Accepted Manuscript

Title: Chemometric regression techniques as emerging, powerful tools in genetic association studies

Author: Gerard G. Dumancas, Sindhura Ramasahayam, Ghalib Bello, Jeff Hughes, Richard Kramer

PII: S0165-9936(15)00228-9
DOI: http://dx.doi.org/doi: 10.1016/j.trac.2015.05.007
Reference: TRAC 14509

To appear in: *Trends in Analytical Chemistry*

Please cite this article as: Gerard G. Dumancas, Sindhura Ramasahayam, Ghalib Bello, Jeff Hughes, Richard Kramer, Chemometric regression techniques as emerging, powerful tools in genetic association studies, *Trends in Analytical Chemistry* (2015), http://dx.doi.org/doi: 10.1016/j.trac.2015.05.007.

Chemometric regression techniques as emerging, powerful tools in genetic association studies

Gerard G. Dumancas [a,] *, Sindhura Ramasahayam* [b], *Ghalib Bello* [c], *Jeff Hughes* [d], *Richard Kramer* [e]

[a] *Department of Chemistry, Wood Science Building, Oklahoma Baptist University, Shawnee, Oklahoma, USA 74804*
[b] *Department of Hematology and Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA, 19104*
[c] *Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, OKC, USA, 73104*
[d] *School of Applied Science, Royal Melbourne Institute of Technology University, Melbourne VIC 3001, Australia*
[e] *Applied Chemometrics, Inc., Sharon, MA, USA 02067*

HIGHLIGHTS

- Applications of some chemometric techniques to genetic epidemiology
- Advantages of chemometric techniques over conventional techniques
- Role of chemometrics in the future of genetic association studies

ABSTRACT

The field of chemometrics has its origin in chemistry and has been widely applied to the evaluation of analytical chemical data and quantitative structure-activity relationships. Chemometric techniques apply statistical and algorithmic methods to extract information from analytical multivariate data, including fused, heterogeneous data. These techniques are now widely applied across fields as varied as food technology, environmental chemistry, process control, medical diagnostics, and metabolomics. In the mid-1980s, cross-disciplinary interaction between genetics and epidemiology led to the emergence of genetic epidemiology as a new discipline. Chemometric techniques are extremely appropriate for, and have been widely applied to, this discipline. Here, we present a broad review of the application of chemometric techniques to the fields of genetic epidemiology and statistical genetics. We also consider some future directions. We focus on chemometrics-based regression methodologies in genetic association studies.

Keywords:
Chemometrics
Genetic epidemiology
Genome-wide association studies
Multivariate data
Multivariate regression technique
Partial least squares
Principal-component regression
Ridge regression
Single-nucleotide polymorphism

Statistical genetics

Abbreviations: (to be set in normal style)

ANOVA-simultaneous component analysis	ASCA
Backward interval PLS	biPLS
Basic Local Alignment Search Tool	BLAST
Bayesian regression	Bayes-R
Canonical correlation analysis	CCA
Expression quantitative trait loci	eQTL
False discovery rate	FDR
Fixed regression-least squares	FR-LS
Genome-wide association studies	GWAS
Genetic algorithm	GA
Heteroscedastic effects model	HEM
Interval PLS	*i*PLS
Joint Genetic Association of Multiple Phenotypes	JAMP
Linkage disequilibrium	LD
Molecular breeding value	MBV
Multivariate analysis of variance	MANOVA
Multivariate sparse partial least squares	M-SPLS
National Center for Biotechnology Information	NCBI
National Human Genome Research Institute	NHGRI
Ordinary least squares	OLS
Orthogonal projections to latent structures	OPLS
Partial least squares	PLS
Principal component analysis	PCA
Principal Component of Heritability Association Test	PCHAT
Principal component regression	PCR
Recursive weighted PLS	rPLS
Quantitative trait loci	QTL
Random regression best linear unbiased prediction	RR-BLUP
Rare variant	RV
Ride regression	RR
Root mean square error of prediction	RMSEP
Selectivity ratio	SR
Single nucleotide polymorphism	SNP
Singular value decomposition	SVD
Support vector regression	SVR
Synergy interval PLS	siPLS
Trait-based Association Test that uses Extended Simes	TATES
Variable importance for projection	VIP

* Corresponding author. Tel.: +1 (405) 730-8752.
E-mail address: gerard.dumancas@okstate.edu (G. Dumancas)

1. Introduction

Genetic epidemiology involves the study of the interaction of genes and the environment and how these factors influence disease in human populations and their patterns of inheritance in families [1,2]. In recent years, the field of genetic epidemiology has been broadly applied in a wide array of research fields. Here, we focus on biallelic single-nucleotide polymorphisms (SNPs), point mutations in the genome that can take on one of two possible alleles. In other

words, two allelic variants are segregating in the population [3]. For diploid organisms, this implies three possible genotypes at each polymorphic site. For example, an SNP with alleles A (adenine) and G (guanine) would lead to three possible genotypes in a diploid organism: AA, AG, and GG [4]. Statistical analysis of the effect of this type of polymorphism on a phenotype of interest would therefore involve representing the SNP as a three-category variable, though we discuss other modes of representation in this article. In general, the central goal of genetic epidemiology is the identification of SNPs (and consequently the genes in which they are located) known to be associated with a phenotype (e.g., disease) of interest. This process is carried out via genome-wide association studies (GWAS). With the discovery of these genes, scientists have gained a deeper understanding of the etiology of diseases suspected to have a genetic component. This has facilitated the development of drugs and treatments to counteract such diseases.

The steadily increasing availability of genomic data has motivated the search for new and/or improved ways to uncover relationships between genotypes and phenotypes. The standard approach is GWAS, wherein the associations of a large number of SNPs (typically of the order of thousands or millions) with a phenotype of interest are tested separately for each SNP. This sidesteps the seemingly intractable high-dimensional problem by breaking it down into a series of univariate regressions, each of which tests the association of a particular SNP with the phenotype. Appropriate *post hoc* adjustments are then made for multiple testing.

Despite an inconsistent record of success, GWAS continue to be the "gold standard" in this area, and few studies have attempted to exploit the potential advantages of analyzing the multivariate relationships inherent in genotype-phenotype data [5]. Multivariate analyses have proved to be effective when working with complex datasets and have several advantages over univariate approaches. A common reason for employing a multivariate model is the ability to use multiple measurements of one underlying construct in order to achieve better construct validity. One advantage of using multivariate analysis is that Type I error rates are better controlled as compared to the inflation of Type I error that results when carrying out a series of univariate statistical tests. A second advantage of the multivariate approach is that it often has more power than the univariate approach, because the latter tends to focus on only marginal effects [6].

As mentioned above, the most common approach in GWAS is to analyze one SNP at a time, in an attempt to perform a genetic dissection of complex diseases in a holistic manner. However, this approach does not fully utilize the potential of GWAS to identify multiple causal variants in order to more fully predict the risk of disease. Several methods for joint analysis of GWAS data tend to miss causal SNPs that are by themselves weakly correlated with disease, and also suffer from a high false-discovery rate (FDR) [7].

Chemometrics is a multivariate data analysis approach with the primary aim of concentrating the significant variance in a mixed data structure (e.g., concentration data) onto a relatively small number of orthogonal (in the case of PCA) or nearly orthogonal (in the case of PLS) principal components or factors. In order to reduce a relatively large number of collinear variables that, in general, are expected to exhibit considerable mutual intercorrelation, to a smaller number of orthogonal or nearly orthogonal factors, the concentration data are treated by multivariate data analytical methods [8,9]. In this article, we review several chemometric regression techniques, which have been applied to genetic data. We focus on implementation of these techniques to a generic GWAS problem and discuss its pros and cons.

2. Regression-modeling approaches for the analysis of genetic data

2.1. Univariate regression

In a univariate regression model, the dependent variable (or response) is modeled by a single independent (explanatory) variable, *x*. The equation below illustrates a univariate generalized regression model:

$$g(E(\mathrm{y})) = \beta_0 + \beta_1 x \qquad (1)$$

Here, β_0 is the intercept, y is the dependent variable and $g(.)$ is known as a link function, the exact formulation of which depends on the distribution of *y*, e.g., if *y* is normally distributed, then $g(.)$ is an identity function, and if *y* is binary/dichotomous (i.e., follows a Bernoulli distribution), then $g(.)$ may be a logit or probit function.

GWAS utilize univariate regression for testing the association of each SNP with the trait or disease of interest [10]. As an illustration, typical GWAS will utilize a univariate regression model similar in form to Equation (1) above, wherein *x* is a particular SNP and *y* is the phenotype of interest. A separate univariate regression model is fit for each SNP in the GWAS [11]. Typically, GWAS would have hundreds of thousands to millions of SNPs, and hence the same number of univariate regression models.

It is worthwhile to discuss exactly how SNPs are represented in statistical models utilized in GWAS. These representations are based on assumptions about the effect of the polymorphism on the phenotype. The most common representation assumes a dosage effect [i.e., the effect of the risk allele on the phenotype is additive/cumulative, so individuals who are heterozygous for the risk allele (possessing 1 copy) would have lower risk than those who are homozygous for the risk allele (possessing 2 copies)]. Conversely, those homozygous for the alternate allele (i.e., possessing 0 copies of the risk allele) would have non-existent risk. In this dosage effect model, the SNP is represented using a categorical coding of 0, 1, and 2, representing the allele counts of the risk allele. Other models (e.g., dominance or genotype) could be utilized. If a "risk" allele is not known *a priori*, a reference allele can be chosen.

The statistical models covered here are by no means the only tools used in GWAS. For example, categorical data-analysis techniques, such as the Cochran-Armitage trend test, Pearson's chi-square test and Fisher's exact test, are ubiquitous in genetic epidemiology studies. The Cochran-Armitage trend test, in particular, has become a standard tool for association testing in genetic case-control studies, although logistic regression (particularly the Wald test) has been proposed as superior [12,13].

Pearson's chi-square test is one of the oldest statistical tests; typically applied to categorical data organized in contingency tables, the test examines the difference between the observed frequencies and their corresponding expected frequencies under the null hypothesis. The resulting statistic follows a chi-square distribution, which is used to determine the p-value. The Pearson's chi-square test typically performs poorly with small sample sizes.

Fisher's exact test is also useful for categorical data; it is a permutation-based test for evaluating the association between two dichotomous variables. It is a particularly good alternative to Pearson's chi-square test when sample size is small. In principle, Fisher's exact test can be applied to any sample size, but, due to its computational demands, it is typically reserved as a small-sample test. There have been suggestions that this test is conservative, and Barnard's

exact test [14] has been proposed as a similar, but less conservative, alternative.

Although standard GWAS problems utilize SNPs as independent variables, it is genes, not SNPs, that are the functional units in the genome [15,16]. While this review primarily focuses on SNP-based analysis, we note here that there are also gene-based analyses for the univariate trait setting {e.g., GATES, VEGAS, and JAG, [16-18]}. This class of techniques offers another option to avoid the multiple testing problems inherent in univariate SNP-based analyses [11].

As discussed above, since many genome-wide studies consist of a large number of SNPs (often on the order of millions), a large number of univariate tests (one for each SNP) must be performed. Performing such a large number of univariate tests has potential pitfalls because the greater the number of tests, the higher the likelihood of erroneously rejecting the null hypothesis when it is true. Special statistical procedures therefore have to be used to adjust for multiple testing, which will control the false-positive rate. [19]. Examples are the Bonferroni procedure, the Benjamini-Hochberg procedure, and the Tukey's procedure. As a consequence, these univariate approaches require stringent significance thresholds (due to the large number of tests being undertaken) to control the false-positive rate (the adjusted α significance level for GWAS is typically a value of $\alpha = 5 \times 10^{-8}$ [11]. This strict threshold is sometimes difficult to meet, and this is a key disadvantage of univariate approaches. As mentioned above, univariate techniques also fail to take into account the combined effects of multiple SNPs for it is possible that the interactions among multiple genetic variants could contribute to the phenotype of interest [20,21]. Lastly, traits of interest are often multivariate in nature (i.e., multiple phenotypes are measured to cover the full extent of a trait). For example, cognitive ability is usually measured through batteries of tests covering various cognitive abilities (e.g., memory and vocabulary). In the GWAS context, this multivariate information is generally collapsed to a univariate score (i.e., a univariate full-scale IQ score, or a binary case-control index) [11].

In order to circumvent the limitations of univariate regression techniques in genetic epidemiology, recent studies began to apply other approaches, including multiple linear regression (MLR), ridge regression (RR), principal-component regression (PCR), and partial least squares (PLS). These techniques are multivariate in nature, so they take into account the combined effects of multiple SNPs and also control for confounding variables [19].

2.2. Multivariate regression

While univariate regression has certain advantages (e.g., simplicity and computational feasibility), it is inadequate for modeling more complex relationships in genetic studies. In situations where the phenotype is best modeled by multiple SNPs and environmental factors, multivariate regression is the more appropriate tool to utilize. For this sort of situation, if phenotype *Y* is normally distributed, then a MLR model can be utilized:

$$\mathbb{E}[\mathrm{Y}] = \beta_O + \beta_X Z + \beta_G G, \qquad (1)$$

where β_G is the parameter of interest quantifying the association between a matrix of genotypes G and the mean of the phenotype Y. The matrix G is typically an ($n \times m$) matrix where n is the number of individuals and m is the number of SNPs genotyped in these individuals. Further, Z is a matrix of p covariates, representing variables to be adjusted (e.g., for age and gender). Denote $X = (1, Z, G)$ and $\beta = (\beta_O, \beta_X, \beta_G)$. As long as the model (1) for the expected value of the outcome is correct, the ordinary least squares estimate $\hat{\beta} = (X^T X)^{-1} X^T \mathrm{Y}$ is an unbiased estimator of

β [22]. It can be roughly thought of as a weighted average of the model outcome Y with weights that depend on the covariate set X [22].

Binary outcomes often arise in biomedical data, where they may, for example, represent cases and controls with respect to a disease or condition. Such phenotypes are modeled using logistic regression models. The typical logistic regression model has the following form:

$$\log_e\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = a + b_1 x_{1,i} + b_2 x_{2,i} + \ldots\ldots + b_p x_{p,i}$$

Here, $x_{1,i},\ldots,x_{p,i}$ are the explanatory/predictor variables (e.g., SNPs or environmental variables) for the i^{th} individual. Y_i is a Bernoulli random variable representing the outcome or phenotype for individual i. The outcome is usually coded numerically as 1 or 0 (e.g., cases or controls). Therefore $P(Y_i = 1)$ is the probability of the phenotype being equal to 1 (e.g., a case). Thus, the logistic regression simply relates the probability that the i^{th} individual is a case to the predictor/explanatory variables:

$$P(Y_i = 1 \mid \boldsymbol{x}_i) = p_i = \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)} \qquad (3)$$

where β is a $(p \times 1)$ vector of parameters to be estimated.

Traits of interest are often multivariate in nature, as mentioned in sub-section 2.1, and also genes, and not SNPs, are the functional unit in a genome [15,16]. In order to circumvent such limitations, a tool called MGAS was recently developed that allows gene-based testing of multivariate phenotypes in unrelated individuals. MGAS allows researchers to conduct their multivariate gene-based analyses efficiently, without the loss of power that is often associated with incorrectly specified genotype-phenotype models [11].

A joint analysis of multiple, potentially correlated traits (i.e., a multivariate analysis) offers a number of advantages over the univariate approach [23]. Primarily, a multivariate analysis may have more power to detect association in situations where there is genetic correlation among different traits. This extra information that is provided by the cross-trait covariance is ignored in a typical univariate analysis [13, 24]. Second, most multivariate procedures offer the ability to perform a single test for association with a set of traits. This consequently reduces the number of tests performed and alleviates the multiple testing burden that applies when analyzing all traits separately [13,25] Most importantly, multivariate analysis is helpful in cases of pleiotropy, where a single genetic variant influences multiple phenotypes [26]. Multivariate methods for the genome-wide SNP-based analysis include MultiPhen, canonical correlation analysis (CCA), i.e. MANOVA (multivariate analysis of variance) with the SNP-effect treated as covariate, TATES (Trait-based Association Test that uses Extended Simes procedure) and JAMP (Joint Genetic Association of Multiple Phenotypes) [27–29]. Some popular software packages for performing multivariate GWAS include methods that are implemented in PLINK (the multivariate test of association MQFAM) [27], SNPTEST (a Bayesian multiple phenotype test) [30], BIMBAM (a Bayesian model comparison and model averaging for multivariate regression) [31], and PCHAT (Principal Component of Heritability Association Test) [25].

Although MLR has been successfully applied in the context of GWAS, it is unreliable in the presence of multicollinearity, which occurs when there are strong correlations among the

independent variables. In such situations, multiple regression analysis produces inaccurate and unstable estimates, and generally fails to clarify the relationships among the predictor and the response variables [32]. Multiple logistic regression is equally susceptible to issues arising from multicollinearity [33]. Lastly, standard linear regression models are based on a number of assumptions {e.g., normality and constant variance of errors [homoscedasticity]}, which, if violated, may result in inefficient or biased estimates [34].

2.3. Principal-component regression (PCR)

Another robust chemometric approach applied in the area of genetic epidemiology is PCR. PCR is a powerful extension of standard ordinary least squares regression. It is ideal for situations where multicollinearity exists among independent variables. The basic idea behind PCR is to resolve the multicollinearity problem by performing an orthogonal transformation of the independent variables into a number of principal components. Once this is accomplished, multicollinearity is easily detected by examining the eigenvalues, and components associated with low eigenvalues can be removed (other methods exist for determining which components to exclude). Since principal components are, by definition, mutually uncorrelated, regressing the response vector (Y) unto them would solve the multicollinearity problem. The results can then be transformed back to the original scale of the independent variables.

We now outline the formal procedure in mathematical terms. We begin with the standard regression model:

$$Y = X^T b + e \tag{6}$$

where Y is the vector of responses, X is the matrix of independent variables (e.g., a matrix of SNPs) and β is the vector of unknown regression coefficients. The first step in PCR is to standardize the independent variables so that they have zero mean and unit standard deviation. Then, the resulting data matrix X can be factorized using singular value decomposition (SVD):

$$X = UWV^T \tag{7}$$

In the above expression, U and V are orthogonal matrices and W is a diagonal matrix with non-negative real numbers along its diagonal representing the singular values of X.

X^TX can therefore be expressed as:

$$\begin{aligned} X^TX &= (UWV^T)^T(UWV^T) \\ &= VW^TU^TUWV^T \\ &= VWU^TUWV^T \\ &= VW^2V^T \\ &= VDV^T \end{aligned} \tag{8}$$

The matrix VDV^T is essentially the Eigen decomposition of X^TX and the columns v_j of V are the PCA loadings of X. PCA is the most popular multivariate statistical technique that helps to analyze datasets with highly related predictors. PCA is a data-compression technique that

reduces a larger set of predictor variables to a smaller set with a minimal loss of information [35]. The objective of the PCA is to extract the important information from the data table, which represents the observations reported as dependent variables, and expresses this information as a set of new orthogonal variables called principal components [36]. PCA is used in dietary studies, physical medicine, and in human kinematics and biomechanics studies [37]. The diagonal elements of matrix D are the eigenvalues of X^TX. Each eigenvalue is equal to the total variance modelled by each respective principal component. PCA concentrates the significant variance onto the earlier components. After the significant variance is substantially modelled by the earlier components, the remaining variance that is modeled by the later components represents only noise and the respective eigenvalues for those components will be close to zero and the components with which they correspond can be discarded from the model. After this step, the remaining q columns of V can be combined into a matrix V_q and a new data matrix $Z_q = XV_q$ can be obtained. The columns of this derived matrix are orthogonal so, using it in lieu of the original data matrix X would eliminate the multicollinearity problem. Ordinary least squares can then be applied to obtain an estimate using the derived data matrix Z_q:

$$\hat{a} = (Z_q^T Z_q)^{-1} Z_q^T Y \tag{9}$$

The PCR estimate of β can then be obtained by transforming the above estimate back to the original scale of the X data matrix:

$$\hat{b} = V_q \hat{a} \tag{10}$$

PCR provides an elegant solution to the problem of multicollinearity, so it is particularly useful in genomic association studies where SNPs tend to be highly intercorrelated. For example, Wang and Abbott [79] demonstrated how PCR can be used as a dimensionality-reduction technique in genomic association studies, and applied it to the problem of testing the association between expression of the gene CHI3L2 and SNPs within that gene. PCR was able to confirm significant associations of SNPs that had previously been reported in other studies.

In a comparative study, Ballard et al. [38] assessed the performance of seven multi-marker association tests and found that PCR was the most powerful among them.

Gauderman et al. [39] demonstrated through simulations that PCR is typically at least as powerful as other genotype- or haplotype-based techniques. Using PCR to test the association between a select group of SNPs (within the Glutathione-S-Transferase P1 gene) and childhood chronic bronchitis, they observed stronger evidence of association than what was observed using the more traditional genotype- and haplotype-based methods.

2.4. Partial least squares (PLS)

PCR is commonly used as an alternative to PLS [40]. PCR and PLS are both known to have no significant differences in the prediction errors except in cases when artificial constraints are placed on the number of latent variables retained [41,42], but PLS is a more commonly used standard tool in chemometrics [43]. Further PLS may concentrate the significant variance in the independent data onto fewer latent variables than PCR, but this does not appear to influence predictive ability [41,42]. For PLS in the context of genetic epidemiology, the association

between a phenotype vector *y* and the genotypes *X* is assumed to be explained by the following linear model:

$$E(y) = X'\beta \quad (4)$$

where β is a $p \times 1$ vector of regression coefficients. PLS then performs a simultaneous decomposition of *X* and *y* with the constraint that these components ("latent factors") explain as much of the covariance between *X* and *y* as possible. The regression coefficients in the above model can be estimated via the PLS method, as follows [5]:

$$\hat{\hat{\beta}} = \hat{W}(\hat{p}'\hat{W})^{-1}\hat{q} \quad (5)$$

where $\hat{p}$ is the $p \times k$ matrix of *X* loadings, $\hat{q}$ is the $k \times 1$ vector of *y* loadings and $\hat{W}$ is the $p \times k$ matrix of loadings, as defined in references [8,9].

The superiority of PLS over the classical methods is not surprising, since it was designed in situations where a large (relative to sample size) number of correlated predictors exist [44]. It is also known for its ability to handle huge data matrices, and is less apt to be overwhelmed by "noisy" variables. PLS is able to isolate the informative variance in the data by using only the significant latent variables. PLS retains only that variance in the independent variables that exhibits linear correlation to the dependent variables. The resulting reduction in the noise level in the data may improve predictive stability and produce more accurate models [45]. In a study conducted by Cassel et al. [46], PLS was tested in the presence of the following three inadequacies, and the algorithm demonstrated remarkable robustness in face of them:

(i) skewed (asymmetric) distributions for-observed variables;

(ii) multicollinearity within blocks of observed (manifest) variables and among latent variables; and,

(iii) misspecification of the structural model (by omission of regressors).

A number of studies have utilized PLS in the analysis of genetic data. PLS has been applied to predict molecular breeding values (MBVs) using genotypes at 7372 SNPs and very accurately estimated breeding values of 1945 dairy bulls. The algorithm produced MBV prediction (for genomic selection) accuracies that were similar to those of Bayesian regression (Bayes-R) [47–49], random regression best linear unbiased prediction (RR-BLUP) [47–49], and non-parametric support vector regression (SVR) [50–52]. Fixed regression-least squares (FR-LS) [47,48,53] yielded poor performance in comparison [54].

A similar application of the PLS in genomic selection was carried out by Colombani and colleagues, and involved computing a prediction equation from the estimated effects of a large number of DNA markers based on a limited number of genotyped animals with observed phenotypes [55]. PLS and sparse PLS were used with a reference population of 940 genotyped and phenotyped French Holstein bulls and 39,738 polymorphic SNP markers. Correlations between observed phenotypes and phenotypes predicted by PLS and sparse PLS were similar, but sparse PLS highlighted some genomic regions more clearly. Both PLS and sparse PLS were more accurate than pedigree-based BLUP and generally provided lower correlations between observed and predicted phenotypes than did genomic BLUP.

Sarkis and colleagues utilized PLS to locate causal markers by treating the phenotype as the dependent variable and the genotype as the independent variable. Their study showed that the results obtained using their PLS-based approach were in good agreement with those obtained by

more standard techniques (e.g., χ^2 and trend-regression tests). Also, PLS achieved higher accuracy, thus demonstrating the successful application of chemometrics to a problem in human genetics [56].

Another interesting application of PLS was proposed by Turkmen and colleagues, who introduced two PLS approaches to aggregate the signals of many SNPs within a gene to reveal possible genetic effects related to rare variants (RVs). The proposed methods were able to identify some rare SNPs that were missed by the standard SNP-based analysis [57].

PLS has also been applied to mining for genotype-phenotype relations specifically from genomic sequences. Mehmood and colleagues introduced a methodology based on the Basic Local Alignment Search Tool (BLAST) approach for extracting information from genomic sequences and soft-threshold PLS for mapping genotype-phenotype relations. BLAST and PLS-based multivariate approach produced results that showed good agreement with known yeast phylogeny and gene ontology. This confirmed that the methodology extracts a set of fast-evolving genes that accurately capture the phylogeny of the yeast strains [5]. BLAST is a program designed to compare the similarity between nucleotide or protein sequences with the databases and calculates its match. It interprets relationship between sequences and also helps identify members of gene families. BLAST is popular due to its free availability on the Web through a large server at the National Center for Biotechnology Information (NCBI) [58].

2.5. Ridge regression (RR)

Although less commonly used in chemometrics [59], RR is another chemometric technique that, in recent years, received considerable attention in quantitative genetics [60]. It is considered to be a particularly promising alternative regression procedure for the analysis of data in which the predictor variables are highly correlated [61]. PLS and PCR can be viewed as shrinkage methods, so they share interesting connections with RR. All three techniques shrink the regression coefficients away from the directions of low variation [62]. RR has been successfully applied to analyze genetic data especially in cases where the SNPs are in high linkage disequilibrium [63]. RR controls the variance inflation arising from multicollinearity by introducing a degree of bias to the regression estimates. The bias is introduced by imposing constraints on the regression parameters. It is therefore a very simple variant of standard MLR. As a reminder, below is the standard linear regression model:

$$Y = X\beta + \varepsilon \tag{11}$$

As discussed above, matrix *X* could represent a matrix of genotypes, while vector *y* could be a vector of phenotypes for subjects in the sample. The ordinary least squares estimator of β is then given by:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{12}$$

In RR, $\hat{\beta}$ in Equation (12) is replaced by RR estimator $\hat{\beta}^{\lambda}$ using the following modification to the OLS estimate [19]:

$$\hat{\beta}^{\lambda} = (X'X + \lambda I)^{-1}X'y \tag{13}$$

where λ is a positive number (usually 0 <*k* <1) and I is the *m x m* identity matrix. Comparison of this expression with Equation (12) reveals that a constant is added to the diagonal elements of the X'X matrix of the normal equations. Setting λ = 0, we retrieve the standard OLS estimate given in Equation (12).

Hoerl and Kennard [64] suggested selecting a value of λ by examining a ridge trace, which is a plot of the regression coefficients for different values of the ridge parameter. The value of λ ideally should be chosen at a point where the regression coefficients begin to stabilize and the root mean square error of prediction (RMSEP) begins to decrease. Other methods exist for choosing optimal values of the ridge parameter {e.g., Cule et al. proposed a simple modification to the trace plot that utilizes p-values of the regression coefficients rather than the coefficients themselves [19]}.

Shen et al. [65] recently introduced a generalized RR method for large *p*, small *n* problems (i.e., problems wherein the number of parameters significantly exceeds the number of observations). Such problems are typically encountered in genomic studies (e.g., GWAS, whole-genome sequencing). Their novel generalized RR method, referred to as the heteroscedastic effects model (HEM), is a more flexible version of standard RR that allows ridge parameter λ to vary across variables (SNPs in this case). The performance of HEM was compared to standard RR (which the authors demonstrated is equivalent to a linear mixed model/BLUP) using simulated datasets and real data. The latter was a publicly available genomic dataset consisting of 84 inbred *Arabidopsis* lines with ~200,000 SNPs, with the phenotype of interest being a bacteria-hypersensitive trait. HEM was shown to demonstrate superior performance to standard RR (i.e. SNP-BLUP) with respect to QTL mapping. The improvement in performance was attributed to HEM allowing λ to vary across SNPs, as opposed to SNP-BLUP, which assumes a common value. In addition to its superior performance, HEM was also shown to be highly computationally efficient. The computation of all SNP effects required <10 s using a single 2.7-GHz core. This reduction in run time could be particularly advantageous for whole-genome models requiring computationally-intensive permutation tests. HEM is implemented in R package *bigRR* [65].

3. Model validation and variable selection

3.1. Model validation

Validation of a multivariate calibration model in chemometrics is a vital step that must be performed prior to widespread adoption and use of the calibration model for routine analysis. The purpose of model validation is to determine the reproducibility of a multivariate calibration [66]. Similarly, in the context of genetic association studies, model validation is understood to be an indispensable step for establishing the scientific credibility of the results [67].

"Validation" has a number of different uses within the wider context of genetic association studies. For example, before any genetic association is investigated in itself, quality control of the data plays a critical role. Here, one step is the validation of the marker genotypes (i.e., the genetic information) [67]. Validation of genotypes is often understood as the concordance between different genotyping methods [67,68]. However, our focus in this article is within the context of validating statistical models.

The gold standard for model validation of any genetic study is replication in additional independent samples. Replication of a GWAS result should be thought of as the replication of a specific statistical model (i.e., a given SNP predicting a specific phenotype effect). The National Human Genome Research Institute (NHGRI) has outlined several criteria for establishing a positive replication [69], some of which are discussed below.

Sufficient sample size is critical in replication studies to detect the effect of the susceptibility allele. With replication, it is vital for the study to be well-powered to identify spuriously associated SNPs, or, in other words, to call the initial GWAS finding a false-positive result confidently [70].

Another important criterion for a good replication set is that it should be from the same population (i.e., same race/ethnicity) as that of the original GWAS,. Once the effect detected in the original GWAS has been confirmed in a similar but independent replication cohort, cohorts from other populations (e.g., races/ethnicities different from the original GWAS population) can be tested in order to determine if the effect is ethnicity-specific or not. Further, an identical phenotype should be used in both the original study and the replication studies. Lastly, a similar effect should also be seen in the replication set from the same SNP, or an SNP in high linkage disequilibrium (LD) with the GWAS-identified SNP. Overall, the general strategy for a replication study is to repeat the ascertainment and the design of the GWAS as closely as possible and then, consequently, to examine the specific genetic effects that were found to be significant in the GWAS [70]. Conforming to these general recommendations will facilitate both investigators to avoid bias and readers to evaluate reports (31).

3.2. Variable selection

One critical aspect in chemometrics is variable selection. As discussed in sub-section 2.2, for MLR, all available variables x_1, x_2, x_3,… x_m were used to build a linear model for the prediction of the *y* variable. This approach is useful as long as the number, *m*, of regressor variables is small (e.g., < 10). However, often in chemometrics, one has to deal with several hundred regressor variables. This consequently leads to problems, since ordinary least squares (OLS) regression is no longer computable in cases where the regressor variables are highly correlated, or where the number of objects is lower than the number of variables [71].

Although the widely-used regression methods, such as PCR and PLS, can handle such data without problems, there are several arguments as to why all available regressor variables should not be used. First, a regression model with large number of regressor variables is practically impossible to interpret. Second, reduction of the regressor variables can avoid the effects of overfitting and can lead to an improved prediction performance by removing irrelevant, noisy or unreliable variables [72]. Lastly, using a small number of regressor variables can considerably reduce the computational time [71]. There is much empirical evidence, which suggests that variable selection is a very important step when using methods such as PCR and PLS [73].

A number of variable-selection methods for regression-based calibration models are available, including genetic algorithms (GAs), interval PLS (*i*PLS*)*, PLS for discrimination, jack-knifing, variable importance for projection (VIP), forward-interval PLS, backward-interval PLS (biPLS), synergy-interval PLS (siPLS), selectivity ratio (SR), and LASSO-type methods [74–76].

Over the years, PLS has evolved to include several variable-selection variants, such as the jack-knifed PLS and multivariate-sparse PLS (M-SPLS) methods [77]. Jack-knifed PLS regression was utilized by Bjornstad and colleagues for a quantitative trait-loci (QTL) study in

which two matrices *X* (genetic markers) and *Y* (phenotypes) were decomposed into latent variables (PLS components or principal components) in a manner that enabled statistically sound and graphically interpretable model building. A very good feature of this approach is that it allows the simultaneous analysis of several traits, and the direct visualization of individuals with desirable marker genotypes [78].

Another variant of PLS-variable selection called M-SPLS was recently used for expression quantitative trait-loci (eQTL) analysis. Simulation studies were performed in order to assess the feasibility and the performance of M-SPLS as a dimension-reduction method for analyzing gene-expression and genomic marker data, particularly in the presence of multicollinearity. The results indicated that the technique was able to control type-I error adequately, while demonstrating increased power for the analysis of multiple transcripts by multiple response regression. It also had higher computation efficiency than standard techniques [79].

While the intent of this article is to focus on the applications of chemometric regression models in the context of chemometrics, we recommend that readers refer to several manuscripts [74,77] regarding mathematical details and characteristics of the different variable-selection methods. In general, all these variable-selection methods have been found to improve model performance and can be used to visualize which parts of the data are assessed as important and which parts are not [74]. Overall, in the context of GWAS, variable-selection methods are particularly useful to determine multiple variants (i.e., SNPs) or interactions between variants known to be causes of genetic susceptibility to a particular disease [80–83].

4. Conclusions, perspectives, and future directions

The applications of chemometric regression techniques in genetic epidemiology are relatively recent but growing in number. In the past few years, technological innovations led to increased feasibility of large-scale genetic association studies. Though densely-typed genetic markers (i.e., SNPs) can be generated using SNP arrays, next-generation technologies and imputation, SNPs typed using these techniques tend to be highly correlated due to linkage disequilibrium. Thus, standard MLR techniques are often inadequate for analyzing these data due to the inherent high dimensionality and complex correlation structure [19].

Since the advent of GWAS, thousands of common alleles and variants have been implicated in disease susceptibility. However, for most heritable diseases and complex traits, the common variants identified so far collectively explain only a small fraction of the total inferred genetic variance. This is the so-called “missing heritability problem” [84]. One proposed source of the missing heritability are RVs [84]. In the context of SNPs, RVs are those with a low allele frequency (typically <1%) in the population of interest. It is believed that such rare alleles may have a large impact on disease susceptibility; thus, over the last few years, there was a push to elucidate the role of rare alleles in disease susceptibility [84]. However, one key problem is that standard GWAS techniques are unsuited for analyzing variants with such low frequency so various novel methods were recently developed for testing RV association. Classical multivariate chemometric methods are promising for RV analysis, but they have received little attention. Xu and colleagues recently compared some chemometric techniques for testing RV associations. They found that RR, PCR, PLS, and sparse PLS are all adequately powered to detect associations of rare genetic variants with disease susceptibility, and can even substantially outperform several popular methods for RV analysis (e.g., burden tests) [85]. Studies have shown that most human genetic variants are rare, and that RVs are (statistically) more likely than common variants to

affect disease susceptibility [86]. With the increasing feasibility and affordability of whole-genome sequencing that captures a wide spectrum of rare human genetic variation, RV analysis will continue to be an area of intense interest in genetic epidemiology. Chemometric methods have been shown to provide excellent performance for RV analysis and are a promising avenue of further research.

Newer trends in chemometrics, such as the ANOVA-simultaneous component analysis (ASCA), orthogonal projections to latent structures (OPLS), recursive weighted PLS (rPLS) and data-fusion methods, can potentially have significant applications in GWAS. Similar to PCA, ASCA [87] is another interesting technique that has potential applications in genetic epidemiology, as in cases of modeling the relationship between genotype and/or samples to detect group differences. Such cases are characterized by datasets containing underlying factors, such as time, doses, or combinations thereof [87] The application of OPLS to genetic data analysis has started to evolve. OPLS-discriminant analysis (OPLS-DA) [88], in particular, has been useful for the selection of a set of gene transcripts discriminating two different types of mutated cells in a certain disease [89]. The main advantage of using OPLS is that the model results in a reduced complexity while retaining prediction ability [88].

rPLS is another chemometric method that has an interesting application in genetic epidemiology. It involves iteratively reweighing the variables using the regression coefficients calculated by PLS. In contrast to other variable-selection methods, it has the advantage in that only one parameter needs to be estimated: the number of latent factors used in the PLS model [90]. In the context of GWAS, this is particularly useful in converging to a very limited set of variables (e.g., genetic markers) that is useful for interpretation and prediction of a specific disease phenotype. Lastly, data fusion is a subclass of chemometrics procedures, wherein data or models are combined into one contiguous fused entity, which has proved useful in many disciplines. The value of data fusion can be attributed to its ability to fuse together different sources of information that may be measuring different parts of a process [91]. In the context of GWAS, data fusion is particularly useful in incorporating genetic, environmental, proteomics and transcriptomic data for use in predicting a specific disease phenotype. This can provide a holistic model predictive of an individual's disease state.

It is clear that chemometrics is rapidly growing and has major applications in GWAS. By integrating genetic epidemiology with chemometrics, it is anticipated that novel genetic variants with high phenotypic manifestations will be unveiled. Further, the growing application of well-known chemometric techniques, such as PLS and PCR, to GWAS promises more robust approaches for genotype-phenotype disease modeling.

This review highlights a number of chemometric regression algorithms that have been successfully applied to the problem of testing genotype-phenotype associations, with the ultimate goal of unraveling the etiology of specific diseases by identifying the genes that increase susceptibility to them. GWAS are the standard and popular approach to testing genotype-phenotype relationships, but these studies are complicated by multiple statistical issues as a consequence of having more predictors/variables than units of observations. The use of separate univariate tests to analyze each predictor is the common solution to this particular problem, and other related problems arising from high dimensionality. However, this simplistic approach has multiple disadvantages that we discussed in this review. Chemometric regression algorithms, such as PLS and its variants, offer powerful alternatives, and have been successfully employed in multiple studies. These algorithms generally outperform the univariate regression techniques used in GWAS. More importantly, they are able to exploit multivariate information (e.g., the

effect of gene-gene interactions on the phenotype), which eludes the common univariate approaches. Further, such algorithms are computationally efficient and reduce noise that would typically confound standard regression techniques.

With the recent development of more powerful supercomputers and next-generation sequencing techniques, it is reasonable to anticipate ever-increasing efforts to develop and to refine increasingly sophisticated techniques for mining relationships between genotypes and phenotypes. These efforts should lead to the discovery of rare and novel disease-associated genes or biomarkers that have so far eluded identification by traditional techniques. This may be expected to provide a more complete mapping of the genetic architecture of many diseases. The eventual exploitation of this knowledge to reveal new therapeutic avenues should have an immense positive impact on the lives of individuals affected by (or susceptible to) these diseases.

The application of chemometric techniques in GWAS is a relatively new area of pursuit in genetic epidemiology. The future of chemometrics as a distinct discipline has been called into question, particularly its application to the physical sciences [92]. However, the recent and growing applications of chemometric techniques in the area of genetic epidemiology indicate that it may be poised to play a more prominent role in the future of genomics.

References

[1] D.C. Thomas, *Overview of genetic epidemiology*, in *Statistical methods in genetic epidemiology*. 2004, Oxford University Press: New York.

[2] J.A. Last, *A dictionary of epidemiology*. 1993, Oxford: Oxford University Press.

[3] T. Casci, *Population genetics: SNPs that come in threes.* Nature Reviews Genetics, 2010. **11**(8).

[4] V.M. Lourenco, A.M. Pires, and M. Kirst, *Robust linear regression methods in association studies.* Bioinformatics, 2011. **27**(6): p. 815-21.

[5] T. Mehmood, H. Martens, S. Saebo, J. Warringer, and L. Snipen, *Mining for genotype-phenotype relations in Saccharomyces using partial least squares.* BMC Bioinformatics, 2011. **12**: p. 318.

[6] J.J. Hox, *Multivariate multilevel regression models*, in *Multilevel analysis: techniques and applications*. 2010, Routledge: Great Britain. p. 188.

[7] Q. He and D.Y. Lin, *A variable selection method for genome-wide association studies.* Bioinformatics, 2011. **27**(1): p. 1-8.

[8] H. Martens and T. Naes, *Multivariate calibration*. 1989, New York: Wiley.

[9] R. Kramer, *Chemometric Techniques for Quantitative Analysis*. 1998, New York: Taylor and Francis/Marcel Dekker.

[10] C. Wellcome Trust Case Control, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

[11] S. Van der Sluis, C.V. Dolan, J. Li, Y. Song, P. Sham, D. Posthuma, and M.X. Li, *MGAS: a powerful tool for multivariate gene-based genome-wide association analysis.* Bioinformatics, 2015. **31**(7): p. 1007-15.

[12] S. Wellek and A. Ziegler, *Cochran-Armitage test versus logistic regression in the analysis of genetic association studies.* Hum Hered, 2012. **73**(1): p. 14-7.

[13] W. Zhu and H. Zhang, *Why Do We Test Multiple Traits in Genetic Association Studies?* Journal of the Korean Statistical Society, 2009. **38**(1): p. 1-10.

[14] R. Fisher, *A New Test for 2 x 2 Tables.* Nature, 1945. **156**.
[15] H. Huang, P. Chanda, A. Alonso, J.S. Bader, and D.E. Arking, *Gene-based tests of association.* PLoS Genet, 2011. **7**(7): p. e1002177.
[16] M.X. Li, H.S. Gui, J.S. Kwan, and P.C. Sham, *GATES: a rapid and powerful gene-based association test using extended Simes procedure.* Am J Hum Genet, 2011. **88**(3): p. 283-93.
[17] J.Z. Liu, A.F. McRae, D.R. Nyholt, S.E. Medland, N.R. Wray, K.M. Brown, A. Investigators, N.K. Hayward, G.W. Montgomery, P.M. Visscher, N.G. Martin, and S. Macgregor, *A versatile gene-based test for genome-wide association studies.* Am J Hum Genet, 2010. **87**(1): p. 139-45.
[18] D. Ruano, G.R. Abecasis, B. Glaser, E.S. Lips, L.N. Cornelisse, A.P. de Jong, D.M. Evans, G. Davey Smith, N.J. Timpson, A.B. Smit, P. Heutink, M. Verhage, and D. Posthuma, *Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability.* Am J Hum Genet, 2010. **86**(2): p. 113-25.
[19] E. Cule, P. Vineis, and M. De Iorio, *Significance testing in ridge regression for genetic data.* BMC Bioinformatics, 2011. **12**: p. 372.
[20] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, J.H. Cho, A.E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C.N. Rotimi, M. Slatkin, D. Valle, A.S. Whittemore, M. Boehnke, A.G. Clark, E.E. Eichler, G. Gibson, J.L. Haines, T.F. Mackay, S.A. McCarroll, and P.M. Visscher, *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
[21] K. Xia, Y. Yu, M. Ahn, H. Zhu, F. Zou, J.H. Gilmore, and R.C. Knickmeyer, *Environmental and genetic contributors to salivary testosterone levels in infants.* Front Endocrinol (Lausanne), 2014. **5**: p. 187.
[22] P. Buzkova, *Linear regression in genetic association studies.* PLoS One, 2013. **8**(2): p. e56976.
[23] 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, 2014. **9**(4): p. e95923.
[24] D.B. Allison, B. Thiel, P. St Jean, R.C. Elston, M.C. Infante, and N.J. Schork, *Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages.* American journal of human genetics, 1998. **63**(4): p. 1190-201.
[25] L. Klei, D. Luca, B. Devlin, and K. Roeder, *Pleiotropy and principal components of heritability combine to increase power for association analysis.* Genetic epidemiology, 2008. **32**(1): p. 9-19.
[26] S. Chavali, F. Barrenas, K. Kanduri, and M. Benson, *Network properties of human disease genes with pleiotropic effects.* BMC systems biology, 2010. **4**: p. 78.
[27] M.A. Ferreira and S.M. Purcell, *A multivariate test of association.* Bioinformatics, 2009. **25**(1): p. 132-3.
[28] P.F. O'Reilly, C.J. Hoggart, Y. Pomyen, F.C. Calboli, P. Elliott, M.R. Jarvelin, and L.J. Coin, *MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS.* PLoS One, 2012. **7**(5): p. e34861.

[29] S. van der Sluis, D. Posthuma, and C.V. Dolan, *TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies.* PLoS genetics, 2013. **9**(1): p. e1003235.

[30] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nature genetics, 2007. **39**(7): p. 906-13.

[31] Y. Guan and M. Stephens, *Practical issues in imputation-based association mapping.* PLoS genetics, 2008. **4**(12): p. e1000279.

[32] R. Berk, *Regression analysis: A constructive critique*. 2004, Thousand Oaks, CA: Sage.

[33] P.D. Allison, *Logistic regression using SAS: theory and application*. 2nd ed. 2012, Cary, NC: SAS Institute, Inc,.

[34] R.J. Rossi, *Assumptions of the simple linear regression model*, in *Applied biostatistics for the health sciences*, R.J. Rossi, Editor. 2010, John Wiley and Sons: NJ. p. 343.

[35] K.L. Sainani, *Introduction to principal components analysis.* PM & R : the journal of injury, function, and rehabilitation, 2014. **6**(3): p. 275-8.

[36] H. Abdi and L.J. Williams, *Principal component analysis.* Wiley Interdisciplinary Reviews: Computational Statistics, 2010. **2**(4): p. 433-459.

[37] G. Laffaye, B.G. Bardy, and A. Durey, *Principal component structure and sport-specific differences in the running one-leg vertical jump.* International journal of sports medicine, 2007. **28**(5): p. 420-5.

[38] D.H. Ballard, J. Cho, and H. Zhao, *Comparisons of multi-marker association methods to detect association between a candidate region and disease.* Genet Epidemiol, 2010. **34**(3): p. 201-12.

[39] W.J. Gauderman, C. Murcray, F. Gilliland, and D.V. Conti, *Testing association between disease and multiple SNPs in a candidate gene.* Genet Epidemiol, 2007. **31**(5): p. 383-95.

[40] C. Preda and G. Saporta, *PLS regression on a stochastic process.* Computational Statistics & Data Analysis, 2005. **48**(1): p. 149-158.

[41] P.D. Wentzell and L.V. Montoto, *Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures.* Chemometrics and Intelligent Laboratory Systems, 2003. **65**(2): p. 257-279.

[42] P. Yaroshchyk, D.L. Death, and S.J. Spencer, *Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS.* Journal of Analytical Atomic Spectrometry, 2012. **27**(1): p. 92-98.

[43] S. Wold, M. Sjostrom, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 109-130.

[44] H. Abdi, W.W. Chin, V.E. Vinzi, G. Russolillo, and L. Trinchera, *New perspectives in partial least squares and related methods*. 2013: Springer.

[45] P. Geladi and B.R. Kowalski, *Partial Least-Squares Regression - a Tutorial.* Analytica Chimica Acta, 1986. **185**: p. 1-17.

[46] C. Cassel, P. Hackl, and A. Westlund, *Robustness of partial least-squares method for estimating latent variable quality structures.* Journal of Applied Statistics, 1999. **26**(4): p. 435-446.

[47] T.H. Meuwissen, B.J. Hayes, and M.E. Goddard, *Prediction of total genetic value using genome-wide dense marker maps.* Genetics, 2001. **157**(4): p. 1819-29.

[48] D. Habier, R.L. Fernando, and J.C. Dekkers, *The impact of genetic relationship information on genome-assisted breeding values.* Genetics, 2007. **177**(4): p. 2389-97.

[49] S. Xu, *Estimating polygenic effects using markers of the entire genome.* Genetics, 2003. **163**(2): p. 789-801.

[50] V. Vapnik, *Statistical Learning Theory*. 1998, New York: John Wiley & Sons.

[51] D. Gianola, R.L. Fernando, and A. Stella, *Genomic-assisted prediction of genetic value with semiparametric procedures.* Genetics, 2006. **173**.

[52] J. Bennewitz, T. Solberg, and T. Meuwissen, *Genomic breeding value estimation using nonparametric additive regression models.* Genet Sel Evol, 2009. **41**: p. 20.

[53] R. Crump, B. Tier, G. Moser, J. Solkner, R. Kerr, A. Woolaston, K. Zenger, M. Khatkar, J. Cavanagh, and H. Raadsma, *Genomewide selection in dairy cattle: use of genetic algorithms in the estimation of molecular breeding values.* Proc Assoc Advmt Anim Breed Genet, 2007. **17**.

[54] G. Moser, B. Tier, R.E. Crump, M.S. Khatkar, and H.W. Raadsma, *A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers.* Genet Sel Evol, 2009. **41**: p. 56.

[55] C. Colombani, P. Croiseau, S. Fritz, F. Guillaume, A. Legarra, V. Ducrocq, and C. Robert-Granie, *A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle.* J Dairy Sci, 2012. **95**(4): p. 2120-31.

[56] M. Sarkis, K. Diepold, and F. Westad, *A new algorithm for gene mapping: Application of partial least squares regression with cross model validation*, in *Genomic Signal Processing and Statistics*. 2006, IEEE: College Station, TX. p. 89-90.

[57] A.S. Turkmen and S. Lin, *Gene-based partial least-squares approaches for detecting rare variant associations with complex traits.* BMC Proc, 2011. **5 Suppl 9**: p. S19.

[58] D.W. Mount, *Using the Basic Local Alignment Search Tool (BLAST).* CSH protocols, 2007. **2007**: p. pdb top17.

[59] M.J. Adams, *Chemometrics in analytical spectroscopy*. 2004: Royal Society of Chemistry.

[60] R. de Vlaming and P.J. Groenen, *The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics.* 2014.

[61] D.L. Massart, B.G. Vandeginste, L. Buydens, P. Lewi, and J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: Part A*. 1997: Elsevier Science Inc.

[62] R. Wehrens, *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. 2011: Springer Science & Business Media.

[63] N. Malo, O. Libiger, and N.J. Schork, *Accommodating linkage disequilibrium in genetic-association analyses via ridge regression.* Am J Hum Genet, 2008. **82**(2): p. 375-85.

[64] A.E. Hoerl and R.W. Kennard, *Ridge Regression - Biased Estimation for Nonorthogonal Problems.* Technometrics, 1970. **12**(1): p. 55-&.

[65] X. Shen, M. Alam, F. Fikse, and L. Ronnegard, *A novel generalized ridge regression method for quantitative genetics.* Genetics, 2013. **193**(4): p. 1255-68.

[66] P. Gemperline, *Practical guide to chemometrics*. 2006: CRC press.

[67] I.R. Konig, *Validation in genetic association studies.* Briefings in bioinformatics, 2011. **12**(3): p. 253-8.

[68] S. Dong, E. Wang, L. Hsie, Y. Cao, X. Chen, and T.R. Gingeras, *Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation.* Genome research, 2001. **11**(8): p. 1418-24.

[69] N.-N.W.G.o.R.i.A. Studies, S.J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D.J. Hunter, G. Thomas, J.N. Hirschhorn, G. Abecasis, D. Altshuler, J.E. Bailey-Wilson, L.D. Brooks, L.R. Cardon, M. Daly, P. Donnelly, J.F. Fraumeni, Jr., N.B. Freimer, D.S. Gerhard, C. Gunter, A.E. Guttmacher, M.S. Guyer, E.L. Harris, J. Hoh, R. Hoover, C.A. Kong, K.R. Merikangas, C.C. Morton, L.J. Palmer, E.G. Phimister, J.P. Rice, J. Roberts, C. Rotimi, M.A. Tucker, K.J. Vogan, S. Wacholder, E.M. Wijsman, D.M. Winn, and F.S. Collins, *Replicating genotype-phenotype associations.* Nature, 2007. **447**(7145): p. 655-60.

[70] W.S. Bush and J.H. Moore, *Chapter 11: Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.

[71] K. Varmuza and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*. 2009: CRC press.

[72] C.M. Andersen and R. Bro, *Variable selection in regression—a tutorial.* Journal of Chemometrics, 2010. **24**(11-12): p. 728-737.

[73] R. Leardi, *Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks: genetic algorithms and artificial neural networks*. Vol. 23. 2003: Elsevier.

[74] C. Andersen and R. Bro, *Variable selection in regression - a tutorial.* J Chemom, 2010. **24**(11-12).

[75] M. Barker and W. Rayens, *Partial least squares for discrimination.* J Chemom, 2003. **17**(3): p. 166-173.

[76] L. Norgaard, A. Saudland, J. Wagner, J. Nielsen, L. Munck, and S. Engelsen, *nterval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy.* Applied Spectroscopy, 2000. **54**(3): p. 413-419.

[77] I. Karaman, E. Qannari, H. Martens, M. Hedemann, K. Knudsen, and A. Kohler, *Comparison of Sparse and Jack-knife partial least squares regression methods for variable selection.* Chemometrics and Intelligent Laboratory Systems, 2013. **122**: p. 65-77.

[78] A. Bjornstad, F. Westad, and H. Martens, *Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (PLSR).* Hereditas, 2004. **141**(2): p. 149-65.

[79] H. Chun and S. Keles, *Expression quantitative trait loci mapping with multivariate sparse partial least squares regression.* Genetics, 2009. **182**(1): p. 79-90.

[80] M. Mooney, B. Wilmot, T. Bipolar Genome Study, and S. McWeeney, *The GA and the GWAS: using genetic algorithms to search for multilocus associations.* IEEE/ACM Trans Comput Biol Bioinform, 2012. **9**(3): p. 899-910.

[81] S. Hong, Y. Kim, and T. Park, *Practical issues in screening and variable selection in genome-wide association analysis.* Cancer Inform, 2014. **13**(Suppl 7): p. 55-65.

[82] P. Waldmann, G. Meszaros, B. Gredler, C. Fuerst, and J. Solkner, *Evaluation of the lasso and the elastic net in genome-wide association studies.* Front Genet, 2013. **4**: p. 270.

[83] K.A. Le Cao, S. Boitard, and P. Besse, *Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems.* BMC Bioinformatics, 2011. **12**: p. 253.

[84] E.T. Cirulli and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nature Review Genetics, 2010. **11**(6): p. 415-425.

[85] C. Xu, M. Ladouceur, Z. Dastani, J.B. Richards, A. Ciampi, and C.M. Greenwood, *Multiple regression methods show great potential for rare variant association tests.* PLoS One, 2012. **7**(8): p. e41694.

[86] M.R. Nelson, D. Wegmann, M.G. Ehm, D. Kessner, P. St. Jean, C. Verzilli, J. Shen, Z. Tang, S. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M.D. Hall, K. Nangle, J. Wang, G. Abecasis, L.R. Cardon, and S. Zollner, *An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People.* Science, 2012. **337**(6090): p. 100-104.

[87] A.K. Smilde, J.J. Jansen, H.C. Hoefsloot, R.J. Lamers, J. van der Greef, and M.E. Timmerman, *ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data.* Bioinformatics, 2005. **21**(13): p. 3043-8.

[88] J. Trygg and S. Wold, *Orthogonal projections to latent structures (O-PLS).* Journal of Chemometrics, 2002. **16**(3): p. 119-128.

[89] G. Musumarra, D.F. Condorelli, and C.G. Fortuna, *OPLS-DA as a suitable method for selecting a set of gene transcripts discriminating RAS- and PTPN11-mutated cells in acute lymphoblastic leukaemia.* Comb Chem High Throughput Screen, 2011. **14**(1): p. 36-46.

[90] A. Rinnan, M. Andersson, C. Ridder, and S.B. Engelsen, *Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS.* Journal of Chemometrics, 2014. **28**(5): p. 439-447.

[91] C. Ovalles and C. Rechsteiner, *Analytical Methods in Petroleum Upstream Applications*. 2015: CRC Press. 337.

[92] P. Geladi and P.K. Hopke, *Editorial: Is there a future for chemometrics? Are we still needed?* Journal of Chemometrics, 2008. **22**: p. 289-290.

[93] N.J. Samani, J. Erdmann, A.S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R.J. Dixon, T. Meitinger, P. Braund, H.E. Wichmann, J.H. Barrett, I.R. Konig, S.E. Stevens, S. Szymczak, D.A. Tregouet, M.M. Iles, F. Pahlke, H. Pollard, W. Lieb, F. Cambien, M. Fischer, W. Ouwehand, S. Blankenberg, A.J. Balmforth, A. Baessler, S.G. Ball, T.M. Strom, I. Braenne, C. Gieger, P. Deloukas, M.D. Tobin, A. Ziegler, J.R. Thompson, H. Schunkert, Wtccc, and C. the Cardiogenics, *Genomewide association analysis of coronary artery disease.* N Engl J Med, 2007. **357**(5): p. 443-53.

[94] K. Askland, C. Read, and J. Moore, *Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission.* Hum Genet, 2009. **125**(1): p. 63-79.

[95] A. Coster and M.P. Calus, *Partial least square regression applied to the QTLMAS 2010 dataset.* BMC Proc, 2011. **5 Suppl 3**: p. S7.

[96] F. Zhang, X. Guo, and H.W. Deng, *Multilocus association testing of quantitative traits based on partial least-squares analysis.* PLoS One, 2011. **6**(2): p. e16739.

[97] T. Mehmood, J. Warringer, L. Snipen, and S. Saebo, *Improving stability and understandability of genotype-phenotype mapping in Saccharomyces using regularized variable selection in L-PLS regression.* BMC bioinformatics, 2012. **13**: p. 327.

[98] I.R. White, K. Patel, W.T. Symonds, A. Dev, P. Griffin, N. Tsokanas, M. Skehel, C. Liu, A. Zekry, P. Cutler, M. Gattu, D.C. Rockey, M.M. Berrey, and J.G. McHutchison, *Serum proteomic analysis focused on fibrosis in patients with hepatitis C virus infection.* J Transl Med, 2007. **5**: p. 33.

[99] H. Mei, W. Chen, A. Dellinger, J. He, M. Wang, C. Yau, S.R. Srinivasan, and G.S. Berenson, *Principal-component-based multivariate regression for genetic association studies of metabolic syndrome components.* BMC Genet, 2010. **11**: p. 100.

[100] S.D. Pant, F.S. Schenkel, C.P. Verschoor, Q. You, D.F. Kelton, S.S. Moore, and N.A. Karrow, *A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle.* Genomics, 2010. **95**(3): p. 176-82.

[101] K. Wang and D. Abbott, *A principal components regression approach to multilocus genetic association studies.* Genet Epidemiol, 2008. **32**(2): p. 108-18.

[102] D.P. Hibar, N. Jahanshad, J.L. Stein, O. Kohannim, A.W. Toga, S.E. Medland, N.K. Hansell, K.L. McMahon, G.I. de Zubicaray, G.W. Montgomery, N.G. Martin, M.J. Wright, and P.M. Thompson, *Alzheimer's disease risk gene, GAB2, is associated with regional brain volume differences in 755 young healthy twins.* Twin research and human genetics : the official journal of the International Society for Twin Studies, 2012. **15**(3): p. 286-95.

[103] D.P. Hibar, J.L. Stein, O. Kohannim, N. Jahanshad, A.J. Saykin, L. Shen, S. Kim, N. Pankratz, T. Foroud, M.J. Huentelman, S.G. Potkin, C.R. Jack, Jr., M.W. Weiner, A.W. Toga, and P.M. Thompson, *Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects.* NeuroImage, 2011. **56**(4): p. 1875-91.

[104] N. Malo, O. Libiger, and N.J. Schork, *Accommodating linkage disequilibrium in genetic-association analyses via ridge regression.* American journal of human genetics, 2008. **82**(2): p. 375-85.

[105] E. Pimentel, S. Queiroz, R. Carvalheiro, and L. Fries, *Use of ridge regression for the prediction of early growth performance in crossbred calves.* Genetics and Molecular Biology, 2007. **30**(3): p. 536-544.

[106] O. Kohannim, D.P. Hibar, J.L. Stein, N. Jahanshad, J. Jack, C.R., M.W. Weiner, A.W. Toga, and P.M. Thompson, *Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression.* Biomedical Imaging, 2011: p. 1855-1859.

[107] L.S. Chen, C.M. Hutter, J.D. Potter, Y. Liu, R.L. Prentice, U. Peters, and L. Hsu, *Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data.* American journal of human genetics, 2010. **86**(6): p. 860-71.

[108] G. de Los Campos, A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen, *Prediction of complex human traits using the genomic best linear unbiased predictor.* PLoS genetics, 2013. **9**(7): p. e1003608.

[109] V.M. Roso, F.S. Schenkel, S.P. Miller, and L.R. Schaeffer, *Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation.* Journal of animal science, 2005. **83**(8): p. 1788-800.

Table 1.
Concise summary of selected applications of various chemometric regression techniques in statistical genetics

Technique	Application	Field	Phenotype	Variables	Comments	Ref.
Cochran-Armitage test	GWAS of Coronary Artery Disease	Human Disease Genetics	Coronary Artery Disease	SNPs	An additive genetic model was assumed, and the Cochran-Armitage trend test gave conservative results	[93]
Fisher's exact test	Pathway-based whole-genome association study	Human Disease Genetics	Bipolar Disorder	SNPs	Pathway-based analysis were carried out, with comparisons done using Fisher's exact test	[94]
Partial least squares (PLS)	Prediction of molecular breeding values (MBVs)	Animal Breeding / Genetics	Molecular Breeding Value	SNPs	Lowest computational time than linear, Bayesian, and support vector regression	[54],[55]
	Quantitative trait loci (QTL) analysis	Crop Science	Tomato weight	Genetic markers	Provides statistically reliable and graphically interpretable model building	[78],[95]
	Expression QTL mapping (using multi-	Disease Genetics	Obesity & diabetes measures in mice	Gene expression measurements	Computationally efficient method for handling multicollinear	[79]

	variate sparse PLS)				ity and controlling Type I error	
	Multilocus association testing (MLAS) of quantitative traits	Disease Genetics	Lean body mass in humans	SNPs	Improves power of PLS-based MLAS in disease genes mapping relative to other approaches	[96]
	Genotype-phenotype mapping (L-PLS)	Yeast Genetics	Yeast genotype-phenotype mapping	Gene variations	Improves the stability of selected genes and background information	[97]
PLS Discriminant Analysis (PLS-DA)	Proteomic analysis for identification of potential disease biomarkers	Proteomics	Extent of Fibrosis (no/mild vs. advanced)	Protein expression profile	Can be used to rank individual differences in protein expression (or some other quantitative measure) between 2 or more groups.	[98]
	Gene mapping	Genetics	Disease Status (cases versus controls)	SNPs	Combined with cross-model validation, this technique offers a powerful method for identifying causal SNPs	[56]
Principal Component Regression (PCR)	Gene pleiotro	Genetics of Obesity	Obesity-related traits	SNPs	Yields increased power while retaining computational efficiency	[99]
	QTL	Bovine	MAP	SNPs	Provides	[100]

	identification	Genomics	(*Mycobacterium avium* spp. *Paratuberculosis*) infection		superior performance in the presence of linkage disequilibrium	
	Multilocus genetic association	Human Genetics	Gene Expression level	SNPs	Higher power than some popular methods	[101]
	Genetic associations with morphological brain differences	Neuro-genetics	Regional brain volume difference	SNPs	Demonstrates higher statistical power than univariate statistical methods	[102]
	Genetic association study of voxel-level volume differences in the brains of Alzheimer's patients	Imaging Genomics	Voxel-level volume difference	SNPs	Boosts power and reduces number of statistical tests	[103]
Ridge Regression	Genetic association studies	Animal/ Plant genetics	Disease/Trait	SNPs	In the presence of strong linkage disequilibrium (LD) among SNPs, this approach is able to tease out independent effects on the phenotype	[104]
	Gene-environment interactions	Animal Breeding	Pre-weaning average daily gain	Direct and maternal genetic effects, age, sex, date of birth	Resolves multicollinearity and allows accommodates complex interactions	[105]

					among predictor variables	
	GWAS of brain data from Alzheimer's Disease	Neuro-genetics	Hippocampal and temporal lobe volume measures	SNPs	Yields more significant associations than univariate analysis	[106]
	Gene-set analysis of colon cancer data	Human Disease Genetics	Disease status (colon cancer)	eigenSNPs: SNP sets collapsed into gene-based groups	Performs well in situations wherein the number of predictors exceeds the sample size.	[107]
	Prediction of complex traits and diseases in plants, animals and humans	Plant and Animal Breeding	Disease/trait	SNPs	Demonstrates good predictive performance when applied to human populations	[108]
	Genetic effect estimation in multibreed beef cattle evaluation	Animal Science	Pre-weaning weight gain of calves	Fixed genetic effects	More effective than ordinary least squares regression at decreasing multicollinearity	[109]