, and (3) significant markers in two clusters. The simulation settings (2) and (3) are based on our experience with phenotypes where the correlative pattern of nearby markers indicates a QTL region instead of a single QTL marker. Also, in real data analysis, many of the phenotypes of interest are very complex, and genetic and environmental fluctuations hugely influence them. A measure of an individual's genetic variation accountable for differences in their traits is termed as "Heritability" (represented as h^2). It should be noted that the estimate of the heritability in a particular trait is conditional on a specific population and environment. It is highly dynamic (changes over time as circumstances change).

Estimates of heritability can range from zero to one, where a value close to zero indicates that most of the variability in a given trait is due to environmental factors, with very little influence from genetic variations. On the other hand, a heritability score close to one indicates that genetic differences can be attributed to explaining almost all of the variations in a trait, with little contribution from environmental factors. Many genetic disorders caused by variants (also known as mutations) in a single gene have high heritability. In human genomics, many of the complex traits in an individual, such as intelligence and genetic diseases, have an estimated heritability score in the range of 0.4 to 0.55, suggesting that the variability of such trains is due to a combination of genetic and environmental factors. In our simulation we have incorporated a heritability score of 0.2, and 0.5 for each of the three cases to account for difference in trait complexity. Also, we have varied the number of significant markers as 10, 20, and 30 to replicate real data situations. Below are the results from the three different situations as per our interest. In each of the cases, the number of individuals (n) is 599, the number of available markers (p) is 1279, the number of significant markers (p_0) is 10, 20, 30, and the heritability (h^2) is 0.2, and 0.5. For each combination

of significant markers and heritability score under one setup, the univariate response variable is generated as follows:

$$y_i = \sum_{j=1}^{p} x_{i,j}\beta_j + \epsilon_i \tag{2.8}$$

where $\epsilon_i \sim \mathcal{N}(0, 1 - h^2)$ and the marker effects, β_j are modeled via the following mixture model,

$$\beta_j = \begin{cases} \mathcal{N}(0, h^2/10) & \text{if } j \in \text{Significant marker list} \\ 0 & \text{otherwise.} \end{cases}$$

This simulation design was chosen closely following Perez and de los Campos [28]. To compare the predictive performance, we divide the dataset into training and testing framework following the common convention of 80-20 split. For prediction accuracy, we compared three different measure: (1) correlation of predicted response with the signal, i.e., Cor($\hat{y}$, signal), (2) correlation of predicted response with actual y, i.e., Cor($\hat{y}$, y), and (3) Mean Square error (MSE). For an ideal model, one should achieve higher values for the measures (1) and (2), but should attain lower MSE. For the Bayesian methods, we ran $10,000$ iterations, with $1,000$ samples as the burn-in samples. For LASSONET, we chose the optimal λ which minimizes the MSE in the training setup from the LASSO path and used it for the testing dataset. The value of M was set at 10. Below we present one of the regularization path for the LASSONET solution for the equispaced markers with $h^2 = 0.2$ and $p_0 = 10$

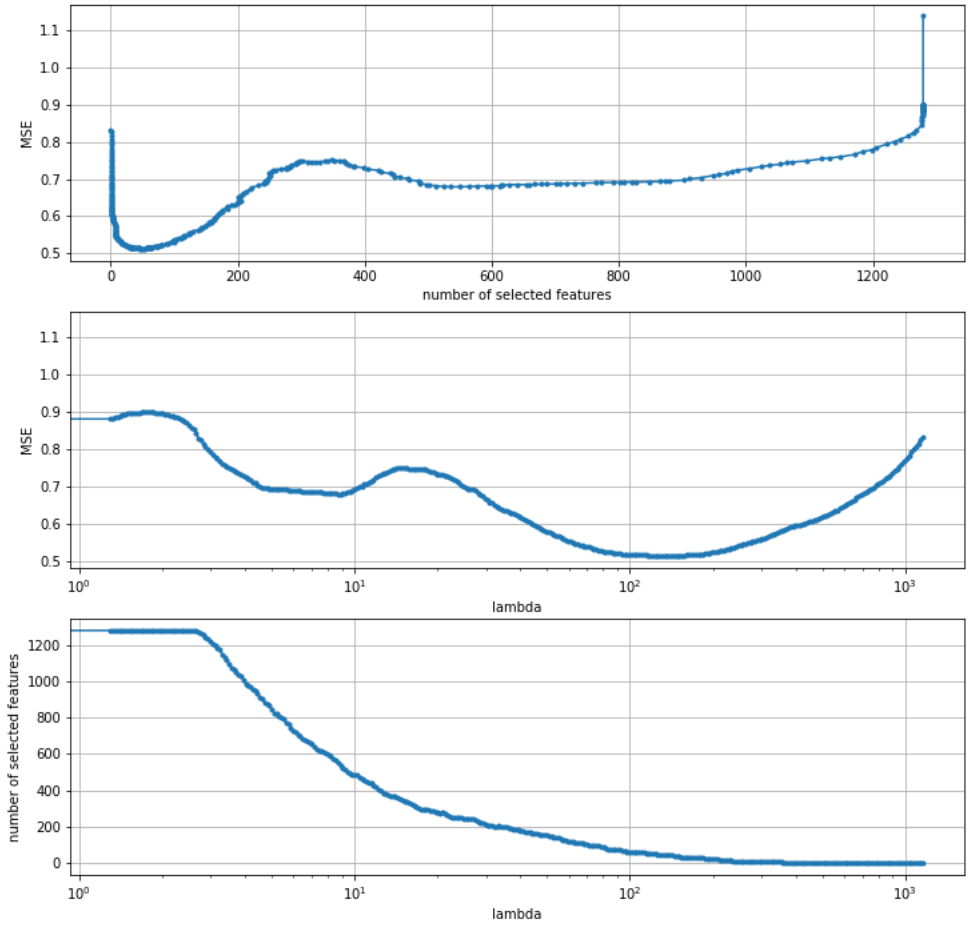

Figure 2.1: Choice of λ and LASSONET path

2.3.1 Simulation setup 1: Significant markers in equispaced locations

In this simplistic setup, we chose the significant markers in the genomic regions as equispaced markers for each choice of $p_0 = 10, 20$ and 30 and the effects of the selected markers are generated following equation 2.8. Table 2.1 highlights the performance accuracy of different methods.

Table 2.1: Prediction performance from simulation setup 1 with heritability score as 0.2 and 0.5. Cor with signal: correlation of predicted response with the signal, Cor with actual y: correlation of predicted response with actual response, MSE: Mean square error

Methods	Cor with signal		Cor with actual y		MSE	
	$h^2 = 0.2$	$h^2 = 0.5$	$h^2 = 0.2$	$h^2 = 0.5$	$h^2 = 0.2$	$h^2 = 0.5$
	0.67	0.51	0.45	0.26	0.59	0.74
BRR	0.68	0.53	0.44	0.23	0.77	1.03
	0.80	0.68	0.61	0.40	0.62	0.88
	0.93	0.86	0.78	0.61	0.31	0.54
BayesA	0.89	0.83	0.78	0.57	0.41	0.84
	0.86	0.70	0.85	0.70	0.32	0.61
	0.95	0.86	0.77	0.61	0.32	0.54
BayesB	0.93	0.65	0.73	0.59	0.44	0.83
	0.89	0.70	0.82	0.69	0.35	0.61
	0.96	0.76	0.76	0.65	0.33	0.53
BayesC	0.93	0.64	0.72	0.61	0.46	0.78
	0.88	0.70	0.82	0.71	0.36	0.60
	0.89	0.71	0.82	0.63	0.28	0.56
BL	0.85	0.62	0.79	0.55	0.41	0.86
	0.85	0.70	0.85	0.69	0.32	0.63
	0.81	0.64	0.80	0.65	0.32	0.54
GBLUP	0.81	0.63	0.76	0.63	0.47	0.77
	0.84	0.70	0.85	0.71	0.32	0.59
	0.94	0.82	0.62	0.36	0.45	0.71
LASSONET	0.83	0.51	0.55	0.20	0.66	1.02
	0.81	0.62	0.60	0.37	0.67	0.92

2.3.2 Simulation setup 2: Significant markers in one cluster

In this setup with Significant markers in one cluster, we chose the significant markers in the genomic regions in the range $(91, 100)$, $(91, 110)$, and $(91, 120)$ for the choice of $p_0 = 10, 20$ and 30 and the effects of the selected markers are generated following equation 2.8. Table 2.2 highlights the performance accuracy of different methods.

Table 2.2: Prediction performance from simulation setup 2 with heritability score as 0.2 and 0.5. Cor with signal: correlation of predicted response with the signal, Cor with actual y: correlation of predicted response with actual response, MSE: Mean square error

Methods	Cor with signal		Cor with actual y		MSE	
	$h^2 = 0.2$	$h^2 = 0.5$	$h^2 = 0.2$	$h^2 = 0.5$	$h^2 = 0.2$	$h^2 = 0.5$
	0.56	0.73	0.25	0.46	0.73	0.56
BRR	0.68	0.80	0.34	0.57	1.01	0.71
	0.59	0.76	0.34	0.57	0.94	0.67
	0.78	0.91	0.62	0.76	0.54	0.31
BayesA	0.74	0.89	0.61	0.80	0.77	0.39
	0.69	0.86	0.72	0.85	0.62	0.32
	0.81	0.95	0.61	0.72	0.54	0.35
BayesB	0.76	0.91	0.63	0.77	0.75	0.43
	0.71	0.88	0.71	0.82	0.62	0.36
	0.77	0.96	0.64	0.70	0.53	0.37
BayesC	0.75	0.92	0.67	0.76	0.72	0.45
	0.67	0.88	0.72	0.82	0.61	0.36
	0.73	0.88	0.63	0.81	0.56	0.28
BL	0.74	0.87	0.66	0.83	0.76	0.36
	0.66	0.85	0.71	0.86	0.64	0.30
	0.70	0.85	0.66	0.80	0.51	0.29
GBLUP	0.72	0.86	0.69	0.82	0.71	0.38
	0.63	0.82	0.72	0.86	0.61	0.32
	0.86	0.94	0.37	0.64	0.68	0.44
LASSONET	0.74	0.84	0.37	0.64	0.96	0.61
	0.61	0.77	0.42	0.26	0.88	0.60

2.3.3 Simulation setup 3: Significant markers in two clusters

In this setup with Significant markers in two clusters, we chose the significant markers in the genomic regions in the range between $(91, \ldots, 95, 701, \ldots, 705)$, $(91, \ldots, 100, 701, \ldots, 710)$, and $(91, \ldots, 105, 701, \ldots, 715)$ for the choice of $p_0 = 10, 20$ and 30 and the effects of the selected markers are generated following equation 2.8. Table 2.3 highlights the performance accuracy of different methods.

Table 2.3: Prediction performance from simulation setup 3 with heritability score as 0.2 and 0.5. Cor with signal: correlation of predicted response with the signal, Cor with actual y: correlation of predicted response with actual response, MSE: Mean square error

Methods	Cor with signal		Cor with actual y		MSE	
	$h^2 = 0.2$	$h^2 = 0.5$	$h^2 = 0.2$	$h^2 = 0.5$	$h^2 = 0.2$	$h^2 = 0.5$
	0.68	0.78	0.34	0.55	0.73	0.58
BRR	0.77	0.86	0.33	0.56	0.97	0.65
	0.40	0.70	0.26	0.53	0.88	0.62
	0.89	0.94	0.62	0.79	0.53	0.31
BayesA	0.78	0.96	0.57	0.77	0.76	0.40
	0.48	0.81	0.70	0.84	0.64	0.31
	0.90	0.97	0.61	0.76	0.53	0.35
BayesB	0.81	0.92	0.59	0.74	0.74	0.43
	0.49	0.84	0.69	0.81	0.64	0.34
	0.85	0.97	0.66	0.74	0.53	0.37
BayesC	0.79	0.92	0.64	0.73	0.71	0.44
	0.49	0.81	0.71	0.84	0.63	0.32
	0.81	0.93	0.65	0.83	0.55	0.27
BL	0.77	0.88	0.63	0.80	0.75	0.36
	0.48	0.79	0.69	0.84	0.66	0.31
	0.78	0.88	0.68	0.82	0.51	0.30
GBLUP	0.74	0.86	0.67	0.80	0.69	0.37
	0.47	0.77	0.72	0.85	0.61	0.31
	0.90	0.95	0.43	0.68	0.68	0.44
LASSONET	0.80	0.91	0.34	0.59	0.92	0.62
	0.49	0.74	0.19	0.52	0.91	0.62

2.4 Discussion

First noticeable difference from comparing the heritability (h^2) from 0.2 to 0.5 is the reduction of our prediction accuracy in all the three simulation setups. That means, the correlation between the signal and actual response with the predicted fell off while the MSE increased. This can be biologically explained as the heritability is measure of how complex the phenotype is, and as the complexity increases, the prediction performs poorly.

Next, in all the three situations BRR performed the worst while there was no conclusive evidence of a best method that outperformed others at least for the cases considered. For simulation setup 1, both the bayesian mixture models, BayesB and BayesC were performing better in terms of prediction accuracy. This result was consistent across the different number of significant markers chosen for the context. For simulation situation two, we found that the mixture based methods, BayesB and BayesC have a better accuracy in terms of the correlation of the predicted response and signal. But, We attained similar performance in terms of correlation with actual response and MSE with Bayesian LASSO and G-BLUP. For heritability score of 0.5, MSE of Bayesian LASSO and G-BLUP were consistently smaller than BayesB and BayesC. Note that, we did not see reasonable prediction from LASSONET so far with the simulated situation 1 and 2. For simulation setup 3, we found similar conclusions as we had obtained for setup 2. But, here we found LASSONET to have improved performance in terms of correlation with signal than the competing methods.

Overall, we found that Bayesian LASSO and Genomic-BLUP were the robust performers to apply for different complexities of the data in terms of prediction purpose. In genetic studies, there are usually two major interests: (a) Prediction, (b) Selection. In this chapter, we have mainly focused on the Prediction aspect, but argue that since we have used Bayesian tools, we can do hypothesis testing based on the credible intervals to find the selected markers. LASSONET is more interpretable in a sense that it does prediction and selection simultaneously. So, we can not disregard it as well. We intend to explore more situations to justify our arguments.

CHAPTER 3

BAYESIAN LATENT FACTOR MODELS TO DIFFERENTIATING GENETIC AND MECHANISTIC BASES OF PHOTOSYNTHESIS

3.1 Background

The term "Natural variations" in photosynthetic processes explains the ability of some phototrophs to transcend others under specific environmental conditions. Inter-dependency between the genetic architecture of an individual (genome) and observable physical or physiological traits or characteristics (phenome) provides an opportunity to harness these natural variations. With the advent of high-throughput phenotyping platforms, it is highly feasible to rapidly measure multiple photosynthetic parameters (phenotypes), which allows us to compare the genotypic variations across numerous traits. Also, rapid advancements in genotyping technologies permitted the production of high-density genetic chips cost-effectively, making the connection from "genome to phenome" possible. Integrating such multi-omics data platforms creates data "hyper-cubes," which involve multidimensional potentially linked traits, environmental variables, and genomic content. Using "co-association" (or co-segregation) maps of such hyper-cubic data sets, we dissect various functional and genetic linkages under photosynthetic machinery. We propose to develop a new class of generative models based on dimensionality reduction methods in the form of a "colocalized" phenotype network. Representation of such dynamics across $Genotype \times Phenotype \times Environmental$ space is crucial to understanding the mechanics of adaptations and facilitating agricultural yield improvement.

One way to associate observed responses (phenotypes) with certain genomic regions is through the familiar Quantitative trait loci (QTL) maps. Using such an association tool over a population of cowpea recombinant inbred lines (RIL), we tested how the tolerance of plants differ when exposed to heat stress imposed under different lighting conditions (light or dark). Furthermore, as with all "omics" approaches, where correlations among multiple traits are informative under

different conditions, it can enable the discovery of potentially causal phenotype interactions, which in turn may shed light on the functions of photosynthetic regulatory pathways. However, we found potential caveats from this analogy as the system's dominant correlations can result from parallel transitive or indirect interactions.

Rapid advancements in multi-omics technologies have led to great deal of interest in the integrated analysis of multi-modal datasets. As the multi-modal datasets provide information on multiple subjects (genotypes) and features from different viewpoints (treatments), the integrated analysis will help to understand biological mechanisms of complex problems and develop tailored treatment for many diseases and health problems. In the past few years, several approaches for integrative analysis have been proposed and applied in diverse field of applications, e.g., in brain imaging, chemical systems biology, single-cell RNA-seq data. One class of models that has been extensively used are based on low-rank matrix factorization such as nonnegative matrix factorization [29, 30], factor analysis [31, 32], canonical correlation analysis methods [33, 34]. Other approaches utilizes clustering framework to obtain interpretation among the multi-omics data , e.g., hierarchical clustering [35], consensus clustering [36], icluster [37]. The basic concept underlying these approaches deals with finding low-dimensional latent factors, which are assumed to carry pertinent information regarding the underlying biological variations across different genotypes and phenotypes and environments. But, these methods are highly unsupervised, which makes the model estimation, inference and interpretation very difficult. Nonetheless, the integrative analysis has proven to be far superior than individual (uni-modal) analysis and there is room for improvement in both methodological and applied research areas.

In this paper, we adopted factor analysis framework to assess differential linkages by generating networks of interactions defined by latent variables (LVs), each of which represents a distinct mode of action of photosynthesis . We then compare the behavior of these networks with the outcomes of hypothetical models operating under different conditions and assess potential associations of genetic components to specific modes of action.

3.2 Materials and methods

Plants (RIL population) being used in this study is taken from University of California, Riverside from the cross between Yacine and 58-77. The Yacine × 58-77 RIL population consisted of 104 lines used to generate the population-specific linkage map, but only 90 RILs were used in the QTL mapping due to limitations in seed stocks for some lines. A total of five different treatments were used for the data analysis, namely: Control (Con), Dark heat stress (DHS), Recovery after Dark heat stress (RecD), Light heat stress (LHS), Recovery after Light heat stress (RecL).

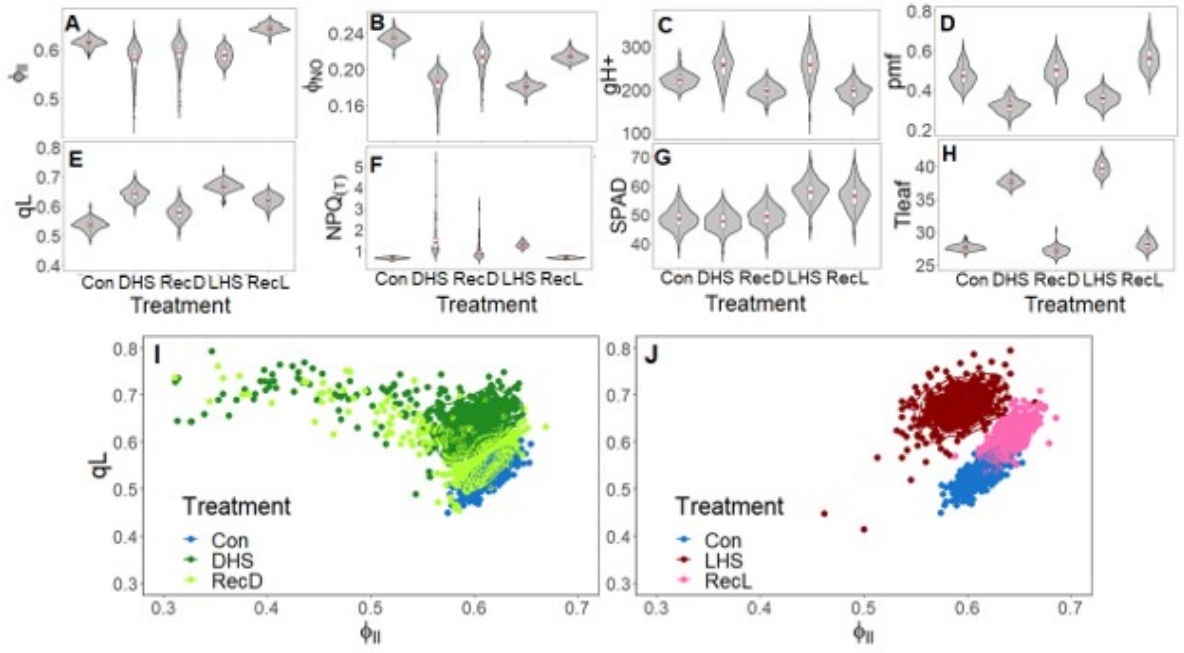

Figure 3.1: Variation in photosynthetic parameters and leaf temperature across the different treatments. A-H) Violin and box plots showing the distribution of various parameters among the RILs and parental lines. The red marker indicates the mean of all genotypes. Con= Control, DHS=dark heat stress, LHS=light heat stress , RecD=recovery after DHS and RecL=recovery after LHS. I-J) Correlation and density plots between qL and ϕ_{II} under Control and I) DHS or J) LHS using the raw data for each treatment.

Data analyses were performed using (R Core Team 2019). Subsequent analyses uses the covariate adjusted effects obtained via the analysis of covariance (ANCOVA) model.

3.2.1 Linakage maps using QTL mapping

This section assesses possible linkages between genomic variations in the RIL population and specific responses to LHS and DHS. For the linakge maps we have used genomic BLUP as discussed in earlier sections. Figures 3.2 and 3.3 show several striking features in QTL maps for ϕ_{II} under control, LHS and DHS, and recovery conditions. First, the control shows significant QTLs on chromosomes 3, 6, 9, and 10 that completely disappeared and were replaced by distinct QTLs during LHS (chromosome 2) and DHS (chromosome 2 and 6). The most likely basis for this "linkage swapping" is that, under control conditions, ϕ_{II} is modulated by one set of genetically-controlled processes and a different set of processes linked to various genetic components under stressful conditions. This interpretation is consistent with genotype-by-environment interaction, whereby genotypes may behave differently depending on the environment, and the roles of "ancillary" components of the organism, that control processes not essential under many conditions but critical under diverse and fluctuating environments. There are many examples in photosynthesis research where knocking out well-conserved genes has little effect under (artificially static) laboratory conditions but shows emergent phenotypes under more severe or rapidly fluctuating environments.

3.3 Bayesian Latent Factor Models

To set the models and notations, assume that the observed data contain N independent units. For each unit Q traits (phenotypes) are observed for S different treatments. We use $X^{(1)} \in \mathbb{R}^{Q\times N}$, $X^{(2)} \in \mathbb{R}^{Q\times N}, \ldots, X^{(S)} \in \mathbb{R}^{Q\times N}$ to denote the collection of S treatments with
dimensionality Q on N independent observations. Let Y be their vertical concatenate, which is of size $P \times N$, where $P = SQ$,

$$Y^{P\times N} = [X^{(1)}, X^{(2)}, \ldots, X^{(S)}]^T$$

Our goal is to first find out $K < P$ factors that describe the dependencies between the observed phenotypes across the data sets encompassing different treatments. In other words, the problem can be described as a set of K latent factors which contain a projection for each of the S treatments

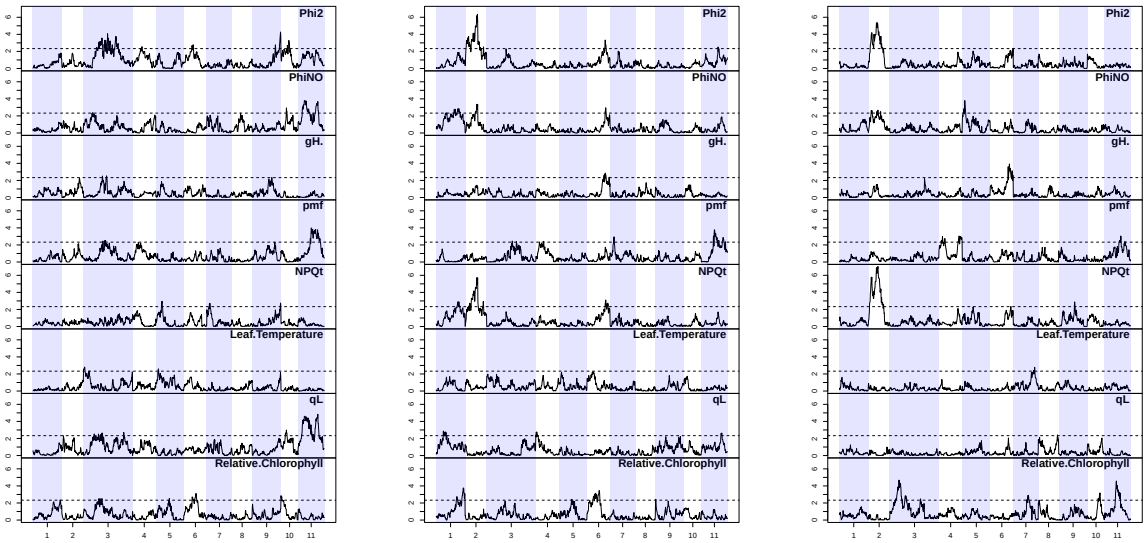

Figure 3.2: Genetic and phenotypic linkages among multiple photosynthetic processes. LOD scores for different parameters are presented for Control/Pre-stress (Con) (left most panel), DHS (middle panel), LHS (right most panel).Chromosomes are separated by transparent colors with faint lines for borders.

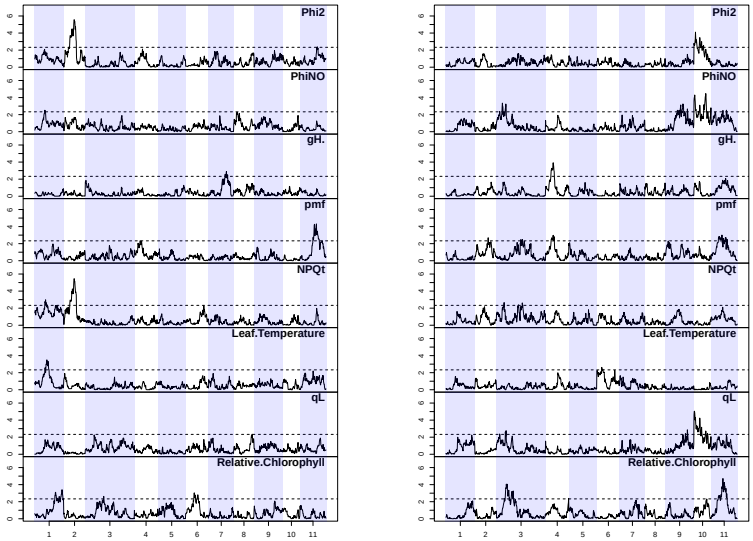

Figure 3.3: Genetic and phenotypic linkages among multiple photosynthetic processes. LOD scores for different parameters are presented for Recovery after DHS (RecD) (left panel) and Recovery after LHS (RecL) (right panel). Chromosomes are separated by transparent colors with faint lines for borders.

with a non-zero weight for each factor. Note that, one would like to put sparsity over the weights for added interpretations.

Our data challenge is very similar to the Factor analysis (FA), which explains a multivariate dataset $X \in \mathbb{R}^{Q \times N}$ in terms of $K < Q$ latent factors for defining the dependencies between the N observed samples of dimensionality Q. In FA, the underlying latent factors are connected to the observable variables through factor weights, which are collected in the loading matrix $W \in \mathbb{R}^{Q \times D}$. One can add sparsity to the individual entry of W to obtain straightforward interpretations. In our setup, we can apply FA to each of the S treatment conditions and estimate the loading matrix and factor scores for every condition. Since the loading matrix W connects the set of observed phenotypes in terms of a smaller group of latent variables (LV), one can expect that each LV identified by FA can, under appropriate conditions, represent the physical modes of interactions among photosynthetic processes, which result in distinct patterns of interactions between measured phenotypes. In a single genotype, the observed LVs will likely reflect environmental or developmental effects on mechanistic interconnections. Here, though, we consider a population of genetically distinct plants under a single environmental condition, where the LV will likely reflect the actions of genetic polymorphisms that alter the behavior of the co-regulatory network among phenotypes.

A hypothetical illustration describes two genetic components that alter photosynthesis's responses to HS in distinct ways. Consider a functional mechanism (mechanism A) representing the correlations expected for pmf-controlled photoprotection. For example, a genetic variation that decreases the activity of the ATP syntheses can result in decreases in gH+, increases in pmf, and decreases ϕ_{II}. Using FA analysis one can describe this behavior by LV1 in Figure 3.4, as a network of correlations. In this visualization, if the same colored lines connect two measured parameters, they are positively correlated, whereas different colors signify negatively correlated. From our model, decreases in gH+ should increase NPQ(t) but decreases in ϕ_{II}. Hence, gH+ is linked through LV1 with differently colored lines to NPQ(t) but with the same colored lines to ϕ_{II}. Suppose that an additional process modulates ϕ_{II} through photoinhibition and repair of PSII, independently of ATP synthesis activity (mechanism B). For example, a genetic component that

results in more rapid PSII repair should increase the content of active PSII and thus ϕ_{II} while decreasing NPQ(t). Here, parameters can be connected through multiple LVs, each describing a different set of correlations, as illustrated through LV2 in Figure 3.4.

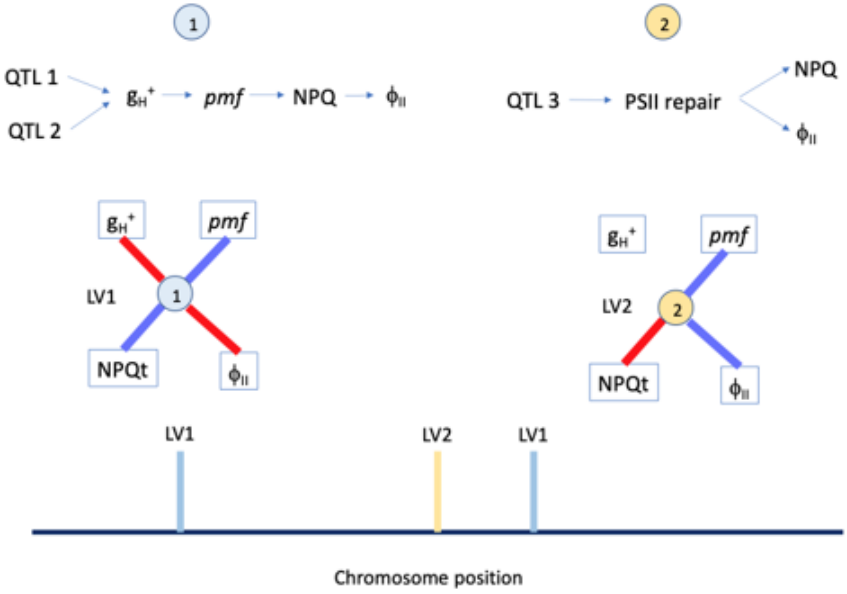

Figure 3.4: A hypothetical illustration of expected relationships between latent variables, correlations among measured parameters, and genetic components

But, FA on individual treatments fails to capture the dependencies across multiple treatment conditions. Also, the LVs under each treatment are independent, and one can not make connections or test separable models based on individual FA analysis. One solution is the Canonical Correlation Analysis (CCA,), which can simultaneously model underlying associations between two sets of treatment conditions. CCA helps identify linear combinations of variables from each modality that maximize their correlation. CCA also suffers from certain caveats. For example, they do not provide an inherent robust inference for statistical associations between phenotypes. Also, the associations between data modalities need to be modeled to capture variations. One possible way to address the caveats is based on the probabilistic interpretation of CCA [38], which allows for the uncertainty estimation of the model parameter. This approach has been extended to more complex situations by adding hierarchical prior distributions as explained in Bayesian CCA [39]. Still, it fails to recover associations among data modalities and is computationally challenging under high-dimensional

problems. One of the limitations is averted by Virtanen and colleagues [40, 41] by removing the irrelevant latent factors and further extending to more than two modalities, namely Group Factor Analysis (GFA) [42, 41]. GFA is a simple extension of the Bayesian FA model with group-wise sparsity, which helps in straightforward interpretations. GFA has proven its applicability in various domains, from genomics, drug discoveries, and task-based FMRI data [43]. To our knowledge, GFA has not been applied to reveal Phenome-Genome interactions from Multi-omics data modalities.

To illustrate the differences between various methods, we have applied Bayesian formulations of FA, CCA, and GFA to our dataset and argued the implications and interpretations under different setups. Our findings include multiple nodes of mechanistic processes representing "positive-negative" associations linking phenotypes to specific patterns under various stress conditions. We conclude from our analysis that due to the flexibility and robustness, the integrative framework of Bayesian GFA can reveal meaningful biological mechanisms previously unknown under heat stress treatments.

3.3.1 Bayesian Factor Analysis (BFA)

The Bayesian version of FA assumes that N observations of Q phenotypes stored in the data matrix $X \in \mathbb{R}^{Q \times N}$ are generated by the latent variable matrix $F \in \mathbb{R}^{K \times N}$, where K represents the number of latent dimensions. In formal notations, suppose that Q dimensional data vector y_n follows a K factor model:

$$\begin{aligned} x_n &= W f_n + \mu + \epsilon_n \\ f_n &\sim \mathcal{N}(0, \mathcal{I}_K) \\ \epsilon_n &\sim \mathcal{N}(0, \psi) \end{aligned} \tag{3.1}$$

Under the Model 3.1, $x_n \sim \mathcal{N}(\mu, \psi + WW^T)$. Without loss of generality, we can assume zero mean data and omit the μ parameter hereafter. To tackle FA model in Bayesian context, we introduce a prior $p(\theta)$ over the model parameters $\theta = (W, \psi)$ with respect to the posterior distribution $p(\theta|Y)$. For simpler inferences, the prior distributions are selected to be conjugate such that the posterior

distribution has the same functional form as the prior distribution. To determine the number of latent dimensions to be included in the model, we incorporate Automatic relevance Determination (ARD) prior over the loading matrix W. This is achieved through a hierarchical prior specifications $p(w|\alpha)$ on the elements of W, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$. Inherently, by pushing some α_k's towards infinity, one can drive the elements of the loading matrix of W to become close to zero. This results in the pruning of the irrelevant latent components k during inference.

$$\begin{aligned} p(W|\alpha) &= \prod_{j=1}^{Q}\prod_{k=1}^{K} \mathcal{N}(w_{j,k}|0, \alpha_k^{-1}), \\ p(\alpha) &= \prod_{k=1}^{K} \Gamma(\alpha_k|a_\alpha, b_\alpha), \\ p(\psi) &= \mathcal{W}^{-1}(\psi|\Lambda_0, v_0), \end{aligned}$$

where, $\Gamma(\cdot)$ denotes a gamma distribution and Λ_0 represents a symmetric positive definite matrix and v_0 is the degrees of freedom of the inverse Wishart distribution ($\mathcal{W}^{-1}(\cdot)$). The joint probabilistic distribution of the model 3.1 is given by,

$$p(X, F, W, \alpha, \psi) = \Big[p(X|F, W, \psi)p(W|\alpha)P(\alpha)p(\psi)\Big]p(F)$$

To estimate the model parameters and the latent variables , we need to evaluate the posterior distribution $p(F, W, \alpha, \psi|X)$ and marginalising the unintended variables. However, the marginalisations are very complex and often analytically intractable. Thus, the posterior distribution needs to be approximated.

3.3.2 Bayesian Canonical Correlation Analysis (BCCA)

In Bayesian CCA, we assume that N observations from two different data modalities, $X^{(1)}$ and $X^{(2)}$ are generated from a common latent variables $F \in \mathbb{R}^{K\times N}$.

Similar to Model 3.1, we can write,

$$\begin{aligned} x_n^{(1)} &= W^{(1)} f_n + \epsilon_n^{(1)} \\ x_n^{(2)} &= W^{(2)} f_n + \epsilon_n^{(2)} \\ f_n &\sim \mathcal{N}(0, \mathcal{I}_{\mathcal{K}}) \\ \epsilon_n^{(1)} &\sim \mathcal{N}(0, \psi^{(1)}) \\ \epsilon_n^{(2)} &\sim \mathcal{N}(0, \psi^{(2)}) \end{aligned} \tag{3.2}$$

Here, in Model 3.2, $W^{(1)} \in \mathbb{R}^{Q \times K}$ and $W^{(2)} \in \mathbb{R}^{Q \times K}$ are the projection matrices which transform the latent variables f_n into the input space of two separate treatments. The joint distribution is given by,

$$\begin{aligned} p(X, F, W, \alpha, \psi) &= \prod_{n=1}^{N} \prod_{s=1}^{2} \left[p(x_n^{(s)} | f_n, w^{(s)}, \psi^{(s)}) p(w^{(s)} | \alpha^{(s)}) P(\alpha^{(s)}) p(\psi^{(s)}) \right] p(f_n), \\ p(w^{(s)} | \alpha^{(s)}) &= \prod_{j=1}^{Q} \prod_{k=1}^{K} \mathcal{N}(w_{j,k}^{(s)} | 0, \alpha_k^{(s)^{-1}}), \\ p(\alpha^{(s)}) &= \prod_{k=1}^{K} \Gamma(\alpha_k^{(s)} | a_\alpha^{(s)}, b_\alpha^{(s)}), \\ p(\psi^{(s)}) &= \mathcal{W}^{-1}(\psi^{(s)} | \Lambda_0^{(s)}, v_0^{(s)}) \end{aligned}$$

Here, the prior distributions are chosen so that the posterior distributions has the same functional form. The prior distribution over the loading matrix is chosen to be ARD priors similar to be BFA setup which helps in recovering the relevant latent factors. The inference of model parameters and latent variables depends on computing the posterior distribution, $p(F, W, \alpha, \psi | X)$ which is analytically intractable and should be approximated. Following Chong Wang [2007], one can use the mean field variational Bayes, or the Gibbs sampling. Even there the inference becomes unusually cumbersome in the presence of high dimensional data. To overcome this, Virtanen er al (2011) proposed to impose modality wise sparsity. A further extension has been proposed by the same authors which generalizes the same idea to more than two data modalities, which is known as the Bayesian group factor analysis.

3.3.3 Bayesian Group Factor Analysis (BGFA)

For the group factor analysis, we assume that there are S many data modalities, where the s'th data modality is being represented as $X^{(s)} \in \mathbb{R}^{Q \times N}$, $s = 1, \ldots, S$. Now, equivalent to the latent factor components discussed in BFA and BCCA, BGFA tries to find the optimal set of K latent factors which can separate between-group associations from within-group associations. Mathematically, one can write the data from the s'th group generated as follows,

$$\begin{aligned} f_n &\sim \mathcal{N}(0, \mathcal{I}_{\mathcal{K}}) \\ x_n^{(s)} &= W^{(s)} f_n + \epsilon_n^{(s)} \\ \epsilon_n^{(s)} &\sim \mathcal{N}(0, T^{(s)^{-1}}) \end{aligned} \tag{3.3}$$

Where $T^{(s)^{-1}}$ denotes a diagonal covariance matrix, with $T^(s) = diag(\tau_1^{(s)}, \ldots, \tau_Q^{(s)})$ as the inverse of the error variances of the s'th group. The structure of the loading matrix, W, and the latent structures, F, are automatically learned by imposing group-wise sparsity through the independent ARD priors. The automatic pruning of the unimportant latent components is achieved by putting a separate ARD prior to the elements of $W^{(s)}$,

$$\begin{aligned} p(w^{(s)}|\alpha^{(s)}) &= \prod_{j=1}^{Q} \prod_{k=1}^{K} \mathcal{N}(w_{j,k}^{(s)}|0, \alpha_k^{(s)^{-1}}), \\ p(\alpha^{(s)}) &= \prod_{k=1}^{K} \Gamma(\alpha_k^{(s)}|a_\alpha^{(s)}, b_\alpha^{(s)}), \\ p(\tau^{(s)}) &= \Gamma(\tau^{(s)}|a_\tau^{(s)}, b_\tau^{(s)}) \end{aligned}$$

We have chosen the hyperparameters $a_\alpha^{(s)}, b_\alpha^{(s)}), a_\tau^{(s)}, b_\tau^{(s)}$ to be very small number (e.g., 10^{-14}) in order to get uninformative priors. Finally, we can write the joint distribution as,

$$p(X, F, W, \alpha, \tau) = \prod_{n=1}^{N} \prod_{s=1}^{S} \left[p(x_n^{(s)}|f_n, w^{(s)}, \tau^{(s)}) p(w^{(s)}|\alpha^{(s)}) P(\alpha^{(s)}) p(\tau^{(s)}) \right] p(f_n)$$

Note that, the posterior calculations are often analytically intractable and it needs to be approximated through mean field variational approximation.

3.3.4 Mean Field Variational Approximation

In Bayesian settings, the calculations regarding the posterior distributions are computationally challenging. The way around is to approximate the true posterior with a suitable factorized distribution through Variational Bayesian (VB) setting. Let the model parameters are denoted by θ and our goal is to approximate the true posterior, $p(\theta|X)$ with the help of the variational distribution, $q(\theta)$. The main idea in VB is to minimize the dissimilarity, $D(q;p)$ between $q(\theta)$ and $p(\theta|X)$. The most used dissimilarity measure in such cases is the Kullback–Leibler divergence (KL-divergence) which makes this minimization tractable. In theory, the KL-divergence is written as,

$$D_{KL}(q||p) = \int q(\theta)\, ln\, \frac{p(\theta|X)}{q(\theta)}\, d\theta$$

Following Bishop [44], the marginal log-likelihood is written as,

$$\begin{aligned} \mathcal{L}(q) &= \int q(\theta)\, ln\, \frac{p(X,\theta)}{q(\theta)}\, d\theta \\ ln\, p(X) &= \mathcal{L}(q) + D_{KL}(q||p), \end{aligned}$$

where $\mathcal{L}(q)$ is the lower bound of the marginal log likelihood. Note that, $ln\, p(X)$ is constant and this implies that maximizing the Evidence lower bound (ELBO) $\mathcal{L}(q)$ is equivalent to the minimization of the KL-divergence $D_{KL}(q||p)$. We assume that $q(\theta)$ can be factorized as $q(\theta) = \prod_i q_i(\theta_i)$ and $\mathcal{L}(q)$ is maximized with respect to all possible $q_i(\theta_i)$,

$$ln\, q_i(\theta_i) = < ln\, p(X,\theta) >_{j\neq i} + constant,$$

where $< \cdot >$ represents the expectation taken with respect to $\prod_i q_i(\theta_i)$ for all $j \neq i$. In BGFA, we can approximate the full posterior by the following variational distribution,

$$q(\theta) = q(F) \prod_{s=1}^{S} q(W^{(s)}) q(\alpha^{(s)}) q(\tau^{(s)}),$$

where $\theta = \{F, W, \alpha, \tau\}$. Since we have assigned conjugate priors, we can obtain the following analytically tractable solutions after optimizing $q(\theta)$.

$$\begin{aligned} q(F) &= \prod_{n=1}^{N} \mathcal{N}(f_n|\mu_{f_n}, \Sigma_{f_n}), \\ q(W^{(s)}) &= \prod_{j=1}^{Q} \mathcal{N}(W_{j,*}^{(s)}|\mu_{W_{j,*}^{(s)}}, \Sigma_{W_{j,*}^{(s)}}), \\ q(\alpha^{(s)}) &= \prod_{k=1}^{K} \Gamma(\alpha_k^{(s)}|\tilde{a}_{\alpha^s}^{(k)}, \tilde{b}_{\alpha^s}^{(k)}), \\ q(\tau^{(s)}) &= \prod_{j=1}^{Q} \Gamma(\tau_j^{(s)}|\tilde{a}_{\tau^s}^{(j)}, \tilde{b}_{\tau^s}^{(j)}), \end{aligned} \tag{3.4}$$

where the j'th row of $W^{(s)}$ is denoted by $W_{j,*}^{(s)}$. For the optimization, we can follow the variational Bayes Expectation-Maximization (VBEM) scheme. For the convergence, we assign the relative change of ELBO, $\mathcal{L}(q)$ to fall below a preassigned small value (e,g., 10^{-5}). This is essentially a sequential procedure where the parameters are updated sequentially. We have only listed the optimization scheme for BGFA. Procedures for BFA and BCCA follow similar strategies.

3.4 Results

Although the comparative LOD profiles and QTL linkages shown in Figure 3.2 and 3.3 reflect genetic linkage between the measured traits, one can not comprehend the complex nature of the interaction between these photosynthesis regulatory partners. Furthermore the question remains if there exist multiple interaction co-occurring withing a treatment conditions and modulated through different genetic components. For example, the QTL maps show apparent linkages between different subsets of measurable parameters under different conditions. Under DHS we observed one set of overlapping QTLs on chromosome 6 between ϕ_{II}, gH+, NPQt, ϕ_{NO} and ϕ_{NPQ}, and a distinct set on chromosome 2 with linkages between ϕ_{II} and ϕ_{NPQ}, as well as NPQt (in LHS) but not the other parameters. These complexities likely reflect the pleiotropic, time-and condition-dependent interactions among processes and genetic loci. To mitigate this problem we have performed BFA on each treatment conditions. We observed several trends in the BFA analyses and associated QTL

maps, as described in the following.

First, different LVs were linked to distinct sets of QTLs. For example, under DHS, LV1 mapped to a QTL on chromosome 6, LV2 to a QTL on chromosome 1 and LV4 mapped to a QTL on chromosome 4. The segregation of LV with distinct QTLs was consistently observed across all conditions, suggesting that BFA was able to partition the observed variation into distinct modes of behavior that are influenced by different sets of genetic components.

Second, most QTLs associated with LVs were also observed from QTL analyses of individual parameters. When multiple parameters were linked through a LV, we also observed overlapping QTL from maps of the individual parameters (Figure 3.2 and 3.3), providing further support that EFA coupled to QTL mapping was able to identify possible mechanistic and genetic linkages between traits associated with distinct biochemical/physiological behaviors. For example, leaf temperature (Tleaf) and relative chlorophyll content (SPAD) showed genetic associations, but only weak functional connections to other parameters, suggesting that genetic variations in the traits do not, under our conditions, strongly influence the photosynthetic control mechanisms.

Third, some LV showed functional trends but did not show measurable genetic linkages. For example, the linkage between ϕ_{NO} and gH+ on LV1 during DHS and the link between pmf and gH+, on LV3 during recovery from LHS (Figure 3.5 and 3.6) showed only small associations with genomic loci. This behavior may indicate that the observed variations were controlled through many small effect loci that did not result in measurable associations using our current techniques.

Next, BFA revealed changes in mechanistic interactions and genetic control modes under different environmental challenges. Under each treatment, both the patterns of correlations among parameters and the QTL linkages for the contributing LVs were distinct. For example, a key distinction between Control and DHS was the change in the sign of the correlations between traits for LV1. In the control, ϕ_{NO} and qL were negatively and positively linked to ϕ_{II} on LV1 respectively. Under DHS, ϕ_{NO} and qL became positively and negatively linked respectively at LV1, with no change in the directionality of the linkage for ϕ_{II} (Figure 3.5). These changes imply that the effects of genetic variations on functional/regulatory interactions are distinct under the different

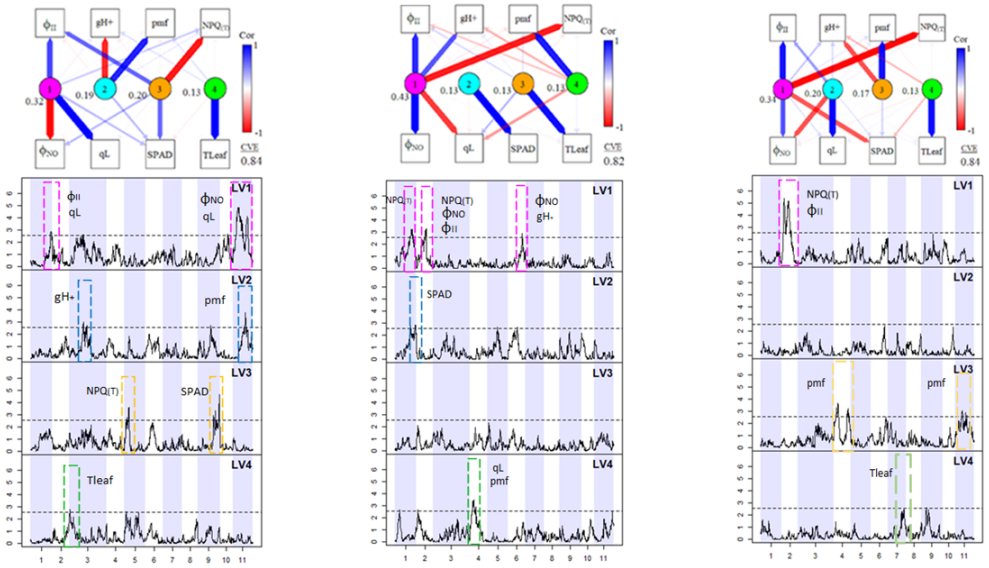

Figure 3.5: Bayesian Factor Analysis coupled with QTL mapping for genetic linkage between photosynthetic traits under Control/Pre-stress(Con) (left most panel), DHS (middle panel), LHS (right most panel).

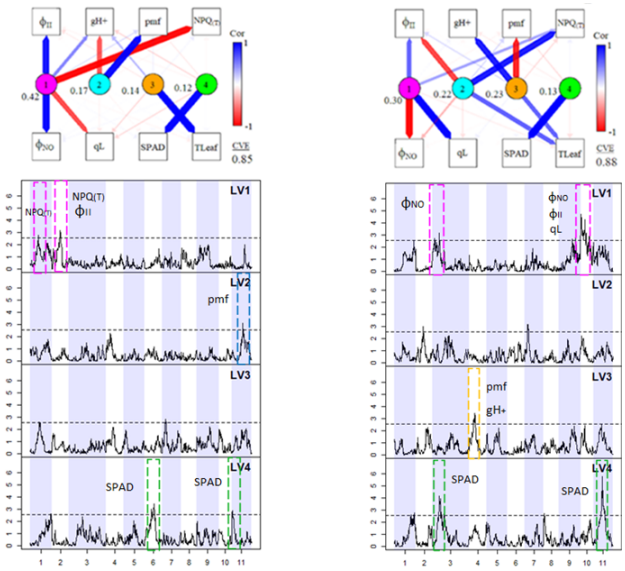

Figure 3.6: Bayesian Factor Analysis coupled with QTL mapping for genetic linkage between photosynthetic traits under Recovery after dark heat stress (RecD) (left panel), and Recovery after light heat stress (RecL) (right panel).

treatments. Moreover, even when the functional linkages among parameters were similar, e.g. comparing control and LHS Recovery (Figure 3.6), the LV loading factors mapped to distinct sets of loci, suggesting that different processes and loci are involved in maintaining and reestablishing photosynthetic responses before and after LHS. Overall, these changes in functional and genomic linkages suggest that different sets of genetic components influence these behaviors under different conditions.

Finally, BFA may resolve distinct but overlapping QTLs. When comparing QTLs of individual parameters, we observed an apparent linkage between pmf and qL under Control condition (Figure 3.2), whereas BFA suggested that these two parameters are controlled by different interaction networks (LVs) (Figure 3.5). While we cannot rule out the possibility that this separation was caused by the limitations of our approach, it suggests that BFA may distinguish between distinct but closely-linked QTLs as long as they control distinct patterns of behaviors.

One of the shortcoming of using BFA in our framework is we can not make inference from combining two or more treatment conditions. Similar to clustering approach, BFA can only explain within group variation. This results in a lack of connection between the latent factors from different data modalities. For example, the LV1 from Control condition is not comparable with the LV1 from DHS condition. Consequently, the QTL maps from LVs can not fully resolve the colocalization between interlinked parameters from multiple treatments.

To mitigate the between group association and possible mechanistic linkages among the different treatments we incorporated the between group interactions into our analysis through BCCA and BGFA. In BCCA, we conducted pairwise comparisons among treatments with different combinations of photosynthetic parameters. In Figure 3.7, we compared the two treatment combinations, Control and Dark heat stress from the photosynthetic parameters, ϕ_{II}, pmf, qL, and NPQt. The upper right panel shows log of the estimated ARD matrix shown as Hinton diagrams where blue segment corresponds to active components while red segment corresponds to inactive ones. To compute the number of latent factors, We compared different Hinton diagrams and ground truthed it based on QTL from each latent factors. Here, we found three Latent factors to be optimal with LV1

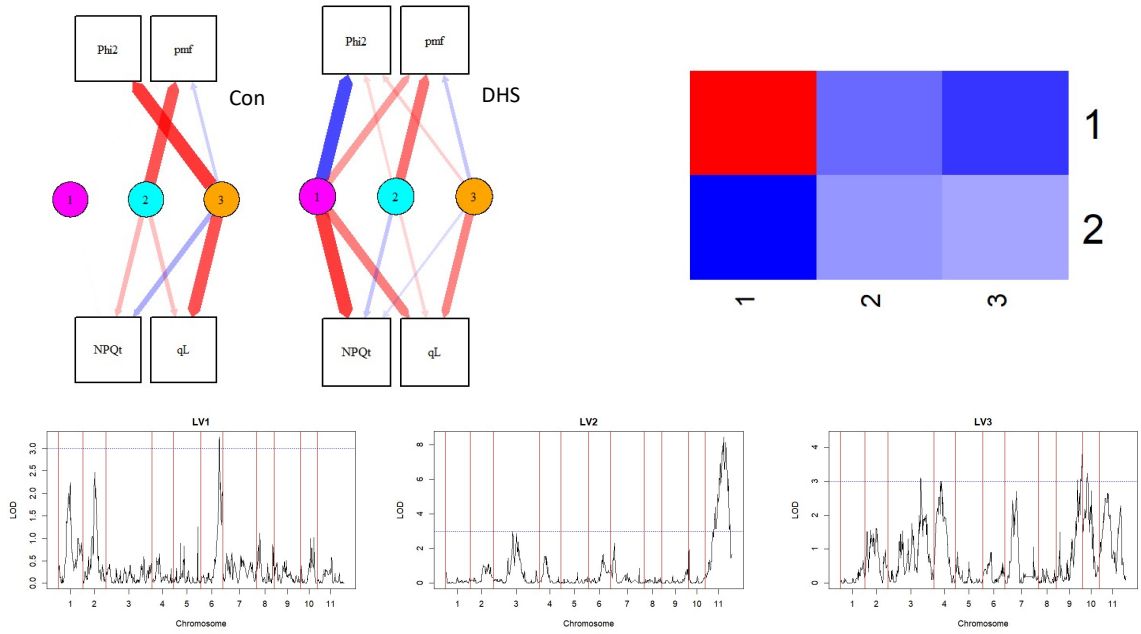

Figure 3.7: Bayesian Canonical Correlation analysis coupled with QTL mapping for genetic linkage between photosynthetic traits under Control (Con) and dark heat stress (DHS).

is specific to DHS, LV3 is specific to Control. One interesting distinction between the interactions among the parameters was that, under Control condition, ϕ_{II} and qL were found to be positively correlated whereas under DHS, ϕ_{II} and qL were negatively correlated with a strong association with NPQt. Also, this LV1 was found to be mapped with the QTLs found with pmf under DHS (Figure 3.2).

One of our findings from this approach is that in the case where multiple trait interact in opposite directions (negative correlations), their affect will cancel out and the resulting LV will not detect the genetic linkages found from individual QTL maps. This is because the LVs are linear combinations of the individual weights obtained for each trait. For example, in our heterogeneous population, the observed negative correlations between ϕ_{II} and NPQt was reflected as a "damped out" effect in

LV3 and the apparent linkage in Chromosome 2 in missing.

We further applied BCCA on the treatment conditions, DHS and LHS on the same set of photosynthetic parameters (Figure 3.8). We found two latent factors to be optimal in this setup with LV1 specific to DHS and LV2 specific to LHS. The notable difference in the mechanistic connection between the two treatments is through the lack of connection between NPQt and pmf under LHS. Also, LV2 mapped out to chromosome 2, which co-localized with the ϕ_{II} and NPQt under LHS (Figure 3.2). In order to explore the differences and associations between a set of

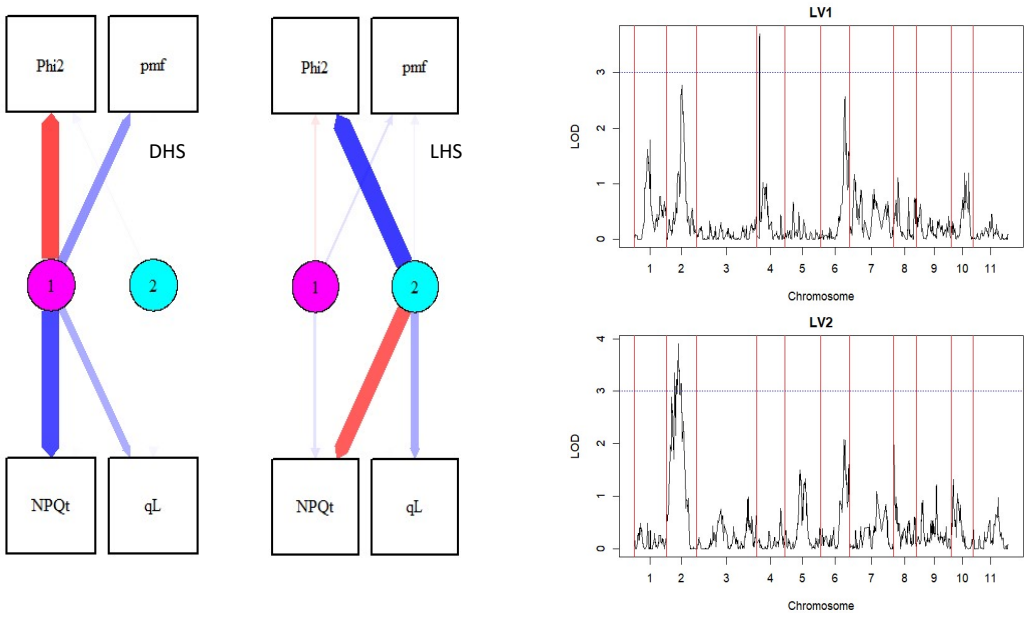

Figure 3.8: Bayesian Canonical Correlation analysis coupled with QTL mapping for genetic linkage between photosynthetic traits under dark heat stress (DHS), and light heat stress (LHS)

measured phenotypes across five treatments, we applied BGFA by concatenating the observed data matrices vertically. As we have discussed earlier, the resulting loading matrix (W) provided the connections with measured phenotypes with the auxiliary latent variables and the latent factors (f_n) were mapped with QTL mapping to show different genetic linkages. The number of latent

variables optimal under this case were optimally chosen to be five. In Figure 3.9, we plotted out the loading matrices corresponding to each treatments and Figure 3.10 showed the genetic linkages corresponding to each latent factors. We found that LV2 was specific to Control condition, and LV1 was specific to LHS. LV4 was shared by DHS and RecD, whereas LV3 was shared by LHS and RecD. LV5 was shared by all the treatment conditions barring LHS. One of the key mechanistic linakages we found is based on the LV4, where the connection of pmf with LV4 is missing under RecD, but present in DHS. Also, we found the canonical connection between NPQt and pmf was missing under LHS, which is consistent across the BCCA and BGFA analysis. Since the LVs are comparable across treatments, we can test for the amount of interactions between traits being modulated by any particular LV. For example, LV5 which is shared by Con, DHS, RecD and RecL has a different extent of interactions across treatments. The connection between ϕ_{II}, ϕ_{NO}, pmf and qL is significantly stronger in Control from others.

From the genetic linkages obtained from the LVs we can confirm the colocalization of ϕ_{II}, ϕ_{NO} and NPQt being modulated by a QTL regions at chromosome 2. We found a QTL peak at chromosome from LV4 which could be mechanistically linking phenotypes of interest. Also, with LV5 and LV3, we found QTL peaks at chromosome 7 which were not found from individual QTL msps or with BFA or BCCA.

3.5 Discussion

We aimed to extend the analyses of biophysical measurements of photosynthesis to understand how nature has tweaked key processes to respond to changing environmental conditions. This is possible because of the availability of inexpensive genomic sequencing and the development of rapid and detailed phenotyping that combines measurements of photosynthetic regulatory networks at multiple points. The combined data can give a more resolved view of the interplay of biophysical processes in vivo and the genetic components that control them. However, methods to handle such hyperdimensional data sets are still being developed. While it is possible to make predictions from such data sets using ML, the need to generate and test specific mechanistic hypotheses is

essential to the scientific method. The methods described representing first-order attempts to use these combined tools to compress hyperdimensional data into usable forms (LVs) and use these to generate and test hypothetical models for how genetic polymorphisms impact the regulatory network of photosynthesis. We also show that latent factors can provide a deeper analysis of more complex interacting networks, by teasing apart distinct modes of interactions and specific genetic components that control them. The results strongly support the view that the regulatory network is

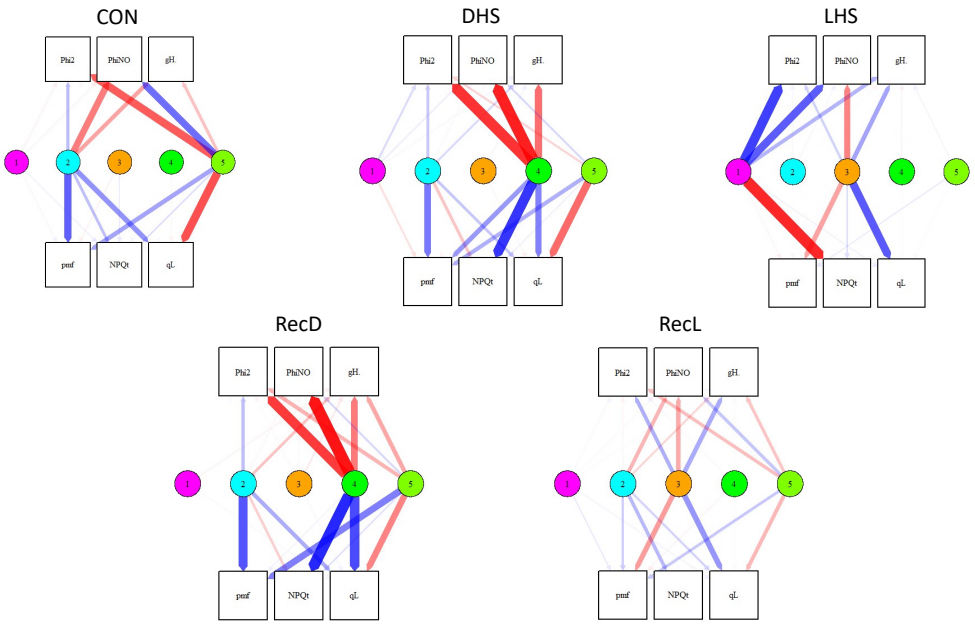

Figure 3.9: Bayesian Group factor analysis of photosynthetic traits under Control (Con), dark heat stress (DHS), light heat stress (LHS), Recovery after dark heat stress (RecD) , and Recovery after light heat stress (RecL).

highly flexible and controlled by distinct sets of ancillary genetic components depending on specific environmental challenges, consistent with the genotype-by-environment interaction paradigm.

We conclude that, in the cowpea diversity panels we used and under the conditions of our

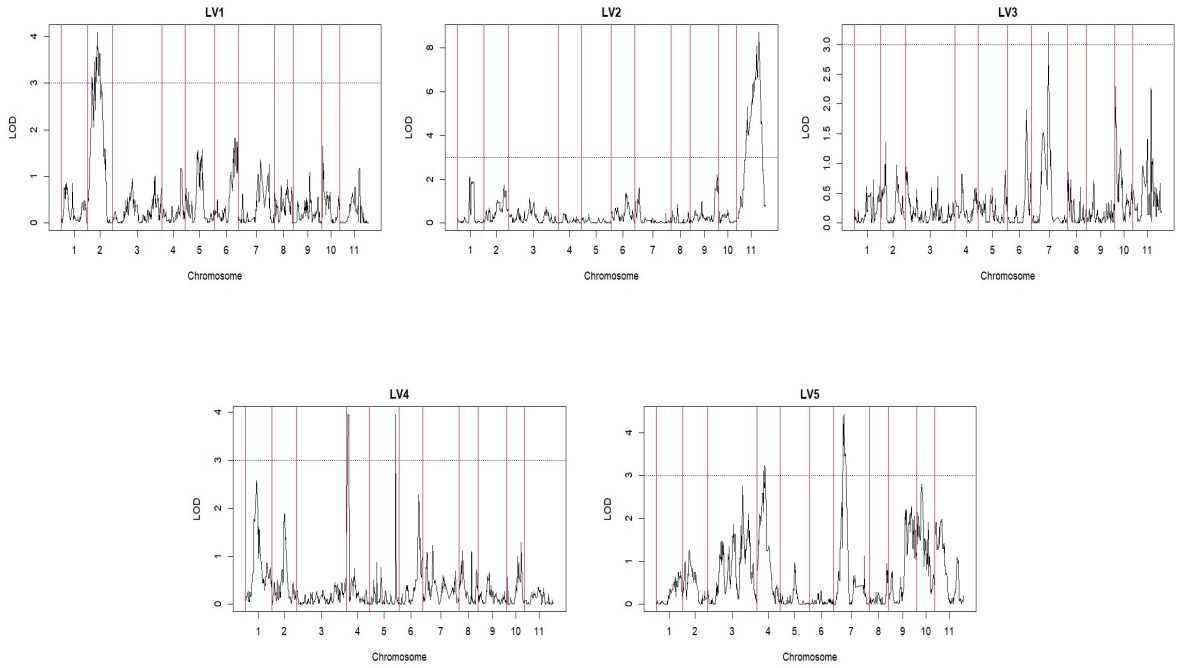

Figure 3.10: QTL analysis of resulting LVs from BGFA

experiments, genetic variations observed in leaf movements do not lead to measurable variations in photoprotection under low-temperature stress. Further, responses to DHS and LHS are governed by distinct genetic variations that broadly impact non-qE-dependent NPQ mechanisms, but not qE or transpiration rates. In addition, genetic variation in transpiration-induced cooling did not influence the tolerance of PSII to LHS and DHS. the methods will no doubt be advanced by increasing the diversity and resolution of genetic variants, the numbers of specific processes measured, and the sophistication of the modeling, including the use of machine learning.

These observations suggest that latent factors can be helpful in applications towards generating hypothetical models from genetic diversity experiments that measure multiple, functionally-related phenotypes. Because we used the results of latent factors for subsequent analyses, i.e., QTL

mapping, combining the two marks and interpreting one in light of the other provides some confidence in the conclusions. His approaches should also be useful for crop improvement efforts, especially in identifying specific mechanisms and genetic components that modulate photosynthetic efficiency and resilience under diverse environmental challenges.

The results also emphasize certain caveats that need to be considered for immediate applications and improved methods of development. Some of these issues can be alleviated by introducing functions to linearizing parameters or adding additional measurements that discriminate between possible mechanisms. other issues, including the simplified assumptions of linear interdependencies and compensation between parameters, multiple interpretations of correlational data, etc., will require the development of next-level approaches, such as extended methods like clustering algorithms to determine and constrain possible LV structures. Also, using FA, we explored the possibility of deciding possible latent space in the phenotypic area which can regulate specific phenotypic interactions. They intend to extend this knowledge by possibly backtracking the phenotypic interactions by exploring the latent factors in the genomic space. While incorporating the gene regulatory network in our desired data, we expect to observe the gene-driven pathways controlling the phenotypic interactions.

LV structures help explain the mechanistic bases of biophysical mechanisms corresponding to causal pathways (domain-specific) in phenotype interactions. In fact, our empirically motivated LV proposes a new research theme for understanding the interdependence across the $Genotype \times Phenotype \times Environment$ space. Methodological and practical innovations for quantifying such pathways provide scientific grounds for the functions of photosynthetic regulatory pathways. However, dominant correlations in a system can result from parallel transitive or indirect interactions. We show certain classes of hypotheses can be generated and tested using simple comparisons of QTL maps. But still the question remains how we can model the interactions or correlations among the measured phenotypes with a given set of predictors. We model the correlations among multiple traits with a selected number of predictors in the following chapter.

CHAPTER 4

CMPLE TO DECODE PHOTOSYNTHESIS USING THE MINORIZE-MAXIMIZE ALGORITHM

4.1 Motivation

4.1.1 General background

Understanding photosynthesis, how solar energy transduction enables and limits the energy productivity of crops is critical for improving the quality and resilience of agriculture in a rapidly changing world. Abiotic stress factors, e.g., high light intensities, high or low temperatures, lack of water, inhibit the ability to use light energy productively and lead to photodamage to the photosynthetic machinery [45]. Plants can maintain photochemistry to adapt to the challenges of non-ideal environments using a range of mechanisms, where several photosynthetic responses can contribute to this maintenance of yield, and it is possible to harness these variations to improve crop performance. However, the dynamics of photosynthetic responses may include complex interactions among species, genotypes, developmental stages, or other environmental conditions. The recent development of high-throughput phenotyping platforms [46, 47] can rapidly and non-invasively measure multiple, potentially related, photosynthetic traits and environmental parameters. Analyzing such voluminous data with complex interactions among multiple traits, genotypes, and environmental variables requires computationally efficient and interpretable statistical models that can potentially explore the mechanistic bases of useful or adaptive photosynthetic processes.

Several methods have been suggested for investigating the statistical association between measured traits and genetic markers, including genome-wide association studies (GWAS) and whole-genome regression (WGR) approaches, which produce familiar quantitative trait loci (QTL) map [48, 49]. Standard QTL mapping has mainly been used to analyze the genetic association with individual traits. Nevertheless, alterations in genetic loci can affect the associations between multiple characteristics, classified as meaningful biological mechanisms. This is particularly important

when addressing important but complex traits such as photosynthetic efficiency or crop yield, which can be affected by multiple processes under different conditions. It is thus essential to interpret associations between traits using genetic markers and determine if variations at different genetic markers affect the inter-relationships among traits through similar or distinct mechanisms. A natural choice for multiple-trait analysis is to extend single-trait GWAS or WGR methods directly to the multiple-trait domain [50, 51]. But, characterization of such methods to elucidate the association among multiple traits remains challenging [52, 53]. We address a few of the challenges below.

Multiple-trait analysis tools do not exploit the information in the correlation matrix of related traits and thus cannot connect them with genetic and environmental predictors. Pleiotropy, the effect of genetic diversity on multiple traits, plays a significant role under different abiotic stresses [54, 55]. Without modeling the correlation matrix, one can not fully express the occurrence of pleiotropy in real-world applications. Also, dimension reduction procedures, where a multivariate response is summarized into a univariate score using principal component (PC) analysis, have limited usage due to its lack of interpretability. To address this stated need, we propose an interpretable model of the variance-covariance matrix in terms of the predictor variables and related inference.

Pourahmadi [56] used the Cholesky decomposition, and expressed the entries of the variance-covariance matrix in terms of the unrestricted parameters and guaranteed positive-definiteness of the variance-covariance matrix. Although one could model these unrestricted parameters in terms of the predictor variables, the regression parameters do not have any easy interpretation. Alternatively, one can model the covariance matrix as a parsimonious quadratic function of predictor variables [57]. For modelling the variance-covariance in terms of predictor variables, Zou et al. [58] proposed to use a regression model for the second moments of the response variable. The authors then imposed a positivity restriction on the resulting eigenvalues to ensure positive definiteness of the variance-covariance matrix. Unfortunately, for these methods, model parameters lack direct interpretation when correlations among responses are of utmost interest.

A downside of correlation modeling is the computational burden to estimate many parameters [59]. If there are p predictors and q traits, the correlation and standard deviation modeling involve

at least $(p+1)q(q+1)/2$ model parameters. The estimation of so many parameters is challenging and computationally expensive.

These limitations motivate us to develop a framework to model the correlations and standard deviations among the responses in terms of several predictor variables. We use the pairwise composite likelihood method for statistical inference. For efficient estimation of the parameters we develop an Minorize-Maximize (MM) algorithm. The method is abbreviated as CMPLE for Correlation Modeling under Pairwise Likelihood Estimation. Specifically, by comparing the impacts of genetic variations on the correlations among a set of related phenotypes, we can distinguish between certain classes of (well-defined) hypothetical biological models and determine whether combinations of genetic variations and environmental conditions affect similar or distinct mechanisms. We show that it is possible to distinguish between classes of hypothetical models under certain conditions, leading to new biological discoveries. This analysis has direct application in plant breeding research. We predict that by applying CMPLE to diversity panels from different species, we can reveal additional mechanisms of adaptation and will guide the breeding and engineering of photosynthesis for higher, more climate-resilient productivity.

4.1.2 Contributions to the literature

Finding the possible genetic variations and environmental conditions that dictate the photodamage or photoprotection is a critical step in improving photosynthetic yield and productivity. We believe that modeling the pairwise correlations through the genetic and environmental predictors is the best way to explore the dynamic nature of the problem stated. With this goal in mind, we have developed CMPLE, where the correlations among different traits are subjectively modeled and estimated using a pairwise composite likelihood framework. The pairwise-composite likelihood method has, in the past, been used in different contexts. For example, Lele and Taper [60] used it in the estimation of variance components, Gao et al. [61] used it in genome-wide association studies, and Bai et al. [62] used it in spatial-clustered data. However, to the best of our knowledge, the pairwise-composite likelihood method has not been used to model pairwise correlations. Our

work directly models the correlations and standard deviation in terms of predictor variables.

Instead of the pairwise likelihood approach if one tries the conventional full likelihood based inference using the q variate response, then the parameters of the standard deviations and correlations need to be estimated in such a way that the resulting $q \times q$ variance-covariance matrix is positive definite. Without any doubt, this is an exceptionally hard optimization problem and difficult to interpret in practical situations. Pairwise likelihood approach allows the modeling of pairwise correlation between q multivariate responses avoiding the requirement of the $q \times q$ variance covariance matrix to be positive definite. In real life settings, biologists need to address how the pairwise correlations are related to the predictors, not the entire variance-covariance matrix. Our estimated model parameters have easy interpretations which can be directly applied in various situations.

Our approach also mitigates the problem of the computational burden. To alleviate the computational issue, we develop a Minorize-Maximimize (MM) algorithm [63] for parametric estimation. Although the MM algorithm has been successfully used in different areas [64, 65, 66], it has never been used in correlation modeling. The critical aspect of the MM algorithm is to find a suitable minorizing function that helps optimize a complex objective function (aka logarithm of the pairwise composite likelihood, in our case). There is no standard recipe to obtain a minorizing function. It is very much problem-specific and requires innovative use of mathematical inequalities. Nevertheless, our numerical studies show that the use of the MM algorithm can reduce the computation time manifold. It has also demonstrated superior performance while handling a large number of parameters. We have developed an R function, called CMPLE which can be readily applied while modeling correlations between multiple responses in terms of predictors (both continuous and categorical).

4.2 Models and notations

4.2.1 Background

To set the models and notations assume that the observed data are collected from n independent units/subjects. For each unit, q traits (phenotypes) and p features (candidate genes) are observed. Let $Y_{i,j}$ and $X_{i,r}$ be the jth trait and the rth feature of the ith unit, $j = 1, \ldots, q$, $r = 1, \ldots, p$, and $i = 1, \ldots, n$. The goal is to study the correlation between any pair of phenotypes and investigate how this correlation is regulated by a set of features. Let us assume that conditional on the covariate $X_i = (X_{i,1}, \ldots, X_{i,p})^T$, $Y_i = (Y_{i,1}, \ldots, Y_{i,q})^T$ follows a multivariate normal distribution with mean $\mu_i = 0$, and variance-covariance matrix Σ_i. The goal is understanding the correlation and its behavior with respect to the features. The variance-covariance matrix can be presented as $\Sigma_i = \text{Diag}(\sigma_{i,1}, \ldots, \sigma_{i,q})\, R_i\, \text{Diag}(\sigma_{i,1}, \ldots, \sigma_{i,q})$, where $R_i = ((\rho_{i,j,k}))$ is the $q \times q$ correlation matrix for the q phenotypes from the ith subject, and the variance of $Y_{i,j}$ is denoted by $\sigma^2_{i,j}$.

4.2.2 Correlation modeling

To achieve that goal, $\rho_{i,j,k}$, the pairwise correlation between $Y_{i,j}$ and $Y_{i,k}$, is written as $\rho_{i,j,k} = g^{-1}(\eta_{i,j,k})$, where $g : (-1, 1) \to (-\infty, \infty)$ is a known link function to transform the correlation to the linear predictor defined as $\eta_{i,j,k} = \delta_{j,k,0} + \sum_{r=1}^{p} \delta_{j,k,r} X_{i,r}$, where $\delta_{j,k} = (\delta_{j,k,0}, \delta_{j,k,1}, \ldots, \delta_{j,k,p})^T$ is the regression parameter. Observe that $\eta_{i,j,k} = g(\rho_{i,j,k})$, and we require that g to be a one-to-one function. There are many popular choices for the link function g. For the convenience, we take $g(\bullet) = \log\{(1 + \bullet)/(1 - \bullet)\}$. This results in

$$\rho_{i,j,k} \;=\; g^{-1}(\eta_{i,j,k}) = 1 - \frac{2}{1 + \exp(\eta_{i,j,k})} = 1 - \frac{2}{1 + \exp(\delta_{j,k,0} + \sum_{r=1}^{p} \delta_{j,k,r} X_{i,r})}. \qquad (4.1)$$

The regression coefficient $\delta_{j,k,r}$ has a monotone linear relation with the correlation. Hence, we can interpret a predictor's effect on the correlation via the regression parameters $\delta_{j,k,r}$. Specifically, if $\delta_{j,k,r} > 0$ ($\delta_{j,k,r} < 0$), then the correlation between the jth and kth phenotype increases (decreases) with the rth feature while other features remains unchanged.

Although any model is just an approximation of the truth, we can use the model to compute another interpretable measure, such as the average marginal effect (AME) [67, 68, 69]. In general, AME on the mean is defined as the change in the conditional mean of an outcome variable with respect to a single feature. Likewise, the AME of the rth feature on the (j,k)th pairwise correlation can be defined as the average change of the correlation for a change in the rth feature. Let us denote the $(p-1)$ component vector $(X_{i,1},\ldots,X_{i,r-1},X_{i,r+1},\ldots,X_{i,p})^T$ by $X_{i,(-r)}$. Then, for a binary feature X_r, the AME is defined as $AME_r = E\{\rho_{i,j,k}|X_{i,r}=1,X_{i,(-r)}\} - E\{\rho_{i,j,k}|X_{i,r}=0,X_{i,(-r)}\} = E\{\varphi_{r,(j,k)}(X_i,\theta)\}$, where θ denotes all the parameters and

$$\varphi_{r,(j,k)}(X_i,\theta) = 2\left\{\frac{1}{1+\exp(\delta_{j,k,0}+\sum_{s\neq r}^{p}\delta_{j,k,s}X_{i,s})} - \frac{1}{1+\exp(\delta_{j,k,0}+\delta_{j,k,r}+\sum_{s\neq r}^{p}\delta_{j,k,s}X_{i,s})}\right\}.$$

For a continuous feature X_r, $AME_r = E\{\varphi_{r,(j,k)}(X_i,\theta)\}$, where

$$\varphi_{r,(j,k)}(X_i,\theta) = \left(\frac{\partial\rho_{i,j,k}}{\partial X_{i,r}}\right) = 2\delta_{j,k,r}\frac{\exp(\delta_{j,k,0}+\sum_{s=1}^{p}\delta_{j,k,s}X_{i,s})}{\{1+\exp(\delta_{j,k,0}+\sum_{s=1}^{p}\delta_{j,k,s}X_{i,s})\}^2}.$$

Let $\widehat{\theta}$ be the estimator of θ and S denotes the estimated variance-covariance matrix of $\widehat{\theta}$. Then the estimator of AME_r is $\widehat{AME}_r = (1/n)\sum_{i=1}^{n}\varphi_{r(j,k)}(X_i,\widehat{\theta})$. Applying the delta method, we obtain the standard error of AME_r as

$$\sqrt{[\nabla_\theta\frac{\sum_{i=1}^{n}\varphi_{r(j,k)}(X_i,\theta)}{n}]^\top_{\theta=\widehat{\theta}}\,S\,[\nabla_\theta\frac{\sum_{i=1}^{n}\varphi_{r(j,k)}(X_i,\theta)}{n}]_{\theta=\widehat{\theta}}},$$

where $\nabla_\theta(\bullet) \equiv \partial(\bullet)/\partial\theta$.

4.2.3 Standard deviation modeling

The log-linear function is used to model the standard deviation of the phenotypes in terms on the features. Specifically, for the jth response and the ith experimental unit, the standard deviation is modeled as

$$\log(\sigma_{i,j}) = \alpha_{j,0} + \sum_{r=1}^{p}\alpha_{j,r}X_{i,r}. \tag{4.2}$$

The α parameters measure the effect of the features on the standard deviation. Like the correlation, AME can be used to measure the effect of the features.

4.3 Estimation methodology

4.3.1 Composite likelihood

As mentioned previously, in our pairwise modeling, there is no guaranty that the correlation matrix R_i is positive definite. Thus, the model parameters cannot be estimated by maximizing the multivariate normal density function. With this, we propose to estimate the model parameters via the pairwise-composite likelihood method. Now define $\theta = (\alpha^T,\ \delta^T)^T$, where $\alpha = (\alpha_1^T, \ldots, \alpha_q^T)^T$ and $\delta = (\delta_{1,2}^T, \delta_{1,3}^T, \ldots, \delta_{q-1,q}^T)^T$. The pairwise composite likelihood function for q responses is

$$CL(\theta) \ = \ \prod_{j=1}^{q-1} \prod_{k=j+1}^{q} \mathcal{L}_{j,k}(\theta),$$

where $\mathcal{L}_{j,k}(\theta) = \prod_{i=1}^{n} f(Y_{i,j}, Y_{i,k}|X_i)$ denotes the pairwise likelihood function for the jth and kth responses, and

$$f(Y_{i,j}, Y_{i,k}|X_i) = \frac{1}{2\pi\sigma_{i,j}\sigma_{i,k}\sqrt{1-\rho_{i,j,k}^2}}$$
$$\times \exp\left\{ -\frac{1}{2(1-\rho_{i,j,k}^2)}\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^2} - \frac{2\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^2}\right)\right\}. \tag{4.3}$$

The estimator of θ is defined as $\widehat{\theta} = \text{argmax}_\theta \ell(\theta)$, where $\ell(\theta) = \log\{CL(\theta)\}$. Note that the length of the θ-vector is $n_\theta = q \times (p+1) + \binom{q}{2} \times (p+1) = \frac{(p+1)q(q+1)}{2}$. For a scenario with two features ($p = 2$) and four phenotypes ($q = 4$), n_θ is 30. For the scenario of $p = 6$ and $q = 4$, n_θ is 70. Thus, applying the standard Newton-Raphson method or its variant is very time-consuming as it will require repeated inversion of a large matrix. Therefore, we develop an MM algorithm which is more computationally efficient than direct maximization of $\ell(\theta)$ using the Newton-Raphson method.

4.3.2 The MM algorithm

The MM algorithm squarely depends on finding a suitable minorization function for the log of the composite likelihood, $\ell(\theta)$. Note that $\ell(\theta) = \sum_{j<k} \ell_{j,k}(\alpha, \delta)$, where $\ell_{j,k}(\alpha, \delta)$ is the logarithm of

the pairwise likelihood function

$$\begin{aligned}\ell_{j,k}(\theta) &= -\frac{1}{2}\sum_{i=1}^{n}\Bigg\{\log(\sigma_{i,j}^2)+\log(\sigma_{i,k}^2)+\log(1-\rho_{i,j,k}^2)\\ &\quad+\frac{1}{(1-\rho_{i,j,k}^2)}\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^2}-\frac{2\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}}+\frac{Y_{i,k}^2}{\sigma_{i,k}^2}\right)\Bigg\}.\end{aligned}$$

Now, we state the main result based on which our analysis is based on.

Theorem 1. *For any θ and θ_0 in the parameter space,*

$$\ell^*(\theta|\theta^{(0)})=\sum_{j=1}^{p}g_1(\alpha_j|\theta^{(0)})+\sum_{j<k}\sum g_2(\delta_{j,k}|\theta^{(0)})+g_3(\theta^{(0)})$$

is a minorization function of $\ell(\theta)$ such that and $\ell(\theta)\geq\ell^*(\theta|\theta^{(0)})\ \forall\theta,\theta_0$ and $\ell(\theta)=\ell^*(\theta|\theta)$, where $g_1(\alpha_j|\theta^{(0)})=\sum_{s:s<j}\psi_{1,s,j}(\alpha_j-\alpha_j^{(0)},j|\theta^{(0)})+\sum_{s:s>j}\psi_{1,j,s}(\alpha_j-\alpha_j^{(0)},j|\theta^{(0)})$ for $j=1,\ldots,q$, $g_2(\delta_{j,k}|\theta^{(0)})=\psi_{2,j,k}(\rho_{j,k}|\theta^{(0)})$ for $j\neq k$, and $g_3(\theta^{(0)})=\sum\sum_{j<k}\psi_{3,j,k}(\theta^{(0)})$, with

$$\begin{aligned}\psi_{1,j,k}(\alpha_r-\alpha_r^{(0)},r|\theta^{(0)})=\sum_{i=1}^{n}\Bigg[&\left\{1+\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}+\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\right\}Z_i^T(\alpha_r^{(0)}-\alpha_r)\\ &-\frac{Y_{i,r}^2}{4\sigma_{i,r}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\exp\{4Z_i^T(\alpha_r^{(0)}-\alpha_r)\}\\ &-\left\{\frac{(Y_{i,j}^2+Y_{i,k}^2)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}+\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\right\}\exp\{3Z_i^T(\alpha_r^{(0)}-\alpha_r)\}\Bigg],\end{aligned}$$

$$\psi_{2,j,k}(\rho_{j,k}|\theta^{(0)}) = \sum_{i=1}^{n}\Bigg[-\frac{1}{2}\log(1-\rho_{i,j,k}^2) - \left\{\frac{Y_{i,j}^2}{4\sigma_{i,j}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})} + \frac{Y_{i,k}^2}{4\sigma_{i,k}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\right\}\times$$
$$\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)^2 + \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\log\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)$$
$$-\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)^3 - \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\left(\frac{1+\rho_{i,j,k}^{(0)}}{1+\rho_{i,j,k}}\right)^3$$
$$+\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\log\left(\frac{1+\rho_{i,j,k}^{(0)}}{1+\rho_{i,j,k}}\right)\Bigg],$$

and

$$\psi_{3,j,k}(\theta^{(0)}) = \sum_{i=1}^{n}\left\{\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})} + \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})} - \frac{1}{2}\log(\sigma_{i,j}^{(0)^2}\sigma_{i,k}^{(0)^2})\right\}.$$

Proof of Theorem 1: Conditional on the covariate X_i, $Y_i = (Y_{i,1},\ldots,Y_{i,q})^T$ follows a multivariate normal distribution with mean 0 and variance-covariance matrix Σ_i. As defined in Section 4.3.1, the pairwise likelihood for the (j,k)th response is $\mathcal{L}_{j,k}(\theta) = \prod_{i=1}^{n} f(Y_{i,j},Y_{i,k}|X_i)$, with $f(Y_{i,j},Y_{i,k}|X_i)$ is given in (4.3). The logarithm of $\mathcal{L}_{j,k}(\theta)$ is

$$\ell_{j,k}(\alpha,\delta) = -\frac{1}{2}\sum_{i=1}^{n}\Bigg[\log(\sigma_{i,j}^2)+\log(\sigma_{i,k}^2)+\log(1-\rho_{i,j,k}^2)$$
$$+\frac{1}{(1-\rho_{i,j,k}^2)}\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^2} - \frac{2\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^2}\right)\Bigg].$$

To derive the minorization function for $\ell_{j,k}$, we consider each term separately. Consider the following term

$$-\frac{Y_{i,j}^2}{\sigma_{i,j}^2(1-\rho_{i,j,k}^2)} = -\frac{Y_{i,j}^2}{\sigma_{i,j}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\times\frac{\sigma_{i,j}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}{\sigma_{i,j}^2(1-\rho_{i,j,k}^2)}$$
$$\geq -\frac{Y_{i,j}^2}{2\sigma_{i,j}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\left\{\left(\frac{\sigma_{i,j}^{(0)}}{\sigma_{i,j}}\right)^4 + \left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)^2\right\}.$$

The above inequality follows from the AM-GM inequality. Similarly, we have

$$-\frac{Y_{i,k}^2}{\sigma_{i,k}^2(1-\rho_{i,j,k}^2)} = -\frac{Y_{i,k}^2}{\sigma_{i,k}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})} \times \frac{\sigma_{i,k}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}{\sigma_{i,k}^2(1-\rho_{i,j,k}^2)}$$
$$\geq -\frac{Y_{i,k}^2}{2\sigma_{i,k}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\left\{\left(\frac{\sigma_{i,k}^{(0)}}{\sigma_{i,k}}\right)^4 + \left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)^2\right\}.$$

Next, after replacing $\rho_{i,j,k}$ by $1 - 2/\{1+\exp(\delta_{j,k,0} + \sum_{r=1}^p \delta_{j,k,r}X_{i,r})\}$, in the term $\rho_{i,j,k}Y_{i,j}Y_{i,k}/\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)$, we obtain

$$\begin{aligned}
\frac{\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)} &= \frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)} \\
&\quad - \frac{2Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)\{1+\exp(\delta_{j,k,0}+\sum_{r=1}^p \delta_{j,k,r}X_{i,r})\}} \\
&= \frac{\left(Y_{i,j}+Y_{i,k}\right)^2 - \left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)} \\
&\quad - \frac{\left(Y_{i,j}+Y_{i,k}\right)^2 - \left(Y_{i,j}^2+Y_{i,k}^2\right)}{\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)\{1+\exp(\delta_{j,k,0}+\sum_{r=1}^p \delta_{j,k,r}X_{i,r})\}} \qquad (4.4) \\
&= B_1 + B_2 + B_3 + B_4.
\end{aligned}$$

Now,

$$B_1 = \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)} \geq \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\left\{1+\log\left(\frac{\sigma_{i,j}^{(0)}}{\sigma_{i,j}}\right)+\left(\frac{\sigma_{i,k}^{(0)}}{\sigma_{i,k}}\right)+\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)\right\},$$

and this inequality follows due to the fact that for any generic $x > 0$, $x \geq \{1+\log(x)\}$ and equality holds when $x = 1$. Next, using the AM-GM inequality we have

$$B_2 = -\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}\sigma_{i,k}(1-\rho_{i,j,k}^2)} \geq -\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\left\{\left(\frac{\sigma_{i,j}^{(0)}}{\sigma_{i,j}}\right)^3+\left(\frac{\sigma_{i,k}^{(0)}}{\sigma_{i,k}}\right)^3+\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)^3\right\}.$$

After replacing $1+\exp(\delta_{j,k,0}+\sum_{r=1}^p \delta_{j,k,r}X_{i,r})$ by $2/\left(1-\rho_{i,j,k}\right)$ in (4.4), we have

$$B_3 + B_4 = -\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}\sigma_{i,k}(1+\rho_{i,j,k})} + \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}\sigma_{i,k}(1+\rho_{i,j,k})}.$$

Now,

$$B_3 = -\frac{\left(Y_{i,j} + Y_{i,k}\right)^2}{2\sigma_{i,j}\sigma_{i,k}(1+\rho_{i,j,k})}$$
$$\geq -\frac{\left(Y_{i,j} + Y_{i,k}\right)^2}{6\sigma^{(0)}_{i,j}\sigma^{(0)}_{i,k}(1+\rho^{(0)}_{i,j,k})}\left\{\left(\frac{\sigma^{(0)}_{i,j}}{\sigma_{i,j}}\right)^3 + \left(\frac{\sigma^{(0)}_{i,k}}{\sigma_{i,k}}\right)^3 + \left(\frac{1+\rho^{(0)}_{i,j,k}}{1+\rho_{i,j,k}}\right)^3\right\},$$

and

$$B_4 = \frac{\left(Y^2_{i,j} + Y^2_{i,k}\right)}{2\sigma_{i,j}\sigma_{i,k}(1+\rho_{i,j,k})}$$
$$\geq \frac{\left(Y^2_{i,j} + Y^2_{i,k}\right)}{2\sigma^{(0)}_{i,j}\sigma^{(0)}_{i,k}(1+\rho^{(0)}_{i,j,k})}\left\{1 + \log\left(\frac{\sigma^{(0)}_{i,j}}{\sigma_{i,j}}\right) + \log\left(\frac{\sigma^{(0)}_{i,k}}{\sigma_{i,k}}\right) + \log\left(\frac{1+\rho^{(0)}_{i,j,k}}{1+\rho_{i,j,k}}\right)\right\},$$

and these two inequalities follow from the AM-GM inequality and $x \geq 1+\log(x)$ for any generic $x > 0$.

We further define

$$
\begin{aligned}
&\psi_{1,j,k}(\alpha_r - \alpha_r^{(0)}, r|\theta^{(0)}) \\
= &\sum_{i=1}^{n}\Bigg\{\log\left(\frac{\sigma_{i,r}^{(0)}}{\sigma_{i,r}}\right) - \frac{Y_{i,r}^2}{4\sigma_{i,r}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\left(\frac{\sigma_{i,r}^{(0)}}{\sigma_{i,r}}\right)^4 + \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\log\left(\frac{\sigma_{i,r}^{(0)}}{\sigma_{i,r}}\right) \\
&-\frac{(Y_{i,j}^2+Y_{i,k}^2)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\left(\frac{\sigma_{i,r}^{(0)}}{\sigma_{i,r}}\right)^3 - \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\left(\frac{\sigma_{i,r}^{(0)}}{\sigma_{i,r}}\right)^3 \\
&+\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\log\left(\frac{\sigma_{i,r}^{(0)}}{\sigma_{i,r}}\right)\Bigg\}, \\
= &\sum_{i=1}^{n}\Bigg[Z_i^T(\alpha_r^{(0)}-\alpha_r) - \frac{Y_{i,r}^2}{4\sigma_{i,r}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\exp\{4Z_i^T(\alpha_r^{(0)}-\alpha_r)\} \\
&+\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}Z_i^T(\alpha_r^{(0)}-\alpha_r) \\
&-\left\{\frac{(Y_{i,j}^2+Y_{i,k}^2)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})} + \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\right\}\exp\{3Z_i^T(\alpha_r^{(0)}-\alpha_r)\} \\
&+\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}Z_i^T(\alpha_r^{(0)}-\alpha_r)\Bigg] \\
= &\sum_{i=1}^{n}\Bigg[\left\{1+\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})} + \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\right\}Z_i^T(\alpha_r^{(0)}-\alpha_r) \\
&-\frac{Y_{i,r}^2}{4\sigma_{i,r}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\exp\{4Z_i^T(\alpha_r^{(0)}-\alpha_r)\} \\
&-\left\{\frac{(Y_{i,j}^2+Y_{i,k}^2)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})} + \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\right\}\exp\{3Z_i^T(\alpha_r^{(0)}-\alpha_r)\}\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
&\psi_{2,j,k}(\rho_{j,k}|\theta^{(0)}) \\
&= \sum_{i=1}^{n}\Bigg[-\frac{1}{2}\log(1-\rho_{i,j,k}^2) - \left\{\frac{Y_{i,j}^2}{4\sigma_{i,j}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})} + \frac{Y_{i,k}^2}{4\sigma_{i,k}^{(0)^2}(1-\rho_{i,j,k}^{(0)^2})}\right\}\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)^2 \\
&\quad + \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\log\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right) \\
&\quad - \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\left(\frac{1-\rho_{i,j,k}^{(0)^2}}{1-\rho_{i,j,k}^2}\right)^3 - \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\left(\frac{1+\rho_{i,j,k}^{(0)}}{1+\rho_{i,j,k}}\right)^3 \\
&\quad + \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\log\left(\frac{1+\rho_{i,j,k}^{(0)}}{1+\rho_{i,j,k}}\right)\Bigg].
\end{aligned}
$$

Since,

$$
\begin{aligned}
\rho_{i,j,k} &= 1 - 2/\{1+\exp(\delta_{j,k}^T Z_i)\}, \\
1+\rho_{i,j,k} &= 2\exp(\delta_{j,k}^T Z_i)/\{1+\exp(\delta_{j,k}^T Z_i)\}, \\
\rho_{i,j,k}^2 &= \{\exp(\delta_{j,k}^T Z_i)-1\}^2/\{\exp(\delta_{j,k}^T Z_i)+1\}^2, \\
1-\rho_{i,j,k}^2 &= 4\exp(\delta_{j,k}^T Z_i)/\{\exp(\delta_{j,k}^T Z_i)+1\}^2.
\end{aligned}
$$

Now using these terms, we obtain

$$
\begin{aligned}
&\psi_{2,j,k}(\rho_{j,k}|\theta^{(0)}) \\
&= \sum_{i=1}^{n}\Bigg(-\frac{\log(4)}{2} - 0.5\delta_{j,k}^T Z_i + \log\{\exp(\delta_{j,k}^T Z_i)+1\} \\
&\quad -\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(0)^2}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^{(0)^2}}\right)\left(1-\rho_{i,j,k}^{(0)^2}\right) \times \frac{\{\exp(\delta_{j,k}^T Z_i)+1\}^4}{64\exp(2\delta_{j,k}^T Z_i)} \\
&\quad + \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})}\left[\log\left(1-\rho_{i,j,k}^{(0)^2}\right) - \log(4) - \delta_{j,k}^T X_i + 2\log\{\exp(\delta_{j,k}^T Z_i)+1\}\right] \\
&\quad - \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}}\left(1-\rho_{i,j,k}^{(0)^2}\right)^2 \times \frac{\{1+\exp(\delta_{j,k}^T Z_i)\}^6}{64\exp(3\delta_{j,k}^T Z_i)} \\
&\quad - \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}}\left(1+\rho_{i,j,k}^{(0)}\right)^2 \times \frac{\{1+\exp(\delta_{j,k}^T Z_i)\}^3}{8\exp(3\delta_{j,k}^T Z_i)} \\
&\quad + \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})}\left[\log\left(1+\rho_{i,j,k}^{(0)}\right) - \log(2) - \delta_{j,k}^T Z_i + \log\{1+\exp(\delta_{j,k}^T Z_i)\}\right]\Bigg).
\end{aligned}
$$

Also,

$$
\psi_{3,j,k}(\theta^{(0)}) = \sum_{i=1}^{n}\left\{\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1-\rho_{i,j,k}^{(0)^2})} + \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,k}^{(0)}(1+\rho_{i,j,k}^{(0)})} - \frac{1}{2}\log(\sigma_{i,j}^{(0)^2}\sigma_{i,k}^{(0)^2})\right\}.
$$

Now, the minorization of the composite log-likelihood is

$$
\begin{aligned}
\ell^*(\theta|\theta^{(0)}) &= \sum_{j=1}^{(q-1)}\sum_{k=(j+1)}^{q}\Big\{\psi_{1,j,k}(\alpha_j-\alpha_j^{(0)}, j|\theta^{(0)}) + \psi_{1,j,k}(\alpha_k-\alpha_k^{(0)}, k|\theta^{(0)}) \\
&\qquad + \psi_{2,j,k}(\rho_{j,k}|\theta^{(0)}) + \psi_{3,j,k}(\theta^{(0)})\Big\} \\
&= \sum_{j=1}^{q} g_1(\alpha_j|\theta^{(0)}) + \sum_{j<k}\sum g_2(\delta_{j,k}|\theta^{(0)}) + g_3(\theta^{(0)}),
\end{aligned}
$$

where $g_1(\alpha_j|\theta^{(0)}) = \sum_{s:s<j}\psi_{1,s,j}(\alpha_j-\alpha_j^{(0)}, j|\theta^{(0)}) + \sum_{s:s>j}\psi_{1,j,s}(\alpha_j-\alpha_j^{(0)}, j|\theta^{(0)})$ for $j = 1,\ldots,q$, $g_2(\delta_{j,k}|\theta^{(0)}) = \psi_{2,j,k}(\rho_{j,k}|\theta^{(0)})$ for $j \neq k$, $g_3(\theta^{(0)}) = \sum\sum_{j<k}\psi_{3,j,k}(\theta^{(0)})$.

In the MM algorithm, we maximize the minorizing function ℓ^* rather than ℓ. The minorizing function ℓ^* is expressed as a summation of $g_1(\alpha_1|\theta^{(0)}), \ldots, g_1(\alpha_p|\theta^{(0)})$, and $g_2(\delta_{1,2}|\theta^{(0)}), \ldots, g_2(\delta_{p-1,p}|\theta^{(0)})$, this results in the separation of the parameters. Separation of the parameter has a great advantage when optimizing a function with respect to a high-dimensional argument (θ in our case). The functions g_1, g_2 are all differentiable functions, and satisfy standard regularity conditions, and these function are used in updating the parameter values. The parameter estimates are obtained by the gradient MM algorithm [63]. Let $\theta^{(t)}$ be the parameter value at the tth iteration, then $\theta^{(t+1)}$ is obtained by one step Newton-Raphson method

$$\theta^{(t+1)} = \theta^{(t)} - \left\{ \frac{\partial^2}{\partial\theta\partial\theta^T} \ell^*(\theta|\theta^{(t)}) \right\}^{-1}_{\theta=\theta^{(t)}} \left\{ \frac{\partial}{\partial\theta} \ell^*(\theta|\theta^{(t)}) \right\}_{\theta=\theta^{(t)}}. \tag{4.5}$$

The above step is repeated until the estimate converges. Specifically, we stop the above iteration when $1^T(|\theta^{(t+1)} - \theta^{(t)}|/|\theta^{(t)}|) < \epsilon_0$, a prespecified small number.

Observe that in Equation (4.5), rather than the log-composite likelihood $\ell(\theta)$, the minorization function $\ell^*(\theta|\theta^{(t)})$ is used. Next, $\partial\ell^*(\theta|\theta^{(t)})/\partial\alpha_j = \partial g_1(\alpha_j|\theta^{(t)})/\partial\alpha_j$, a function of α_j only, and $\partial\ell^*(\theta|\theta^{(t)})/\partial\delta_{j,k} = \partial g_2(\delta_{j,k}|\theta^{(t)})/\partial\delta_{j,k}$, a function of $\delta_{j,k}$ only. Consequently $\partial^2\ell^*(\theta|\theta^{(t)})/\partial\theta\partial\theta^T$ is a block-diagonal matrix, and each block is a matrix of order $(p+1)\times(p+1)$ and this greatly enhances computational efficiency.

Specifically, the complexity of the inversion of each block matrix is in the order $O((p+1)^3)$. Thus, the complexity of one update of the MM algorithm is $O(nn_\theta + n(p+1)^2q(q+1)/2 + (p+1)^3q(q+1)/2)$. In other words, the complexity is $O(nn_\theta + n(p+1)n_\theta + (p+1)^2n_\theta)$, where $n_\theta = (1+p)q(1+q)/2$. On the other hand, the complexity of a direct optimization of $\ell(\theta)$ using the Newton-Raphson method is $O(nn_\theta + nn_\theta^2 + n_\theta^3)$. Alternatively, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm can be used to optimize $\ell(\theta)$. This algorithm avoids large matrix inversion, so it has a square order complexity. Although the proportionality constant of the complexity order is unknown, the order of this complexity for BFGS is still larger than the complexity of the MM algorithm as long as $(p+1) < q(q+1)/2$, and this holds for our real data and the simulation scenarios.

The terms of Equation (4.5) are $[\partial\ell^*(\theta|\theta^{(t)})/\partial\theta]_{\theta=\theta^{(t)}} = [\partial\ell^*(\theta|\theta^{(t)})/\partial\alpha_1, \ldots,$ $\partial\ell^*(\theta|\theta^{(t)})/\partial\alpha_q, \partial\ell^*(\theta|\theta^{(t)})/\partial\delta_{1,2}, \ldots, \partial\ell^*(\theta|\theta^{(t)})/\partial\delta_{q-1,q}]_{\theta=\theta^{(t)}}$, and

$$\begin{aligned}
\left(\frac{\partial\ell^*(\theta|\theta^{(t)})}{\partial\alpha_j}\right)_{\theta=\theta^{(t)}} &= \sum_{s:s<j}\sum_{i=1}^{n}\left\{-1+\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,s,j}^{(t)^2})}-\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho^{(t)}{}_{i,s,j}^2)}\right.\\
&\quad\left.+\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,s,j}^{(t)})}\right\}Z_i+\sum_{s:s>j}\sum_{i=1}^{n}\left\{-1+\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})}\right.\\
&\quad\left.-\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho^{(t)}{}_{i,j,s}^2)}+\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,j,s}^{(t)})}\right\}Z_i,\ \text{for } j=1,\ldots,q,\\
\left(\frac{\partial\ell^*(\theta|\theta^{(t)})}{\partial\delta_{j,k}}\right)_{\theta=\theta^{(t)}} &= \sum_{i=1}^{n}\left\{\frac{\rho_{i,j,k}^{(t)}}{1-\rho_{i,j,k}^{(t)^2}}-\frac{Y_{i,j}^2\rho_{i,j,k}^{(t)}}{\sigma^{(t)}{}_{i,j}^2(1-\rho_{i,j,k}^{(t)^2})^2}-\frac{Y_{i,k}^2\rho_{i,j,k}^{(t)}}{\sigma^{(t)}{}_{i,k}^2(1-\rho_{i,j,k}^{(t)^2})^2}\right.\\
&\quad\left.+\frac{2Y_{i,j}Y_{i,k}\rho_{i,j,k}^{(t)}}{\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})^2}+\frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1+\rho_{i,j,k}^{(t)})^2}\right\}\frac{(1-\rho_{i,j,k}^{(t)^2})}{2}Z_i,
\end{aligned}$$

for $j < k = 1, \ldots, q$ and $Z_i = (1\ X_i^T)^T$. Furthermore, let $A = \{\partial^2\ell^*(\theta|\theta^{(t)})/\partial\theta\partial\theta^T\}_{\theta=\theta^{(t)}}$, then $A = \text{Diag}(A_1, \ldots, A_q, A_{1,2}^\dagger, \ldots, A_{q-1,q}^\dagger)$, where

$$\begin{aligned}
A_j &= \left(\frac{\partial^2\ell^*(\theta|\theta^{(t)})}{\partial\alpha_j\partial\alpha_j^T}\right)_{\theta=\theta^{(t)}}\\
&= -\sum_{i=1}^{n}\left(\sum_{s:s<j}\left[\frac{4Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,s,j}^{(t)^2})}+\frac{3}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}}\left\{\frac{(Y_{i,j}^2+Y_{i,s}^2)}{(1-\rho_{i,s,j}^{(t)^2})}+\frac{(Y_{i,j}+Y_{i,s})^2}{(1+\rho_{i,s,j}^{(t)})}\right\}\right]\right.\\
&\quad\left.+\sum_{s:s<j}\left[\frac{4Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})}+\frac{3}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}}\left\{\frac{(Y_{i,j}^2+Y_{i,s}^2)}{(1-\rho_{i,j,s}^{(t)^2})}+\frac{(Y_{i,j}+Y_{i,s})^2}{(1+\rho_{i,j,s}^{(t)})}\right\}\right]\right)Z_iZ_i^T,
\end{aligned}$$

$$
\begin{aligned}
A^{\dagger}_{j,k} &= \left(\frac{\partial^2 \ell^*(\theta|\theta^{(t)})}{\partial \delta_{j,k} \partial \delta^T_{j,k}} \right)_{\theta=\theta^{(t)}} \\
&= \sum_{i=1}^{n} \Bigg[\frac{1+\rho^{(t)^2}_{i,j,k}}{(1-\rho^{(t)^2}_{i,j,k})^2} - \left(\frac{Y^2_{i,j}}{\sigma^{(t)^2}_{i,j}} + \frac{Y^2_{i,k}}{\sigma^{(t)^2}_{i,k}} \right) \times \frac{1+5\rho^{(t)^2}_{i,j,k}}{(1-\rho^{(t)^2}_{i,j,k})^3} \\
&\quad + \frac{1}{\sigma^{(t)}_{i,j}\sigma^{(t)}_{i,k}(1-\rho^{(t)^2}_{i,j,k})^3} \left\{ (Y_{i,j}+Y_{i,k})^2(1+\rho^{(t)^2}_{i,j,k}) - (Y^2_{i,j}+Y^2_{i,k})(1+7\rho^{(t)^2}_{i,j,k}) \right\} \\
&\quad + \frac{Y^2_{i,j}+Y^2_{i,k}-4(Y_{i,j}+Y_{i,k})^2}{2\sigma^{(t)}_{i,j}\sigma^{(t)}_{i,k}(1+\rho^{(t)}_{i,j,k})^3} \Bigg] \frac{(1-\rho^{(t)^2}_{i,j,k})^2}{4} Z_i Z_i^T \\
&\quad + \sum_{i=1}^{n} \Bigg[\frac{\rho^{(t)}_{i,j,k}}{1-\rho^{(t)^2}_{i,j,k}} - \left(\frac{Y^2_{i,j}}{\sigma^{(t)^2}_{i,j}} + \frac{Y^2_{i,k}}{\sigma^{(t)^2}_{i,k}} \right) \frac{\rho^{(t)}_{i,j,k}}{(1-\rho^{(t)^2}_{i,j,k})^2} \\
&\quad - \frac{Y_{i,j}Y_{i,k}}{\sigma^{(t)}_{i,j}\sigma^{(t)}_{i,k}} \left\{ \frac{2\rho^{(t)}_{i,j,k}}{(1-\rho^{(t)^2}_{i,j,k})^2} + \frac{1}{(1+\rho^{(t)}_{i,j,k})^2} \right\} \Bigg] \frac{\rho^{(t)}_{i,j,k}(1-\rho^{(t)^2}_{i,j,k})}{2} Z_i Z_i^T .
\end{aligned}
$$

The above expressions are derived in the following manner.

Observe that $\partial \ell^*(\theta|\theta^{(t)})/\partial \alpha_j = \partial g_1(\alpha_j|\theta^{(t)})/\partial \alpha_j = \sum_{s:s<j} \partial \psi_{1,s,j}(\alpha_j - \alpha^{(t)}_j, j|\theta^{(t)})/\partial \alpha_j + \sum_{s:s>j} \partial \psi_{1,j,s}(\alpha_j - \alpha^{(t)}_j, j|\theta^{(t)})/\partial \alpha_j$ for $j = 1, \ldots, q$. Now,

$$
\begin{aligned}
\psi_{1,j,s}(\alpha_j - \alpha^{(0)}_j, j|\theta^{(0)}) = \sum_{i=1}^{n} \Bigg[& \left\{ 1 + \frac{\left(Y_{i,j}+Y_{i,s} \right)^2}{2\sigma^{(0)}_{i,j}\sigma^{(0)}_{i,s}(1-\rho^{(0)^2}_{i,j,s})} + \frac{\left(Y^2_{i,j}+Y^2_{i,s} \right)}{2\sigma^{(0)}_{i,j}\sigma^{(0)}_{i,s}(1+\rho^{(0)}_{i,j,s})} \right\} Z_i^T(\alpha^{(0)}_j - \alpha_j) \\
& - \frac{Y^2_{i,j}}{4\sigma^{(0)^2}_{i,j}(1-\rho^{(0)^2}_{i,j,s})} \exp\{4Z_i^T(\alpha^{(0)}_j - \alpha_j)\} \\
& - \left\{ \frac{(Y^2_{i,j}+Y^2_{i,s})}{6\sigma^{(0)}_{i,j}\sigma^{(0)}_{i,s}(1-\rho^{(0)^2}_{i,j,s})} + \frac{\left(Y_{i,j}+Y_{i,s} \right)^2}{6\sigma^{(0)}_{i,j}\sigma^{(0)}_{i,s}(1+\rho^{(0)}_{i,j,s})} \right\} \exp\{3Z_i^T(\alpha^{(0)}_j - \alpha_j)\} \Bigg],
\end{aligned}
$$

so

$$\frac{\partial \psi_{1,j,s}(\alpha_j - \alpha_j^{(0)}, j|\theta^{(0)})}{\partial \alpha_j} = \sum_{i=1}^{n}\Bigg[-\Bigg\{1 + \frac{\left(Y_{i,j} + Y_{i,s}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,s}^{(0)}(1-\rho_{i,j,s}^{(0)^2})} + \frac{\left(Y_{i,j}^2 + Y_{i,s}^2\right)}{2\sigma_{i,j}^{(0)}\sigma_{i,s}^{(0)}(1+\rho_{i,j,s}^{(0)})}\Bigg\}$$
$$+ \frac{Y_{i,j}^2}{\sigma_{i,j}^{(0)^2}(1-\rho_{i,j,s}^{(0)^2})}\exp\{4Z_i^T(\alpha_j^{(0)} - \alpha_j)\}$$
$$+ \Bigg\{\frac{(Y_{i,j}^2 + Y_{i,s}^2)}{2\sigma_{i,j}^{(0)}\sigma_{i,s}^{(0)}(1-\rho_{i,j,s}^{(0)^2})} + \frac{\left(Y_{i,j} + Y_{i,s}\right)^2}{2\sigma_{i,j}^{(0)}\sigma_{i,s}^{(0)}(1+\rho_{i,j,s}^{(0)})}\Bigg\}\times$$
$$\exp\{3Z_i^T(\alpha_j^{(0)} - \alpha_j)\}\Bigg]Z_i,$$

and

$$\left(\frac{\partial \psi_{1,j,s}(\alpha_j - \alpha_j^{(t)}, j|\theta^{(t)})}{\partial \alpha_j}\right)_{\theta=\theta^{(t)}} = \sum_{i=1}^{n}\Bigg[-\Bigg\{1 + \frac{\left(Y_{i,j} + Y_{i,s}\right)^2}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho_{i,j,s}^{(t)^2})} + \frac{\left(Y_{i,j}^2 + Y_{i,s}^2\right)}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,j,s}^{(t)})}\Bigg\}$$
$$+ \frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})}$$
$$+ \Bigg\{\frac{(Y_{i,j}^2 + Y_{i,s}^2)}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho_{i,j,s}^{(t)^2})} + \frac{\left(Y_{i,j} + Y_{i,s}\right)^2}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,j,s}^{(t)})}\Bigg\}\Bigg]Z_i,$$
$$= \sum_{i=1}^{n}\Bigg\{-1 + \frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})} - \frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho^{(t)^2}_{i,j,s})}$$
$$+ \frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,j,s}^{(t)})}\Bigg\}Z_i.$$

Likewise,

$$\left(\frac{\partial \psi_{1,s,j}(\alpha_j - \alpha_j^{(t)}, j|\theta^{(t)})}{\partial \alpha_j}\right)_{\theta=\theta^{(t)}} = \sum_{i=1}^{n}\Bigg\{-1 + \frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,s,j}^{(t)^2})} - \frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho^{(t)^2}_{i,s,j})}$$
$$+ \frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,s,j}^{(t)})}\Bigg\}Z_i.$$

Adding the above two expressions, we obtain

$$\begin{aligned}\left(\frac{\partial \ell^*(\theta|\theta^{(t)})}{\partial \alpha_j}\right)_{\theta=\theta^{(t)}} &= \sum_{s:s<j}\sum_{i=1}^{n}\Bigg\{-1+\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,s,j}^{(t)^2})}-\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho^{(t)}{}_{i,s,j}^{2})}\\ &+\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,s,j}^{(t)})}\Bigg\}Z_i+\sum_{s:s>j}\sum_{i=1}^{n}\Bigg\{-1+\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})}\\ &-\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1-\rho^{(t)}{}_{i,j,s}^{2})}+\frac{Y_{i,j}Y_{i,s}}{\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}(1+\rho_{i,j,s}^{(t)})}\Bigg\}Z_i.\end{aligned}$$

Next consider,

$$\begin{aligned}\frac{\partial^2\psi_{1,j,s}(\alpha_j-\alpha_j^{(t)},j|\theta^{(t)})}{\partial\alpha_j\partial\alpha_j^T} &= -\sum_{i=1}^{n}\Bigg[\frac{4Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})}\exp\{4Z_i^T(\alpha_j^{(t)}-\alpha_j)\}\\ &+\frac{3}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}}\left\{\frac{(Y_{i,j}^2+Y_{i,s}^2)}{(1-\rho_{i,j,s}^{(t)^2})}+\frac{(Y_{i,j}+Y_{i,s})^2}{(1+\rho_{i,j,s}^{(t)})}\right\}\exp\{3Z_i^T(\alpha_j^{(t)}-\alpha_j)\}\Bigg]Z_iZ_i^T,\end{aligned}$$

and

$$\begin{aligned}\left(\frac{\partial^2\psi_{1,j,s}(\alpha_j-\alpha_j^{(t)},j|\theta^{(t)})}{\partial\alpha_j\partial\alpha_j^T}\right)_{\theta=\theta^{(t)}} &= -\sum_{i=1}^{n}\Bigg[\frac{4Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})}+\frac{3}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}}\times\\ &\left\{\frac{(Y_{i,j}^2+Y_{i,s}^2)}{(1-\rho_{i,j,s}^{(t)^2})}+\frac{(Y_{i,j}+Y_{i,s})^2}{(1+\rho_{i,j,s}^{(t)})}\right\}\Bigg]Z_iZ_i^T.\end{aligned}$$

Similarly,

$$\begin{aligned}\left(\frac{\partial^2\psi_{1,s,j}(\alpha_j-\alpha_j^{(t)},j|\theta^{(t)})}{\partial\alpha_j\partial\alpha_j^T}\right)_{\theta=\theta^{(t)}} &= -\sum_{i=1}^{n}\Bigg[\frac{4Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,s,j}^{(t)^2})}+\frac{3}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}}\times\\ &\left\{\frac{(Y_{i,j}^2+Y_{i,s}^2)}{(1-\rho_{i,s,j}^{(t)^2})}+\frac{(Y_{i,j}+Y_{i,s})^2}{(1+\rho_{i,s,j}^{(t)})}\right\}\Bigg]Z_iZ_i^T.\end{aligned}$$

Combining the above two expressions, we obtain

$$\begin{aligned}\left(\frac{\partial^2\ell^*(\theta|\theta^{(t)})}{\partial\alpha_j\partial\alpha_j^T}\right)_{\theta=\theta^{(t)}} &= -\sum_{s:s<j}\sum_{i=1}^{n}\Bigg[\frac{4Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,s,j}^{(t)^2})}+\frac{3}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}}\left\{\frac{(Y_{i,j}^2+Y_{i,s}^2)}{(1-\rho_{i,s,j}^{(t)^2})}\right.\\ &+\left.\frac{(Y_{i,j}+Y_{i,s})^2}{(1+\rho_{i,s,j}^{(t)})}\right\}\Bigg]Z_iZ_i^T-\sum_{s:s<j}\sum_{i=1}^{n}\Bigg[\frac{4Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}(1-\rho_{i,j,s}^{(t)^2})}\\ &+\frac{3}{2\sigma_{i,j}^{(t)}\sigma_{i,s}^{(t)}}\left\{\frac{(Y_{i,j}^2+Y_{i,s}^2)}{(1-\rho_{i,j,s}^{(t)^2})}+\frac{(Y_{i,j}+Y_{i,s})^2}{(1+\rho_{i,j,s}^{(t)})}\right\}\Bigg]Z_iZ_i^T.\end{aligned}$$

Next, observe that $\partial \ell^*(\theta|\theta^{(t)})/\partial \delta_{j,k} = \partial g_2(\delta_{j,k}|\theta^{(t)})/\partial \delta_{j,k} = \partial \psi_{2,j,k}(\delta_{j,k}|\theta^{(t)})/\partial \delta_{j,k}$.

Recall that,

$$
\begin{aligned}
\psi_{2,j,k}(\rho_{j,k}|\theta^{(t)}) = \sum_{i=1}^{n} \Bigg(& -\frac{\log(4)}{2} - 0.5\delta_{j,k}^T Z_i + \log\{\exp(\delta_{j,k}^T Z_i) + 1\} \\
& - \left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^{(t)^2}} \right) \left(1 - \rho_{i,j,k}^{(t)^2}\right) \times \frac{\{\exp(\delta_{j,k}^T Z_i) + 1\}^4}{64 \exp(2\delta_{j,k}^T Z_i)} \\
& + \frac{\left(Y_{i,j} + Y_{i,k}\right)^2}{2\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)} (1 - \rho_{i,j,k}^{(t)^2})} \Bigg[\log\left(1 - \rho_{i,j,k}^{(t)^2}\right) - \log(4) \\
& - \delta_{j,k}^T Z_i + 2\log\{\exp(\delta_{j,k}^T Z_i) + 1\} \Bigg] \\
& - \frac{\left(Y_{i,j}^2 + Y_{i,k}^2\right)}{6\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)}} \left(1 - \rho_{i,j,k}^{(t)^2}\right)^2 \times \frac{\{1 + \exp(\delta_{j,k}^T Z_i)\}^6}{64 \exp(3\delta_{j,k}^T Z_i)} \\
& - \frac{\left(Y_{i,j} + Y_{i,k}\right)^2}{6\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)}} \left(1 + \rho_{i,j,k}^{(t)}\right)^2 \times \frac{\{1 + \exp(\delta_{j,k}^T Z_i)\}^3}{8 \exp(3\delta_{j,k}^T Z_i)} \\
& + \frac{\left(Y_{i,j}^2 + Y_{i,k}^2\right)}{2\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)} (1 + \rho_{i,j,k}^{(t)})} \Bigg[\log\left(1 + \rho_{i,j,k}^{(t)}\right) - \log(2) \\
& - \delta_{j,k}^T Z_i + \log\{1 + \exp(\delta_{j,k}^T Z_i)\} \Bigg] \Bigg)
\end{aligned}
$$

Further simplifying, we get

$$
\begin{aligned}
\psi_{2,j,k}(\rho_{j,k}|\theta^{(t)}) = \sum_{i=1}^{n}\Bigg(&-\frac{\log(4)}{2} - 0.5\delta_{j,k}^{T}Z_i + \log\{\exp(\delta_{j,k}^{T}Z_i)+1\} \\
&- \left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^{(t)^2}}\right)\left(1-\rho_{i,j,k}^{(t)^2}\right) \times \frac{1}{64}\left\{\exp(\delta_{j,k}^{T}Z_i/2) + \exp(-\delta_{j,k}^{T}Z_i/2)\right\}^4 \\
&+ \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})}\Bigg[\log\left(1-\rho_{i,j,k}^{(t)^2}\right) - \log(4) \\
&- \delta_{j,k}^{T}Z_i + 2\log\{\exp(\delta_{j,k}^{T}Z_i)+1\}\Bigg] \\
&- \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1-\rho_{i,j,k}^{(t)}\right)^2 \times \frac{1}{64}\left\{\exp(\delta_{j,k}^{T}Z_i/2) + \exp(-\delta_{j,k}^{T}Z_i/2)\right\}^6 \\
&- \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1+\rho_{i,j,k}^{(t)}\right)^2 \times \frac{1}{8}\left\{1+\exp(-\delta_{j,k}^{T}Z_i)\right\}^3 \\
&+ \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1+\rho_{i,j,k}^{(t)})}\Bigg[\log\left(1+\rho_{i,j,k}^{(0)}\right) - \log(2) \\
&- \delta_{j,k}^{T}Z_i + \log\{1+\exp(\delta_{j,k}^{T}Z_i)\}\Bigg]\Bigg).
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\frac{\partial \psi_{2,j,k}(\rho_{j,k}|\theta^{(t)})}{\partial \delta_{j,k}} \\
&= \sum_{i=1}^{n}\Bigg[-0.5 + \frac{\exp(\delta_{j,k}^T Z_i)}{1+\exp(\delta_{j,k}^T Z_i)} \\
&- \left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^{(t)^2}}\right)\left(1-\rho_{i,j,k}^{(t)^2}\right) \times \frac{1}{32}\left\{\exp(\delta_{j,k}^T Z_i/2) + \exp(-\delta_{j,k}^T Z_i/2)\right\}^3 \\
&\times \left\{\exp(\delta_{j,k}^T Z_i/2) - \exp(-\delta_{j,k}^T Z_i/2)\right\} \\
&+ \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})}\left\{-1 + \frac{2\exp(\delta_{j,k}^T Z_i)}{1+\exp(\delta_{j,k}^T Z_i)}\right\} \\
&- \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1-\rho_{i,j,k}^{(t)^2}\right)^2 \times \frac{3}{64}\left\{\exp(\delta_{j,k}^T Z_i/2) + \exp(-\delta_{j,k}^T Z_i/2)\right\}^5 \\
&\times \left\{\exp(\delta_{j,k}^T Z_i/2) - \exp(-\delta_{j,k}^T Z_i/2)\right\} \\
&- \frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1+\rho_{i,j,k}^{(t)}\right)^2 \times \frac{3}{8}\left\{1+\exp(-\delta_{j,k}^T Z_i)\right\}^2\left\{1-\exp(-\delta_{j,k}^T Z_i)\right\} \\
&+ \frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1+\rho_{i,j,k}^{(t)})}\left\{-1 + \frac{\exp(\delta_{j,k}^T Z_i)}{1+\exp(\delta_{j,k}^T Z_i)}\right\}\Bigg] Z_i,
\end{aligned}
$$

and

$$\begin{aligned}\left(\frac{\partial \psi_{2,j,k}(\rho_{j,k}|\theta^{(t)})}{\partial \delta_{j,k}}\right)_{\theta=\theta^{(t)}} &= \sum_{i=1}^{n}\Bigg[-0.5+\frac{\exp(Z_i^T\delta_{j,k}^{(t)})}{1+\exp(Z_i^T\delta_{j,k}^{(t)})} \\ &\quad -\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}}+\frac{Y_{i,k}^2}{\sigma_{i,k}^{(t)^2}}\right)\left(1-\rho_{i,j,k}^{(t)^2}\right)\times\frac{1}{32}\Bigg\{\exp(Z_i^T\delta_{j,k}^{(t)}/2) \\ &\quad +\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\Bigg\}^3\Bigg\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)-\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\Bigg\} \\ &\quad +\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})}\left\{-1+\frac{2\exp(Z_i^T\delta_{j,k}^{(t)})}{1+\exp(Z_i^T\delta_{j,k}^{(t)})}\right\} \\ &\quad -\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1-\rho_{i,j,k}^{(t)^2}\right)^2 \\ &\quad \times\frac{3}{64}\Bigg\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)+\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\Bigg\}^5 \\ &\quad \times\Bigg\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)-\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\Bigg\} \\ &\quad -\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1+\rho_{i,j,k}^{(t)}\right)^2 \\ &\quad \times\frac{3}{8}\Bigg\{1+\exp(-Z_i^T\delta_{j,k}^{(t)})\Bigg\}^2\Bigg\{1-\exp(-Z_i^T\delta_{j,k}^{(t)})\Bigg\} \\ &\quad +\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1+\rho_{i,j,k}^{(t)})}\left\{-1+\frac{\exp(Z_i^T\delta_{j,k}^{(t)})}{1+\exp(Z_i^T\delta_{j,k}^{(t)})}\right\}\Bigg]Z_i.\end{aligned}$$

Simplifying further, we obtain

$$\begin{aligned}\left(\frac{\partial \ell^*(\theta|\theta^{(t)})}{\partial \delta_{j,k}}\right)_{\theta=\theta^{(t)}} &= \sum_{i=1}^{n}\Bigg\{\frac{\rho_{i,j,k}^{(t)}}{1-\rho_{i,j,k}^{(t)^2}}-\frac{Y_{i,j}^2\rho_{i,j,k}^{(t)}}{\sigma^{(t)}{}_{i,j}^2(1-\rho_{i,j,k}^{(t)^2})^2}-\frac{Y_{i,k}^2\rho_{i,j,k}^{(t)}}{\sigma^{(t)}{}_{i,k}^2(1-\rho_{i,j,k}^{(t)^2})^2} \\ &\quad +\frac{2Y_{i,j}Y_{i,k}\rho_{i,j,k}^{(t)}}{\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})^2}+\frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1+\rho_{i,j,k}^{(t)})^2}\Bigg\}\frac{(1-\rho_{i,j,k}^{(t)^2})}{2}Z_i.\end{aligned}$$

Next,

$$\begin{aligned}
\frac{\partial^2 \psi_{2,j,k}(\rho_{j,k}|\theta^{(t)})}{\partial \delta_{j,k} \partial \delta_{j,k}^T} = \sum_{i=1}^{n} \Bigg(& \frac{\exp(\delta_{j,k}^T Z_i)}{\{1+\exp(\delta_{j,k}^T Z_i)\}^2} \\
& - \left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^{(t)^2}} \right) \left(1 - \rho_{i,j,k}^{(t)^2}\right) \times \frac{1}{64} \Big[3\{\exp(\delta_{j,k}^T Z_i/2) + \exp(-\delta_{j,k}^T Z_i/2)\}^2 \\
& \times \{\exp(\delta_{j,k}^T Z_i/2) - \exp(-\delta_{j,k}^T Z_i/2)\}^2 + \{\exp(\delta_{j,k}^T Z_i/2) + \exp(-\delta_{j,k}^T Z_i/2)\}^4 \Big] \\
& + \frac{\left(Y_{i,j} + Y_{i,k}\right)^2}{2\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)} (1 - \rho_{i,j,k}^{(t)^2})} \times \frac{2\exp(\delta_{j,k}^T Z_i)}{\{1+\exp(\delta_{j,k}^T Z_i)\}^2} \\
& - \frac{\left(Y_{i,j}^2 + Y_{i,k}^2\right)}{6\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)}} \left(1 - \rho_{i,j,k}^{(t)^2}\right)^2 \times \frac{3}{128} \Big[5\{\exp(\delta_{j,k}^T Z_i/2) + \exp(-\delta_{j,k}^T Z_i/2)\}^4 \\
& \times \{\exp(\delta_{j,k}^T Z_i/2) - \exp(-\delta_{j,k}^T Z_i/2)\}^2 + \{\exp(\delta_{j,k}^T Z_i/2) + \exp(-\delta_{j,k}^T Z_i/2)\}^6 \Big] \\
& - \frac{\left(Y_{i,j} + Y_{i,k}\right)^2}{6\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)}} \left(1 + \rho_{i,j,k}^{(t)}\right)^2 \times \frac{3}{8} \Big[2\{1+\exp(-\delta_{j,k}^T Z_i)\}\{1-\exp(-\delta_{j,k}^T Z_i)\}^2 \\
& + \{\exp(\delta_{j,k}^T Z_i) + \exp(-\delta_{j,k}^T Z_i)\}^3 \Big] \\
& + \frac{\left(Y_{i,j}^2 + Y_{i,k}^2\right)}{2\sigma_{i,j}^{(t)} \sigma_{i,k}^{(t)} (1 + \rho_{i,j,k}^{(t)})} \times \frac{\exp(\delta_{j,k}^T Z_i)}{\{1+\exp(\delta_{j,k}^T Z_i)\}^2} \Bigg) Z_i Z_i^T .
\end{aligned}$$

Subsequently,

$$\left(\frac{\partial^2\psi_{2,j,k}(\rho_{j,k}|\theta^{(t)})}{\partial\delta_{j,k}\partial\delta_{j,k}^T}\right)_{\theta=\theta^{(t)}} = \sum_{i=1}^{n}\Bigg(\frac{\exp(Z_i^T\delta_{j,k}^{(t)})}{\{1+\exp(Z_i^T\delta_{j,k}^{(t)})\}^2}$$
$$-\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}}+\frac{Y_{i,k}^2}{\sigma_{i,k}^{(t)^2}}\right)\left(1-\rho_{i,j,k}^{(t)^2}\right)\times\frac{1}{64}\Bigg[3\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)$$
$$+\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\}^2\times\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)-\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\}^2$$
$$+\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)+\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\}^4\Bigg]$$
$$+\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})}\times\frac{2\exp(Z_i^T\delta_{j,k}^{(t)})}{\{1+\exp(Z_i^T\delta_{j,k}^{(t)})\}^2}$$
$$-\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1-\rho_{i,j,k}^{(t)^2}\right)^2\times\frac{3}{128}\Bigg[5\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)$$
$$+\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\}^4\times\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)-\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\}^2$$
$$+\{\exp(Z_i^T\delta_{j,k}^{(t)}/2)+\exp(-Z_i^T\delta_{j,k}^{(t)}/2)\}^6\Bigg]$$
$$-\frac{\left(Y_{i,j}+Y_{i,k}\right)^2}{6\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left(1+\rho_{i,j,k}^{(t)}\right)^2\times\frac{3}{8}\Bigg[2\{1+\exp(-Z_i^T\delta_{j,k}^{(t)})\}$$
$$\times\{1-\exp(-Z_i^T\delta_{j,k}^{(t)})\}^2$$
$$+\{\exp(Z_i^T\delta_{j,k}^{(t)})+\exp(-Z_i^T\delta_{j,k}^{(t)})\}^3\Bigg]+\frac{\left(Y_{i,j}^2+Y_{i,k}^2\right)}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1+\rho_{i,j,k}^{(t)})}$$
$$\times\frac{\exp(Z_i^T\delta_{j,k}^{(t)})}{\{1+\exp(Z_i^T\delta_{j,k}^{(t)})\}^2}\Bigg)Z_iZ_i^T.$$

Simplifying further, we obtain,

$$\begin{aligned}\left(\frac{\partial^2 \ell^*(\theta|\theta^{(t)})}{\partial\delta_{j,k}\partial\delta_{j,k}^T}\right)_{\theta=\theta^{(t)}} &= \sum_{i=1}^{n}\Bigg[\frac{1+\rho_{i,j,k}^{(t)^2}}{(1-\rho_{i,j,k}^{(t)^2})^2}-\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^{(t)^2}}+\frac{Y_{i,k}^2}{\sigma_{i,k}^{(t)^2}}\right)\times\frac{1+5\rho_{i,j,k}^{(t)^2}}{(1-\rho_{i,j,k}^{(t)^2})^3}\\ &+\frac{1}{\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})^3}\Bigg\{(Y_{i,j}+Y_{i,k})^2(1+\rho_{i,j,k}^{(t)^2})\\ &-(Y_{i,j}^2+Y_{i,k}^2)(1+7\rho_{i,j,k}^{(t)^2})\Bigg\}\\ &+\frac{Y_{i,j}^2+Y_{i,k}^2-4(Y_{i,j}+Y_{i,k})^2}{2\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}(1+\rho_{i,j,k}^{(t)})^3}\Bigg]\frac{(1-\rho_{i,j,k}^{(t)^2})^2}{4}Z_iZ_i^T\\ &+\sum_{i=1}^{n}\Bigg[\frac{\rho_{i,j,k}^{(t)}}{1-\rho_{i,j,k}^{(t)^2}}-\left(\frac{Y_{i,j}^2}{\sigma^{(t)^2}_{i,j}}+\frac{Y_{i,k}^2}{\sigma^{(t)^2}_{i,k}}\right)\frac{\rho_{i,j,k}^{(t)}}{(1-\rho_{i,j,k}^{(t)^2})^2}\\ &-\frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}^{(t)}\sigma_{i,k}^{(t)}}\left\{\frac{2\rho_{i,j,k}^{(t)}}{(1-\rho_{i,j,k}^{(t)^2})^2}+\frac{1}{(1+\rho_{i,j,k}^{(t)})^2}\right\}\Bigg]\frac{\rho_{i,j,k}^{(t)}(1-\rho_{i,j,k}^{(t)^2})}{2}Z_iZ_i^T.\end{aligned}$$

4.4 Inference

Let θ_0 be the true parameter lies in an open subset of multidimensional Euclidean space. Assume that all predictors are in a compact subset of multidimensional Euclidean space, $\sum_{i=1}^{n} X_i X_i^\top$ has full rank, and other regularity conditions hold. Then following the standard asymptotic results [70], we obtain $\sqrt{n}(\widehat{\theta}-\theta_0) \longrightarrow N_{n_\theta}(0, \mathcal{G}^{-1})$ in distribution, where N_{n_θ} denotes the n_θ-variate multivariate normal distribution. The Godambe information $\mathcal{G}$ is $\mathcal{G}(\theta_0) = \mathcal{H}(\theta_0)\mathcal{J}^{-1}(\theta_0)\mathcal{H}^T(\theta_0)$, where $\mathcal{H}(\theta) = E[-\nabla_\theta \mathcal{U}(\theta; D)]$, $\mathcal{J}(\theta) = \text{var}[\mathcal{U}(\theta; D)]$, where D denotes the data from randomly chosen subject or experimental unit, and $\mathcal{U}(\theta; D)$ denotes the score function corresponding to this randomly chosen subject. The information $\mathcal{G}(\theta_0)$ is consistently estimated by $\widehat{\mathcal{H}}\widehat{\mathcal{J}}^{-1}\widehat{\mathcal{H}}^T$, where

$\widehat{\mathcal{H}} = -(1/n)\sum_{i=1}^{n} \nabla_\theta \mathcal{U}(\theta; D_i)|_{\widehat{\theta}}$, $\widehat{\mathcal{J}} = (1/n)\sum_{i=1}^{n} \mathcal{U}(\theta; D_i)\mathcal{U}(\boldsymbol{\theta}; D_i)^T|_{\widehat{\theta}}$,

$$\begin{aligned}
\mathcal{U}(\theta; D_i) &= \frac{\partial}{\partial\theta}\sum_{j}\sum_{j<k}\log\{f(Y_{i,j}, Y_{i,k}|X_i)\} \\
&= \frac{\partial}{\partial\theta}\Bigg[-\frac{1}{2}\sum_{j=1}^{q}(q-1)\log(\sigma_{i,j}^2) - \frac{1}{2}\sum_{j<k}\sum\log(1-\rho_{i,j,k}^2) \\
&\quad -\frac{1}{2}\sum_{j<k}\sum\frac{1}{(1-\rho_{i,j,k}^2)}\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^2} - \frac{2\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^2}\right)\Bigg],
\end{aligned}$$

and detailed expressions of $\mathcal{H}$ and $\mathcal{J}$ are given below. Thus, the standard error of the rth component of $\widehat{\theta}$ is the square root of the rth diagonal element of the inverse of $\widehat{\mathcal{H}}\widehat{\mathcal{J}}^{-1}\widehat{\mathcal{H}}^T$. This standard error helps compute the Wald confidence interval for the parameter and can also be used for hypothesis testing. Let $\ell_{i,j,k}(\alpha, \delta) = \log\{f(Y_{i,j}, Y_{i,k}|X_i)\}$, and $\ell_i(\alpha, \delta) = \sum_j \sum_{j<k} \log\{f(Y_{i,j}, Y_{i,k}|X_i)\}$. Then,

$$\begin{aligned}
\ell_{i,j,k}(\alpha, \delta) = -\frac{1}{2}\Bigg[&\log(\sigma_{i,j}^2) + \log(\sigma_{i,k}^2) + \log(1-\rho_{i,j,k}^2) \\
&+ \frac{1}{(1-\rho_{i,j,k}^2)}\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^2} - \frac{2\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^2}\right)\Bigg]
\end{aligned}$$

and

$$\begin{aligned}
\ell_i(\alpha, \delta) &= -\frac{1}{2}\sum_{j=1}^{q-1}\sum_{k=j+1}^{q}\Bigg[\log(\sigma_{i,j}^2) + \log(\sigma_{i,k}^2) + \log(1-\rho_{i,j,k}^2) \\
&\quad + \frac{1}{(1-\rho_{i,j,k}^2)}\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^2} - \frac{2\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^2}\right)\Bigg] \\
&= -\frac{1}{2}\sum_{j=1}^{q}(q-1)\log(\sigma_{i,j}^2) - \frac{1}{2}\sum_{j<k}\sum\log(1-\rho_{i,j,k}^2) \\
&\quad -\frac{1}{2}\sum_{j<k}\sum\frac{1}{(1-\rho_{i,j,k}^2)}\left(\frac{Y_{i,j}^2}{\sigma_{i,j}^2} - \frac{2\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}} + \frac{Y_{i,k}^2}{\sigma_{i,k}^2}\right).
\end{aligned}$$

We need to calculate the score functions $\mathcal{U}(\theta; D_i) = \partial \ell_i(\alpha, \delta)/\partial\theta$. For this derivation we use $\partial\sigma_{i,j}/\partial\alpha_j = \sigma_{i,j}Z_i$ and $\partial\rho_{i,j,k}/\partial\delta_{j,k} = 0.5(1-\rho^2_{i,j,k})Z_i$. For $j = 1, \ldots, q$,

$$\begin{aligned}
\frac{\partial \ell_i(\alpha,\delta)}{\partial \alpha_j} &= -\frac{(q-1)}{\sigma_{i,j}}\frac{\partial\sigma_{i,j}}{\partial\alpha_j} + \sum_{k=1,k\neq j}^{q}\frac{Y^2_{i,j}}{(1-\rho^2_{i,j,k})\sigma^3_{i,j}}\frac{\partial\sigma_{i,j}}{\partial\alpha_j} - \sum_{k=1,k\neq j}^{q}\frac{\rho_{i,j,k}Y_{i,j}Y_{i,k}}{(1-\rho^2_{i,j,k})\sigma_{i,k}\sigma^2_{i,j}}\frac{\partial\sigma_{i,j}}{\partial\alpha_j} \\
&= -(q-1)Z_i + \sum_{k=1,k\neq j}^{q}\frac{Y^2_{i,j}}{(1-\rho^2_{i,j,k})\sigma^2_{i,j}}Z_i - \sum_{k=1,k\neq j}^{q}\frac{\rho_{i,j,k}Y_{i,j}Y_{i,k}}{(1-\rho^2_{i,j,k})\sigma_{i,k}\sigma_{i,j}}Z_i, \\
\frac{\partial \ell_i(\alpha,\delta)}{\partial \delta_{j,k}} &= \frac{\rho_{i,j,k}}{(1-\rho^2_{i,j,k})}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}} - \left(\frac{Y^2_{i,j}}{\sigma^2_{i,j}} + \frac{Y^2_{i,k}}{\sigma^2_{i,k}}\right)\frac{\rho_{i,j,k}}{(1-\rho^2_{i,j,k})^2}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}} \\
&\quad + \frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}}\left((1-\rho^2_{i,j,k})^{-1}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}} + \frac{2\rho^2_{i,j,k}}{(1-\rho^2_{i,j,k})^2}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}}\right) \\
&= \frac{\rho_{i,j,k}}{2}Z_i - \frac{1}{2}\left(\frac{Y^2_{i,j}}{\sigma^2_{i,j}} + \frac{Y^2_{i,k}}{\sigma^2_{i,k}}\right)\frac{\rho_{i,j,k}}{(1-\rho^2_{i,j,k})}Z_i + \frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}}\left(\frac{Z_i}{2} + \frac{\rho^2_{i,j,k}}{(1-\rho^2_{i,j,k})}Z_i\right),
\end{aligned}$$

for $j < k$. To calculate the sensitivity matrix, we need to calculate the double derivatives of the above two expressions. That is for $j = 1, \ldots, q$

$$\begin{aligned}
\frac{\partial^2 \ell_i(\alpha,\delta)}{\partial\alpha_j\partial\alpha_j^T} &= \sum_{k=1,k\neq j}^{q} -2\left(\frac{Y^2_{i,j}}{(1-\rho^2_{i,j,k})\sigma^3_{i,j}}\right)Z_i\frac{\partial\sigma_{i,j}}{\partial\alpha_j^T} + \sum_{k=1,k\neq j}^{q}\left(\frac{\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma^2_{i,j}\sigma_{i,k}}\right)Z_i\frac{\partial\sigma_{i,j}}{\partial\alpha_j^T} \\
&= -2\sum_{k=1,k\neq j}^{q}\left(\frac{Y^2_{i,j}}{(1-\rho^2_{i,j,k})\sigma^2_{i,j}}\right)Z_iZ_i^T + \sum_{k=1,k\neq j}^{q}\left(\frac{\rho_{i,j,k}Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}}\right)Z_iZ_i^T, \\
\frac{\partial^2 \ell_i(\alpha,\delta)}{\partial\alpha_j\partial\alpha_k^T} &= \frac{\rho_{i,j,k}Y_{i,j}Y_{i,k}}{(1-\rho^2_{i,j,k})\sigma_{i,j}\sigma_{i,k}}Z_iZ_i^T,
\end{aligned}$$

and for $j < k$

$$\begin{aligned}
\frac{\partial^2 \ell_i(\alpha,\delta)}{\partial\alpha_j\partial\delta_{j,k}^T} &= \left\{\frac{Y^2_{i,j}\rho_{i,j,k}}{\sigma^2_{i,j}(1-\rho^2_{i,j,k})} - \frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}}\left(\frac{1}{2} + \frac{\rho^2_{i,j,k}}{1-\rho^2_{i,j,k}}\right)\right\}Z_iZ_i^T, \\
\frac{\partial^2 \ell_i(\alpha,\delta)}{\partial\delta_{j,k}\partial\delta_{j,k}^T} &= \frac{(1-\rho^2_{i,j,k})Z_iZ_i^T}{4} - \frac{Z_i}{2}\left(\frac{Y^2_{i,j}}{\sigma^2_{i,j}} + \frac{Y^2_{i,k}}{\sigma^2_{i,k}}\right)\left((1-\rho^2_{i,j,k})^{-1}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}^T} + \frac{2\rho^2_{i,j,k}}{(1-\rho^2_{i,j,k})^2}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}^T}\right) \\
&\quad + \frac{Y_{i,j}Y_{i,k}Z_i}{\sigma_{i,j}\sigma_{i,k}}\left(\frac{2\rho^3_{i,j,k}}{(1-\rho^2_{i,j,k})^2}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}^T} + \frac{2\rho_{i,j,k}}{(1-\rho^2_{i,j,k})}\frac{\partial\rho_{i,j,k}}{\partial\delta_{j,k}^T}\right) \\
&= \left\{\frac{(1-\rho^2_{i,j,k})}{4} - \frac{1}{2}\left(\frac{Y^2_{i,j}}{\sigma^2_{i,j}} + \frac{Y^2_{i,k}}{\sigma^2_{i,k}}\right)\left(\frac{1}{2} + \frac{\rho^2_{i,j,k}}{(1-\rho^2_{i,j,k})}\right)\right. \\
&\quad \left. + \frac{Y_{i,j}Y_{i,k}}{\sigma_{i,j}\sigma_{i,k}}\left(\frac{\rho^3_{i,j,k}}{(1-\rho^2_{i,j,k})} + \rho_{i,j,k}\right)\right\}Z_iZ_i^T.
\end{aligned}$$

4.5 Simulation studies

4.5.1 Simulation design

Three different scenarios were considered. The number of phenotypes, q, was set to four for all scenarios. For scenarios 1 and 2, the number of predictors p was set to 2, and the sample size n was set to 261 and 600, respectively. For the third scenario, p was set to 6 and n to 500. All these numbers were chosen by closely following real datasets [11]. Each dataset contained information on X and Y from n independent units. The predictor X had p components for every unit, and each component was independently simulated from the Bernoulli(0.5) distribution. Next, Y was generated from $N_4(0, \Sigma_i)$, where $\Sigma_i = \text{Diag}(\sigma_{i,1}, \ldots, \sigma_{i,q}) R_i \text{Diag}(\sigma_{i,1}, \ldots, \sigma_{i,q})$, where $R_i = ((\rho_{i,j,k}))$. The true values of the parameter θ are given in the simulation tables.

4.5.2 Method of analysis

For each scenario $\rho_{i,j,k}$'s and $\sigma_{i,j}$'s were modelled according to Equations (4.1) and (4.2) with respectively. Under each scenario 500 datasets were generated. Each dataset was analyzed by two approaches, 1) the proposed MM algorithm, and 2) the direct method where parameter estimates were obtained by directly maximizing the log-composite likelihood function. Under approach 2), we used the `optim` function of `R` and chose to optimize using the L-BFGS-B, a variant of the Broyden–Fletcher–Goldfarb–Shanno algorithm.

For both the approaches, the initial values for θ parameters were randomly generated from Normal(0, 0.10). Since our proposed method is an iterative optimization, we used the sum of the absolute relative difference between the parameter estimates in subsequent iterations to be less than 0.001 as the stopping criteria for the convergence.

4.5.3 Results

Results for scenarios 1 and 2 are presented in Tables 4.1 and 4.2. Results for scenario 3 are presents in Tables 4.3 and 4.4.

Table 4.1: Results of the simulation study for scenario 1 with $n = 261$, $p = 2$, $q = 4$. All entries of the table except for the true parameter values are multiplied by 100. Par: Parameter, SD: standard deviation, SE: standard error, CP: 95% coverage probability, RMSE: root mean squared error

Par	True	Bias	SD	SE	CP	RMSE	Par	True	Bias	SD	SE	CP	RMSE
$\alpha_{1,0}$	−1.9	−0.9	8.0	7.4	93.8	8.0	$\alpha_{3,0}$	−1.3	−0.1	7.8	7.5	93.2	7.8
$\alpha_{1,1}$	−0.4	0.1	8.9	8.5	93.6	8.9	$\alpha_{3,1}$	−0.2	0.1	8.6	8.6	96.2	8.6
$\alpha_{1,2}$	0.3	−0.1	9.1	8.5	93.0	9.1	$\alpha_{3,2}$	0	−0.9	9.1	8.6	94.0	9.1
$\alpha_{2,0}$	−1.7	0.3	7.8	7.4	92.4	7.8	$\alpha_{4,0}$	−1.4	−0.2	7.9	7.4	92.2	7.9
$\alpha_{2,1}$	−0.4	−1.0	8.8	8.5	93.8	8.9	$\alpha_{4,1}$	0	0	9.2	8.5	92.4	9.2
$\alpha_{2,2}$	0	−0.6	8.9	8.5	93.6	8.9	$\alpha_{4,2}$	0	−0.4	8.9	8.6	93.4	8.9
$\delta_{1,2,0}$	−0.7	−0.4	21.2	20.9	94.2	21.2	$\delta_{2,3,0}$	0	−0.1	22.3	21.2	93.6	22.3
$\delta_{1,2,1}$	−0.8	0.2	24.6	24.5	95.8	24.6	$\delta_{2,3,1}$	0	0.5	26.9	24.5	92.0	26.8
$\delta_{1,2,2}$	0	0.1	25.0	24.1	94.8	25.0	$\delta_{2,3,2}$	0	0.3	25.2	24.6	93.8	25.2
$\delta_{1,3,0}$	1.2	1.3	21.3	21.4	94.6	21.3	$\delta_{2,4,0}$	1.1	1.5	20.9	20.9	93.8	20.9
$\delta_{1,3,1}$	0	−0.9	24.9	24.6	94.6	24.9	$\delta_{2,4,1}$	0	−1.5	24.8	24.0	95.0	24.8
$\delta_{1,3,2}$	0	−0.1	25.5	24.6	94.4	25.5	$\delta_{2,4,2}$	−0.9	−0.2	24.9	24.5	95.6	24.9
$\delta_{1,4,0}$	0	−0.1	22.2	21.1	93.6	22.1	$\delta_{3,4,0}$	0	0.8	22.4	21.2	94.4	22.4
$\delta_{1,4,1}$	0	0.7	25.5	24.3	91.8	25.4	$\delta_{3,4,1}$	0	−1.3	26.1	24.6	93.0	26.1
$\delta_{1,4,2}$	0.6	0.1	24.9	24.6	95.0	24.9	$\delta_{3,4,2}$	0	−0.2	25.4	24.6	95.0	25.4

Table 4.2: Results of the simulation study for scenario 2 with $n = 600$, $p = 2$, $q = 4$. All entries except for the true parameter values of the table are multiplied by 100. Par: Parameter, SD: standard deviation, SE: standard error, CP: 95% coverage probability, RMSE: root mean squared error

Par	True	Bias	SD	SE	CP	RMSE	Par	True	Bias	SD	SE	CP	RMSE
$\alpha_{1,0}$	−1.0	−0.1	5.1	4.9	94.8	5.1	$\alpha_{3,0}$	1.0	−0.6	5.1	4.9	94.8	5.1
$\alpha_{1,1}$	1.0	0.1	5.7	5.6	94.2	5.7	$\alpha_{3,1}$	0.3	0.4	5.4	5.7	97.0	5.4
$\alpha_{1,2}$	0.2	0	5.7	5.7	94.6	5.7	$\alpha_{3,2}$	0.1	0.3	5.7	5.7	95.0	5.7
$\alpha_{2,0}$	2.0	−0.7	5.0	4.9	93.6	5.1	$\alpha_{4,0}$	−1.0	−0.4	5.3	5.0	92.4	5.3
$\alpha_{2,1}$	0.2	0.2	5.7	5.6	94.4	5.7	$\alpha_{4,1}$	−0.5	0.2	5.4	5.7	96.4	5.4
$\alpha_{2,2}$	−0.5	0.5	5.8	5.7	94.2	5.9	$\alpha_{4,2}$	1.0	0	5.8	5.7	95.4	5.8
$\delta_{1,2,0}$	0.2	0.1	14.0	13.9	93.4	14.0	$\delta_{2,3,0}$	−0.1	0.4	13.7	14.0	94.4	13.7
$\delta_{1,2,1}$	0.5	−0.7	15.6	15.9	95.0	15.6	$\delta_{2,3,1}$	1.0	1.0	15.9	16.2	94.8	16.0
$\delta_{1,2,2}$	1.0	1.2	16.8	16.2	95.6	16.8	$\delta_{2,3,2}$	−0.2	−1.0	15.9	15.8	94.4	15.9
$\delta_{1,3,0}$	0.2	0.7	13.7	14.1	95.6	13.7	$\delta_{2,4,0}$	0.2	0.4	15.1	14.1	94.2	15.1
$\delta_{1,3,1}$	0.2	−1.1	16.3	16.2	94.2	16.3	$\delta_{2,4,1}$	0.5	−0.5	16.6	15.9	95.0	16.6
$\delta_{1,3,2}$	0.5	−0.7	16.4	16.2	94.8	16.4	$\delta_{2,4,2}$	−1.0	0	16.6	16.2	94.2	16.6
$\delta_{1,4,0}$	0.2	0.1	14.0	14.2	95.4	14.0	$\delta_{3,4,0}$	−0.1	0.1	14.2	14.1	94.2	14.2
$\delta_{1,4,1}$	0.2	−0.4	16.2	16.2	94.4	16.2	$\delta_{3,4,1}$	0.2	−0.1	16.5	16.2	95.6	16.5
$\delta_{1,4,2}$	−0.5	1.2	15.9	16.3	96.6	16.0	$\delta_{3,4,2}$	−0.2	0.7	16.4	16.2	94.8	16.4

Table 4.3: Results of α parameters from the simulation study for scenario 3 with $n = 500$, $p = 6$, $q = 4$. All entries except for the true parameter values of the table are multiplied by 100. Par: Parameter, SD: standard deviation, SE: standard error, CP: 95% coverage probability, RMSE: root mean squared error

Par	True	Bias	SD	SE	CP	RMSE	Par	True	Bias	SD	SE	CP	RMSE
$\alpha_{1,0}$	−1.0	−0.8	8.7	8.0	92.4	8.8	$\alpha_{3,0}$	1.0	−1.7	8.7	8.1	92.6	8.8
$\alpha_{1,1}$	1.0	0.5	6.0	6.0	94.2	6.0	$\alpha_{3,1}$	0.3	0.4	6.5	6.1	92.4	6.5
$\alpha_{1,2}$	0.2	0.2	6.1	6.0	94.8	6.1	$\alpha_{3,2}$	0.1	−0.3	6.5	6.1	93.8	6.5
$\alpha_{1,3}$	−0.4	−0.4	6.2	6.0	95.2	6.2	$\alpha_{3,3}$	0	0.3	6.4	6.1	93.2	6.4
$\alpha_{1,4}$	0.3	−0.1	6.0	6.0	95.4	6.0	$\alpha_{3,4}$	−0.2	0.1	6.6	6.1	92.6	6.6
$\alpha_{1,5}$	0	0.5	6.2	6.0	93.4	6.2	$\alpha_{3,5}$	0.1	0.5	6.5	6.1	93.8	6.5
$\alpha_{1,6}$	−0.5	−0.1	6.5	6.0	93.8	6.5	$\alpha_{3,6}$	−0.2	0.6	6.3	6.1	93.8	6.3
$\alpha_{2,0}$	2.0	−0.4	8.4	8.1	93.4	8.4	$\alpha_{4,0}$	−1.0	−0.7	8.4	8.0	92.8	8.4
$\alpha_{2,1}$	0.2	−0.1	6.3	6.0	93.8	6.2	$\alpha_{4,1}$	−0.5	−0.2	6.5	6.0	92.0	6.5
$\alpha_{2,2}$	−0.5	0	6.2	6.0	94.6	6.2	$\alpha_{4,2}$	0	0	6.6	6.1	91.2	6.6
$\alpha_{2,3}$	−0.4	−0.3	6.7	6.2	92.8	6.7	$\alpha_{4,3}$	0.3	−0.6	6.3	6.0	93.4	6.3
$\alpha_{2,4}$	0	0	6.4	6.0	93.8	6.4	$\alpha_{4,4}$	−0.2	0.3	6.2	6.0	93.6	6.2
$\alpha_{2,5}$	0.3	−0.5	6.1	6.0	93.4	6.1	$\alpha_{4,5}$	−0.2	0.1	6.1	6.0	95.0	6.1
$\alpha_{2,6}$	0	−0.2	6.2	6.1	93.8	6.2	$\alpha_{4,6}$	0.2	0.2	6.2	6.0	94.4	6.2

We present the bias, the standard deviation of the estimates (SD), the estimated standard error (SE), the empirical coverage probability of the 95% Wald 's confidence intervals, and the root mean squared error (RMSE) of the estimates for the proposed MM algorithm. The second approach's results are qualitatively similar to the MM algorithm. Hence they are not presented in the tables.

The important take-way messages are 1) the biases of the parameters are negligible for different sample sizes and different p, 2) the SEs are very close to the SDs, indicating that the asymptotic standard deviation of the estimators is captured well by the SE, 3) the empirical coverage probabilities are pretty close to 0.95. All of these indicate that the method of estimation works well, and asymptotic properties of the estimator hold. The SD and SE decrease with the sample size (Tables 4.1 and 4.2 in the Supplementary Materials). Even for the scenario of a large number of parameters (Table 4.3, 4.4), the performance of the MM algorithm is extremely satisfactory.

In general, the bias and SD (also RMSE) are considerably larger for the δ parameters than the α parameters, indicating more uncertainties (less information) in the correlation parameters than the standard deviations.

Table 4.4: Results of δ parameters from the simulation study for scenario 3 with $n = 500$, $p = 6$, $q = 4$. All entries except for the true parameter values of the table are multiplied by 100. Par: Parameter, SD: standard deviation, SE: standard error, CP: 95% coverage probability, RMSE: root mean squared error

Par	True	Bias	SD	SE	CP	RMSE	Par	True	Bias	SD	SE	CP	RMSE
$\delta_{1,2,0}$	0.2	1.2	23.5	23.1	94.8	23.5	$\delta_{2,3,0}$	0	1.1	24.5	23.5	92.8	24.5
$\delta_{1,2,1}$	0.5	−2.4	17.7	17.3	95.6	17.8	$\delta_{2,3,1}$	0	0.3	19.3	17.8	93.4	19.3
$\delta_{1,2,2}$	0	−4.4	17.2	17.2	94.6	17.8	$\delta_{2,3,2}$	−0.2	3.1	18.5	17.9	93.6	18.7
$\delta_{1,2,3}$	−0.7	3.3	18.3	17.4	94.4	18.6	$\delta_{2,3,3}$	0	−2.5	18.9	17.8	93.8	19.0
$\delta_{1,2,4}$	0.3	−2.7	16.9	17.2	95.2	17.1	$\delta_{2,3,4}$	0	1.7	17.0	17.8	95.0	17.1
$\delta_{1,2,5}$	0	1.8	17.6	17.2	93.0	17.7	$\delta_{2,3,5}$	0	−1.1	17.8	17.8	94.4	17.8
$\delta_{1,2,6}$	−0.8	3.0	18.0	17.5	93.4	18.2	$\delta_{2,3,6}$	0	−2.6	18.4	17.8	93.0	18.6
$\delta_{1,3,0}$	0.5	−3.4	25.0	23.4	93.0	25.2	$\delta_{2,4,0}$	0.3	1.7	23.3	23.0	94.4	23.4
$\delta_{1,3,1}$	0.2	−0.5	17.9	17.6	94.2	17.9	$\delta_{2,4,1}$	0	2.7	18.6	17.2	93.0	18.8
$\delta_{1,3,2}$	0.5	−1.4	17.7	17.7	95.0	17.8	$\delta_{2,4,2}$	−1.0	4.1	18.14	17.6	93.6	18.6
$\delta_{1,3,3}$	0	1.9	19.3	17.5	92.0	19.3	$\delta_{2,4,3}$	0.4	−4.5	18.0	17.2	93.2	18.5
$\delta_{1,3,4}$	0.2	−3.9	19.2	17.5	92.4	19.6	$\delta_{2,4,4}$	−0.3	1.4	17.0	17.2	94.4	17.1
$\delta_{1,3,5}$	−0.5	5.4	18.8	17.7	93.6	19.5	$\delta_{2,4,5}$	0.5	−1.4	17.6	17.2	93.6	17.7
$\delta_{1,3,6}$	0	1.0	18.4	17.5	93.8	18.4	$\delta_{2,4,6}$	0.3	−6.0	18.4	17.2	92.4	19.3
$\delta_{1,4,0}$	1.0	−1.4	26.8	23.5	90.8	26.8	$\delta_{3,4,0}$	−1.0	0.5	24.4	23.4	92.6	24.4
$\delta_{1,4,1}$	0.2	−2.1	19.6	17.7	92.0	19.7	$\delta_{3,4,1}$	0.2	1.3	18.3	17.6	95.2	18.3
$\delta_{1,4,2}$	−0.5	−0.1	18.1	17.8	94.8	18.1	$\delta_{3,4,2}$	−0.2	2.2	16.9	17.7	95.4	17.0
$\delta_{1,4,3}$	−0.2	0.1	18.9	17.7	93.8	18.9	$\delta_{3,4,3}$	−0.1	1.0	18.5	17.6	94.0	18.5
$\delta_{1,4,4}$	0.5	−2.8	18.3	17.8	94.2	18.5	$\delta_{3,4,4}$	0.2	2.9	17.8	17.7	95.2	18.0
$\delta_{1,4,5}$	0.1	4.6	18.5	17.7	92.6	19.1	$\delta_{3,4,5}$	0.6	−2.8	17.9	17.7	95.2	18.1
$\delta_{1,4,6}$	−0.1	−0.9	18.0	17.7	95.0	18.0	$\delta_{3,4,6}$	−0.1	−1.4	18.6	17.6	92.8	18.6

4.5.4 Computational advantage

We have extended Scenario 3 from the simulation design by varying the number of predictor variables. Specifically, we set the number of phenotypes, q to 4, and the sample size, n to 500. We used four different values of the predictor variable, p= 2, 3, 4, 5. This resulted in the number of unknown parameters in our setting as 30, 40, 50, 60 respectively. The multivariate phenotype response, Y, and the design matrix, X were generated exactly as Section 4.5. Under each scenario, we performed 100 simulations. Figure 4.1 shows the average computation time (in seconds) of the MM algorithm

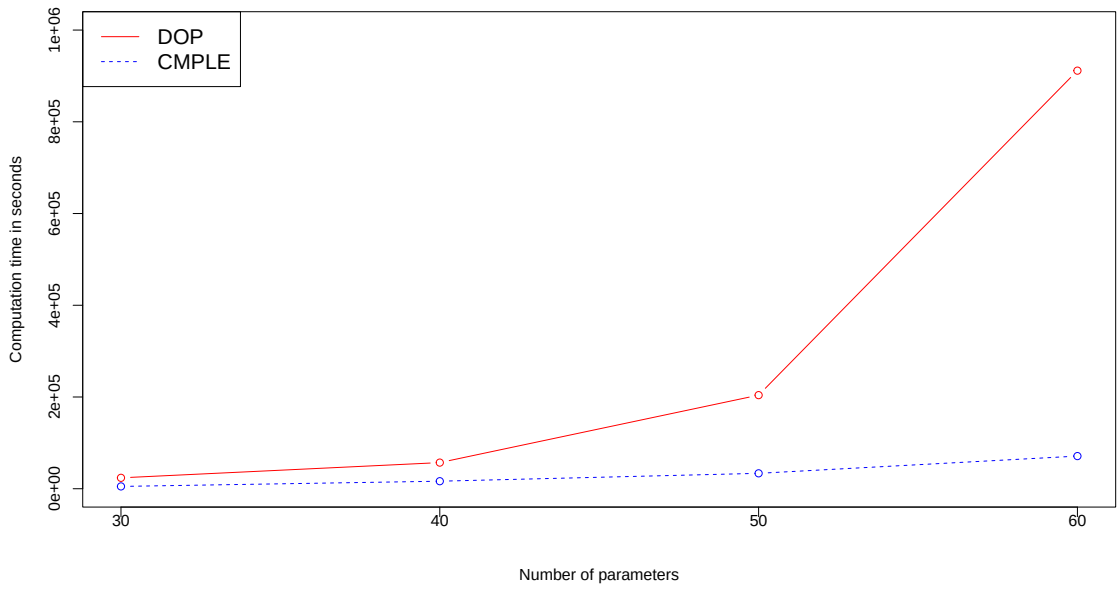

Figure 4.1: Average computational time comparison between CMPLE and direct optimization method (DOP) for 100 simulations

and direct optimization (DOP) via the optim function with the "L-BFGS-B" method. Both techniques were used to maximize the pairwise composite likelihood function. The numerical results seem to indicate that the DOP method has an exponential time complexity, and CMPLE has linear time complexity with respect to the number of parameters.

Our method has a clear advantage in terms of computation time over the direct optimization method. The proposed method is at least four times faster than the direct method (see Table 4.5). All simulations were done on a 2.8Ghz Intel Xeon E5-1603 processor.

Table 4.5: Average computation time (in seconds) using the MM algorithm and direct optimization (DOP) via the optim function with the "L-BFGS-B" method for 100 simulations under different scenarios.

	Simulation scenario		
	1	2	3
MM	5021	16394	33576
DOP	23782	56970	204036

4.6 Data Example

4.6.1 Background

We analyzed a population of cowpea (*Vigna unguiculata. (L.)Walp.*) recombinant inbred lines (RILs) which has a high level of genetic diversity and significantly variable phenotypic response to fluctuating environments. Previous studies have demonstrated strong genetic variation on photosynthetic responses in cowpea that co-regulates the light reactions of photosynthesis [3]. We were particularly interested in assessing the phenotype associations in terms of previously identified candidate genes under two environmental conditions: (1) CT, control temperature 29°C/19°C (day/night), and (2) LT, low or suboptimal temperature (chilling stress) 19°C/13°C (day/night). The responses consisted of $q = 4$ phenotypes, namely (1) steady-state PS II quantum yields, ϕ_{II}, (2) non-photochemical quenching, NPQ_t, (3) Q_A redox state PS II center opened, q_L, and (4) thylakoid pmf (proton motive force), ECS_t. These phenotypes were measured using MultispeQ 2.0 hand-held instruments as described in [47].

For this experiment, $n = 470$ observations were used which originated from a cross between a tolerant cultivar *California Blackeye 27 (CB27)* bred by the University of California, Riverside and a sensitive breeding line *24-125B-1* developed by Institute de Recherche Agricole pour le Développement (IRAD, Cameroon). Single nucleotide polymorphism (SNP) markers of genotype data of *CB27* × *24-125B-1*, based on EST sequences produced by [71]. Individuals of the RIL population are homozygous for each marker in the two parental lines, as indicated by the designations of either AA, having the allele from *CB27* (tolerant, maternal line), or BB, having the allele from *24-125B-1* (sensitive, paternal line). To incorporate them in our analysis, we have used dummy coding to transform them into binary (0, 1) features, where 0 (1) characterize the AA (BB) allele at a given marker locus.

First, we performed individual QTL analysis on these four phenotypes using the Multiple QTL Mapping (MQM) model using the `Rqtl` package [72]. LOD thresholds were determined using a permutation analysis implemented with the `mqmpermutation` and `mqmscan` functions with the

number of permutations set at 1000 and a nominal significance cutoff of $p < 0.05$. Results from the QTL analysis are presented in Figure 4.2. We found two candidate loci under QTL peaks at chromosome 4 (59.64 cm) and chromosome 9 (86.93 cm) that are the common significant SNP's under both conditions. These loci were also predicted by pseudomolecules through BLAST in early release genomes in Phytozome and those are annotated by Pfam, Panther, EuKaryotic Orthologous Groups (KOG), Kyoto Encyclopedia of Genes and Genomes (KO), Gene Ontology (GO) and best-hit of Arabidopsis gene. For the subsequent analysis, we have used these $q = 4$ phenotypes with $p = 3$ predictors (two candidate loci and one environmental variable).

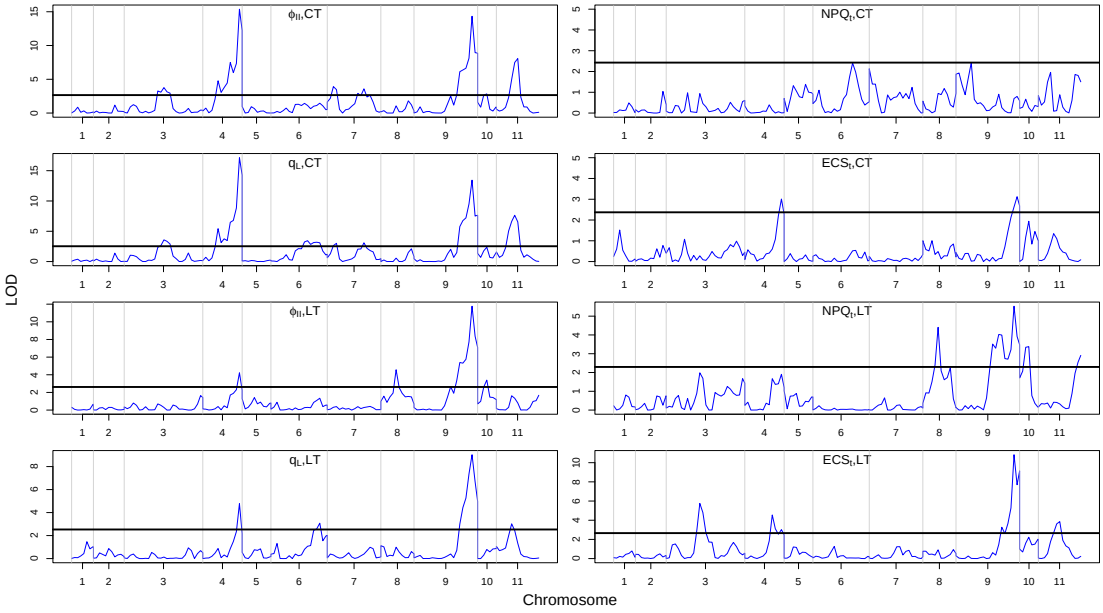

Figure 4.2: QTL plot of different phenotypes used in the Cowpea RIL data. LOD threshold for each phenotype is marked by the bold horizontal line. QTL with a LOD higher than that can be considered significant. Chromosomes are marked by the vertical lines.

4.6.2 Method of analyses

We fit the following model to $\sigma_{\phi_{II}}$, the standard deviation of phenotype ϕ_{II}, in terms of the predictors, $\sigma_{\phi_{II}} = \exp(\alpha_{1,0} + \alpha_{1,1}\text{Marker 1} + \alpha_{1,2}\text{Marker 2} + \alpha_{1,3}\text{Environment})$. Similar model was fit to the standard deviation of the other phenotypes. Simultaneously, we fit the following model to the pairwise correlations between phenotypes ϕ_{II} and NPQ_t, $\rho_{\phi_{II}\&NPQ_t} = 1 - 2/\{1 +$

$\exp(\delta_{1,2,0}+\delta_{1,2,1}\text{Marker 1}+\delta_{1,2,2}\text{Marker 2}+\delta_{1,2,3}\text{Environment})\}$. Similarly, the remaining five pairwise correlations, $\rho_{\phi_{II}\&q_L}, \rho_{\phi_{II}\&ECS_t}, \rho_{NPQ_t\&q_L}, \rho_{NPQ_t\&ECS_t}, \rho_{q_L\&ECS_t}$ were modelled in terms of the predictors.

Before applying the MM algorithm, we subtracted respective mean from the four phenotypes to have mean zero. For the above described models, there were a total of 40 parameters including α parameters and δ parameters. Note that $\theta = (\alpha^\top, \delta^\top)^\top$. We set the initial value of θ to the random numbers generated from Normal(0, 0.15), and used the sum of the absolute relative difference between subsequent iterations to be less than $\epsilon_0 = 0.001$ as the stopping criteria for the iterative algorithm 4.5.

4.6.3 Interpretation

The results of our analyses are placed in Tables 4.6 and 4.7. Specifically, Table 4.6 contains

Table 4.6: Parameter estimates and the 95% confidence interval in parentheses of the parameters of the standard deviation model for the measured phenotypes from the cowpea dataset.

	ϕ_{II}		NPQ_t		qL		ECS_t	
Intercept	−2.61	(−2.73, −2.49)	−0.37	(−0.51, −0.23)	−2.17	(−2.29, −2.05)	−3.18	(−3.43, −2.93)
Marker 1	−0.04	(−0.16, 0.08)	−0.43	(−0.61, −0.25)	0.03	(−0.13, 0.19)	0.02	(−0.16, 0.20)
Marker 2	0.08	(−0.04, 0.20)	0.45	(0.29, 0.61)	−0.07	(−0.21, 0.07)	−0.24	(−0.42, −0.06)
Environment	0.19	(0.07, 0.31)	0.45	(0.29, 0.61)	−0.14	(−0.30, 0.02)	0.16	(0.01, 0.32)

Table 4.7: Parameter estimates and 95% confidence interval in parentheses of the parameters of pairwise correlation among the measured phenotypes from the cowpea dataset.

	Intercept		Marker 1		Marker 2		Environment	
$\phi_{II}\&NPQ_t$	−1.35	(−1.72, −0.98)	−0.45	(−0.82, −0.08)	0.12	(−0.25, 0.49)	−0.71	(−1.10, −0.32)
$\phi_{II}\&qL$	1.82	(1.49, 2.15)	0.29	(−0.02, 0.60)	−0.32	(−0.63, −0.01)	−0.30	(−0.63, 0.03)
$\phi_{II}\&ECS_t$	0.16	(−0.23, 0.55)	−0.65	(−1.02, −0.28)	0.38	(−0.01, 0.77)	0.17	(−0.18, 0.52)
$NPQ_t\&qL$	0.35	(−0.12, 0.82)	−0.54	(−0.95, −0.13)	0.23	(−0.18, 0.64)	−0.42	(−0.75, −0.09)
$NPQ_t\&ECS_t$	0.82	(0.45, 1.19)	0.13	(−0.24, 0.50)	−0.14	(−0.55, 0.27)	−1.05	(−1.44, −0.66)
$qL\&ECS_t$	1.01	(0.58, 1.44)	−0.52	(−0.95, −0.09)	−0.01	(−0.42, 0.40)	−0.47	(−0.88, −0.06)

the estimate and 95% CI of the α parameters involved in the standard deviation modelling, whereas Table 4.7 corresponds to the δ parameters involved in the pairwise correlation modelling. We made several key observations from our analysis.

- The α parameters measure how the standard deviation of individual phenotypes changes with different candidate loci or the environmental factor. The α_j-parameters (j is the index for phenotype) can be viewed as the conditional effect of each predictor variable on j'th phenotype. For example, the conditional effect of Marker 1 on the standard deviation of NPQ_t was estimated to be -0.43. This means that in our population, if there is a change from allele AA to allele BB at Marker 1, the conditional standard deviation of NPQ_t will decrease by 35% while all other predictors remain unchanged. Likewise, if there is a change from allele AA to allele BB at Marker 2, then the conditional standard deviation of NPQ_t will increase by 57% while all other predictors remain unchanged. Like Marker 2, if the temperature changes from control (CT) to low (LT), then the conditional standard deviation of NPQ_t will increase by 57% while all other predictors remain unchanged. These are the most noteworthy changes in the standard deviation of the phenotypes. The standard deviation of ϕ_{II} and ECS_t seem to be affected by environment, and Marker 2 and environment, respectively. Similarly, in Table 4.7, we collect $\delta_{j,k}$ estimates that can be used to calculate the conditional effect of each predictor on the correlation between (j, k) phenotype pair. For example, the estimated regression parameter of Marker 1 on the pairwise correlation of ϕ_{II} and NPQ_t was -0.45. This means in our population, if the Marker 1 allele changes from AA to BB, the conditional pairwise correlation between ϕ_{II} and NPQ_t will decrease by 0.13 (using the Equation 4.1) when Marker 2 is at allele AA and environment is set at the control condition.

  Quantification of the changes in correlations based on the predictors has a profound significance in the photosynthetic experiments as it indicates the change in biological processes which plant adapts. As example, the estimated regression parameter of the Environment variable on NPQ_t and ECS_t was -1.05. This indicates in our population, if the Environment variable changes from control temperature(CT) to low temperature(LT), the conditional pairwise correlation between NPQ_t and ECS_t will decrease by 0.50 (using the Equation 4.1) when both the markers are at allele AA.

- Intercept terms, after appropriate transformation, represent the baseline conditional standard

deviation and baseline pairwise correlation among phenotypes when all the markers are fixed at allele AA and the environment variable is fixed at the control temperature. For example, in Table 4.6, an intercept of -2.61 under column ϕ_{II} implies that the standard deviation of the phenotype ϕ_{II} is $\exp(-2.61) = 0.07$ when all the predictors are at their baseline. Likewise, in Table 4.7, the estimated intercept -1.35 under the column $\phi_{II}\&NPQt$ implies that the estimated correlation between the phenotypes ϕ_{II} and $NPQt$ is $1 - 2/\{1 + \exp(-1.35)\} = -0.59$ when all the predictors are at their baseline. Consequently, the 95% CI of the intercept $(-1.72, -0.98)$ implies that the 95% CI of the correlation between the phenotypes ϕ_{II} and $NPQt$ is $(-0.70, -0.45)$ when all the predictors are at their baseline.

- We have also estimated the average marginal effects and the corresponding 95% confidence intervals for candidate loci on correlations between phenotypes in Table 4.8. For example, the average marginal effect of Marker 1 on the correlation between ϕ_{II} and $ECSt$ is estimated to be -0.32 (95% CI: $-0.50,\ -0.13$) when the marker allele changes from AA to BB. Likewise, the average marginal effect of the environment variable on the correlation between $NPQt$ and $ECSt$ is estimated to be -0.51 (95% CI: $-0.69,\ -0.32$).

Table 4.8: Average marginal effect estimates and 95% confidence interval in parentheses of the pairwise correlations between phenotypes based on predictors.

	Marker 1		Marker 2		Environment	
$\phi_{II}\&NPQ_t$	-0.11	$(-0.27, 0.04)$	0.03	$(-0.06, 0.12)$	-0.16	$(-0.25, -0.07)$
$\phi_{II}\&qL$	0.08	$(-0.07, 0.23)$	-0.08	$(-0.17, -0.01)$	-0.08	$(-0.17, 0.01)$
$\phi_{II}\&ECS_t$	-0.32	$(-0.50, -0.13)$	0.18	$(0.02, 0.37)$	0.08	$(-0.11, 0.27)$
$NPQ_t\&qL$	-0.26	$(-0.48, -0.05)$	0.11	$(-0.09, 0.31)$	-0.20	$(-0.41, -0.01)$
$NPQ_t\&ECS_t$	0.06	$(-0.11, 0.23)$	-0.06	$(-0.23, 0.11)$	-0.50	$(-0.69, -0.32)$
$qL\&ECS_t$	-0.24	$(-0.40, -0.08)$	0.01	$(-0.20, 0.19)$	-0.22	$(-0.41, -0.03)$

- The signs of the coefficients in Tables 4.6 and 4.7 indicate the direction in which the conditional (or marginal) standard deviation and correlation between the phenotypes change with respect to predictors. Different directionality can be biologically explained as different regulatory pathways inside a photosynthesis system. For example, the estimated regression parameter for Marker 1 on the correlation between ϕ_{II} and qL is 0.29, whereas the regres-

sion parameter for Marker 2 on the exact correlation is −0.32. This explains two different relationship between ϕ_{II} and qL asserted by Marker 1 and Marker 2, respectively.

Table 4.9: Pairwise correlation estimates and 95% confidence interval in parentheses of the measured phenotypes from all genetic combinations of Marker 1 and Marker 2 from the cowpea dataset under $Control$ temperature.

	$AAAA$		$AABB$		$BBAA$		$BBBB$	
$\phi_{II}\&NPQ_t$	−0.59	(−0.70, −0.45)	−0.55	(−0.68, −0.37)	−0.72	(−0.78, −0.64)	−0.69	(−0.78, −0.57)
$\phi_{II}\&qL$	0.72	(0.63, 0.79)	0.64	(0.53, 0.72)	0.78	(0.73, 0.83)	0.71	(0.64, 0.78)
$\phi_{II}\&ECS_t$	0.08	(−0.11, 0.27)	0.27	(0.08, 0.43)	−0.24	(−0.38, −0.09)	−0.06	(−0.26, 0.15)
$NPQ_t\&qL$	0.17	(−0.06, 0.39)	0.28	(0.13, 0.42)	−0.10	(−0.23, 0.04)	0.02	(−0.19, 0.23)
$NPQ_t\&ECS_t$	0.39	(0.22, 0.54)	0.33	(0.11, 0.52)	0.44	(0.34, 0.54)	0.38	(0.18, 0.55)
$qL\&ECS_t$	0.47	(0.29, 0.62)	0.46	(0.33, 0.58)	0.24	(0.04, 0.42)	0.23	(0.01, 0.43)

Table 4.10: Pairwise correlation estimates and 95% confidence interval in parentheses of the measured phenotypes from all genetic combinations of Marker 1 and Marker 2 from the cowpea dataset under Low temperature.

	$AAAA$		$AABB$		$BBAA$		$BBBB$	
$\phi_{II}\&NPQ_t$	−0.77	(−0.84, −0.68)	−0.75	(−0.81, −0.67)	−0.85	(−0.90, −0.78)	−0.83	(−0.88, −0.77)
$\phi_{II}\&qL$	0.64	(0.53, 0.73)	0.54	(0.40, 0.65)	0.72	(0.63, 0.79)	0.63	(0.51, 0.73)
$\phi_{II}\&ECS_t$	0.16	(−0.04, 0.36)	0.34	(0.18, 0.49)	−0.16	(−0.31, 0.03)	0.03	(−0.16, 0.21)
$NPQ_t\&qL$	−0.04	(−0.25, 0.18)	0.03	(−0.12, 0.18)	−0.30	(−0.42, −0.16)	−0.19	(−0.38, −0.02)
$NPQ_t\&ECS_t$	−0.19	(−0.30, −0.09)	−0.18	(−0.31, −0.05)	−0.05	(−0.24, 0.15)	−0.03	(−0.25, 0.19)
$qL\&ECS_t$	0.26	(0.04, 0.46)	0.26	(0.04, 0.45)	0.01	(−0.15, 0.16)	0.01	(−0.23, 0.23)

- Using the results presented in Tables 4.6 and 4.7, we estimated correlations among the different pairs of phenotypes and their associated 95% confidence intervals for all possible combinations of genetic variations and environmental conditions (see Table 4.9 and 4.10). This resulted in eight possible combinations, revealing biologically relevant patterns among the phenotypes. For the row corresponding to $NPQt\&ECSt$, a positive association was found under control temperature (CT), whereas a negative association was predominant under low temperature (LT). Under the control temperature, genetic variations in chromosomes 4 and 9 modulated photochemistry mainly through the qE mechanism, while under the low temperature, they modulated photochemistry predominantly through the qI mechanism. Also, under LT, we have found that the combinations AAAA and AABB produced negative correlations between $NPQt$ and $ECSt$, while the combinations BBAA and BBBB resulted in uncorrelated $NPQt$ and $ECSt$. This suggests that the genetic variations of chromosome 4 are

more likely to modulate the qI mechanism. To illustrate further, we looked into the estimated correlations among $NPQt$ and qL for the different groups under the low temperature. For the AAAA and AABB combinations, $NPQt$ and qL were uncorrelated, but for the BBAA and BBBB combinations, $NPQt$ and qL were negatively correlated. This can be explained as follows. If the downstream processes are blocked, electrons get accumulated in Q_A, ensuring Q_A to be more reduced (qL goes down), which increases qI. As qI builds up, the slope gradient between $NPQt$ and qL gradually increases to a point where the negative associations between $NPQt$ and qL break down. For example, under low temperatures, within AAAA and AABB, we have found no association between $NPQt$ and qL.

4.7 Discussion

Analyzing high-dimensional voluminous datasets generated by high-throughput phenotyping and genome sequencing is of paramount interest for adaptive plant breeding. However, as the process involves complex interactions among multiple traits, genotypes, and environmental variables, suitable statistical models and efficient computational techniques are required to identify appropriate mechanisms. Therefore, we have developed the CMPLE workflow (overall workflow presented in Figure B.1) to bridge the gap in the phenotype-genotype-environment association studies by exploiting the correlation structure among phenotypes based on genetic and environmental variables. This is an important step toward solving different applications arising from the integration of multi-omics datasets.

Standard quantitative genomics experiments aim to determine what genetic variations contribute to individual phenotypes. On the other hand, interactions among various phenotypes signify *pleiotropy*, i.e., markers having multi-trait effects. Our method, CMPLE, is possibly the first tool in the quantitative genetics literature that explains pleiotropy by incorporating the pairwise correlations of multiple traits. The proposed methodology helps recover pertinent information regarding different regulatory pathways associated with genetic variations.

With our experimental data on photosynthesis, we have explained a possible hypothesis that

genetic variations alone are not responsible for photodamage. Instead, they condition the photosynthetic system to respond differently, favoring photoprotection or photodamage. With the given population, we identified a trade-off in photosynthesis machinery between these two processes being regulated through combinations of genetic and environmental predictors. We sparse out the genetic marker effects of specific SNPs in chromosomes 4 and 9, which under the control temperature favor photoprotective mechanisms, whereas, under low temperature, they are more consistent with modulation of qI. Also, under low-temperature conditions, where the photosynthetic machinery favors the qI mechanism, we have identified the correlations between $NPQt$ and qL to be changing with specific genetic configurations, e.g., uncorrelated $NPQt$ and qL (for combinations AAAA and AABB) and negatively correlated $NPQt$ and qL (for combinations BBAA and BBBB). This provides evidence for a subsequent hypothesis that the gradient between $NPQt$ and qL can further modulate the qI mechanism under low temperatures. This can be interpreted as one example of reflecting *epistasis*, where multiple genetic components interact in complex ways to further modulate regulatory pathways inside a system.

CHAPTER 5

IMPACT OF THIS DISSERTATION

In this thesis, we have proposed novel statistical machine learning methodology and computationally efficient algorithms, tools, and techniques to detect genetic markers for multivariate phenotypes and estimate the network structure from high-dimensional genomics datasets in multidisciplinary research. Through the analysis of "massive" datasets consisting of multiple phenotypes and many genetic markers we were to reveal new insights into how genetic diversity may have tuned biological processes to enhance fitness under diverse conditions.

The thesis tackled numerous applications from the perspective of plant physiology and genetics as a whole. In a nutshell, we have explored model-based clustering tools to identify environmental conditions affecting different phenotypes and assessed their interactions to reveal a new limiting behavior that plants adapt to in the real world. We provided a comparison of different statistical tools for genome-enabled analysis. Next, we implemented Bayesian latent factor analysis to discover and test possible mechanistic bases of such variations by assessing cosegregation (or lack thereof) between genetic diversity and multiple traits. We found that these latent factors under appropriate conditions represent the physical modes of interactions among phenotypes, which led to the identification of quantitative trait loci (QTLs), i.e., genetic polymorphisms altering the co-regulatory network among phenotypes. A significant conclusion from our work is that standard QTL mapping on individual traits fails to address the associations between multivariate phenotypes. One should model the interactions/ correlations among phenotypes through genetic markers to affirm meaningful biological mechanisms. To this end, we proposed to model the correlations among multiple complex phenotypes as a function of genetic and environmental explanatory variables (weighted graph estimation through correlation regression model). We have developed the “state-of-the-art” estimation methodology called Correlation Modeling under Pairwise Likelihood Estimation (CMPLE), aided by a novel Minorize-Maximize (MM) algorithm, and provided a technique for statistical inference.

In-plant breeding, a key aspect is to evaluate the genetic merit of candidate markers for artificial selection and predict the expected yield for phototrophs. Using CMPLE, we can provide genome-enabled predictions for correlation between multiple traits. In practice, plant breeders can use our tool to screen plants to detect the participation of distinct response mechanisms in different species, under diverse environments, at different development stages. Further, it can guide the breeding of varieties with improved responses to other environmental conditions, most notably for application to climate-resilient agriculture.

We want to stress that mean regression modeling/analysis is not a substitute for correlation modeling/analysis. These two provide different aspects of association in the phenotype-genotype space. Therefore, irrespective of whether we work with the residual responses (residual obtained after regressing the phenotypes on the candidate genes) or the mean zero responses (obtained after subtracting the crude means from respective phenotypes), the results of the correlation analysis remain somewhat unchanged. This work represents a significant advance in modeling pairwise correlation and standard deviation in terms of predictor variables. The modeling is also accompanied by a novel estimation technique that boosts the optimization problem involving many parameters. Besides methodology development, we have shown that joint inference of standard deviations and correlations among phenotypes can be used to test co-segregation of genetically-resolved association between different traits and improves the precision of phenotype network structure (Figure: B.2 and B.3)

The approach can be extended for different applications. The focus of this paper was purely on the modeling of the correlation and standard deviation. This can be relaxed by modeling both the mean and variance-covariance and developing problem-specific MM algorithms and minorizing functions. The current proof-of-concept approach was developed for a moderate number of predictors. Generally, a regularized estimation is recommended for many predictors, and creating a statistical method for a regularized analysis of the correlation will be an exciting topic of future research. Another possible way for extending our work is through the simultaneous selection (of genetic predictors) and estimation of pairwise correlation in the context of high-dimensional

datasets.

In a nutshell, this dissertation has argued uniqueness of photosynthetic mechanisms under abiotic stress (heat and cold temperatures). Next, we have demonstrated new genetic controls by incorporating the interactions between biological traits. Finally, we have offered novel statistical methodology and computationally efficient algorithm: CMPLE for Multi-omics platforms. All of these taken together can be used for creating climate adaptive plants for the betterment of mankind.

This work was supported by the DOE Office of Science, Basic Energy Sciences under Awards DEFG02-91ER20021 and DE-SC0007101 and NSF-DMS 1945824 and 1924724.

APPENDICES

APPENDIX A

SUPPLEMENT FOR PHENOME-BY-GENOME-BY-ENVIRONMENT INTERACTIONS AND THE SCOPE OF DATA SCIENCE

A.1 Clustering on the Light-potential experiment

The experiment, examining light-induced changes in chlorophyll fluorescence and absorbance changes at ambient photosynthetically- active radiation (PAR), following 10s of PAR equivalent to full sunlight, and following 10s of darkness, yields an estimate of the rapid Light Potentials of linear electron flow (LEF), nonphotochemical quenching (NPQ) and related processes. (Figure: A.1)

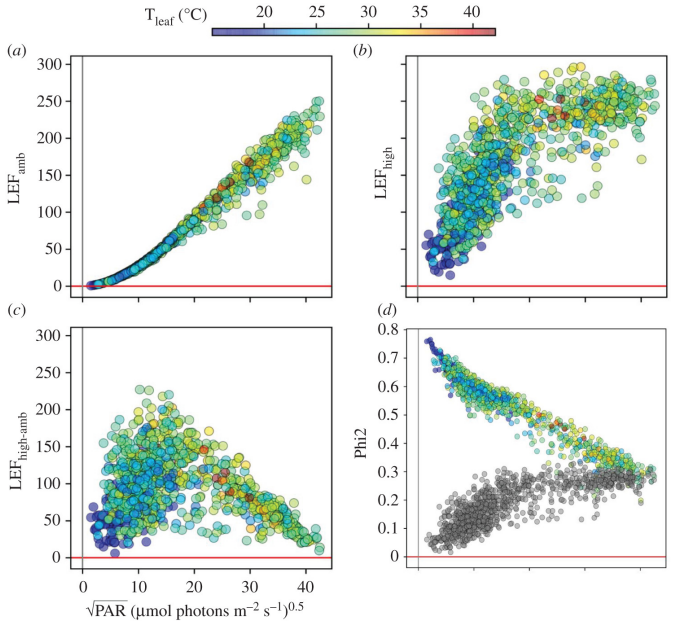

Figure A.1: Light and temperature effects on LEF and photosystem II quantum efficiency (ϕ_{II}). Each parameter was plotted as a function of the square root of the ambient photosynthetically active radiation (PARamb, X-axis) and leaf temperature (Tleaf, coloration of points). (a) Dependencies of LEF measured at PARamb; (b) LEF measured at 10 s high light (LEF_{high}); (c) the high light-induced differences in LEF ($LEF_{high-amb}$); (d) the PSII quantum efficiencies measured under ambient PAR ($Phi2_{amb}$, points coloured by Tleaf) and at 10 s high light ($Phi2_{high}$, grey points).

As shown in Figure A.2, GMM analysis of LEFamb, PARamb, and Tleaf, found six distinct, compact clusters that differed in the mode of interaction among the photosynthetic and environmen-

tal parameters. Encompassing points with lower PARamb showed moderate (Cluster 5) to strong (Clusters 1,2, and 4) dependence of LEFamb on PARamb, with little contributions from Tleaf.

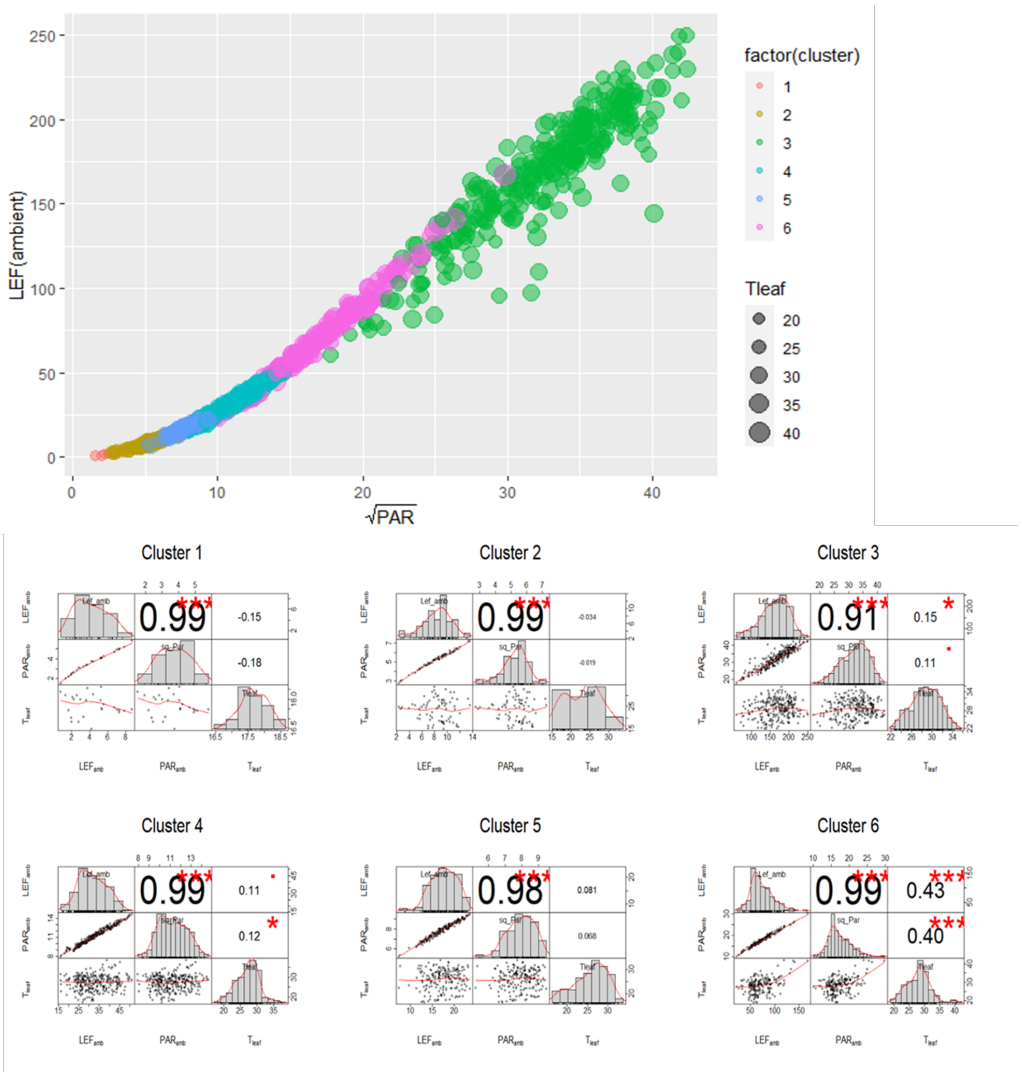

Figure A.2: Gaussian Mixture Model (GMM) clustering of LEFamb (Panel A) and correlation matrixes between LEFamb, PARamb and leaf temperature (Tleaf) for each cluster (Panel B).

By contrast, two clusters (3 and 6), which included points at higher PARamb, showed substantial dependencies on both PARamb and Tleaf. These results are consistent with LEF being predominantly light-limited at low ambient PAR but increasingly limited by temperature-dependent processes at higher PAR. These two cluster classes indicate that PARamb and Tleaf are likely to affect LEFamb in independent ways. The fact that the shapes of the clusters were not determined with individual slicing under the individual parameters for PARamb and Tleaf but with a co-dependence

on both PARamb and Tleaf suggests that, under some conditions, these effects interact, e.g., Tleaf may affect the dependence of LEFamb on PARamb.

GMM identified five distinct clusters for interactions among LEFhigh, PARamb, and Tleaf (Figure A.3). In contrast to the results on LEFamb, sets at lower PARamb (1, 2, and 4) showed LEFhigh dependencies on both Tleaf and PARamb, while Cluster 3 showed correlations with Tleaf but not with PARamb. The stronger dependence on Tleaf of LEFhigh compared to LEFamb implies that the exposure to high light revealed additional rate limitations in LEFhigh that were more strongly controlled by both Tleaf and PARamb and that, at least under some conditions, these effects were independent of each other.

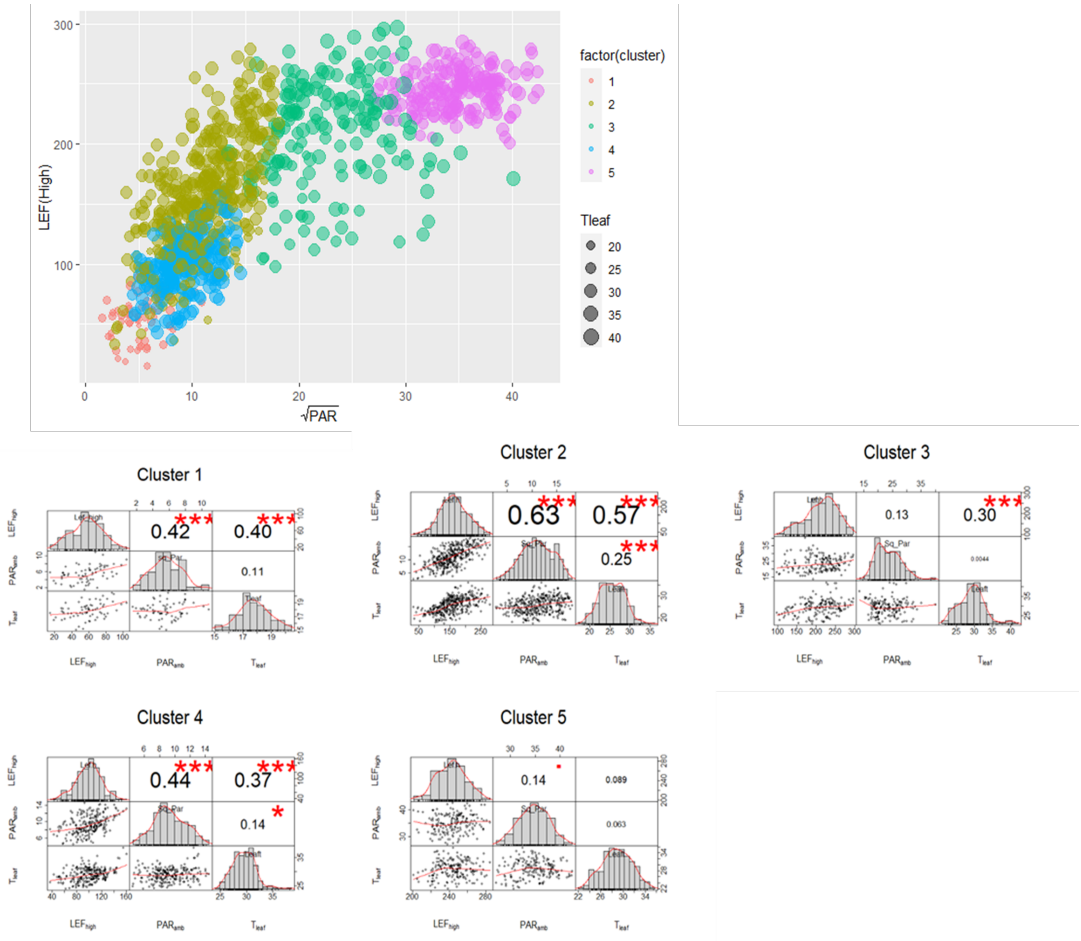

Figure A.3: Gaussian Mixture Model (GMM) clustering of LEFhigh (Panel A) and correlation matrixes between LEFhigh, PARamb and leaf temperature (Tleaf) for each cluster (Panel B).

APPENDIX B

SUPPLEMENT FOR CMPLE TO DECODE PHOTOSYNTHESIS USING THE MINORIZE-MAXIMIZE ALGORITHM

B.1 CMPLE workflow

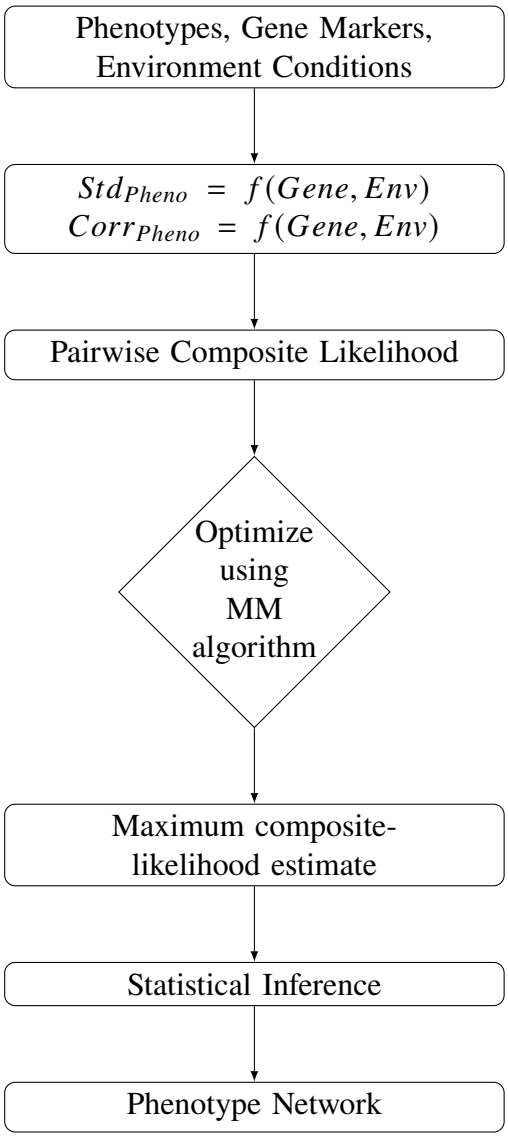

Figure B.1: Correlation Modeling Under Pairwise Likelihood Estimation (CMPLE) workflow

B.2 Additional Simulation

We have also performed an additional simulation with n= 1000, p= 10, and q= 4. The total number of parameters estimated here is 110.

Table B.1: Simulation results for α parameters under scenario 4 with n = 1000, p = 10, q = 4. All entries except for the true parameter values of the table are multiplied by 100. Par: Parameter, SD: standard deviation, DOP: direct optimization, MM: minorize-maximize

		DOP		MM				DOP		MM	
Par	True	Bias	SD	Bias	SD	Par	True	Bias	SD	Bias	SD
$\alpha_{1,0}$	−1.0	−0.2	9.2	−0.3	9.2	$\alpha_{3,0}$	0.0	−0.3	9.8	−0.3	9.8
$\alpha_{1,1}$	1.0	0.6	6.0	0.6	6.0	$\alpha_{3,1}$	0.3	0.8	6.2	0.8	6.2
$\alpha_{1,2}$	0.2	0.6	6.5	0.7	6.5	$\alpha_{3,2}$	0.1	−0.2	7.3	−0.2	7.3
$\alpha_{1,3}$	−0.4	0.3	6.5	0.3	6.5	$\alpha_{3,3}$	0.0	0.4	6.3	0.4	6.3
$\alpha_{1,4}$	0.3	−0.5	7.2	−0.5	7.2	$\alpha_{3,4}$	−0.2	−0.2	6.4	−0.1	6.4
$\alpha_{1,5}$	0.0	0.0	6.6	0.0	6.6	$\alpha_{3,5}$	0.1	0.2	5.8	0.2	5.8
$\alpha_{1,6}$	−0.5	0.5	5.4	0.6	5.4	$\alpha_{3,6}$	−0.2	−0.6	6.6	−0.6	6.6
$\alpha_{1,7}$	−0.3	0.5	7.0	0.5	6.9	$\alpha_{3,7}$	1.0	0.3	6.7	0.3	6.7
$\alpha_{1,8}$	0.1	−0.3	6.3	−0.2	6.3	$\alpha_{3,8}$	−1.0	0.4	5.5	0.4	5.5
$\alpha_{1,9}$	−0.4	0.6	6.5	0.6	6.5	$\alpha_{3,9}$	−0.1	1.0	6.1	1.0	6.1
$\alpha_{1,10}$	1.0	0.3	6.7	0.3	6.7	$\alpha_{3,10}$	0.2	0.1	6.9	0.2	6.9
$\alpha_{2,0}$	2.0	1.6	9.1	1.6	9.1	$\alpha_{4,0}$	−1.0	0.0	9.1	0.0	9.1
$\alpha_{2,1}$	0.2	−0.7	5.4	−0.7	5.4	$\alpha_{4,1}$	−0.5	1.3	5.7	1.2	5.7
$\alpha_{2,2}$	−0.5	0.0	7.4	0.1	7.4	$\alpha_{4,2}$	0.0	0.3	6.0	0.3	5.9
$\alpha_{2,3}$	−0.4	−0.3	6.3	−0.3	6.3	$\alpha_{4,3}$	0.3	1.2	6.0	1.2	6.0
$\alpha_{2,4}$	0.0	−0.9	5.9	−0.9	5.9	$\alpha_{4,4}$	−0.2	0.1	7.4	0.1	7.4
$\alpha_{2,5}$	0.3	−0.4	5.9	−0.4	5.9	$\alpha_{4,5}$	−0.2	0.2	6.4	0.2	6.4
$\alpha_{2,6}$	0.0	0.0	6.1	0.0	6.1	$\alpha_{4,6}$	0.2	0.0	5.7	0.0	5.7
$\alpha_{2,7}$	0.3	0.4	5.8	0.4	5.8	$\alpha_{4,7}$	0.0	−0.2	6.8	−0.1	6.8
$\alpha_{2,8}$	0.2	0.6	6.1	0.6	6.1	$\alpha_{4,8}$	0.0	0.8	5.9	0.8	5.8
$\alpha_{2,9}$	−0.4	−0.2	6.3	−0.2	6.3	$\alpha_{4,9}$	0.6	−0.7	6.2	−0.7	6.3
$\alpha_{2,10}$	0.0	0.2	5.5	0.2	5.6	$\alpha_{4,10}$	−0.5	−0.5	6.6	−0.5	6.6

Table B.2: Simulation results for δ parameters under scenario 4 with $n = 1000$, $p = 10$, $q = 4$. All entries except for the true parameter values of the table are multiplied by 100. Par: Parameter, SD: standard deviation, DOP: direct optimization, MM: minorize-maximize

		DOP		MM				DOP		MM	
Par	True	Bias	SD	Bias	SD	Par	True	Bias	SD	Bias	SD
$\delta_{1,2,0}$	0.2	0.1	30.3	−0.2	29.9	$\delta_{2,3,0}$	0.0	0.8	30.6	0.6	30.5
$\delta_{1,2,1}$	0.5	0.4	19.0	0.5	19.0	$\delta_{2,3,1}$	0.0	−4.6	19.3	−4.5	19.4
$\delta_{1,2,2}$	0.0	2.2	18.8	2.3	18.7	$\delta_{2,3,2}$	−0.2	−0.1	19.3	−0.1	19.3
$\delta_{1,2,3}$	−0.7	−2.3	18.2	−2.2	18.2	$\delta_{2,3,3}$	0.0	−0.2	19.4	−0.2	19.5
$\delta_{1,2,4}$	0.3	1.1	17.5	1.2	17.5	$\delta_{2,3,4}$	0.0	−0.1	18.8	−0.1	18.9
$\delta_{1,2,5}$	0.0	−3.4	17.3	−3.4	17.5	$\delta_{2,3,5}$	0.0	−3.6	17.7	−3.6	17.7
$\delta_{1,2,6}$	−0.8	0.3	20.8	0.3	20.9	$\delta_{2,3,6}$	0.0	5.3	19.9	5.4	19.9
$\delta_{1,2,7}$	0.1	5.2	19.7	5.2	19.7	$\delta_{2,3,7}$	−0.3	2.8	19.3	2.6	19.5
$\delta_{1,2,8}$	0.0	−0.4	15.5	−0.4	15.5	$\delta_{2,3,8}$	0.2	−3.8	16.3	−3.6	16.4
$\delta_{1,2,9}$	0.1	−1.5	19.9	−1.5	20.0	$\delta_{2,3,9}$	0.0	0.1	20.3	0.2	20.3
$\delta_{1,2,10}$	0.2	−1.1	20.7	−0.9	20.7	$\delta_{2,3,10}$	0.0	−1.0	20.1	−0.9	20.1
$\delta_{1,3,0}$	0.5	1.8	30.3	1.4	29.9	$\delta_{2,4,0}$	0.3	0.5	30.6	0.4	30.8
$\delta_{1,3,1}$	0.2	−0.5	20.9	−0.4	21.0	$\delta_{2,4,1}$	0.0	−1.8	19.5	−1.9	19.6
$\delta_{1,3,2}$	0.5	5.1	17.7	5.2	17.7	$\delta_{2,4,2}$	1.0	−2.9	19.9	−3.0	19.7
$\delta_{1,3,3}$	0.0	−4.5	19.3	−4.3	19.2	$\delta_{2,4,3}$	0.4	0.2	18.0	0.3	17.8
$\delta_{1,3,4}$	0.2	1.9	21.0	2.1	21.0	$\delta_{2,4,4}$	−0.3	−1.5	15.6	−1.6	15.7
$\delta_{1,3,5}$	−0.6	−5.9	19.1	−5.8	19.2	$\delta_{2,4,5}$	0.5	2.9	17.6	3.0	17.5
$\delta_{1,3,6}$	0.0	−3.4	17.8	−3.3	17.7	$\delta_{2,4,6}$	0.3	0.5	19.9	0.6	20.1
$\delta_{1,3,7}$	−0.3	0.0	19.2	0.1	19.2	$\delta_{2,4,7}$	0.1	1.9	18.4	1.9	18.4
$\delta_{1,3,8}$	0.2	−3.0	18.8	−2.9	18.8	$\delta_{2,4,8}$	−0.6	−2.1	17.6	−1.9	17.5
$\delta_{1,3,9}$	0.4	2.1	18.4	2.3	18.2	$\delta_{2,4,9}$	0.6	−1.6	18.4	−1.5	18.3
$\delta_{1,3,10}$	−0.2	2.0	18.2	2.0	18.2	$\delta_{2,4,10}$	−0.1	1.0	17.3	0.8	17.1
$\delta_{1,4,0}$	1.0	1.2	30.3	1.1	29.9	$\delta_{3,4,0}$	−1.0	2.7	30.4	3.2	30.0
$\delta_{1,4,1}$	0.2	0.0	18.9	−0.1	18.9	$\delta_{3,4,1}$	0.2	−4.9	19.4	−5.1	19.1
$\delta_{1,4,2}$	−0.5	4.5	21.1	4.3	21.0	$\delta_{3,4,2}$	−0.2	0.1	19.4	−0.2	19.5
$\delta_{1,4,3}$	−0.2	−0.4	20.2	−0.3	20.0	$\delta_{3,4,3}$	−0.1	0.0	18.2	−0.1	18.2
$\delta_{1,4,4}$	0.5	3.1	16.5	3.3	16.7	$\delta_{3,4,4}$	0.2	−3.1	18.0	−3.2	18.0
$\delta_{1,4,5}$	0.1	−0.5	19.0	−0.5	19.2	$\delta_{3,4,5}$	0.6	1.1	19.4	1.0	19.4
$\delta_{1,4,6}$	−0.1	−0.4	20.8	−0.4	21.1	$\delta_{3,4,6}$	−0.1	0.1	21.6	0.1	21.5
$\delta_{1,4,7}$	−0.1	−1.2	20.5	−1.1	20.5	$\delta_{3,4,7}$	0.0	1.3	19.9	0.9	19.4
$\delta_{1,4,8}$	0.0	1.5	17.9	1.5	17.9	$\delta_{3,4,8}$	0.2	−1.0	20.3	−0.8	20.0
$\delta_{1,4,9}$	−0.1	−3.5	18.6	−3.5	18.6	$\delta_{3,4,9}$	0.3	−0.9	20.3	−0.9	20.1
$\delta_{1,4,10}$	−0.2	0.3	19.1	0.4	19.1	$\delta_{3,4,10}$	−0.2	1.6	19.0	1.4	18.8

B.3 CMPLE application on Heat Stress treatments

Using CMPLE on the DHS dataset based on the two selected SNP markers in chromosome 2 we have found distinguishable correlation pattern among the selected set of phenotypes.

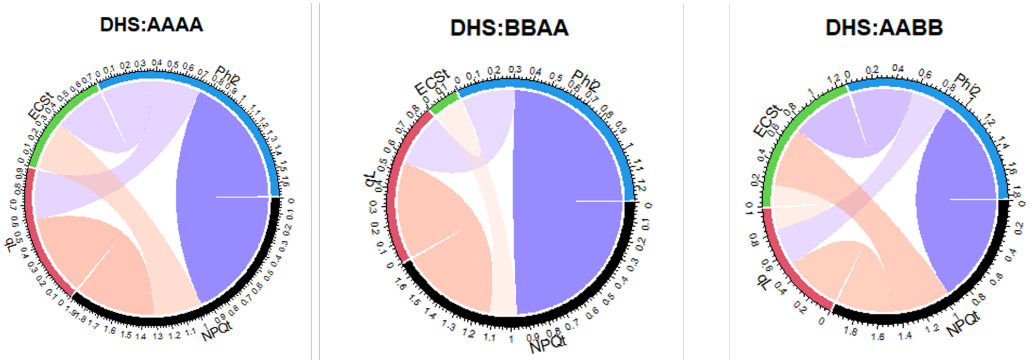

Figure B.2: Correlation network under DHS

Again with the genetic configuration of BBAA for the two selected SNPs under DHS and LHS, we can identify distinct phenotypic network under the various condions specified.

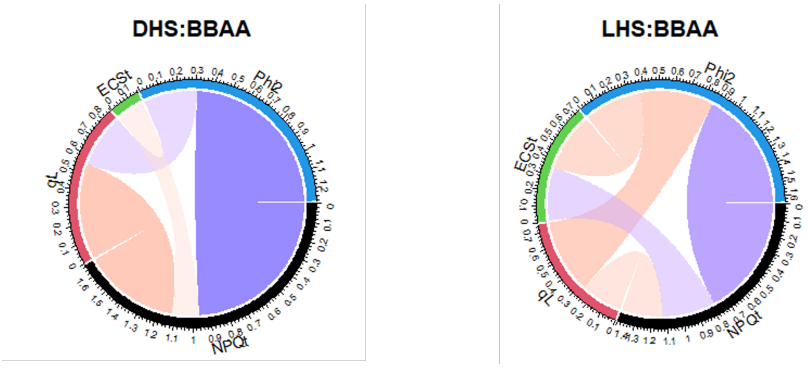

Figure B.3: Correlation network under DHS and LHS

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] R. E. Blankenship, *Molecular Mechanisms of Photosynthesis*. John Wiley & Sons, 2021.

[2] D. M. Kramer, G. Johnson, O. Kiirats, and G. E. Edwards, "New fluorescence parameters for the determination of qa redox state and excitation energy fluxes," *Photosynthesis research*, vol. 79, no. 2, pp. 209–218, 2004.

[3] T. J. Avenson, J. A. Cruz, A. Kanazawa, and D. M. Kramer, "Regulating the proton budget of higher plant photosynthesis," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 9709–9713, Jul 2005.

[4] A. Kanazawa, A. Chattopadhyay, S. Kuhlgert, H. Tuitupou, T. Maiti, and D. M. Kramer, "Light potentials of photosynthetic energy storage in the field: what limits the ability to use or dissipate rapidly increased light energy?," *R. Soc. Open Sci.*, vol. 8, p. 211102, 2021.

[5] N. Keren and A. Krieger-Liszkay, "Photoinhibition: molecular mechanisms and physiological significance," *Physiol Plant.*, vol. 142, no. 1, pp. 1–5, 2011.

[6] E. Tyystjärvi, "Photoinhibition of photosystem II," *Int Rev Cell Mol Biol.*, vol. 300, pp. 243–303, 2013.

[7] B. Demmig-Adams, "Carotenoids and photoprotection in plants: a role for the xanthophyll zeaxanthin," *Biochim Biophys Acta.*, vol. 1020, no. 1, pp. 1–24, 1990.

[8] K. K. Niyogi, O. Björkman, and A. R. Grossman, "The roles of specific xanthophylls in photoprotection," *Proc. Natl. Acad. Sci. USA*, vol. 94, no. 25, pp. 14162–14167, 1997.

[9] S. Tietz, C. C. Hall, J. A. Cruz, and D. M. Kramer, "$NPQ_{(T)}$: a chlorophyll fluorescence parameter for rapid estimation and imaging of non-photochemical quenching of excitons in photosystem-ii-associated antenna complexes," *Plant Cell Environ.*, vol. 40, pp. 1243–1255, 2017.

[10] N. R. Baker, J. Harbinson, and D. M. Kramer, "Determining the limitations and regulation of photosynthetic energy transduction in leaves," *Plant, Cell & Environment*, vol. 30, no. 9, pp. 1107–1125, 2007.

[11] D. Hoh, I. Osei-Bonsu, A. Chattopadhyay, *et al.*, "Genetic variation in photosynthetic responses to chilling modulates proton motive force, cyclic electron flow and photosystem ii photoinhibition." https://doi.org/10.22541/au.163422290.08126533/v1, October 14 2021. *Authorea*, October 14, 2021.

[12] J. A. Raven, "The cost of photoinhibition," *Physiol Plant.*, vol. 142, no. 1, pp. 87–104, 2011.

[13] D. M. Kramer, J. A. Cruz, and A. Kanazawa, "Balancing the central roles of the thylakoid proton gradient," *Trends Plant Sci.*, vol. 8, no. 1, pp. 27–32, 2003.

[14] T. H. Meuwissen, B. J. Hayes, and M. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[15] J. Crossa, G. d. l. Campos, P. Pérez, D. Gianola, J. Burgueno, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker, J. Yan, *et al.*, "Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers," *Genetics*, vol. 186, no. 2, pp. 713–724, 2010.

[16] P. VanRaden, C. Van Tassell, G. Wiggans, T. Sonstegard, R. Schnabel, J. Taylor, and F. Schenkel, "Invited review: Reliability of genomic predictions for north american holstein bulls," *Journal of dairy science*, vol. 92, no. 1, pp. 16–24, 2009.

[17] P. M. Visscher, J. Yang, and M. E. Goddard, "A commentary on 'common snps explain a large proportion of the heritability for human height'by yang et al.(2010)," *Twin Research and Human Genetics*, vol. 13, no. 6, pp. 517–524, 2010.

[18] G. de Los Campos, J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus, "Whole-genome regression and prediction methods applied to plant and animal breeding," *Genetics*, vol. 193, no. 2, pp. 327–345, 2013.

[19] R. Bellman, "On the approximation of curves by line segments using dynamic programming," *Communications of the ACM*, vol. 4, no. 6, p. 284, 1961.

[20] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[21] G. Wu, Y. Xie, H. Chen, M. Zhong, R. Liu, B. Shi, Q. Li, X. Wang, T. Wu, Y. Yan, *et al.*, "Superconductivity at 56 k in samarium-doped srfeasf," *Journal of Physics: Condensed Matter*, vol. 21, no. 14, p. 142203, 2009.

[22] I. Lemhadri, F. Ruan, and R. Tibshirani, "Lassonet: Neural networks with feature sparsity," in *International Conference on Artificial Intelligence and Statistics*, pp. 10–18, PMLR, 2021.

[23] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[24] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[25] D. Habier, R. L. Fernando, K. Kizilkaya, and D. J. Garrick, "Extension of the bayesian alphabet for genomic selection," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–12, 2011.

[26] D. Gianola, S. Van Petegem, M. Legros, S. Brandstetter, H. Van Swygenhoven, and K. Hemker, "Stress-assisted discontinuous grain growth and its effect on the deformation behavior of nanocrystalline aluminum thin films," *Acta Materialia*, vol. 54, no. 8, pp. 2253–2263, 2006.

[27] L. Janss, G. de Los Campos, N. Sheehan, and D. Sorensen, "Inferences from genomic models in stratified populations," *Genetics*, vol. 192, no. 2, pp. 693–704, 2012.

[28] P. Pérez and G. de Los Campos, "Genome-wide regression and prediction with the bglr statistical package," *Genetics*, vol. 198, no. 2, pp. 483–495, 2014.

[29] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[30] J. Vendrow, J. Haddock, E. Rebrova, and D. Needell, "On a guided nonnegative matrix factorization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3265–32369, IEEE, 2021.

[31] H. H. Harman, *Modern factor analysis*. University of Chicago press, 1976.

[32] R. J. Rummel, "Understanding factor analysis," *Journal of conflict resolution*, vol. 11, no. 4, pp. 444–480, 1967.

[33] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[34] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, pp. 1247–1255, PMLR, 2013.

[35] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, 2008.

[36] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, *et al.*, "Sc3: consensus clustering of single-cell rna-seq data," *Nature methods*, vol. 14, no. 5, pp. 483–486, 2017.

[37] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, M. Ladanyi, and C. Sander, "Integrative subtype discovery in glioblastoma using icluster," *PloS one*, vol. 7, no. 4, p. e35236, 2012.

[38] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *The Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, 2006.

[39] S. Virtanen, A. Klami, and S. Kaski, "Bayesian cca via group sparsity," in *ICML*, 2011.

[40] A. Klami, G. Bouchard, and A. Tripathi, "Group-sparse embeddings in collective matrix factorization," *arXiv preprint arXiv:1312.5921*, 2013.

[41] S. Virtanen, A. Klami, S. Khan, and S. Kaski, "Bayesian group factor analysis," in *Artificial Intelligence and Statistics*, pp. 1269–1277, PMLR, 2012.

[42] A. Klami, "Polya-gamma augmentations for factor models," in *Asian Conference on Machine Learning*, pp. 112–128, PMLR, 2015.

[43] X. Zhuang, Z. Yang, K. R. Sreenivasan, V. R. Mishra, T. Curran, R. Nandy, and D. Cordes, "Multivariate group-level analysis for task fmri data with canonical correlation analysis," *NeuroImage*, vol. 194, pp. 25–41, 2019.

[44] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[45] G. R. Cramer, K. Urano, S. Delrot, M. Pezzotti, and K. Shinozaki, "Effects of abiotic stress on plants: a systems biology perspective," *BMC Plant Biol*, vol. 11, no. 1, pp. 1–14, 2011.

[46] J. A. Cruz, L. J. Savage, R. Zegarac, C. C. Hall, M. Satoh-Cruz, G. A. Davis, W. K. Kovac, J. Chen, and D. M. Kramer, "Dynamic environmental photosynthetic imaging reveals emergent phenotypes," *Cell Syst.*, vol. 2, no. 6, pp. 365–377, 2016.

[47] S. Kuhlgert, G. Austic, R. Zegarac, I. Osei-Bonsu, D. Hoh, M. I. Chilvers, M. G. Roth, K. Bi, D. TerAvest, P. Weebadde, and D. M. Kramer, "MultispeQ Beta: a tool for large-scale plant phenotyping connected to the open PhotosynQ network," *R. Soc. Open Sci.*, vol. 3, p. 160592, 2016.

[48] M. Ritchie, E. Holzinger, R. Li, S. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nat Rev Genet*, vol. 16, pp. 85–97, 02 2015.

[49] D. Gianola and R. L. Fernando, "A multiple-trait bayesian lasso for genome-enabled analysis and prediction of complex traits," *Genetics*, vol. 214, no. 2, pp. 305–331, 2020.

[50] Y. Jia and J.-L. Jannink, "Multiple-trait genomic selection methods increase genetic value prediction accuracy," *Genetics*, vol. 192, no. 4, pp. 1513–1522, 2012.

[51] T. E. Galesloot, K. Van Steen, L. A. Kiemeney, L. L. Janss, and S. H. Vermeulen, "A comparison of multivariate genome-wide association methods," *PLoS ONE*, vol. 9, no. 4, p. e95923, 2014.

[52] E. Schadt, J. Lamb, X. Yang, and et al., "An integrative genomics approach to infer causal associations between gene expression and disease," *Nat Genet*, vol. 37, pp. 710—717, 2005.

[53] D. C. Kulp and M. Jagalur, "Causal inference of regulator-target pairs by gene mapping of expression phenotypes," *BMC Genomics*, vol. 7, p. 125, May 2006.

[54] F. W. Stearns, "One hundred years of pleiotropy: A retrospective," *Genetics*, vol. 186, no. 3, pp. 767–773, 2010.

[55] E. D. Schifano, L. Li, D. C. Christiani, and X. Lin, "Genome-wide association analysis for multiple continuous secondary phenotypes," *Am J Hum Genet.*, vol. 92, no. 5, pp. 744–759, 2013.

[56] M. Pourahmadi, "Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation," *Biometrika*, vol. 86, no. 3, pp. 677–690, 1999.

[57] P. D. Hoff and X. Niu, "A covariance regression model," *Stat Sin.*, vol. 22, no. 2, pp. 729–753, 2012.

[58] T. Zou, W. Lan, H. Wang, and C.-L. Tsai, "Covariance regression analysis," *J Am Stat Assoc.*, vol. 112, pp. 266–281, 2017.

[59] K. Meyer and M. Kirkpatrick, "Better estimates of genetic covariance matrices by "bending" using penalized maximum likelihood," *Genetics*, vol. 185, no. 3, p. 1097, 2010.

[60] S. Lele and M. L. Taper, "A composite likelihood approach to (co)variance components estimation," *J Stat Plan Inference*, vol. 103, no. 1, pp. 117–135, 2002.

[61] B. Gao, C. Yang, J. Liu, and X. Zhou, "Accurate genetic and environmental covariance estimation with composite likelihood in genome-wide association studies," *PLoS Genet*, vol. 17, pp. 1–25, 01 2021.

[62] Y. Bai, J. Kang, and P. X. Song, "Efficient pairwise composite likelihood estimation for spatial-clustered data," *Biometrics*, vol. 70, no. 3, pp. 661–670, 2014.

[63] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am Stat.*, vol. 58, pp. 30–37, 2004.

[64] X. Huang, J. Xu, and G. Tian, "On profile MM algorithms for Gamma frailty survival models," *Stat Sin.*, vol. 29, pp. 895–916, 2019.

[65] H. Zhou, L. Hu, J. Zhou, and K. Lange, "MM algorithms for variance components models," *J Comput Graph Stat.*, vol. 28, pp. 350–361, 2019.

[66] D. R. Hunter and L. Runze, "Variable selection using MM algorithms," *Ann Stat.*, vol. 33, pp. 1617–1642, 2005.

[67] T. Leeper, "Interpreting regression results using average marginal effects with r's margins." https://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf, 2021.

[68] A. Hugues, "A perspective on interaction effects in genetic association studies," *Genet Epidemiol.*, vol. 40, pp. 678–688, 2016.

[69] W. H. Greene, *Econometric Analysis*. New York, NY: Pearson, 1997.

[70] C. Varin, N. Reid, and D. Firth, "An overview of composite likelihood methods," *Stat Sin.*, vol. 21, pp. 5–42, 2011.

[71] W. Muchero, N. N. Diop, P. R. Bhat, R. D. Fenton, S. Wanamaker, M. Pottorff, S. Hearne, N. Cisse, C. Fatokun, J. D. Ehlers, P. A. Roberts, and T. J. Close, "A consensus genetic map of cowpea [Vigna unguiculata (L) Walp.] and synteny based on EST-derived SNPs," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 18159–18164, Oct 2009.

[72] K. W. Broman and S. Sen, *A Guide to QTL Mapping with R/qtl*. New York, NY: Springer, 2009.