5-7-2002

Principal Components Regression with Data-Chosen Components and Related Methods

J. T. Hwang
Cornell University

Dan Nettleton
Iowa State University, dnett@iastate.edu

Principal Components Regression with Data-Chosen Components and Related Methods

Abstract

Multiple regression with correlated explanatory variables is relevant to a broad range of problems in the physical, chemical, and engineering sciences. Chemometricians, in particular, have made heavy use of principal components regression and related procedures for predicting a response variable from a large number of highly correlated variables. In this paper we develop a general theory for selecting principal components that yield estimates of regression coefficients with low mean squared error. Our numerical results suggest that the theory also can be used to improve partial least squares regression estimators and regression estimators based on rotated principal components. Although our work has been motivated by the statistical genetics problem of mapping quantitative trait loci, the results are applicable to any problem where estimation of regression coefficients for correlated explanatory variables is of interest.

Disciplines

Genetics | Statistical Theory | Statistics and Probability

Principal Components Regression with Data-Chosen Components and Related Methods

J. T. Gene Hwang

Department of Mathematics
Cornell University
Ithaca, NY 14853-4201
(*hwang@math.cornell.edu*)

Dan Nettleton

Department of Statistics
Iowa State University
Ames, IA 50011-1210
(*dnett@iastate.edu*)

May 7, 2002

Multiple regression with correlated explanatory variables is relevant to a broad range of problems in the physical, chemical, and engineering sciences. Chemometricians, in particular, have made heavy use of principal components regression and related procedures for predicting a response variable from a large number of highly correlated variables. In this paper we develop a general theory for selecting principal components that yield estimates of regression coefficients with low mean squared error. Our numerical results suggest that the theory also can be used to improve partial least squares regression estimators and regression estimators based on rotated principal components. Although our work has been motivated by the statistical genetics problem of mapping quantitative trait loci, the results are applicable to any problem where estimation of regression coefficients for correlated explanatory variables is of interest.

KEY WORDS: Biased regression; Eigenvalues; Mean squared error; Multicollinearity; Partial least squares; Quantitative trait loci; Varimax rotation.

1. INTRODUCTION

It is well known that the ordinary least squares regression coefficient estimator may perform poorly when there are near multicollinearities in $\mathbf{X}$ the matrix of explanatory variables. The variance of the ordinary least squares estimator becomes inflated when one or more eigenvalues of the matrix of explanatory variables are close to zero. This results in an estimator that may have low probability of being close to the true value of the vector of regression coefficients $\boldsymbol{\beta}$.

There are a wealth of proposals in the statistics literature for combating this problem. Principal components regression (PCR) and partial least squares regression (Wold, 1966) are two related families of methods that are often used in chemometrics. Both involve selecting a subspace of the column space of $\mathbf{X}$ on which to project the response vector $\mathbf{Y}$. The two families of methods differ in the subspaces that they consider. Principal components regression considers subspaces spanned by subsets of the principal components of $\mathbf{X}$. Partial least squares regression considers subspaces spanned by subsets of the partial least squares components, which depend on both $\mathbf{Y}$ and $\mathbf{X}$.

It is common in each method to regress $\mathbf{Y}$ against the first k components where k is determined by leave-one-out cross validation. For partial least squares regression (PLSR), this seems sensible intuitively because the first k PLS components are by design the ones most relevant to $\mathbf{Y}$. The first k principal components, however, correspond to the largest k eigenvalues and are constructed independently of $\mathbf{Y}$. Restricting attention to principal components with the largest eigenvalues helps to control variance inflation but can introduce high bias by discarding components with small eigenvalues that may be most associated with $\mathbf{Y}$. Jollife (1982) provided several real-life examples where the principal components corresponding to small eigenvalues had high correlation with $\mathbf{Y}$. Hadi and Ling (1998) provided an example where only the principal component associated with the smallest eigenvalue was correlated with $\mathbf{Y}$.

An alternative approach to PCR, called the inferential approach, uses only the set of principal components whose regression coefficients are significantly different from zero. See, for example, Massey (1965)

and Mason and Gunst (1985). In this paper we present methods for choosing a subset of components that attempt to minimize the mean squared error (MSE) of the estimator of $\boldsymbol{\beta}$. We show that a component's regression coefficient being significantly different from zero is not sufficient to warrant its use for estimating $\boldsymbol{\beta}$. Rather a component's coefficient must be far enough from zero to ensure that its bias reduction benefit will outweigh its variance inflation liability. Using this criterion we show how to improve upon several types of estimators and develop a new PCR estimator that exhibits substantially lower MSE than other commonly used methods based on principal components and partial least squares components.

In Section 2 of this paper we develop a general theory for component selection to minimize the MSE of the $\boldsymbol{\beta}$ estimator. Given a particular set of basis vectors for the column space of $\mathbf{X}$ (i.e., components), we derive the subset of the basis vectors that leads to an estimator of $\boldsymbol{\beta}$ with lowest MSE. This optimal subspace depends on the unknown error variance σ^2 and the unknown regression coefficients that we are attempting to estimate. We investigate several methods of using the data to approximate the optimal subspace and thereby approximate the optimal estimator. We explore the implications of our general theory for regression on principal components, varimax rotation of principal components, and partial least squares components in Sections 3, 4 and 5, respectively. Section 6 contains a description of several estimators whose performance is examined via a simulation study described in Section 7.

Our work is motivated by a problem in statistical genetics where estimates of regression coefficients on correlated explanatory variables provide information about the location and effect of genetic regions associated with quantitative traits. In this problem prediction of the response is far less important than estimating $\boldsymbol{\beta}$ well. We discuss the application in greater detail in Section 8. The paper concludes with a discussion of results.

2. GENERAL DEVELOPMENT

Consider the standard regression model $\mathbf{Y}^0 = \alpha\mathbf{1} + \mathbf{X}^0\boldsymbol{\beta} + \boldsymbol{\varepsilon}^0$, where $\mathbf{Y}^0$ is an n-dimensional response vector, $\mathbf{X}^0$ is an $n \times p$ predictor matrix, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression parameters,

ε^0 is a random vector satisfying $E(\varepsilon^0) = \mathbf{0}$ and $\mathrm{Var}(\varepsilon^0) = \sigma^2 I$ for some unknown $\sigma^2 > 0$, and $\mathbf{1}$ represents the n-dimensional column vector of ones. We center each variable via $\mathbf{Y} = \mathbf{Y}^0 - \mathbf{1}\mathbf{1}'\mathbf{Y}^0/n$ and $\mathbf{X} = \mathbf{X}^0 - \mathbf{1}\mathbf{1}'\mathbf{X}^0/n$ (so the sum of $\mathbf{Y}$ and the sum of each column $\mathbf{X}$ is zero), and write the model as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where $\varepsilon = \varepsilon^0 - \mathbf{1}\mathbf{1}'\varepsilon^0/n$.

Our goal is to estimate $\boldsymbol{\beta}$. Let r denote the rank of $\mathbf{X}$. The full-column-rank case ($r = p$) is most relevant because all components of $\boldsymbol{\beta}$ are estimable only when $r = p$. For the sake of completeness we allow $r < p$ except where noted.

Let $\mathbf{Z}$ denote a $n \times r$ matrix whose columns are an orthonormal basis for the column space of $\mathbf{X}$. In subsequent sections we will consider special cases where the columns of $\mathbf{Z}$ are standardized principal components, rotated principal components, or partial least squares components. Rather than projecting $\mathbf{Y}$ onto the column space of $\mathbf{X}$ as in ordinary least squares regression (OLSR), we wish to consider projecting $\mathbf{Y}$ onto a subspace of the column space of $\mathbf{X}$ spanned by a subset of the columns of $\mathbf{Z}$. Let $\mathcal{C}$ consist of k distinct integers chosen from $1, \ldots, r$ that represent the indices of selected columns of $\mathbf{Z}$. Let $\mathbf{C}$ denote the $r \times k$ matrix consisting of columns of the $r \times r$ identity matrix corresponding to $j \in \mathcal{C}$. An estimator of $\boldsymbol{\beta}$ based on the orthonormal basis $\mathbf{Z}$ and the chosen set $\mathcal{C}$ is given by $\hat{\boldsymbol{\beta}}_{\mathcal{C}} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{C}\mathbf{C}'\mathbf{Z}'\mathbf{Y}$, where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_r] \equiv \mathbf{X}'\mathbf{Z}$. To see this, note that $\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathcal{C}} = \mathbf{Z}\mathbf{W}'\hat{\boldsymbol{\beta}}_{\mathcal{C}} = \mathbf{Z}\mathbf{C}\mathbf{C}'\mathbf{Z}'\mathbf{Y} = (\mathbf{Z}\mathbf{C})[(\mathbf{Z}\mathbf{C})'(\mathbf{Z}\mathbf{C})]^{-1}(\mathbf{Z}\mathbf{C})'\mathbf{Y}$. Thus, when $\mathcal{C} = \{1, \ldots, r\}$, $\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ is the unique projection of $\mathbf{Y}$ on the column space of $\mathbf{Z}$ or, equivalently, the unique projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$. It follows that $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ is the unique and unbiased least squares estimator of $\boldsymbol{\beta}$ when $\mathcal{C} = \{1, \ldots, r\}$ and $r = p$. When $\mathcal{C} \subset \{1, \ldots, r\}$, $\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ is the projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbf{Z}$ indexed by $\mathcal{C}$.

General expressions for the expectation and variance of $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ are given by

$$E(\hat{\boldsymbol{\beta}}_{\mathcal{C}}) = \mathbf{A}\mathbf{C}\mathbf{C}'\mathbf{W}'\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathcal{C}}) = \sigma^2 \mathbf{A}\mathbf{C}\mathbf{C}'\mathbf{A}' = \sigma^2 \sum_{j \in \mathcal{C}} \mathbf{a}_j \mathbf{a}_j', \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_r] \equiv \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}$. When $\mathcal{C} \subset \{1, \ldots, r\}$, expression (1) shows that $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ is a potentially biased estimator of $\boldsymbol{\beta}$. It can be shown that $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ will be unbiased for $\boldsymbol{\beta}$ if and only if $\boldsymbol{\beta}$ is in the

column space of $\mathbf{W}$ and $\mathbf{w}'_j\boldsymbol{\beta} = 0$ for any $j \notin \mathcal{C}$. Expression (1) shows that the variance of any component of $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ will be no larger than the variance of the corresponding component of the least squares estimator. To balance the virtue of lower variance with the cost of higher bias, we seek the set $\mathcal{C}$ that will yield the estimator with the lowest MSE; i.e., we wish to find $\mathcal{C}$ so that

$$\text{MSE}(\mathcal{C}) \equiv E \parallel \hat{\boldsymbol{\beta}}_{\mathcal{C}} - \boldsymbol{\beta} \parallel^2 = E \parallel \hat{\boldsymbol{\beta}}_{\mathcal{C}} - E(\hat{\boldsymbol{\beta}}_{\mathcal{C}}) \parallel^2 + \parallel E(\hat{\boldsymbol{\beta}}_{\mathcal{C}}) - \boldsymbol{\beta} \parallel^2 \tag{2}$$

is minimized. Note that $\text{MSE}(\mathcal{C})$ is the trace of the MSE matrix $E\{(\hat{\boldsymbol{\beta}}_{\mathcal{C}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{\mathcal{C}} - \boldsymbol{\beta})'\}$. A variety of other performance measures based on the MSE matrix can be used to judge the quality of an estimator. We use the trace of the MSE matrix because it is mathematically and intuitively appealing to find the estimator that will minimize the expected squared Euclidean distance of the estimator from the estimand. In the motivating application discussed in Section 8, all the explanatory variables naturally have the same scale. When this is not the case, it may be necessary to divide each explanatory variable by its standard deviation to avoid a situation in which the trace of the MSE matrix is dominated by a small subset of its diagonal elements.

The first term on the right hand side of (2) is equal to

$$\text{Trace}\{\text{Var}(\hat{\boldsymbol{\beta}}_{\mathcal{C}})\} = \sigma^2 \text{Trace}(\mathbf{A}\mathbf{C}\mathbf{C}'\mathbf{A}') = \sigma^2 \text{Trace}(\mathbf{C}'\mathbf{A}'\mathbf{A}\mathbf{C}) = \sigma^2 \sum_{j \in \mathcal{C}} \mathbf{a}'_j \mathbf{a}_j. \tag{3}$$

Using the identity $\mathbf{W}\mathbf{A}'\mathbf{A} = \mathbf{A}$, the second term on the right hand side of (2) equals

$$\parallel \mathbf{A}\mathbf{C}\mathbf{C}'\mathbf{W}'\boldsymbol{\beta} - \boldsymbol{\beta} \parallel^2 = \boldsymbol{\beta}'\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{W}\mathbf{C}\mathbf{C}'\mathbf{A}'\mathbf{A}\mathbf{C}\mathbf{C}'\mathbf{W}'\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{W}\mathbf{A}'\mathbf{A}\mathbf{C}\mathbf{C}'\mathbf{W}'\boldsymbol{\beta}$$

$$= \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 (\sum_{j \in \mathcal{C}} \theta_j \mathbf{a}_j)'(\sum_{j \in \mathcal{C}} \theta_j \mathbf{a}_j) - 2\sigma^2 (\sum_{j=1}^{r} \theta_j \mathbf{a}_j)'(\sum_{j \in \mathcal{C}} \theta_j \mathbf{a}_j), \tag{4}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)' \equiv \mathbf{W}'\boldsymbol{\beta}/\sigma$. By (3) and (4), minimizing $\text{MSE}(\mathcal{C} \mid \boldsymbol{\beta}, \sigma^2)$ with respect to $\mathcal{C}$ is equivalent to minimizing

$$g(\mathcal{C} \mid \boldsymbol{\theta}) \equiv \sum_{j \in \mathcal{C}} \mathbf{a}'_j \mathbf{a}_j + (\sum_{j \in \mathcal{C}} \theta_j \mathbf{a}_j)'(\sum_{j \in \mathcal{C}} \theta_j \mathbf{a}_j) - 2(\sum_{j=1}^{r} \theta_j \mathbf{a}_j)'(\sum_{j \in \mathcal{C}} \theta_j \mathbf{a}_j) \tag{5}$$

with respect to $\mathcal{C}$. If we let $\mathcal{C}^*$ denote the subset of $\{1, \ldots, r\}$ that minimizes (5), the estimator $\hat{\boldsymbol{\beta}}_{\mathcal{C}^*}$ has the lowest total mean squared error among all estimators of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$, i.e., $\hat{\boldsymbol{\beta}}_{\mathcal{C}^*}$ has the lowest total mean

squared error among all estimators based on a projection $\mathbf{Y}$ onto a space spanned by a subset of the columns of $\mathbf{Z}$.

This result has somewhat limited usefulness since $\boldsymbol{\theta}$ is unknown in realistic regression problems, but it does suggest the following iterative algorithm that might yield an estimator of $\boldsymbol{\beta}$ with low MSE. Let $\hat{\boldsymbol{\theta}}^{(0)}$ denote an initial estimate of $\boldsymbol{\theta}$. (Usually $\hat{\boldsymbol{\theta}}^{(0)}$ will be of the form $\mathbf{W}'\hat{\boldsymbol{\beta}}^{(0)}/\hat{\sigma}^{(0)}$, where $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\sigma}^{(0)}$ are initial estimates of $\boldsymbol{\beta}$ and σ obtained through conventional means, but that form is not necessary.)

1. Given $\hat{\boldsymbol{\theta}}^{(t)}$, let $\mathcal{C}(t)$ denote the subset of $\{1, \ldots, r\}$ that minimizes $g(\mathcal{C} \mid \hat{\boldsymbol{\theta}}^{(t)})$.

2. Let $\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}_{\mathcal{C}(t)}$.

3. Let $\hat{\sigma}^{(t+1)} = \| \mathbf{Y} - \bar{Y}\mathbf{1} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(t+1)} \| / \sqrt{n - k - 1}$, where k denotes the cardinality of the set $\mathcal{C}(t)$.

4. Let $\hat{\boldsymbol{\theta}}^{(t+1)} = \mathbf{W}'\hat{\boldsymbol{\beta}}^{(t+1)}/\hat{\sigma}^{(t+1)}$.

5. Set t to $t + 1$ and return to step 1 until $\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(s)}$ for some $s \leq t$.

In general, step 1 of the algorithm requires the computation of $g(\mathcal{C} \mid \hat{\boldsymbol{\theta}}^{(t)})$ for $\mathcal{C}$ equaling all 2^r subsets of $\{1, \ldots, r\}$. We will show, however, that computation is greatly simplified for the special case of principal components regression. Although we have presented this algorithm as an iterative procedure, we have adopted the convention of stopping the algorithm after a single iteration. Our experience with the algorithm suggests that there is little gain beyond a single iteration. In the vast majority of the cases considered, the procedure converged immediately after the initial estimate was changed, i.e., $\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(2)}$. Thus, to simplify computing, we use $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\sigma}^{(1)}$ as the final estimates of $\boldsymbol{\beta}$ and σ for those methods that rely on the algorithm. We demonstrate the usefulness of this algorithm for several special cases in the simulations of Section 7. The next three sections examine the application of the work in this section to regression on principal components, regression on rotated principal components, and regression on partial least squares components, respectively.

3. APPLICATION TO PRINCIPAL COMPONENTS REGRESSION

Let $d_1 > d_2 > \cdots > d_r > 0$ where d_j^2 denotes the jth non-zero eigenvalue of $\mathbf{X}'\mathbf{X}$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ denote the corresponding unit-length eigenvectors of $\mathbf{X}'\mathbf{X}$. The sample principal components corresponding to the non-zero eigenvalues of $\mathbf{X}'\mathbf{X}$ are $\mathbf{X}\mathbf{v}_1, \ldots, \mathbf{X}\mathbf{v}_r$. In principal components regression (PCR), $\mathbf{Y}$ is regressed on a subset of the sample principal components. The estimated regression coefficients for the principal components in the chosen subset are used to obtain regression coefficients for the original columns of $\mathbf{X}$. For example, suppose $\mathbf{Y}$ is regressed against the first and third sample principal components to obtain

$$\hat{Y} = \hat{\delta}_1 \mathbf{X}\mathbf{v}_1 + \hat{\delta}_3 \mathbf{X}\mathbf{v}_3 = \mathbf{X}(\hat{\delta}_1 \mathbf{v}_1 + \hat{\delta}_3 \mathbf{v}_3),$$

where $\hat{\delta}_j$ denotes the ordinary least squares estimate of δ_j, the regression coefficient for the jth principal component. Then a principal components estimator of $\boldsymbol{\beta}$ is given by $\hat{\delta}_1 \mathbf{v}_1 + \hat{\delta}_3 \mathbf{v}_3$.

Principal components regression estimators, like $\hat{\delta}_1 \mathbf{v}_1 + \hat{\delta}_3 \mathbf{v}_3$, are of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ described in Section 2. The singular value decomposition of $\mathbf{X}$ implies $\mathbf{X} = \mathbf{S}\mathbf{D}\mathbf{V}'$, where

$$\mathbf{S} \equiv [\mathbf{X}\mathbf{v}_1/d_1, \ldots, \mathbf{X}\mathbf{v}_r/d_r], \ \mathbf{D} \equiv \mathrm{Diag}(d_1, \ldots, d_r), \ \text{and} \ \mathbf{V} \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_r].$$

Note that the orthonormal columns of $\mathbf{S}$ are the sample principal components of $\mathbf{X}$, scaled to unit length. We may equate $\mathbf{S}$ and $\mathbf{V}\mathbf{D}$ with $\mathbf{Z}$ and $\mathbf{W}$, respectively. Furthermore $\theta_j = d_j \mathbf{v}_j' \boldsymbol{\beta}/\sigma$. The estimator $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ simplifies to $\mathbf{V}\mathbf{D}^{-1}\mathbf{C}\mathbf{C}'\mathbf{S}'\mathbf{Y}$. In the simple example considered previously, $\mathbf{C}$ consists of the first and third columns of the $r \times r$ identity matrix, and we have

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathcal{C}} &= \mathbf{V}\mathbf{D}^{-1}\mathbf{C}\mathbf{C}'\mathbf{S}'\mathbf{Y} = \mathbf{V}\mathbf{D}^{-2}\mathbf{D}\mathbf{C}\mathbf{C}'\mathbf{C}\mathbf{C}'\mathbf{S}'\mathbf{Y} = \mathbf{V}\mathbf{C}\mathbf{C}'\mathbf{D}^{-2}\mathbf{C}\mathbf{C}'\mathbf{D}\mathbf{S}'\mathbf{Y} \\
&= \mathbf{V}\mathbf{C}(\mathbf{C}'\mathbf{D}^2\mathbf{C})^{-1}\mathbf{C}'\mathbf{D}\mathbf{S}'\mathbf{Y} = \mathbf{V}\mathbf{C}[(\mathbf{S}\mathbf{D}\mathbf{C})'(\mathbf{S}\mathbf{D}\mathbf{C})]^{-1}(\mathbf{S}\mathbf{D}\mathbf{C})'\mathbf{Y} \\
&= [\mathbf{v}_1, \mathbf{v}_3]([\mathbf{X}\mathbf{v}_1, \mathbf{X}\mathbf{v}_3]'[\mathbf{X}\mathbf{v}_1, \mathbf{X}\mathbf{v}_3])^{-1}[\mathbf{X}\mathbf{v}_1, \mathbf{X}\mathbf{v}_3]'\mathbf{Y} = \hat{\delta}_1 \mathbf{v}_1 + \hat{\delta}_3 \mathbf{v}_3.
\end{aligned}
$$

It can be shown that the set minimizing (5) is

$$\mathcal{C}^* = \{j : |\theta_j| = d_j |\mathbf{v}_j' \boldsymbol{\beta}|/\sigma > 1\} \tag{6}$$

7

for the special case of principal components regression.

The variance of $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ in principal components regression is $\sum_{j \in \mathcal{C}} d_j^{-2} \mathbf{v}_j \mathbf{v}_j'$, so excluding the principal components with the smallest eigenvalues from $\mathcal{C}$ can greatly reduce the variances of the components of $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$. The criterion for component selection suggested by (6) clearly discourages the use of principal components with small eigenvalues, but eigenvalue size is not the only consideration. Even when d_j is small, the jth component will be selected when $|\mathbf{v}_j' \boldsymbol{\beta}| = |\delta_j|$ is sufficiently large. This is similar in spirit to the standard inferential PCR approach where the goal is to retain the jth component if and only if $\delta_j \neq 0$. Note, however, that $\delta_j \neq 0$ is not the right criterion for minimizing MSE according to (6). A component for which $\delta_j \neq 0$ should be discarded unless δ_j is far enough from zero ($|\delta_j| > \sigma/d_j$) to counteract the variance inflating effect of a small eigenvalue d_j^2.

The minimum-MSE PCR estimator indicated by (6) depends on the unknown value of $\boldsymbol{\theta}$. The iterative algorithm described in Section 2 can be used to approximate the minimum-MSE PCR estimator. Step 1 simplifies to

1. Given $\hat{\boldsymbol{\theta}}^{(t)}$, let $\mathcal{C}(t) = \{j : |\hat{\theta}_j^{(t)}| > 1\}$.

The performance of the procedure for a variety of starting values is investigated in the simulations of Section 7.

Belinfante and Coxe (1986) advocate a component selection strategy that is equivalent to using our iterative algorithm with the initial estimate of $\boldsymbol{\theta}$ obtained from the OLSR estimates of $\boldsymbol{\beta}$ and σ^2. This procedure, however, does not account for the error in estimating $\boldsymbol{\theta}$. An alternative procedure can be based on tests of the hypotheses $H_{0j} : \theta_j^2 \leq 1$ for $j = 1, \ldots, r$. The jth component is used to estimate $\boldsymbol{\beta}$ if and only if the hypothesis H_{0j} is rejected at a prespecified level. If we assume that the error terms are normally distributed, these hypothesis tests can be carried out on $\hat{\theta}_j^2 \equiv (d_j \mathbf{v}_j' \hat{\boldsymbol{\beta}})^2 / \hat{\sigma}^2$, which is distributed as a noncentral F random variable when $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the ordinary least squares estimates of $\boldsymbol{\beta}$ and σ^2, respectively. A p-value for testing H_{0j} can be found by computing the probability that an F random variable

with 1 and $n - r - 1$ degrees of freedom and noncentrality parameter 1 exceeds the observed value of $\hat{\theta}_j^2$. This procedure yields an estimator that exhibits substantially lower MSE than all component-based procedures in the simulations of section. We use this procedure in Section 8 to investigate the value of PCR methods for estimating the locations and effects of genomic regions associated with quantitative traits. A similar procedure that results in the inclusion of more components deletes the jth component if and only if $H_{0j}' : \theta_j^2 \geq 1$ is rejected. This procedure appears to behave much like the OLSR estimator for the simulation settings we considered in Section 7.

4. APPLICATION TO REGRESSION ON ROTATED COMPONENTS

Although principal components have many nice properties, a given component is often difficult to interpret as a linear combination of the original explanatory variables. Many methods of rotating components to improve their interpretability have been proposed. The varimax rotation, due to Kaiser (1958, 1959), is the most well known of the orthogonal rotation methods. Chapter 8 of Jackson (1991) contains a brief description of varimax along with several other rotation methods and relevant references.

In this section, we consider the following problem: Given a set of orthogonally transformed components, find the subset of the transformed components that corresponds to an estimator of $\boldsymbol{\beta}$ with minimum MSE among all estimators that are based on the projection of $\mathbf{Y}$ onto a subspace spanned by a subset of the given transformed components. Such a problem is of interest to a researcher who wishes to both (i) study the relationship between the response variable and a particular set of interpretable components and (ii) accurately estimate the regression coefficients for the original variables using the set of interpretable components.

Recall that the sample principal components, scaled to unit length, are the columns of $\mathbf{S} = \mathbf{X}\mathbf{V}\mathbf{D}^{-1}$. Suppose $\mathbf{T}$ is an $r \times r$ orthogonal matrix such that the columns of $\mathbf{S}\mathbf{T}'$ are interpretable components. We make no attempt to give a formal definition of "interpretable component". Interested readers might see Thurstone's (1947) concept of simple structure summarized in Harmon (1976) and Jackson (1991). The basic idea is that the columns of $\mathbf{S}\mathbf{T}'$ should be dominated by relatively simple linear combinations of the

original explanatory variables rather than the potentially complex linear combinations found in the columns of $\mathbf{S}$. Because $\mathbf{X} = \mathbf{ST'TDV'}$ and $\mathbf{TS'ST'} = \mathbf{I}$, we may associate $\mathbf{ST'}$ and $\mathbf{VDT'}$, respectively, with $\mathbf{Z}$ and $\mathbf{W}$ of Section 2. The estimator $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ simplifies to $\mathbf{VD}^{-1}\mathbf{T'CC'TS'Y}$ in this case.

In practice, we must rely on the algorithm of Section 2 to provide a low-MSE estimator based on the projection of $\mathbf{Y}$ onto the space spanned by a set of orthogonally transformed components. The example in Section 7 and the Simulations of Section 8 examine the performance of such an estimator when the principal components are subjected to the varimax transformation.

5. APPLICATION TO PARTIAL LEAST SQUARES REGRESSION

Partial least squares regression (PLSR) is a method that has been developed and used primarily by chemometricians for predicting a response variable (or vector) from an often large number of multicollinear explanatory variables. Although the main focus of PLSR has been on prediction of the response variable, the method can be used to produce an estimate of $\boldsymbol{\beta}$ with low MSE. The origins of the technique can be traced to Wold (1966). Many authors have made important contributions to the development and understanding of PLSR. Papers by Stone and Brooks (1990) and Frank and Friedman (1993) (and the accompanying discussions) examine PLSR in a broad context and provide references to much of the relevant literature.

A PLSR estimate of $\boldsymbol{\beta}$ can be characterized as an estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ when $\mathbf{Z}$ is taken to be a particular orthonormal basis of the column space of $\mathbf{X}$. This orthonormal basis can be developed sequentially as follows. Let $\mathbf{q}_0 = \mathbf{0}$. For $j = 1, \ldots, r$; define

$$\mathbf{q}_j = \arg\max_{\mathbf{q} \in \mathcal{S}_j} \mathbf{Y'Xq}, \text{ where } \mathcal{S}_j \equiv \{\mathbf{q} : \mathbf{q'q} = 1, \mathbf{q}_k'\mathbf{X'Xq} = 0 \text{ for } k < j\}.$$

Then $\mathbf{Xq}_1, \ldots, \mathbf{Xq}_r$ are linear combinations of the columns of $\mathbf{X}$ constructed so that each linear combination has maximal sample covariance with $\mathbf{Y}$ among all linear combinations that are orthogonal to the previous linear combinations and have column weights whose squares sum to 1. For $j = 1, \ldots, r$; we define the jth partial least squares component as $\mathbf{z}_j = \mathbf{Xq}_j / \parallel \mathbf{Xq}_j \parallel$. These partial least squares components

10

form an orthonormal basis for the column space of $\mathbf{X}$. With $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_r]$ and $\mathbf{W} = \mathbf{X}'\mathbf{Z}$, the PLSR estimator of $\boldsymbol{\beta}$ is of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$, where $\mathcal{C}$ is a subset of $\{1, \ldots, r\}$ that contains $j - 1$ whenever it contains j.

The partial least squares basis for the column space of $\mathbf{X}$ is constructed as a function of both $\mathbf{Y}$ and $\mathbf{X}$. This is in contrast to the orthonormal bases discussed in Sections 3 and 4 whose construction depends only on $\mathbf{X}$. The theoretical development in Section 2 implicitly assumes the matrices $\mathbf{Z}$ and $\mathbf{W}$ are fixed. Thus the result of Section 2 is not directly applicable to the PLSR estimator of $\boldsymbol{\beta}$, even though the PLSR estimator is of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$. Nonetheless, the simulations in Section 7 suggest that the MSE of the PLSR estimator can be improved by using the iterative algorithm of Section 2 to select the number of partial least squares components utilized by the PLSR estimator. The computational requirements of step 1 of the algorithm are greatly reduced in PLSR. Rather than computing $g(\mathcal{C} \mid \hat{\boldsymbol{\theta}}^{(t)})$ for all 2^r subsets $\{1, \ldots, r\}$, $g(\mathcal{C} \mid \hat{\boldsymbol{\theta}}^{(t)})$ need be evaluated for only the $r + 1$ subsets of $\{1, \ldots, r\}$ satisfying $j - 1 \in \mathcal{C}$ whenever $j \in \mathcal{C}$. Estimators corresponding to other subsets of $\{1, \ldots, r\}$ are not PLSR estimators.

By the Cauchy-Schwarz inequality, $\mathbf{q}_1 = \mathbf{X}'\mathbf{Y} / \parallel \mathbf{X}'\mathbf{Y} \parallel$. Computation of subsequent partial least squares components is more complex. Helland (1988) shows the equivalence of two popular partial least squares algorithms and provides a third algorithm that can be used to compute a PLSR estimate of $\boldsymbol{\beta}$. Denham (1995) provides guidance on implementing these algorithms in FORTRAN, Matlab, and Splus. We have used a variation on the *orthogonal scores* code provided by Denham (1995) to determine $\mathbf{z}_1, \ldots, \mathbf{z}_r$ with Splus.

6. INVESTIGATED METHODS FOR ESTIMATING $\boldsymbol{\beta}$

Simulation studies that compare a variety of biased regression methods for estimating $\boldsymbol{\beta}$ are common in the statistics literature. Some examples include Dempster, Schatzoff, and Wermuth (1977), Gunst and Mason (1977), and Frank and Friedman (1993). All these papers consider OLSR, one or more variants of PCR, and one or more variants of ridge regression. The latter paper examines the performance of PLSR

11

as well. The next section of our paper describes a simulation study of methods for estimating β that are motivated by the development in Section 2. The primary goal of the simulation study is to obtain some measure of the usefulness of the results of Section 2 for selecting components (principal, rotated, or PLS) that yield estimators of β with low MSE. We will focus on the ability of the selection criteria described in Sections 3, 4, and 5 to improve upon a variety of initial estimates of β. The methods that we will consider in the simulations of Section 7 are described as follows.

1. PCR Estimators.

    (a) The OLSR estimator, i.e., $\hat{\beta}_{\mathcal{C}}$ for $\mathcal{C} = \{1, \ldots, r\}$.

    (b) The estimator of the form $\hat{\beta}_{\mathcal{C}}$, where $\mathcal{C}$ includes j if and only if δ_j is significantly different from 0 at the 0.05 level (i.e., the standard inferential approach).

    (c) The estimator of the form $\hat{\beta}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with the OLSR estimator as a starting value. This estimator is recommended by Belifante and Coxe (1986).

    (d) The estimator of the form $\hat{\beta}_{\mathcal{C}}$, where $\mathcal{C}$ is chosen through leave-one-out cross validation from the $r + 1$ subsets of $\{1, \ldots, r\}$ satisfying $j - 1 \in \mathcal{C}$ whenever $j \in \mathcal{C}$.

    (e) The estimator of the form $\hat{\beta}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with starting value provided by method 1(d).

    (f) The estimator of the form $\hat{\beta}_{\mathcal{C}}$, where $\mathcal{C}$ is chosen through leave-one-out cross validation from the $r + 1$ subsets of $\{1, \ldots, r\}$ satisfying $j' \in \mathcal{C}$ whenever $j \in \mathcal{C}$ and the sample correlation between $\mathbf{X}\mathbf{v}_{j'}$ and $\mathbf{Y}$ is greater than the sample correlation between $\mathbf{X}\mathbf{v}_j$ and $\mathbf{Y}$.

    (g) The estimator of the form $\hat{\beta}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with starting value provided by method 1(f).

(h) The estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ that includes the jth component in $\mathcal{C}$ if and only if $H_{0j} : \theta_j^2 \leq 1$ is rejected at significance level 0.05. (This estimator performs best in the simulation of Section 7.)

(i) The estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ that excludes the jth component from $\mathcal{C}$ if and only if $H'_{0j} : \theta_j^2 \geq 1$ is rejected at significance level 0.05.

(j) The pseudo-estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with true value of $\boldsymbol{\theta}$ as the starting value. (We use the term pseudo-estimator to emphasize that the computed quantity is a function of the unknown parameter $\boldsymbol{\beta}$.)

2. Estimators Based on the Varimax Rotation of the Principal Components.

(a) The OLSR estimator, i.e., $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ for $\mathcal{C} = \{1, \ldots, r\}$.

(b) The estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with the OLSR estimator as a starting value.

(c) The pseudo-estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with true value of $\boldsymbol{\theta}$ as the starting value. (We use the term pseudo-estimator to emphasize that the computed quantity is a function of the unknown parameter $\boldsymbol{\beta}$.)

3. PLSR Estimators.

(a) The OLSR estimator, i.e., $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ for $\mathcal{C} = \{1, \ldots, r\}$.

(b) The estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with the OLSR estimator as a starting value.

(c) The estimator of the form $\hat{\boldsymbol{\beta}}_{\mathcal{C}}$, where $\mathcal{C}$ is chosen through leave-one-out cross validation from the $r + 1$ subsets of $\{1, \ldots, r\}$ satisfying $j - 1 \in \mathcal{C}$ whenever $j \in \mathcal{C}$.

(d) The estimator of the form $\hat{\beta}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with starting value provided by method 3(c).

(e) The pseudo-estimator of the form $\hat{\beta}_{\mathcal{C}}$ obtained through the algorithm described in Section 2 with true value of θ as the starting value. (We use the term pseudo-estimator to emphasize that the computed quantity is a function of the unknown parameter β.)

The pseudo-estimators produced by methods 1(j) and 2(c) are optimal in the sense of minimizing MSE as described in Section 2. Methods 1(j), 2(c), and 3(e) cannot be used in practice because the true value of θ will be unknown. These pseudo-estimators have been included in the simulation study to gauge the impact of using imperfect starting values in the algorithm.

7. A SIMULATION STUDY

A simulation study was conducted to investigate the effectiveness of the proposed estimators at reducing MSE. Several useful methods, including all ridge regression procedures, are not considered in the simulation study. We focus only on the estimators of the form $\hat{\beta}_{\mathcal{C}}$ that are described in Section 6. The main goal of the study is to determine if the component selection methods suggested by Sections 2 through 5 can be used to reduce MSE. We also examine how well methods perform when true values of β and σ are used to select components by including methods 1(j), 2(c), and 3(e) in the study, even though these are not estimators.

The simulation study was conducted as a completely randomized design with three factors: the matrix of explanatory variables, the true value of β, and the signal to noise ratio. To simulate near multicollinearity in the matrix of explanatory variables, the rows of each of the four $\mathbf{X}$ matrices considered in the study were independently drawn from the $N_5(\mathbf{0}, \Sigma)$ distribution, where the elements of Σ are 1 on the diagonal and 0.9 on the off diagonal. The true values of β used in the study are $\beta_1 \equiv (1, 0, 0, 0, 0)'$, $\beta_2 \equiv (1, 1, 1, 1, 1)'$, $\beta_3 \equiv (1, 2, 3, 4, 5)'$, and $\beta_4 \equiv (1, 4, 9, 16, 25)'$. Signal to noise ratio, defined by $S/N = \sqrt{\beta'\Sigma\beta}/\sigma$, was set at 0.5 and 2.0. One hundred $\mathbf{Y}$ vectors were generated for each of the 32 simulation settings. The squared

14

distance from the true value of $\boldsymbol{\beta}$ to each of the estimators described in Section 6 was computed for all 3,200 data sets. The means of the 100 squared distances (empirical MSE estimates) were computed for each of the estimators described in Section 6 and each of the 32 simulation settings. For each of the 32 simulation settings, the OLSR estimator along with the other estimators discussed in Section 6 were ranked according to their empirical MSE values with lowest ranks corresponding to lowest empirical MSE. The median of the 32 ranks for each method is provided in Table 4.

The OLSR estimator performed quite poorly in this simulation study. This comes as no surprise because the simulation was designed to study the behavior of the estimators when OLSR estimation is expected to be deficient. The OLSR estimator was strongest relative to the other estimators when signal to noise ratio was high and the true value of $\boldsymbol{\beta}$ was $(1, 0, 0, 0, 0)'$. The median rank for the OLSR estimator was worst (15.5) among the 16 procedures ranked. The OLSR estimator ranked last for half of the 32 simulation settings. Method 1(i) offered only a slight improvement over OLSR with a median rank of 15.

Method 1(c), the PCR estimator that uses algorithm of Section 2 with the OLSR estimate as a starting value, exhibited lower MSE than the OLSR estimator for 31 of 32 simulation settings. The one exception occurred when signal to noise ratio was 2.0 and $\boldsymbol{\beta}$ was $(1, 0, 0, 0, 0)'$. Method 2(b), the varimax estimator that uses the OLSR estimate as a starting value for the algorithm of Section 2, improved upon the OLSR estimator for 20 of the 32 simulation settings. The analogous PLSR estimator 3(b) had lower empirical MSE than the OLSR estimator for 30 of the 32 simulation settings. Among estimators that use the OLSR estimate as a starting value, the PLSR estimator 3(b) had the best median rank of 12. The PCR estimator 1(c) and the varimax estimator 2(b) exhibited similar performance with median ranks of 13 and 14, respectively.

Although the algorithm of Section 2 does tend to reduce the empirical MSE of the OLSR estimator for the majority of the situations we examined, the improvement at any particular simulation setting is seldom dramatic. The same can be said for the use of the algorithm with other starting values. It is interesting, however, to note that the algorithm seems to reduce the MSE of the estimator supplying the starting value,

15

regardless of which estimator is used to provide the starting value. For example, method 1(e) had lower empirical MSE than 1(d) for 29 of the 32 simulation settings. Method 1(g) had lower empirical MSE than 1(f) for 26 of the 32 settings. Method 3(d) outperformed 3(c) for 30 of the 32 settings. The value of the algorithm of Section 2 can also be seen by comparing the median ranks for the estimators that use the algorithm to the median ranks of the estimators used as starting values for the algorithm. Each estimator that uses the algorithm of Section 2 earned a better median rank than the estimator used to produce its starting value.

The PLSR estimator 3(d) with the cross-validated PLSR estimate as a starting value had the best median rank among the estimators that utilize the algorithm of Section 2. The two PCR estimators that used cross validation to produce starting values, 1(e) and 1(g), had the next best median ranks, with 1(e) performing slightly better than 1(g). This group of estimators – 3(d), 1(e), and 1(g) – performed substantially better than the estimators that used the OLSR estimate as a starting value for many of the simulation settings.

Methods 1(b) and 1(h) exhibited the most dramatic drops in MSE among the procedures that can be used in practice. Method 1(h), in particular, was impressive with empirical MSE values sometimes less than a third the empirical MSE of method 1(b) and less than a tenth the MSE of the OLSR estimator. For 28 of the 32 simulation settings, method 1(h) had the lowest empirical MSE among methods that did not use the true parameter values to select components. Method 1(h) was occasionally the worst among all estimators, but this occurred in situations when all estimators had relatively low MSE. Overall method 1(h) looks to be the best method among those considered for the analysis of the simulation data. Judging by the performance of the estimators that use the algorithm of Section 2, an estimator that uses the algorithm with an estimate from method 1(h) as the starting value might have slightly lower MSE than method 1(h).

Methods 1(j), 2(c), and 3(e) used the algorithm of Section 2 with the true value of θ as a starting value. The empirical MSE values were always extremely low for methods 1(j) and 3(e). These methods shared the best median rank of 2.0. The empirical MSE of method 2(c) was generally higher than the empirical

16

MSE of either method 1(j) or 3(e). For the majority of the simulation settings the varimax estimator 2(c) estimated $\boldsymbol{\beta}$ as $\mathbf{0}$ for all 100 randomly generated data sets. This often led to values of empirical MSE that were lower than empirical MSE values for the OLSR estimator and the other methods that used the data to select components.

Figure 1 illustrates the estimated root mean squared error (RMSE) for the four PCR methods with the lowest median ranks. The results have been averaged over the four randomly generated $\mathbf{X}$ matrices and reported as a proportion of the estimated RMSE for the OLSR estimator to simplify presentation. These results clearly show that none of the estimators that use data to select components approach the performance of the minimum-MSE PCR method 1(j). Method 1(h) comes the closest to matching the ideal, but even this method has estimated RMSE values that are sometimes more than double the values for the optimal method. Detailed examination of the simulation results indicates that type I errors in the test of $H_{0j} : \theta_j^2 \leq 1$ can sometimes result in individual 1(h) estimates that are far from the true $\boldsymbol{\beta}$. It is likely that performance of the estimator 1(h) could be improved by adjusting the significance levels at which the tests of H_{0j} are conducted, with perhaps smaller significance levels assigned to tests corresponding to smaller eigenvalues. We leave this as an issue for further study.

We ran a second simulation study with the number of explanatory variables p increased from 5 to 20. To reduce computational demands, only the top PCR estimators from the $p = 5$ simulation (1(b) and 1(h)) along with the OLSR estimator 1(a) and the minimum-MSE method 1(j) were considered. For all 32 simulation settings the OLSR estimator 1(a) had the highest estimated RMSE followed by 1(b), 1(h), and 1(j) with average estimated RMSEs at 44%, 28%, and 5% of OLSR, respectively.

8. APPLICATION TO MAPPING QUANTITATIVE TRAIT LOCI

Quantitative trait loci (QTL) are genomic regions that affect the quantitative characteristics of plants and animals. The main goals in QTL mapping are to locate QTL on the genetic map of an organism and to estimate the effect of each QTL on the quantitative trait of interest. The review article by Doerge, Zeng, and

17

Weir (1997) describes many statistical methods for locating QTL. Kao, Zeng, and Teasdale (1999) provide references to more recent work.

In QTL mapping, $\mathbf{Y}$ is a vector of quantitative trait values (e.g., blood pressures, lean percentages, or yields) with one entry for each of n units (e.g., mice, hogs, or plants). The explanatory data $\mathbf{X}$ consist of the genotypes of the units at each of p molecular markers. In general the explanatory data is categorical and, in many cases, binary. For simplicity we will assume that each marker is one of two genotypes so that the entries in the matrix $\mathbf{X}$ can be coded as 0 or 1. (Such data are derived from crosses between inbred lines in backcross, recombinant inbred line, recombinant inbred chromosome line, and doubled haploid designs among others.) We can order the columns of $\mathbf{X}$ to match the physical order of the markers on the genome. As the distance between markers on the same chromosome decreases, the correlation between the columns of $\mathbf{X}$ associated with the markers increases. It is common to see nearby markers agreeing for all but a few of the units, resulting in very high correlations between adjacent columns of $\mathbf{X}$.

Whittaker, Thompson, and Visscher (1996) show how coefficient estimates from a regression of quantitative trait on marker genotype can be used to estimate the locations and effects of multiple QTL when there is sufficient marker coverage over relevant portions of the genome. Regression based estimates have been shown to be remarkably similar to estimates obtained through more computationally intensive methods for QTL mapping (Haley and Knott, 1992). Because of the high correlation among markers on any given chromosome, the variance of the least-squares estimator used by Whittaker, et al. (1996) is likely to be inflated. In this section we describe how principal components regression methods can be used to obtain estimates of QTL location and effect with lower MSE than the commonly used OLSR estimators.

Whittaker, Thompson, and Visscher (1996) derive their results for a situation in markers and QTL can have one of three genotypes. Piepho and Gauch (2001) give analogous results for a two-genotype case. We briefly summarize the key points here.

(i) $E(\mathbf{Y} \mid \mathbf{X}) = \alpha + \mathbf{X}\boldsymbol{\beta}$, where $\beta_j = 0$ for all j such that there is at least one marker between marker $\mathbf{X}_j$

and all QTL on its chromosome.

(ii) When there are at least two markers between any pair of QTL on a single chromosome, the position and effect of any particular QTL can be determined from $\boldsymbol{\beta}$ and a genetic map giving the locations of the markers. The relevant equations are

$$\rho_l = 0.5 \left(1 - \sqrt{\frac{(1-2\rho)\{\beta_l + \beta_r(1-2\rho)\}}{\beta_r + \beta_l(1-2\rho)}} \right) \qquad \gamma = \frac{\rho(1-\rho)(1-2\rho)\beta_l}{(\rho-\rho_l)(1-\rho-\rho_l)}. \qquad (7)$$

In (ii) β_l and β_r denote the regression coefficients of the markers immediately flanking the QTL on the left and right, respectively. The parameter γ denotes the difference between mean trait values for units with QTL genotype 1 and units with QTL genotype 0 (i.e., γ is the effect of the QTL). The parameter ρ denotes the recombination fraction (i.e., a genetic distance) between the markers immediately flanking the QTL. The parameter ρ_l denotes the recombination fraction between the left flanking marker and its QTL (i.e., the location of the QTL). A distance between markers ρ is typically assumed to be known, even though in practice it is estimated from the marker data. The consequence of (i) and (ii) is that markers with non-zero regression coefficients are exactly those markers flanking QTL, and the values of the nonzero regression coefficients determine the effects of QTL and their locations relative to the marker positions.

To illustrate the use of principal components regression in the context of QTL mapping, we applied PCR procedure 1(h) to barley data available on the Web at http://wheat.pw.usda.gov/ggpages/SxM/. Hayes et al. (1993) used 1991 barley data averaged over five environments to find evidence for QTL affecting various agronomic traits. For α-amylase (a quantitative characteristic related to malting quality), Hayes et al. (1993) mapped QTL to nine marker intervals on barley's seven chromosomes. We used the 1991 data available on the Web to obtain estimates of the effects and locations of the QTL within the nine intervals. In this example, $n = 150$ (150 doubled-haploid barley lines were planted in each field) and $p = 18$ (two markers for each of nine QTL).

Nearly 4% of the $np = 2700$ marker genotype values are missing. Fortunately, the high correlation

among neighboring markers allows for reliable random imputation of missing genotypes. We computed the OLSR and PCR estimator for each of 1000 $\mathbf{X}$ matrices with missing values filled in by random imputation. The OLSR and PCR estimates of the regression coefficients, averaged over the 1000 imputations, are provided in Table 5. The PCR estimator was most often based on the first five principal components. There was some variation, however, due to differences among imputed $\mathbf{X}$ matrices. For example, component 18 was included with the first five components in one case. We used equations (7) to convert the regression coefficients to estimates of QTL locations and effects. The results are displayed in Table 6.

The signs of the estimated regression coefficients obtained through OLSR disagree for marker intervals 1, 2, and 9. The OLSR estimates for these intervals have not been converted to estimates of ρ and γ because equations (7) do not necessarily give meaningful results when regression coefficients differ in sign. Whittaker et al. (1996) show that the regression coefficients of markers flanking QTL must agree in sign. Both Whittaker et al. (1996) and Peipho and Gauch (2001) assume that markers intervals whose estimated regression coefficients disagree in sign are devoid of QTL. Hwang and Nettleton (2002), however, show that high correlation among adjacent markers can cause the *estimated* regression coefficients to disagree in sign with high probability even when the true regression coefficients agree in sign. Thus it is not surprising that we see some sign disagreements for the OLSR estimator in this example. The regression coefficients estimated with the PCR method agree in sign for all nine intervals. Its good performance in this one example is not proof of better performance in general, but simulations not reported here indicated that the PCR estimator has far fewer problems with sign disagreement.

Hayes et al. (1993) used a method developed by Haley and Knott (1992) to estimate the effects of the QTL. The effect estimates computed using the PCR approach agree reasonably well with the effect estimates reported by Hayes et al. The correlation between the two sets of estimated effects is 0.816, and both sets of estimates suggest that the Morex genotype is associated with higher mean α-amylase values for each QTL. The PCR estimates tended to be closer to zero than the original estimates reported by Hayes et al. When the

signs of the estimated regression coefficients agreed, the OLSR-based effect estimates were also similar to those reported by Hayes et al. The negative OLSR-based effect estimate for interval 7, however, suggests that the Morex genotype is associated with lower mean α-amylase at the 7th QTL. This conflicts with both the original estimate and the PCR-based estimate.

In the presentation of this example, we have focused specifically on the problem of estimating the locations and effects of QTL, given the marker intervals in which the QTL are contained. In practice, a subset of the markers believed to flank QTL must be selected because the number of QTL and their marker intervals are unknown. Marker selection is an important aspect of the QTL mapping problem that is essentially a special case of variable selection in multiple regression. Piepho and Gauch (2001) describe a stepwise procedure for choosing flanking markers and then estimating locations and effects of QTL through OLSR. Our work suggests that it may be better to replace OLSR with PCR at the estimation stage of the procedure. Of course any estimates are subject to selection bias if the same data used to select flanking markers are also used to produce estimates. Problems with selection bias can be managed by using cross validation techniques. Utz et al. (2001) discuss cross validation in the context of QTL mapping. The PCR regression estimator 1(h) could play an important role in the estimation stage of a cross validation procedure that partitions the data into independent sets for marker selection and effect/location estimation.

9. DISCUSSION

Several papers have been written on the topic of selecting components for principal components regression. The texts by Jackson (1991) and Jollife (1986) provide discussions of selection procedures and a road map to the relevant literature. We have considered a more general component-selection problem that encompasses the selection of principal components, rotated principal components, and partial least squares components. Our goal has been to develop methods for component selection that lead to regression coefficient estimators with low MSE. Our theoretical development in Section 2 and the performance of the pseudo estimators 1(j), 2(c), and 3(e) in the simulations of Section 7 suggest that there exists a subset of components

21

that will yield an estimator with very low MSE in most problems. Using the data to select the best set is challenging.

The iterative algorithm described in Section 2 appears to reduce the MSE of a variety of estimators used as starting values for the algorithm in the simulation study of Section 7. The improvements, however, are minimal for the most part, and the empirical MSE of the estimators was typically orders of magnitude higher than the MSE of the idealized pseudo-estimators. Although our attempts to approximate the optimal subset of components fall short of the ideal, our methods do provide a substantial reduction in the MSE of the regression coefficient estimator in many cases.

In particular, a new principal components regression estimator has emerged from this work that appears to maintain relatively low MSE when there is a high degree of correlation among explanatory variables. This estimator – denoted 1(h) in Sections 6, 7, and 8 – selects only those components exhibiting a special form of significant correlation with the response vector. In this sense, the procedure is a specific implementation of one of Massy's (1965) suggestions – delete the components that are relatively unimportant as predictors of the dependent variable.

The procedure is similar in spirit to the common practice of using only those principal components whose regression coefficient estimates are statistically significant (method 1(b) in Sections 6 and 7). The new method 1(h) outperformed the more common method 1(b) for all but 2 of the 32 simulation settings and, more often than not, had an estimated MSE of less than half that of the more common method. Method 1(h) also performed better than the PLSR estimators. The series of tests used to select the principal components in method 1(h) are derived under the assumption that the error distribution is normal. More work is required to determine how sensitive the procedure is to the normality assumption.

REFERENCES

Belinfante, A. and Coxe, K. L. (1986). Principal components regression – selection rules and application. *Proc. Bus. & Econ. Sec., Amer. Stat. Assoc.*, 429-431.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Assoc.* **72**, 77-106.

Denham, M. C. (1995). Implementing partial least squares. *Statistics and Computing*. **5**, 191-202.

Doerge, R.W., Zeng, Z-B., and Weir, B.S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **12**, 195-219.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Techno-metrics*. **35**, 109-148.

Gunst, R. F. and Mason, R. L. (1977). Biased estimation in regression: an evaluation using mean squared error. *J. Amer. Statist. Assoc.* **72**, 616-628.

Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*. **52**, 15-19.

Haley, C.S. and Knott, S.A. (1992). A simple method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315-324.

Harmon, H. H. (1976). *Modern Factor Analysis*, 3rd Ed. University of Chicago Press, Chicago.

Hayes, P. M., Liu, B. H., Knapp, S. J., Chen, F., Jones, B., Blake, T., Franckowiak, J., Rasmusson, D., Sorrells, M., Ullrich, S. E., Wesenberg, D., Kleinhofs, A. (1993). Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *Theor. Appl. Genet.* **87**, 392-401.

Helland, I. S. (1988). On the structure of partial least squares regression. *Commun. Statist. – Simula.* **17**, 581-607.

Hwang, J. T. G. and Nettleton, D. (2002). Investigating the probability of sign inconsistency in the regression coefficients of markers flanking QTL. *Genetics*. **160**, 1697-1705.

Jackson, J. E. (1991). *A User's Guide to Principal Components*. Wiley & Sons, New York.

Jollife, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*. **31**, 300-303.

Jollife, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*. **23**, 187-200.

Kaiser, H. F. (1959). Computer program for varimax rotation in factor analysis. *Educ. & Psych. Meas.* **19**, 413-420.

Kao, C-H., Zeng, Z-B., and Teasdale, R.D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216.

Mason, R. L. and Gunst, R. F. (1985). Selecting principal components in regression. *Stat. & Prob. Lett*. **3**, 299-301.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Stat. Assoc.* **60**, 234-256.

Piepho, H. P. and Gauch, H. G. (2001). Marker pair selection for mapping quantitative trait loci. *Genetics*. **157**, 433-444.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. Roy. Statist. Soc. ser. B*. **52**, 237-269.

Thurstone, L. L. (1947). *Multiple Factor Analysis*, University of Chicago Press, Chicago.

Utz, H. F., Melchinger, A. E., and Schön, C. C. (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics*. **154**, 1839-1849.

Whittaker, J.C., Thompson, R., and Visscher, P.M. (1996). On the mapping of QTL by the regression of phenotype on marker-type. *Heredity*. **77**, 23-32.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*. Ed. P.R. Krishnaiah. New York: Academic Press, 391-420.

Table 4. Median of the Ranks of the Estimators

Method	1-3(a)	1(b)	1(c)	1(d)	1(e)	1(f)	1(g)	1(h)	1(i)	1(j)
Median Rank	15.5	5.5	13	9.5	8	10	9.5	4	15	2

Method	2(b)	2(c)	3(b)	3(c)	3(d)	3(e)
Median Rank	14	3	12	7	5	2

*For each of the 32 simulation settings, the 16 methods were ranked. Methods with the lowest estimated MSE received the smallest ranks. The median of the 32 ranks assigned to each method is reported. Other than the pseudo estimators 1(j), 2(c), and 3(e) that depend on unknown parameters, 1(h) ranks the best.

Table 5. Estimated Regression Coefficients for the QTL Mapping Example

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
OLSR	2.06	-1.20	3.43	-0.61	1.11	1.02	0.85	0.10	1.02	1.31
PCR 1(h)	0.81	1.01	1.03	0.94	0.74	0.81	0.73	0.66	1.07	1.16

Method	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}
OLSR	0.77	0.68	-0.24	-0.17	1.39	0.13	-0.12	0.90
PCR 1(h)	0.63	0.57	0.27	0.31	0.47	0.50	0.17	0.21

Table 6. Estimated QTL Locations and Effects

Method	Parameter	Marker Interval								
		1	2	3	4	5	6	7	8	9
OLSR	ρ	-	-	0.042	0.012	0.085	0.040	0.070	0.004	-
PCR 1(h)	ρ	0.075	0.033	0.046	0.053	0.079	0.040	0.088	0.025	0.058
OLSR	γ	-	-	2.15	0.96	2.35	1.46	-0.42	1.52	-
PCR 1(h)	γ	1.84	1.98	1.56	1.40	2.26	1.21	0.59	0.96	0.38

Figure 1 : Estimated root mean squared error (RMSE) as a proportion of OLSR estimated

RMSE for the best PCR methods.