Australian & New Zealand Journal of Statistics

*Aust. N. Z. J. Stat.* **54**(3), 2012, 359–373 doi: 10.1111/j.1467-842X.2012.00679.x

# SHRINKAGE AND PENALTY ESTIMATORS OF A POISSON REGRESSION MODEL

SHAKHAWAT HOSSAIN[1,*] AND EJAZ AHMED[2]

*University of Winnipeg and Brock University*

## Summary

In this paper we propose Stein-type shrinkage estimators for the parameter vector of a Poisson regression model when it is suspected that some of the parameters may be restricted to a subspace. We develop the properties of these estimators using the notion of asymptotic distributional risk. The shrinkage estimators are shown to have higher efficiency than the classical estimators for a wide class of models. Furthermore, we consider three different penalty estimators: the LASSO, adaptive LASSO, and SCAD estimators and compare their relative performance with that of the shrinkage estimators. Monte Carlo simulation studies reveal that the shrinkage strategy compares favorably to the use of penalty estimators, in terms of relative mean squared error, when the number of inactive predictors in the model is moderate to large. The shrinkage and penalty strategies are applied to two real data sets to illustrate the usefulness of the procedures in practice.

*Key words*: asymptotic distributional bias and risk; likelihood ratio test; Monte Carlo simulation; penalty estimators; Poisson regression; shrinkage estimators.

## 1. Introduction

Variable selection is fundamental in statistical modelling. Initially, there may be many variables to consider as candidates for predictors in the model. Some of these variables may not be active and should therefore be excluded from the final model so as to achieve the goal of good prediction accuracy. Researchers are often interested in finding an active subset of predictors that represents a sparsity pattern in the predictor space. In the next step, they may consider this information and use it either in the full model or in the reduced model. The procedure in this paper is inspired by Stein's result that in a dimension greater than two, efficient estimators can be obtained by shrinking full model estimators in the direction of sub-model estimators.

The Poisson regression model is an important member of the generalized linear model family. It is widely used to study count data in medicine, economics, and social sciences. This model assumes the response variable to have a Poisson distribution, and also that the logarithm of its expected value can be modelled by a linear combination of unknown parameters. Many practitioners prefer to work directly with this model. For this reason, it is treated

*Author to whom correspondence should be addressed.
[1]Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, R3B 2E9, Canada.
e-mail: sh.hossain@uwinnipeg.ca
[2]Department of Mathematics, Brock University, St. Catharines, L2S 3A1, Ontario, Canada.
*Acknowledgements.* We would like to thank the referees, the editor and an associate editor for their valuable suggestions in the revision of this paper. The research of Shakhawat Hossain was supported by a grant from the University of Winnipeg and that of S. Ejaz Ahmed was supported by the Natural Sciences and Engineering Research Council of Canada.

as an independent model in its own right. Some authors dedicate entire chapters to Poisson regression in their books; for example, Dupont (2003) and Zelterman (2010). Sapra (2003) considered the pretest method to estimate the parameters of this model.

In this paper, we consider the problem of estimating the Poisson regression model for the purpose of predicting a response variable that may be affected by several potential predictor variables, some of which may be inactive. The prior information about the inactive variables may be incorporated into the estimation procedure to obtain shrinkage estimators. The existing literature shows that the shrinkage estimators significantly improve upon the classical estimators.

Shrinkage estimation has been studied by several authors: Ahmed & Saleh (1999), Judge & Mittelhammaer (2004), Ahmed, Hussein & Sen (2006), Ahmed *et al.* (2007), Hossain, Doksum & Ahmed (2009), and Ahmed, Hossain & Doksum (2012). These authors developed shrinkage estimation strategies for parametric, semiparametric, and nonparametric linear models. This paper extends the shrinkage estimation strategy to Poisson regression by combining ideas from the recent literature (see, Huang, Ma and Zhang 2008 and Kim, Choi and Oh 2008) on sparsity patterns and compares the resulting estimators to the full and sub-model estimators, as well as to a version of the penalty estimators appropriate for count data. Thus, the goal of this paper is to analyze some of the issues involved in parameter estimation for a Poisson regression model. For example, in genomics research, it is common practice to test a subset of genetic markers for association with a disease. If the subset is found in a certain population after doing genome-wide association studies, then the subset is tested for disease association in a new population. In this new population, it is possible that genetic markers may be discovered that cannot be found in the first population associated with the disease.

Another example can be found in Cameron and Trivedi (1998), who observed that the number of visits to a doctor may be related to sex, age, income, illness, number of reduced activity days, general health questionnaire scores, number of chronic conditions, and dummy variables for two levels (levyplus and freerepa) of health insurance coverage. Since prior information was not available in this case, the shrinkage method uses a two-step approach. In the first step, a set of covariates (the number of reduced activity days, illness, health questionnaire scores, age, sex, and levyplus) are selected based on the best subset selection procedure and traditional model selection criteria, such as AIC and BIC. The effects of other covariates may be inactive. We then use these inactive variables or linear combinations of them to create a linear subspace of the full parameter space for $\boldsymbol{\beta}$. That is, we consider a linear subspace where an unknown $k$-dimensional parameter vector $\boldsymbol{\beta}$ satisfies a set of $q$ linear restrictions

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{h},$$

where $\mathbf{R}$ is a $q \times k$ matrix of rank $q \leq k$ and $\mathbf{h}$ is a given $q \times 1$ vector of constants. Because $\mathbf{R}$ has rank $q$, the $q$ equations do not contain any redundant information about $\boldsymbol{\beta}$. In the second step, we combine the sub and full model estimators in an optimal way in order to achieve an improved estimator for the remaining active parameters. This approach can be implemented for moderate values of $k$. For large values of $k$, one can use modern penalty estimation methods, such as LASSO and its variants, to obtain sub-models and then apply our suggested shrinkage estimation strategy to obtain efficient estimators.

A family of penalized likelihood methods, such as the least absolute shrinkage and selection operation (LASSO) (Tibshirani 1996), the smoothly clipped absolute deviation method (SCAD) (Fan & Li 2001), and adaptive LASSO (Zou 2006) were proposed for linear and generalized linear models. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously. Friedman, Hastie & Tibshirani (2010) developed an efficient algorithm for the estimation of a generalized linear model with a convex penalty which efficiently computes the solution at a given regularization parameter. Thus, the whole process is repeated for typically 100 different regularization parameters to construct a piecewise linear approximation of the true nonlinear solution path. Park & Hastie (2007) proposed an algorithm (called *glmpath*) that generates the coefficient paths for the $L_1$ regularization problems as in LASSO, but in which the loss function is replaced by the negative log-likelihood of any distribution in the exponential family. This method features both shrinkage and variable selection due to the nature of the constraint region, which often results in coefficients of several predictor variables becoming identically zero. However, it does not, possess the oracle properties (Fan & Li 2001). To overcome the inefficiency of traditional variable selection procedures, Fan & Li (2001) proposed SCAD to select variables and estimate the coefficients of variables automatically and simultaneously. This method not only retains the good features of both subset selection and ridge regression, but also produces sparse solutions, and ensures continuity of the selected models. It also gives unbiased estimates for large coefficients. Zou (2006) modified the LASSO penalty by using adaptive weights on $L_1$ penalties on different regression coefficients. Such a modified method was referred to as an adaptive LASSO. It has been shown theoretically that the adaptive LASSO estimator is able to identify the true model consistently, and the resulting estimator is as efficient as oracle. The above three methods have been extensively reported in the literature; for example, Efron *et al.* (2004), Tibshirani *et al.* (2005), Yuan & Lin (2006), Zou (2006), Wang & Leng (2007), Huang *et al.* (2008), Kim *et al.* (2008), and others. In this paper, we provide a unified estimation strategy which implements both shrinkage and penalty methods for estimating the parameters.

The rest of this paper is organized as follows. The model and suggested estimators are introduced in Section 2. The asymptotic properties of the proposed estimators and their asymptotic distributional biases and risks are presented in Section 3. The results of a simulation study that includes a comparison with three penalty methods are reported in Section 4. Application to real data and a comparison of our methods are described in Section 5. Finally, concluding remarks are given in Section 6.

## 2. Estimation strategies

Suppose that $y_i$, given the vector of regressors $\mathbf{x}_i$, is independently distributed as Poisson with probability function:

$$f(y_i|\mathbf{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, \ldots; \quad i = 1, 2, \ldots, n$$

and mean parameter

$$\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ is a $k \times 1$ vector of covariates for the $i$th subject, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression parameters.

Under the assumption of independent observations, the log-likelihood function is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) - \ln y_i!). \tag{1}$$

The derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ is obtained by the chain rule:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \big(y_i - \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\big)\mathbf{x}_i = 0. \tag{2}$$

### 2.1. The unrestricted and restricted maximum likelihood estimators

The unrestricted maximum likelihood estimator (UE) $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by solving the score equation (2) which is non-linear in $\boldsymbol{\beta}$, and can be solved by using an iterative method such as, Newton-Raphson.

Under the usual regularity conditions (see, Santos & Neves 2008), it can be shown that $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal with a variance-covariance matrix $(\mathbf{I}(\boldsymbol{\beta}))^{-1}$, where $\mathbf{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n} e^{\mathbf{x}_i \boldsymbol{\beta}} \mathbf{x}_i \mathbf{x}_i^\top$.

The restricted maximum likelihood estimator (RE) $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be obtained by maximizing the log-likelihood function (1) under the linear restriction $\mathbf{R}\boldsymbol{\beta} - \mathbf{h} = 0$.

### 2.2. The shrinkage and positive shrinkage estimators

The shrinkage estimators are based on the likelihood ratio statistic $D$ for testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{h}$. If $l(\widehat{\boldsymbol{\beta}})$ and $l(\widetilde{\boldsymbol{\beta}})$ are the values of the log-likelihood at the unrestricted and restricted estimates respectively, then

$$\begin{aligned} D &= 2\Big(l(\widehat{\boldsymbol{\beta}}; y_1, \ldots, y_n) - l(\widetilde{\boldsymbol{\beta}}; y_1, \ldots, y_n)\Big), \\ &= (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{h})^\top (\mathbf{R}\mathbf{I}^{-1}(\boldsymbol{\beta})\mathbf{R}^\top)^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{h}) + o_p(1). \end{aligned}$$

Under $H_0$, the distribution of $D$ converges to $\chi^2$ with $q$ degrees of freedom as $n \to \infty$.

The shrinkage estimator (SE) combines the sample and non-sample information in a way that improves the precision of the estimation process or the quality of subsequent predictions. It is easy to implement and adapt to maximum likelihood and other classical estimators. The existing literature shows that the shrinkage estimator has a lower risk than the maximum likelihood estimator in the classical regression models (Ahmed *et al.* 2007). This estimator can be defined as

$$\widehat{\boldsymbol{\beta}}^S = \widetilde{\boldsymbol{\beta}} + \big(1 - (q-2)D^{-1}\big)(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}), \quad q \geq 3.$$

The above estimator is not a convex combination of unrestricted and restricted estimators. However, the shrinkage estimator may have the opposite sign to UE. To avoid this strange behavior of the shrinkage estimator, we truncate this estimator, which leads to a convex combination of the unrestricted and restricted estimators and is called a positive-part shrinkage estimator (PSE). This estimator can be defined as

$$\widehat{\boldsymbol{\beta}}^{S+} = \widetilde{\boldsymbol{\beta}} + \left(1 - (q-2)D^{-1}\right)^{+}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}),$$

where $z^{+} = max(0, z)$.

### 2.3. Penalty estimators

The LASSO for the Poisson regression model was originally proposed by Park & Hastie (2007). It is a popular technique for simultaneous estimation and variable selection. It computes the coefficients that minimize the negative log-likelihood subject to an $L_1$ penalty on $\boldsymbol{\beta}$. The LASSO estimator of $\boldsymbol{\beta}_\lambda$, is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_\lambda^{LASSO} &= \arg\min_{\boldsymbol{\beta}}(-l(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1) \\ &= \arg\min_{\boldsymbol{\beta}}\left(-\sum_{i=1}^{n}\left(y_i\mathbf{x}_i^{\top}\boldsymbol{\beta} - \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta}) - \ln y_i!\right) + \lambda\|\boldsymbol{\beta}\|_1\right),\end{aligned}$$

where $\lambda > 0$ is the tuning parameter. For large values of $\lambda$, this technique produces shrunken estimates of $\boldsymbol{\beta}$, often with many components equal to zero. Park & Hastie (2007) introduced an algorithm that implements the predictor-corrector method to determine the entire path of the coefficient estimates as $\lambda$ varies from 0 to $\infty$. Starting from $\lambda = \infty$, this algorithm computes a series of solutions that estimates the coefficients with a smaller $\lambda$ each time based on previous estimates. The final estimator is denoted here as the LASSO type penalty estimator.

The adaptive LASSO is the solution of

$$\widehat{\boldsymbol{\beta}}_\lambda^{ALASSO} = \arg\min_{\boldsymbol{\beta}}\left(-\sum_{i=1}^{n}(y_i\mathbf{x}_i^{\top}\boldsymbol{\beta} - \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta}) - \ln y_i!) + \lambda\sum_{i=1}^{k}|\beta_i|w_i\right),$$

where $w_i$'s are adaptive weights defined as $w_i = |\widehat{\boldsymbol{\beta}}_i|^{-\tau}$ for some positive $\tau$, and $\widehat{\beta}_i$ is the maximizer of the log likelihood $l(\boldsymbol{\beta})$. The idea of the adaptive LASSO is to give large weights to inactive variables, and thus to heavily shrink their associated coefficients. On the other hand, it gives small weights to active variables, and thus slightly shrinks their associated coefficients. Theoretically, adaptive LASSO enjoys oracle properties (Fan & Li 2001) and LASSO does not. When $k$ is fixed and $n \to \infty$, with some selected $\lambda$, the adaptive LASSO selects the true model with probability tending to one.

Fan & Li (2001) proposed the SCAD method for linear and generalized linear models. This method selects variables and estimates the parameter $\boldsymbol{\beta}$ simultaneously by maximizing the penalized likelihood function

$$\widehat{\boldsymbol{\beta}}_\lambda^{SCAD} = \arg\min_{\boldsymbol{\beta}}\left(-\sum_{i=1}^{n}(y_i\mathbf{x}_i^{\top}\boldsymbol{\beta} - \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta}) - \ln y_i!) + \lambda\sum_{i=1}^{k}p_\lambda(|\beta_i|)\right),$$

where $p_\lambda(\cdot)$ is the SCAD penalty with a tuning parameter $\lambda$ that is to be selected by a data-driven method. The penalty $p_\lambda(\cdot)$ satisfies $p_\lambda(0) = 0$, and its first-order derivative

$$p'_\lambda(\theta) = \lambda\left(I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda}I(\theta > \lambda)\right),$$

where $(t)_+ = tI(t > 0)$ is the hinge loss function, and $a$ is some constant usually taken to be 3.7 (Fan & Li 2001). This method consistently identifies inactive variables by

producing zero solutions for their associated regression coefficients. The tuning parameter $\lambda$ is selected using a cross validation technique.

The output of the above three penalty methods looks like a shrinkage method in both shrinking and deleting coefficients. However, the output of the penalty methods is different from the shrinkage estimation procedure in that it weighs all the covariate coefficients equally. It does not use a specified linear subspace with $\mathbf{R}\boldsymbol{\beta} = \mathbf{h}$.

## 3. Asymptotic results and comparison

In this section, we obtain expressions for the asymptotic distributional bias and risk of the proposed estimators. Suppose that $\boldsymbol{\beta}^*$ is any estimator of $\boldsymbol{\beta}$ and $\mathbf{Q}$ is a positive semi-definite matrix, then the quadratic loss function is

$$\mathcal{L}(\boldsymbol{\beta}^*; \mathbf{Q}) = \big(n^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\big)^\top \mathbf{Q}\big(n^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\big).$$

For theoretical results, we use a general weight $\mathbf{Q}$ on the asymptotic variances and covariances of the estimators. A common choice of $\mathbf{Q}$ is the identity matrix. This is what we use in the simulation study in Section 5.

We now investigate the properties of the estimators under local alternatives. Let $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_q) \in \Re^q$ and consider the following local alternatives:

$$K_{(n)} : \mathbf{R}\boldsymbol{\beta} = \mathbf{h} + \frac{\boldsymbol{\delta}}{n^{1/2}}.$$

The asymptotic distribution function of $\boldsymbol{\beta}^*$ under $K_{(n)}$ is given by

$$G(\mathbf{y}) = \lim_{n\to\infty} P\Big(n^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \le \mathbf{y} | K_{(n)}\Big),$$

where $G(\mathbf{y})$ is a nondegenerate distribution function. We define the *asymptotic distributional risk* (ADR) by

$$\begin{aligned} R(\boldsymbol{\beta}^*; \mathbf{Q}) &= \int \cdots \int \mathbf{y}^\top \mathbf{Q}\mathbf{y} dG(\mathbf{y}), \\ &= \text{trace}(\mathbf{Q}\mathbf{Q}^*), \end{aligned}$$

where $\mathbf{Q}^* = \int \cdots \int \mathbf{y}\mathbf{y}^\top dG(\mathbf{y})$ is the dispersion matrix of $G(\mathbf{y})$.

Note that under nonlocal (fixed) alternatives, all the estimators are asymptotically equivalent to $\widehat{\boldsymbol{\beta}}$, while $\widetilde{\boldsymbol{\beta}}$ has an unbounded risk. In order to make an interesting comparison and to obtain a non-degenerate asymptotic distribution $G(\mathbf{y})$, we need the local alternatives in (3).

We define the *asymptotic distributional bias* (ADB) of an estimator $\boldsymbol{\beta}^*$ as

$$\text{ADB}(\boldsymbol{\beta}^*) = \lim_{n\to\infty} E\big(n^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\big) = \int \cdots \int \mathbf{y} dG(\mathbf{y}),$$

where the second equality can be established under our model assumptions.

Two key results for the study of the ADR and ADB of the SE and PSE are given in the following theorem.

**Theorem 1.** *Under the local alternatives $K_{(n)}$ in (3) and the usual regularity conditions, as $n \to \infty$,*

(i) $n^{1/2}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{h}) \xrightarrow{\mathcal{L}} N(\boldsymbol{\delta}, \mathbf{R}\mathbf{B}^{-1}\mathbf{R}^{\top})$, *where* $\mathbf{B}_{k\times k} = \lim_{n\to\infty} \frac{I(\boldsymbol{\beta})}{n}$ *is nonsingular.*

(ii) *The test statistic* D *converges to a non-central chi-squared distribution* $\chi^2_q(\Delta)$ *with* q *degrees of freedom and non-centrality parameter* $\Delta = \boldsymbol{\delta}^{\top}(\mathbf{R}\mathbf{B}^{-1}\mathbf{R}^{\top})^{-1}\boldsymbol{\delta}$.

Using this theorem, we can obtain the main results of this section. We present (without derivation) the ADB and ADR results. Detailed proofs are available in Hossain, Shrinkage pretest and lasso estimates in parametric and semi-parametric linear models, (unpub. PhD Disc), (2008).

**Theorem 2.** *Under the local alternatives $K_{(n)}$ in (3) and the conditions of Theorem 1, the ADBs of the estimators are*

$$\begin{aligned}
\mathrm{ADB}(\widehat{\boldsymbol{\beta}}) &= \mathbf{0},\\
\mathrm{ADB}(\widetilde{\boldsymbol{\beta}}) &= -\mathbf{J}\boldsymbol{\delta}, \quad \mathbf{J} = \mathbf{B}^{-1}\mathbf{R}^{\top}(\mathbf{R}\mathbf{B}^{-1}\mathbf{R}^{\top})^{-1},\\
\mathrm{ADB}(\widehat{\boldsymbol{\beta}}^{S}) &= -(q-2)\mathbf{J}\boldsymbol{\delta}E\big(\chi^{-2}_{q+2}(\Delta)\big),\\
\mathrm{ADB}(\widehat{\boldsymbol{\beta}}^{S+}) &= -(q-2)\mathbf{J}\boldsymbol{\delta}\big(E\big(\chi^{-2}_{q+2}(\Delta)\big) - E\big(\chi^{-2}_{q+2}(\Delta)I(\chi^2_{q+2}(\Delta) < q-2)\big)\big)\\
&\quad - \mathbf{J}\boldsymbol{\delta}\Psi_{q+2}(q-2,\Delta),
\end{aligned}$$

*where $\Psi_q(.,\Delta)$ is the distribution function of the $\chi^2_q(\Delta)$ distribution.*

**Theorem 3.** *Under the local alternatives $K_{(n)}$ and the assumptions of Theorem 1, the ADRs of the estimators are*

$$\begin{aligned}
R(\widehat{\boldsymbol{\beta}};\mathbf{Q}) &= \mathrm{trace}(\mathbf{Q}\mathbf{B}^{-1}),\\
R(\widetilde{\boldsymbol{\beta}};\mathbf{Q}) &= R(\widehat{\boldsymbol{\beta}};\mathbf{Q}) - \mathrm{trace}(\mathbf{Q}\mathbf{J}\mathbf{R}\mathbf{B}^{-1}) + \boldsymbol{\delta}^{\top}(\mathbf{J}^{\top}\mathbf{Q}\mathbf{J})\boldsymbol{\delta},\\
R(\widehat{\boldsymbol{\beta}}^{S};\mathbf{Q}) &= R(\widehat{\boldsymbol{\beta}};\mathbf{Q}) - 2(q-2)\mathrm{trace}(\mathbf{Q}\mathbf{J}\mathbf{R}\mathbf{B}^{-1})\big(2E\big(\chi^{-2}_{q+2}(\Delta)\big).\\
&\quad - (q-2)E\big(\chi^{-4}_{q+2}(\Delta)\big)\big) + (q-2)\boldsymbol{\delta}^{\top}(\mathbf{J}^{\top}\mathbf{Q}\mathbf{J})\boldsymbol{\delta}\big(2E\big(\chi^{-2}_{q+2}(\Delta)\big)\\
&\quad - 2E\big(\chi^{-4}_{q+2}(\Delta)\big) + (q-2)E\big(\chi^{-4}_{q+4}(\Delta)\big)\big),\\
R(\widehat{\boldsymbol{\beta}}^{S+};\mathbf{Q}) &= R(\widehat{\boldsymbol{\beta}}^{S};\mathbf{Q}) - \boldsymbol{\delta}^{\top}(\mathbf{J}^{\top}\mathbf{Q}\mathbf{J})\boldsymbol{\delta}E\big(\big(1-(q-2)\chi^{-2}_{q+4}(\Delta)\big)^2 I\big(\chi^2_{q+4}(\Delta) < q-2\big)\big)\\
&\quad - \mathrm{trace}(\mathbf{Q}\mathbf{J}\mathbf{R}\mathbf{B}^{-1})E\big(\big(1-(q-2)\chi^{-2}_{q+2}(\Delta)\big)^2 I\big(\chi^2_{q+4}(\Delta) < q-2\big)\big)\\
&\quad + 2\boldsymbol{\delta}^{\top}(\mathbf{J}^{\top}\mathbf{Q}\mathbf{J})\boldsymbol{\delta}E\big(\big(1-(q-2)\chi^{-2}_{q+4}(\Delta)\big)I(\chi^2_{q+4}(\Delta) < q-2)\big).
\end{aligned}$$

In order to elucidate the performance of the estimators reviewed in this section, we next report a simulation study which compares the performance of the estimators and the penalty estimators for finite sample sizes.

## 4. Simulation results

In this section, we use a Monte Carlo simulation experiment to examine the risk (namely MSE) performance of the proposed estimators. Our simulation is based on a Poisson regression model with (i) sample size $n = 60$ for the low dimensional setting ($n \geq k$); and (ii) sample sizes $n = 10$, 15 for the high dimensional setting ($n \leq k$). In this study, we simulate the response from the following model:

$$\log y_i = \mathbf{x}_i^\top \boldsymbol{\beta},$$

where the covariates $\mathbf{x}_i{}^\top = (x_{i1}, x_{i2}, \ldots, x_{in})$ have been drawn from a multivariate standard normal distribution.

We consider the hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$, where the first $p$ columns of $\mathbf{R}$ are zeros and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. Thus the reduced parameter space has $\boldsymbol{\beta}_2 = \mathbf{0}$. The coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $p \times 1$ and $q \times 1$ vectors, respectively with $k = p + q$. We set the true value of $\boldsymbol{\beta}$ at $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \mathbf{0})^\top$ with $\boldsymbol{\beta}_1 = (0.2, -1.2, 0.1)$ and the weight matrix $\mathbf{Q} = \mathbf{I}_1$, where $\mathbf{I}_1$ is the identity matrix. We define

$$\Delta = \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\|^2,$$

where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1^\top, \mathbf{0})^\top$ and $\|\cdot\|$ is the Euclidian norm. Samples were generated using $\Delta$ between 0 and 2. We provide detailed results for $(p, q) = (3, 3)$, $(3, 5)$, $(3, 8)$, $(3, 11)$, $(3, 14)$, and $(3, 17)$.

The number of replications in the simulation was varied initially but it was determined that 2,000 was adequate for each combination of parameters because a further increase in the number did not change the results significantly.

Based on the simulated data, we estimated the MSE of all the estimators studied in this paper. We consider the unrestricted estimator $\widehat{\boldsymbol{\beta}}$ as the 'benchmark' estimator, and thus the performance of the estimators is evaluated in terms of the MSE relative to the MSE of $\widehat{\boldsymbol{\beta}}$ (RelMSE). For any estimator $\widehat{\boldsymbol{\beta}}^*$, the simulated RelMSE of $\widehat{\boldsymbol{\beta}}^*$ to $\widehat{\boldsymbol{\beta}}$ is defined by

$$\text{RelMSE}(\widehat{\boldsymbol{\beta}}^*) = \frac{\text{simulated MSE}(\widehat{\boldsymbol{\beta}})}{\text{simulated MSE}(\widehat{\boldsymbol{\beta}}^*)},$$

keeping in mind that an RelMSE larger than 1 indicates the degree of superiority of the estimator $\widehat{\boldsymbol{\beta}}^*$ over $\widehat{\boldsymbol{\beta}}$.

### 4.1. Restricted, shrinkage, and penalty estimators when $\mathbf{R}\boldsymbol{\beta} = \mathbf{h}$ is correct ($\Delta = 0$) and $n \geq k$

In this case, the penalty estimators can be expected to provide a better estimate of $\boldsymbol{\beta}_1$ by choosing $\lambda$ so that many of the components of $\boldsymbol{\beta}_2$ will be set to zero. Similarly the positive shrinkage estimator can be expected to do well by giving almost all weight to the restricted estimator $\widetilde{\boldsymbol{\beta}}$. Here we investigate how these two procedures compare with each other and with the restricted and unrestricted estimators for low dimensional data.

In Table 1, we give relative MSEs of restricted, shrinkage, and the three penalty estimators (LASSO, adaptive LASSO, and SCAD) with respect to the unrestricted maximum likelihood estimator for $n = 60$ when 3 out of 20 coefficients are not zero. The

TABLE 1

*Simulated RelMSE of RE, SE, PSE, LASSO, adaptive LASSO, and SCAD with respect to* $\widehat{\boldsymbol{\beta}}$ *when the restricted parameter space* $\mathbf{R}\boldsymbol{\beta} = \mathbf{h}$ *is correct* ($\Delta = 0$).

| | $n = 60$ | | | | | |
|---|---|---|---|---|---|---|
| Method | $q = 3$ | $q = 5$ | $q = 8$ | $q = 11$ | $q = 14$ | $q = 17$ |
| RE | 1.69 | 1.78 | 3.48 | 5.97 | 8.58 | 10.07 |
| SE | 1.17 | 1.37 | 2.12 | 3.09 | 4.14 | 5.52 |
| PSE | 1.23 | 1.45 | 2.48 | 4.05 | 4.86 | 6.80 |
| LASSO | 1.28 | 1.51 | 2.11 | 2.77 | 3.97 | 5.78 |
| Adaptive LASSO | 1.31 | 1.57 | 2.13 | 2.78 | 4.07 | 5.92 |
| SCAD | 1.33 | 1.73 | 2.54 | 3.79 | 4.77 | 6.64 |

simulation results are summarized in Table 1 for $\Delta = 0$. The tuning parameter $\lambda$ of the three penalty methods is estimated using 10-fold cross validation.

First, we note that the relative efficiency of all the estimators increases as the number of inactive variables increases. Moreover, at $\Delta = 0$, as we would expect, RE is the best and all the estimators are superior to UE.

Table 1 reveals that the penalty methods perform better than the shrinkage strategy when the number of inactive predictors $q$ in the model is small. We also see from Table 1 that, when $(p, q) = (3, 5)$, the relative efficiency of the LASSO, adaptive LASSO, and SCAD estimators with respect to UE is higher than that of PSE. On the other hand, PSE outshines the LASSO, adaptive LASSO, and SCAD estimators for larger values of $q$. This is somewhat surprising because the penalty estimators are known to be especially effective when there are a small number of active predictors among many candidates. The PSE is even more effective for this scenario and the model considered here.

Generally speaking, in the presence of a relatively large number of inactive predictors in the model, the shrinkage strategy does well relative to the penalty estimators. However, the adaptive shrinkage estimator PSE is preferred when the number of inactive predictors is relatively large.

### 4.2. Restricted and shrinkage estimators when $\mathbf{R}\boldsymbol{\beta} = \mathbf{h}$ is correct and incorrect ($\Delta \geq 0$) and $n \geq k$

The penalty estimators are not included in the $\Delta > 0$ case because they cannot be expected to do well when $\boldsymbol{\beta}_2 \neq \mathbf{0}$ and $n \geq k$. The shrinkage estimators do well by adapting to the $\boldsymbol{\beta}_2 \neq \mathbf{0}$ case. Here we investigate how the shrinkage estimators improve on the unrestricted estimator for low dimensional data. The results for $n = 60$ are presented in Figure 1 and Table 2. The findings are summarized as follows:

(i) The restricted estimator $\widetilde{\boldsymbol{\beta}}$ outshines all the estimators at and near $\Delta = 0$. On the contrary, as $\Delta$ becomes larger than zero, the relative efficiency of $\widetilde{\boldsymbol{\beta}}$ decreases and becomes unbounded whereas the relative efficiency of all the other estimators remains bounded and approaches 1. Thus, severe departure from the restriction is fatal to $\widetilde{\boldsymbol{\beta}}$, but it has little impact on the shrinkage estimators. Furthermore, our numerical finding is in agreement with the asymptotic results of Section 4.

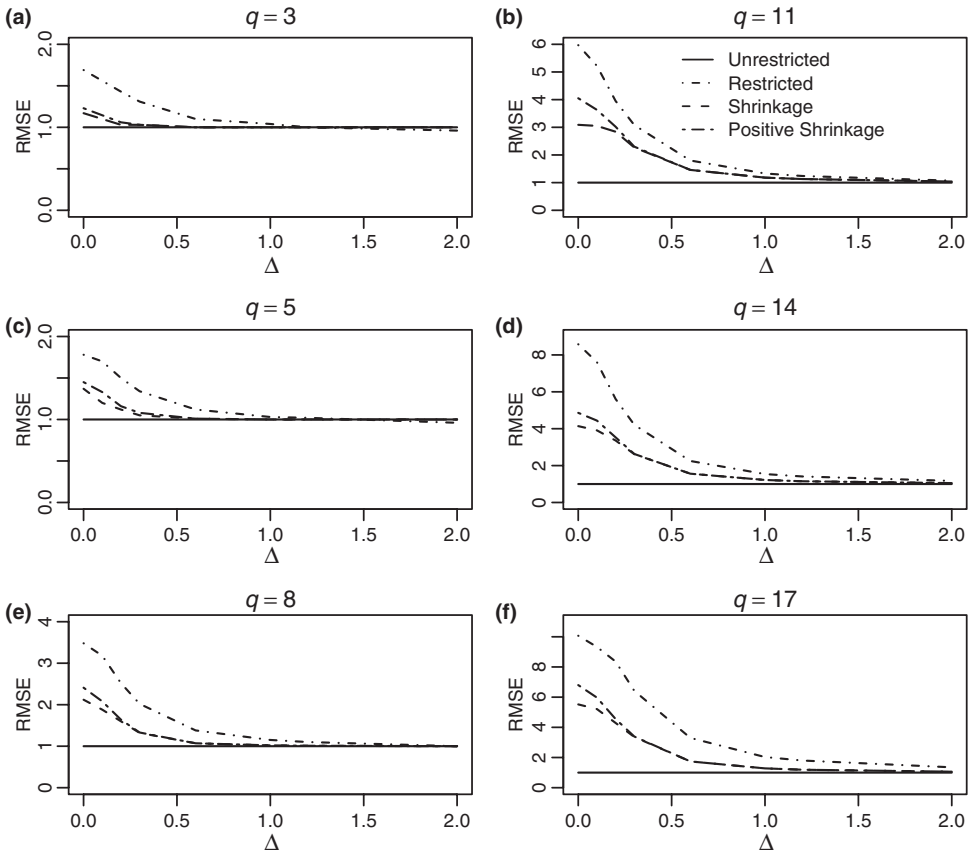


Figure 1. RelMSE with respect to $\widehat{\boldsymbol{\beta}}$ of the estimators when the restricted parameter space is correct and incorrect ($\Delta \geq 0$) for different numbers of inactive predictors. Here $p = 3$, $q = 3, 5, 8, 11, 14, 17$, and $n = 60$.

TABLE 2

*Simulated RelMSE of RE, SE, and PSE with respect to $\widehat{\boldsymbol{\beta}}$ when the restricted parameter space is correct and incorrect ($\Delta \geq 0$) for n = 60 and q = 8.*

| $\Delta$ | $\widetilde{\boldsymbol{\beta}}$ | $\widehat{\boldsymbol{\beta}}^{S}$ | $\widehat{\boldsymbol{\beta}}^{S+}$ |
|---|---|---|---|
| 0.00 | 3.48 | 2.12 | 2.41 |
| 0.32 | 3.18 | 1.88 | 2.08 |
| 0.45 | 2.52 | 1.60 | 1.64 |
| 0.54 | 2.02 | 1.32 | 1.33 |
| 0.77 | 1.38 | 1.07 | 1.07 |
| 1.00 | 1.15 | 1.02 | 1.02 |
| 1.10 | 1.10 | 1.01 | 1.01 |
| 1.41 | 1.00 | 1.00 | 1.00 |

TABLE 3

*Simulated RelMSE of adaptive LASSO and SCAD with respect to LASSO when $n \leq k$.*

| | $n = 10$ | | |
|---|---|---|---|
| Method | $k = 20$ | $k = 25$ | $k = 30$ |
| Adaptive LASSO | 1.12 | 1.17 | 1.19 |
| SCAD | 1.32 | 1.33 | 1.39 |
| | $n = 15$ | | |
| Adaptive LASSO | 1.09 | 1.12 | 1.13 |
| SCAD | 1.19 | 1.24 | 1.36 |

TABLE 4

*Estimates (first row) and standard errors (second row) of the coefficients for number of reduced activity days, illness, health questionnaire scores, age, sex, levyplus on the number of visits to a doctor. The RelMSE column gives the relative mean square error of the estimators with respect to the unrestricted maximum likelihood estimator.*

| Estimators | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | RMSE |
|---|---|---|---|---|---|---|---|
| UE | 0.070 | 0.026 | 0.008 | 0.015 | 0.012 | 0.070 | 1.000 |
| | 0.010 | 0.021 | 0.018 | 0.003 | 0.082 | 0.013 | |
| RE | 0.070 | 0.023 | 0.009 | 0.014 | 0.013 | 0.084 | 1.897 |
| | 0.010 | 0.030 | 0.018 | 0.002 | 0.096 | 0.091 | |
| SE | 0.070 | 0.023 | 0.009 | 0.015 | 0.012 | 0.073 | 1.064 |
| | 0.010 | 0.030 | 0.018 | 0.003 | 0.098 | 0.125 | |
| PSE | 0.070 | 0.025 | 0.009 | 0.015 | 0.012 | 0.073 | 1.067 |
| | 0.010 | 0.030 | 0.018 | 0.002 | 0.098 | 0.125 | |
| LASSO | 0.064 | 0.025 | 0.007 | 0.008 | 0.012 | 0.061 | 1.113 |
| | 0.097 | 0.030 | 0.010 | 0.018 | 0.003 | 0.132 | |
| Adaptive LASSO | 0.061 | 0.016 | 0.005 | 0.007 | 0.011 | 0.053 | 1.136 |
| | 0.056 | 0.022 | 0.010 | 0.012 | 0.001 | 0.075 | |
| SCAD | 0.058 | 0.013 | 0.003 | 0.005 | 0.008 | 0.042 | 1.174 |
| | 0.053 | 0.022 | 0.010 | 0.011 | 0.002 | 0.092 | |

(ii) If the number of inactive variables $q = 3$ and the sample size is 60, the RelMSEs of the shrinkage and positive shrinkage estimators are 1.17 and 1.23 when the restriction holds, and they increase as the number of variables $q$ increases which is in agreement with the theoretical results.

For example, for $q = 11$ and $n = 60$, the RelMSEs of these estimators are 3.09 and 4.05, respectively, indicating the outstanding performances of the proposed estimators. On the other hand, the RelMSE falls sharply as $\Delta$ moves away from 0 and converges to one irrespective of $p$ and $q$. Figure 1 shows that all the estimators dominate $\widehat{\boldsymbol{\beta}}$ for small values of $\Delta$ and shrinkage and positive shrinkage estimators work better in case of large $q$.

Simulation studies for many other possible combinations of $n$, $p$, $q$, and $\Delta$ were carried out and showed similar results. As mentioned before, the penalty estimators are relatively inefficient when $\Delta > 0$ because they are not designed for non-sample information. These results are not reported here, but are available from the authors.

### 4.3. Penalty estimators when $n \leq k$

In this sub-section we consider the penalty estimators when the data are high dimensional ($n \leq k$). The maximum likelihood method and the shrinkage estimators for Poisson regression model are not applicable in this Subsection. Here we investigate the relative MSEs of adaptive LASSO and SCAD with respect to LASSO when the sample sizes are $n = 10, 15$ and the predictors are $k = 20, 25, 30$.

Table 3 shows the relative MSEs of the adaptive LASSO and SCAD estimators when three predictor variables are active and all other predictors are inactive. The relative MSEs of adaptive LASSO and SCAD are larger than the LASSO estimator, which implies that the adaptive LASSO and SCAD outperform the LASSO estimator.

## 5. Real data examples

### 5.1. Australian health survey data

A detailed description of the data is available in Cameron & Trivedi (1998). We apply the proposed estimation strategies to this data set. A total of 5,190 individuals over 18 years of age answered all of the essential questions that are recorded in this data set. The main objective of this survey was to study the relationship between the number of consultations with a doctor and the type of health insurance, health status, and socioeconomic indicators. The response variable of interest was the number of visits to a doctor that were made during a two week interval, and the covariates of interest were sex, age, income, illness, number of reduced activity days, general health questionnaire scores, number of chronic conditions, and dummy variables for two levels (levyplus = 1 if the respondent is covered by private health insurance, 0 otherwise, freerepa = 1 if the respondent is covered for free by the government, 0 otherwise) of health insurance coverage.

The maximum likelihood theory shows that the number of reduced activity days ($\beta_1$), illness ($\beta_2$), health questionnaire scores ($\beta_3$), age ($\beta_4$), sex ($\beta_5$), levyplus ($\beta_6$) are the active predictors of the number of visits to a doctor, while the other three variables income ($\beta_7$), number of chronic conditions ($\beta_8$), and freerepa ($\beta_9$) are inactive predictors. In a future study, a natural linear subspace would omit these three variables. In this example, our null hypothesis has $\boldsymbol{\beta}_2 = (\beta_7, \beta_8, \beta_9) = (0, 0, 0)$, $k = 9$, $p = 6$, $q = 3$, and $n = 5 = 190$.

We bootstrap from the data to examine the performance of the proposed estimation strategies for estimating coefficients of the other six variables. We draw bootstrap samples of size $n = 1500$ by 1000 times drawing 1500 rows with replacement from the data matrix $(y_i, x_{i,j})$ to examine the point estimates, standard errors, and relative efficiencies of the proposed estimators. The results in Table 4 are consistent with the simulation findings. Thus, the penalty estimators perform well relative to the shrinkage estimators when the number of inactive variables is relatively small. On the other hand, the adaptive positive shrinkage estimator performs well when there are moderate or relatively large numbers of inactive predictors in the model. However, the restricted estimator outperforms the penalty methods because the inactive predictors that are deleted in the sub-model are indeed irrelevant, or nearly irrelevant, for the response.

TABLE 5

*Estimates (first row) and standard errors (second row) of the coefficients for the effect of bid price, management invitation, and total book value of assets on the number of takeover bids. The RelMSE column gives the relative mean square error of the estimators with respect to the unrestricted maximum likelihood estimator.*

| Estimators | $\beta_1$ | $\beta_2$ | $\beta_3$ | RMSE |
|---|---|---|---|---|
| UE | −0.798 | 0.498 | 0.045 | 1.000 |
| | 0.320 | 0.118 | 0.040 | |
| RE | −0.710 | 0.565 | 0.042 | 2.270 |
| | 0.333 | 0.128 | 0.034 | |
| SE | -0.401 | 0.360 | 0.029 | 1.224 |
| | 0.355 | 0.135 | 0.036 | |
| PSE | −0.405 | 0.361 | 0.028 | 1.238 |
| | 0.350 | 0.134 | 0.036 | |
| LASSO | −0.710 | 0.565 | 0.042 | 1.509 |
| | 0.328 | 0.118 | 0.037 | |
| Adaptive LASSO | −0.761 | 0.527 | 0.044 | 1.557 |
| | 0.327 | 0.118 | 0.037 | |
| SCAD | −0.807 | 0.582 | 0.048 | 1.742 |
| | 0.314 | 0.102 | 0.035 | |

### 5.2. Takeover bids data

Again, we apply our estimation strategies to a takeover bids data set provided by Cameron & Trivedi (1998). The data set includes the number of bids received by 126 U.S. firms that were targets of takeover offers during the period between 1978 and 1985. These firms were actually taken over within 52 weeks of the initial offer. The response count variable is the number of bids after the initial bid received by the target firm. The covariates include defensive actions taken by management of the target firm, firm-specific characteristics, and government intervention. The defensive actions taken by the target firm include indicator variables for legal defense by lawsuit, proposed changes in asset structure, proposed changes in ownership structure, and management invitations for friendly third-party bids. The firm-specific characteristics are bid price divided by the price that prevails 14 working days before the bid, percentage of stock held by institutions, and total book value of assets in billions of dollars. Federal regulators are indicator variables for Department of Justice interventions.

Based on maximum likelihood inference, percentage of stock held by institutions ($\beta_4$), legal defense by lawsuit ($\beta_5$), proposed changes in asset structure ($\beta_6$), proposed changes in ownership structure ($\beta_7$), and government intervention ($\beta_8$) are not significantly associated with the number of takeover bids received by targeted firms. To examine the effectiveness of our procedure, we regard the information on these variables as non-sample information and use the penalty and shrinkage methods to evaluate the effect of bid price ($\beta_1$), management invitation ($\beta_2$), and total book value of assets on the takeover bids ($\beta_3$). In this example, our null hypothesis has $\boldsymbol{\beta}_2 = (\beta_4, \beta_5, \beta_6, \beta_7, \beta_8) = (0, 0, 0, 0, 0)$, $k = 8$, $p = 3$, $q = 5$, and $n = 126$. From Table 5, we can make similar conclusions as we made from the previous example.

## 6. Discussion and conclusion

In the low-dimensional simulation setting, we compared the performance of shrinkage estimators, penalty estimators, and the maximum likelihood estimator in the context of a Poisson regression model with potentially inactive predictors when we have prior knowledge of a linear subspace. We explored the risk properties of the shrinkage estimators via asymptotic distributional risk and Monte Carlo experiments. The properties of the penalty estimators were evaluated by Monte Carlo simulation studies and it was found that they are competitive when there are many inactive predictors in the model for $n \geq k$. On the other hand, the adaptive positive shrinkage estimator performs better when the number of inactive predictors is moderate or relatively large. Further, the shrinkage estimators with data based weights perform well. In fact, the shrinkage estimators outperformed the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ in the entire parameter space for $q > 2$.

In the high-dimensional simulation setting, we compared the relative performance of adaptive LASSO and SCAD with the LASSO estimator. It shows that SCAD and adaptive LASSO perform better than the LASSO estimator.

Finally, we applied the proposed strategies to two real data sets to evaluate the relative performance of the suggested estimators. These results are consistent with our analytical and simulated findings. One of the reviewers suggested that the theoretical and numerical results can be extended to the entire class of generalized linear models. We are currently working on this project.

## *References*

AHMED, S.E., HUSSEIN A.A. & SEN, P.K. (2006). Risk comparison of some shrinkage M-estimators in linear models. *J. Nonparametr. Statist.* **18**, 401–415.

AHMED, S. E. & SALEH, A.K.M.E. (1999). Estimation of regression coefficients in an exponential regression model with censored observation. *J. Japan Statist. Soc.* **29**, 55–64.

AHMED, S.E., HOSSAIN, S. & DOKSUM, K.A. (2012). LASSO and shrinkage estimation in Weibull censored regression models. *J. Statist. Plann. Inference.* **142**, 1273–1284.

AHMED, S.E., DOKSUM, K.A., HOSSAIN, S. & YOU, J. (2007). Shrinkage, pretest and LASSO estimators in partially linear models. *Aust. N. Z. J. Stat.* **49**, 461–471.

CAMERON, A.C. & TRIVEDI, P.K. (1998). *Regression Analysis of Count Data.* Cambridge, UK: Cambridge University Press.

DUPONT, W.D. (2003). *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data.* Cambridge, UK: Cambridge University Press.

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–451.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softw.* **33**, 1–22.

HOSSAIN, S., DOKSUM, K.A. & AHMED, S.E. (2009). Positive-part shrinkage and absolute penalty estimators in partially linear models. *Linear Algebra Appl.* **430**, 2749–2761.

HUANG, J., MA, S., & ZHANG, C. (2008). Adaptive LASSO for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603–1618.

JUDGE, G.G. & MITTELHAMMAER, R.C. (2004). A semiparametric basis for combining estimation problem under quadratic loss. *J. Amer. Statist. Assoc.* **99**, 479–487.

KIM, Y., CHOI, H. & OH, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.* **103**, 1665–1673.

PARK, M.Y. & HASTIE, T. (2007). An $L_1$ regularization-path algorithm for generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 659–677.

SANTOS, J.A. & NEVES, M.M. (2008). A local maximum likelihood estimator for poisson regression. *Metrika.* **68**, 257–270.

SAPRA, S. K. (2003). Pre-test estimation in poisson regression model. *Econom. Lett.* **10**, 541–543.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused LASSO. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 91–108.

WANG, H. & LENG, C. (2007). Unified LASSO estimation via least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039–1048.

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 49–68.

ZELTERMAN, D. (2010). *Applied Linear Models: with GLM*. Cambridge, UK: Cambridge University Press.

ZOU, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.