A modified-likelihood-type shrinkage estimator for the sparse-response Poisson and Binomial GLM

By
Bei Dong
Lu Zhang

Supervised by
Moudud Alam

Submitted to the school of Economics and Social Sciences
in partial fulfillment of the requirements
for the degree of
Master of Science
June 2009
A modified-likelihood-type shrinkage estimator for the sparse-response Poisson and Binomial GLM

Bei Dong*, Lu Zhang*
Supervisor: Moudud Alam

June, 2009

Abstract

We propose the shrinkage estimation based on modified likelihood, and put forward the algorithms especially for the Binomial and Poisson generalized linear models. Our simulation studies suggest that this method enjoys some of the favorable properties such as decreasing the mean square error and increasing the stability of the estimation. The application of this shrinkage estimator is illustrated by using several real data sets and all of the performance showing that shrinkage estimation is a valuable tool when dealing with sparse response.

Keywords: shrinkage estimation, generalized linear models, sparse response

*Author e-mails: Bei Dong: h08beido@du.se, Lu Zhang: h08luzha@du.se
Properties of the shrinkage estimator for sparse-response Poisson and Binomial GLM

Contents

1 Introduction 3

2 Estimators of Modified Likelihood 3

3 Algorithms 4
   3.1 Modified-IWLS 4
   3.2 Newton-Raphson 4
   3.3 Comparison of The Two Algorithms 4
   3.4 Selection of Shrinkage Estimator \( \lambda \) 5

4 Simulation Study 6
   4.1 Nonsparse Binomial Case 6
   4.2 Sparse Binomial Case 7

5 Application with Real Data 9
   5.1 Survival of Heart Transplant 9
   5.2 Survival of Snails 10
   5.3 Wave Damage to Cargo Ships 11

6 Concluding Discussion 12

Appendix 13
   Poisson Case 13
   R-code of Modified-IWLS Algorithm 14
   R-code of Newton-Raphson Algorithm for Binomial Case 15
   R-code of Newton-Raphson Algorithm for Poisson Case 16

References 18

Bei Dong, Lu Zhang
1 Introduction

Shrinkage estimation was initially developed by Charles Stein in 1956 (Stein 1956). Since then considerable research has been done focusing on shrinkage estimation of location parameters (An, Nkurunziza, Fung, Krewski & Luginaah 2009). The shrinkage estimation has been widely applied to increase robustness and accuracy of estimation.

In the research of rank evaluation of Swedish hospitals (Andersson, Carling & Mattson 1998), it is shown that ranking these hospitals by mortality rate may result in smaller units having the highest and lowest ranks, because they are more susceptible to sampling variation. Shrinkage estimation can make some adjustment such that the random variation plays less role in the estimation (Andersson et al. 1998).

Up till now, lots of methods of shrinkage estimation have been posed, such as James-Stein-type estimation (James&Stein 1962), Lasso (Tibshirani 1996), Penalized likelihood (Hastie&Tibshirani 1986). However, most of the researches focused on the general linear model, while few researches have dealt with the generalized linear case. Therefore, it is important to explore the performance of shrinkage estimation in generalized linear models.

In Binomial and Poisson generalized linear models, sparse response situation often appears in analysis of rare events if the mean parameter is very close to (or numerically indistinguishable from) 0 for Binomial and Poisson cases and 1 (or n when Binomial numerator n > 1) for binomial case. It makes the estimation procedure numerically unstable. Though the problem is easily detected by big estimate of the standard error, there is no straightforward solution to this problem. In order to overcome the above problems, this paper examines the applicability of a shrinkage estimation method in the Binomial and Poisson generalized linear model by using the following type of modified likelihood,

\[
L = \exp \left[ \sum_{i=1}^{n} \left\{ \frac{y_i h(\eta_i) - b(h(\eta_i))}{a(\phi)} - c(y_i, \varphi) \right\} - \sum_{k=1}^{q} \lambda \beta_k^2 \right]
\]

(1.1)

where \( \eta_i = \sum_{j=1}^{p} x_{ij} \beta_j + \sum_{k=1}^{q} \tilde{x}_{ik} \tilde{\beta}_k \) with \( \beta_j's \) being non-shrinkage estimators and \( \tilde{\beta}_k's \) being shrinkage estimators, \( x_{ij} \) and \( \tilde{x}_{ik} \) are covariates, \( y_i's \) are true responses, \( h(\cdot) \) is the function to transform the linear predictor, and \( \lambda \) denotes the shrinkage parameter (or penalty factor). The above modified likelihood is the ordinary likelihood with an extra penalty. It is just proportional to the Bayesian posterior of a GLM with \( N(0, \frac{1}{\lambda}) \) prior for all the regression parameters (Dey, Ghosh & Mallick 2000).

For a Gaussian model it provides the Ridge estimator (Fredman, Hasti & Tibsirani 2001). The shrinkage estimator should be derived by maximizing this modified likelihood. By adding a penalty into the likelihood, we expect to sacrifice a little bit of bias in order to reduce the variance of the predicted values and then improve the overall prediction accuracy.

The paper aims at exploring the advantages and disadvantages of estimating sparse Binomial and Poisson model using the shrinkage estimator of the above equation.

The frame of this paper is as follows. In section 2, we simply introduce the modified likelihood and derive the analytical form of the estimator. We summarize and modify the algorithms for solving the shrinkage estimator in section 3. Also the selection of shrinkage parameter \( \lambda \) is discussed. Then the simulation study is done in section 4, both with non-sparse simulated data and sparse simulated data. In section 5 we apply the shrinkage estimation in several real data sets. The conclusions are included in section 6 and we also pose some discussions in that section.

2 Estimators of Modified Likelihood

The modified log-likelihood is expressed as below,

\[
\log L = \sum_{i=1}^{n} \left\{ \frac{y_i h(\eta_i) - b(h(\eta_i))}{a(\phi)} - c(y_i, \varphi) \right\} - \sum_{k=1}^{q} \lambda \beta_k^2
\]

(2.1)

In this paper we use the canonical link of Binomial and Poisson generalized linear models, so the function \( h(\cdot) \) is identity. With the scale parameter \( a(\varphi) = 1 \), which is natural for Binomial and Poisson models, equation (2.1) for a single observation is expressed as,

\[
l = y \eta - b(\eta) - c(y, \varphi) - \sum_{k=1}^{q} \lambda \beta_k^2
\]

(2.2)

where \( l \) is the log-likelihood function (Olsson 2002). We denote the canonical parameter \( \theta = \eta = \sum_{j=1}^{p} x_j \beta_j + \sum_{k=1}^{q} \tilde{x}_k \tilde{\beta}_k \). Differentiation of \( l \) with respect to one element of the parameter, using the chain rule, yields

\[
\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \eta} \frac{\partial \eta}{\partial \mu} \frac{\partial \mu}{\partial \beta} - 2 \lambda \beta_1 \in (\bar{\beta}_1, \ldots, \bar{\beta}_q)
\]

(2.3)

where \( \beta \) is a scalar and \( 1_{\beta(\bar{\beta}_1, \ldots, \bar{\beta}_q)} \) is an indicator.

Since for generalized linear models with exponential distribution, \( b'(\theta) = \mu \), and \( b''(\theta) = V \), where \( V \) is the variance function. Thus, \( \frac{\partial \eta}{\partial \mu} = V \). Also from the expression of the linear predictor, \( \frac{\partial \eta}{\partial \beta} = x \) is derived. Under the chain rule, differentiation of \( l \) with respect to the nonshrinkage...
subset yields,
\[ \frac{\partial l}{\partial \beta_j} = (y - \mu) \frac{1}{V} \frac{d\mu}{d\eta} x_{ij} \] (2.4)

Summing over the observations, the likelihood equation for one non-shrinkage estimator \( \hat{\beta}_j \) is given by
\[ \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij} = 0 \] (2.5)

Also differentiation of \( l \) with respect to the shrinkage subset can be derived the same way as below,
\[ \sum_{i=1}^{n} [(y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} \tilde{x}_{ik}] - 2\lambda \hat{\beta}_k = 0 \] (2.6)

Given \( \mu_i \)'s are functions of parameters \( \beta_j \) and \( \hat{\beta}_k \), we can solve (2.5) and (2.6) with respect to \( \beta_j \) and \( \hat{\beta}_k \) to get the non-shrinkage estimators and shrinkage estimators of the modified likelihood. The details of the algorithm to get the parameters are introduced in the next section.

3 Algorithms

From the modified log-likelihood above in section 2, to attain estimates directly seems impossible. However, we can modify the iteratively reweighted least squares approach and use Newton-Raphson fitting algorithm to get the numerical solutions. In this section, the two algorithms are stated and then we compare them by implementing them in the R functions using simple simulated data.

3.1 Modified-IWLS

Referring to the iteratively reweighted least squares approach (McCullagh & Nelder 1989), we modify its fourth step and implement it to solve (2.5) and (2.6) numerically. The modified algorithm is as follows,

1. linearize the link function \( \theta = g(\cdot) \) by using the first order Taylor series approximation \( g(y) \approx g(\mu) + (y - u)g'(\mu) = z \).

2. let \( \hat{\eta}_0 \) be the current estimate of the linear predictor, and let \( \hat{\mu}_0 \) be the corresponding fitted value derived from the link function \( \eta = g(\mu) \). From the adjusted dependent variate \( z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0) \cdot \frac{d\mu}{d\mu} \) where the derivative of the link is evaluated at \( \hat{\mu}_0 \) and \( z_0, \hat{\eta}_0 \) and \( \hat{\mu}_0 \) are all \( n \times 1 \) vectors.

3. Define the weight matrix \( W \) from \( W_0^{-1} = (\frac{d\mu}{d\eta})^2 V \), where \( V \) is the vector of variance function \( V \).

4. Then (2.5) and (2.6) can be expressed as matrix forms,
\[ X^T W_0 (z_0 - \hat{\eta}_0) = 0 \] (3.1)
\[ \tilde{X}^T W_0 (z_0 - \hat{\eta}_0 - 2\lambda q \tilde{\beta}_0) = 0 \] (3.2)

where \( \hat{\eta}_0 = X_0 \beta_0 + \tilde{X} \tilde{\beta}_0 \), \( X \) is a \( n \times p \) design matrix for the nonshrinkage estimator \( \beta(p \times 1) \) and \( \tilde{X} \) is a \( n \times q \) design matrix for the shrinkage estimator \( \tilde{\beta}(q \times 1) \). Hence, \( \tilde{\beta}_1 = (\tilde{X}^T W_0 \tilde{X} - 2\lambda I_{q \times q})^{-1} \tilde{X}^T W_0 (z_0 - \tilde{X} \tilde{\beta}_0) \) and \( \tilde{\beta}_1 = (\tilde{X}^T W_0 \tilde{X} + 2\lambda I_{q \times q})^{-1} \tilde{X}^T W_0 (z_0 - \tilde{X} \tilde{\beta}_0) \)

5. Repeat steps 1-4 until the changes are sufficiently small to attain the numerical solutions of \( \beta \) and \( \tilde{\beta} \).

3.2 Newton-Raphson

Newton-Raphson (Ortega & Rheinboldt 2000) is an iterative method to get the numerical solution of the estimator by calculating Hessian matrix\(^1\). By taking the first derivative of the modified likelihood function, the score functions of nonshrinkage estimator and shrinkage estimator are as bellow,
\[ S(\beta) = X^T y - X^T b'(\eta) \] (3.3)
\[ S(\tilde{\beta}) = \tilde{X}^T y - \tilde{X}^T b'(\hat{\eta}) - 2\lambda \tilde{\beta} \] (3.4)

and the Hessian matrix is given as,
\[ H = \begin{pmatrix} -X^T diag(b''(\eta)) X & -X^T diag(b''(\eta)) \tilde{X} \\ -\tilde{X}^T diag(b''(\eta)) X & -\tilde{X}^T diag(b''(\eta)) \tilde{X} - 2\lambda I_{q \times q} \end{pmatrix} \] (3.5)

where \( diag(b''(\eta)) \) is a diagonal matrix, the diagonal elements are the value of the second derivative of \( b(\eta) \) with respect to \( \eta \). Hence by Newton-Raphson algorithm, we get,
\[ \begin{pmatrix} \beta^{k+1} \\ \tilde{\beta}^{k+1} \end{pmatrix} = \begin{pmatrix} \beta^k \\ \tilde{\beta}^k \end{pmatrix} - H^{-1} \begin{pmatrix} S(\beta^k) \\ S(\tilde{\beta}^k) \end{pmatrix} \] (3.6)

By repeating this procedure until the changes are sufficiently small, the numerical solutions of \( \beta \) and \( \tilde{\beta} \) can be attained.

\(^1\)In mathematics, the Hessian matrix is the square matrix of second-order partial derivatives of a function
3.3 Comparison of The Two Algorithms

For Modified-IWLS algorithm, the convergence procedure is a Fisher scoring approach. Since when adopting the canonical link, the Hessian matrix in Newton-Raphson algorithm contains no random term which essentially the same as Modified-IWLS algorithm. Hence we expect the two algorithms perform the same when fitting the same data. In this subsection, we show the results of the comparison of these two algorithms.

We use a set of simulated data \((x, y)\), where \(x\) is drawn from the standard normal distribution, and \(y\) is distributed as \text{binomial}(1, p)\. The parameter \(p\) is calculated as \(\logit(p) = -1 + 1.2 \times x\). Hence, the model can be presented as,

\[
E(y) = \logit^{-1}(\alpha + x\beta)
\]  

(3.7)

where \(\alpha = -1\) and \(\beta = 1.2\). We compare the above two algorithms by fitting the model with simulated data sets.

We only shrink \(\beta\) with different \(\lambda\) s, the results are shown in table 3.1 and figure 3.1. In figure 3.1, the circles denote the result of Modified-IWLS and stars denote the result of Newton-Raphson. We can see both from table 3.1 and figure 3.1 that the two algorithms perform exactly the same.

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>(\hat{\alpha})</th>
<th>(\hat{\beta})</th>
<th>Iteration</th>
<th>(\hat{\alpha})</th>
<th>(\hat{\beta})</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>-1.1726</td>
<td>1.4482</td>
<td>5</td>
<td>-1.0809</td>
<td>1.1766</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>-1.1726</td>
<td>1.4482</td>
<td>5</td>
<td>-1.0809</td>
<td>1.1766</td>
<td>5</td>
</tr>
</tbody>
</table>

![Figure 3.1 Trace of convergence](image)

3.4 Selection of Shrinkage Estimator \(\lambda\)

As is discussed before, we can reduce the random variation by adding a penalty into the ordinary likelihood which gives a restriction on the shrinkage estimator \(\hat{\beta}\). In the modified likelihood, we choose an arbitrarily small \(\lambda\). We shrink \(\hat{\beta}\) towards 0 and allow a little bias to stabilize the system and provide more accurate estimates. Actually, there is no strictly technical method to choose a exact shrinkage parameter. However, in generalized restricted ridge regression, a widely used method is the general cross validation (GCV) criterion (Golub, Michael & Grace 1979). Since the Modified-IWLS method linearized the link function of generalized linear models, we can extend the GCV criterion to improve the arbi-
trary selection of $\lambda$, which is,

$$V(k) = \frac{\| (I_{n \times n} - H_{\text{hat}}) z \|^2}{\text{tr}(I_{n \times n} - H_{\text{hat}})^2}$$

(3.8)

where $H_{\text{hat}} = X(X^T X + \lambda I)^{-1} X^T$ is hat matrix, $I_{n \times n}$ is a identity matrix and $z$ is a $n \times 1$ vector of working response in the Modified-IWLS method.

In our Modified-IWLS case, the hat matrix is expressed as follows,

$$H_{\text{hat}} = (X \tilde{X}) \left( (X^T X)^{-1} X^T \right) \left( (\tilde{X}^T \tilde{X} + \lambda I)^{-1} \tilde{X}^T \right)$$

(3.9)

The fact that $H_{\text{hat}}$ based on $\lambda$ leads to $V(k)$ depends on $\lambda$. Therefore we can choose a proper value of $\lambda$ by minimizing $V(k)$. The minimization is often done by line search of $\lambda$ (Hawkins & Xin 2002).

4 Simulation Study

We study the performance of shrinkage estimator comparing to that of the MLE through simulation. Since the two algorithms perform the same with canonical link, in this section we just evaluate the Modified-IWLS algorithm with simulated data sets. In the following two subsections we compare the Modified-IWLS algorithm with the function "glm" in R.

4.1 Nonsparse Binomial Case

We simulate the Binomial case with two coefficients by setting two parameters $\alpha = -1$ and $\beta = 1.2$. The covariate $x$ is sampled from the standard normal distribution, and then $\log(\mu) = \alpha + x \beta$. Hence the true response $y$ is sampled from $\text{binomial}(1, \mu)$. We shrink $\beta$ in the model with simulated data by using the Modified-IWLS and "glm" in R, 1000 times for each different value of $\lambda$.

Figure 4.2 shows the estimation results for different values of $\lambda$. The dashed curve represents the density of estimate attained by "glm" and the solid curve represents the density of estimate attained by Modified-IWLS. We can see that when the shrinkage parameter is relatively small, which refers to the case $\lambda = 0.001$ and $\lambda = 0.1$, estimates from the modified likelihood and the ordinary likelihood are almost the same. But when adding some big values of $\lambda$, difference appears. When $\lambda = 1$, the two methods differ from each other obviously. With $\lambda = 10$, estimates from the modified likelihood shrink towards zero intensively.

Figure 4.1 Simulation results of binomial nonsparse response
Table 4.2 gives the results of standard errors and bias getting from the Modified-IWLS algorithm and "glm" in R, in which $V(\hat{\beta}) = \frac{1}{N} \sum_{r=1}^{N} (\hat{\beta}^{(r)} - \frac{1}{N} \sum_{r=1}^{N} \hat{\beta}^{(r)})^2$ and $Bias = \frac{1}{N} \sum_{r=1}^{N} (\hat{\beta}^{(r)} - \beta)$, where $\hat{\beta}^{(r)}$ is the estimate of $\beta$ in $r$ th replication and $r = 1, 2, ..., N$. The results are corresponding to that shown in figure 4.2 which are for small value of shrinkage parameter $\lambda$, both square-root of variance and bias are almost same from the two method. With $\lambda$ increasing, the square-root of variance of Modified-IWLS decreased while the bias increased.

For balancing the square-root of variance and bias, we consider the mean square error of each method for different value of $\lambda$. Table 4.3 shows the MSE results, where $MSE = V(\hat{\beta}) + (Bias)^2$. Since the simulated data is non-sparse, "glm" works perfectly to fit the model. Obviously for non-sparse data a large $\lambda$ cause larger MSE than a small value of $\lambda$.

### Table 4.2 MSE of Binomial nonsparse response

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>0(glm)</th>
<th>0.001</th>
<th>0.1</th>
<th>1</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.0070</td>
<td>0.0070</td>
<td>0.0036</td>
<td>0.0278</td>
<td>0.5340</td>
</tr>
</tbody>
</table>

#### 4.2 Sparse Binomial Case

Here we consider Binomial case with categorical covariates. Assuming there are five categories which represent five hospitals. Referring to the case Kenneth Carling done, we draw five random numbers from 10 to 100 to be the number of observed patients which are 39, 78, 20, 70 and 54. We set the intercept parameter $\alpha = -1$ and a vector $\beta = (0, -1.9, -2.3, -2.5, -2.8)^T$ such that sparse response exists. In the model, $\beta$ is shrunken by Modified-IWLS. The simulation procedure is the same as the non-sparse case above. We repeat the estimation 1000 times for each different values of $\lambda$. 

Bei Dong, Lu Zhang
Figure 4.2 shows the estimation results for different values of $\lambda$. The dashed curve represents the density of estimate attained by “glm” and the solid curve represents the density of estimate attained by Modified-IWLS. Generally speaking, with the shrinkage parameter $\lambda$ increasing, the solid curve shrinks towards zero, appearing high peak. Because for different $\lambda$, the results from “glm” are the same, the changing of dashed curve is all due to the changing of the length of density-axis. It is surprising to observe two peaks in the figure of beta2hat when $\lambda = 0.001$. However, the whole density is in an average low level when $\lambda$ is small, that is to say the estimates value almost uniformly cover an relatively large interval, which behavior high variation and big bias. When using a large shrinkage parameter, the solid curve appeared high peaks around zero, which means the shrinkage is intensified and the variation is decreased.

Table 4.3 shows the squareroot of variance and bias results, in which $\sqrt{V(\hat{\beta})} = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\beta}_j^{(r)} - \frac{1}{N}\sum_{r=1}^{N}\hat{\beta}_j^{(r)})^2}$ and $Bias = \frac{1}{N}\sum_{r=1}^{N}(\hat{\beta}_j^{(r)} - \beta_j)$, where $\hat{\beta}_j^{(r)}$ is the estimate of $\beta_j$ in $r$ th replication and $r = 1, 2, ..., N$. Since for sparse response “glm” always shows high variation and
big bias, we reckon the result as failure if Std.Error by glm is larger than 10. Since the Modified-IWLS could not provide Std.Error values, we reckon the result as failure if $\hat{\beta}_j^{(r)} - \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_j^{(r)} > 4$. From table 4.3 we find that "glm" method failed many times when comes to the sparse response case. With small shrinkage parameter, Modified-IWLS also failed a certain times. But when $\lambda \geq 0.1$, Modified-IWLS did not fail any more. And also the squareroot of variance decreased and bias increased with shrinkage parameter $\lambda$ increasing.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\sqrt{V(\hat{\beta}_2)}$</th>
<th>Bias</th>
<th>Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0(glm)</td>
<td>0.2705</td>
<td>0.1745</td>
<td>0.2074</td>
</tr>
<tr>
<td>0.001</td>
<td>1.2373</td>
<td>1.6627</td>
<td>2.7677</td>
</tr>
<tr>
<td>0.1</td>
<td>0.3478</td>
<td>0.2200</td>
<td>0.2754</td>
</tr>
<tr>
<td>1</td>
<td>0.0493</td>
<td>0.0179</td>
<td>0.0337</td>
</tr>
<tr>
<td>10</td>
<td>0.0012</td>
<td>0.0003</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

From the values of MSE in table 4.4, where $MSE = V(\hat{\beta}) + (Bias)^2$, we can see the shrinkage estimation performs very well with $\lambda = 0.1$. We also evaluated the shrinkage method with simulated data in Poisson case. We draw five random numbers from 20 to 200 to be the number of observations which are 79, 156, 40, 142 and 111. We set the intercept parameter $\alpha = -1.5$ and a vector $\beta = (0, -0.7, -1.6, -2.7, -2.3)^T$. It turns out the results of Poisson case differs rarely from the Binomial case. We show the results in Appendix.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$MSE$</th>
<th>Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0(glm)</td>
<td>0.1080</td>
<td>219</td>
</tr>
<tr>
<td>0.001</td>
<td>1.6227</td>
<td>120</td>
</tr>
<tr>
<td>0.1</td>
<td>0.3269</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2.3764</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3.4648</td>
<td>0</td>
</tr>
</tbody>
</table>

5 Application with Real Data

In this section we applied the shrinkage method to solve some practical issues. We analyze the data sets that already have been researched by others and compare our results with the existing ones.

5.1 Survival of Heart Transplant

From an investigation of the mortality of heart transplants in different hospitals (Albert 2007), 94 hospitals were recorded on their total number of patients and deaths. The data set are available in R called "heart-transplant" in the package of "LearnBayes". In the data, $y$ denotes the number of deaths and $e$ denotes the total number of patients. We model the data by "glm" and shrinkage estimation to compare their performance.

Figure 5.1 shows the GCV value for different $\lambda$, here we choose $\lambda = 0.2$ at which the decreasing of GCV value slows down remarkably.
Properties of the shrinkage estimator for sparse-response Poisson and Binomial GLM

Figure 5.2 shows the coefficients estimated by "glm" and shrinkage estimation. The circle denotes the results by glm and the star denotes the results by Modified-IWLS. We see that the estimates by "glm" are in two clusters, one of which has value more than 20, while the other group are around 0. The abnormal results implies that "glm" fails to give a proper estimation. When using shrinkage estimation, with the shrinkage parameter $\lambda = 0.2$ shrinking the estimates, the values of the coefficients are scattered in the interval near 0, which makes more sense than the results by "glm".

The comparison of proportions for observed value, prediction by "glm" and prediction by Modified-IWLS are shown in figure 5.3. Since the model is a saturated model, the estimation of "glm" just describes the observed data. But with shrinkage parameter equals to 0.2, Modified-IWLS makes some adjustment to shrink the predictions towards zero, which makes the predictions more sensible.

5.2 Survival of Snails

In an experiment studying the survival of snails (Olsson 2002), groups of 20 snails were held for periods of 1, 2, 3 or 4 weeks under controlled conditions, where temperature and humidity were kept at assigned levels. The snails were of two species (A or B). The experiment was a completely randomized design. The variables are described in table 5.2.

<table>
<thead>
<tr>
<th>Species</th>
<th>Snail species A or B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>Exposure in weeks (1, 2, 3 or 4)</td>
</tr>
<tr>
<td>Humidity</td>
<td>Relative humidity (4 levels)</td>
</tr>
<tr>
<td>Temp</td>
<td>Temperature in degrees Celsius (3 levels)</td>
</tr>
<tr>
<td>Deaths</td>
<td>Number of deaths</td>
</tr>
<tr>
<td>N</td>
<td>Number of snail exposed</td>
</tr>
</tbody>
</table>

A common sense is that the Species and Exposure must have some influence on the survivability of snails. Hence we consider Species, Exposure and their interaction as factors to estimate the coefficients with "glm" and Modified-IWLS. We also choose the value of $\lambda$ by applying the GCV criterion. Figure 5.4 shows the GCV value with respect to different value of $\lambda$,
Here we choose $\lambda = 0.5$ at which the decreasing of GCV value slows down remarkably.

As is shown in table 5.3, the estimates by glm are abnormal and the standard error by "glm" are all too large which imply that the estimation results are highly unstable. While the estimates by Modified-IWLS contain no such extreme values as those by glm. Figure 5.5 shows the comparison of proportions for observed values, predictions by "glm" and predictions by Modified-IWLS. We can see with shrinkage parameter equals to 0.5, Modified-IWLS makes some adjustment to shrink the predictions towards zero, which makes the predictions more sensible.

![GCVCriterionValue](image1.png)

![ProportionComparison](image2.png)

**Figure 5.4 GCV values**

**Figure 5.5 Comparision of proportions**

| Coefficients          | Estimate | Std.Error | z value | Pr(> |z|) |
|----------------------|----------|-----------|---------|------|
| Intercept            | -4.474   | 2.071e+03 | -0.011  | 0.992|
| as.factor(Exposure)2 | 0.298    | 2.071e+03 | 0.008   | 0.993|
| as.factor(Exposure)3 | 2.228    | 2.071e+03 | 0.009   | 0.992|
| as.factor(Exposure)4 | 3.208    | 2.071e+03 | 0.010   | 0.992|
| SpeciesB             | 0.115    | 2.928e+03 | -9.77e-15 | 1.000|
| as.factor(Exposure)2:SpeciesB | 0.931 | 2.928e+03 | 4.55e-04 | 1.000|
| as.factor(Exposure)3:SpeciesB | 1.198 | 2.928e+03 | 4.40e-04 | 1.000|
| as.factor(Exposure)4:SpeciesB | 1.016 | 2.928e+03 | 3.74e-04 | 1.000|

**Table 5.3 Estimation results for snail survival experiment**

<table>
<thead>
<tr>
<th>ship type factors, 4 levels</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of construction</td>
<td>factors, 4 levels</td>
</tr>
<tr>
<td>number of damage incidents</td>
<td>numeric</td>
</tr>
<tr>
<td>Period of operation</td>
<td>factors, 4 levels</td>
</tr>
<tr>
<td>Aggregate moths service</td>
<td>numeric</td>
</tr>
</tbody>
</table>

**5.3 Wave Damage to Cargo Ships**

This study concerns a type of damage caused by waves to the forward section of certain cargo-carrying vessels. The data set is taken from (McCullagh & Nelder 1989). In their (McCullagh & Nelder 1989) analysis, a Poisson model was used but no interaction was considered. However, it is reasonable to believe that a ship produced by the same company in different years has different qualities. The variables are described in table 5.5,
Also by GCV criterion, we choose \( \lambda = 0.7 \) to use shrinkage estimation. Then considering the type of ship, constructed year and their interaction as factors we also attain the estimation results by "glm". Figure 5.6 shows the comparison of nonshrinkage and shrinkage estimates. We see the results by "glm" either too large or too small, which shows high instability. While most of the shrunken estimates gather in the interval (-1,2). They are more concentrated than the results of "glm". Figure 5.7 shows the predictions by glm and Modified-IWLS, we can see that after shrunken, the differences among individuals become smaller. This adjustment by shrunken makes the prediction more stable. Therefore, the shrinkage estimation also performs better than "glm" in Poisson case.

6 Concluding Discussion

In this paper we have posed a modified-likelihood-type shrinkage estimator, summarized and modified two algorithms, analyzed the performance of the shrinkage estimator and applied it in some practical issues.

In the algorithm section, we modify the IWLS algorithm and Newton-Raphson algorithm to solve the estimates numerically. We compare the two algorithms and comes to the conclusion that these two methods perform almost the same. In addition, Modified-IWLS is more applicable than Newton-Raphson when used in Binomial and Poisson generalized linear models. We have also discussed the general cross validation criterion to choose a proper shrinkage parameter.

From the simulation section, we get the conclusion that modified-likelihood-type estimation has good properties when comes to sparse response. From the figures we can see that the stability of the estimation is improved when the shrinkage parameter is properly small while the ordinary-likelihood-type estimator may cause problems under the same situation. The essential advantages of this shrinkage estimates are:

- improved stability of estimates (robustness against separation)
- improvement of mean squared error

Then we have applied the shrinkage estimation with real data in three examples. It turns out with sparse response, the modified-likelihood-type shrinkage estimator has more applicability when fitting the data in sparse case than the estimator derived from ordinary likelihood.

There are several limitations in our study. At first, we only consider the canonical link for Binomial and Poisson generalized linear model. Extending the canonical link to other links may extend the applicability of this modified-likelihood-type shrinkage estimator. However, such an extension is not very difficult. Secondly, we just compared the result with ordinary-likelihood-type estimator, it is better to compare it with shrinkage estimators of other types. Furthermore, the theoretical foundation of shrinkage estimator should be deeply researched. At last, the selection of shrinkage parameter \( \lambda \) can be different for different shrinkage estimators in the same model.
Properties of the shrinkage estimator for sparse-response Poisson and Binomial GLM

Appendix

Poisson Case

Figure A.1 Estimation results of Poisson case

Table A.1 Squareroot of variance and bias of Poisson case

<table>
<thead>
<tr>
<th>λ</th>
<th>$\sqrt{V(\hat{\beta}_2)}$</th>
<th>Bias</th>
<th>Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (glm)</td>
<td>0.0587</td>
<td>0.1692</td>
<td>0.1707</td>
</tr>
<tr>
<td>0.001</td>
<td>0.0587</td>
<td>1.7803</td>
<td>1.2652</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0764</td>
<td>0.4176</td>
<td>0.8982</td>
</tr>
<tr>
<td>1</td>
<td>0.0434</td>
<td>0.0392</td>
<td>0.0308</td>
</tr>
<tr>
<td>10</td>
<td>0.0057</td>
<td>0.0016</td>
<td>0.0020</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>$\hat{\beta}_2$ MSE</td>
<td>$\hat{\beta}_2$ Failure</td>
<td>$\hat{\beta}_3$ MSE</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>0 (glm)</td>
<td>0.0034</td>
<td>0</td>
<td>0.0320</td>
</tr>
<tr>
<td>0.001</td>
<td>0.0034</td>
<td>0</td>
<td>3.6854</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1575</td>
<td>0</td>
<td>0.8230</td>
</tr>
<tr>
<td>1</td>
<td>0.2636</td>
<td>0</td>
<td>0.9692</td>
</tr>
<tr>
<td>10</td>
<td>0.6616</td>
<td>0</td>
<td>2.3366</td>
</tr>
</tbody>
</table>

**R-code of Modified-IWLS Algorithm**

```r
MIWLS <- function(y, X1, X2, N1=1, lambda=0.001, beta0.cell=0.1, conv_crit=1e-4, n_maxiter=10000, family){
  N <- length(y)
  beta10 <- rep(beta0.cell,ncol(X1))
  beta20 <- rep(beta0.cell,ncol(X2))
  eta <- X1%*%beta10+X2%*%beta20
  if(family=="binomial"){
    mu <- as.numeric(N1*(exp(eta)/(1+exp(eta))))
    data <- N1/(mu*(N1-mu))
    w <- sqrt(mu*(N1-mu)/N1)
  }
  if(family=="poisson"){
    mu <- as.numeric(exp(eta))
    data <- 1/mu
    w <- sqrt(mu)
  }
  z <- (eta+(y-mu)*deta)
  k1 <- rep(0,ncol(X1))
  k2 <- rep(1,ncol(X2))
  k <- c(k1,k2)
  A <- diag(k)
  X <- cbind(X1,X2)*w
  Z <- z*w
  beta <- solve(X%*%X+2*lambda*A)%*%t(X)%*%Z
  beta1 <- beta[1:ncol(X1)]
  beta2 <- beta[(ncol(X1)+1):(ncol(X1)+ncol(X2))]
  d <- max(abs(beta1-beta10),abs(beta2-beta20))
  if(d<conv_crit) {conv <- T} else {conv <- F}
  n <- 1
  while(n<n_maxiter & d>conv_crit) {
    beta10 <- as.numeric(beta1)
    beta20 <- as.numeric(beta2)
    eta <- X1%*%beta10+X2%*%beta20
    if(family=="binomial"){
      mu <- as.numeric(N1*(exp(eta)/(1+exp(eta))))
      data <- N1/(mu*(N1-mu))
      w <- sqrt(mu*(N1-mu)/N1)
    }
    if(family=="poisson"){
      mu <- as.numeric(exp(eta))
      data <- 1/mu
      w <- sqrt(mu)
    }
    z <- eta+(y-mu)*deta
  }
}
```

Bei Dong, Lu Zhang
R-code of Newton-Raphson Algorithm for Binomial Case

```r
Newtonbin <- function(y, X1, X2, N1=1, lambda=0.001, beta0.cell=0.1, n_maxiter=100, conv_crit=1e-4){
  N <- length(y)
  A <- diag(ncol(X2))
  beta10 <- rep(beta0.cell,ncol(X1))
  beta20 <- rep(beta0.cell,ncol(X2))
  beta0 <- matrix(c(beta10,beta20),ncol(X1)+ncol(X2),1)
  X <- as.matrix(cbind(X1,X2))
  eta <- X%*%beta0
  dbeta1 <- t(X1)%*%(y-N1*exp(eta)/(1+exp(eta))
  dbeta2 <- t(X2)%*%(y-N1*exp(eta)/(1+exp(eta)))-2*lambda*A%*%beta20
  dbeta <- matrix(c(dbeta1,dbeta2),ncol(X1)+ncol(X2),1)
  temp <- diag(as.numeric(N1*exp(eta)/(1+exp(eta))^2)
  d2beta11 <- -t(X1)%*%temp%*%X1
  d2beta22 <- -t(X2)%*%temp%*%X2-2*lambda*A
  d2beta12 <- -t(X1)%*%temp%*%X2
  d2beta21 <- -t(X2)%*%temp%*%X1
  H <- rbind(cbind(d2beta11,d2beta12),cbind(d2beta21,d2beta22))
  beta <- beta0-solve(H)%*%dbeta
  beta1 <- beta[1:ncol(X1)]
  beta2 <- beta[(ncol(X1)+1):(ncol(X1)+ncol(X2))]
  d <- max(abs(beta-beta0))
  if(d<conv_crit) {conv <- T} else {conv <- F}
  if(conv==T) {
    name.beta1 <- name.beta2 <- NULL
    name.beta1 <- c(name.beta1, paste("Intercept",seq=""))
    if(ncol(X1)>1){
      for(i in 2:ncol(X1)) {name.beta1 <- c(name.beta1, paste("beta_nonshrinkage", as.character(i-1), sep=""))
    }
    for(i in 1:ncol(X2)) {name.beta2 <- c(name.beta2, paste("beta_shrinkage", as.character(i), sep=""))
    }
    est <- matrix(c(beta1,beta2),ncol(X1)+ncol(X2),1)
    dimnames(est) <- list(c(name.beta1, name.beta2), "Estimates")
    est <- round(est, digits=6)
    num.iteration <- paste("Iterations converged after ", n, " times.")
    output <- list(ESTIMATES=est, ITERATION=num.iteration, ShrinkageParameter=lambda)
    return(output)
  }
  else {print("Iterations did NOT converge!")
  }
}
```

// Properties of the shrinkage estimator for sparse-response Poisson and Binomial GLM

```r
k1 <- rep(0,ncol(X1))
k2 <- rep(1,ncol(X2))
k <- c(k1,k2)
A <- diag(k)
X <- cbind(X1,X2)*w
Z <- x*w
beta <- solve((X%*%t(X)+2*lambda*A)%*%t(X)%*%Z)
beta1 <- beta[1:ncol(X1)]
beta2 <- beta[(ncol(X1)+1):(ncol(X1)+ncol(X2))]
d <- max(abs(beta-beta0))
n <- n+1
}
if(conv<conv_crit) {conv <- T} else {conv <- F}
if(conv==T) {
  name.beta1 <- name.beta2 <- NULL
  name.beta1 <- c(name.beta1, paste("Intercept",seq=""))
  if(ncol(X1)>1){
    for(i in 2:ncol(X1)) {name.beta1 <- c(name.beta1, paste("beta_nonshrinkage", as.character(i-1), sep=""))
  }
  for(i in 1:ncol(X2)) {name.beta2 <- c(name.beta2, paste("beta_shrinkage", as.character(i), sep=""))
  }
  est <- matrix(c(beta1,beta2),ncol(X1)+ncol(X2),1)
  dimnames(est) <- list(c(name.beta1, name.beta2), "Estimates")
  est <- round(est, digits=6)
  num.iteration <- paste("Iterations converged after ", n, " times.")
  output <- list(ESTIMATES=est, ITERATION=num.iteration, ShrinkageParameter=lambda)
  return(output)
  }
  else {print("Iterations did NOT converge!")
  }
```
if(d < conv_crit) {conv < -T} else{conv < - F}

n < - 1
while(n < n_maxiter & d > conv_crit) {
    beta10 < - as.numeric(beta1)
    beta20 < - as.numeric(beta2)
    beta0 < - as.numeric(beta)
    eta < - X%*%beta0
    dbeta1 < -t(X1)%*%y-t(X1)%*%(N1*exp(eta))/(1+exp(eta))
    dbeta2 < -t(X2)%*%y-t(X2)%*%(N1*exp(eta))/(1+exp(eta))-2*lemda*A%*%beta20
    dbeta < - matrix(dbeta1,dbeta2,ncol(X1)+ncol(X2),1)
    temp < - diag(numeric(N1*exp(eta))/(1+exp(eta))^2)
    dbeta11 < - -t(X1)%*%temp%*%X1
    dbeta22 < - -t(X2)%*%temp%*%X2-2*lemda*A
    dbeta12 < - -t(X1)%*%temp%*%X2
    dbeta21 < - -t(X2)%*%temp%*%X1
    H < - rhomb(chind(dbeta11,dbeta12),chind(dbeta21,dbeta22))
    H < - as.matrix(H)
    beta < - beta0+solve(H)%*%dbeta
    beta1 < beta[1:ncol(X1)]
    beta2 < beta[(ncol(X1)+1):(ncol(X1)+ncol(X2))]
    d < - max(abs(beta-beta0))
    n < - n+1
}

if(d < conv_crit) {conv < - T} else {conv < - F}

if(conv==T) {
    name.beta1 < - names(beta1)
    name.beta2 < - NULL
    est < - matrix(c(beta1,beta2),ncol(X1)+ncol(X2),1)
    dimnames(est) < - list(c(name.beta1, name.beta2), "Estimates")
    est < - round(est, digits=6)
    num.iteration < - paste("Iterations converged after", n, "times."")
    output < - list(ESTIMATES=est, ITERATION=num.iteration, ShrinkageParameter=lemda)
    return(output)
}

else {print("Iterations did NOT converge!")}

R-code of Newton-Raphson Algorithm for Poisson Case

Newtonpoi <- function(y, X1, X2, lemda=0.001, beta0.cell=0.1, n_maxiter=100, conv_crit=1e-4){
    N < - length(y)
    A < - diag(nrow(X2))
    beta10 < rep(beta0.cell,ncol(X1))
    beta20 < rep(beta0.cell,ncol(X2))
    beta0 < matrix(beta10, beta20, ncol(X1)+ncol(X2),1)
    X < - as.matrix(cbind(X1,X2))
    eta < - X%*%beta0
    dbeta1 < -t(X1)%*%(y-exp(eta))/(1+exp(eta))
    dbeta2 < -t(X2)%*%(y-exp(eta))^2*lemda*A%*%beta20
    dbeta < - matrix(dbeta1,dbeta2,ncol(X1)+ncol(X2),1)
    beta1 < beta1[1:ncol(X1)]
    beta2 < beta2[(ncol(X1)+1):(ncol(X1)+ncol(X2))]
    d < - max(abs(beta-beta0))
    n < - n+1
    if(d < conv_crit) {conv < - T} else{conv < - F}
    if(conv==T) {
        name.beta1 < - names(beta1)
        name.beta2 < - NULL
        est < - matrix(c(beta1,beta2),ncol(X1)+ncol(X2),1)
        dimnames(est) < - list(c(name.beta1, name.beta2), "Estimates")
        est < - round(est, digits=6)
        num.iteration < - paste("Iterations converged after", n, "times."")
        output < - list(ESTIMATES=est, ITERATION=num.iteration, ShrinkageParameter=lemda)
        return(output)
    }
    else {print("Iterations did NOT converge!")}
}
Properties of the shrinkage estimator for sparse-response Poisson and Binomial GLM

```r
temp <- diag(as.numeric(exp(eta)))
d2beta11 <- -t(X1)%*%temp%*%X1
d2beta22 <- -t(X2)%*%temp%*%X2-2*lemda*A
d2beta12 <- -t(X1)%*%temp%*%X2
d2beta21 <- -t(X2)%*%temp%*%X1
H <- rbind(cbind(d2beta11,d2beta12),cbind(d2beta21,d2beta22))
##maybe there is some better way to combine the Hessian
H <- as.matrix(H)
beta <- beta0-solve(H)%*%dbeta
beta1 <- beta[1:ncol(X1)]
beta2 <- beta[ncol(X1)+1:(ncol(X1)+ncol(X2))]
d <- max(abs(beta-beta0))
if(d<conv_crit) conv <- T else conv <- F
n <- 1
while(n<=maxiter & d>conv_crit) {
  beta10 <- as.numeric(beta1)
beta20 <- as.numeric(beta2)
beta0 <- as.numeric(beta)
  eta <- X%*%beta0
dbeta1 <- t(X1)%*%y-t(X1)%*%exp(eta)
dbeta2 <- t(X2)%*%y-t(X2)%*%exp(eta)-2*lemda*A%*%beta20
dbeta <- matrix(c(dbeta1,dbeta2),ncol(X1)+ncol(X2),1)
temp <- diag(as.numeric(exp(eta)))
d2beta11 <- -t(X1)%*%temp%*%X1
d2beta22 <- -t(X2)%*%temp%*%X2-2*lemda*A

d2beta12 <- -t(X1)%*%temp%*%X2
d2beta21 <- -t(X2)%*%temp%*%X1
H <- rbind(cbind(d2beta11,d2beta12),cbind(d2beta21,d2beta22))
H <- as.matrix(H)
beta <- beta0-solve(H)%*%dbeta
beta1 <- beta[1:ncol(X1)]
beta2 <- beta[ncol(X1)+1:(ncol(X1)+ncol(X2))]
d <- max(abs(beta-beta0))
  n <- n+1
}
if(conv==T) {
  name.beta1 <- name.beta2 <- NULL
  name.beta1 <- c(name.beta1, paste("Intercept",seq=""))
  if(ncol(X1)>1){
    for(i in 2:ncol(X1)) {name.beta1 <- c(name.beta1, paste("beta_nonshrinkage", as.character(i-1), sep=""))}
  }
  for(i in 1:ncol(X2)) {name.beta2 <- c(name.beta2, paste("beta_shrinkage", as.character(i), sep=""))
  }
  est <- matrix(c(beta1,beta2),ncol(X1)+ncol(X2),1)
dimnames(est) <- list(c(name.beta1, name.beta2), "Estimates")
est <- round(est, digits=6)
num.iteration <- paste("Iterations converged after", n, "times.")
output <- list(ESTIMATES=est, ITERATION=num.iteration, ShrinkageParameter=lemda)
return(output)
}
else print("Iterations did NOT converge!")
```

Bei Dong, Lu Zhang
Properties of the shrinkage estimator for sparse-response Poisson and Binomial GLM

References


