International Journal of Probability and Statistics 2016, 5(1): 10-17
DOI: 10.5923/j.ijps.20160501.02

Performance Assessment of Penalized Variable Selection Methods Using Crop Yield Data from the Three Northern Regions of Ghana

Smart A. Sarpong[1,*], N. N. N. Nsowah-Nuamah[2], Richard K. Avuglah[3], Seungyoung Oh[4], Youngjo Lee[4]

[1]Institute of Research, Innovations and Development - IRID, Kumasi Polytechnic, Ghana
[2]Kumasi Polytechnic, Ghana
[3]Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[4]Department of Statistics, Seoul National University, South Korea

Abstract Time and money can be saved by measuring only relevant predictors. Measuring relevant predictors also ensures a noise free estimation of parameters and preserves some degrees of freedom for a given predictive model. By comparing the performance of SCAD, LASSO and H-Likelihood, this study seeks to select among access to credit, training, study tour, demonstrative practicals, networking events, post-harvest equipments, size of plot cultivated and number of farmers; variables (fixed and interaction terms) that significantly influences crop yield in the three Northern regions of Ghana. Our simulation as well as real life results gives evidence to the fact that H-Likelihood method of penalized variable selection performs best followed by SCAD, with LASSO coming last. It does both selection of significant variables and estimation of their coefficients simultaneously with the least penalize cross-validated errors compared to the SCAD and the LASSO. The study therefore recommends that deliberate efforts be put into strengthening the Agricultural support systems as a form of strategy for increasing crop production in Northern Ghana.

Keywords Penalized, SCAD, LASSO, H-likelihood, PCVE, Variable Selection, Crop yield

1. Introduction

The rate of food production in many parts of sub-Saharan Africa has not kept pace with the rate of population growth. Whereas the estimates of population growth rate increase at about 3 per cent annually, that of food production increases by only 2 per cent (Rosegrant et al., 2001). The sub-region's per capita deficit in grains and cereals according to Rosegrant et al., (2001) is one of the highest in the world. Way back in 1967, the sub-region's cereal imports was 1.5 million tons. However, just within thirty years down the way, this figure increased to 12 million tons in 1997, and projections have it that the sub-region will require about 27 million tons of cereal imports to satisfy demand by 2020 (Rosegrant et al., 2001). In the long run, importation may not be economically feasible to ameliorate food shortages. Thus, there is a need to increase domestic production to guarantee food security.

Ghana is still recognized worldwide as an agriculture-based economy. Agriculture has been the anchor of Ghana's economy throughout post-independence history (McKay and Aryeetey, 2004). While policy and political instability had induced the fall of per capita GDP growth until 1980s, the agricultural sector had been less affected due to less interventions by the government compared to the non-agricultural sector and the fact that its growth is mainly led by smallholder farmers for subsistence purposes. Beyond 1992 when Ghana gained the forth republican political stability, there has been a rapid growth in the nonagricultural sectors; expanding by an average rate of 5.5 percent annually, compared to 5.2 percent for the entire economy (Bogetic et al., 2007).

The analysis presented in this paper suggests that a system of support services; Access to credit facility, Training, Study tour, Demonstrative practical, Networking events and Post-harvest Equipments, plays an important role in determining crop yields even though their individual and interaction effects on yield is not uniform across farmer base organizations. This research focuses primarily on the production of Maize and Soy beans in northern region of Ghana where there exists considerable farming activity. Maize and Soy beans are the very much cultivated in these parts of the country due to their vegetation which supports the growth of grains and cereals. Beyond the numbers and descriptive statistics on yield of such crops, this study tries to

* Corresponding author:
smartsarpong2015@gmail.com (Smart A. Sarpong)
Published online at http://journal.sapub.org/ijps

bring out variables that significantly contribute to yield. We seek to select among access to credit, training, study tour, demonstrative practicals, networking events, post-harvest equipments, size of plot cultivated and number of farmers; variables (fixed and interaction terms) that significantly influences crop yield in the three Northern regions of Ghana.

2. Variable Selection

Variable selection is aims at choosing the "optimum" subset of predictors. If a model is to be used for prediction, time and/or money can be saved by measuring only necessary predictors. Redundant predictors will add noise to the estimation of other quantities of interest and also lead to loss of some degrees of freedom. Choosing which predictors amongst many potential ones to be included in a model is one of the central challenges in regression analysis. Traditionally, stepwise selection and subset selection are the main means of variable selection for many years. Unfortunately, these methods are unstable and ignores the stochastic errors projected by the selection process.

Many techniques, including ridge regression, least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), elastic net (EN) (Zou and Hastie 2005), and adaptive lasso (A-LASSO) (Zou 2006) have been projected to select variables and estimate their regression coefficients simultaneously. All these techniques have common advantages over the traditional selection methods; they are computationally simpler, and the derived distributed estimators are stable, and they enhance higher prediction accuracies. These techniques can be cast in the frame of penalized least squares and likelihood. The central benefit of those techniques is that they choose vital variables and estimate their regression coefficients at the same time. In this paper, the H-likelihood (Lee and Oh, 2009) is projected for its special ability to produce penalty functions for variable selection, allow an oracle variable selection and at the same time improve estimation power.

In Agricultural and particularly crop yield analysis, variable selection for decision making in is mostly guided by expert opinion. Very few studies have tried to statistically evaluate these methods in decision making, or to indicate how they might be made better. Variable selection is especially central in the interpretation of Statistical models, particularly, when the actual fundamental model has a distributed representation. Identifying null predictors improves the prediction performances of the fitted model. Notwithstanding, traditional variable selection techniques have two important challenges. First, when the number of predictors $\boldsymbol{d}$ is large, it is computationally infeasible to perform subset selection. Secondly, subset selection is extremely unreliable due to its inherent discreteness (Breiman, 1996; Fan and Li, 2001).

To overcome these difficulties, several other penalties have been proposed. The L_2-penalty yields ridge regression estimation, but it does not perform variable selection. With the L_1-penalty, specifically, the penalized least squares (PLS) estimator becomes the least absolute shrinkage and selection operator (LASSO), (Tibshirani, 1996). LASSO is a common method for simultaneous estimation and variable selection, ensuring high prediction accuracy, and enabling the discovery of relevant predictive variables. Donoho and Johnstone (1994) selected significant wavelet bases by thresholding based on an L_1-penalty. Prediction accuracy can sometimes be improved by shrinking (Efron and Morris, 1975) or setting some coefficients to zero by thresholding (Donoho and Johnston, 1994).

LASSO has been criticized on the grounds that a single parameter λ is used for both variable selection and shrinkage. It results in choosing a model with too many variables to forestall over shrinkage of the regression coefficients (Radchenko and James, 2008); otherwise, regression coefficients of the chosen variables are often over shrunken. To surmount this challenge, Fan and Li (2001) proposed a variable selection method based on a non-concave penalized likelihood approach called the smoothly clipped absolute deviation (SCAD) penalty. These methods are distinct from traditional techniques of variable selection in that they remove insignificant variables by estimating their coefficients as 0.

Consequently, their approaches simultaneously select significant variables and estimate regression coefficients. Recent related studies include [Fan and Li 2006, Leng et.al, 2006, Zou and Li, 2008]. More recently, Zou (2006) showed that LASSO does not satisfy Fan and Li's (2001) oracle property, and proposed the adaptive LASSO. We demonstrate how the h-likelihood methods overcome such difficulties to allow an oracle variable selection and at the same time improve estimation power.

3. Methods

The idea of penalization was initially introduced in the context of solving integral equation numerically by Tikhonov (1943). As is well known, if $f \in L_2(\mathbb{R})$ and $K(x, y)$ is a smooth kernel, the range of the operator A, $R(A)$, A: $L_2(\mathbb{R}) \rightarrow L_2(\mathbb{R})$ with $(A f)(y) \equiv \int K(x, y) f(x)dx$ is dense in $L_2(\mathbb{R})$ but not onto. Hence, the inverse A^{-1} is ill-posed. The solution to the equation

$$\mathrm{A\ f = g} \tag{1}$$

is difficult to evaluate since approximations to g easily lie beyond $R(A)$. Tikhonov's solution was to replace (1) by the minimization of

$$\|Af - g\|^2 + \gamma W(f)$$

where the Tikhonov's factor $\gamma > 0$ is a regularization parameter and $W(f)$ is a smoothness penalty such as $\int [f'(x)]^2 dx$. Numerical (finite dimensional) approximations to this problem are more stable. Note that unless $\gamma = 0$, the solution will not satisfy (1).

Generally, regularization is the class of methods required

to develop the maximum likelihood to give valid answers in volatile situations. There is a great amount of work in statistics relating to regularization in a broad scope of problems. A thorough survey is beyond the scope of this paper. The central characteristic of most recent data has to do with both size and complexity. The size may allow us to non-parametrically estimate quantities which are 'unstable' and 'discontinuous' rudimentary functions of the distributions of the data, with the density being a typical instance. Complexity of the data, which usually relates to high dimensionality of observations, makes us attempt more and more complex models to accommodate the data. The fitting of models with a large number of parameters is also inherently unstable (Breiman, 1996). Both of these features, compel us to regularize in order to get sensible procedures. For recent discussions of these issues from different aspects, see Donoho (2000) and Fan and Li (2006). We will consider and relate the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and H-Likelihood (Lee and Nelder, 2009).

3.1. Least Absolute Shrinkage and Selection Operator (LASSO)

We consider the setting where we have observed data $(y_1, x_1),\ldots, (y_n, x_n)$ with each y_i a realization of a scalar random variable Y_i, and each $x_i = (x_{i1}, \ldots, x_{ip})^T$ a p-vector of explanatory variables. Let X be a matrix with ith row given by x_i^T. Without loss of generality, we shall require that the columns of X are centred. We assume that

$$Y_i = \mu + (X\beta)_i + \varepsilon_i, \quad (2)$$

where each ε_i is i.i.d $N(0, \sigma^2)$. In the classical linear model, we would assume X has full column rank, and so $p < n$. The tuning parameter λ controls the sparsity of the estimate, with large values of λ resulting in estimates with many components set to 0. Unfortunately, this optimization problem is hard, and to the best of our knowledge, it is computationally intractable for $p > 50$.

The Lasso (Tibshirani, 1996) solves the related problem:

$$\left(\hat{\mu},\ \hat{\beta}(\lambda)\right) = argmin_{m,b}\{\tfrac{1}{2n}\ \|Y - m - Xb\|^2 + \lambda\|b\|_1\} \quad (3)$$

The non-differentiability of the l_1 norm at 0 ensures that the resulting estimator is sparse, and its convexity makes the overall optimization problem convex. There exist very efficient algorithms for solving this problem, even when $p >$ 105 (see for example the R package glmnet).

3.2. The Smoothly Clipped Absolute Deviation (SCAD)

We again analyze the setting where we have (X_i, Y_i), $i = 1, \ldots, n$, as n observations satisfying

$$Y_i = \beta_0 + X_i'\beta + \varepsilon_i,\ i = 1, \ldots, n \quad (4)$$

where $Y_i \in R$ is a response variable, X_i is a $p_n \times 1$ covariates vector and ε_i has mean 0 and variance σ^2. Here the superscripts are used to make it explicit that both the covariates and parameters may change with n. For simplicity, we assume $\beta_0 = 0$. In sparse models. the p_n covariates can be classified into two categories: the important ones whose corresponding coefficients are non-zero and the trivial ones whose coefficients are zero. For notational convenience, we write

$$\beta = (\beta_1', \beta_2')' \quad (5)$$

where $\beta_1' = (\beta_1, \ldots, \beta_{kn})$ and $\beta_2' = (0, \ldots, 0)$. Here $k_n\ (\leq p_n)$ is the number of non-trivial covariates. Let $m_n = p_n - k_n$ be the number of zero coefficients. Let $Y = (Y_1, \ldots, Y_n)'$ and let $X = (X_{ij},\ 1 \leq i \leq n, 1 \leq j \leq pn)$ be the $n \times p_n$ design matrix. According to the partition of β, write $X = (X_1, X_2)$, where X_1 and X_2 and $n \times k_n$ and $n \times m_n$ matrices, respectively. Given $a > 2$ and $\lambda > 0$, the SCAD penalty at θ is

$$p\lambda(\theta; a) = \begin{cases} \lambda|\theta|, & |\theta| \leq \lambda, \\ -\frac{\theta^2 - 2a\lambda|\theta| + \lambda^2}{[2(a-1)]}, & \lambda < |\theta| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & |\theta| > a\lambda \end{cases} \quad (6)$$

More insight into it can be gained through its first derivative:

$$p'\lambda(\theta; a) = \begin{cases} sgn(\theta)\lambda, & |\theta| \leq \lambda, \\ sgn(\theta)(a\lambda - |\theta|)/(a-1), & \lambda < |\theta| \leq a\lambda, \\ 0, & |\theta| > a\lambda \end{cases} \quad (7)$$

The SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but not differentiable at 0. Its derivative vanishes outside $[-a\lambda, a\lambda]$. Consequently, SCAD penalized regression can produce distributed solutions and unbiased estimates for large coefficients. More in dept analysis of this penalty can be found in Fan and Li (2001). The penalized least squares objective function for estimating β with the SCAD penalty is

$$Q_n(b;\lambda_n, a) = \|Y - Xb\|^2 + n\sum_{j=1}^{P_n} p\lambda_n(b_j; a) \quad (8)$$

where $\|.\|$ is the L_2 norm. Given penalty parameters λ_n and a, the LS-SCAD estimator of β is

$$\hat{\beta}_n \equiv \hat{\beta}(\lambda_n; a) = \arg\min Q_n(b; \lambda_n, a) \quad (9)$$

We write $\hat{\beta}_n = (\beta_{1n}', \beta_{2n}')'$ the way we partition β into β_1 and β_2.

3.3. Variable Selection Using the Penalized H-Likelihood

In this section, we consider variable selection of fixed effects β by maximizing a penalized profile h-likelihood hp using a weight $h_\omega^*(\beta, v, \theta)$ and a penalty defined by

$$h_p(\beta, v, \theta) = h_\omega^* - n\sum_{j=1}^{P} J_\gamma(|\beta_j|) \quad (10)$$

where $J\gamma\ (|\cdot|)$ is a penalty function that controls model complexity using the tuning parameter γ. Normally, setting $\gamma = 0$ result in a sub-hazard frailty model, whereas the regression coefficient estimates β approaches 0 as $\gamma \to \infty$ is inclined to choose a complex model (Fan and Lv, 2010).

Although several penalty functions have been applied in the literature section, (Fan and Li, 2001, 2002; Fan and Lv, 2010), this paper primarily analyze the following three penalty functions.

- LASSO (Tibshirani, 1996)

$$J\gamma(|\beta|) = \gamma|\beta| \quad (11)$$

• SCAD (Fan and Li, 2001)

$$J'_\gamma(|\beta|) = \gamma(|\beta| \le \gamma) + \frac{(a\gamma - |\beta|)}{a-1} I(|\beta| > \gamma) \quad (12)$$

where a = 3.7 and x_+ denotes the positive part of x, i.e. x_+ is x if $x > 0$, zero otherwise.

• HL (Lee and Oh, 2009)

$$J_\gamma(|\beta|) \equiv J_{(a,b)}(|\beta|) = log\Gamma\left(\frac{1}{b}\right) + \frac{logb}{b} + \frac{\beta^2}{2au(|\beta|)} + \frac{(b-2)logu(|\beta|)}{2b} + \frac{u(|\beta|)}{b}, \quad (13)$$

where u(|β|) = [{$8b\beta^2/a + (2 + b)^2$ }$^{1/2}$+ 2 − b]/4

An acceptable penalty function should provide estimates that satisfy unbiasedness, sparcity, and continuity (Fan and Li, 2001, 2002). The LASSO penalty in (11) is as general as L_1 penalty, but it does not satisfy these three properties at the same time. Fan and Li (2001) proved that SCAD in (12) meet all the three properties and that it can perform the orcale process in terms of choosing the correct subset model and estimating the true non-zero coefficient, at the same time.

Lee and Oh (2009) proposed a new penalty unbounded at the origin in the structure of a random effect model. The form of the penalty changes from a quadratic shape (b = 0) for ridge regressions to a cusped form (b = 2) for LASSO and then to an unbounded form (b > 2) at the origin. In the case of b = 2, it allows for an uncountable number of gains at zero. The SCAD provides oracle ML estimates (least squares estimators), whereas the HL provides oracle shrinkage estimates. When multi-collinearity exists, shrinkage estimation becomes much better than the ML estimation. Lee et al. (2010, 2011a,b) has shown the importance of the HL approach over LASSO and SCAD methods, with respect to the number of covariates being larger than the sample size (i.e $p > n$). In reality it has an attribute for a variable selection without losing prediction power. Since in (13) it has a greater sensitivity to alter the penalty than b, we also analyze only a few values for b, e.g. b = 2.1, 3, 10, 30, 50 denoting small, medium and large.

3.4. Penalized H-likelihood Procedure

By maximizing the penalized h-likelihood h_p in (10), we need to analyze the variable and estimate their related regression coefficients at the same time. In other words, those variable whose regression coefficients are estimated as zero are automatically removed. To accomplish this goal, by applying h_p, the estimation process of the fixed parameters (β, θ) and random effects ν are needed. First, the maximum penalized h-likelihood (MPHL) estimates of (β, ν), are obtained by finding the joint estimating of β and ν:

$$\frac{\partial h_p}{\partial \beta_j} = \frac{\partial h_w^*}{\partial \beta_j} - n\sum_{j=1}^{P} [J_\gamma(|\beta_j|)]' = 0 \quad (14)$$

and

$$\frac{\partial h_p}{\partial \beta_v} = \frac{\partial h_p^*}{\partial v} = 0 \quad (15)$$

Notice that (14) is an altered estimating equation evoked by adding the penalty term, but (15) is similar to the standard estimating equation without penalty. Notwithstanding, for the three penalty functions considered in (11)-(13), $J\gamma$ in (14) becomes non-differentiable at the origin and it does not have continuous second-order derivatives. To elicit this challenge in solving (14) we use local quadratic approximation (referred to as LQA, Fan and Li, 2001) to such penalty functions. That is, given an initial value of β_0 approaching the true value of β, the penalty function $J\gamma$ can be locally approximated by a quadratic function as

$$[J_\gamma(|\beta_j|)]' = J'_\gamma(|\beta_j|)sgn(|\beta_j|) \approx \{J'_\gamma(|\beta_j^0|)/[|\beta_j^0|]\}\beta_j \quad \text{for } \beta_j \approx \beta_j^0 \quad (16)$$

Then the negative Hessian matrix of β and ν founded on h_p can be explicitly written as a simple matrix from (Ha and Lee, 2003):

$$H(h_p; \beta, v) = \frac{\partial^2 h_p}{\partial(\beta, v)^2} = \begin{pmatrix} X^T W * X + n\sum_\gamma & X^T W * Z \\ Z^T W * X & Z^T W * Z + U \end{pmatrix} \quad (17)$$

Where $\sum_\gamma = diag\{J'_\gamma(\beta j)/\beta j\}$. Here X *and* Z are $n \times q$ and $n \times q*$ model matrices for β and v whose ijth row vectors are X_{ij}^T and X_{ij}^T respectively, $W^* = W^*(\beta, v) = -\partial^2 h_w/\partial\eta^2$ is a form of the symmetric matrix given in Appendix 2 of Ha and Lee (2003) and Ha et al. (2013) $\eta = X\beta + Zv$ and $U = -\partial^2 l_2/\partial v^2$ is a $q^* \times q^*$ matrix that takes a form of $U = BD(\sum^{-1}, ..., \sum^{-1})$ if $v \sim N(0, \sum)$, where $q^* = q \times r$ and BD(·) represents a block diagonal matrix.

Following Ha and Lee (2003) and (15), it can be observed that given θ, the MPHIL estimates of (β, ν) are obtained from the following scores equations:

$$\begin{pmatrix} X^T W^* X + n\sum_\gamma & X^T W^* Z \\ Z^T W^* X & Z^T W^* Z + U \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X^T w \\ Z^T W^* + R^* \end{pmatrix} \quad (18)$$

where $w^* = W^*\eta + (\delta - \mu)$ with $\mu = \exp(\log w + \log \Lambda_{01}^s + \eta)$ and $R^* = Uv + (\partial l/\partial v)$. Here w is the weight w_{ij} and Λ_{01}^s is the baseline cumulative sub-hazard function. The scores of equations (16) are extensions of the already existing estimation processes. For instance, under no penalty (i.e. γ) they become the score equations of Ha et al. (2003) for the standard sub-hazard frailty models, for variable selection under the Fine-Gray model (1999) without frailty. They also change to

$$(X^T W^* X + n\sum_\gamma)\ \hat{\beta} = X^T w^*, \quad (19)$$

suggesting that the new equations (16) allow a special case of the penalized equation (17) for the Fine-Gray model. Observe that, to refrain from some numerical complications, we apply $\sum_{\gamma,\varepsilon} = dia\{J'_\gamma(|\beta_j|)/(|\beta_j| + \varepsilon)\}$, in solving (16) for a small non-negative value of ε (e.g. $\varepsilon = 10^{-8}$), rather than $\sum_\gamma$, to assert the existence of $\sum_{\gamma,\varepsilon}$ (Lee and Oh, 2009). So far as ε is small non-negative value, the diagonal component of $\sum_{\gamma,\varepsilon}$ are similar to that of $\sum_\gamma$. As a matter of fact, this algorithm is similar to that of Hunter and Li (2005) for modifying the LQA; see also Johnson et al. (2008). In this paper, we report $\hat{\beta} = 0$ if all five printed decimals are zero. In the case of, SCAD and HL penalties, there exist many local maximums. Hence, an acceptable initial value is vital to get a proper estimate β. In this paper, a LASSO solution will

be applied as the initial value for the SCAD and HL penalties.

Consequently, we apply an adjusted profile h-likelihood $p_\tau(h_p)$ for the estimation of θ (Ha and Lee, 2003; Lee et al., 2006) which removes (β, *v*) from h_p in (11), defined by

$$p_\tau(h_p) = [h_p - \frac{1}{2}\log det\{\frac{H(h_p;\tau)}{2\pi}\} \quad (20)$$

where $\tau = (\beta^T, v^T)^T$ and $\tau^T = \tau^T(\theta) = ((\hat{\beta}^T(\theta), \hat{v}^T(\theta))^T$. By solving the score equations $\partial p_\tau(hp)/\partial\theta = 0$ as in Ha et al. (2013), the estimates of θ are found. Consequently, we observe that the projected process is easily implemented by a little change to the already existing h-likelihood process (Ha and Lee, 2003; Ha et al., 2011, 2013).

4. Results

4.1. Simulation Analysis

In this section, the performance of the HL method is examined through simulated data, and compared to the LASSO and SCAD. For each method we select optimal tuning parameters that maximize the log-likelihood obtained from an independent validation dataset of size *n* /2, where *n* is the size of the training set. We varied the number of covariates (*p*) and fixed coefficients (*q*) in two simulations. In one simulation we use *n* = 200 while in the other *n* = 100.

For the simulation studies, we consider the following GLM:

$$y/x \sim N(\mu(X'\beta), 2)$$

with linear link function $\mu(X'\beta) = X'\beta$ where the linear predictor $X'\beta = \sum_{j=1}^{-pk} x_{jk}\beta_{jk}$ consist of *p* covariates. To generate covariate X'_{jk} s, we first generate $p = \sum_{j=1}^{k} p_k$ random variables X'_{jk} s independently from the standard normal distribution. Then z'_k s are simulated with a multivariate normal distribution. The covariate X'_{jk} s are generated from

$$X_{kj} = (z_k + \varepsilon_{kj})/\sqrt{2}k = 1,\ldots, p_k \quad (21)$$

where $z = (z_1,\ldots,z_k)^T \sim N(0, \sum)$ with covariance structure $\sum_{kl} = cov(z_k, z_l) = 0.5^{|k-l|}$ and $\varepsilon_{kj} \sim N(0, I_p)$ that of independent of *z*. The true non-zero coefficients are

$$\beta_{kj} = c/j, j = 1, \ldots, q_k, k \neq A$$

where q_k is the number of non-zero coefficients in the *k*th group, and *A* is the set of the non-null groups.

A group is said to be non-null if at least one coefficient in the group is estimated to be non-zero. The constant *c* is chosen so that the signal-to-noise ratio is equal to 5 in the linear model. For each model setting we consider one dimensionality level only, the one with $p < n$. So, overall we have 4 simulation scenarios, where each is replicated 100 times with sample size $n = 200$ and $n = 100$. The cross validation errors which are defined based on independent test sample of size $N = 5000$ forms the basis for performance comparison. For variable selection quality, cross validation errors for the three methods are compared and the method with the smallest penalized cross validation errors is preferred. The results are shown in table below. The HL estimator performs better than the other methods for prediction accuracy as evident by its smallest cv errors comparative to the other methods.

Table 1. Comparative simulation results for penalized variable selection methods

N=100 P= 10 Q =3										
Method	sim 1	sim 2	sim 3	sim 4	sim 5	sim 6	sim 7	sim 8	sim 9	sim 10
LASSO	12.078814	9.154382	8.09983	8.075933	11.001972	11.456325	13.042181	12.49598	8.606529	10.729943
SCAD	12.161675	9.304411	8.193715	8.134833	11.079497	11.558803	13.031146	12.53982	8.662975	11.357576
H-L	11.424703	8.989876	7.700814	9.586048	11.239171	10.810527	12.891827	10.58886	8.532304	10.799194
N=200 P= 10 Q = 3										
LASSO	16.08259	23.72576	18.89797	25.91064	18.28419	22.8221	21.5849	18.24477	24.7486	19.22165
SCAD	16.12418	23.67929	18.88515	25.92822	18.22037	22.82895	21.63649	18.31884	24.75796	19.19523
H-L	15.86426	22.57885	18.33407	23.89222	18.02888	21.61746	20.70557	18.04133	25.58047	18.44424
N=100 P= 8 Q =5										
LASSO	12.477717	9.976822	9.567962	7.672933	11.656413	13.577737	10.798311	16.09166	14.212732	12.474876
SCAD	12.664895	10.003768	9.849049	7.660239	11.845172	13.811124	10.989915	17.00159	14.318873	12.343221
H-L	11.912251	9.440881	9.317837	7.763213	11.043935	16.451458	10.599491	13.9028	13.701678	11.351132
N=200 P= 8 Q =5										
LASSO	23.62724	21.68409	22.38465	20.89009	17.62312	18.37307	22.97981	21.5812	26.90833	25.05369
SCAD	23.64128	21.64646	22.44532	20.86011	17.91219	18.43819	22.95806	21.53924	27.03923	25.16297
H-L	22.78054	21.70741	22.26482	20.29142	17.29375	18.20228	21.64576	21.33577	25.43343	24.61937

4.2. Real Data Analysis (Crop yield data)

We analyze the crop yield datasets obtained from 2013 main season yield measured in kilograms. The data consists of a numeric response variable (yield) and 9 covariates obtained from 790 farmer based organizations in the three northern regions of Ghana. We excluded 10 observations (FBO's) due to missing values. The dataset has 7 categorical covariates crop type (Maize or Soybean), Financial Credit (Acquired or Not), Training (Acquired or Not), Study tour (Acquired or Not), Demonstrative Practical (Acquired or Not), Networking Events (Acquired or Not), Post-harvest Equipment (Acquired or Not)) and 2 continuous variables, including plot size in acres and number of farmers. Beside these 9 fixed effects, 36 two-way interaction terms are also generated as fixed interaction terms. This brings the total number of fixed covariates to 45. To allow possible non-linear effects, a third-degree polynomial is used for each continuous covariate, and dummy variables are used for categorical variables.

The results are obtained by 100 random segments of the data set divide into training (70 percent) and test sets (30 percent). For each random segment, the tuning parameters are chosen by the 10-fold cross validation in the training set, and the prediction errors are calculated on the test set. Table 3 presents averages of cv errors, the number of significant variables and number of insignificant variables.

The HL estimator performs better than the other methods for prediction accuracy as evident by its smallest cv errors comparative to the other methods.

Table 2. Standardized Penalized Coefficients of Crop Yield Data

Selected Variables	LASSO	SCAD	H-L
Crop	-1.39	-2.38	-0.74
Credit			1.23
Training	0.82	1.91	2.29
Study Tour			
Demo. Practical	3.65	5.14	5.04
Networking Events			0.69
Post harvest Equipment	-5.56	-7.21	-7.21
No. of farmers	0.29		1.88
plot size	11.53	12.42	12.83
Crop*Credit	-1.72	-2.41	-2.74
Crop*Training			-0.67
Crop*Study Tour	0.86	1.47	1.53
Crop*Demo. Practical	-1.68	-2.76	-2.63
Crop*Networking Events			
Crop*Post-harvest Equipment	2.74	4.60	4.50
Crop*No. of farmers	-0.96		-2.11
Crop*plot size			
Credit*Training			
Credit*Study Tour	-0.963	-1.14	-1.23
Credit*Demo. Practical	0.20		
Credit*Networking Events	0.89	1.37	1.34
Credit*Post-harvest Equipment			
Credit*No. of farmers	-0.74		-1.5
Credit*plot size	2.80	2.30	2.86
Training*Study Tour	0.81	0.62	0.11
Training*Demo. Practical	-1.10	-1.41	-1.33
Training*Networking Events	-0.06		-0.69
Training*Post-harvest Equipment	0.32		0.48
Training*No. of farmers	-1.91	-3.11	-1.97
Training*plot size	-0.53		-0.82
Study Tour*Demo. Practical	0.14		0.13
Study Tour*Networking Events			
Study Tour*Post-harvest Equipment	-1.38	-1.27	-1.22
Study Tour*No. of farmers	0.46	0.02	1.08
Study Tour*plot size	-0.14		-0.74
Demo. Practical*Networking Events	-0.29		-0.46
Demo. Practical*Post-harvest Equipment	1.65	1.68	1.69
Demo. Practical*No. of farmers			
Demo. Practical*plot size	-3.53	-3.89	-3.58
Networking Events*Post-harvest Equipment			
Networking Events*No. of farmers			
Networking Events*plot size	-1.04	-2.04	-2.09
Post-harvest Equipment*No. of farmers			-0.38
Post-harvest Equipment*plot size	4.43	4.59	5.06
No. of farmers*plot size			-0.67

Table 3. Performance of Penalized methods on Crop Yield Data

Method	LASSO	SCAD	H-L
No. of Significant Variables Selected	31	21	35
No. of Variables Ignored	14	24	10
Cross validated Errors	24.097	24.043	23.543

5. Discussion

In section 4.2, we sort to select significant variables amongst a number of latent ones to be included in the model through penalized methods. We have compare the sparsity and number of significant crop yield variables selected by the three penalized methods; LASSO, SCAD, and H-likelihood both through simulation studies and by the real data (See Table 1 and 2). These techniques have common benefits over the classical selection procedures; they are computationally easy, deriving sparse estimators that are stable, and they aid higher prediction accuracies. We have shown how to choose significant variables amongst common semi-parametric models via penalized methods. We have also shown through numerical studies and data analysis that the projected process with H-Likelihood performs best followed by SCAD, with LASSO coming last (See Table 1 and Table 3).

There has been a number of criticisms against LASSO with some reasons being that a single tuning parameter λ is utilized for variable selection and shrinkage. A model with too many variables is usually chosen to prevent over shrinkage of the regression coefficients (Radchenko and James, 2008); otherwise, regression coefficients of selected variables are often over-shrunken. This assertion is highly confirmed by the results of this in table 1.

To overcome this problem, a method known as the smoothly clipped absolute deviation (SCAD) penalty for oracle variable selection was proposed by Fan and Li (2001). More recently, Zou (2006) showed that LASSO does not satisfy Fan and Li's (2001) oracle property, and proposed the adaptive LASSO. Based on the findings of this study, we also propose the H-likelihood approach by Lee and Nelder (2009), as the best in crop yield variable selection and we do so on the basis that, compared to other forms of penalized methods ie. LASSO and SCAD, the H-likelihood approach (Lee and Nelder 2009) facilitates higher prediction accuracy since it has least estimated penalized cross validated errors (see table 2).

6. Conclusions

H-Likelihood method of penalized variable selection performs best followed by SCAD, with LASSO coming last. It does both selection of significant variables and estimation of their coefficients simultaneously with the least penalize cross-validated errors compared to the SCAD and the LASSO. We recommend that a deliberate effort be put into strengthening the Agricultural support systems as a form of strategy for increasing crop production in Northern Ghana. Access to credit, training, access to post harvest equipments, access to demonstrative practicals and access to large plot size are the physical support services highly recommended by this study.

REFERENCES

[1] Bogetic, Y., M. Bussolo, X. Ye, D. Medvedev, Q. Wodon, and D. Boakye. 2007 (April). Ghana's growth story: How to accelerate growth and achieve MDGs? Background paper for Ghana Country Economic Memorandum. Washington, DC: World Bank.

[2] Breiman, L. (1996) Heuristics of instability and stabilization in model selection, Ann. Statist., 24, 2350-2383.

[3] Brent A Johnson, (2008). ''Penalized Estimating Functions and Variable selection in Semi parametric regressions models", Journal of the American Statistical Association., 12/2006, 350-383.

[4] Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81,425-455.

[5] Donoho, D. L. (2000). High dimensional data analysis: the curses and blessings of dimensionality. In Math Challenges of 21st Centuary (2000). American Mathematical Society. Plenary speaker. Available in: http://www.stat.stanford.edu/ donoho/Lectures/AMS2000/.

[6] Efron, B. and Morris, C. (1975) Data analysis using Stein's estimator and its generalizations, J. Am. Statist. Ass., 70, 311-319.

[7] Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, vol 96, 1348-1360.

[8] Fan, J. and Li, R. (2006) Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In Proc. of the Madrid International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, European Mathematical Society, Zurich, 595-622.

[9] Fan, J. and Lv, J. (2010), "A selective overview of variable selection in high dimensional feature space," Statistica Sinica, 20, 101-148.

[10] Fan, J. and Li, R. (2002), 'Variable selection for Cox's proportional hazards model and frailty model," The Annals of Statistics, 30, 74-99.

[11] FAO (Food and Agriculture Organization of the United

Nations). 2005. Fertilizer use by crop in Ghana. http://www.fao.org/docrep/0C (retrieved June 10, 2012).

[12] FtM AGRA grant Narrative report, 2011. Linking Farmers to Markets (FTM) Project. Prepared for the Alliance for a Green Revolution in Africa (AGRA) by IFDC, 2011.

[13] Gunter, L. 2011. "Variable selection for qualitative interactions" Statistical Methodology, 01-2011, 54-69.

[14] Ha, I. D. and Lee, Y. (2003), "Estimating frailty models via Poisson hierarchical generalized linear models," Journal of Computational and Graphical Statistics, 12, 663-681.

[15] H. Zou and T. Hastie, 2005. "Regularization and Variable Selection via the Elastic Net," Journal of the Royal Statistical Society Series B, Vol. 67, No. 2, 2005, pp 301-320.

[16] Hans C. van Houwelingen, Willi Sauerbrei, 2013. Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited. Open Journal of Statistics, 2013, 3, 79-102 (http://www.scirp.org/journal/ojs).

[17] Lee, Y. and Oh, H. S. (2009), "Random-effect models for variable selection," Department of Statistics, Stanford University, Technical report No. 2009-4, 1-24.

[18] Lee, D., Lee, W., Lee, Y. and Pawitan, Y. (2010), "Super sparse principal component analysis for high-throughput genomic data," BMC Bioinformatics, 11, 296.

[19] Lee, D., Lee, W., Lee, Y. and Pawitan, Y. (2011a), "Sparse partial least-squares regression and its applications to high-throughput data analysis," Chemometrics and Intelligent Laboratory Systems, 109, 1-8.

[20] Lee, Y. and Nelder, J.A. (1999). The robustness of the quasi-likelihood estimator. Canadian Journal of Statistics, 27, 321-327.

[21] Lee, W., Lee, D., Lee, Y. and Pawitan, Y. (2011b), "Sparse canonical covariance analysis for high-throughput data," Statistical Applications in Genetics and Molecular Biology, 10, 1-24.

[22] Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society B, 58, 619-656.

[23] Lee, Y, Nelder, J.A. and Noh, M. (2006). H-likelihood: problems and solutions. Statistics and Computing, revision.

[24] Leng C. L., Y. Lin and G. Wahba, (2006). "A note on LASSO and Related Procedures in Model Selection," Statistica Sinica, Vol. 16, 2006, pp. 1273-1284.

[25] McKay, A., and E. Aryeetey. 2004. Operationalizing pro-poor growth: A country case study on Ghana. A joint initiative of AFD, BMZ (GTZ, KfW Development Bank), DFID, and the World Bank. http://www.dfid.gov.uk/pubs/files/oppgghana.pdf.

[26] Radchenko, P. and James, G. (2008) Variable inclusion and shrinkage algorithms, J. Am. Statist. Ass., 103, 1304-1315.

[27] Rosegrant, M.W., Paisner, M.S., Meijer, S. and Witcover, J. 2001. Global food projections to 2020: emerging trends and alternative futures. International Food Policy Research Institute, Washington, DC.

[28] Tikhonov, A. N. (1943). On the stability of inverse problems. C. R. (Doklady) Acad. Sci. URSS (N.S.), 39:176179.

[29] Tibshirani, R. J. (1996) Regression shrinkage and selection via the LASSO, Journal of Royal Statist. Soc. B, 58, 267-288.

[30] Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418-1429.

[31] Zou, H., Li, R., 2008. One-step sparse estimates in non-concave penalized likelihood models. The Annals of Statistics 36, 1509-1533.

[32] Il Do Ha, Jianxin Pan, Seungyoung Oh & Youngjo Lee (2013): Variable Selection in General Frailty Models using Penalized H-likelihood, Journal of Computational and Graphical Statistics.

[33] Ha, I. D., Sylvester, R., Legrand, C. and MacKenzie, G. (2011), 'Frailty modelling for survival data from multicentre clinical trials," Statistics in Medicine, 30, 2144-159.

[34] Ha, II Do, Minjung Lee, Seungyoung Oh, Jong-Hyeon Jeong, Richard Sylvester, and Youngjo Lee (2014), "Variable selection in sub-distribution hazard frailty models with competing risks data", Statistics in Medicine, 50, 144-159.

[35] Johnson, B. A., Lin, D. Y. and Zeng, D. (2008), "Penalized estimating functions and variable selection in semi-parametric regression models," Journal of the American Statistical Association, 103, 672-680.