Variable Selection in a Log–Linear Birnbaum–Saunders Regression Model for High-Dimensional Survival Data via the Elastic-Net and Stochastic EM

Yukun Zhang and Xuwen Lu
Department of Mathematics and Statistics
University of Calgary
2500 University Drive NW
Calgary, Alberta T2N 1N4, Canada
(yukzhang@ucalgary.ca; lux@math.ucalgary.ca)

Anthony F. Desmond
Department of Mathematics and Statistics
University of Guelph
50 Stone Road East
Guelph, Ontario N1G 2W1, Canada
(tdesmond@uoguelph.ca)

The Birnbaum–Saunders (BS) distribution is broadly used to model failure times in reliability and survival analysis. In this article, we propose a simultaneous parameter estimation and variable selection procedure in a log–linear BS regression model for high-dimensional survival data. To deal with censored survival data, we iteratively run a combination of the stochastic EM algorithm (SEM) and variable selection procedure to generate pseudo-complete data and select variables until convergence. Treating pseudo-complete data as uncensored data via SEM makes it possible to incorporate iterative penalized least squares and simplify computation. We demonstrate the efficacy of our method using simulated and real datasets.

KEY WORDS: Birnbaum–Saunders distribution; Censored data; Coordinate descent; Iterative least squares.

1. INTRODUCTION

Fatigue failure happens when a material is subject to fluctuating stresses and strains. Statistical models for the fatigue process describe randomly varying failure times of fatiguing materials. Birnbaum and Saunders (1969) derived a model coined the BS model showing that fatigue failure is caused by the development and growth of cracks from cyclic loading. Desmond (1985) provided a more general derivation based on a biological model. There are many advantages in using the BS model in reliability analysis. For example, this model fits well within the extremes of the distribution, even when there is little lifetime data.

Let $T$ be a nonnegative random variable. The distribution of $T$ is a Birnbaum–Saunders distribution with shape parameter $\alpha$ and scale parameter $\beta$, that is, $T \sim BS(\alpha, \beta)$, if the cumulative density function (c.d.f) of $T$ is given by

$$F(t) = \Phi \left( \frac{1}{\alpha} \left( \sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}} \right) \right),$$

where $\Phi$ is the standard normal c.d.f. The probability density function (p.d.f) of $T$ is given by

$$f(t; \alpha, \beta) = \frac{1}{2\sqrt{2\pi}\alpha\beta} \left\{ \left( \frac{\beta}{t} \right)^{\frac{1}{2}} + \left( \frac{\beta}{T} \right)^{\frac{1}{2}} \right\} \exp \left\{ -\frac{1}{2\alpha^2} \left( \frac{t}{\beta} + \frac{\beta}{t} - 2 \right) \right\},$$

where $t > 0$, $\alpha > 0$, $\beta > 0$. This p.d.f is unimodal and its median is $\beta$.

For complete data without censoring, Saunders and Birnbaum (1968) derived maximum likelihood estimators (MLEs) of $\alpha$ and $\beta$ and their asymptotic distribution was obtained by Engelhardt, Bain, and Wright (1981). Although the MLEs have several optimal properties, it is difficult to get the closed form for the estimator of $\beta$. For this reason, Ng, Kundu, and Balakrishnan (2003) proposed modified moment estimators (MMEs) for $\alpha$ and $\beta$ and a bias correction technique. They used a jackknife procedure to reduce the bias of the MLEs and MMEs. For censored data, Desmond (1982) studied MLE for both Type I and Type II censored data from the BS distribution. Wang, Desmond, and Lu (2006) considered parameter estimation for the BS distribution under random censoring.

In the regression setting, Rieck and Nedelman (1991) studied the following log–linear model,

$$Y = X'\theta + \epsilon,$$

where $Y = \ln(T)$, $X = (X_1, X_2, \ldots, X_p)'$ is a $p$-dimensional covariate vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ is a $p$-vector of regression coefficients, $\mathbf{a}'$ denotes transpose of an arbitrary vector $\mathbf{a}$ and $\epsilon$ follows the sinh-normal distribution $SN(\alpha, 0, 2)$. If $\theta_1$ is treated as an intercept, then $X_1 \equiv 1$. For complete data without censoring, Saunders and Birnbaum (1968) derived maximum likelihood estimators (MLEs) of $\alpha$ and $\beta$ and their asymptotic distribution was obtained by Engelhardt, Bain, and Wright (1981). Although the MLEs have several optimal properties, it is difficult to get the closed form for the estimator of $\beta$. For this reason, Ng, Kundu, and Balakrishnan (2003) proposed modified moment estimators (MMEs) for $\alpha$ and $\beta$ and a bias correction technique. They used a jackknife procedure to reduce the bias of the MLEs and MMEs. For censored data, Desmond (1982) studied MLE for both Type I and Type II censored data from the BS distribution. Wang, Desmond, and Lu (2006) considered parameter estimation for the BS distribution under random censoring.

In the regression setting, Rieck and Nedelman (1991) studied the following log–linear model,

$$Y = X'\theta + \epsilon,$$

where $Y = \ln(T)$, $X = (X_1, X_2, \ldots, X_p)'$ is a $p$-dimensional covariate vector, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ is a $p$-vector of regression coefficients, $\mathbf{a}'$ denotes transpose of an arbitrary vector $\mathbf{a}$ and $\epsilon$ follows the sinh-normal distribution $SN(\alpha, 0, 2)$. If $\theta_1$ is treated as an intercept, then $X_1 \equiv 1$. For complete data without censoring, Saunders and Birnbaum (1968) derived maximum likelihood estimators (MLEs) of $\alpha$ and $\beta$ and their asymptotic distribution was obtained by Engelhardt, Bain, and Wright (1981). Although the MLEs have several optimal properties, it is difficult to get the closed form for the estimator of $\beta$. For this reason, Ng, Kundu, and Balakrishnan (2003) proposed modified moment estimators (MMEs) for $\alpha$ and $\beta$ and a bias correction technique. They used a jackknife procedure to reduce the bias of the MLEs and MMEs. For censored data, Desmond (1982) studied MLE for both Type I and Type II censored data from the BS distribution. Wang, Desmond, and Lu (2006) considered parameter estimation for the BS distribution under random censoring.
In fact, model (3) can be expressed as a BS regression model as follows:

\[ T = \beta(x) \delta = \exp(x' \theta) \delta, \]

where \( \delta = \exp(\varepsilon) \sim \text{BS}(\alpha, \beta = 1) \). Therefore, model (3) and (4) are equivalent. But model (3) is linear with respect to \( \mathbf{X} \), and is easier to interpret than model (4). Desmond, Rodríguez-Yam, and Lu (2008) implemented an EM algorithm to fit a log–linear regression model with censored data when failure times follow the BS distribution.

In reliability and survival analysis, there are many examples of data where the number of covariates are often greater than the sample size, which is called a small \( n \) and big \( p \) problem. For example, a dataset corresponding to electron-probe X-ray microanalysis of archaeological glass vessels in Janssens et al. (1998) has a sample of size 180 glass vessels and 1920 spectrum frequencies as predictors. The breast cancer dataset in Sorlie et al. (2003) contains 115 female patients, including 77 patients whose survival times were censored during the follow-up, and 549 gene expression measurements as predictors.

We now turn to variable selection methods for these high-dimensional problems with \( p \gg n \). Most variable selection criteria are closely related to penalized least squares and penalized log-likelihoods. Some traditional variable selection criteria such as AIC and BIC requiring subset selection have some drawbacks. They ignore stochastic errors inherent at the stage of variable selection and also lack stability (Breiman 1996). To retain important variables and to avoid the instability of the subset selection, Tibshirani (1996) proposed the LASSO with \( L_1 \) penalty. Efron et al. (2004) further provided deep insights into procedures of the LASSO and least angle regression. Hoerl and Kennard (1970) considered the \( L_2 \) penalty, which yields a ridge regression to handle the ill-conditioned design matrix. Zou and Hastie (2005) combined the \( L_1 \) penalty and the \( L_2 \) penalty into an elastic-net penalty, which thus enjoys flexibility of being a more LASSO-like or more ridge-like solution, depending on the choice of tuning parameters. Fan and Li (2001) suggested the use of smoothly clipped absolute deviation (SCAD) penalty. The SCAD is an improvement over the \( L_0, L_1, \) and \( L_2 \) penalty.

Variable selection has been widely used in various survival regression models. To name a few, Huang, Ma, and Xie (2006) considered the LASSO and the threshold gradient directed regularization for simultaneous variable selection and estimation in the accelerated failure time model with high-dimensional covariates, based on Stute’s weighted least-squares method. Hu and Lian (2013) used the LASSO and SCAD penalties for variable selection in a partially linear proportional hazards model. These variable selection methods can identify the important variables among all the covariates. We are motivated to conduct this research by a few reliability and survival datasets such as the two datasets aforementioned and recent developments of the Birnbaum–Saunders model. To the best of our knowledge, there is no work on variable selection in Birnbaum–Saunders regression models in the literature. The contribution of this article is to extend the elastic-net penalty to the BS regression model.

Censoring is a common problem in survival data analysis. Broadly speaking, censoring occurs when some lifetimes are known to have occurred only within certain intervals. Here, we only consider right censoring. In this case, data can be represented by pairs of random variables \( \{Y_i, \Delta_i\} \), where \( Y_i = \ln[\min(T_i, C_i)] = \log[T_i \Delta_i + C_i(1 - \Delta_i)] \) and \( \Delta_i = [T_i \leq C_i], i = 1, \ldots, n; \Delta_i \) indicates whether lifetime \( T_i \) corresponds to an event (\( \Delta_i = 1 \)) or is censored (\( \Delta_i = 0 \)); \( C_i \) represents censoring time.

The expectation-maximization (EM) algorithm formalized by Dempster, Laird, and Rubin (1977) is a well-known tool to deal with censored data. The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the expectation and maximization steps. However, calculating the conditional expectation required in the E-step of the algorithm may be infeasible, especially when there are multi-dimensional expectations over the censoring ranges. There are also problems in obtaining maxima. The converging point is sensitive to the initial value and sometimes leads to local maxima or saddle points of the log-likelihood function. To overcome the above limitations of EM, Diebolt, Ip, and Olkin (1994) developed the stochastic EM algorithm (SEM) algorithm. The idea of SEM is that the expectation can be estimated by simulation. Since SEM was proposed, several authors have applied it to solve their problems. Tregouet and Tietj (2004) applied SEM to estimate haplotype effects in censored data analysis using a standard Cox proportional hazards formulation. Chauveau (1995) also extended SEM for censored mixture models and studied its asymptotic behavior. Using SEM, we can draw pseudo-samples from their conditional density given the observed data and current estimates of the parameters. Replace the censored data by the pseudo-samples so that we can get pseudo-complete data. Then, the problem can be treated as in uncensored data case.

In this article, we consider simultaneous estimation and variable selection in model (3) for high-dimensional survival data, where the survival or failure time \( T \) may be right censored and the number of covariates \( p \) is greater than the sample size \( n \). Under either case of censored or uncensored data and \( p \gg n \), to the best of our knowledge, simultaneous estimation and variable selection in model (3) has not been studied in the literature. Our research aims to fill this gap.

The remainder of this article is organized as follows. In Section 2, we first review the existing methods for estimation with uncensored data when dimensionality is low. Then, we study the methodology for simultaneous parameter estimation and variable selection. The coordinate descent algorithm and the stochastic EM algorithm for the implementation of the proposed methodology are described in Section 3. Simulation studies and real data examples are presented in Sections 4 and 5, respectively. Some concluding remarks are provided in Section 6. Finally, some additional materials and the R codes for the simulation studies and data analysis are given in the online supplementary material.

2. ESTIMATION AND REGULARIZATION METHODOLOGY

2.1 Parameter Estimation for Uncensored Data When Dimensionality Is Low

In Section 1, we introduced the BS regression model and its equivalent: the log–linear model (3). Here we first discuss the estimation of the parameters \( \theta \) and \( \alpha \) in the log–linear model.
when dimensionality of \( x \) is low and fixed, that is, \( p < n \) and data are uncensored.

Let \((y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\) be \( n \) independent observations from model (3), \( y_i = \log(t_i), t_i \) are uncensored failure times. Then, the log-likelihood function that Rieck and Nedelman (1991) considered for this model is given as follows:

\[
l_0 = \ln[L(\theta)] = \sum_{i=1}^{n} [\ln(W_i) - Z_i^2/2] + \text{constant},
\]

where

\[
W_i = (2/\alpha)\cosh[(y_i - x_i^{\theta})/2],
\]

and

\[
Z_i = (2/\alpha)\sinh[(y_i - x_i^{\theta})/2].
\]

The sinh and cosh represent hyperbolic sine and cosine functions which are defined as \( \sinh(x) = (e^x - e^{-x})/2, \) and \( \cosh(x) = (e^x + e^{-x})/2. \) To get the maximum likelihood estimator, we differentiate the likelihood function \( l_0 \) with respect to \( \theta_j \) and \( \alpha \) and set them equal to zero. Thus, we obtain the following equations:

\[
\frac{\partial l_0}{\partial \theta_j} = \frac{1}{2} \sum_{i=1}^{n} x_{ij} \left( Z_i W_i - Z_i/W_i \right) = 0 \quad \text{for} \quad j = 1, \ldots, p,
\]

where \( p \) is the number of covariates, and

\[
\frac{\partial l_0}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha} \sum_{i=1}^{n} Z_i^2 = 0.
\]

So, the expression for the MLE of \( \alpha^2 \) in terms of the MLE \( \hat{\theta} \) of \( \theta \) by solving Equation (9) is

\[
\hat{\alpha}^2 = \frac{4}{n} \sum_{i=1}^{n} \sinh^2[(y_i - x_i^{\hat{\theta}})/2].
\]

Since there is no closed-form solution for Equation (8), we compute the estimate of \( \theta \) using ordinary least squares (OLS). Rieck and Nedelman (1991) proved that the least-squares estimator of \( \theta \) with uncensored data provides an unbiased estimator for \( \theta \) and is highly efficient for small values of \( \alpha \). They showed that the success of least-squares regression is due to the approximate normality of the sinh-normal when \( \alpha \) is small. They also mentioned that the efficiency decreased as \( \alpha \) increased. However, under many commonly occurring conditions, it was felt that the value for \( \alpha \) would be less than 1. For model (3), \( E(\epsilon_i) = 0 \) and \( \text{var}(\epsilon_i) = 4\alpha(\alpha) (i = 1, \ldots, n). \) There is no closed-form expression for \( \omega(\alpha) \) but it could be estimated from data. If the observations \( y_1, \ldots, y_n \) are independent, then \( \text{cov}(\epsilon_i, \epsilon_j) = 0, i \neq j, (i, j = 1, \ldots, n). \) With these conditions on the \( \epsilon_i \)’s, the OLS of \( \theta \) is the best linear unbiased estimator of \( \theta: \)

\[
\hat{\theta} = (XX)^{-1}XY,
\]

where \( Y \) is the column vector of the \( y_i \)’s and \( X = (x_{11}, \ldots, x_{1n})' \).

An unbiased estimator of \( \omega(\alpha) \) is given by \( \hat{\omega}(\alpha) = \sum_{i=1}^{n}(y_i - x_i^{\hat{\theta}})^2/[4(n - p)]. \)

To improve efficiency, we can use a numerical procedure to calculate the MLE, by taking the least-squares estimate of \( \theta \) as the initial value, and the initial value of \( \alpha \) calculated by (10). Here, we show the procedure by using Fisher’s method of scoring. It is equivalent to an iterative ordinary least-squares procedure. Rieck and Nedelman (1991) obtained the following system of equations at the \( m \)th iteration:

\[
\theta^{(m)} = (XX)^{-1}XY^*,
\]

where

\[
Y^* = X\theta^{(m-1)} + (2/C(\alpha^{(m-1)}))R
\]

is computed at the \((m-1)\)th step. \( C(\alpha) = 2 + 4/\alpha^2 - \pi/\alpha^2 \) \((2\pi/\alpha^2)^{1/2}1 - \text{erf}((2\pi/\alpha^2)^{1/2})\) \(\exp(2/\alpha^2), \) \( \text{erf}(x) \) is the error function defined by \( \text{erf}(x) \equiv (2/\sqrt{\pi}) \int_{-\infty}^{x} e^{-r^2} dr, \) and \( R = (R_1, \ldots, R_n) ', R_i = Z_i W_i - Z_i/W_i. \) For small values of \( \alpha, \) we have \( C(\alpha) \approx 1 + (4/\alpha^2). \)

The iterative equation for \( \alpha \) is

\[
\alpha^{(m)} = [\alpha^{(m-1)}/2] \left[ 1 + \sum_{i=1}^{n} Z_i^2/n \right].
\]

Using these equations, we can use the iterative least-squares method to get the MLEs of \( (\theta, \alpha) \) as follows:

1. Obtain the initial value \( \theta^{(0)} \) from Equation (11). Obtain \( \alpha^{(0)} \) from \( \theta^{(0)} \) using the square root of Equation (10).
2. Set \( m = 1. \)
3. Calculate \( R \) and \( Y^* \) using \( \theta^{(m-1)} \) and \( \alpha^{(m-1)} \).
4. Calculate \( \theta^{(m)} \) and \( \alpha^{(m)} \) by Equations (12) and (13), respectively.
5. If \( |\theta^{(m)} - \theta^{(m-1)}| \) and \( |\alpha^{(m)} - \alpha^{(m-1)}| \) are sufficiently small, stop iteration and proceed to Step 7. Otherwise, go to Step 6.
6. Add 1 to \( m \) and go to Step 3. Continue until convergence is attained.
7. \( \theta^{(m)} \) and \( \alpha^{(m)} \) are the MLEs of \( (\theta, \alpha) \).

The above procedure works well for uncensored data with \( n > p. \) But for high-dimensional data (i.e., \( n < p \)) and censored data, the existing methods are not directly applicable. Instead we propose a regularization method based on the elastic-net penalty for simultaneous variable selection and parameter estimation.

2.2 Elastic-Net Penalty and Regularization Method

In general, parameter estimation and variable selection can be implemented by the penalized log-likelihood as follows:

\[
\arg\min_{\theta} \{-\ln(L(\theta)) + \lambda P_\tau(\theta)\},
\]

where \( P_\tau(\cdot) \) is penalty function and \( \lambda \geq 0 \) is a tuning parameter controlling the degree of regularization. In this article, we consider the elastic-net penalty given by

\[
P_\tau(\theta) = \sum_{i=1}^{p} \left\{ \frac{1}{2}(1 - \tau)\theta_i^2 + \tau | \theta_i | \right\},
\]

where \( \tau \) is another tuning parameter, \( 0 \leq \tau \leq 1. \) The elastic-net penalty provides a compromise between the ridge penalty (\( \tau = 0 \)) and the LASSO penalty (\( \tau = 1 \)). The first term ensures democracy among groups of correlated variables, whereas the second term enforces the sparsity of the solution. By choosing the tuning parameter \( \tau, \) we can tune elastic-net to be more...
ridge-like or more LASSO-like. The elastic-net also incorporates the advantages of LASSO and ridge regression and does group selection. It is particularly useful in the \( p \gg n \) situation, or any situation where there are many correlated predictor variables. Zou and Hastie (2005) used simulation results to demonstrate the good performance of the elastic-net and its superiority over the LASSO. The elastic-net produces a sparse model with good prediction accuracy, while encouraging a grouping effect.

However, computation based on the above penalized log-likelihood is very complicated. A solution to this is to iteratively consider any penalty and variable selection. Specifically, our new method uses the iterative least-squares procedure to construct the penalized least-squares objective function defined as

\[
M_\lambda(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i^* - x_i^T \theta)^2 + \lambda P_r(\theta). \tag{15}
\]

By minimizing Equation (15), we can get a regularized estimate of \( \theta \) at the \( m \)th iteration. Then, we can also update \( \alpha \) by using the estimate of \( \theta \) as we did using (13) in Section 2.1. Since it is a problem of penalized least squares instead of penalized log-likelihood, this idea was used by Rieck and Nedelman (1991) under the case of \( p < n \) and uncensored data, where they did not consider any penalty and variable selection. Specifically, our new method uses the iterative least-squares procedure to construct the penalized least-squares objective function defined as

In summary, the following are the steps of estimating \( \theta \) and selecting variables using SEM and iterative penalized least squares based on elastic-net penalty.

1. First, treat observed data as complete data and use iterative penalized least-squares estimation procedure to get the penalized MLE of \( (\theta, \alpha) \). Use the MLE as the initial values \( \hat{\theta}(0) \) and \( \alpha(0) \).
2. Put \( m = 1 \).
3. Draw samples \( Z^{(m)} = (Z_{r+1}^{(m)}, \ldots, Z_n^{(m)}) \) from Equation (17) given the current estimation of parameters \( \theta^{(m-1)} \) and \( \alpha^{(m-1)} \).
4. With the pseudo-complete sample \( \{T_1, \ldots, T_r, Z_{r+1}^{(m)}, \ldots, Z_n^{(m)}\} \), we update \( \theta^{(m-1)} \) and \( \alpha^{(m-1)} \) to \( \theta^{(m)} \) and \( \alpha^{(m)} \) by the iterative penalized least squares method.
5. Add 1 to \( m \) and repeat Steps 3 and 4.
6. After a burn-in period of \( m_0 \) iterations, the sequences \( \theta^{(m)} \) and \( \alpha^{(m)} \) are expected to be close to its stationary regime and we stop after a sufficiently large number of iterations \( M \).
7. Calculate the average of \( (M - m_0) \) estimates, denoted by \( \bar{\theta} \) and \( \bar{\alpha} \), which are the final estimates of \( \theta \) and \( \alpha \).

In our numerical studies, we take \( M = 300 \) and \( m_0 = 100 \). According to the properties of an ergodic Markov chain, the averaged estimates will converge to the true parameter values with probability one when \( M \to \infty \). The final output from the stochastic EM is a sample from a stationary distribution whose mean is close to the MLE and whose variance reflects the information loss due to missing data. In addition to the iterative penalized least squares, the stochastic EM algorithm for censored data involves another layer of iteration and requires more computing time, but the algorithm is quite simple and avoids a problematic regularized minimization procedure based on penalized log-likelihood function, because the whole problem can be treated as in the complete data case using penalized least squares, which can be easily implemented, for example, by the R package glmnnet. We also want to select the optimal tuning parameter \( \lambda \) to minimize risk in estimation. All these are discussed in the online supplementary material.

3. IMPLEMENTATION

3.1 Stochastic EM Algorithm for Censored Data

For censored data, following the definition given in Section 1, we treat the problem as a typical missing data problem, where \( T = \{T_1, \ldots, T_r, T_{r+1}, \ldots, T_n\} \) is the complete data and \( Y = \{Y_1 = T_1, \ldots, Y_r = T_r, Y_{r+1} = (C_{r+1}, \Delta_{r+1}), \ldots, Y_n = (C_n, \Delta_n)\} \) is the observed data. Denote \( \bar{C} = \{Y_{r+1} = (C_{r+1}, \Delta_{r+1}), \ldots, Y_n = (C_n, \Delta_n)\} \), and let \( Z = \{T_{r+1}, \ldots, T_n\} \) be the \((n - r)\) survival times censored by \( C_i, i = r + 1, \ldots, n \).

The observed data likelihood function is

\[
g(Y \mid \theta) = \int_Z f((Y, Z) \mid \theta) dZ. \tag{16}
\]

The complete data likelihood function, \( L_c(\theta) = f((Y, Z) \mid \theta) \), often admits a closed form expression or simple estimating equation for the corresponding MLE, \( \hat{\theta}_{ML}(Y, Z) \), while the observed data likelihood function given by (16) does not. The information about the missing data is completely embedded in the conditional density of data \( Z \) given \( \bar{C}, X \), and \( \theta^{(m)} \) (Diebolt, Ip, and Olkin 1994):

\[
p(Z, \bar{C}, X, \theta^{(m)}) = f(Z, X, \theta^{(m)}) / S(\bar{C}, X, \theta^{(m)}). \tag{17}
\]

The main idea of the stochastic EM is to “fill-in” the missing data with a single draw from Equation (17), that is, the conditional density given the observed data including covariate \( X \) and \( \theta^{(m)} \), a current guess of the parameter. Specifically, in each iteration, we generate data by SEM to replace the censored observation so that we get pseudo-complete data. This is called the S-step. Once we have a pseudo-complete sample, we can directly maximize its complete data log-likelihood to obtain an updated maximizer \( \theta^{(m+1)} \). This is called the M-step.

In summary, the following are the steps of estimating \( \theta \) and selecting variables using SEM and iterative penalized least squares based on elastic-net penalty.

1. First, treat observed data as complete data and use iterative penalized least-squares estimation procedure to get the penalized MLE of \( (\theta, \alpha) \). Use the MLE as the initial values \( \hat{\theta}(0) \) and \( \alpha(0) \).
2. Put \( m = 1 \).
3. Draw samples \( Z^{(m)} = (Z_{r+1}^{(m)}, \ldots, Z_n^{(m)}) \) from Equation (17) given the current estimation of parameters \( \theta^{(m-1)} \) and \( \alpha^{(m-1)} \).
4. With the pseudo-complete sample \( \{T_1, \ldots, T_r, Z_{r+1}^{(m)}, \ldots, Z_n^{(m)}\} \), we update \( \theta^{(m-1)} \) and \( \alpha^{(m-1)} \) to \( \theta^{(m)} \) and \( \alpha^{(m)} \) by the iterative penalized least squares method.
5. Add 1 to \( m \) and repeat Steps 3 and 4.
6. After a burn-in period of \( m_0 \) iterations, the sequences \( \theta^{(m)} \) and \( \alpha^{(m)} \) are expected to be close to its stationary regime and we stop after a sufficiently large number of iterations \( M \).
7. Calculate the average of \( (M - m_0) \) estimates, denoted by \( \bar{\theta} \) and \( \bar{\alpha} \), which are the final estimates of \( \theta \) and \( \alpha \).

In our numerical studies, we take \( M = 300 \) and \( m_0 = 100 \). According to the properties of an ergodic Markov chain, the averaged estimates will converge to the true parameter values with probability one when \( M \to \infty \). The final output from the stochastic EM is a sample from a stationary distribution whose mean is close to the MLE and whose variance reflects the information loss due to missing data. In addition to the iterative penalized least squares, the stochastic EM algorithm for censored data involves another layer of iteration and requires more computing time, but the algorithm is quite simple and avoids a problematic regularized minimization procedure based on penalized log-likelihood function, because the whole problem can be treated as in the complete data case using penalized least squares, which can be easily implemented, for example, by the R package glmnnet. We also want to select the optimal tuning parameter \( \lambda \) to minimize risk in estimation. All these are discussed in the online supplementary material.

3.2 Variable Selection by Multiple Testing in Stochastic EM

In our numerical analysis with censored data, in each simulation run involving SEM algorithm, we get \( M - m_0 = 200 \) sets of coefficient estimates from 200 iterations which may give us different variable selection results. To select a final set of variables from all these results, we adopt the idea of a hypothesis test for a proportion. For each simulation setting, we calculate the proportion of nonzero estimates for each variable in the 200 iterations; denote the proportions as \( \pi = \pi_1, \ldots, \pi_p \).

For a one-proportion test with null hypothesis \( H_0 : \pi = \pi_0 \) versus alternative hypothesis \( H_a : \pi < \pi_0 \), the test statistic
one-proportion z-test is
\[
Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/m_0}} \sqrt{M - m_0},
\]
where \( \hat{\pi} \) is the estimated proportion.

In our case, we treat all variables whose proportions of nonzero estimates are statistically less than 0.5 as nonsignificant variables. Then, our hypothesis is \( H_0: \pi = 0.5 \) versus \( H_1: \pi < 0.5 \). For each individual test, we calculate the value of the test statistic from (18), which equals \( Z_{\text{calc}} = (\hat{\pi} - 0.5)\sqrt{M - m_0}/0.5 \), and \( p \)-value from \( p_i = P(Z \leq Z_{\text{calc}}) \). We treat the whole test procedure as a multiple testing problem. We use the method of Benjamini and Hochberg (1995) to control the false discovery rate and calculate adjusted \( p \)-values \( \hat{p}_i \). This can be implemented by using the R function \texttt{p.adjust}.

To study departures from the error assumptions as well as presence of outlying observations in the BS regression model, Leiva et al. (2007) considered two kinds of residuals: standard deviance component residuals and standardized martingale-type residuals. We use their method to check if the BS regression model is appropriate in a specific application. We demonstrate this using two real-data examples in Section 5.

3.3 Residual Analysis for Model Checking

To study departures from the error assumptions as well as presence of outlying observations in the BS regression model, Leiva et al. (2007) considered two kinds of residuals: standard deviance component residuals and standardized martingale-type residuals. We use their method to check if the BS regression model is appropriate in a specific application. We demonstrate this using two real-data examples in Section 5.

4. SIMULATION STUDY

4.1 Design of Simulation Study

The purpose of the simulation study is to test the performance of simultaneous parameter estimation and variable selection in the Birnbaum–Saunders regression for high-dimensional uncensored and censored data, respectively. We also simulate grouped or correlated data to see if the elastic-net is a better variable selection procedure than LASSO when covariates are grouped or correlated.

We simulate data from model (3). To generate \( \varepsilon \sim \text{SN}(\alpha, 0, 2) \), according to Rieck (1989), first we generate \( T \sim \text{BS}(\alpha, 1) \), then let \( \varepsilon = \ln(T) \). Following Ng, Kundu, and Balakrishnan (2003), we generate \( T \) from a normal random variable by using the following transformation:

\[
T = \beta(1 + 2V^2 + 2V(1 + V^2)^{1/2}),
\]

where \( V \) is normally distributed with mean zero and variance 1/4\( \alpha^2 \). In generating pseudo-responses from the conditional density (17), we use the acceptance–rejection method. We present the following three examples to illustrate our method under different design matrices \( X \) and different censoring rates.

We first generate high-dimensional uncensored and censored datasets in Example 1. In this example, we use LASSO for variable selection. To observe the difference between the LASSO and the elastic-net, we generate data with grouping effects in Example 2, and data with correlated covariates in Example 3.

Example 1: We simulate uncensored data with \( p = 204 \) covariates and \( n = 50 \). The shape parameter \( \alpha = 0.5 \), \( \theta = (1, 1.5, 5, 3, 2, -0.5, 0, \ldots, 0) \) here the first element 1.5 is the intercept, \( T \sim \text{BS}(\alpha, \beta = \exp(\mathbf{X} \theta)) \) and \( \mathbf{X} \sim \text{multivariate normal} \); its components are independent and follow \( N(0, \mathbf{sd} = 2) \). Censored data are generated with censoring variable \( C \sim \exp(\text{cvalue} + \sqrt{2} \times N(0, 1)) \), so that \( Y = \log(T) \) is randomly censored by \( C \), where \( \text{cvalue} \) is used to control the censoring rate (cr). The censoring rates are chosen as \( \text{cr}=16\%, 29\%, \) and \( 41\% \), corresponding to \( \text{cvalue}=1, 2, 3 \), respectively.

Example 2: We choose \( \theta = (1, 1.5, \ldots, 2, 3, \ldots, 3, 4, \ldots, 10, 199, 199, \ldots, 0)' \) (the first element 1 is the intercept) and \( \alpha = 0.5 \). We simulate correlated \( X \) by setting

\[
X_i = Z_1 + \varepsilon_i^1, \quad Z_1 \sim N(0, 1), \quad i = 3, \ldots, 6,
\]

\[
X_i = Z_2 + \varepsilon_i^2, \quad Z_2 \sim N(0, 1), \quad i = 7, \ldots, 11,
\]

\[
X_i = Z_3 + \varepsilon_i^3, \quad Z_3 \sim N(0, 1), \quad i = 12, \ldots, 16,
\]

\[
X_i \equiv 1, \quad X_2 \sim N(0, \mathbf{sd} = 0.1), \quad X_3 \sim N(0, 1),
\]

where \( \varepsilon_i^1 \) are iid \( N(0, \mathbf{sd} = 0.01), \quad i = 3, \ldots, 16 \). The censoring variable \( C \) is chosen in the same way as in Example 1. Under this setting, the number of predictors is 214 and there are still \( n = 50 \) observations.

Example 3: Based on Example 1, we generate \( X \) from a multivariate normal distribution with mean zero and covariance matrix \( (\rho^{i-j})_{i,j=1,\ldots,n} \); \( \rho \) is set to 0.1, 0.5, and 0.8, respectively, to control the amount of correlation between covariates.

In Section 4.2, we will present the simulation results based on 100 simulation runs for the Examples 1 and 2. If the model and the model assumptions are violated, it is natural to ask whether the proposed method is still reliable. We conduct simulations to assess the effect of model misspecification on the inference. The results indicate that when the model assumptions are violated, the proposed method is still reliable. This part and the results of Example 3 are included in the online supplementary material.

4.2 Examples for Parameter Estimation and Variable Selection

4.2.1 Example 1: Parameter Estimation and Variable Selection With Uncensored and Censored Data. We test the performance under different censoring rates: \( \text{cr} = 0\% \) (i.e., uncensored data), 16\%, 29\%, and 41\% when \( \alpha = 0.5 \) and \( \alpha = 1.5 \),
respectively. We use five measures to evaluate the variable selection performance:

- $\#S$ refers to the average number of selected variables. We want $\#S$ to be close to the true number of nonzero parameters.
- $\#FN$ refers to the average number of incorrectly excluded variables. An incorrectly excluded variable means the variable should be selected, but our result shows that this variable is not significant.
- $L_1$-loss is defined as the $L_1$ norm of the difference between the target value $\theta_0$ and the estimated value $\hat{\theta}$, in the form of $\| \hat{\theta} - \theta_0 \|_1$.
- $L_2$-loss is defined as the $L_2$ norm of the difference between the target value $\theta_0$ and the estimated value $\hat{\theta}$, in the form of $\| \hat{\theta} - \theta_0 \|_2$.
- Mean squared error (MSE) is defined as $\sum_{i=1}^{p} (\hat{\theta}_i - \theta_0)^2 / p$. For each simulation, we will get a mean squared error. We record the median of MSE from all the simulation runs.

For $\#FN$, $L_1$-loss, $L_2$-loss, and median of MSE, we want them to be as small as possible.

Tables 1–3 show the results obtained from Example 1. In Table 1, we only show the estimation of $\alpha$, the intercept $\theta_1$ and nonzero parameters $\theta_2$ to $\theta_5$. Table 2 shows the five measures under different settings. Values reported in the bracket are standard deviations (SD). Rounding all the $\#S$s to the nearest integer, we report all the IDs of selected variables based on the rank of the percentages of being selected under each setting in Table 3.

The estimation for uncensored data ($cr = 0\%$) when $\alpha = 0.5$ is quite good for both $\alpha$ and $\theta$. When data are censored, the estimation bias and standard deviation are larger than those for uncensored data, and increase when the censoring rate increases. When $\alpha = 1.5$, the estimation for $\alpha$ has a large bias. This is reasonable since we mentioned in Section 2.1 that the efficiency of least-squares estimator decreases as $\alpha$ increases. The poor estimate of $\alpha$ also affects the estimate of $\theta$. Rieck and Nedelman (1991) observed the same phenomenon in the case of low-dimensional data.

When $\alpha = 0.5$, the number of selected variables is around 12 while the number of true nonzero variables is 4. According to Table 3, when we rank variables according to the magnitude of proportions of selected variables from the largest to the smallest, the true significant variables $\#2$ to $\#5$ have the highest rank. Thus, although more variables are selected, the true significant variables are most likely to be selected and have the highest rank.

As reported in Table 2, for all the five measures except $\#S$, they are small and have small standard deviations, which shows our method does well in parameter estimation and variable selection. $L_1$-loss and $L_2$-loss increase as the censoring rate increases. This pattern is consistent with the bias shown in Table 1.

When $\alpha = 0.5$ is increased to 1.5, more variables are selected and the bias becomes larger. Unlike the case of parameter estimation, variable selection does not produce noticeably different results compared to that under $\alpha = 0.5$. In fact, all the true significant variables are still selected with high probability and have the highest rank as shown in Table 3.

Comparing variable selection for uncensored and censored data in terms of the results from $cr = 0\%$ and results from other censoring rates, we do not see big differences from the
VARIABLE SELECTION IN A LOG–LINEAR BIRNBAUM–SAUNDERS REGRESSION MODEL

Table 3. Example 1: Selected variables in censored data simulation under different settings

| $\alpha$ | cr (%) | Top $S$ ranked variables |
|---------|--------|--------------------------|
| 0       | 2 3 4 5 120 186 146 149 | 175 22 87 124 |
| 0.5     | 2 3 4 5 10 100 114 178 | 11 15 52 |
| 1.5     | 2 3 4 5 89 42 175 7 | 13 85 101 106 |
| 4       | 2 3 4 5 117 157 27 87 | 147 15 86 107 108 111 116 139 |
| 16      | 2 3 4 5 58 15 20 64 | 77 11 51 118 124 131 165 |

Table 4. Example 2: Selected variables in censored data simulation under different settings

| $\alpha$ | cr (%) | Top $S$ ranked variables |
|---------|--------|--------------------------|
| 33      | LASSO 2 13 15 8 9 11 14 10 7 12 16 3 4 6 5 141 82 173 |
| 43      | LASSO 2 15 7 9 8 12 3 13 10 11 16 6 14 4 5 214 75 82 102 185 |

In conclusion, our method can provide good parameter estimation and variable selection simultaneously in model (3) for high-dimensional censored data. Particularly, when $\alpha$ is smaller than 1, not only variable selection but also parameter estimation has pretty good performance. Both LASSO and elastic-net can select all the significant variables with high probability, but tend to select more variables than true variables. We cannot see any obvious difference between LASSO and elastic-net based on our simulation studies.

5. REAL-DATA ANALYSIS

In this section, we apply our method to two datasets. These datasets represent industrial data and medical survival data, respectively. The first one is the vessel data, which is a high-dimensional uncensored dataset from chemical analysis. We apply the proposed elastic-net method to select important variables. The second one is a high-dimensional censored dataset from DNA microarray experiments. For the censored data, we first generate pseudo-complete data by using the stochastic EM algorithm, then select important variables based on the proposed method for uncensored data analysis.

5.1 Vessel Data

Variable selection is particularly useful in the field of chemometrics, where hundreds or even thousands of spectra need to be analyzed. We examine a situation in which archeological glass vessels from the 16th and 17th centuries were investigated through chemical analysis. The aim of this study was to learn more about the production of these vessels, particularly regarding their origin and possible trade connections between known producers.

Maronna (2011) proposed a robust ridge regression for high-dimensional data. He analyzed this dataset to demonstrate the advantage of his approach but did not consider variable selection. Following Maronna (2011), we consider this dataset corresponding to electron-probe X-ray microanalysis of archeological glass vessels (Janssens et al. 1998), where each of $n = 180$ vessels is represented by a spectrum of 1920 frequencies. For each vessel, the contents of 13 chemical compounds are registered. Since the frequencies below 15 and above 500 have almost null values of $x_{ij}$, we keep frequencies 15–500, so that we have $p = 486$. This dataset does not contain censored observations. We choose the content of the 11th chemical compound as the response $T$. We can treat $T$ as an event time since it is positive. According to our model, the response $Y$ is a logarithmic transformation of the original chemical compound $T$. Our preliminary analysis shows that the frequency covariates in the dataset are highly correlated, hence, it is more desirable to use the elastic-net penalty than the LASSO penalty to select correlated variables. We first use the cv.glmnet function in the R package glmnet to select the tuning parameter $\tau$ based on

| cr (%) | Method | Top $S$ ranked variables |
|--------|--------|--------------------------|
| 33     | LASSO 2 13 15 8 9 11 14 10 7 12 16 3 4 6 5 141 82 173 |
| 43     | LASSO 2 15 7 9 8 12 3 13 10 11 16 6 14 4 5 214 75 82 102 185 |
Table 5. Example 2: Comparisons between LASSO and elastic-net for variable selection in censored data simulation with grouping effects

| cr (%) | Method | #S | FN | \(L_1\)-loss | \(L_2\)-loss | Median of MSE |
|--------|--------|----|----|--------------|--------------|--------------|
| 33     | LASSO  | 17.87 (4.91) | 3.33 (2.17) | 33.13 (12.91) | 9.73 (3.25)  | 0.40         |
|        | Enet   | 21.72 (7.19) | 1.03 (2.02) | 23.75 (19.70) | 7.43 (6.46)  | 0.12         |
| 43     | LASSO  | 19.74 (6.15) | 5.58 (2.65) | 44.47 (22.00) | 13.24 (9.09) | 0.59         |
|        | Enet   | 23.33 (8.15) | 3.13 (3.34) | 34.59 (20.10) | 10.67 (8.13) | 0.38         |

NOTE: Numbers in parentheses are standard deviations.

a linear regression model and obtain \(\tau = 0.1\). This small value of \(\tau\) may be a result of high correlation between the frequency covariates. Using this \(\tau\), we apply the proposed penalized iterative least-squares procedure for variable selection. In total, 145 variables are selected from 486 spectra. This large number of selected variables is probably due to highly correlated covariates, such that they are selected by elastic-net in a group. But the number of selected variables is less than the sample size \(n = 180\), therefore it still gives us feasibility to use a variety of standard statistical methods. Figure 1 shows the normal probability plots for the residuals calculated by the method of Leiva et al. (2007) based on the selected covariates. It suggests that the BS model is a good fit to the vessel data. Compared to other approaches to analysis of this dataset, Maronna’s (2011) robust ridge regression does not reduce the dimension of the covariate space, while our method can select important variables and estimate parameters simultaneously. Consequently, it is preferred in applications.

5.2 Breast Cancer Gene Expression Data

This dataset was studied by Sørlie et al. (2003) and contains 549 gene expression measurements, exit time and exit status in a study of breast cancer among 115 women. Seventy-seven observations are censored, resulting in about a 67% censoring rate.

Using the proposed methodology, we implement the SEM algorithm and variable selection by the elastic-net simultaneously. We use the `cv.glmnet` function in the R package `glmnet` to select the tuning parameter in each SEM iteration step of the 300 iterations, and obtain an average \(\tau = 0.998\) after removing 100 burn-in samples, which indicates LASSO is preferred in variable selection in this dataset. Using the multiple testing method given in Section 3.2, we select 71 significant genes out of 549; the estimate of the shape parameter is \(\hat{\alpha} = 0.070\). By the method of Leiva et al. (2007) based on the selected covariates, the normal probability plots for residuals shown in Figure 2 indicate that the BS regression model is an adequate fit to the data. Gorst-Rasmussen and Scheike (2011) analyzed this dataset using LASSO in the additive hazards model. They selected only four variables. Interestingly, we do not have any selected variables in common. When we fitted a BS regression model with those four covariates, it ended up with a poor fit and a large amount of variation in the response variable cannot be explained by the four covariates. Our method selected more variables, improved goodness of fit and increased model predictability. Hence, it is an adequate approach in analyzing this dataset.

6. DISCUSSION AND CONCLUSION

We have proposed a regularization method for simultaneous parameter estimation and variable selection in the log–linear
Birnbaum–Saunders regression model. When survival times are censored, we successfully impute the censored values by the stochastic EM algorithm. Our methodology shows good performance in both simulation and data analysis. The dimension can be efficiently reduced and the parameters can be adequately estimated. This valuable approach can be broadly applied to high-dimensional industrial and survival data which may contain a censored survival response. The elastic-net variable selection method can deal with data with or without grouping effects. Moreover, the simulation studies suggest that the proposed method is robust to model misspecification and can improve efficiency and accuracy in estimation when the BS model is correctly specified.

In future research, we can improve our method from three perspectives: (i) From the method perspective, we can bring in more variable selection penalty functions and compare them with elastic-net and LASSO, such as the smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) and the smooth integration of counting and absolute deviation (SICA) (Lv and Fan 2009). These existing methods, including LASSO and elastic-net, are designed for either variable selection or shrinking regression coefficients for highly correlated covariates in a group, but not really for group selection. In the context of uncensored data, Huang et al. (2009) proposed a group bridge approach that is capable of simultaneous selection at both the group and within-group individual variable levels. Their approach uses a specially designed group bridge penalty. We speculate that for variable selection in the BS model with censored data, the group bridge has superior performance in group and individual variable selection to the LASSO and elastic-net methods. For some commonly used survival models, this has been shown to be true, for example, in the Cox proportional hazards model investigated by Huang et al. (2014). We will investigate this issue for the BS model in our future research. (ii) From the simulation perspective, we may find alternative methods to deal with censored data other than the SEM algorithm. Since there are lots of iterations within SEM computation which needs a long time to get the results, more efficient approaches would be desirable. The ascent-based EM or ascent-based MCEM algorithm by Caffo, Jank, and Jones (2005) seems an ideal approach. However, like generalized EM or MCEM, the ascent-based EM or MCEM algorithm operates on the so-called Q-function or an approximate Q-function, which requires a calculation of the expected value of a complicated log-likelihood function, conditioning on the observed data and current estimates of the parameters. The variable selection could not be accomplished directly by the existing software package such as glmnet. In our future research, we will investigate the ascent-based EM or MCEM based on the Q-function of the BS model. A new computing method needs to be developed for this. (iii) From our simulated examples, we did not see a big difference between the elastic-net and the LASSO. Zou and Hastie (2005) included a typical example to show the important differences between these two methods. They generated groups of covariates with different levels of correlation. In this way, they got variables with high within-group correlations and low between-group correlations. Their simulation results showed that elastic-net performed better than LASSO and behaved with “oracle”-like properties. There may be also typical examples to compare elastic-net and LASSO in our situation. All these are worthy of further investigation in our future research.

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and two anonymous referees for their constructive comments that have greatly improved an earlier version of this article. The authors acknowledge with gratitude the support for this research via Discovery Grants from National Science and Engineering Council (NSERC) of Canada. They are grateful to Professor Maronna for providing them the vessel data.

[Received October 2014. Accepted December 2015.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” Journal of the Royal Statistical Society, Series B, 57, 289–300. [387]
Birnbaum, Z., and Saunders, S. C. (1969), “A New Family of Life Distributions,” Journal of Applied Probability, 6, 319–327. [383]
Breiman, L. (1996), “Heuristics of Instability and Stabilization in Model Selection,” Annals of Statistics, 24, 2530–2583. [384]
Caffo, B. S., Jank, W., and Jones, G. L. (2005), “Ascent-Based Monte Carlo Expectation–Maximization,” Journal of the Royal Statistical Society, Series B, 67, 235–251. [391]
Chauveau, D. (1995), “A Stochastic EM Algorithm for Mixtures With Censored Data,” Journal of Statistical Planning and Inference, 46, 1–25. [384]
Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood From Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society, Series B, 39, 1–38. [384]
Desmond, A. F. (1982), Local Maxima of Stationary Stochastic Processes and Stochastic Modelling of Fatigue, unpublished Ph.D. Thesis, University of Waterloo. [383]
——— (1985), “Stochastic Models of Failure in Random Environments,” The Canadian Journal of Statistics, 13, 171–183. [383]
Desmond, A. F., Rodríguez-Yam, G. A., and Lu, X. (2008), “Estimation of Parameters for a Birnbaum–Saunders Regression Model With Censored Data,” Journal of Statistical Computation and Simulation, 78, 983–997. [384]
Diebolt, J., Ip, E., and Olkin, I. (1994), “A Stochastic EM Algorithm for Approximating the Maximum Likelihood Estimate,” Technical Report 301, Department of Statistics, Stanford University, California. [384,386]
Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least Angle Regression,” The Annals of Statistics, 32, 407–499. [384]
Engelhardt, M., Bain, L. J., and Wright, F. (1981), “Inferences on the Parameters of the Birnbaum–Saunders Fatigue Life Distribution Based on Maximum Likelihood Estimation,” Technometrics, 23, 251–256. [383]
Fan, J. Q., and Li, R. Z. (2001), “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” Journal of the American Statistical Association, 96, 1348–1360. [384,391]
Gorst-Rasmussen, A., and Scheike, T. (2011), “Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model,” Technical report, Department of Mathematical Sciences, Aalborg University. [390]
Hoel, A. E., and Kennard, R. W. (1970), “Ridge Regression: Biased Estimation of Regression Coefficients and Its Oracle Properties,” Technometrics, 12, 55–67. [384]
Hu, Y., and Lian, H. (2013), “Variable Selection in a Partially Linear Proportional Hazards Model With a Diverging Dimensionality,” Statistics and Probability Letters, 83, 61–69. [384]
Huang, J., Liu, L., Liu, Y., and Zhao, X. (2014), “Group Selection in the Cox Model With a Diverging Number of Covariates,” Statistica Sinica, 24, 1787–1810. [391]
Huang, J., Ma, S., and Xie, H. (2006), “Regularized Estimation in the Accelerated Failure Time Model With High-Dimensional Covariates,” Biometrics, 62, 813–820. [384]
——— (2009), “A Group Bridge Approach For Variable Selection,” Biometrika, 96, 339–355. [391]
Janssens, K. H., Deraedt, I., Schalm, O., and Veekmans, J. (1998), Composition of 15–17th Century Archaeological Glass Vessels Excavated in Antwerp, Belgium: Springer. [384,389]
Leiva, V., Barros, M., Paula, G. A., and Galea, M. (2007), “Influence Diagnostics in Log-Birnbaum–Saunders Regression Models With Censored Data,” Computational Statistics & Data Analysis, 51, 5694–5707. [387,390]
Lv, J., and Fan, Y. (2009), “A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares,” *The Annals of Statistics*, 37, 3498–3528. [391]

Maronna, R. A. (2011), “Robust Ridge Regression for High-Dimensional Data,” *Technometrics*, 53, 44–53. [389]

Ng, H., Kundu, D., and Balakrishnan, N. (2003), “Modified Moment Estimation for the Two-Parameter Birnbaum–Saunders Distribution,” *Computational Statistics & Data Analysis*, 43, 283–298. [383,387]

Rieck, J. R. (1989), “Statistical Analysis for the Birnbaum–Saunders Fatigue Life Distribution,” Unpublished Ph.D. Thesis, Department of Mathematical Statistics, Clemson University. [387]

Rieck, J. R., and Nedelman, J. R. (1991), “A Log–Linear Model for the Birnbaum–Saunders Distribution,” *Technometrics*, 33, 51–60. [383,385,386,388]

Saunders, S. C., and Birnbaum, Z. (1968), “Estimation for a Family of Life Distributions With Applications to Fatigue,” Technical report, DTIC Document. [383]

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003), “Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Datasets,” *Proceedings of the National Academy of Sciences*, 100, 8418–8423. [384,390]

Tibshirani, R. (1996), “Regression Shrinkage and Selection via the LASSO,” *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [384]

Tregouet, D., and Tiret, L. (2004), “Cox Proportional Hazards Survival Regression in Haplotype-based Association Analysis Using the Stochastic-EM Algorithm,” *European Journal of Human Genetics*, 12, 971–974. [384]

Wang, Z., Desmond, A., and Lu, X. (2006), “Modified Censored Moment Estimation for the Two-Parameter Birnbaum–Saunders Distribution,” *Computational Statistics & Data Analysis*, 50, 1033–1051. [383]

Zou, H., and Hastie, T. (2005), “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [384,386,389,391]