November 1996
Volume 38, Number 4

Technometrics

A Journal of Statistics for the Physical, Chemical, and Engineering Sciences

ASQC and American Statistical Association

Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression

Brian D. MARX

Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803-5606

I extend the concept of partial least squares (PLS) into the framework of generalized linear models. A spectroscopy example in a logistic regression framework illustrates the developments. These models form a sequence of rank 1 approximations useful for predicting the response variable when the explanatory information is severely ill-conditioned. Iteratively reweighted PLS algorithms are presented with various theoretical properties. Connections to principal-component and maximum likelihood estimation are made, as well as suggestions for rules to choose the proper rank of the final model.

KEY WORDS: Biased estimation; Cross-validation; Ill-conditioned information; Latent variables; Principal components.

1. INTRODUCTION

I focus on modeling an exponential family response through generalized linear regression (GLR) with p (standardized) explanatory variables in the matrix $X_{N\times p}$, with rows x_i^T such that

$$g(\mu_i) = \beta_0 + x_i^T\beta = \eta_i, \tag{1}$$

where $i = 1, \ldots, N$. Details and notation of the generalized linear model follow in Section 2. Despite the popularity of maximum likelihood (ML) parameter estimation in (1), the effects of ill-conditioning in training data can be nontrivial, and approaches are needed for reducing the effects of these dependencies. Perhaps more important in the area of chemometrics is that often the number of explanatory variables far exceeds the number of observations; thus we start with an ill-posed estimation problem. Frank and Friedman (1993) provided an excellent overview of regression tools for (approximately) Normal response data, useful for the chemometric community. One of these methods is partial least squares (PLS), due to Wold (1975). PLS was initially developed for social-science problems having scarce information, but more recently it has received a great amount of attention in the chemometrics literature. Wold, Ruhe, Wold, and Dunn (1984) provided an alternate construction of PLS. Later, Helland (1988) provided a nice overview and summary of PLS.

Response data may be discrete, however—for example, presence/absence or counts of a phenomenon. For instance, a researcher may need to predict a discrete or non-Normal physical or chemical composition of a substance in which the explanatory variables consist of several (hundred) signals of (collinear) wavelengths from spectroscopy. Furthermore, one can imagine several other potential datasets using spectroscopy to model, say, the presence/absence of certain animal parts in ground meat, presence/absence of preservative in food products, or perhaps to construct a model used to predict counts of a certain discrete constituent within a product. I propose an iteratively reweighted partial least squares (IRPLS) estimation technique for GLR. I will demonstrate that the standard PLS algorithm can be made to work in a wider GLR sense. IRPLS forms a sequence of rank 1 approximations useful to predict gamma, binomial, Poisson, or other response variables in the exponential family. Many biased estimation techniques for GLR's have surfaced in the last decade, but nearly all of these efforts addressed alternatives to ML estimation when the information matrix is *near* singular. Often the motivation was to combat (weighted) collinearity in the more complex GLR setting. Apart from producing alternative models through variable subset selection (VSS), research efforts in biased estimation include the lasso (Tibshirani 1996), ridge (Le Cessie and van Houwelingen 1992; Marx, Eilers, and Smith 1992), and (iteratively reweighted) principal component (Marx and Smith 1990), among other penalized likelihood approaches. Both the PLS and the IRPLS algorithms are variants of the conjugate gradient method of finding generalized inverses (Hestenes and Stiefel 1952) that construct noninterfering directions to solve a maximum (minimum) of a multidimensional function.

I revisit a dataset that is an application of near-infrared reflectance (NIR) spectroscopy used to predict the probability that a freesia's bud will flower. There were 100 sources yielding $N = 100$ branch bunches. Within a branch bunch, several branches (in most cases 20) were collected, then split into two groups: The first group (usually consisting of 14 branches) was put into a vase and monitored for successful budding, and the other group (usually 6) branches was used for NIR spectroscopy. For a given source, the total combined number of buds (m_i) in the vase were counted, $i = 1, \ldots, 100$. For a given bunch, the researcher is in-

TECHNOMETRICS, NOVEMBER 1996, VOL. 38, NO. 4

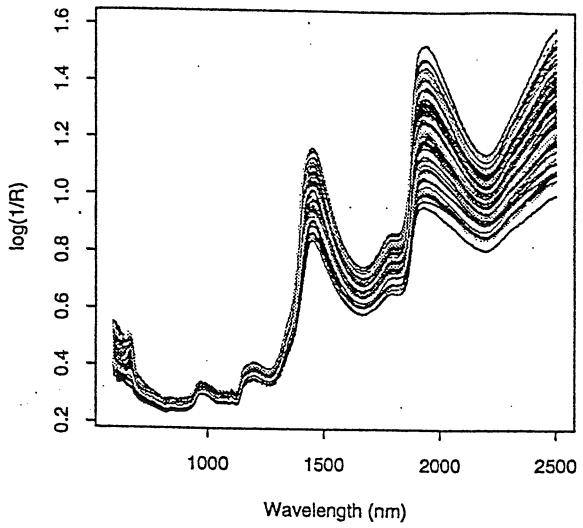

Figure 1. Spectra Readings Over the N = 100 Freesia Sources.

terested in how the spectrum is related to the number of buds that produce flowers in the vase, $y_i \in \{0, \ldots, m_i\}$. As mentioned, the total number of buds (m_i) may vary from vase to vase, and I assume $y_i \sim$ binomial(m_i, p_i), with p_i unknown. Each bunch has an NIR spectra consisting of 476 $\log(R^{-1})$ readings at wavelengths ranging from 600 nanometers (nm) to 2,500 nm in equal steps of 4 nm. Corresponding to these NIR spectra, there also exists additional lab information. Figure 1 displays the spectra $\log(R^{-1})$ readings over the 476 wavelengths, or explanatory variables, ranging from 600 nm to 2,500 nm from the $N = 100$ freesia sources. As mentioned, we wish to use this spectra information as explanatory information to predict the binary response of *successful flowering* (1) or *unsuccessful flowering* (0) *bud*. The model of interest is

$$\log \frac{p_i}{1-p_i} = \beta_0 + \sum_{j=1}^{476} x_{ij}\beta_j, \tag{2}$$

where x_{ij} is the $\log(R^{-1})$ for the jth wavelength of the ith observation, $i = 1, \ldots, 100$ and $j = 1, \ldots, 476$. The parameter p_i is the probability of success (1) of the ith observation. The problem is terribly ill posed with $N = 100$ and $p = 476$. The first three eigenvalues of the estimated information matrix are 2.652e11, 7.8489e4, and 2.0419e3, and together they account for at least 99.99999% of the total variance. Prediction equations and parameter estimation using IRPLS and PC techniques follow in Section 6. First I provide the details.

2. BACKGROUND AND NOTATION FOR GENERALIZED LINEAR REGRESSION

Using a variety of interesting examples, Dobson (1990) provided an extremely clear introductory presentation of how many statistical methods involving a linear predictor can be united through generalized linear models. I start with a (spectra) matrix of explanatory variables $X^* = (x_1^*, x_2^*, \ldots, x_p^*)$ of dimension $N \times p$. Consistent with chemometrics analyses, we center and scale (autoscale) the columns of X^*, but with a weighted mean and weighted sum of squares, respectively. Throughout the remainder of this article, I will work with the autoscaled explanatory variables, denoted by X of dimension $N \times p$, which does not contain a column vector of ones. The weights are defined as V, and the use of such weights in the standardization will become more apparent as the details and motivation for the algorithm are explicitly given in Section 3. Again, I denote x_i^T a $1 \times p$ row vector of X. The matrix $(1, X)$ of dimension $N \times (p+1)$ will be used when an intercept is included into the regression model. One attractive and convenient consequence of the preceding standardization scheme is that the estimation of the regression intercept coefficient is uncorrelated with the other explanatory variable (spectra) estimated coefficients. In fact, we will see that the IRPLS algorithm simply estimates the intercept coefficient with the weighted mean of the (iterated and adjusted) dependent variable. If one wishes, such an intercept term can be made more transparent in the algorithm by internally autoscaling the (adjusted) dependent variable at each iteration. Due to the unintuitive nature of the (adjusted) dependent variable and for sake of comparison to existing GLR principal-component (PC) algorithms, I choose to present a standard GLR model, only autoscaling the X explanatory variables. It is well known that there exist simple linear transformations connecting various equivalent representations, and in all cases the resulting solutions can further be transformed back into the variables' natural metric. Assume that the random response vector, $Y_{N\times 1}$, has independent entries Y_i following a distribution in the exponential family and is expressed as $f(y;\theta,\phi) = \exp[\{yb(\theta) + c(\theta)\} + d(y)]$, where b, c, d are known functions. This is referred to the canonical form of the exponential family. Furthermore, $b(\theta)$ is referred to the natural parameter of the distribution. In many applications $b(\theta) = \theta$; then $E(Y) = \mu = c'(\theta)$, providing the crucial connection between θ and μ. Any nuisance parameter ϕ is assumed constant over i.

In constructing the joint distribution for the Y_i, we find as many θ_i to estimate as there are observations. Given the set of p explanatory variables, GLR uses the relationship

$$g(\mu_i) = \beta_0 + x_i^T\beta = \eta_i, \tag{3}$$

satisfying (a) $\mu_i = E(Y_i)$; (b) g is a monotone, twice differentiable link function with a unique inverse, $h := g^{-1}$; (c) x_i^T is a $p \times 1$ row vector of autoscaled explanatory variables; (d) β is the $p \times 1$ unknown parameter vector and β_0 is the unknown intercept; (e) the estimation of β or β_0 does not depend on having knowledge of any nuisance parameter, ϕ. When $g(\mu_i) = \theta_i = \eta_i$, we have the canonical link. In cases in which $p < N$, (3) can reduce the dimensionality of estimation. For a thorough overview and standard theory of the generalized linear model, refer to McCullagh and Nelder (1989).

Referring to the relationship in (3), the log-likelihood equation can be expressed, using the canonical link, as

$$l(\beta_0, \beta; X) = \sum_{i=1}^{N} \{[y_i\eta_i + c(\eta_i)] + d(y_i)\}. \tag{4}$$

The ML estimation of the parameters is typically based on maximizing (4) through the method-of-scoring iterative equations, which simplifies to

$$\tilde{\eta}_t = \tilde{\beta}_{0,t-1}1 + X(X^T\tilde{V}_{t-1}X)^{-1}X^T\tilde{V}_{t-1}\tilde{y}^*_{t-1}, \qquad (5)$$

where, if convergence is attained, the estimated information matrix $\tilde{\Phi} = X^T\tilde{V}X, \tilde{V} = \text{diag}(\tilde{v}_{ii}) = \text{diag}[\{h'(\tilde{\eta}_i)\}^2/\text{var}(Y_i)], \tilde{y}^*_i = \tilde{\eta}_i + \tilde{e}_i/h'(\tilde{\eta}_i)$, and $\tilde{e}_i = y_i - \tilde{\mu}_i$. As mentioned, the scalar intercept $\tilde{\beta}_0$ is the weighted mean of the adjusted dependent vector $\tilde{y}^*$ using the weights in $\tilde{V}$ and is uncorrelated with the estimation of β. Here the estimates of V and y^* must be updated at each iteration step until convergence because they are a function of the iterated $\tilde{\eta}_{t-1}$.

3. IRPLS ESTIMATION FOR GENERALIZED LINEAR REGRESSION

Much like (iteratively reweighted) principal-component (IRPC) estimation in GLR, IRPLS estimation also produces a sequence of constructed or latent variables that are linear combinations of the autoscaled explanatory variables and are useful to predict the response variable of interest. In addition, both the IRPC and IRPLS constructed variables form an orthogonal sequence in a weighted metric. In Section 4, IRPC will be revisited, and specific algorithmic differences between these two techniques will be revealed, and then Section 5 will provide suggestions for choosing the optimal number of constructed variables to be used in the final model. One key difference between IRPC and IRPLS is the mechanism for choosing the loadings associated with the linear combination constructed variables. IRPC uses the eigenvector loadings associated with components of the information matrix that have a large variance, whereas IRPLS chooses loadings based on the strength of linear correlation of an explanatory variable with the (adjusted) dependent variable.

Perhaps what distinguishes IRPLS even further from IRPC estimation is that, once IRPLS constructs the first latent variable, it is immediately related to the (adjusted) dependent variable. Furthermore, the second latent variable is constructed, through residuals, in a subspace that is orthogonal to this first latent variable. Specifically, we work with the residuals after regressing out the first latent variable from both the (adjusted) dependent variable and then from the matrix of explanatory variables. The loadings for the second latent variable now depend on the strength of linear correlation between this residual (adjusted) dependent variable and the residual explanatory variables. This process continues for as many components as are needed in the model. (IR)PLS regression is sometimes referred to as *criss-cross* regression because it sequentially regresses latent variables from the (adjusted) dependent variable and then from the (residual) explanatory variable matrix. Critics of the PLS estimation approach point out its highly nonlinear features and claim it as more of an algorithm than a linear model. Helland (1988), however, elegantly removed much of this algorithmic armor and showed equivalence between various PLS algorithms. These equivalencies also hold for IRPLS.

The beauty of (IR)PLS is that only two (iterated) matrix multiplications are needed for each desired rank estimate, and moreover the (iterated) moment matrix calculations are not needed. Thus (IR)PLS can be an alternative estimation method when other techniques are prohibited by large matrix inversion or diagonalization. IRPLS borrows features of both the PLS algorithm and the GLR method-of-scoring algorithm. In addition to iterating the observation weights and the adjusted dependent vector, the IRPLS algorithm also simultaneously iterates the latent variables, their loadings, along with their relationship to the response variable, until specified convergence. As mentioned, the explanatory variable space is carved out into orthogonal latent variables, in a weighted metric. The analog to the dependent variable in the PLS algorithm is the iterated adjusted dependent vector in the IRPLS algorithm. An important feature of IRPLS is that the following two decompositions, of the data matrix and of the adjusted dependent vector, are carried out together:

$$E_0 \equiv X = \sum_{j=1}^{K} t_j p_j^T + E_K \qquad (6)$$

and

$$f_0 \equiv y^* = \sum_{j=1}^{K} q_j t_j + f_K, \qquad (7)$$

where the t_j are N-vector latent variables, p_j are the loadings, and E_K is a residual matrix. When $K = R =$ column rank$(\tilde{\Phi})$, we have $E_R = 0$. The q_j are scalar coefficients, and f_K is an N vector of residuals. The uniqueness of the t_j's and p_j's comes from imposing conditions of orthogonality.

I now provide one form of the IRPLS algorithm for GLR. It may be useful to reference the clear presentation of the PLS algorithm of Martens and Næs (1989). I present my algorithm in four parts:

1. Line 1 of the following algorithm provides one suggestion for the initializations of the algorithm. It should be clear that $\hat{E}_0$ is the autoscaled X (spectra) matrix and that $\hat{f}_0$ is the usual method of scoring the adjusted dependent variable, which must be iterated. The initial values for this adjusted dependent are usually based on a suitably transformed version of the observed y, denoted as $\psi(y)$. Care must be taken, however, to avoid infinite values of the transformed version. For example, $\psi_P(y) = \ln(y + .5)$ and $\psi_B(y) = (y + .5)/2$ work well for Poisson and Bernoulli responses, respectively.

2. Lines 2(a)–2(b) of the algorithm iterate and construct the latent variables.

3. The method-of-scoring portion of the algorithm is given in lines 2(c)–2(f).

4. Once the estimated latent variables are constructed and converged, an appropriate GLR is performed in lines 3–4.

Algorithm IRPLS.

1. Initialize $\hat{E}_0 \leftarrow X; \hat{f}_0 \leftarrow \psi(y); \hat{V} \leftarrow \{h'[\psi(y)]\}^2/\text{var}(Y)$
2. Iterate until $\Delta\hat{\eta}$ small
   (a) For $k = 1$ to R
      i. $\hat{w}_k \leftarrow (\hat{f}_{k-1}^T\hat{V}\hat{E}_{k-1}\hat{E}_{k-1}^T\hat{V}\hat{f}_{k-1})^{0.5}\hat{E}_{k-1}^T\hat{V}\hat{f}_{k-1}$ #(unit length) orthog. loadings
      ii. $\hat{t}_k \leftarrow \hat{E}_{k-1}\hat{w}_k$ #latent variables such that $\hat{V}^{1/2}\hat{t}_k$ orthogonal
      iii. $\hat{t}_k \leftarrow \text{scale}\{\hat{t}_k, \text{center} = \text{wt.mean}(\hat{t}_k, \text{wt} = \hat{V}), \text{scale} = SS(\hat{t}_k)\}$
      iv. $\hat{q}_k \leftarrow$ coefficient lsfit ($\hat{f}_{k-1}$ on $\hat{t}_k$, wt $= \hat{V}$, no intercept)
      v. $\hat{f}_k \leftarrow \hat{f}_{k-1} - \hat{t}_k\hat{q}_k$
      vi. $\hat{p}_k \leftarrow$ coefficients lsfit ($\hat{E}_{k-1}$ on $\hat{t}_k$, wt $= \hat{V}$, no intercept)
      vii. $\hat{E}_k \leftarrow$ residuals lsfit ($\hat{E}_{k-1}$ on $\hat{t}_k$, wt $= \hat{V}$, no intercept)
   (b) end For
   (c) $\hat{\eta} \leftarrow$ wt.mean ($\hat{f}_0$, wt $= \hat{V}$) $+ \sum_{k=1}^{R}\hat{q}_k\hat{t}_k$
   (d) $\hat{V} \leftarrow \{h'(\hat{\eta})\}^2/\text{var}(Y)$
   (e) $\hat{f}_0 \leftarrow \hat{\eta} + \text{diag}\{1/h'(\hat{\eta}_i)\}\{y - h(\hat{\eta})\}$
   (f) $\hat{E}_0 \leftarrow \text{scale}\{X, \text{center} = \text{wt.mean}(X, \text{wt} = \hat{V}), \text{scale} = SS(X)\}$
3. Choose $s \ni \|\hat{f}_{s+1}\|$ small, $s \leq R$
4. $\text{glm}(y \sim \hat{t}_1 \cdots \hat{t}_s)$

I now focus on the second portion, or lines 2(a)–2(b), of this algorithm, while moving from step $k-1$ to step $k, k = 1, \ldots, R = \text{column rank}(\tilde{\Phi})$. As seen from line 2(a)i, the adjusted dependent vector residuals, $\hat{f}_{k-1}$ (in step $k-1$), are partially regressed on the explanatory variable residuals, $\hat{E}_{k-1}$. This partial regression consists of computing the weighted covariance and using this vector to construct latent variables [line 2(a)ii]. Next, the adjusted dependent vector residuals (in step $k-1$) are regressed on the current latent variable (in step k) [line 2(a)iv]. The result of this fitted value is then subtracted from the residuals (in step $k-1$) to form the next sequence of adjusted dependent vector residuals (step k) [line 2(a)v]. The explanatory variables residuals (in step k) are formed by subtracting from the residuals $\hat{E}_{k-1}$ its (weighted) projection on the estimated kth latent variable [line 2(a)vii]. An alternative, but equivalent IRPLS algorithm can be extended from the work of Martens (1985) and is provided in the Appendix. The orthogonal loading vectors of the alternative algorithm are now found by multiple regression. Arguments for the equivalence of the two algorithms can be borrowed from Helland (1988, theorem 2.1) by treating the iterated adjusted dependent vector as the current *dependent* variable, in a weighted metric.

In addition to orthogonal loading vectors w_k, I mentioned that the vectors $\hat{V}^{1/2}\hat{t}_k, (k = 1, 2, \ldots)$ build up an orthogonal basis of the matrix of explanatory variables $X = \hat{E}_0$. The matrix $\hat{E}_k$ is the projection of $X = \hat{E}_0$ orthogonal to $\hat{V}^{1/2}\hat{t}_k$. Perhaps this is best seen through the following recursive formulation:

$$\hat{E}_k = \hat{E}_{k-1} - \frac{\hat{t}_k\hat{t}_k^T\hat{V}\hat{E}_{k-1}}{\hat{t}_k^T\hat{V}\hat{t}_k} = \prod_{i=1}^{k}\left(I - \frac{\hat{t}_i\hat{t}_i^T\hat{V}}{\hat{t}_i^T\hat{V}\hat{t}_i}\right)X, \quad (8)$$

since $\hat{t}_i^T\hat{V}\hat{t}_{i'} = 0$ for $i \neq i'$. Note that $\hat{t}_k$ in step 2(a)ii of the algorithm is a gradient step, $\partial l/\partial\beta$. This is particularly enlightening because it draws a connection between IRPLS and the conjugate gradient algorithm applied to the GLR normal equations, $X^T\{y - h(\eta)\} = 0$. A similar recursive formula to (8) exists for the adjusted dependent vector $\hat{y}^* = \hat{f}_0$; that is,

$$\hat{f}_k = \prod_{i=1}^{k}\left(I - \frac{\hat{t}_i\hat{t}_i^T\hat{V}}{\hat{t}_i^T\hat{V}\hat{t}_i}\right)\hat{y}^*. \quad (9)$$

Recursive relationships additionally exist for the GLR weight vectors (see Helland 1988, eq. 3.3). Given converged estimates of $\tilde{w}, \tilde{V}$, and the converged adjusted dependent vector $\tilde{y}_k^{*\text{PLS}}$ (based on $k \leq R$ components), I have

$$\tilde{w}_{k+1} = \{I - \tilde{\Phi}\tilde{W}_k(\tilde{W}_k^T\tilde{\Phi}\tilde{W}_k)^{-1}\tilde{W}_k^T\}X^T\tilde{V}\tilde{y}_k^{*\text{PLS}}, \quad (10)$$

where $\tilde{W}_k = (\tilde{w}_1\|\cdots\|\tilde{w}_k)$ is the matrix of the first s orthogonal loading vectors from step 2(a)i of the algorithm and $X_{N\times p}$ is the autoscaled matrix of explanatory variables. The key to (10) is to notice that

$$\begin{aligned}\tilde{w}_{k+1} &= cE_k^T\tilde{V}\tilde{f}_k \\ &= cX^T\{I - P_{S(k)}\}\tilde{V}\{I - P_{S(k)}\}\tilde{y}_k^{*\text{PLS}} \\ &= cX^T\tilde{V}\{I - P_{S(k)}\}\tilde{y}_k^{*\text{PLS}},\end{aligned}$$

where we define $S(k) \equiv \text{span}(\hat{t}_1, \ldots, \hat{t}_k) = \text{span}(X\hat{w}_1, \ldots, X\hat{w}_k)$ and $P_{S(k)}$ as the projection operator onto the space spanned by $S(k)$. The scalar c gives unit length.

These recursive formulations, in conjunction with the span equivalence stated previously, allow us to reexpress the IRPLS algorithm as an analog of the familiar method-of-scoring algorithm; that is, for $k = S$,

$$\tilde{\eta}_s^{\text{PLS}} = \tilde{\beta}_0 1 + X\tilde{W}_s(\tilde{W}_s^T\tilde{\Phi}\tilde{W}_s)^{-1}\tilde{W}_s^TX^T\tilde{V}\tilde{y}_s^{*\text{PLS}}. \quad (11)$$

Notice the resemblance of (11) to (5). Equivalently, on convergence

$$\begin{aligned}\tilde{\beta}_s^{\text{PLS}} &= \tilde{W}_s(\tilde{W}_s^T\tilde{\Phi}\tilde{W}_s)^{-1}\tilde{W}_s^TX^T\tilde{V}\tilde{y}_s^{*\text{PLS}} \\ &= \tilde{W}_s(\tilde{P}_s^T\tilde{W}_s^T)^{-1}\tilde{q}_s, \quad (12)\end{aligned}$$

where $\tilde{P} = (\tilde{p}_1\|\cdots\|\tilde{p}_s)$ and $\tilde{q}_s = (\tilde{q}_1, \ldots, \tilde{q}_s)^T$ are both given in the IRPLS algorithm. Given a future or new observation x_{new}^{*T}, prediction could be obtained by standardizing x_{new}^{*T} into x_{new}^T, then constructing $\tilde{\eta}_{\text{new}} = \tilde{\beta}_0 + x_{\text{new}}^T\tilde{\beta}_s^{\text{PLS}}$ and using $\tilde{\mu}_{\text{new}} = h(\tilde{\eta}_{\text{new}})$. The last equivalence in (12) is due to the relationship provided by Helland (1988, sec. 3.3), $\tilde{T}_s = (\tilde{t}_1\|\cdots\|\tilde{t}_s) = X\tilde{W}_s(\tilde{P}_s\tilde{W}_s)^{-1}$. I shall now expand on this connection.

4. CONNECTING IRPLS TO PRINCIPAL-COMPONENT ESTIMATION

An interesting connection exists between ML, PC, and

IRPLS for GLR parameter estimation. Putting these various algorithms in a similar form provides a useful standard of comparison. First, it will be useful to define GLR principal components for each observation, $Z = XM$, where the (i,j)th element of Z is the score of the jth principal component for the ith observation. Define M as the $p \times p$ matrix whose jth column is the jth eigenvector of the information matrix (without the intercept), $\Phi = X^T V X$. Hence, M is an orthogonal matrix and $M^T \Phi M = \text{diag}(\lambda_j) = \Lambda$, where λ_j are the corresponding eigenvalues of Φ. In many chemometrics applications, Φ can be semi-positive definite. Denote Q as the number of iterations until convergence. The method of scoring ML algorithms in (5) can be reexpressed as

$$\tilde{\eta}^{\text{ML}} = \sum_{t=1}^{Q} \left\{ \tilde{\beta}_0 1 + X \sum_{j=1}^{p} \tilde{\lambda}_j^{-1} \tilde{m}_j \tilde{m}_j^T X^T \tilde{V} \tilde{y}^* \right\}_t, \quad (13)$$

which is undefined in the presence of zero eigenvalues. Again $\tilde{\beta}_0$ is the iterated weighted mean of the GLR-adjusted dependent vector, and $\tilde{V}$ is the usual updated GLR weight matrix. A truncated ML based on the full set of nonnull components can be defined

$$\tilde{\eta}^{\text{TR}} = \sum_{t=1}^{Q} \left\{ \tilde{\beta}_0 1 + X \sum_{j=1}^{R} \tilde{\lambda}_j^{-1} \tilde{m}_j \tilde{m}_j^T X^T \tilde{V} \tilde{y}^* \right\}_t. \quad (14)$$

Based on these R components, it will be useful to define the matrices $\tilde{\lambda}_R$ and $\tilde{M}_R$ corresponding to the diagonal matrix of nonnull eigenvalues and the associated matrix of eigenvectors, respectively, of the converged truncated ML estimate of the information matrix, if they exist. Marx and Smith (1990) provided an argument using Taylor series approximations that $\tilde{V}$ estimates V relatively well because $\text{var}(\tilde{V}_{ii})$ is not affected for training data covariate patterns. This argument carries over to the truncated ML estimate.

One of several strategies to reduce the effects of ill-conditioned information is to further delete, in sequence, terms in the sum corresponding to the $r = R - s$ smallest nonnull $\tilde{\lambda}_j$. Deletion rules in this GLR PC estimation context were also presented by Marx and Smith (1990). A candidate strategy, taking the response into consideration, is to delete components if $|\tilde{\alpha}_j \tilde{\lambda}_j^{1/2}|$ exceeds a critical value of a t distribution on $N - R - 1$ df, where $\tilde{\alpha} = \tilde{M}_R^T \tilde{\beta}^{TR}$. Using results of the converged truncated ML solution—for example, $\tilde{\beta}_0$ and $\tilde{\Phi}$—the PC estimation technique, based on $s \le R$ components, can be expressed as

$$\hat{\eta}_s^{\text{PC}} = \left\{ \tilde{\beta}_0 1 + X \sum_{j=1}^{s} \tilde{\lambda}_j^{-1} \tilde{m}_j \tilde{m}_j^T X^T \tilde{V} \sum_{t=1}^{Q} \hat{y}_{s,t}^{*\text{PC}} \right\}, \quad (15)$$

where $\hat{y}_s^{*\text{PC}}$ is again the adjusted dependent vector but this time updated using only s terms in $\hat{\eta}_{s,(t-1)}^{\text{PC}}$. If all R terms are used in PC estimation, then truncated ML estimation is achieved.

Interestingly enough, IRPLS estimation can also be less taxing if information is borrowed from the converged truncated ML solution. Based on results given in (11) with s latent variables, one candidate IRPLS iterative scheme can be expressed as

$$\hat{\eta}_s^{\text{PLS}} = \left\{ \tilde{\beta}_0 1 + X \tilde{W}_s \sum_{j=1}^{s} \tilde{\phi}_j^{-1} \tilde{\gamma}_j \tilde{\gamma}_j^T \tilde{W}_s^T X^T \tilde{V} \sum_{t=1}^{Q} \hat{y}_{s,t}^{*\text{PLS}} \right\}, \quad (16)$$

where similar to PC estimation, $\hat{y}_s^{*\text{PLS}}$ is the adjusted dependent vector using $s \le R$ latent variables in $\hat{\eta}_s^{\text{PLS}}$ and again $\tilde{W}_s = (\tilde{w}_1 \| \cdots \| \tilde{w}_s)$. The $\tilde{\gamma}_j$ and the $\tilde{\phi}_j$ correspond to the eigenvectors and eigenvalues of $\tilde{W}_s^T \tilde{\Phi} \tilde{W}_s$, respectively.

5. CHOOSING MODEL RANK

In chemometric applications requiring the identity link and approximately Normal errors, the choice of the meta parameter K is commonly made through ordinary cross-validation (CV) (Stone 1974),

$$\hat{K}^{\text{CV}} = \arg \min_{0 \le K \le R} \sum_{i=1}^{N} (y_i - h(\hat{\eta}_{K \setminus i})^2, \quad (17)$$

where $\hat{\eta}_{K \setminus i}$ is the estimate of the linear predictor using the training sample without the ith observation. In the GLR framework, the square residual is replaced with the ith observation deviance, and an analog of CV can be constructed using first-order approximations. It is often convenient to work with Akaike's information criterion (AIC), in which I now choose K based on

$$\hat{K}^{\text{AIC}} = \arg \min_{0 \le K \le R} \{\text{deviance}(\hat{\eta}_K) + 2 \dim(\beta_0, \beta; K)\}. \quad (18)$$

One potential problem with using AIC, however, is that it is unclear how to assign dimension to (18). I have constructed the latent variables; therefore, it may be improper to simply use the number of linear parameters in (3). I can also choose the *meta parameter* s based on the estimation criterion

$$\hat{K}^* = \arg \min_{1 \le s \le R} \text{trace}\{\text{MSE}(\hat{\beta}_s^{\text{PLS}})\} \quad (19)$$

or the prediction criterion

$$\hat{K}^{**} = \arg \min_{1 \le s \le R} \text{MSE}(c^T \hat{\beta}_s^{\text{PLS}}) \quad (20)$$

for all nonnull c of proper dimension. Note that the variance component in (19) can be expressed asymptotically, $\sum_{j=1}^{p} \text{var}(\hat{\beta}_{s,j}^{\text{PLS}}) = \text{trace}\{(\tilde{W}_s^T \tilde{\Phi} \tilde{W}_s)^{-1} \tilde{W}_s^T \tilde{W}_s\}$. Frank and Friedman (1993) pointed to several other model-selection criteria. Generalized cross-validation (GCV), Bayesian information (Schwartz's) criterion, and Mallow's *Cp* statistic, among others, could all be useful for (IR)PLS model selection. AIC, GCV, and *Cp* are all equal to the first order.

6. AN ILLUSTRATIVE EXAMPLE

Recall the freesia NIR spectra data introduced in Section 1. I now wish to use this spectra explanatory information to calibrate the *probability of a successful flowering bud* for a given branch source. I have assumed that

the number of successful flowering buds in a given vase, $y_i \sim$ binomial$(m_i, p_i), p_i$ is unknown, $i = 1, \ldots, 100$. I explore IRPLS estimation in a logistic regression setting. I hope that this proposed example will be useful in demonstrating the mechanics of IRPLS analysis in the GLR framework while modeling a response that is not necessarily from a symmetric or continuous distribution. The $s = 1, 2, 3$ component models were used for both IRPLS and IRPC estimation with the model provided in (2). All of these techniques converged rather quickly, between six or eight iterations. It is reasonable to compute the percent correct classification of the training data where estimated probabilities greater (less) than .5 are classified as a 1 (0). The IRPLS ($s = 1, 2, 3$) techniques had roughly the same performance with 73.51%, 74.37%, and 73.68% correct classification, respectively. All three IRPC techniques converged with 73.51% correct classification. Ideally CV should be performed; however, as mentioned, this can be computationally taxing, especially with the PC approach because the eigenstructure has to be iteratively and repetitively reconstructed. It should be pointed out that generalizations to ridge estimation can also be explored in this GLR setting. Ridge estimation can require large amounts of memory, however; in this case a 476×476 matrix must be iteratively inverted. Again the beauty of IRPLS is that it produces a sequence of rank 1 approximations useful for prediction, avoiding potentially unwieldy matrix inversions or singular value decompositions.

Despite the fact that IRPLS and PC estimation produce roughly the same percent correct classification, they produce quite different vectors of parameter estimates. Figure 2 displays the IRPLS as well as the IRPC estimated coefficients associated with each of the 476 explanatory variable wavelengths. All of these coefficient plots correspond to the autoscaled X matrix, using weights associated with the estimated diagonal elements of V. Recall that IRPLS-estimated coefficients are chosen to maximize the correlation between the explanatory wavelength information and the GLR-adjusted dependent variable in a weighted metric. Accordingly, the result of such maximization can yield a nonsystematic pattern of parameter estimation. Note, however, that there appears to be less information in the spectrum beyond 1,500 nm. On the other hand, IRPC estimation is seeking coefficients associated with linear combinations of the wavelength information that maximizes variance in the weighted explanatory wavelength space. Unlike IRPLS, the maximization associated with IRPC estimation is independent of the adjusted dependent (apart from the estimated weights). The key difference between the IRPLS and PC estimation techniques stems from differences in spectral decomposition provided in (15) and (16). The lower portion of Figure 2 displays IRPC estimation. IRPC ($s = 1$) estimation results in roughly constant influence of the 476 wavelengths on prediction of probability of successful flowering, whereas IRPC ($s = 3$) has a rough decaying cyclical influence, greatest near 650 nm. Figure 3 is a plot of the binomial variance associated with each of the 100 observations for each of the estimation techniques. This figure helps confirm the stability of estimation at the original datapoint locations.

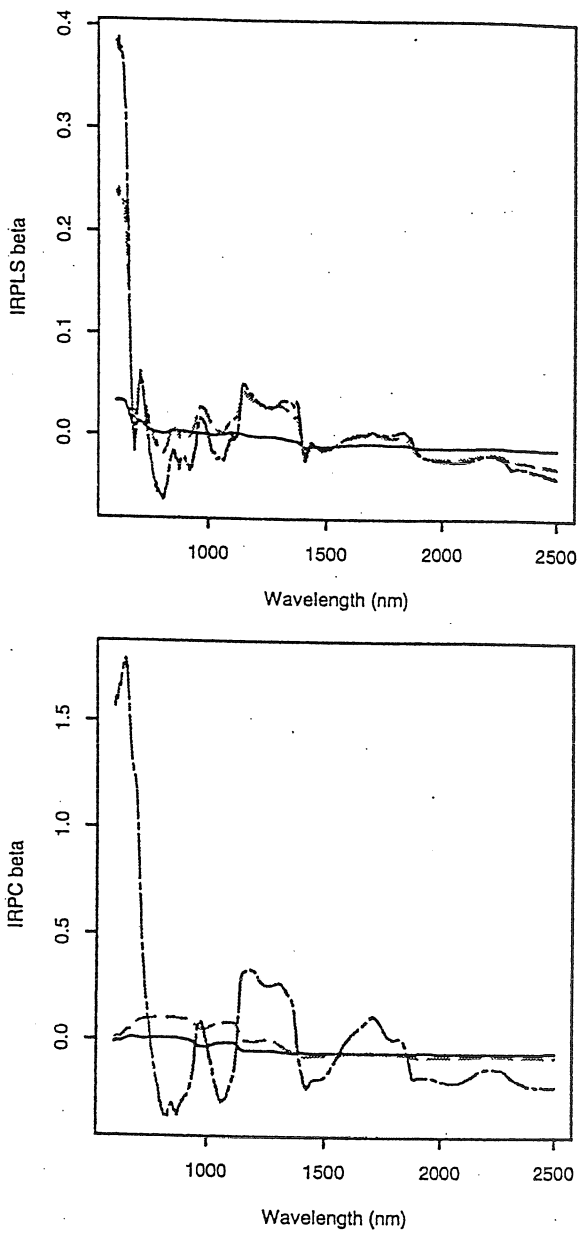

Figure 2. IRPLS and IRPC Logistic Regression Parameter Estimates Using s = 1, 2, 3 Component Models: Top Panel: ———, IRPLS (s = 1); — —, IRPLS (s = 2); ——— —, IRPLS (s = 3): Bottom Panel: ———, IRPC (s = 1); — —, IRPC (s = 2); ——— —, IRPC (s = 3).

7. DISCUSSION AND SOME AREAS OF FUTURE RESEARCH

We have extended the concept of PLS estimation into the generalized linear model framework. Using the identity link function with Normal $(0, \sigma^2 I)$ errors reduces to the work of Wold et al. (1984). As in most estimation problems, several alternative approaches exist, and rarely is any one a clear choice under all experimental conditions. PLS estimation has been criticized for being highly nonlinear and algorithmic in nature; IRPLS does not circumvent this feature. As seen in previous sections, however, IRPLS does have elegant properties and theoretical underpinnings to the

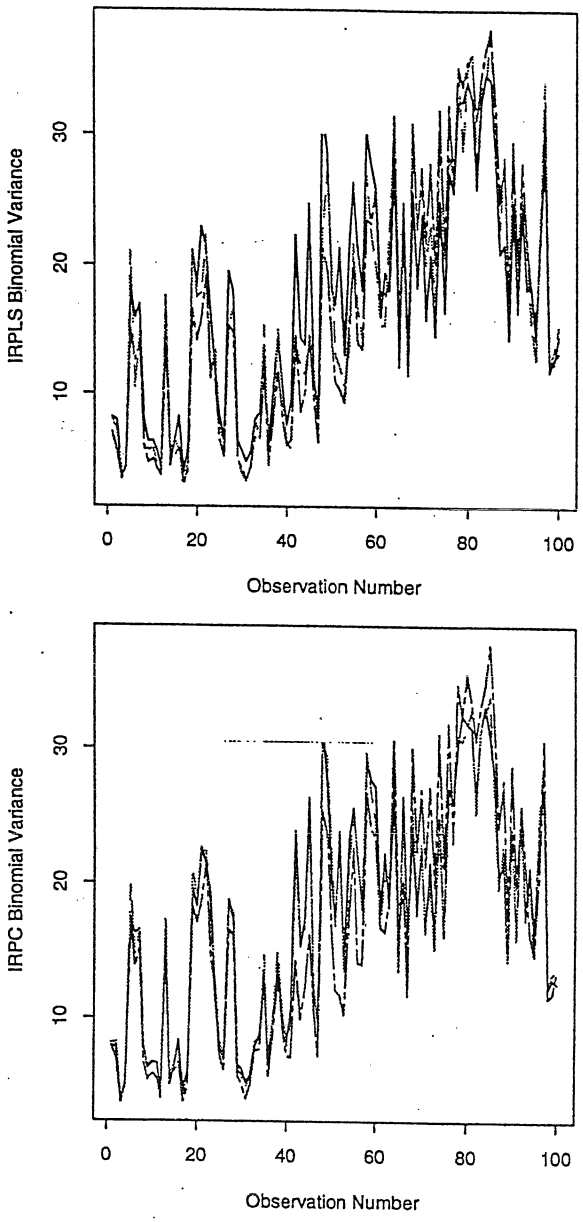

Figure 3. IRPLS and IRPC Binomial Variances Using s = 1, 2, 3 Component Models: Top Panel: ———, IRPLS (s = 1); — —, IRPLS (s = 2); ——— —, IRPLS (s = 3): Bottom Panel: ———, IRPC (s = 1); — —, IRPC (s = 2); ——— —, IRPC (s = 3).

conjugate gradient algorithm. Furthermore, it is not uncommon in spectroscopy or chemometric examples that ML solutions do not exist, and IRPLS can be computationally more efficient than PC or ridge approaches. Further research is needed to compare these estimation techniques under a variety of settings. Currently, research is surfacing that produces alternatives to PLS through fitting piecewise constants, adaptively, through penalized least squares (Land and Friedman 1994). Penalized likelihood extensions of adaptive schemes in the GLR setting appear promising but also are highly nonlinear and computationally intensive. Last, I have only considered GLR problems with the conditional distribution of the response variable, given the matrix of explanatory variables. In an effort to get a meaningful population interpretation, Helland (1990) and Næs and Helland (1993) extended the standard PLS algorithm to accommodate a joint covariance structure of the explanatory variables and the response. Future research could examine this population approach together with the wider IRPLS algorithm presented in this article.

ACKNOWLEDGMENTS

I extend my thanks to Rob Tibshirani for his helpful suggestions and to Hilko van der Voet for permission to use this data. I am also grateful for the thorough and constructive review of this work by the editor Max D. Morris, the anonymous associate editor, and two anonymous referees.

APPENDIX: ALTERNATIVE IRPLS ALGORITHM

1. Initialize $\hat{E}_0^* \leftarrow X; \hat{f}_0^* \leftarrow \psi(y); \hat{V} \leftarrow \{h'[\psi(y)]\}^2/\text{var}(Y)$
2. Iterate until $\Delta\hat{\eta}^*$ small
   (a) For $k = 1$ to R
      i. $\hat{w}_k^* \leftarrow \hat{E}_{k-1}^{*T}\hat{V}\hat{f}_{k-1}^*$ #orthogonal vector loadings
      ii. $\hat{t}_k^* \leftarrow \hat{E}_{k-1}^*\hat{w}_k^*/\|\hat{w}_k^*\|$ #latent variables
      iii. $\hat{t}_k^* \leftarrow \text{scale}\{\hat{t}_k^*, \text{center} = \text{wt. mean}(\hat{t}_k^*, \text{wt} = \hat{V}), \text{scale} = SS(\hat{t}_k^*)\}$
      iv. $T_k^* = (\hat{t}_1^* \| \cdots \| \hat{t}_k^*)$
      v. $\hat{q}_k^* \leftarrow$ coefficients lsfit ($\hat{f}_0^*$ on $\hat{T}_k^*$, wt $= \hat{V}$, no intercept)
      vi. $\hat{f}_k^* \leftarrow \hat{f}_0^* - \hat{T}_k^*\hat{q}_k^*$
      vii. $\hat{E}_k^* \leftarrow \hat{E}_{k-1}^* - \hat{t}_k^*\hat{w}_k^{*T}$
   (b) end For
   (c) $\hat{\eta}^* \leftarrow \text{wt.mean}\ (\hat{f}_0^*, \text{wt} = \hat{V}) + \hat{T}_R^*\hat{q}_R^*$
   (d) $\hat{V} \leftarrow \{h'(\hat{\eta}^*)\}^2/\text{var}(Y)$
   (e) $\hat{f}_0^* \leftarrow \hat{\eta}^* + \text{diag}\{1/h'(\hat{\eta}_i)\}\{y - h(\hat{\eta}^*)\}$
   (f) $\hat{E}_0^* \leftarrow \text{scale}\{X, \text{center} = \text{wt.mean}(X, \text{wt} = \hat{V}), \text{scale} = SS(X)\}$
3. Choose $s \ni \|\hat{t}_{s+1}^*\|$ small, $s \leq R$
4. $\text{glm}(y \sim \hat{t}_1^* \cdots \hat{t}_s^*)$

[Received February 1995. Revised January 1996.]

REFERENCES

Dobson, A. J. (1990), *An Introduction to Generalized Linear Models*, London: Chapman and Hall.

Frank, I. E., and Friedman, J. H. (1993), "A Statistical Review of Some Chemometrics Regression Tools" (with comments), *Technometrics*, 35, 109–148.

Helland, I. S. (1988), "On the Structure of Partial Least Squares Regression," *Communications in Statistics, Part B—Simulation and Computation*, 17, 581–607.

——— (1990), "Partial Least Squares Regression and Statistical Models," *Scandinavian Journal of Statistics*, 17, 97–114.

Hestenes, M. R., and Stiefel, E. (1952), "Methods of Conjugate Gradients for Solving Linear Systems," *Journal of Research, National Bureau of Standards*, 49, 409–436.

Land, S. R., and Friedman, J. H. (1994), "Adaptive Signal Regression," in *Proceedings of Statistical Computing Section, American Statistical Association*, pp. 100–105.

Le Cessie, S., and van Houwelingen, J. C. (1992), "Ridge Estimators in Logistic Regression," *Applied Statistics*, 41, 191–201.

Martens, H. (1985), "Multivariate Calibration," unpublished doctoral thesis, Technical University of Norway, Trondheim.

Martens, H., and Næs, T. (1989), *Multivariate Calibration*, New York: John Wiley.

Marx, B. D., Eilers, P. H. C., and Smith, E. P. (1992), "Ridge Likelihood Estimation for Generalized Linear Regression," in *Statistical Modelling*, eds. P. van der Heijden, W. Jansen, B. Francis, and G. Seeber, Amsterdam: North-Holland (Elsevier), pp. 227–237.

Marx, B. D., and Smith, E. P. (1990), "Principal Component Estimation for Generalized Linear Regression," *Biometrika*, 77, 23–31.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), New York: Chapman and Hall.

Næs, T., and Helland, I. S. (1993), "Relevant Components in Regression," *Scandinavian Journal of Statistics*, 20, 239–250.

Stone, M. (1974), "Cross-Validatory Choice and Assess of Statistical Descriptions" (with comments), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Wold, H. (1975), "Soft Modelling by Latent Variables: The Nonlinear Iterative Partial Least Squares Approach," in *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett*, ed. J. Gani, London: Academic Press.

Wold, S., Ruhe, A., Wold, H., and Dunn, W. J., III (1984), "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM Journal on Scientific and Statistical Computing*, 5, 735–743.