Marx:

On Ill-Conditioned Generalized Estimating Equations and Toward Unified Biased Estimation

Sonderforschungsbereich 386, Paper 182 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

On Ill-Conditioned Generalized Estimating Equations and Toward Unified Biased Estimation

Brian D. Marx
Institüt für Statistik
Ludwig-Maximilians-Universität München
Ludwigstaße 33/II
80539 München, Germany
brian@stat.lsu.edu

Abstract: I address the issue of ill-conditioned regressors within generalized estimating equations (GEEs). In such a setting, standard GEE approaches can have problems with: convergence, large coefficient variances, poor prediction, deflated power of tests, and in some extreme cases, e.g. functional regressors, may not even exist. I modify the quasi-likelihood score functions, while presenting a variety of biased estimators that simultaneously address the issues of (severe) ill-conditioning and correlated response variables. To simplify the presentation, I attempt to unite or link these estimators as much as possible. Some properties, as well as some guidelines for choosing the *meta* or penalty parameters are suggested.

Keywords: Longitudinal data, partial least squares, principal components, quasi-likelihood, repeated measures, signal regression.

1 Introduction

It is clear that generalized linear models (GLM) (Nelder and Wedderburn, 1972) have become a well established tool for a variety of applications. Much statistical research has firmly reversed the notion of bending the data to fit the *linear* model to an outlook of providing extremely flexible models to accommodate the data at hand. Dobson (1990) is an excellent introduction for readers who are unfamiliar with the GLM, and McCullagh and Nelder (1989) is a classic reference. Naturally the modeling arena has now been broadened even further in many directions, one with the pioneering work in generalized estimating equations (GEE) by Liang and Zeger (1986).

GEEs consider the multivariate setting generated from correlated or clustered response variables that can arise either from longitudinal studies or by sampling within several clusters of units. In the former, responses are often repeated measures over time on units, whereas the latter may represent sampling within various clusters, e.g. telephone area codes, families, or tree stands. Of course even more complex situations can be imagined: for example units may be given a variety of (correlated) questions that represent different response variables, where regressor coefficients are expected to have differing magnitude for different questions. In the spirit of the GLM, non-normal response variables are allowed

such as: binomial successes, Poisson counts, Gamma realizations, or any response that is a member of exponential family of distributions. Regressors are available and adjustments are made for GLM parameter estimation through a modified score function (generalized estimating equations) to account for the presence of correlation. Unlike the GLM, full likelihood parameter estimation is not usually feasible. GEEs utilize quasi-likelihood that depends on only the mean and covariance structure of the response variable (WEDDERBURN, 1974). A nice introduction to correlated data for logit models is FAHRMEIR and TUTZ (1994). An important reference for longitudinal data analysis and GEEs is DIGGLE, LIANG, and ZEGER (1995).

In view of these extensions, there also has been an increased awareness and understanding of how collinear data problems extend from standard regression particularly into the GLM through an ill-conditioned Fisher information matrix (see MACKINNON and PUTERMAN, 1992). LESAFFRE and MARX (1993) identified further problematic sources in GLM estimation, namely *ML*-collinearity: data patterns that force maximum likelihood (ML) parameter estimates to the boundary of the parameter space via a deficient rank GLM weight matrix. This latter type of collinearity is related to quasi-complete separation in logistic regression (ALBERT and ANDERSON, 1984). CLARKSON and JENNRICH (1991) further considered *extended* ML estimation in the GLM setting when some parameter estimates are infinite (at the boundary). Despite the popularity of ML parameter estimation in the GLM, ill-conditioned information can be responsible for lack of convergence, large estimated coefficient variances, poor prediction in certain regions, as well as deflating power for hypotheses concerning model assessment. As in standard multiple regression applications, the effects of ill-conditioning in training data can be nontrivial and approaches are needed for reducing the effects of these dependencies.

Producing partial models through variable subset selection (VSS) is a popular approach, but as in the standard multiple regression model tends to work best in situations characterized by true coefficient vectors with components consisting of very few (relatively) large (absolute) values (see FRANK and FRIEDMAN, 1993). Further, one should perhaps consider the research efforts extending biased estimation to the GLM framework to alleviate consequences caused by collinear data. Alternatives to VSS in GLMs have surfaced in the literature, for example: the bridge (FU, 1998), the lasso (TIBSHIRANI, 1996), iteratively reweighted partial least squares (MARX, 1996), the garrote (BREIMAN, 1995), ridge (LE CESSIE and VAN HOUWELINGEN, 1992), principal component (MARX and SMITH, 1990), among other penalized likelihood approaches. It is true that different biased techniques lead to different parameter estimation. However many of them try to achieve the same goal: to bias the solution coefficient vector away from directions for which the projected sample weighted predictor variables have small variance.

Little or no work has been done to my knowledge focusing on the detrimental features of a (nearly) singular *working* Fisher information matrix and alternative estimation within the GEE framework. In addition to addressing this issue, I aim to show that many asymptotically biased estimators are members of a broader class of shrinkage estimators for the GLM, and these can be transplanted into the GEE framework. I divide these alternative estimators into two main groups: generalized fractional principal component estimators (GFPCE) and penalized quasi-likelihood estimators (PQLE). GFPC estimation is accomplished by taking a general weighting of the principal component variables, whereas PQL estimation in many cases broadens ridge type estimation with a variety of clever penalizations. I link these two groups together. Estimation unification is not a new

idea, e.g. see HOCKING, SPEED and LYNN (1976), LEE and BIRCH (1988), STONE and BROOKS (1990), and FRANK and FRIEDMAN (1993). Section 2 gives a brief overview of the GEE, as well as notational details. Some specifics of quasi-likelihood are provided in Section 3. In Section 4, a GEE template is presented for biased estimation. Section 5 provides some common GFPC estimation techniques extended from the GLM into the GEE framework. Some PQL estimators are given in Section 6. An iteratively reweighted partial least squares estimation algorithm is suggested in Section 7. Lastly, some suggestions for GEE estimation are proposed for functional or other extremely high dimensional regressors.

2 Marginal Models: Background and Notation

Marginal models are used in situations where the primary research interest is to analyze the *marginal* mean of the response given the explanatory variables. The association between the responses is often of secondary interest or even perhaps a nuisance. Consider m_i repeated measures on unit i, $i = 1, \cdots, n$. To simplify further presentation, the reader is free to alternatively imagine m_i samples from cluster i. Let $y_i = (y_{i1}, \cdots, y_{im_i})'$ be the vector of (exponential family) responses and $X_i = (x_{i1}, \cdots, x_{im_i})'$ be the corresponding $m_i \times (p+1)$ matrix of regressors (including an intercept term). These regressors can either vary (e.g. age) or remain constant (e.g. ethnicity) within units in such studies. The specifications for marginal models in a non-normal setting are presented in FAHRMEIR and TUTZ (1994, Section 3.5.2) and can be outlined as:

1. The marginal means
$$\mu_{ij} = E(y_{ij}|x_{ij}) = h(\eta_{ij}), \tag{1}$$
where $h(\cdot)$ is the (monotone and twice differentiable) inverse link function, $\eta_{ij} = x'_{ij}\beta$ and β is the unknown coefficient vector. Some care should be taken in defining both x_{ij} and β in (1):

   (a) The coefficients may truly be *population averaged* in the sense that they are common across units (samples) and time (cluster) yielding $\mu = h(X\beta)$, where X is the $N \times (p+1)$ matrix ($N = \sum_{i=1}^{n} m_i$) and β is $(p+1) \times 1$;

   (b) The assumption that the coefficients are homogeneous across units (samples) and time (cluster) can be relaxed. Unit-specific parameters can be proposed when the number of repeated measures for each unit is large enough. In such a setting define $\beta^* = (\beta'_1, \cdots, \beta'_n)'$ of dimension $n(p+1) \times 1$ (each β_i of dimension $(p+1) \times 1$). The design matrix X^* (of dimension $N \times n(p+1)$) is block diagonal where the n blocks are of dimension $m_i \times (p+1)$, $i = 1, \cdots, n$;

   (c) Similar to (b) above, when the number of subjects is large enough and $m_i = m$ for all units, then time-specific parameters can be proposed. Now $\beta^{**} = (\beta'_1, \cdots, \beta'_m)'$ of dimension $m(p+1) \times 1$ and the design matrix X^{**} (of dimension $N \times m(p+1)$) is block diagonal where the m blocks are of dimension $n \times (p+1)$.

   I mainly consider case (a) in this article. Case (b) can arise in subject-specific modeling. Case (c) might occur for example when units are asked a few (correlated) questions and thus the explanatory variables may have differing significance across questions.

2. The marginal variance is function of μ_{ij}

$$\text{var}(y_{ij}|x_{ij}) = \sigma^2(\mu_{ij}) = \phi v(\mu_{ij})/\omega_{ij}, \tag{2}$$

where ϕ and $v(\mu_{ij})$ represents the scale parameter and variance function, respectively, determined by the specific exponential family member. The weights, ω_{ij}, can be depend on data grouping or be assigned to zero for missing values.

3. To account for within unit dependence, the covariance between y_{ij} and $y_{ij'}$ is a function of the marginal means and possibly an additional association parameters θ. For a known function ζ,

$$\text{cov}(y_{ij}, y_{i'j'}) = \zeta(\mu_{ij}, \mu_{i'j'}, \theta) \qquad \text{for } i = i', \tag{3}$$

and uncorrelated for $i \neq i'$. Thus for unit i, a $m_i \times m_i$ *working* covariance matrix is defined $\text{cov}(y_i) = \Sigma_i(\beta, \theta)$. As we will see in the next section, it is convenient to express Σ $(m_i \times m_i)$ in terms of the correlation matrix R, i.e. $\Sigma_i = A_i^{1/2} R(\theta) A_i^{1/2}$ with $A_i = \text{diag}(\sigma_{i1}^2, \cdots, \sigma_{im_i}^2)$. Several choices of R are provided in Liang and Zeger (1986): uncorrelated repeated observations (independence), fully unspecified, exchangable, or auto-regressive. The choice can be based on both simplicity and efficiency.

3 Score Functions and Quasi-likelihood

Analogous to quasi-likelihood (Wedderburn, 1974), the generalized estimating equations for β can be expressed (for fixed θ)

$$s_\beta(\beta, \theta) = \sum_{i=1}^{n} X_i' D_i(\beta) \Sigma_i^{-1}(\beta, \theta)(y_i - h(\eta_i)). \tag{4}$$

The matrices X_i and $D_i = \text{diag}(h'(\eta_{ij}))$ are of dimension $m_i \times (p+1)$ and $m_i \times m_i$, respectively. In general, some alternating estimation between $(\theta,\ \phi)$ and β iterations is needed in (4). Typically $(\theta,\ \phi)$ are estimated by either method of moments (Liang and Zeger, 1986) or, in some special logit models, by a second GEE (Liang, Zeger, and Qaqish, 1992). Given current estimates, say $(\hat{\theta},\ \hat{\phi})$, (4) is set to zero and solved by

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (\hat{F}^{(t)})^{-1} \hat{s}^{(t)}, \tag{5}$$

where the estimated *working* Fisher matrix is defined as,

$$\hat{F}^{(t)} = \sum_{i=1}^{n} X_i' D_i(\hat{\beta}^{(t)}) \Sigma_i^{-1}(\hat{\beta}^{(t)}, \hat{\theta}, \hat{\phi}) D_i(\hat{\beta}^{(t)}) X_i, \tag{6}$$

and $\hat{s}^{(t)} = s(\hat{\beta}^{(t)}, \hat{\theta}, \hat{\phi})$.

Under regularity conditions and with $\hat{\theta}$ fixed, (5) produces consistent estimates of β, and asymptotically $\hat{\beta} \sim N(\beta, F^{-1} V F^{-1})$ for any of the mentioned choices of $R(\theta)$. The matrices

$$F = \sum_{i=1}^{n} X_i' D_i \Sigma_i^{-1} D_i X_i = X' \Omega X,$$

$V = \sum_{i=1}^{n} X_i' D_i \Sigma_i^{-1} \mathrm{cov}(y_i) \Sigma_i^{-1} D_i X_i$, Ω = block diagonal(Ω_i), and $\Omega_i = D_i \Sigma_i^{-1} D_i$. In practice, consistent estimation for β is achieved by using $\hat{V}$ for V: substituting converged $\hat{\beta}$ into D_i, $(\hat{\theta}, \hat{\phi})$ into Σ_i, and using $(y_i - h(\hat{\eta}_i))(y_i - h(\hat{\eta}_i))'$ for $\mathrm{cov}(y_i)$. Thus the estimated covariance matrix, often referred to as the *sandwich matrix*, is

$$\hat{H} = \hat{F}^{-1} \sum_{i=1}^{n} X_i' \hat{D}_i \hat{\Sigma}_i^{-1} (y_i - h(\hat{\eta}_i))(y_i - h(\hat{\eta}_i))' \hat{\Sigma}_i^{-1} \hat{D}_i X_i \hat{F}^{-1}, \tag{7}$$

which is useful for constructing confidence intervals and Wald-statistics for β. If the working covariance structure is correctly specified, then estimation is asymptotically efficient. It can be useful to re-express (5) as

$$\hat{\beta}^{(t+1)} = (\hat{F}^{(t)})^{-1} X' \hat{\Omega}^{(t)} \hat{y}^{\star(t)}, \tag{8}$$

where $y^\star = \eta + D^{-1}(y - h(\eta))$ is the adjusted dependent vector and $D = \mathrm{diag}(D_i)$.

It should be pointed out that in the case of the independence working model ($R(\theta) = I$), the working covariance does not have an association parameter θ and is of the form $\Sigma_i(\beta) = D_i$ for the canonical link function. The GEE equations have the GLM score form as if the observations were independent. Despite some loss of efficiency, McDonald (1993) recommends an independence working structure (in the logit model) whenever correlation is regarded as a nuisance.

4 A Unifying GEE Template

For the developments to follow, it will be useful to work with the principal components for each observation, Z, where the $((i,j),k)$th element of Z is the score of the kth principal component for the (i,j)th observation. Define the principal components $Z = X_{-1}M$, where X_{-1} is the $N \times p$ matrix of regressors (case a), and M is the $p \times p$ matrix whose kth column is the kth eigenvector of the information matrix F_{-1} constructed without the intercept, $k = 1, \cdots, p$. Hence M is an orthogonal matrix and $M' F_{-1} M = \mathrm{diag}(\lambda_k) = \Lambda$, where λ_k are the corresponding eigenvalues of F_{-1}. In light of this decomposition, it is clear that near-singular $\hat{F}$ can have a dramatic impact on the stability, and even existence, of the sandwich covariance matrix of $\hat{\beta}$ in (7), as well as the iterative algorithm in (8). Marx and Smith (1990) presented details of some consequences.

There is great controversy regarding standardization of the regressors; it is not my objective to solve these disputes. For the methods to follow, similar derivations can be made for alternatives such as: standardization of F (Belsley, 1991), quasi-standardizations (Marx, 1992), or lack of standardization when the explanatory variables have the same units. In any case, solutions can be transformed back into the natural metric.

I recommend to standardize such that estimation of the intercept coefficient is uncorrelated with estimation of the other coefficients. In general this can be achieved by centering and scaling X_{-1} (without intercept) using a weighted mean and weighted sum of squares, respectively. The $N = \sum_{i=1}^{n} m_i$ vector of weights can be computed (iteratively) by defining the (ith, jth) weight as the jth column total τ_{ij} of Ω_i. With such a choice the intercept term is trivially the weighted mean of the adjusted dependent variable $y^\star$.

I rewrite the linear predictor as

$$\eta_{ij} = \beta_0 + z_{ij}' \alpha, \tag{9}$$

where z'_{ij} is the row of Z corresponding to the (ith, jth) observation, $\alpha = M'\beta$, and β_0 is the intercept term. Equation (1) provides estimation for the orthogonally transformed full principal component model.

However, to introduce a class of estimators, we generalize the model further through a matrix Γ. Let

$$\begin{aligned} \eta^*_{ij} &= \beta_0 + z'_{ij}\Gamma\alpha \\ &= \beta_0 + z'_{ij}\alpha^*, \end{aligned} \tag{10}$$

where $\alpha^* = \Gamma\alpha$. The matrix $\Gamma = \text{diag}(\gamma_k)$ is a diagonal weight matrix with γ_k usually contained in the closed unit interval.

For given Γ, generalized estimating strategies may be applied to estimate α^* in (10). One potentially expensive candidate iterative scheme can be defined as

$$\begin{aligned} \tilde{\eta}^{*(t+1)} &= \tilde{\beta}_0^{(t)}\mathbf{1}_N + \tilde{Z}^{(t)}\tilde{\alpha}^{*(t)} + \tilde{Z}^{(t)}\Gamma\tilde{\Lambda}^{-1(t)}\tilde{Z}^{(t)'}\tilde{\Omega}^{(t)}\tilde{D}^{(t)}\tilde{e}^{*(t)} \\ &= \tilde{\beta}_0^{(t)}\mathbf{1}_N + \tilde{Z}^{(t)}\Gamma\tilde{\Lambda}^{-1(t)}\tilde{Z}^{(t)'}\tilde{\Omega}^{(t)}\tilde{D}^{(t)}\tilde{y}^{\star(t)}, \end{aligned} \tag{11}$$

where $\tilde{e}^* = y - h(\tilde{\eta}^*)$ and the entries of the adjusted dependent variable are $\tilde{y}^\star = \tilde{\eta}^* + \text{diag}(h'(\tilde{\eta}^*))^{-1}\tilde{e}^*$.

Equation (11) is particularly taxing because new eigen-decomposition is required at each iteration through $\tilde{\Lambda}$ and $\tilde{Z} = X_{-1}\tilde{M}$. Some relief is possible if the converged GEE estimated parameter estimates and working Fisher information matrix both exist from (5) and (6) respectively. Define the matrices $\hat{M}$ and $\hat{\Lambda}$ to be the respective eigenvector matrix and diagonal matrix of eigenvalues of the converged $\hat{F} = X'_{-1}\hat{\Omega}X_{-1}$. The intercept $\hat{\beta}_0$ is the weighted ($\hat{\tau}_{ij}$) mean of the converged adjusted dependent variable $\hat{y}^\star$. Whenever possible I recommend substitution of $\hat{\beta}_0$, $\hat{Z}$, $\hat{\Lambda}$, $\hat{\Omega}$, and $\hat{D}$ into (11):

$$\begin{aligned} \tilde{\eta}^{*(t+1)} &= \hat{\beta}_0\mathbf{1}_N + \hat{Z}\tilde{\alpha}^{*(t)} + \hat{Z}\hat{\Gamma}\hat{\Lambda}^{-1}\hat{Z}'\hat{\Omega}\hat{D}\tilde{e}^{*(t)} \\ &= \hat{\beta}_0\mathbf{1}_N + \hat{Z}\hat{\Gamma}\hat{\Lambda}^{-1}\hat{Z}'\hat{\Omega}\hat{D}\tilde{y}^{\star(t)}. \end{aligned} \tag{12}$$

In (12) only $\tilde{\alpha}^*$, $\tilde{e}^*$ and $\tilde{y}^\star$ are iterated. As (12) also shows, the weight matrix Γ is not always known in practice. We will see that it too can be useful to use $\hat{\Gamma}$ with converged GEE estimators.

Justification of (12) is based on variance arguments of $\hat{\eta}$ showing that GEEs produce relatively stable estimates at the *original* data point locations, even in the presence of ill-conditioned information. As a demonstration, consider $\text{var}(\hat{\eta}_{ij})$ when $\Gamma = I_p$ (identity). Apart from the intercept term, $\text{var}(\hat{\eta}_{ij}) \propto \sum_{k=1}^{p} z^2_{ijk} m_k m'_k \lambda_k^{-1}$. However, by nature of principal components, small λ_k cannot generally co-exist with large entries in the kth column of Z, z_k.

A conversion can be made to the X metric using

$$\tilde{\beta} = \hat{M}\tilde{\alpha}^*, \tag{13}$$

with $\text{var}(\tilde{\beta}) = M'\Gamma MHM\Gamma M'$. If Γ is chosen a priori, then $\tilde{\beta}$ is also approximately normally distributed. Standard approaches can be taken to uncenter and unscale the regression coefficients to their natural units.

A simple and useful approximation to the above iterative approach is to directly shrink the converged GEE estimate of α, i.e. $\hat{\alpha} = \hat{M}'\hat{\beta}$. However GEE convergence must be met. A *one-step* estimator for α can be constructed as

$$\tilde{a}^* = \hat{\Gamma}\hat{\alpha}. \tag{14}$$

See SCHAEFER (1986). Thus $\tilde{b} = \hat{M}\hat{\Gamma}\hat{\alpha}$ is the *one-step* estimate of β. The essential difference between $\tilde{a}^*$ in (14) and $\tilde{\alpha}^*$ in (11) is the residual in the adjusted dependent variable. In fact, the two estimators are identical except that $\tilde{a}^*$ utilizes the residual $\hat{e}_i$, whereas $\tilde{\alpha}^*$ uses $\tilde{e}_i^*$.

5 Some Generalized Fractional PC Estimators

5.1 GEE Estimation

When $\Gamma = I$, we have $\eta^* = \eta$, $\tilde{\beta} = \hat{M}\tilde{\alpha}^* = \tilde{b} = \hat{\beta}$ as in (5).

5.2 Fractional Estimation

One of several strategies to reduce the effects of ill-conditioned information is to delete, in sequence, terms in the sum corresponding to the $r = p - s$ smallest $\hat{\lambda}_j$ (or t-like statistics, $\hat{\alpha}_k \hat{\lambda}_k^{-1/2}$). Hence η_i^* is of the form given in (10) with,

$$\Gamma^{pc} = \begin{pmatrix} I_{s-1} & 0 & 0 \\ 0 & \rho & 0 \\ 0 & 0 & 0_r \end{pmatrix}. \tag{15}$$

The quantity $0 < \rho \leq 1$ is the fraction of the sth principal component used in parameter estimation, commonly $\rho = 1$.

Hence a principal component parameter estimate based on s $(= 1)$ components can be expressed as $\tilde{\beta}_s^{pc} = \hat{M}\tilde{\alpha}_s^{pc}$. The asymptotic reduction in variance is given as,

$$\mathrm{var}(\tilde{\beta}_s^{pc}) = F_{-1}^{-1} - M_r \Lambda_r^{-1} M_r',$$

which can be substantial with the deletion of components associated with small eigenvalues. The asymptotic bias can further be quantified as

$$E(\tilde{\beta}_s^{pc}) = \beta - M_r \alpha_r.$$

We have an approximate result that

$$\tilde{\beta}_s^{pc} \sim N(M_s \alpha_s,\ M_s \Lambda_s^{-1} M_s').$$

Alternatively selection of s can be based on cross-validation, some estimation criterion, e.g.

$$s^* = \arg\min_{0 \leq s \leq p} \left[\mathrm{tr}\{\mathrm{MSE}(\tilde{\beta}_s^{pc})\} - \mathrm{tr}\{\mathrm{MSE}(\hat{\beta})\} \right], \tag{16}$$

or some prediction criterion, e.g.

$$s^{**} = \arg\min_{0 \leq s \leq p} \{\mathrm{MSE}(c'\tilde{\beta}_s^{pc}) - \mathrm{MSE}(c'\hat{\beta})\}, \tag{17}$$

for all nonnull c of proper dimension. Calculations of (16) and (17) would involve the sandwich matrix H presented in (7).

5.3 Stein Estimation

The biased technique based on STEIN (1960) can be used within GEE by scaling parameter estimates $\hat{\beta}$ by a constant $0 \leq \gamma \leq 1$. This in effect shrinks its inflated norm caused by ill-conditioning. The constant γ can be chosen by a variety of methods, but popular choices minimize a version of the asymptotic MSE. One choice could be

$$\gamma_0 = \arg \min_{0 \leq \gamma \leq 1} E(\gamma\hat{\beta} - \beta)'(\gamma\hat{\beta} - \beta), \tag{18}$$

with the solution

$$\hat{\Gamma}_0^S = \text{diag}\left(\frac{\hat{\alpha}'\hat{\alpha}}{\hat{\alpha}'\hat{\alpha} + \sum_{k=1}^{p} \hat{\lambda}_k^{-1}}\right). \tag{19}$$

Alternatively a choice based on

$$\gamma_1 = \arg \min_{0 \leq \gamma \leq 1} E(\gamma\hat{\beta} - \beta)'H(\gamma\hat{\beta} - \beta) \tag{20}$$

produces the solution

$$\hat{\Gamma}_1^S = \text{diag}\left(\frac{\hat{\alpha}'\hat{\Lambda}\hat{M}\hat{V}^{-1}\hat{M}'\hat{\Lambda}\hat{\alpha}}{\hat{\alpha}'\hat{\Lambda}\hat{M}\hat{V}^{-1}\hat{M}'\hat{\Lambda}\hat{\alpha} + p}\right). \tag{21}$$

5.4 Sclove Estimation

SCLOVE (1968) proposed improved estimators for coefficients in linear regression. These developments are particularly useful when explanatory variables are ordered, as in principal component regression, hence aligned with the framework of this section. Sclove suggested only shrinking a subset of the components. The analogue of this concept in the GEE context results in

$$\Gamma^{SC} = \begin{pmatrix} I_s & 0 \\ 0 & \gamma I_r \end{pmatrix}, \tag{22}$$

where $0 \leq \gamma \leq 1$. Note when $\gamma = 0$, the Sclove estimator reduces to a principal component estimator ($\rho = 1$). Sections 5.2 and 5.3 provide some guidelines for choosing s and γ.

6 Penalized Quasi-likelihood Approaches

Consider modifying (4) such that penalized quasi-likelihood estimators can be generally expressed as

$$\tilde{\beta}_\kappa^{PQL} = \mathcal{Z}_\beta\{s(\beta, \theta) - \kappa\frac{\partial P(\beta)}{\partial \beta}\},$$

where $\mathcal{Z}\{\cdot\}$ is the zero solution, $P(\beta)$ is a penalty function for the coefficient vector and κ is the non-negative regularization parameter. Below some penalized likelihood GLM estimators are extended to the GEE framework through PQL.

6.1 Penalizing Adjacent Coefficients

There exist experimental situations when the set of regressors have some ordering, and it is reasonable to assume that adjacent coefficients cannot differ too much from each other. Eilers and Marx (1996) imposed a penalization scheme to B-spline coefficients in a variety of smoothing applications using penalized likelihood (P-splines). Such notions of penalized estimation can be extended into the GEE setting. The penalty matrix is constructed using the dth order difference operator Δ^d, $d = 0, 1, \cdots, p-1$ (Whittaker, 1923). Define

$$\begin{aligned} \Delta^1(\beta_k) &= \beta_k - \beta_{k-1} \\ \Delta^2(\beta_k) &= \Delta^1(\Delta^1(\beta_k)) = \beta_k - 2\beta_{k-1} + \beta_{k-2}, \end{aligned} \tag{23}$$

for $k = d+1, \cdots, p$. Higher order differences can be found by induction and in general can be expressed as $P^d\beta$, where P^d is the $(p-d) \times p$ banded matrix constructed by taking d row differences of I_p. Estimation involves penalizing the score in (4) for fixed $(\hat{\theta}, \hat{\phi})$,

$$\tilde{\beta}^D = \mathcal{Z}_\beta\{s(\beta, \theta) - \kappa P^{d\prime} P^d \beta\}, \tag{24}$$

where $\kappa \geq 0$. The PQL-solution results in modifying (8) as

$$\tilde{\beta}^{D(t+1)} = (\tilde{F}^{(t)} + \kappa P^{d\prime} P^d)^{-1} X' \tilde{\Omega}^{(t)} \tilde{y}^{\star(t)}. \tag{25}$$

Upon convergence, the covariance matrix simplifies: $\tilde{\text{cov}}(\tilde{\beta}^D) = (\tilde{F} + \kappa P^{d\prime} P^d)^{-1} \tilde{V} (\tilde{F} + \kappa P^{d\prime} P^d)^{-1}$. If needed, the intercept term can be left unpenalized by augmenting a column 0_{p-d} to P^d.

Linking (25) back to (12) would require a non-diagonal Γ matrix. The corresponding GFPC estimator could use

$$\hat{\Gamma}^D = (\hat{\Lambda} + \kappa \hat{M}' P^{d\prime} P^d \hat{M})^{-1} \hat{\Lambda}, \tag{26}$$

when the regressors are consistently centered and scaled. Note that (26) penalizes differences of adjacent β_k, not the α_k. The penalized solution is given upon convergence as $\tilde{\beta}^D_\kappa = \hat{M}\tilde{\alpha}^D$.

The above result routinely allows smoothing in GEEs using penalized B-splines (regression splines). Further this smoothing approach is closely related to the Demmler-Reinsch basis (Demmler and Reinsch, 1975), which would orthogonally rotate a B-spline basis, $B(= X_{-1})$ using a spectral decomposition of $F^{-1/2} P^{d\prime} P^d F^{-1/2}$, rather than F. Such smooth bases do not need to be centered and scaled, and moreover have a span which includes the intercept hence β_0 can be set to zero.

6.2 Ridge Component Estimation

Ridge regression for the GLM (le Cessie and van Houwelingen, 1992 and Marx, Eilers and Smith, 1992) was introduced into the statistical literature as a restricted or penalized ML estimate for stabilizing regression coefficients in the presence of ill-conditioned information. The penalty is function with the penalty proportional to the square norm of β, $\|\beta\|^2 \leq \kappa$. Notice that when $d = 0$ in the above section, we have P^0 equal to the

identity matrix of dimension p, and the difference operator reduces to exactly the ridge regression solution for the GEE.

There is a vast literature on techniques for choosing the nonnegative ridge regularization parameter. One choice could be to minimize the error of prediction through cross-validation (CV). Le Cessie and van Houwelingen considered a variety of prediction error criteria useful for logistic regression. Marx, Eilers and Smith considered extensions of the ridge C_p statistic.

NYQUIST (1991) demonstrated that the above ridge estimator (in the GLM) is a special case of equality constraint estimation of the form $M'\beta = \alpha$, where α_k iid $(0, \kappa^2)$. Some care has to be taken when considering a Bayesian connection. Although it may be tempting to, for example, take a Normal prior on $\beta \sim N(0, Q)$ (which specializes to a penalized likelihood), the full likelihood often does not exist. For that matter the quasi-likelihood does not always exist.

6.3 Regression Shrinkage via the Lasso, Garrote, and Bridge

Other more nonlinear shrinkage estimators exist, one being the lasso technique proposed by TIBSHIRANI (1996). The idea is similar to ridge estimation except we now solve

$$\tilde{\beta}^L = \mathcal{Z}_\beta\{s(\beta, \theta) - \kappa \sum_{k=1}^{p} |\beta_k|\}. \tag{27}$$

One beauty of this technique is that is combines desirable features from both variable subset selection and ridge estimation, in that it shrinks some estimated coefficients and sets others to zero. For standard multiple regression, Tibshirani presented a lasso algorithm that utilizes the least squares problem with 2^p inequality constraints, and showed it must converge in a finite number of steps. Extensions and applications to generalized estimating equations appear promising, but more work needs to be done in this area. Tibshirani also proposed solving the constrained problem by iterative application of the lasso algorithm, within the method of scoring algorithm. Convergence is not guaranteed, but can be well behaved in logistic regression. Fu (1998) proposed another nonlinear bridge estimator, which is a competitor to the lasso.

The motivation for the lasso comes from non-negative garrote estimation (BREIMAN, 1995). An extension of Breiman's work into the GEE setting would now involve solving for

$$\tilde{\beta}^G = \mathcal{Z}_\beta\{s(c\beta, \theta)\} \quad \text{subject to } c_k \geq 0, \quad \sum_{k=1}^{p} c_k \leq \kappa, \tag{28}$$

where c is a p vector of non-negative constants. However this generalized garrote estimator is likely to suffer when the F matrix is severely ill-conditioned since it would depend on both the signs and magnitudes of the GEE solution.

7 GEE (Iteratively Reweighted) Partial Least Squares

Related to principle component estimation, partial least squares (PLS) (WOLD, 1975) produces a sequence of models $\{\hat{\eta}_K^{PLS}\}_1^R$, where $R = \text{rank}(\hat{F})$. Applications of PLS often

arise with high dimensional regression, e.g. $p \gg N$. Sometimes PLS is used with functional regressors, such as spectra, time series, but it should be pointed out that PLS does not use any of the ordering information among these regressors. In the next section below, I will propose an additional approach for functional regressors in the GEE setting.

A key feature of partial least squares estimation, unlike principle components, is that a latent regressor variable is constructed using response variable information. Given a latent regressor, it is then removed from the remainder of the regressor space. Marx (1996) presented the algorithmic details and properties for iterative reweighted partial least squares estimation (IRPLS) in the GLM setting. (IR)PLS estimation only needs two iterated matrix multiplications for each desired rank estimate, and moment matrix calculations are not needed. These features can be a strength in prohibitive situations involving a large matrix inversion or diagonalization. We will see that IRPLS can be further transplanted into the GEE setting, while iterating: the weight matrix (Ω), the adjusted dependent variable, and all R latent variables (until specified convergence).

The regressor subspace is carved out into R orthogonal components, in a weighted metric, i.e. the latent variables. The following two decompositions, of the data matrix and adjusted dependent variable, are carried out together,

$$E_0 \equiv X_{-1} = \sum_{k=1}^{K} t_k p_k' + E_K \tag{29}$$

$$f_0 \equiv \hat{y}^\star = \sum_{k=1}^{K} q_k t_k + f_K, \tag{30}$$

where the t_k are N-vector latent variables, p_k are K-vector loadings, E_K is a residual matrix. When $K = R$, we have $E_R = 0$. The q_k are scalar coefficients, and f_K is a N-vector of residuals. The uniqueness of the t_k's and p_k's come from imposing conditions of orthogonality.

I next provide one form of the IRPLS algorithm for GEEs. The algorithm is presented in five parts:

I. Line 1 of the algorithm below provides one suggestion for the initializations of the algorithm. It should be clear that $\hat{E}_0$ is the X_{-1} matrix (e.g. spectra) which is autoscaled, and that $\hat{f}_0$ is the usual adjusted dependent variable, which must be iterated. The initial values for this adjusted dependent variable are usually based on a suitably transformed version of the observed y, denoted as $\psi(y)$. However care must be taken to avoid infinite values of the transformed version. For example, $\psi_P(y) = \ln(y + 0.5)$ and $\psi_B(y) = (y + 0.5)/2$ work well for Poisson and Bernoulli responses, respectively;

II. Lines 2(a)-2(b) of the algorithm below iterate and construct the latent variables;

III. The ingredients for the GEE scoring portion of the algorithm are given in lines 2(c), 2(e)-2(g);

IV. Line 2(d) updates the association parameter for the working covariance matrix;

V. Once the estimated latent variables are constructed and converged, final estimates of $\hat{\beta}_s^{PLS}$ and $\hat{\eta}_s^{PLS}$ ($s \leq R$) are given in lines 3-4 below.

GEE IRPLS Algorithm

1. Initialize $\hat{E}_0 \leftarrow X_{-1}$; $\hat{f}_0 \leftarrow \psi(y)$; $\hat{\eta} \leftarrow h(\psi(y))$; $\hat{\theta} \leftarrow \theta(\hat{\eta})$ e.g., method of moments;
   $\hat{\Omega} \leftarrow$ block diag$\{\hat{D}_i(\hat{\eta}_{ij})\hat{\Sigma}_i^{-1}(\hat{\eta}_{ij}, \hat{\theta})\hat{D}_i(\hat{\eta}_{ij})\}$; $\hat{\tau}=$ column totals of $\hat{\Omega}$
2. Iterate until $\Delta\hat{\eta}$ small
   (a) For k=1 to R
      i. $\hat{w}_k \leftarrow \hat{E}'_{k-1}\hat{\Omega}\hat{f}_{k-1}/\text{sqrt}(\hat{f}^{*'}_{k-1}\hat{\Omega}\hat{E}_{k-1}\hat{E}'_{k-1}\hat{V}\hat{f}^{*}_{k-1})$ #(unit length) orthog loadings
      ii. $\hat{t}_k \leftarrow \hat{E}_{k-1}\hat{w}_k$ #latent variables such that $\hat{\Omega}^{1/2}\hat{t}_k$ orthogonal
      iii. $\hat{t}_k \leftarrow$ scale$\{\ \hat{t}_k$, center = wt.mean$(\hat{t}_k$, wt $= \hat{\tau})$, scale = SS$(\hat{t}_k)\}$
      iv. $\hat{q}_k \leftarrow$ coefficient of gls fit$(\ \hat{f}_{k-1}$ on $\hat{t}_k$, wt matrix $=\ \hat{\Omega}$, no intercept)
      v. $\hat{f}_k \leftarrow \hat{f}_{k-1} - \hat{t}_k\hat{q}_k$
      vi. $\hat{p}_k \leftarrow$ coefficients gls fit$(\ \hat{E}_{k-1}$ on $\hat{t}_k$, wt matrix $=\ \hat{\Omega}$, no intercept)
      vii. $\hat{E}_k \leftarrow$ residuals gls fit$(\ \hat{E}_{k-1}$ on $\hat{t}_k$, wt matrix $= \hat{\Omega}$, no intercept)
   (b) end For
   (c) $\hat{\eta} \leftarrow$ wt.mean$(\hat{f}_0$, wt $=\hat{\tau}) + \sum_{k=1}^{R} \hat{q}_k\hat{t}_k$
   (d) $\hat{\theta} \leftarrow \theta(\hat{\eta})$, e.g., method of moments
   (e) $\hat{\Omega} \leftarrow$ block diag$\{\hat{D}_i(\hat{\eta}_{ij})\hat{\Sigma}_i^{-1}(\hat{\eta}_{ij}, \hat{\theta})\hat{D}_i(\hat{\eta}_{ij})\}$
   (f) $\hat{\tau} \leftarrow$ column totals of $\hat{\Omega}$
   (g) $\hat{f}_0 \leftarrow \hat{\eta} + D^{-1}(\hat{\eta})\{y - h(\hat{\eta})\}$
   (h) $\hat{E}_0 \leftarrow$ scale $\{\ X$, center = wt.mean$(X$, wt $= \hat{\tau})$, scale = SS$(X)\ \}$
3. Choose $s \ni \|\hat{f}_{s+1}\|$ small, $s \leq R$
4. $\hat{\beta}_s^{PLS} \leftarrow \hat{W}_s(\hat{W}'_s X'_{-1}\hat{\Omega}X_{-1}\hat{W}_s)^{-1}\hat{W}'_s X'_{-1}\hat{\Omega}\hat{f}_0$ and
   $\hat{\eta}_s^{PLS} \leftarrow$ wt.mean$(\hat{f}_0$, wt $=\hat{\tau}) + X_{-1}\hat{\beta}_s^{PLS} \leftarrow$ wt.mean$(\hat{f}_0$, wt $= \hat{\tau}) + \sum_{k=1}^{s} \hat{q}_k\hat{t}_k$,
   where $\hat{W}_s = \{\hat{w}_1, \cdots, \hat{w}_s\}$

The generalized least squares (gls) estimates in lines 2(a):iv,v,vii can be handled routinely with software. For example users are allowed to specify a general form for the correlation matrix $R(\theta)$ that can further build the necessary block diagonal structure by the specified *unit* variable. Care has to be taken since $R^{-1}(\theta)$ cannot substitute for Ω. Since $A_i \neq \sigma^2 I$, an additional weight variable is needed using the diagonal elements of $D^{-1/2}$. Perhaps this can be best seen by noting since cov$(D^{-1/2}y) = R(\theta)$.

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, $k = 1, \cdots,\ R =$ column rank$(\hat{F})$. As seen from line 2(a)i above, the adjusted dependent variable residuals, $\hat{f}_{k-1}$ (in step $k-1$), are partially regressed on the regressor 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 variable 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 variable 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]. As MARX (1996) pointed out, other variants of this algorithm exist.

7.1 Connecting PLS to PC

An interesting connection exists between principal component and iteratively reweighted partial least squares in this GEE setting. Denote Q as the number of iterations until convergence. The GEE scoring solution can be expressed as

$$\hat{\eta}^{GEE} = \sum_{t=1}^{Q} \{\hat{\beta}_0 \mathbf{1}_N + X_{-1} \sum_{k=1}^{R} \hat{\lambda}_k^{-1} \hat{m}_k \hat{m}_k' X_{-1}' \hat{\Omega} \hat{y}^\star\}^{(t)}, \tag{31}$$

where R truncates the null components. Again $\hat{\beta}_0$ is the iterated weighted ($\hat{\tau}$) mean of the adjusted dependent variable, and $\hat{\Omega}$ is the updated GEE weight matrix. Based on these R components, it will be useful to define the matrices $\hat{\Lambda}_R$ and $\hat{M}_R$ corresponding to the diagonal matrix of nonnull eigenvalues and the associated matrix of eigenvectors, respectively, of the converged GEE $\hat{F}$ matrix, if they exist. As argued in Section 4, the PC estimator can use results of the converged GEE solution. Suppose that components are deleted in sequence from the sum associated with the $r = R - s$ smallest nonnull $\hat{\lambda}_k$. The principle component estimator can be expressed as

$$\tilde{\eta}_s^{PC} = \{\hat{\beta}_0 \mathbf{1}_N + X_{-1} \sum_{k=1}^{s} \hat{\lambda}_k^{-1} \hat{m}_k \hat{m}_k' X_{-1}' \hat{\Omega} \sum_{t=1}^{Q} \tilde{y}_s^{\star PC(t)}\}, \tag{32}$$

where $\hat{y}_s^{\star PC}$ is again the adjusted dependent variable but this time updated using only s terms in $\tilde{\eta}_s^{PC(t)}$. IRPLS estimation can also be less taxing if information is borrowed from the converged GEE solution. With s latent variables, one iterative scheme can be expressed as

$$\hat{\eta}_s^{PLS} = \{\hat{\beta}_0 \mathbf{1}_N + X_{-1} \hat{W}_s \sum_{k=1}^{s} \hat{\phi}_k^{-1} \hat{\xi}_k \hat{\xi}_k' \hat{W}_s' X_{-1}' \hat{\Omega} \sum_{t=1}^{Q} \hat{y}_s^{\star PLS(t)}\}, \tag{33}$$

where similarly to PC estimation, $\hat{y}_s^{\star PLS}$ is the adjusted dependent variable utilizing $s \leq R$ latent variables in $\hat{\eta}_s^{PLS}$ and again $\hat{W}_s = (\hat{w}_1, \cdots, \hat{w}_s)$. The $\hat{\xi}_k$ and the $\hat{\phi}_k$ correspond to the eigenvectors and eigenvalues of $\hat{W}_s' \hat{F} \hat{W}_s$, respectively.

8 Functional Regressors

As indicated in the previous section, regression problems can be high dimensional. Consider situations when modern technology generates regressors (near infrared spectra, brain waves, log-periodograms of spoken syllables, time series, etc.). Such information comes in the form of hundreds or thousands of discrete digitizations of some signal, often resulting in $p \gg N$. Moreover there exists some ordered structure, e.g. along wavelength. An application could be measuring several units of a particular food product over time, where at each time a near infrared spectra and a response (say % constituent) is collected. In such a setting, we have high dimensional functional regressors and responses within units that are correlated (perhaps AR(1)). Ramsay and Silverman (1997) provided an excellent overview of applications involving functional data. James and Hastie (1999) provided an interesting subject-specific mixed modeling approach using principal component techniques for sparse functional data.

Consider rewriting (1) as

$$\mu = h(\beta_0 + X_{-1}\beta) = h(\eta). \tag{34}$$

This problem is highly ill-conditioned, and the only hope to get a sensible result is by constraining β in some way. MARX and EILERS (1999) proposed a (GLM) P-spline modeling strategy, as a competitor to (IR)PLS, that forces β to be smooth. The dimension of the *signal* coefficient vector is reduced initially by projecting it onto a B-basis, B (of smooth functions): $\beta_{p\times 1} = B_{p\times q}\delta_{q\times 1}$, where $q < \min(N, p)$. Notice that this approach takes advantage of the spatial or temporal information along the signal and has an attractive linear nature of β, where δ is a relatively low dimensional vector of B-spline coefficients.

P-splines take one step further: use a moderate number of equally spaced B-spline knots (say 10 to 40) and further increase smoothness by imposing a difference penalty on δ (as in Section 6.1). Notice that (34) can be rewritten as

$$\mu = h(\beta_0 + X_{-1}B\delta) = h(\beta_0 + U\delta),$$

where we can define a new full rank regression matrix $U_{N\times q} = X_{-1}B$. Now re-express (24) as

$$\tilde{\delta}^D = \mathcal{Z}_\delta\{s(\delta,\theta) - \kappa P^{d\prime}P^d\delta\},$$

where

$$s(\delta,\theta) = U'\Omega(\delta,\theta)D^{-1}(\delta)(y - h(U\delta)), \tag{35}$$

and $\kappa \geq 0$. The information matrix is $F_U = U'\Omega U$ (with $M_U'F_UM_U = \Lambda_U$), and (25) can be implemented to estimate $\tilde{\delta}^D$. The penalized solutions are given upon convergence as $\tilde{\beta} = B\tilde{\delta}_\kappa$. Suggestions for optimization of κ are given in Marx and Eilers using cross-validation and information criterion. Similar arguments to those presented with (26) can link this setting to GFPC estimation.

Simple data management tricks can make fitting such GEE models trivial in existing (GLM) GEE software and macros that allow a variety of correlation structures, by the *unit* variable. Consider constructing a $p \times q$ B-spline matrix using a modest number of equally spaced knots along the indexing domain (e.g. frequency) of the signal. Thus U is accessible. Instead of passing the y response and the signal matrix X_{-1} into the GEE fitting algorithm, use the augmented matrix $U_{\text{aug}} = \text{rbind}(U_{+1}, (0, \kappa P^d))$ and the augmented adjusted dependent variable $y^\star_{\text{aug}} = \text{rbind}(y^\star, 0_{q-d})$, where U_{+1} includes the (unpenalized) intercept. Thus the software can automatically provide a penalized estimate of δ_κ, and $\tilde{\eta} = U_{+1}\tilde{\delta}_\kappa$. Notice also that $H_{\tilde{\beta}} = BH_{\tilde{\delta}}B'$.

9 Summary and Discussion

I have addressed the issue of ill-conditioned GEEs, which to my knowledge has been lacking in the literature. In doing so, I have pointed out some of the consequences of such a setting while transplanting a variety of biased estimation techniques into situations with correlated responses. To simplify the presentation, I attempted to link and unify these estimators as much as possible, borrowing work of others as mentioned in Section 1. Of course there are trade-offs in taking such a broad approach; naturally a disadvantage is that all of the many specific issues cannot be addressed here. Much more research is needed

to investigate additional analytic properties of the proposed estimators, as well as the most effective techniques for choosing the *meta* or regularization parameters. Future research could investigate comparisons of these estimators under a variety of regressor collinearity, response correlation, link and random component experimental settings. Certainly other GFPC weighting schemes exist, as well as other penalty choices for the quasi-likelihood. Biased estimation is only one tool and is not always the final answer. However as seen in the previous sections, such estimators can provide a practical solution to a complicated problem: degenerate or extreme ill-conditioning and general correlations structure among the responses.

Acknowledgments

I would like to thank Ludwig Fahrmeir, Lynn LaMotte and Rob Tibshirani for valuable discussions.

References

Albert, A. and Anderson, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **71**, 1-10.

Belsley, D.A. (1991). *Conditioning Analysis*. Wiley, New York.

Breiman, L. (1995). Better Subset Selection Using the Non-negative Garrote. *Technometrics*, **37**, 373-384.

Clarkson, D.B. and Jennrich, R.I. (1991). Computing Extended Maximum Likelihood Estimates for Linear Parameter Models. *Journal of the Royal Statistical Society, B*, **53**, 417-426.

Demmler, A. and Reinsch (1975). Oscillation Matrices with Spline Smoothing. *Numerische Mathematik*, **24**, 375-382.

Diggle, P.J., Liang, K., and Zeger, S.L. (1995). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

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

Eilers, P.H.C. and Marx, B.D. (1996). Flexible Smoothing Using B-Splines and Penalized Likelihood (with comments and rejoinder). *Statistical Science* **11**(2): 89-121.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

Frank, I.E. and Friedman, J.H. (1993). A Statistical Review of Some Chemometrics Regression Tools (with discussion). *Technometrics* **35**(2), 109-148.

Fu, W.J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, **7**, 397-416.

James, G. and Hastie, T. (1999). Principal Component Models for Sparse Functional Data. Technical Report, Department of Statistics, Stanford University.

Hocking, R.R., Speed, F.M., and Lynn, M.J. (1976). A Class of Biased Estimators in Linear Regression. *Technometrics* **18**(4), 425-437.

le Cessie, S. and van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. *Applied Statistics* **41**(1), 191-201.

Lee, W. and Birch, J.B. (1988). Fractional Principal Component Regression: A General Approach to Biased Estimators. *Communications in Statistics- Simulation*, **17**(3), 713-727.

Lesaffre, E. and Marx, B.D. (1993). Collinearity in Generalized Linear Regression. *Communications in Statistics: Theory and Methods* **22**(7): 1933-1952.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika* **73**: 13-22.

Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate Regression Analyses for Categorical Data. *Journal of the Royal Statistical Society, B* **54**: 3:40.

MacKinnon, M.J. and Puterman, M.L. (1989). Collinearity in Generalized Linear Models. *Communications in Statistics* **18**(9), 3463-3472.

Marx, B.D. (1992). A Continuum of Principal Component Generalized Linear Regressions. *Computational Statistics and Data Analysis* **13**, 385-393.

Marx, B.D. (1996). Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. *Technometrics*, **38**, 374-381.

Marx, B.D. and Eilers, P.H.C. (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. *Technometrics* **41**: 1-13.

Marx, B.D., Eilers, P.H.C. and Smith, E.P. (1992). Ridge Likelihood Estimation for Generalized Linear Regression. *In: Statistical Modelling*, North Holland Publishing Company (Elseviers), 227-237.

Marx, B.D. and Smith, E.P. (1990). Principal Component Estimators for Generalized Linear Regression. *Biometrika* **77**(1), 23-31.

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

McDonald, B.W. (1993). Estimating Logistic Regression Parameters for Bivariate Binary Data. *Journal of the Royal Statistical Society, B*, **55**, 391-397.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, A*, **135**: 370-384.

Nyquist, H. (1991). Restricted Estimation of Generalized Linear Models. *Applied Statistics* **40**(1), 133-141.

Ramsay, J. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag, New York.

Schaefer, R.L. (1986). Alternative Estimators in Logistic Regression when the Data are Collinear. *Journal of Statistical Computation and Simulation* **25**, 75-91.

Sclove, S.L. (1968). Improved Estimators for Coefficients in Linear Regression. *Journal of the American Statistical Association* **63**, 596-606.

Stein, C.M. (1960). Multiple Regression. In: *Contributions to Probability and Statistics*, Essays in Honor of Harold Hoteling, Ed. I. Olkin, pp. 424-43. Stanford University Press.

Stone, M. and Brooks, R.J. (1990). Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Component Regression. *Journal of the Royal Statistical Society*, B **52**(2), 237-269.

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

Whitakker, E.T. (1923). On a New Method of Graduation. In: *Proceedings of the Edinburgh Mathematical Society* **43**, 63-75.

Wedderburn, R.W.M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika* **61**, 439-447.

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, Academic Press, London.