Abstract

Nowadays, $L_1$ penalized likelihood has absorbed a high amount of consideration due to its simplicity and well developed theoretical properties. This method is known as a reliable method in order to apply in a broad range of applications including high-dimensional cases. On the other hand, $L_1$ driven methods, precisely lasso dependent regularizations, suffer the loss of sparsity when the number of observations is too low. In this paper we address a new differentiable approximation of lasso that can produce the same results as lasso and ridge and also can produce smooth results. We prove the theoretical properties of the model as well as its computation complexity. Due to differentiability, proposed method can be implemented by means of the majority of convex optimization methods in literature. That means a higher accuracy in situations where true coefficients are close to zero that is a major issue of LARS. Simulation study as well as flexibility of the method show that the proposed approach is as reliable as lass and ridge and can be used in both situations.

Keywords: Lasso, Smooth approximation, Ridge regression,

Contents

1 Introduction 2
2 SmoothABS Regression 6
3 Theoretical properties of the model 8
4 Algorithm 10
5 Computations and complexity order 10
6 Simulation study 11
7 Conclusion 13
8 Appendix: Proofs 13

1 Mathematics department, Brunel University London, UK
hamed.haselimashhadi@brunel.ac.uk


1 Introduction

The idea of a smooth lasso than a solid sparse estimation is a new approach that came to literature by some authors like (Hebiri et al., 2011), (Hebiri, 2008), (Schmidt et al., 2007a) and (Zou and Hastie, 2005). The idea behind is considering a smooth loss function instead of a solid absolute function in lasso (Tibshirani, 1996). That means to add a differentiable term to the likelihood. Adding this term results in a trade off between sparsity and increasing the number of covariates (Pourahmadi, 2013, p 99). Elastic net (Zou and Hastie, 2005) and smooth lasso (Hebiri et al., 2011) are two well known examples of these kinds of models. On the other hand strictly smooth models like ridge regression (Hoerl and Kennard, 1970) suffers the enough sparsity and models like elastic net and smooth lasso require a challenging term and a separate tuning parameter to be regularized. In this paper we address a special case of penalized likelihoods that can be used in smooth situations as well. The main idea is to replace the absolute function in lasso with its smooth approximation. Latter means a convex and differentiable loss function that satisfies smoothness definition in (Ramirez et al., 2014). Proposed approximation let the analytic to choose from a very smooth function to a very sharp penalty term. Furthermore, it reduces the optimization problem to an ordinal minimization problem that can be implemented by a broad range of algorithms in literature without any extra effort.

For a brief introduction we consider the problem of approximating an absolute function, \( f(x) = |x| \). The absolute function can be approximated using at least three functions listed below

\[
|x| \approx \sqrt{x^2 + \epsilon}, \quad \epsilon \in \mathbb{R}_+, \quad (1)
\]

\[
|x| \approx \frac{2}{\sqrt{\pi}} x \int_0^{x/s} e^{-t^2} dt, \quad s \in \mathbb{R}_+, \quad (2)
\]

\[
\frac{x^2}{\sqrt{x^2 + u^2}} \leq |x| \leq \sqrt{x^2 + u^2}, \quad u \in \mathbb{R}_+. \quad (3)
\]

Equation (1) is studied in details in (Ramirez et al., 2014). The right hand side of (2) is so called error function and can be considered as a probability distribution (refer to (Olver et al., 2010) for a rich discussion of error function) and the accuracy of approximation increases as \( s \) tends to zero. Also,

\[
\lim_{x \to 0} \frac{2}{\sqrt{\pi}} x \int_0^{x/s} e^{-t^2} dt \to \frac{2}{\sqrt{\pi}} \frac{x^2}{s} e^{-\left(\frac{x}{s}\right)^2} = \frac{x^2}{s} \phi\left(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}\right),
\]
by setting $s = \frac{2}{\sqrt{\pi}}$, latter equation results in $\sqrt{\pi}x^2\phi(\frac{\sqrt{\pi}x}{2}, 0, \frac{1}{\sqrt{2}})$. On the other hand, $\phi(\frac{\sqrt{\pi}x}{2}, 0, \frac{1}{\sqrt{2}}) \approx \frac{1}{\sqrt{\pi}}$ and finally $\sqrt{\pi}x^2\phi(\frac{\sqrt{\pi}x}{2}, 0, \frac{1}{\sqrt{2}}) \approx x^2$ where is very close to $x^2$. Latter result, as illustrated in figure (1), shows that the approximation has a similar behaviour as $x^2$ for the values close to zero. In addition, to reduce computation time of $\frac{2}{\sqrt{\pi}} = 1.128379$ it is possible to set $s = 1$. On the other hand, for large values of $s$,

$$\lim_{s \to \infty} \frac{2}{\sqrt{\pi}} x \int_0^{x/s} e^{-t^2} dt = 0.$$ 

That is, a horizontal line.

Equation (1) is a special case of (3) and it is straightforward to prove that the length of (3) is always less than $u$ (Nesterov, 2005). Recently (Schmidt et al., 2007b) have proposed a smooth approximation of the form of $|x| \approx \frac{1}{\alpha}[\log(1 + e^{-\alpha x}) + \log(1 + e^{\alpha x})] = |x|_\alpha$ and its application in $L_1$ penalized likelihoods. Notice that latter approximation is twice differentiable, $|x| = \lim_{\alpha \to \infty} |x|_\alpha$ and the maximum error of this approximation is $||x| - |x|_\alpha| \leq 2\log(2)\alpha$. 

For simplicity we name $\epsilon, s, u$ and $\alpha$ as precision values. Figure (2)
compares the approximations for different precision values. It is obvious from this figure that the accuracy of the second approximation increases faster than other opponents at the same rate of changing precision values. Technically speaking, it converges at the speed faster than its opponents. In addition, approximation (2) passes the origin regardless of the value of precision value that is of interest when make a use of this approximation in loss functions. Flexibility, convergence and convexity of this function motivate us to make a use of this function in linear regularization problems to replace absolute value function.

The rest of the paper is organized as follows. After proving the smoothness of the functions, a penalized likelihood based on error function is defined in section §2. Some interesting theoretical properties of this model are studied in section §3. Then, a discussion about computation complexity of model in section §5 provides enough support to apply model in real situations. In section §6, a simulation study accompanies the theoretical results. Finally, conclusion in section §7 and detailed technical proofs are provided in the appendix.

To show the smoothness of the approximations, we adopt the definition of smoothness in (Ramirez et al., 2014) as following,

**Definition 1** A function \( f : \mathbb{R} \to \mathbb{R} \) is a smooth approximation of \(|x|\) if this function is differentiable and satisfies two limit properties as following.

\[
\lim_{x \to \pm \infty} \frac{f(x)}{|x|} = 1, \quad \lim_{x \to \pm \infty} \frac{f'(x)}{\text{sign}(x)} = 1.
\]

It is straightforward to show that all approximations above are smooth approximations of \(|x|\). In the next section we concentrate on the approximation (2) and define an \( l(s) \) penalized problem with respect to this approximation.
Figure 2: Comparing different approximations for the absolute value function. From up-left to down-right the precision values decrease at the same rate.
2 SmoothABS Regression

Assume $X = (x_1, x_2, x_3, \ldots, x_m)$ is a known sparse matrix of data and $x_i$, $i = 1, 2, \ldots, m$ are independent column vectors of the length $n$ that can be much less than $m$. Consider a linear model of the form of $Y = X\beta + e$ where $\beta_i, i = 1, 2, 3, \ldots, m$ are regression coefficients and $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Then maximizing likelihood is equivalent to minimizing the following function,

$$L(\beta) = \frac{1}{2} (y - X\beta)'(y - X\beta).$$

If $n \leq m$, this problem reduces to a regular quadratic optimization problem. On the other hand, if $n \ll m$ then a standard approach is to make a use of a penalty terms preferably $L_1$ norm of the form of

$$L(\beta) = \frac{1}{2} (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{m} |\beta_i|,$$  \hspace{1cm} (4)

where $\lambda$ is a tuning parameter and determines the sparsity of the optimum solution. Replacing the absolute value function in equation (4) with smooth approximation in equation (2) results in

$$L(\beta) = \frac{1}{2} (y - X\beta)'(y - X\beta) + \lambda 2 \sqrt{\frac{\pi}{2}} \int_{0}^{\beta_i/s} e^{-t^2} dt, \ s > 0, \lambda \geq 0. \hspace{1cm} (5)$$

Remind that $\frac{\sqrt{2}}{\sigma \sqrt{\pi}} e^{-z^2}, \ z \geq 0$ is called half normal density (Ahsanullah et al., 2014, p 18) and is shown by $HN(0, \sigma)$. Denote cumulative distribution function of this distribution by $CHN(Z, 0, \sigma)$. Then $\frac{2}{\sqrt{\pi}} \int_{0}^{\beta_i/s} e^{-t^2} dt$ can be denoted by $CHN(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}})$. Substituting the latter result in (5) results in,

$$L(\beta) = (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{m} \beta_i CHN(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}), \ s > 0, \lambda \geq 0. \hspace{1cm} (6)$$

On the other hand we have,

$$CHN(z, 0, \sigma) = \frac{\sqrt{2}}{\sigma \sqrt{\pi}} \int_{0}^{z} e^{-t^2} dt = 2 \int_{0}^{z} \frac{1}{\sqrt{2\pi} \sigma} e^{-t^2} dt = 2\Phi(z, 0, \sigma) - 1, \hspace{1cm} (7)$$

where $\Phi(z, 0, \sigma)$ denotes cumulative normal distribution with mean of zero and variance equals to $\sigma^2$. Then, $\frac{2}{\sqrt{\pi}} \int_{0}^{\beta_i/s} e^{-t^2} dt = CHN(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) =$
2Φ(z, 0, \frac{\sqrt{2}}{\sqrt{2}}) - 1. Substituting (7) in (6) and setting \( \sigma = \frac{1}{\sqrt{2}} \) and \( \lambda^* = \frac{3}{2} \) results in,

\[
L(\beta) = \frac{1}{2} (y - X\beta)'(y - X\beta) + \lambda^* \sum_{j=1}^{m} \beta_i \left( 2\Phi \left( \frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) s > 0, \lambda^* \geq 0. \tag{8}
\]

Now the problem is minimizing (8) with respect to \( \beta \). Following remark gives the derivation of the likelihood.

**Remark 1** Setting \( X'X = I \), the likelihood of the form of (8) is differentiable at zero and its derivation can be determined by,

\[
f'(\beta) = -X'y + (1 + \frac{\lambda^*}{\sqrt{\pi} s})\beta - \lambda^* \left( \frac{5}{s} \frac{1}{\sqrt{\pi} s^2} \right) \beta^3 = \beta \left( \frac{3}{\sqrt{\pi}} \lambda^* + 1 \right) - \frac{(5/3)\lambda^* \beta^3}{\sqrt{\pi} s^2} - Xiy. \tag{9}
\]

Equation (9) has three roots where two of them are complex and one of them is real. That is,

\[
\hat{\beta} = \sqrt{3 \pi s^6 \left( \frac{5\lambda^*}{\sqrt{\pi}} + 2\frac{3\lambda^* s^2}{\sqrt{\pi} s^2} + \frac{s}{\sqrt{\pi}} \right)^3 \pi (X'y)^2 s^6 + \frac{27\lambda^* (s + \frac{3}{4})^3 + \lambda^* (X'y)^2 s^3 - \sqrt{\pi} (X'y)s^3}{4\lambda^* (s + \frac{3}{4})^2 - \frac{2\lambda^* (s + \frac{3}{4})}{2\lambda^* (s + \frac{3}{4})}}}
\]

\[
\hat{\beta} = \sqrt{3 \pi s^6 \left( \frac{5\lambda^*}{\sqrt{\pi}} + 2\frac{3\lambda^* s^2}{\sqrt{\pi} s^2} + \frac{s}{\sqrt{\pi}} \right)^3 \pi (X'y)^2 s^6 + \frac{27\lambda^* (s + \frac{3}{4})^3 + \lambda^* (X'y)^2 s^3 - \sqrt{\pi} (X'y)s^3}{4\lambda^* (s + \frac{3}{4})^2 - \frac{2\lambda^* (s + \frac{3}{4})}{2\lambda^* (s + \frac{3}{4})}}}
\]

Or : \( \hat{\beta} = \frac{s(X'y)}{\lambda^*(s+2) - s}, \text{ if } \frac{\lambda^*(1 + \frac{2}{s})}{\sqrt{\pi} s^2} = 0. \tag{10} \)
Notice that (10) only happens when \( s \to \infty \) and \( \lambda^* \to 0 \) at the speed faster than \( s \) or simply \( \lambda^* = 0 \). Moreover, if \( \lambda^* = 0 \) then, equation (9) reduces to the well-known solution for LSE that is,

\[
\hat{\beta} = X' y,
\]

that is coincided by the result of equation (10).

### 3 Theoretical properties of the model

In this section we concentrate on theoretical properties of the model as well as some comparisons to lasso and log-approximation. First we show that our approach reaches a similar amount of bias as lasso. To this end we follow (Knight and Fu, 2000) and define a loss function and claim that this function reaches its minimum at estimations, \( \hat{\beta} \).

**Lemma 1** For any \( u \in \mathbb{R}^m \), \( \lambda \geq 0 \) and \( s > 0 \) define

\[
k(u, s) = L(\beta + u) - L(\beta),
\]

where \( L(\beta) \) is the likelihood of the form of equation (8). Then,

\[
\lim_{s \to 0} k(u, s) = u' X' X u - 2 u' N\left(0, \sigma^2(X' X)\right)
\]

\[
+ \lambda^* \sum_{i=1}^{m} \left( |u_i| I(\beta_i = 0) + u_i \text{sign}(\beta_i + u_i) \right),
\]

where \( N \) denotes normal distribution.

Result of Lemma (1) is similar to lasso (refer to (Knight and Fu, 2000, Theorem 1)). Then fixing \( s \) to a value close to zero will guarantee that the results are similar to lasso. On the other hand, fixing \( s \) to a higher value, results in a decrease in sparsity and an increase in the number of significant variables in model. So, determining a threshold for \( s \) is of interest. In the next theorem it is shown that the minimum speed of \( s \) that guarantee the similarity of results to lasso is \( n^{-1/2-\epsilon} \) for \( \epsilon > 0 \). Furthermore, in remark (1) we will show that the sparsity of the model can be controlled by carefully choosing a precision vector, \( s^* \). Diagram (1) shows a summary of the theorem (1) and remark (1).
Data: Import data
Result: Penalized least square estimations

if needs high sparsity then
    $s < 1/\sqrt{n}$
else
    $s \geq 1/\sqrt{n}$
end

while met a criteria do
    Minimize $(y - X\beta)'(y - X\beta) + \lambda^* \sum_{j=1}^{m} \beta(2\Phi(\frac{\beta}{s}, 0, 1/\sqrt{2}) - 1)$
    Renew criteria
    Renew $\lambda^*$
end

Algorithm 1: ABSlasso algorithm using sparse controlling trigger

Theorem 1 If $s_n = s/(n^{1/2+\epsilon}) \to 0$, $\epsilon > 0$, $\lambda_n^*/\sqrt{n} \to \lambda_0 \geq 0$, $\max_{1 \leq i \leq n} x_i x'_i < \infty$ and $X'X/n \to \Sigma$ where $\Sigma$ is nonsingular and also sparsity of coefficients, then $\sqrt{n}(\hat{\beta}_n - \beta) \to \arg\min(k)$ where

$$k(u) = -2u'N + u'\Sigma u + \lambda_0 \sum_{i=1}^{p} u_i \text{sign}(\beta_i)I(\beta_i \neq 0) + |u_i|I(\beta_i = 0),$$

and $N \sim N(O, \sigma^2\Sigma)$ and it guarantees sparse solutions.

Proof of this theorem is provided in appendixes. Then, to get the same results as lasso, it is enough to set $s_n$ to a value less than $n^{-1/2}$. For instance, $s_n = \frac{1}{\sqrt{nm}}, m \to \infty$ can be used in high-dimensional cases. Latter theorem decreases the computation complexity significantly by fixing the amount of memory that must be assigned to algorithm and can be considered as an advantage of this method over Log-approximation in (Schmidt et al., 2007a).

In contrast, assume $s_n = s^*/\sqrt{n} \to q > 0$, where $s^*$ denotes a vector of precisions, then we prove that it provides a tool to control the smoothness of the estimations by controlling zero coefficients.

Remark 2 Under the similar conditions to theorem (1) but $s_n = s^*/\sqrt{n} \to q > 0$, and given a well-designed vector of precisions values $s^*$ so that,

$$\frac{u'_2}{s^*} \phi(\frac{u_2}{s^*}, 0, \frac{1}{\sqrt{2}}) \to 0$$
and $u_2$ represents near-zero estimations. Then, model (5) can be used to generate a less sparse estimations.

Proof of this theorem is provided in appendixes.

For the last property, we prove that $||\beta| - \beta(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1)| \leq 4s\phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}})$. 

**Corollary 1** For any $(\beta,s) \in (\mathbb{R}, \mathbb{R}^+)$, then

$$||\beta| - \beta(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1)| \leq 4s\phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}).$$

Proof of this corollary is given in appendix. Obviously, latter equation tends to zero as $s \to 0$.

### 4 Algorithm

In order to maximize the likelihood, we propose an unconstrained optimization approach mainly because the likelihood is continuous and differentiable at zero. For example Newton iteration of the form of $\beta^{(k+1)} = \beta^{(k)} - t \frac{\partial^2 L(\beta^{(k)})}{\partial \beta^2}$ for suitable steps of length $t$ can be performed on the likelihood. Then, using remark (1),

$$\beta^{(k+1)} = \beta^{(k)} - t \frac{-X'y + \beta^{(k)} + \lambda^* \left(2\Phi(\frac{\beta^{(k)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + \frac{\beta^{(k)}}{s} \phi(\frac{\beta^{(k)}}{s}, 0, \frac{1}{\sqrt{2}})\right)}{1 + \lambda^* \left(3\phi(\frac{\beta^{(k)}}{s}, 0, \frac{1}{\sqrt{2}}) + \frac{\beta^{(k)}}{s} \frac{d\phi}{d\beta} \phi(\frac{\beta^{(k)}}{s}, 0, \frac{1}{\sqrt{2}})\right)},$$

that can be used to optimize likelihood iteratively.

### 5 Computations and complexity order

The only challenging term in the likelihood proposed in this paper is computing the error function of the form of $erf(x) \propto \int_0^x e^{-t^2} dt$. Taylor expansion
makes it possible to approximate \( \text{erf}(x) \) as following

\[
\text{erf}(x) = \int_0^x e^{-u^2} du \\
\approx \frac{2x}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j}}{j!(2j+1)}, \tag{11}
\]

or

\[
\text{erf}(x) \approx 1 - \frac{e^{-x^2}}{x\sqrt{\pi}} \sum_{j=0}^{k} \frac{(-1)^j (2j)!}{j! \cdot (2x)^{-2j}}. \tag{12}
\]

Equation (11) and (12) can approximate error function for small values of \( x \) while equation (13) is used to approximate large values of \( x \). For small \(|x|\), the series (11) is slightly faster than the series (12) because there is no need to compute an exponential. On the other hand, the series (12) is preferable to (11) for moderate \(|x|\) because it involves no cancellation.

Furthermore, for \( \text{erf}(x) \) there are fast enough algorithms that can be implemented by arbitrary high precision, see for example (Vazquez-Leal et al., 2012), (Olver et al., 2010), (Chevillard and Revol, 2008) (Cody, 1969), (Press, 1992), (Lee, 1992), (Cody, 1990), (Borjesson et al., 1979). In order to approximate lasso, extra precision is negligible and then using fast algorithms like

\[
\text{erf}(x) \approx \tanh \left( \frac{39x}{2\sqrt{\pi}} - \frac{111}{2} \arctan\left( \frac{35x}{111\sqrt{\pi}} \right) \right),
\]

can provide reliable results. For smoother results a hybrid method include equation (11), (12) and (13) can provide accurate enough results.

6 Simulation study

In order to compare the performance of the model, we generate a dataset of 100 sparse variables and 50 data points. Without loss of generality we assume the first five coefficients are non-zero and all the rests are zero. We estimate coefficients using the proposed model and repeat this procedure 25 times and compare the results of Lasso, Log-approximation and our method with respect to coefficients accuracy, error variance, BIC and computational
time. According to table (1), proposed method shows a better results with respect to BIC, execution time and other parameters than log-approximation but it takes a bit more time than lasso.

Table 1: Comparing new approximation, Lasso and log-approximation with respect to BIC, execution time, MSE of $\hat{\beta}$, $\hat{e}$ and the number of zero and non-zero coefficients.

| Model                | Precision value | BIC     | Execution time(s) | MSE($\hat{\beta}$) |
|----------------------|-----------------|---------|-------------------|---------------------|
| New-Approximation    | .01             | 540.36  | 4.01              | <.01                |
| Lasso                | -               | 540.95  | 3.16              | <.01                |
| Log-Approximation    | 100             | 932.87  | 26.05             | <.01                |

| Model                | Estimation error | #Zero Coefficients | #Non-Zero Coefficients |
|----------------------|------------------|--------------------|------------------------|
| New-Approximation    | 1.58             | 94                 | 6                      |
| Lasso                | 1.51             | 93                 | 7                      |
| Log-Approximation    | 2.06             | 95                 | 5                      |

On the other hand, we follow the same procedure but for the small coefficients, precisely for the coefficients in $(-1,1)$ interval, and compare ridge regression and proposed method for $s = 1$. Table (2) compares these results.

Table 2: Comparing new approximation and Ridge estimations for small coefficients.

| Model                | Precisions Value | BIC     | MSE($\beta$) | Execution time(s) |
|----------------------|------------------|---------|--------------|-------------------|
| New-Approximation    | 1                | 135     | 0.228        | 0.90              |
| Ridge                | -                | 135     | 0.229        | 0.85              |

Similarly, table (3) compares LARS and new approximation for small true coefficients.

Table 3: Comparing new approximation and LARS for small values of coefficients.

| Model                | Precisions Value | BIC     | MSE($\beta$) | Execution time(s) |
|----------------------|------------------|---------|--------------|-------------------|
| New-Approximation    | 0.01             | 133     | <0.01        | 1.02              |
| LARS                 | -                | 136     | 6.66         | 0.53              |
7 Conclusion

In this paper we addressed a differentiable replacement for $L_1$ penalized likelihoods that provides well properties and is faster than the previous opponents. New method also can be used in smooth situations in order to select more variables than lasso. We showed that the proposed method is feasible and follows well-defined theoretical properties. Furthermore, using a simulation study we showed that the proposed method outperform LARS for small coefficients that is a major issue of LARS. Finally, proposed method lets the analytic to use a broad range of optimization methods in literature that can be considered as an advantage of model.

8 Appendix: Proofs

Proof 1 (Remark 1) Consider that the likelihood is differentiable and continues. Then, the first derivation is,

$$
\frac{\partial L}{\partial \beta_i} = -X_i'(y - X\beta) + \lambda^* \left( 2\Phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + \frac{\beta_i}{s} \phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) \right),
$$

If $X$ is normalized so that $X'X = I$ then,

$$
\frac{\partial L}{\partial \beta_i} = -X_i'y + \beta + \lambda^* \left( 2\Phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + \frac{\beta_i}{s} \phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) \right),
$$

$$
\frac{\partial^2 L}{\partial \beta_i^2} = 1 + \lambda^* \left( \frac{3}{s} \phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) + \frac{\beta_i}{s} \frac{d}{d\beta_i} \phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) \right),
$$

$$
\frac{\partial^k L}{\partial \beta_i^k} = \lambda^* \left( \frac{k+1}{s} \frac{d^{k-2}}{d\beta_i^{k-2}} \phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) + \frac{\beta_i}{s} \frac{d^{k-1}}{d\beta_i^{k-1}} \phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) \right) \quad k \geq 3.
$$

On the other hand,

$$
\phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) = \frac{1}{\sqrt{\pi}} e^{-\left(\frac{\beta_i}{s}\right)^2} = \frac{1}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{(-1)^j (\frac{\beta_i}{s})^{2j}}{j!}. \quad (15)
$$
Derivation of (15) with respect to $\beta_i$, $i = 1, 2, 3, \ldots, m$ is of the form of

$$
\frac{d^k}{d^k\beta_i} \left( \frac{1}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{(-1)^j (\frac{\beta_i}{s})^{2j}}{j!} \right) =
\frac{1}{\sqrt{\pi}} \sum_{j=1}^{\infty} \left( \prod_{k=1}^{k}(2j - (i - 1)) \frac{(-1)^j (\beta_i s)^{2j-k}}{(j!)^2} \right).
$$

(16)

Latter function is zero at $\beta_j = 0$ for odd values of $k$. Note that (14) to (16) are defined at zero, then using Taylor expansion around zero we get,

$$
f(\beta) = f(0) + f'(0)\beta + \frac{f''(0)}{2!} \beta^2 + \cdots + \frac{f^{(n)}(0)}{n!} \beta^n + \cdots
$$

$$
= \frac{1}{2} y'y - (X'y)\beta + (1 + \lambda^* \frac{3}{\sqrt{\pi} s}) \frac{\beta^2}{2} + \lambda^* \frac{5}{\sqrt{\pi} s} \frac{-2}{4!} \beta^4
$$

$$
+ \lambda^* \left( \frac{12}{\sqrt{\pi} s^4} \frac{7}{6!} \right) \beta^6 - \lambda^* \left( \frac{120}{\sqrt{\pi} s^8} \frac{9}{8!} \right) \beta^8 + \cdots
$$

$$
= \frac{1}{2} y'y - (X'y)\beta +
(1 + \lambda^* \frac{3}{\sqrt{\pi} s}) \frac{\beta^2}{2} + \sum_{j=1}^{\infty} \left( \frac{\lambda^*(-1)^j \frac{2j+3}{s}}{\sqrt{\pi} s^{2j}} \prod_{k=1}^{j} \left( 2 + 4(k-1) \right) \frac{\beta^{2j+2}}{(2j+2)!} \right),
$$

and

$$
f'(\beta) = -X'y + (1 + \lambda^* \frac{3}{\sqrt{\pi} s}) \beta +
\sum_{j=1}^{\infty} \left( \frac{\lambda^*(-1)^j \frac{2j+3}{s}}{\sqrt{\pi} s^{2j}} \prod_{k=1}^{j} \left( 2 + 4(k-1) \right) \frac{\beta^{2j+1}}{(2j+1)!} \right).
$$

Proof is completed.

**Proof 2 (Lemma 1)** Take the equation (8) as following,

$$
L(\beta) = \frac{1}{2} (y - X\beta)'(y - X\beta) + \lambda^* \sum_{j=1}^{m} \beta_j \left( 2\Phi(\frac{\beta_j}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \right)
$$

$s > 0, \lambda^* \geq 0,$
and assume $u \in \mathbb{R}^m$. Then,

$$
\lim_{s \to 0} k(u) = \lim_{s \to 0} \left( L(\beta + u) - L(\beta) \right)
= \frac{1}{2} (e - Xu)'(e - Xu) - \frac{1}{2} e'e + \lim_{s \to 0} \left\{ 2\lambda^* \sum_{i=1}^m \beta_i \int_{\beta_i s}^{\beta_i + u_i s} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \right. \\
+ \lambda^* \sum_{i=1}^m u_i \left( 2\Phi \left( \frac{\beta_i + u_i s}{s}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) \left\}
\right.
\rightarrow \frac{1}{2} u'X'X u - u'N \left( 0, \sigma^2(X'X) \right)
+ \lambda^* \sum_{i=1}^m \left( 2\beta_i \frac{u_i s}{s} \phi \left( \frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}} \right) + \lim_{s \to 0} \left\{ u_i \left( 2\Phi \left( \frac{\beta_i + u_i s}{s}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) \right\} \right)
\rightarrow \frac{1}{2} u'X'X u - u'N \left( 0, \sigma^2(X'X) \right)
+ \lambda^* \sum_{i=1}^m \left\{ \\
\left. \begin{array}{ll}
\beta_i = 0 \\
\beta_i + u_i > 0 & \beta_i + u_i > 0 \\
\beta_i + u_i < 0 & \beta_i + u_i < 0 \\
\end{array} \right\}
\right.$$

Proof is completed.

**Proof 3 (Theorem 1)** Assume $\lambda_n^* / \sqrt{n} \to \lambda_0 \geq 0$, $X'X/n \to \Sigma$, $s_n =$
\[ s/n^{1/2+\epsilon} \to 0, s > 0, \epsilon > 0 \text{ and } \max_{1 \leq i \leq n} x_ix_i' < \infty \text{ then define} \]

\[ k_n(u) = L(\beta + \frac{u}{\sqrt{n}}) - L(\beta) \]

\[ = \frac{1}{2}(e - X\frac{u}{\sqrt{n}})'(e - X\frac{u}{\sqrt{n}}) - \frac{1}{2} e'e + 2\lambda_n^* \sum_{i=1}^{m} \beta_i \int_{\frac{\beta_i + u_i}{s_n}}^{\frac{\beta_i + u_i}{s_n}} \frac{1}{\sqrt{2\pi}} e^{-t^2} dt \]

\[ + \lambda_n^* \sum_{i=1}^{m} \frac{u_i}{\sqrt{n}} \left( 2\Phi\left( \frac{\beta_i + u_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) \]

\[ = \frac{1}{2} u'\Sigma u - u'N(0, \sigma^2 \Sigma) + 2\lambda_n^* \sum_{i=1}^{m} \beta_i \int_{\frac{\beta_i + u_i}{s_n}}^{\frac{\beta_i + u_i}{s_n}} \frac{1}{\sqrt{2\pi}} e^{-t^2} dt \]

\[ + \lambda_n^* \sum_{i=1}^{m} \frac{u_i}{\sqrt{n}} \left( 2\Phi\left( \frac{\beta_i + u_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) \]

\[ \to \frac{1}{2} u'\Sigma u - u'N(0, \sigma^2 \Sigma) + 2\lambda_0 \sum_{i=1}^{m} \beta_i \frac{u_i}{s_n \sqrt{\pi}} e^{-\left( \frac{\beta_i u_i}{s_n^2} \right)^2} \]

\[ + \lambda_0 \sum_{i=1}^{m} u_i \left( 2\Phi\left( \frac{\beta_i + u_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) \]

\[ = \frac{1}{2} u'\Sigma u - u'N(0, \sigma^2 \Sigma) \]

\[ + \begin{cases} \lambda_0 \sum_{i=r+1}^{m} u_i \left( 2\Phi\left( \frac{\beta_i + u_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) & \beta = 0 \\ 2\lambda_0 \sum_{i=1}^{r} \beta_i \frac{u_i}{s_n \sqrt{\pi}} e^{-\left( \frac{\beta_i u_i}{s_n^2} \right)^2} + \lambda_0 \sum_{i=1}^{r} u_i \left( 2\Phi\left( \frac{\beta_i + u_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) & \beta \neq 0 \end{cases} \]

Suppose some coefficients are zero. Then, without loss of generality we assume \( r \)-first coefficients are zero and \( m - r \) coefficients are non-zero and

\[ \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad N = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \]

where \( \Sigma_{11} \) is \( r \times r \), \( \Sigma_{22} \) is \( (m - r) \times (m - r) \), \( \Sigma_{21} = \Sigma_{12} \) and \( u_1 \) and \( N_1 \) are
vectors of the length $r$. If $k_n(u)$ is minimized at $u_2 = 0$ then

$$
\frac{\partial}{\partial u} k_n(u) = \frac{\partial}{\partial u} \left( \frac{1}{2} u'\Sigma u - u' N + 2\lambda_0 \sum_{i=1}^{r} \frac{u_i}{s_n \sqrt{\pi}} e^{- (\frac{\beta_i}{s_n})^2} dt \right)
+ \lambda_0 \sum_{i=1}^{r} u_i \left( 2\Phi \left( \frac{\beta_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) \bigg|_{u_2=0} = 0
= u'\Sigma - N(0, \sigma^2 \Sigma)
+ \frac{\partial}{\partial u} \left\{ \begin{array}{ll}
\lambda_0 \sum_{i=r+1}^{m} u_i \left( 2\Phi \left( \frac{\beta_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) & \beta = 0 \\
2\lambda_0 \sum_{i=1}^{r} \beta_i \frac{u_i}{s_n \sqrt{\pi}} e^{- (\frac{\beta_i + u_i}{s_n})^2} + \lambda_0 \sum_{i=1}^{r} u_i \left( 2\Phi \left( \frac{\beta_i + u_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) & \beta \neq 0
\end{array} \right.,
$$

but $s_n \to 0$ then above equation results in sparse estimations because,

$$
\frac{\partial}{\partial u} k(u) = u'\Sigma - N(0, \sigma^2 \Sigma)
+ \frac{\partial}{\partial u} \left\{ \begin{array}{ll}
\lambda_0 \sum_{i=r+1}^{m} u_i \left( 2\Phi \left( \frac{\beta_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) & \beta = 0 \\
2\lambda_0 \sum_{i=1}^{r} \beta_i \frac{u_i}{s_n \sqrt{\pi}} e^{- (\frac{\beta_i + u_i}{s_n})^2} + \lambda_0 \sum_{i=1}^{r} u_i \left( 2\Phi \left( \frac{\beta_i + u_i}{s_n}, 0, \frac{1}{\sqrt{2}} \right) - 1 \right) & \beta \neq 0
\end{array} \right.,
$$

$$
\to u'\Sigma - N(0, \sigma^2 \Sigma) + \frac{\partial}{\partial u} \left\{ \begin{array}{ll}
\lambda_0 \sum_{i=r+1}^{m} |u_i| & \beta = 0 \\
\lambda_0 \sum_{i=1}^{r} u_i & \beta > 0 \\
\lambda_0 \sum_{i=1}^{r} -u_i & \beta < 0
\end{array} \right.,
$$

and

$$
u'_{11} \Sigma_{11} - N_1 + \frac{\lambda_0}{2} \text{Sign}(\beta_{1:r}) = 0, \quad v'_{12} \Sigma_{12} - N_2 \pm \frac{\lambda_0}{2} 1 = 0
$$

Latter is similar to \citep{Knight and Fu, 2000, equations 9, 10}. Then, solving the equations above with respect to $u_1$ results in a positive probability mass at zero for $u_2$.

Proof of theorem is completed.
Proof 4 (Remark 2) Similar to theorem 1, define \( k_n(u) \) as following

\[
k_n(u) = \frac{1}{2} (e - X \frac{u}{\sqrt{n}})'(e - X \frac{u}{\sqrt{n}}) - \frac{1}{2} e'e + 2 \lambda_n^* \sum_{i=1}^{m} \beta_i \int_{\frac{2u}{\sqrt{n}}}^{\frac{2u_i}{\sqrt{n}}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt + \lambda_n^* \sum_{i=1}^{m} \frac{u_i}{\sqrt{n}} \left( 2 \Phi \left( \frac{u_i}{s^* \sqrt{2}} \right) - 0.5 \right)
\]

\[
= \frac{1}{2} u'\Sigma - u'N(0, \sigma^2 \Sigma) + \lambda_0 \sum_{i=r+1}^{m} \beta_i \left( 2 \Phi(\frac{u_i}{s^*}, 0, \frac{1}{\sqrt{2}}) - 1 \right)
\]

Dividing coefficients into zero, \( u_2 \), and non-zero, \( u_1 \), we have

\[
u_1' = \left( \frac{\lambda_0}{2} \text{sign}(\beta) + N_1 \right) \Sigma^{-1}
\]

\[
u_1' \Sigma_12 - N_2 = \lambda_0 \frac{\partial}{\partial u_2} u_2' \left( \Phi(\frac{u_2}{s^*}, 0, \frac{1}{\sqrt{2}}) - 0.5 \right)
\]

In addition,

\[
\frac{\partial}{\partial u_2} u_2' \left( \Phi(\frac{u_2}{s^*}, 0, \frac{1}{\sqrt{2}}) - 0.5 \right) = \left( \Phi(\frac{u_2}{s^*}, 0, \frac{1}{\sqrt{2}}) - 0.5 \right) + \frac{u_2'}{s^*} \phi(\frac{u_2}{s^*}, 0, \frac{1}{\sqrt{2}}),
\]

where the first term in right hand side is always bounded in \((-0.5, 0.5)\). Then,

\[
|\nu_1' \Sigma_12 - N_2 - \lambda_0 \frac{u_2'}{s^*} \phi(\frac{u_2}{s^*}, 0, \frac{1}{\sqrt{2}})| < \frac{\lambda_0}{2}.
\]

As an example, if \( s^* > \frac{\inf f(u')}{3} \) then, \( \phi(\frac{u'}{s^*}, 0, \frac{1}{\sqrt{2}}) \) is very close to zero because \( \phi(3, 0, \frac{1}{\sqrt{2}}) \approx 0 \). Then, given a set of sparse coefficients and a vector of precisions \( s^* \), it is possible to control the sparsity of the estimations.

The proof is completed.

Proof 5 (Corollary 1) Proof of this corollary makes a use of (Abramowitz and Stegun, 2012) inequalities for complementary error function, \( \Phi^c(x, \mu, \sigma) \),

\[
\Phi^c \left( \frac{\beta}{s}, 0, \frac{1}{\sqrt{2}} \right) < \left( \frac{1}{2(\frac{\beta}{s})} \right) \frac{1}{\sqrt{\pi}} e^{-\left( \frac{\beta}{s} \right)^2},
\]

\[
18
\]
and
\[
\frac{1}{\sqrt{\pi}} \frac{e^{-(\frac{\beta}{s})^2}}{\sqrt{\frac{\beta}{s}} + \sqrt{(\frac{\beta}{s})^2 + 2}} < \Phi^s(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) \leq \frac{1}{\sqrt{\pi}} \frac{e^{-(\frac{\beta}{s})^2}}{\sqrt{\frac{\beta}{s}} + \sqrt{(\frac{\beta}{s})^2 + \frac{4}{\pi}}}.
\]

For \( \beta = 0 \) proof is obvious. Then for \( \beta \neq 0 \) and by using the inequalities above we have,
\[
\beta > 0 \Rightarrow |\beta - \beta(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1)| = |2\beta(1 - \Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}))|
\]
\[
= 2\beta(1 - \Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}))
\]
\[
\leq \frac{2\beta}{\sqrt{\frac{\beta}{s}} + \sqrt{(\frac{\beta}{s})^2 + \frac{4}{\pi}}} \frac{e^{-(\frac{\beta}{s})^2}}{1 + \sqrt{1 + \frac{4s^2}{\pi\beta^2}}}
\]
\[
= \Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) \frac{2s}{1 + \sqrt{1 + \frac{4s^2}{\pi\beta^2}}}
\]
\[
\leq \frac{2s}{\sqrt{\pi}} e^{-(\frac{\beta}{s})^2} = 2s\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}).
\]

Similar method can be applied for \( \beta < 0 \). Then,
\[
||\beta| - \beta(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1)| \leq 4\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) \frac{s}{1 + \sqrt{1 + \frac{4s^2}{\pi\beta^2}}}
\]
\[
\leq 4s\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}).
\]

Proof is completed.

References

Abramowitz, M. and I. Stegun (2012). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Books on Mathematics. Dover Publications.
Ahsanullah, M., B. Kibria, and M. Shakil (2014). *Normal and Student’s t Distributions and Their Applications*. Atlantis Studies in Probability and Statistics. Atlantis Press.

Borjesson, P. S. et al. (1979). Simple approximations of the error function q (x) for communications applications. *Communications*.

Chevillard, S. and N. Revol (2008). Computation of the error function erf in arbitrary precision with correct rounding. *JD Bruguera et M. Daumas (éditeurs): RNC 8*, 27–36.

Cody, W. J. (1969). Rational chebyshev approximations for the error function. *Mathematics of Computation 23*(107), 631–637.

Cody, W. J. (1990, March). Performance evaluation of programs for the error and complementary error functions. *ACM Trans. Math. Softw. 16*(1), 29–37.

Hebiri, M. (2008). Regularization with the smooth-lasso procedure. *arXiv preprint arXiv:0803.0668*.

Hebiri, M., S. van de Geer, et al. (2011). The smooth-lasso and other l1+l2-penalized methods. *Electronic Journal of Statistics 5*, 1184–1226.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, 1356–1378.

Lee, C.-I. C. (1992). On laplace continued fraction for the normal integral. *Annals of the Institute of Statistical Mathematics 44*(1), 107–120.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical programming 103*(1), 127–152.

Olver, F., N. I. of Standards, and T. (U.S.) (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*. Wiley Series in Probability and Statistics. Wiley.

Press, W. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Number v. 4. Cambridge University Press.
Ramirez, C., R. Sanchez, et al. (2014). $\sqrt{x^2+m}$ is the most computationally efficient smooth approximation to $|x|$. *Journal of Uncertain Systems* 8.

Schmidt, M., G. Fung, and R. Rosales (2007a). Fast optimization methods for $l_1$ regularization: A comparative study and two new approaches. In *Machine Learning: ECML 2007*, pp. 286–297. Springer.

Schmidt, M., G. Fung, and R. Rosales (2007b). Fast optimization methods for $l_1$ regularization: A comparative study and two new approaches. In J. Kok, J. Koronacki, R. Mantaras, S. Matwin, D. Mladenić, and A. Skowron (Eds.), *Machine Learning: ECML 2007*, Volume 4701 of *Lecture Notes in Computer Science*, pp. 286–297. Springer Berlin Heidelberg.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Vazquez-Leal, H., R. Castaneda-Sheissa, U. Filobello-Nino, A. Sarmiento-Reyes, and J. Sanchez Orea (2012). High accurate simple approximation of normal distribution integral. *Mathematical problems in engineering* 2012.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), 301–320.