Conference on Applied Statistics in Agriculture 2010 - 22nd Annual Conference Proceedings

INTRODUCTION TO SELECTING SUBSETS OF TRAITS FOR QUANTITATIVE TRAIT LOCI ANALYSIS

Tilman Achberger

James C. Fleet

David E. Salt

R. W. Doerge

See next page for additional authors

Recommended Citation

Achberger, Tilman; Fleet, James C.; Salt, David E.; and Doerge, R. W. (2010). "INTRODUCTION TO SELECTING SUBSETS OF TRAITS FOR QUANTITATIVE TRAIT LOCI ANALYSIS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1071

Author Information

Tilman Achberger, James C. Fleet, David E. Salt, and R. W. Doerge

INTRODUCTION TO SELECTING SUBSETS OF TRAITS FOR QUANTITATIVE TRAIT LOCI ANALYSIS

Tilman Achberger[1], James C. Fleet[2], David E. Salt[3], and R. W. Doerge[1,4*]

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907
[2]Department of Foods and Nutrition, Purdue University, West Lafayette, IN 47907
[3]Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907
[4]Department of Agronomy, Purdue University, West Lafayette, IN 47907
*Corresponding author: doerge@stat.purdue.edu

ABSTRACT: Quantitative trait loci (QTL) mapping is a popular statistical method that is often used in agricultural applications to identify genomic regions associated with phenotypic traits of interest. In its most common form, a QTL analysis tests one phenotypic trait at a time using a variety of research hypotheses that depend on the application. When multiple traits are available, there are considerable benefits to analyzing subsets of biologically related traits in a multiple-trait QTL mapping framework. Determining the most informative subset(s) of traits is the critical challenge that we address in this work. We present our approach, as well as simulations that demonstrate the performance. We also discuss an application of our approach as applied to an *Arabidopsis thaliana* data set.

Keywords: quantitative trait loci (QTL), sparse principal component analysis (sparse PCA), variable selection, ionomic phenotype mapping

1 Introduction

It is well understood that complex traits are rarely associated with a single gene. Instead, complex traits (e.g., yield, disease resistance, etc.) are often associated with multiple genes whose combination lends itself to the manifestation of the trait. Finding genetic determinants of these complex traits remains a fundamental challenge for the scientific community. A powerful approach for identifying regions of the genome that are associated with these traits can be achieved via quantitative trait locus (QTL) mapping, a statistical procedure that searches for associations between phenotypes (traits) and genotypes (genetic markers) in experimental populations. A number of different QTL mapping methods have been developed over the years. Simple methods such as single marker analysis (LUO and KEARSEY, 1989) independently test each trait and each genetic marker for a statistically significant association using simple linear regression. Interval mapping (IM: LANDER and BOTSTEIN, 1989; HALEY and KNOTT, 1992; JANSEN, 1993, 1994) takes advantage of additional information from a genetic map to locate QTL relative to an interval defined by two genetic markers. Composite interval mapping (CIM: ZENG, 1993, 1994) extends IM by including selected genetic markers as cofactors in the model to remove variation from other QTL outside the testing region. Some further advances have been made by accounting for the possibility of multiple QTL associated with one trait; these methods include multiple-interval-mapping (MIM: KAO *et al.*, 1999) and multiple-QTL-models (MQM: JANSEN, 1993, 1994).

While single marker analysis, IM, and CIM are all powerful methods for detecting and locating QTL, they are limited to the analysis of independent traits, even though it is well known that traits are often not independent. When multiple (non-independent) traits are analyzed jointly in multiple-trait QTL mapping procedures (JIANG and ZENG, 1995; KOROL *et al.*, 1995, 1998; HENSHALL and GODDARD, 1999; KNOTT and HALEY, 2000; HACKETT *et al.*, 2001; XU *et al.*, 2005), the power of QTL identification increases, as does the number of biologically interesting hypothesis tests that can be evaluated. Two notable tests that can be performed using joint analysis are tests for genotype-environment interactions, and tests for co-localization of QTL when compared to closely linked QTL for two or more traits.

While conducting a (joint) QTL mapping analysis on multiple traits is useful in developing a more complete and well-rounded understanding of the genetic processes underlying the traits as a whole, simply analyzing all traits together is not recommended, as the interpretation of the results is difficult, and the model often becomes over parameterized. In fact, the limiting factor in performing these types of joint analyses is the lack of direction on which traits should be analyzed together. Certainly, it is possible to rely on prior biological knowledge about related traits as the primary method for selecting the traits, but often this information is not available. We are motivated by this last scenario in which a statistical procedure is needed for selecting groups of traits having non-independent relationships among themselves.

Several different approaches can be used to define and select a desired group of traits to take forward into a joint QTL analysis. One such approach is to jointly analyze traits that have similar QTL results from their single trait analyses. Unfortunately, doing so has notable drawbacks. First, since the traits were selected based on similar QTL results, the corresponding test statistics from the joint analysis may be inflated. To overcome this, a large multiple testing correction must be applied to accommodate the traits being selected from the set of all possible combinations of traits. Second, since several possible definitions can be proposed to assess whether or not a set of traits are similar in their QTL results, variants of these definitions may result in different sets of traits being selected. Third, when traits are selected based on their QTL results alone, minor QTL results may be missed. That is, if two or more traits have only small effect QTL in the same location, but these QTL are not statistically significant based on single trait mapping, the traits would not be selected for joint mapping. Therefore, the opportunity to find small effect QTL via joint mapping would be lost. Finally, while one can imagine testing all possible combinations of trait subsets, exhaustive testing can be computationally prohibitive when the number of traits under consideration is large. Ignoring the empty set and the sets consisting of only one trait, there are $2^p - p - 1$ possible combinations of traits to consider, where p is the number of traits measured in an experiment.

Since selecting traits based upon their single trait results has many drawbacks, we instead consider the correlation structure of the trait data. A high correlation among groups of traits does not necessarily imply that they share common QTL. It does suggest, however, the possibility that the traits are biologically related in some way, and such non-independent traits may be well suited for joint analysis in multiple-trait QTL mapping procedures. In recent years, a number of sparse principal component analysis (sparse PCA) procedures have been developed by different research groups (ZOU *et al.*, 2006; SHEN and HUANG, 2008; WITTEN *et al.*, 2009; JOURNEE *et al.*, 2010) with the goal of finding linear combinations of variables that explain the maximum amount of variability in the data, while restricting the number of variables in the linear combination to be relatively small. In the context of QTL mapping, sparse PCA provides two opportunities for studying genetic determinants of the complex (non-independent) relationships between traits. First, sparse PCA selects traits (i.e., the traits receiving non-zero weight in the linear combination) that are related in some way with one another. These traits can then be analyzed in a multiple-trait QTL mapping framework to test if the traits are associated with one common QTL, or multiple linked QTL. Second, the linear combination of the selected traits provides an estimate of the unobservable latent variable associated with the co-variation in the trait data. This estimated latent variable can be considered as a new trait which can then be analyzed in single trait QTL mapping to identify genomic regions associated with the co-variation in the selected traits.

The purpose of this paper is to present a novel statistical approach for selecting groups of correlated traits for QTL analysis. Our focus will be on the sparse PCA methodology, rather than on QTL mapping analysis itself. This said, we will rely on single trait IM as the eventual QTL mapping methodology since it has the advantage of allowing straightforward comparisons between the results of individual trait mapping and those of the approach being presented.

2 Methods

2.1 Experimental Designs for QTL Mapping

QTL mapping studies can be conducted for a variety of experimental breeding designs. Common studies include backcross (BC), F_2 and recombinant inbred line (RIL) populations (see Figure 1). These designs start with two homozygous parental lines differing in a set of genetic markers, and typically differing in the quantitative trait(s) of interest. The parental lines are crossed and give rise to an F_1 generation. A backcross (BC) individual is obtained by mating an F_1 individual with one of the parental lines. An F_2 individual is obtained by self mating an F_1 individual. A RIL individual is obtained by self mating an F_2 individual, and then repeatedly self mating each subsequent offspring until approximately the F_8 generation, at which point almost all heterozygosity has been eliminated, resulting in a nearly homozygous individual. RIL populations have several advantages for QTL mapping: an increased number of recombination events, giving rise to more precise QTL localization; a nearly homozygous population eliminates dominance and focuses on only additive QTL effects; and since the RIL population is homozygous at nearly all loci, individuals can be self mated to produce genetically identical individuals, allowing replicates to be grown within an experiment, as well as allowing different laboratories to study the same population.

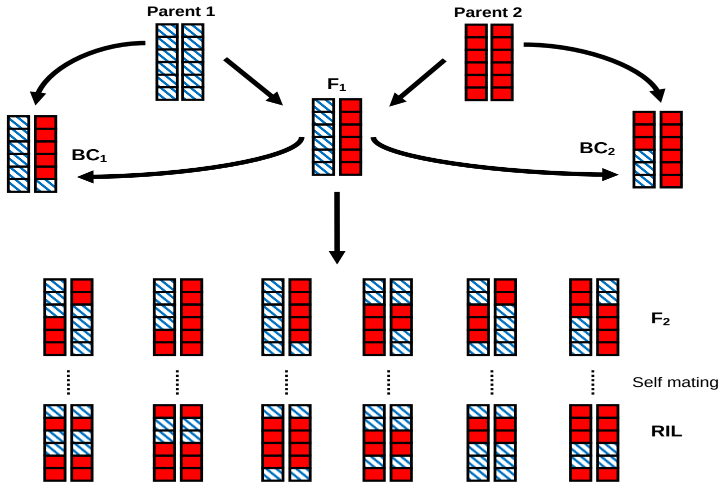


Figure 1: Breeding design for a simple diploid organism with one chromosome. The blue (striped) and red (solid) boxes represent the alleles at different genetic markers from Parent 1 and 2, respectively. Backcross (BC), F_1, F_2 and recombinant inbred line (RIL) populations are shown.

2.2 Sparse Principal Component Analysis

Principal component analysis (PCA: PEARSON, 1901; HOTELLING, 1933a,b; JOLLIFFE, 2002) is a popular statistical method for reducing the dimensionality of multivariate data by finding linear combinations of the observed variables (in this application, traits) that explain the maximum amount of variability in the data. Provided that the traits are not all uncorrelated (i.e., independent), PCA typically summarizes a large amount of the overall variability in the data using relatively few principal components (PCs). Let X denote the $n \times p$ data matrix consisting of p traits measured on n individuals, where all traits (columns) are standardized to have zero means and unit variances. Let V denote the $p \times K$ orthonormal matrix whose K columns $[V_1, \cdots, V_K]$ contain the loadings (coefficients) for the linear combinations of the p traits. PCA seeks to maximize the criterion

$$\begin{aligned} \arg\max_V \text{Var}\,(XV) \; &= \; \arg\max_V V^T X^T X V \\ \text{where} \;\; V^T V = I_{K \times K} \;\; &\text{and} \;\; \text{Var}\,(XV_1) \geq \cdots \geq \text{Var}\,(XV_K) \;\; . \end{aligned} \tag{1}$$

The matrix V in equation (1) is typically found by eigenvalue-eigenvector decomposition of $X^T X$, however this is an ill-conditioned problem when the number of traits (p) is larger than the number of individuals (n). To address this challenge, singular value decomposition (SVD) can be used to sequentially find the column vectors $[V_1, \cdots, V_K]$ of V. The SVD of X can be written as

$$\begin{aligned} X = UDV^T = U_1 d_{1,1} V_1^T + \cdots + U_K d_{K,K} V_K^T \\ \text{where} \;\; V^T V = I_{K \times K} \;, \;\; U^T U = I_{K \times K} \;\; \text{and} \;\; d_{1,1} \geq \cdots \geq d_{K,K} \end{aligned} \tag{2}$$

where X and V are as before, U is an $n \times K$ orthonormal matrix whose columns $[U_1, \cdots, U_K]$ are the standardized PC scores having zero means and unit variances, and D is a diagonal matrix whose elements $d_{i,i}$ are the standard deviation of the i^{th} PC. The i^{th} PC of the data matrix X is estimated by $XV_i = U_i d_{i,i}$. The leading PC of X is obtained by finding the best rank one approximation of X by minimizing the criterion

$$\underset{U_1, d_{1,1}, V_1}{\arg\min} \quad \left\| X - U_1 d_{1,1} V_1^T \right\|_F^2 = \underset{V_1}{\arg\min} \quad \left\| X - X V_1 V_1^T \right\|_F^2 \tag{3}$$

where the squared Frobenius norm $\| \cdot \|_F^2$ is the sum of the squared elements of a matrix. Additional loading vectors V_i for $i > 1$ are obtained by finding the best rank one approximation of the residual matrix after the previous $i - 1$ PCs have been removed. For example, V_i is found by performing SVD on $X - U_1 d_{1,1} V_1^T - \cdots - U_{i-1} d_{i-1,i-1} V_{i-1}^T$.

In the context of multiple-trait QTL mapping, LAN *et al.* (2003) and others have proposed reducing the dimensionality of the trait data by performing PCA and then analyzing each of the PCs using single trait QTL mapping. Here the PCs represent the unobserved latent variables associated with the co-variation in the traits, which may or may not be associated with QTL. One notable drawback of this procedure is that the results are often difficult to interpret since the linear combinations in PCA typically include all traits. To address this issue, we consider a variant of PCA called sparse PCA (ZOU *et al.*, 2006; SHEN and HUANG, 2008; WITTEN *et al.*, 2009; JOURNEE *et al.*, 2010) which finds linear combinations that are comprised of a small number of important traits. The sparse PCA criterion can be formulated using a number of different penalties on the loading vector V_i, including the elastic net penalty (ZOU and HASTIE, 2005) and the lasso penalty (TIBSHIRANI, 1996). Here we use the sparse PCA via regularized SVD with the lasso penalty (sPCA-rSVD-soft) criterion from SHEN and HUANG (2008). This criterion uses the rescaled singular vectors $\widetilde{V_1}$ and $\widetilde{U_1}$ where the right singular vector $\widetilde{V_1}$ is free of any scale constraint while the left singular vector is constrained to have $\|\widetilde{U_1}\|_2 = 1$ (where $\|\widetilde{V_1}\|_2 = \sum \widetilde{V_1}^2$). The sPCA-rSVD-soft criterion minimizes

$$\underset{V_1}{\arg\min} \quad \left\| X - \widetilde{U_1} \widetilde{V_1}^T \right\|_F^2 + \lambda_1 \|\widetilde{V_1}\|_1 \tag{4}$$

where $\|\widetilde{V_1}\|_1 = \sum |\widetilde{V_1}|$. After obtaining $\widetilde{V_1}$, the loading vector for sparse PCA is given by $V_1 = \widetilde{V_1} / \|\widetilde{V_1}\|_2$. Additional details on the algorithms used in the sPCA-rSVD-soft procedure are included in SHEN and HUANG (2008). The L_1 (lasso) penalty on V_1 results in the smaller magnitude loadings in V_1 (corresponding to traits that contribute little to the linear combination) being forced to zero, while keeping the larger (presumably more important) magnitude loadings as non-zero. Sparse PCA selects groups of correlated traits (the traits having non-zero weight in V_1) that are hypothesized to be associated with an unobserved latent variable, which is estimated by XV_1. When the penalty term λ_1 is large, sparse PCA retains only a small number of non-zero elements in V_1; when λ_1 = 0, sparse PCA reduces to PCA. SHEN and HUANG (2008) recommend that the degree of sparsity (number of zero loadings in V_1) be used as the tuning parameter rather than specifying the value of λ_1. Separate degrees of sparsity can be specified for each estimated PC, meaning the number of selected traits in the first PC does not necessarily equal the number in the second PC, and so on. In this publication, our interest is restricted to only the first PC (extensions will follow in later work).

2.3 Estimation of the Sparsity Parameter λ_1

A k-fold cross-validation (CV: MOSTELLER and WALLACE, 1963; STONE, 1974; WOLD, 1978) procedure is used to estimate the appropriate degree of sparsity for sparse PCA. More specifically, the rows (individuals) of X are partitioned into k equally sized groups, where $k-1$ groups are used as the training set $X_{(train)}$ and the remaining group is used as the testing set $X_{(test)}$. The columns of $X_{(train)}$ and $X_{(test)}$ are independently standardized such that each has zero mean and unit variance. The loading vector $V_{1(train)}$ is found by maximizing the sparse PCA criterion (4) on the training data $X_{(train)}$ for all degrees of sparsity from 0 (PCA) to $p-2$ (only two non-zero loadings in the linear combination). The variance of the linear combination of traits in the CV testing data is evaluated by Var $\left(X_{(test)}V_{1(train)}\right)$. The CV procedure is repeated k times, so that each of the k partitions are used as the testing set exactly once. The degree of sparsity that maximizes the variance of the linear combination of traits in the testing data for the mean of the k CVs is chosen as the degree of sparsity to be used in sparse PCA on the full data.

Performing k-fold CV can become computationally expensive when the number of individuals in the sample (n), number of folds (k), or the number of traits (p) is large. Computational complexity can be reduced by performing only one CV sample, however, the degree of sparsity estimate will generally be less precise. If computation time is not a major concern, the k-fold CV procedure can be independently repeated to obtain additional CV samples, thereby allowing more accurate estimation of the appropriate degree of sparsity. Alternatively, the k-fold CV procedure can be skipped entirely when prior biological knowledge is available on the appropriate degree of sparsity.

2.4 Mapping the Estimated Latent Variable

As previously stated, the groups of correlated traits selected by sparse PCA may be well suited for analysis in a multiple-trait QTL mapping framework. Additionally, sparse PCA provides at least one more opportunity to study the genetic determinants of the complex relationships between traits. The PC scores $XV_1, \cdots, XV_K$ from sparse PCA estimate the unobserved latent variables in the data which can be analyzed in single trait QTL mapping to identify genomic regions associated with the co-variation in the selected traits. While this is similar to the PCA approach used in LAN *et al.* (2003), the interpretation of the results will be less difficult when sparse PCA is used since the PCs are comprised of a small number of traits rather than all traits.

3 Results

3.1 Simulations

A simulation study was conducted to assess the performance of sparse PCA with respect to selecting groups of correlated traits and identifying QTL for the estimated latent variable associated with the selected traits. A RIL population was employed for this simulation. We assumed 5 chromosomes of length 100 centimorgans (cM) with marker spacing of 10cM. A single QTL was simulated at 50cM from the left end of chromosome 1 (see Figure 2). A total of 100 traits were simulated, of which 10 traits (X_1 to X_{10}) were associated with the QTL (additive effect of an allele substitution equal to 0.50) while the remaining 90 traits (X_{11} to X_{100}) were not associated with the QTL (additive effect of an allele substitution equal to 0).

In addition to major QTL, additional variability in trait values commonly arise from other small effect QTL, environmental effects, gene-by-environment interactions, and measurement errors. For these simulations, the variability induced by these sources was simulated as independent identically distributed normal random variables with a mean of zero and a variance of σ_x^2. The performance of sparse PCA was evaluated under 15 different simulations settings. The variance term σ_x^2 was simulated under the five settings of 0.375, 0.5833, 1.00, 1.4167 and 2.25, corresponding to 40%, 30%, 20%, 15% and 10% heritability for the traits associated with the QTL (X_1 to X_{10}) and 0% heritability for the traits that are not associated with the QTL (X_{11} to X_{100}). The size of the RIL population was simulated under the three settings of 100, 200, and 300 individuals. Each setting was simulated 1000 times. The appropriate degree of sparsity was estimated independently for each simulation using 5-fold CV. All possible degrees of sparsity from 0 (PCA) to 98 (only two non-zero loadings in sparse PCA) were considered. The degree of sparsity which maximized the variance of the linear combination of traits in the 5-fold CV testing data was then used in sparse PCA on the complete set of RILs.

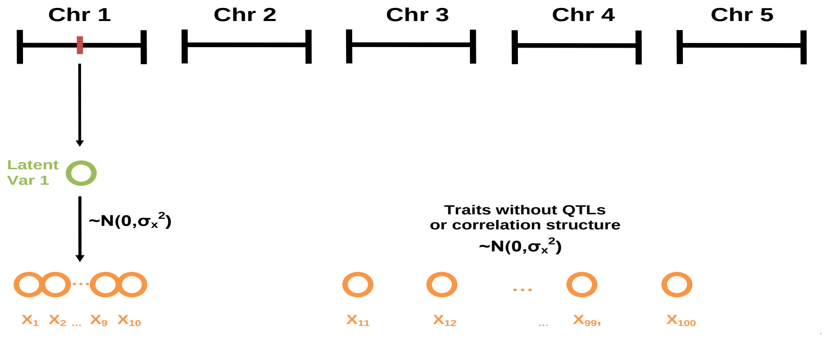
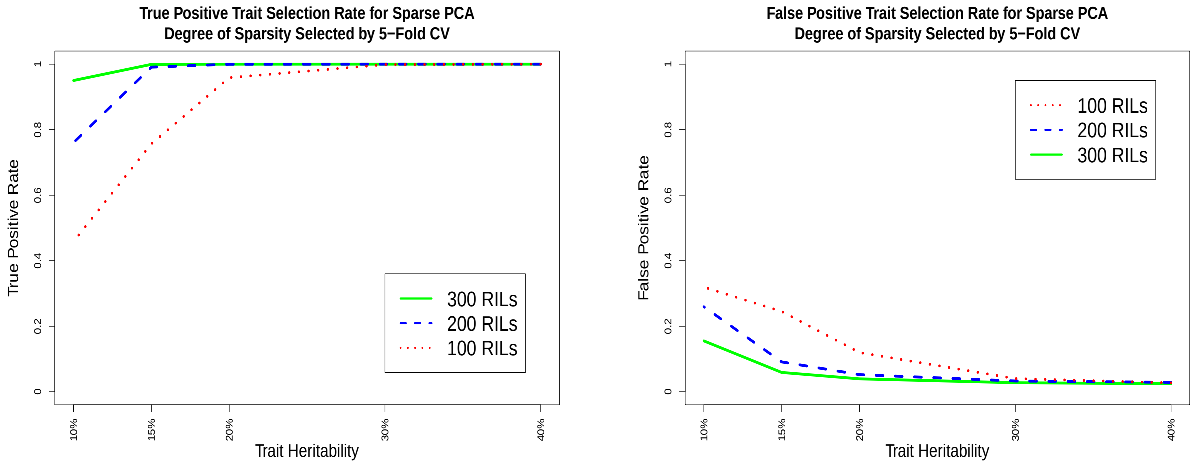


Figure 2: Simulation setting consisting of 100 traits (X_1 to X_{100}) measured on a diploid RIL population with 5 chromosomes. One QTL was simulated in the center of chromosome 1 (represented with a red line) which was associated with the traits X_1 to X_{10}. Random variation in the trait values was simulated from independent normal distributions with a mean of zero and a variance of σ_x^2.

The trait selection performance of sparse PCA is presented in Figure 3 and the QTL mapping performance is presented in Figure 4. The results from these simulations suggest that the trait selection power is highest when the number of individuals in the RIL population is large and the variance of the added noise is small (i.e., the heritability is large). These simulations also demonstrate a notable increase in the QTL detection power and improved localization of the QTL when the estimated latent variable from either PCA or sparse PCA is used, compared to mapping each of the traits X_1 to X_{10} individually. In all 15 simulation settings, the estimated latent variable from sparse PCA had more significant results and more accurate estimates of the QTL location than the estimated latent variable from PCA. This suggests that in addition to improving the interpretability of the results, sparse PCA provides better estimates of the underlying latent variable in the data. This is consistent with the motivation for our interest in sparse PCA, which seeks to create linear combinations consisting of only the traits that are truly associated with a common source of co-variation, unlike PCA which includes all traits in the linear combination.

More complex simulation studies were also considered but are not presented here. These simulations included multiple QTL, multiple latent variables, non-disjoint groups of traits, additional sources of variation influencing the latent variable structures, varying numbers of traits in each group and varying total numbers of traits. As a general statement, sparse PCA is able to identify multiple disjoint latent variable structures, but has difficulty separating overlapping structures

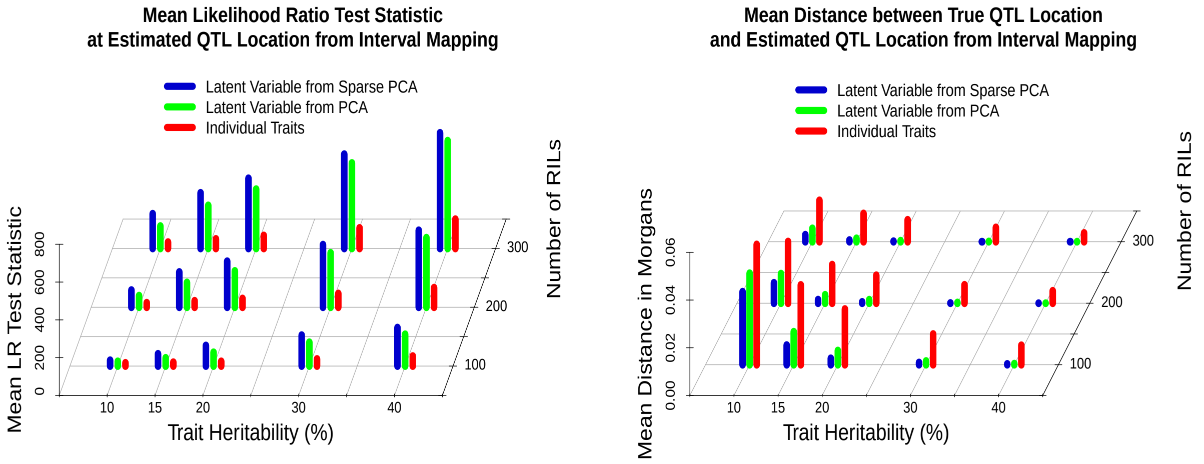


(a) (b)

Figure 3: Sparse PCA performance by trait heritability and population size (number of RILs) is presented based on 1000 simulations. **(a)** The vertical axis represents the group identification power of sparse PCA, which is defined as the proportion of the 10 truly correlated traits X_1 to X_{10} that were correctly identified as belonging to the group of correlated traits (i.e., true positives). These simulations show that the group identification power of sparse PCA increases as the number of RILs and the trait heritability increase. **(b)** The vertical axis represents the false positive rate of sparse PCA, which is defined as the proportion of the 90 truly uncorrelated traits X_{11} to X_{100} that were incorrectly identified as belonging to the group of correlated traits (i.e., false positives). These simulations show that the false positive rate of sparse PCA decreases as the number of RILs and the trait heritability increase.

where traits are associated with multiple latent variables, or where latent variables are not independent due to sharing identical or closely linked QTL. Simulating additional sources of variation (not associated with the major QTL) that cause traits within a group to be more correlated, improves the group identification accuracy of sparse PCA, however, this also reduces the QTL mapping performance, as this introduces additional variability to the traits. The variable selection and QTL mapping performance of sparse PCA typically improve as the number of traits associated with the QTL increases.

All simulations, sparse PCA analyses, and QTL mapping analyses were performed using the R programming environment (R Development Core Team, 2010).



(a) (b)

Figure 4: 1000 simulations were performed for each of the 5 heritability settings (horizontal axis) and 3 population sizes (depth axis). Single trait IM was performed separately on each of the individual traits X_1 to X_{10} (red) associated with the QTL on chromosome one, as well as the estimated latent variable from PCA (green) and sparse PCA (blue). All IM results with significant QTL on chromosome 1 were included in this figure. **(a)** The mean likelihood ratio (LR) test statistic at the peak QTL location is presented on the vertical axis. In all 15 simulation settings the estimated latent variable from sparse PCA had the highest mean LR test statistic, the estimated latent variable from PCA had the second highest mean LR test statistic, and the results from analyzing traits X_1 to X_{10} independently had the smallest mean LR test statistic. **(b)** The mean absolute distance in morgans from the estimated to the true QTL location is presented on the vertical axis. In all 15 simulation settings the mean location estimate from the estimated latent variable in sparse PCA was the most accurate, the mean location estimate from the estimated latent variable in PCA was the second most accurate, and the mean location estimate from analyzing traits X_1 to X_{10} independently was the least accurate.

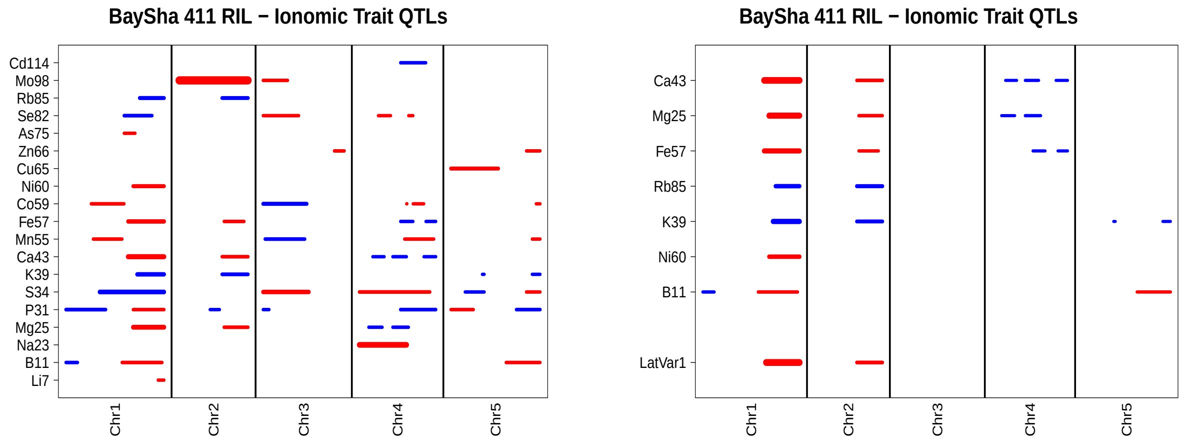
3.2 Real Data Analysis

The "Ionome" is the elemental composition of an organism, tissue or cell. Understanding the genetic basis that regulates the abundance of elements in individuals is important for understanding several biological processes (BAXTER *et al.*, 2007; GHANDILYAN *et al.*, 2009). We focus on an ionomic QTL mapping experiment (BAXTER *et al.*, 2007; BUESCHER *et al.*, 2010) based on a population of 411 *Arabidopsis thaliana* RILs from a cross between two distant ecotypes, Bay-0 and Shahdara (LOUDET *et al.*, 2002). Two biological replicates of each RIL were grown, as well as several replicates of the Bay-0 and Shahdara parental lines. Ionomic measurements from 19 elemental isotopes (Li7, B11, Na23, Mg25, P31, S34, K39, Ca43, Mn55, Fe57, Co59, Ni60, Cu65, Zn66, As75, Se82, Rb85, Mo98, Cd114) were measured on all viable plants. Plants were grown in trays consisting of approximately 70 plants, where the temperature, humidity, light, water and soil were controlled to be as uniform as possible across trays. A total of 15 growing trays were used, each of which contained approximately 7 Bay-0 and 8 Shahdara parental lines, along with 55 different RILs. To remove effects associated with differences in growing conditions across trays, the median trait value from each parental line was used to normalize the RIL data in each tray. Certain ionomic trait measurements were prone to occasional outliers due to machine spikes from the inductively coupled plasma mass spectroscopy (ICP-MS) technology, and therefore outliers were removed prior to tray normalization. After normalization, the mean of the biological replicates was taken as the observation for each RIL.

The genotypic data used for QTL mapping consisted of 38 microsatellite markers from LOUDET *et al.* (2002). The genetic map was estimated using QTL Cartographer software (BASTEN *et al.*, 2005, version 1.17). IM was performed on all 19 ionomic traits separately using Zmapqtl (BASTEN *et al.*, 2005, version 1.17) with model 3 (i.e., IM) and a walking speed of 2cM. The significance threshold for the likelihood ratio test statistic in IM was estimated by performing 1000 permutations for each trait (CHURCHILL and DOERGE, 1994) at the $\alpha = 0.05$ level of significance.

From the single trait IM results (Figure 5(a)), it is evident that several ionomic traits share similar QTL patterns, thus suggesting the possibility of a common latent variable structure influencing certain traits. Inspection of the correlation matrix further suggests that several of these traits may be biologically related due to the number of large correlations. The sparse PCA procedure was conducted using the R programming environment (R DEVELOPMENT CORE TEAM, 2010) to identify groups of correlated traits. A 5-fold CV sample was taken to estimate the appropriate degree of sparsity for these data. Sparse PCA was then performed on the complete set of individuals with the selected value for the degree of sparsity. Seven ionomic traits were selected as being correlated. The locations of significant QTL from single trait IM on these seven traits are shown in Figure 5(b), along with the locations of significant QTL for the estimated latent variable.

Biologically validating this grouping of seven ionomic traits selected by sparse PCA is difficult since the complex interactions among traits are not well understood. This aside, the biological information that is available does suggest that these traits may be related, and may be involved in similar biological processes. Their grouping is further supported by the similarities in the locations



(a) (b)

Figure 5: **(a)** Locations of significant QTL from single trait interval mapping (IM) on all 19 ionomic traits in the population of 411 Bay × Sha RILs. The genetic position of the QTL is represented along the horizontal axis, and the ionomic traits are represented along the vertical axis. For a given trait, all genetic positions having a likelihood ratio (LR) score above the significance threshold are plotted. Higher trait values that are associated with Bay alleles are represented in red, and higher traits values associated with Sha alleles are represented in blue. Positions with higher LR scores (more significant results) are graphed with thicker lines. **(b)** Locations of significant QTL from single trait IM on the seven ionomic traits selected by sparse PCA, as well as the estimated latent variable from the linear combination of traits. Traits are ordered based on the magnitude of their weight in the linear combination. The trait contributing the largest amount to the estimated latent variable is Ca43, and the trait contributing the least is B11.

of significant QTL, and the magnitude of the correlations among these traits. Although this is not conclusive evidence of the validity of this grouping, it does suggest that these traits would be a suitable selection for analysis in multiple-trait QTL mapping.

4 Discussion

When biological knowledge about the relationships between traits is not available, statistical procedures are needed to identify groups of traits for joint analysis. Performing multiple-trait QTL mapping on related (non-independent) traits increases the QTL detection power and provides more informative tests than can be obtained by analyzing each trait separately. In this work, we propose a novel application of sparse PCA to search for groups of correlated traits, as these traits may share underlying genetic determinants. Additionally, sparse PCA estimates the unobserved latent variable associated with these traits, which can in turn be used for QTL mapping, resulting in improved QTL detection power and more accurate estimates of the QTL location than mapping on each trait independently. The procedures presented here address only the first principal component (PC), however, subsequent components can be found by removing variation associated with previous PCs and using the residual data matrix to search for additional PCs.

Correlation structures in phenotypic traits can arise from various sources, including common QTL, closely linked QTL, environmental effects and gene-environment interactions. In the simulation study presented in Section 3.1, only correlations induced by a common QTL were considered. In simulations where additional sources of shared variation were included (e.g., common QTL, environmental effects, etc.), the group identification performance of sparse PCA was greatly increased, reaching nearly 100% detection power when these effects were large.

More complex simulation studies were also performed in addition to the one QTL and one latent variable structure presented in Section 3.1. One notable trend from these simulations was the frequent failure to separate multiple latent variable structures with one or more traits in common. Instead of identifying the groups of traits as being associated with separate PCs, the traits were often selected as belonging to one large group. This insight may help to explain the results of the ionomic trait analysis. Careful inspection of Figure 5(b) reveals that at least three different patterns exist in the results from single trait IM on the traits selected by sparse PCA. First, B11 and Ni60 both have significant QTL on chromosome 1, but none on chromosome 2. Second, Ca43, Mg25 and Fe57 have similar QTL results on chromosome 1 and 2, as well as significant QTL on chromosome 4. Lastly, Rb 85 and K39 both have similar QTL patterns on chromosome 1 and 2, but unlike Ca43, Mg25 and Fe57, they do not have significant QTL on chromosome 4. This may indicate that these ionomic traits are indeed a combination of multiple overlapping latent variable structures, rather than one large group. Further biological research is required to validate this hypothesis, and further statistical research is needed to develop methods to identify when the selected traits may indeed be a combination of multiple overlapping groups.

The simulations also showed that the k-fold CV procedure frequently estimated degrees of sparsity for sparse PCA that resulted in too many traits being selected (i.e., groups that included several truly uncorrelated traits). Performing additional CV samples can improve the accuracy of the estimates; however, the problem does not fully diminish even when the number of CV samples is large. Additional research is needed to improve the degree of sparsity estimates. Alternative evaluation measurements may be considered to supplement or replace the method presented in Section 2.3.

Finally, further research is needed to extend the trait selection methods presented here to situations where one wants to combine traits across different phenotypic data sets. For example, in addition to the ionomic trait data analyzed here, additional phenotypes have been collected by different laboratories using the same Bay × Sha RIL population. Searching for relationships between traits from different data sets (rather than only within a given data set) has the potential to provide insight into the genetic architecture governing relationships among apparently different phenotypes. This problem will be addressed in subsequent publications.

5 Summary

The procedures detailed in this paper present a novel application of sparse principal component analysis (sparse PCA) to select groups of phenotypic traits for quantitative trait loci (QTL) mapping. The proposed procedures perform dimension reduction and variable selection on the phenotypic trait data, followed by QTL mapping on the selected traits and the estimated latent variable associated with these traits. The simulation studies presented demonstrate that the proposed procedures are capable of identifying groups of traits associated with a common QTL. Additionally, these simulations show the improvement in QTL detection power when the estimated latent variable from sparse PCA is used compared to analyzing each trait separately. The analysis of ionomic trait data from the Bay × Sha RIL population in *Arabidopsis thaliana* provides evidence of the method's performance on real data, where correlated traits were identified and significant QTL were found to be associated with the estimated latent variable.

6 Acknowledgements

We thank the Kansas State University Department of Statistics for the student travel grant allowing TRA to present this work at the conference. We thank The Center for Gene-Environment Interactions at Purdue University, and the Purdue University Department of Statistics for their support of this research. We thank Michael Zhu, Mary Ellen Bock, Malgorzata Bogdan, Hyonho Chun, Ivan Baxter, Doug Crabill, My Truong, and the RWD Research Group for their continued support and encouragement. We thank an anonymous reviewer for helpful feedback, and we thank Andrea Rau for her help in editing this manuscript. This work is funded in part by a NSF Plant Genome grant (DBI-0733857) in part to RWD.

References

BASTEN, C. J., B. S. WEIR, and Z.-B. ZENG, 2005 *QTL Cartographer: version 1.17*. Department of Statistics, North Carolina State University.

BAXTER, I., M. OUZZANI, S. ORCUN, B. KENNEDY, S. S. JANDHYALA, *et al.*, 2007 Purdue ionomics information management system. an integrated functional genomics platform. Plant Physiology **143**: 600–611.

BUESCHER, E., T. ACHBERGER, I. AMUSAN, A. GIANNINI, C. OCHSENFELD, *et al.*, 2010 Natural genetic variation in selected populations of arabidopsis thaliana is associated with ionomic differences. PLoS ONE **5**: e11081.

CHURCHILL, G. A., and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138**: 963–971.

GHANDILYAN, A., L. BARBOZA, S. TISN, C. GRANIER, M. REYMOND, *et al.*, 2009 Genetic analysis identifies quantitative trait loci controlling rosette mineral concentrations in arabidopsis thaliana under drought. New Phytologist **184**: 180–192.

HACKETT, C. A., R. C. MEYER, and W. T. B. THOMAS, 2001 Multi-trait qtl mapping in barley using multivariate regression. Genetical Research **77**: 95–106.

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–324.

HENSHALL, J. M., and M. E. GODDARD, 1999 Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. Genetics **151**: 885–894.

HOTELLING, H., 1933a Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology **24**: 417–441.

HOTELLING, H., 1933b Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology **24**: 498–520.

JANSEN, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135**: 205–211.

JANSEN, R. C., 1994 Controlling the type i and type ii errors in mapping quantitative trait loci. Genetics **138**: 871–881.

JIANG, C., and Z. B. ZENG, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140**: 1111–1127.

JOLLIFFE, I. T., 2002 *Principal Component Analysis*. Springer Series in Statistics. Springer New York, 2 edition.

JOURNEE, M., Y. NESTEROV, P. RICHTRIK, and R. SEPULCHRE, 2010 Generalized power method for sparse principal component analysis. Journal of Machine Learning Research **11**: 517–553.

KAO, C.-H., Z.-B. ZENG, and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

KNOTT, S. A., and C. S. HALEY, 2000 Multitrait least squares for quantitative trait loci detection. Genetics **156**: 899–911.

KOROL, A. B., Y. I. RONIN, and V. M. KIRZHNER, 1995 Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics **140**: 1137–1147.

KOROL, A. B., Y. I. RONIN, E. NEVO, and P. M. HAYES, 1998 Multi-interval mapping of correlated trait complexes. Heredity **80**: 273–284.

LAN, H., J. P. STOEHR, S. T. NADLER, K. L. SCHUELER, B. S. YANDELL, *et al.*, 2003 Dimension reduction for mapping mrna abundance as quantitative traits. Genetics **164**: 1607–1614.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping mendelian factors underlying quantitative traits using rflp linkage maps. Genetics **121**: 185–199.

LOUDET, O., S. CHAILLOU, C. CAMILLERI, D. BOUCHEZ, and F. DANIEL-VEDELE, 2002 Bay-0 shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in arabidopsis. Theoretical & Applied Genetics **104**: 1173–1184.

LUO, Z. W., and M. J. KEARSEY, 1989 Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. Heredity **63**: 401–408.

MOSTELLER, F., and D. L. WALLACE, 1963 Inference in an authorship problem. Journal of the American Statistical Association **58**: 275–309.

PEARSON, K., 1901 On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6 **2**: 559–572.

R DEVELOPMENT CORE TEAM, 2010 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SHEN, H., and J. Z. HUANG, 2008 Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis **99**: 1015 – 1034.

STONE, M., 1974 Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological) **36**: 111–147.

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

WITTEN, D. M., R. TIBSHIRANI, and T. HASTIE, 2009 A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**: 515–534.

WOLD, S., 1978 Cross-validatory estimation of the number of components in factor and principal components models. Technometrics **20**: 397–405.

XU, C., Z. LI, and S. XU, 2005 Joint mapping of quantitative trait loci for multiple binary characters. Genetics **169**: 1045–1059.

ZENG, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America **90**: 10972–10976.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136**: 1457–1468.

ZOU, H., and T. HASTIE, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B **67**: 301–320.

ZOU, H., T. HASTIE, and R. TIBSHIRANI, 2006 Sparse principal component analysis. Journal of Computational and Graphical Statistics **15**: 265–286.