Computational Statistics & Data Analysis 13 (1992) 385–393
North-Holland

A continuum of principal component generalized linear regressions

Brian D. Marx

Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803, USA

Abstract: The motivation of the paper is to present an option to maximum likelihood estimation when the information matrix is ill-conditioned. The impacts and diagnosis of ill-conditioned information are revisited. We combine ideas from Good and Smith [4] and Marx and Smith [10] to develop a class of principal component estimators, for generalized linear regression, defined by a scaling parameter. The additional parameter allows a spectrum of standardized explanatory variables which can yield interpolation between correlation and covariance matrices. We show that choice of the scaling parameter depends on the researcher's objectives for the model. A unit scaling parameter produces results of Marx and Smith [10]. If further restrictions of normal response data and the identity link function are imposed with an unit scaling parameter, we have traditional principal component multiple regression (Webster et al. [16]). We discuss the appropriateness of principal component regression. An illustrative example using Poisson response data and the log link function demonstrates the usefulness of the scaling parameter for a generalized linear regression with severely ill-conditioned information.

Keywords: Quasistandardization, Scaling parameter, Weighted multicollinearity.

1. Introduction

We focus on problems in the generalized linear model (GLM) (Nelder and Wedderburn, [13]) regression setting with all continuous explanatory variables. In particular, when variable deletion is not an option, we aim to alleviate many of the detriments associated with an ill-conditioned information matrix. The GLM's regression parameters are typically estimated via an iterative maximum likelihood process, but in the presence of ill-conditioned information, Marx and Smith [10] suggested an asymptotically biased principal component estimator. The proposed principal component estimator has an additional scaling parameter which can accommodate a spectrum of explanatory variable standardizations. The choice of the scaling parameter can be made on a variety of statistical criteria. Section 2 defines the notion of quasistandardized explanatory variables and Section 3 gives a brief overview of generalized linear regression. Section 4 explains the effects of a near-singular information matrix and is thus a motivation for principal component analysis. Section 5 develops the continuum of principal component estimators by incorporating quasistandardization. Section 6

0167-9473/92/$05.00 © 1992 – Elsevier Science Publishers B.V. All rights reserved

discusses some implications that a scaling parameter can have on inference. An illustrative example is provided to demonstrate the usefulness of the continuum of estimators for a Poisson regression.

2. Quasistandardization of explanatory variables

There is much controversy regarding centering and scaling in regression problems; we consider a general approach. Let $X^0 = (\boldsymbol{x}_1^0, \boldsymbol{x}_2^0, \ldots, \boldsymbol{x}_p^0)$ be a $N \times p$ matrix of continuous regression explanatory variables in their natural units. Consider centering and scaling X^0 similar to that presented in Good and Smith [4] and Marx [8]. Define

$$x_{\alpha ij} = q_j^{-\alpha}(n-1)^{-1/2}\left(x_{ij}^0 - \bar{x}_j^0\right), \qquad \text{where} \tag{1}$$

$$q_j^2 = (n-1)^{-1}\sum_{i=1}^{N}\left(x_{ij}^0 - \bar{x}_j^0\right)^2. \tag{2}$$

Let $X_\alpha = \{x_{\alpha ij}\}$. The parameter α allows a spectrum of scaling. Notice for $\alpha = 1$ and $\alpha = 0$, we have $X_\alpha' X_\alpha$ simplifying to the sample correlation and covariance matrices respectively. Scaling parameter values between zero and one lead to an interpolation between correlation and covariance matrices. Good and Smith [4] point out that in practice it may seem unnatural to use parameter values outside the unit interval.

3. Background and notation for generalized linear regression

Augment the matrix X_α with a constant vector of ones. Denote $\boldsymbol{x}_{\alpha i}'$ as a row vector of X_α. Let $\boldsymbol{Y}$ be a $N \times 1$ random response vector. Each entry in Y_i follows the same distribution contained in the exponential family having canonical link function θ_i and dispersion parameter ϕ_i. Given the set of p continuous explanatory variables, generalized linear regression utilizes the relationship,

$$g(\mu_i)_\alpha = \boldsymbol{x}_{\alpha i}'\boldsymbol{\beta}_\alpha = \eta_{\alpha i}, \tag{3}$$

satisfying: $\mu_i = E(Y_i)$; g is a monotone, twice differentiable link function with $g^{-1} = h$; $\boldsymbol{x}_{\alpha i}'$ is a quasistandardized $(p+1) \times 1$ row vector of continuous regressor variables, including the constant; $\boldsymbol{\beta}_\alpha$ is the unknown parameter vector; the estimation of $\boldsymbol{\beta}_\alpha$ does not depend on having an estimate of ϕ; and ϕ is constant for all Y_i. Standard theory of the generalized linear model can be found in McCullagh and Nelder [11], Dobson [3], and Aitkin et al. [1].

For given α, denote the information matrix $\Phi_\alpha = X_\alpha' K_\alpha^{-1} X_\alpha$, where

$$K_\alpha^{-1} = \operatorname{diag}\{k_{\alpha ii}^{-1}\} = \operatorname{diag}\left\{\left[h'(\eta_{\alpha i})\right]^2 / \operatorname{Var}(Y_i)\right\}. \tag{4}$$

The method of scoring maximum likelihood (ML) can be expressed as

$$\hat{\boldsymbol{\beta}}^{ml*}_{-\alpha,t} = \left[\hat{\Phi}^{-1}X'\hat{K}^{-1}\boldsymbol{w}\right]_{\alpha,t-1}, \tag{5}$$

where $w_{\alpha,i} = [\hat{\eta}_i + (y_i - \hat{\mu}_i)(\partial\eta_i/\partial\mu_i)]_\alpha$. Note that in the most general setting, the estimate of K_α^{-1} and $\boldsymbol{w}_\alpha$ must be updated at each iteration step until convergence of the parameter estimate since they are a function of the iterated $\boldsymbol{\eta}_{\alpha,t-1}$. We often refer to $\boldsymbol{w}_\alpha$ as the vector of working variables. Denote $\hat{\boldsymbol{\beta}}^{ml}$ as the corresponding uncentered and unscaled parameter estimates of (5).

4. Ill-conditioned information

As presented in Belsley et al. [2] with normal response data and the identity link function, problems of multicollinearity among the columns of the X^0 matrix usually leads to an investigation of the correlation matrix for explanatory variables. Various options for model building are well documented for the standard multiple regression setting (Stein [15]; Hoerl and Kennard [5]; Webster et al. [16]). Schaefer [14] and Mackinnon and Puterman [7] discussed GLM multicollinearity. Marx and Smith [10] emphasized that more detrimental to the generalized linear model, than multicollinearity among the columns of X^0, is an information matrix which is nearly deficient in rank. Such ill-conditioning of information can result in highly unstable iteration steps in the method of scoring, inflated sum of maximum likelihood coefficient, poor prediction in certain regions, and low power for certain tests. Let $\Phi = X^{0\prime}K^{-1}X^0$ be the information matrix of the uncentered and unscaled explanatory variables. We consider Φ to be ill-conditioned if it has a small singular value relative to the largest singular value. Denote $\gamma_0 \geq \gamma_1 \geq \cdots \geq \gamma_p$ as the singular values of information. Define the information condition indices, $\kappa_j = \gamma_0/\gamma_j$ for $j = 0, \cdots, p$. Traditionally a $\kappa_j > 30$ signifies ill-conditioning. Notice also that ill-conditioning of Φ_α is the result of nearly linear dependent columns of $K_\alpha^{-1/2}X_\alpha$ and is only equivalent to multicollinearity among the columns of X_α for $K_\alpha^{-1/2} \cong cI$, $c > 0$ constant.

5. Continuum of principal component estimation

Marx [9] developed a continuum of principal component estimators for the classical model having a normal error distribution and the identity link function. For a general exponential family response variable, denote $\hat{\boldsymbol{\beta}}^{ml*}_{\alpha=1}$ as the converged ML parameter estimates using the explanatory variables standardized with $\alpha = 1$, if it exists. For generalized linear regression, the one-step principal component estimator developed in Schaefer [14] and Marx and Smith [10] can be expressed as

$$\hat{\boldsymbol{\beta}}^{pc*}_{\alpha=1} = \hat{\Phi}^{-1}_{\mathscr{R}}\hat{\Phi}\hat{\boldsymbol{\beta}}^{ml*}_{\alpha=1} = \left(I - \sum_{j\in\mathscr{D}} \hat{\boldsymbol{m}}_j\hat{\boldsymbol{m}}_j'\right)\hat{\boldsymbol{\beta}}^{ml*}_{\alpha=1}, \tag{6}$$

where I is the identity matrix, $\mathscr{D}$ and $\mathscr{R}$ are the sets of deleted and retained components, respectively, $\hat{\Phi}_{\mathscr{R}}^{-1} = \sum_{j \in \mathscr{R}} \hat{\lambda}_j^{-1} \hat{\boldsymbol{m}}_j \hat{\boldsymbol{m}}_j'$, $\hat{\lambda}_j$ and $\hat{\boldsymbol{m}}_j$ are the eigenvalues and eigenvectors, respectively, of the maximum likelihood information matrix $\hat{\Phi}_{\alpha=1} = (X'\hat{K}^{-1}X)_{\alpha=1}$ using $\hat{\boldsymbol{\beta}}_{\alpha=1}^{ml*}$.

Define $M_\alpha' \Phi_\alpha M_\alpha = \text{diag}\{\lambda_{\alpha j}\} = \Lambda_\alpha$ as the spectral decomposition of information with scaling parameter α. Consider rewriting (3) as

$$g(\mu_i)_\alpha = (\boldsymbol{x}_{\alpha i}' M_\alpha)(M_\alpha' \boldsymbol{\beta}_\alpha) = \boldsymbol{z}_{\alpha i}' \boldsymbol{\gamma}_\alpha. \tag{7}$$

A generalization of (6) can be useful for (7) if Φ_α is ill-conditioned. Consider

$$\hat{\boldsymbol{\beta}}_\alpha^{pc*} = \left(I - \sum_{j \in \mathscr{D}} \hat{\boldsymbol{m}}_{\alpha j} \hat{\boldsymbol{m}}_{\alpha j}' \right) \hat{\boldsymbol{\beta}}_\alpha^{ml*}, \tag{8}$$

where $\hat{\boldsymbol{\beta}}_\alpha^{ml*}$ is the maximum likelihood estimator using the quasistandardized data with parameter α. Standard methods to uncenter and unscale $\hat{\boldsymbol{\beta}}_\alpha^{pc*}$ to $\hat{\boldsymbol{\beta}}_\alpha^{pc}$ can be taken. Note that if $\mathscr{D}$ is the null set, then $\hat{\boldsymbol{\beta}}_\alpha^{pc} = \hat{\boldsymbol{\beta}}^{ml}$, for all α. The dilemma of which components to delete or retain is still present and a decision should be based on standard diagnostics and regression criteria. The reader is referred to Marx and Smith [10] for an overview of the development and properties of principal component generalized linear regression. In the above paper, both iterative and one-step principal component estimators are presented. The continuum of one-step estimators in (8) approximates the possible continuum of iterative estimates. However, even with severe ill-conditioned information, the one-step estimator behaves similarly to the iterative estimator.

We further include the following result which can be applied to the continuum of principal component estimators.

Proposition 1. *The unique maximum likelihood estimator for the generalized linear regression parameters, when it exists, provides the minimum deviance.*

Proof. Let $\hat{\boldsymbol{\beta}}^0$ be any estimator, other than the unique maximum likelihood estimator $\hat{\boldsymbol{\beta}}^{ml}$, of $\boldsymbol{\beta}$. The deviance using $\hat{\boldsymbol{\beta}}^0$ is $D(\hat{\boldsymbol{\beta}}^0) = 2[l(\hat{\boldsymbol{\beta}}^{\max}) - l(\hat{\boldsymbol{\beta}}^0)]$, where l denotes the loglikelihood function and $\hat{\boldsymbol{\beta}}^{\max}$ represents the saturated model. Hence $\Delta D = D(\hat{\boldsymbol{\beta}}^{ml}) - D(\hat{\boldsymbol{\beta}}^0) = 2[l(\hat{\boldsymbol{\beta}}^0) - l(\hat{\boldsymbol{\beta}}^{ml})] < 0$, by definition of ML estimation. □

6. Surgical principal component estimation

Along with the continuum of principal component estimators, there exists a corresponding continuum of eigenstructures for the quasistandardized information matrix. In certain experimental settings, there may be several windows of scaling α such that the deviance and other regression criteria are acceptable to the researcher. In such cases, the researcher may be tempted to choose among various α to yield coefficients of desired sign or magnitude. Further, perhaps the researcher has a preference toward dependence (independence) of scale and

chooses α in the proximity of zero (one). Moreover, if for example $\alpha = 1$ seems to work best, then it may be reasonable to experiment with a scaling parameter value near but greater than unity. One could argue that we are, in one way or another, optimizing over the range of α; perhaps we should lose a degree of freedom for various inferences.

The implications of this window of estimators can be controversial. Consider, for an extreme example, that we wish to choose an α such that when the component associated with the smallest eigenvalue is deleted, the corresponding standardized eigenvector has the kth entry near, but less than unity. By (6), the kth entry of $\hat{\boldsymbol{\beta}}_{\alpha}^{pc*}$ will be shrunk nearly to zero, whereas the remaining p entries of $\hat{\boldsymbol{\beta}}_{\alpha}^{pc*}$ will be relatively close to the maximum likelihood estimation. We refer to this situation as surgical shrinkage. Future work will be devoted to data configurations that yield an ill-conditioned information matrix with a surgical eigenstructure.

7. Illustrative example

Myers ([12], Example 7.10) presents data which consists of $N = 44$ observations on mines in the coal fields of the Appalachian region of western Virginia. Effects of $p = 4$ continuous explanatory variables were analyzed for roles contributing to the number of injuries or fractures that occur in the upper seams of these mines. We consider, as Myers did, a generalized linear regression assuming a Poisson response and the natural log link function. The square root and identity links were also observed and dismissed based on deviance. The regression model of interest is

$$\log \lambda_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}, \tag{9}$$

where λ_i is the expected number of upper seam injuries or fractures in the ith coal mine area, X_{1i} is the corresponding inner burden thickness (the shortest distance, in feet, between seam floor and lower seam), X_{2i} is the percent extraction of the lower previously mined seam, X_{3i} is the lower seam height (feet), and X_{4i} is the length of time that the mine has been open (years).

The maximum likelihood model for (9) is adequate for prediction based on analysis of deviance and will yield parameter estimates for all explanatory variables of interest. However, the condition indices of the information matrix are 1.000, 3.505, 11.665, 32.998 and 2093.331, respectively, indicating severe ill-conditioning and justifying an optional estimation technique. Using (8), one-step principal component estimates (for various acceptable α) are provided along with maximum likelihood estimates in Table 1. All methods converged in 13 iterations. Standard approaches have been taken to uncenter and unscale the parameter estimates. Table 2 displays the corresponding uncentered and unscaled asymptotic standard errors.

Figure 1 demonstrates that acceptable principal component models, based on deviance, depends on the choice of scaling parameter and the number of

Table 1
Maximum likelihood, ML, and principal component, PC, parameter estimation for coal mine example. Link, log; error, Poisson; $N = 44$, $p + 1 = 5$.

	ML	$\alpha = 0$		$\alpha = 0.10$		$\alpha = 0.25$	
		PC(−1)	PC(−2)	PC(−1)	PC(−2)	PC(−1)	PC(−2)
Dev.	37.8560	40.4904	49.0182	37.8765	43.5432	42.2857	42.2858
$\hat{\beta}_0$	−3.5931	-4.8236	−5.5316	−3.6967	−4.6770	−2.3025	−2.3368
$\hat{\beta}_1$	−0.0014	−0.0017	−0.0022	−0.0014	−0.0019	−0.0013	−0.0013
$\hat{\beta}_2$	0.0623	0.0773	0.0838	0.0636	0.0741	0.0458	0.0462
$\hat{\beta}_3$	−0.0021	−0.0017	−0.0037	−0.0020	−0.0034	−0.0033	−0.0033
$\hat{\beta}_4$	−0.0308	−0.0438	0.0010	−0.0323	−0.0014	−0.0050	−0.0045

Table 2
Estimated asymptotic standard errors for coal mine example.

	ML	$\alpha = 0$		$\alpha = 0.10$		$\alpha = 0.25$	
		PC(−1)	PC(−2)	PC(−1)	PC(−2)	PC(−1)	PC(−2)
SE $(\hat{\beta}_0)$	1.0257	0.7337	0.6976	0.7293	0.6051	0.7910	0.4165
SE $(\hat{\beta}_1)$	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008
SE $(\hat{\beta}_2)$	0.0123	0.0086	0.0084	0.0084	0.0072	0.0089	0.0044
SE $(\hat{\beta}_3)$	0.0051	0.0051	0.0050	0.0051	0.0050	0.0050	0.0050
SE $(\hat{\beta}_4)$	0.0163	0.0144	0.0003	0.0128	0.0003	0.0097	0.0005

deleted components. The principal component model minus one dimension (PC(−1)) has acceptable deviance for scaling parameter in the neighborhood of $0 \leq \alpha \leq 0.6$, whereas the minus two dimension model (PC(−2)) has a narrower acceptable range, $0 \leq \alpha \leq 0.4$. Figure 1 also illustrates that if we are willing to

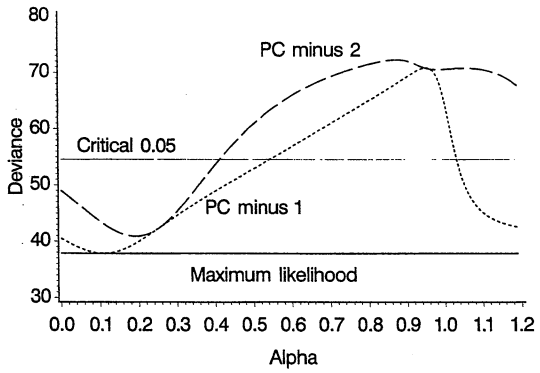

Fig. 1. Deviance for coal mine example.

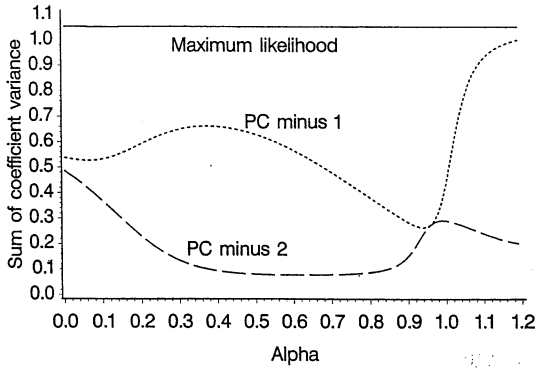

Fig. 2. Sum of coefficient variance for coal mine example.

extrapolate, then the PC(−1) model again becomes acceptable, based on deviance, for α values near but greater than unity. Notice that, despite the ill-conditioned ML model, the traditional PC models, with $\alpha = 1$, would be deemed unacceptable based on deviance. The maximum likelihood model will always achieve a lower deviance than the principal component models, but for α values near 0.1, the PC(−1) model achieves a deviance nearly equivalent to the ML model. Recall Proposition 1 in Section 5. Further, we are also guaranteed ordering in deviance among the iterative PC(−1) and PC(−2) models. However, notice in Figure 1 that the deviance for PC(−1) is greater than the deviance for PC(−2) at α values near 0.96. This anomaly is due to the fact the one-step estimators from (8) only approximate the iterative ones and does not reflect an error in programming.

Moreover, Figure 2 clearly shows the reduction in sum of asymptotic coefficient variance in both the PC(−1) and PC(−2) models, for all α, when compared to the ML model. For any given α, there is a guaranteed consistent decreasing order for the asymptotic sum of coefficient variance in ML, PC(−1), and PC(−2) techniques, respectively. Since the true coefficients are unknown, the author admits that we cannot accurately evaluate the asymptotic coefficient bias associated with the principal component models. Future research will be devoted to this area. However, we do see from this example that the optimal choice of α depends on the researcher's criterion for model selection.

The PC(−1) model with $\alpha = 0.1$ appears to be one of many reasonable candidates since it nearly achieves the deviance of the ML model, yet has approximately half of the sum of coefficient variance of the ML model, and is less biased than the corresponding PC(−2) model. Notice that the model suggests that the mean number of injuries or fractures decreases with increased inner burden thickness, lower seam height or number of years the mine has been opened, whereas the mean injuries increase with increased percent extraction. This choice also yields coefficients very similar, in sign and magnitude (with decreased standard errors and deviance), to Myers' ([12], Page 338) ML model choice containing X_1, X_2, and X_4.

8. Discussion

One consequence of using principal component estimation in the framework of the GLM is the asymptotically biased nature of the parameter estimates. However, we attain asymptotic decreases in the sum of variances of parameter estimates and prediction variances. If the error distribution is normal, the link function is the identity with unity scaling parameter, then the above results are completely consistent with Webster et al. [16]. There is some controversy regarding principal component estimation. Some researcher's argue that in experimental settings where a researcher is not constrained to a theoretical model, then perhaps variable selection is adequate.

Given an ill-conditioned information matrix and variable deletion is not an option, the continuum of principal component techniques can be useful for parameter estimation as an alternative to maximum likelihood. One contention is that the quasistandardized principal component techniques yield linear combinations of explanatory variables in an unintuitive metric. In fact, Jolliffe ([6] Chapter 8) suggested that the information matrix should be proportional to the correlation matrix. However, this unintuitive metric is akin to transformations. In the end, we backtransform to the original coordinates and, in some cases, have a more stable regression than maximum likelihood. Note that if all the explanatory variables are in the same units, then one could argue for a covariance, correlation or quasistandardized approach. The author suggests that the quasistandardized principal component procedure should be used for exploratory purposes.

References

[1] Aitkin, M., D. Anderson, B. Francis and J. Hinde, *Statistical Modelling in GLIM* (Clarendon Press, Oxford, 1989).

[2] Belsley, D.A., E. Kuh and R.E. Welsch, *Regression Diagnostics: Influential Data and Sources of Collinearity* (Wiley, New York, 1980).

[3] Dobson, A.J., *An Introduction to Statistical Modelling* (Chapman and Hall, London, 1983).

[4] Good, I.J. and E.P. Smith, A continuum of principal component methods, *Journal of Statistical Computation and Simulation*, **22** (1985) 136–142.

[5] Hoerl, A.E. and R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12** (1970) 55–68.

[6] Jolliffe, I.T., *Principal Component Analysis* (Springer-Verlag, New York, 1986).

[7] Mackinnon, M.J. and M.L. Puterman, Collinearity in generalized linear models, *Communications in Statistics – Theory and Methods*, **18** (9) (1989) 3463–3472.

[8] Marx, B.D., A correction for C240, Good and Smith, JSCS 22 136–142, *Journal of Statistical Computation and Simulation*, **36** (2+3) (1990) 193.

[9] Marx, B.D., A continuum of principal component regression methods, *Journal of Statistical Computation and Simulation* **37** (3+4) (1990) 234–236.

[10] Marx, B.D. and E.P. Smith, Principal component estimators for generalized linear regression, *Biometrika*, **77** (1) (1990) 23–31.

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

[12] Myers, R.H., *Classical and Modern Regression with Applications*, 2nd ed. (Duxbury Press, Boston, 1990).
[13] Nelder, J.A. and R.W.M. Wedderburn, Generalized linear models, *Journal of the Royal Statistical Society* A, **135** (3) (1972) 370–384.
[14] Schaefer, R.L., Alternative estimators in logistic regression when the data are collinear, *Journal of Statistical Computation and Simulation*, **25** (1986) 75–91.
[15] Stein, C.M., Multiple regression, in: I. Olkin (Ed.), *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling* (Stanford University Press, 1960) 424–443.
[16] Webster, J.T., R.F. Gunst and R.L. Mason, Latent root regression analysis, *Technometrics,* **16** (1974) 513–522.