Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS

Variable Selection Via Subtle Uprooting

Xiaogang SU

This article proposes a variable selection method termed “subtle uprooting” for linear regression. In this proposal, variable selection is formulated into a single optimization problem by approximating cardinality involved in the information criterion with a smooth function. A technical maneuver is then employed to enforce sparsity of parameter estimates while maintaining smoothness of the objective function. To solve the resulting smooth nonconvex optimization problem, a modified Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm with established global and super-linear convergence is adopted. Both simulated experiments and an empirical example are provided for assessment and illustration. Supplementary materials for this article are available online.

Key Words: BFGS; Information criteria; Linear regression; Regularization.

1. INTRODUCTION

With training data \(\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}\) that consist of \(n\) iid copies of a continuous response \(Y\) and \(p\) predictors \(X_1, \ldots, X_p\), we consider the variable selection issue with the Gaussian linear regression model

\[ y_i = \mu_i(\mathbf{x}_i) + \varepsilon_i, \quad \text{where} \quad \mu_i = \mu_i(\mathbf{x}_i) = \mathbf{x}_i^T \beta = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad (1) \]

where \(\varepsilon_i \sim \mathcal{N}(0, \sigma^2)\) for \(i = 1, \ldots, n\). WLOG, we assume \(y_i\)’s are centered and \(x_{ij}\)’s are standardized so that the intercept term has been omitted from (1). The true \(\beta\) is often sparse in the sense that some of its components are zeros.

One standard method is the best subset selection. In this approach, an information criterion such as Akaike information criterion (AIC; Akaike 1974) or Bayesian information criterion (BIC; Schwarz 1978) is employed to compare and select models. The approach can be formulated as

\[ \min_{\beta, \sigma^2} -2 \cdot L(\beta, \sigma^2) + \lambda_0 \cdot \|\beta\|_0, \quad (2) \]

where \(L(\beta, \sigma^2)\) denotes the log-likelihood function of the model; the penalty parameter \(\lambda_0\) is fixed as 2 in AIC or \(\log(n)\) in BIC; and \(\|\beta\|_0 = \text{card}(\beta) = \sum_{j=1}^{p} 1{\{\beta_j \neq 0\}}\) denotes

Xiaogang Su, Associate Professor, Department of Mathematical Sciences, University of Texas, El Paso, TX 79968
(E-mail: xsu@utep.edu).

© 2015 American Statistical Association, Institute of Mathematical Statistics,
and Interface Foundation of North America
Journal of Computational and Graphical Statistics, Volume 24, Number 4, Pages 1092–1113
DOI: 10.1080/10618600.2014.955176
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jcgs.
the cardinality or the number of nonzero components in $\beta$. Plugging the estimate of $\sigma^2$, $\hat{\sigma}^2 = \| y - X\hat{\beta} \|^2 / n$, $L(\beta, \sigma^2)$ can be simply denoted as $L(\beta)$. Thus, (2) becomes

$$
\min_{\beta \in B} n \log \| y - X\beta \|^2 + \lambda_0 \cdot \| \beta \|_0. \tag{3}
$$

up to some constant, where $B$ denotes the parameter set. Owing to the discrete and non-convex nature of cardinality, optimization of (3), known as an NP-hard problem (Natarajan 1995), is proceeded in two steps: first maximize the log-likelihood for every model choice with known sparsity in $\beta$ and then compare the resultant information criterion $-2 \cdot L(\hat{\beta}) + \lambda_0 \cdot \| \hat{\beta} \|_0$ across all model choices, where $\hat{\beta}$ denotes the maximum likelihood estimate (MLE) of $\beta$ and $L(\hat{\beta})$ is the maximized log-likelihood. Although faster algorithms for optimizing (3) are available by resorting to combinatorial optimization techniques such as the branch-and-bound method (Furnival and Wilson 1974), this best subset selection method is not feasible for even moderately large $p$.

Both ridge regression (Hoerl and Kennard 1970) and least absolute shrinkage and selection operator (LASSO; Tibshirani 1996) can be viewed as convex relaxation of the optimization problem presented in (3). They specifically solve

$$
\min_{\beta} -2 \cdot L(\beta) + \lambda \cdot \| \beta \|_r, \tag{4}
$$

where $\| \beta \|_r = \sum_{j=1}^p |\beta_j|^r$ for some $r > 0$. One key difference from (3) is that the penalty parameter $\lambda \geq 0$ is no longer fixed. With tunable $\lambda$, (4) is equivalent to a penalized least-square problem

$$
\min_{\beta} \frac{1}{2} \| y - X\beta \|^2 + \lambda \cdot \| \beta \|_r. \tag{5}
$$

To encourage convexity, typically $r \geq 1$ is used. For example, $r = 1$ in LASSO or $\ell_1$ regularization and $r = 2$ in ridge regression. While the ridge estimator is explicitly available, LASSO enjoys an attractive property that it simultaneously enforces sparsity. Since the proposal of LASSO, a vast literature has been devoted to the $\ell_1$ regularization and numerous variants have been developed for enhancement and expansion. Efficient algorithms, that is, the homotopy (Osborne, Presnell, and Turlach 2000) or Least Angle Regression (LARS) algorithm (Efron et al. 2004), and coordinate descent (Fu 1998; Friedman et al. 2007; Breheny and Huang 2011), are available and its statistical properties have been explored in depth. In particular, $\ell_1$ regularization may not provide satisfactory variable selection especially when predictors are highly correlated. Nonconvex penalties such as smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) and minimax concave penalty (MCP) (Zhang 2010) are proposed for improvement. An up-to-date literature survey, which is not attempted here, can be found in Zhang (2010) and Mazumder, Friedman, and Hastie (2011) and references therein.

However, regularization incurs a lost track of $\lambda$ and its choice has to be tuned. As a result, the common practice of regularization involves two steps as well: first compute the whole regularization path, that is, solutions of $\beta$ for every $\lambda \geq 0$; and then select the best $\lambda$ according to a model selection criterion, often aided with cross-validation. While tuning has been a convention in regularization methods, it is worth noting that the model selection criteria themselves have a penalized function form similar to regularization. It becomes
apparent that there have been duplicated efforts in the current practice of regularization. Moreover, it is noteworthy that the regularization path is merely a one-dimensional curve in the $p$-dimensional parameter space $\mathcal{B}$. Thus, these regularization methods essentially seek to optimize AIC or BIC over the reduced search space.

In this article, we reformulate variable selection into one single nonlinear programming problem by making a closer connection between regularization and model selection criteria. The key idea is to approximate cardinality $\| \beta \|_0$ with a smooth function, which results in a smoothed version of the information criterion for optimization. This conceptually simplifies the formulation of the variable selection problem and greatly reduces the computational burden. The proposed method can be viewed a natural extension of the classical maximum likelihood principle. It brings model selection and parameter estimation under the common framework of optimization and accomplishes both in one step. One immediate advantage of this approach is that it avoids the two-step approaches involved in either best subset selection or regularization. In addition, we devise a method, termed “uprooting,” to enforce sparsity in parameter estimates while maintaining smoothness of the objective function. The remainder of the article is organized as follows. In Section 2, the subtle uprooting method is introduced in details. Section 3 addresses the algorithmic issues. Section 4 presents simulation results and an illustrating example with real data. Section 5 ends the article with a brief discussion.

2. SUBTLE UPROOTING

We wish to borrow strength from both regularization and information criteria. This entails an extended wish list for the desired approach. From the perspective of $\ell_1$ regularization, a desirable penalty function is expected to help achieve three key goals as spelled out by Fan and Li (2001): unbiasedness in estimating nonzero parameters, sparsity in terms of enforcing zero estimates, and continuity in terms of the model spectrum under consideration. At the same time, for computational concerns, we hope to have $\lambda_0$ fixed by using known choices supplied by information criteria. In addition, we wish the penalty function to possess certain level of smoothness. This would allow us to capitalize further on established results from both optimization theory and statistical estimation theory.

To tackle these issues, a method termed “subtle uprooting” is proposed. In its final form, subtle uprooting simply solves

$$\min_{\beta} n \log \left\{ \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij} \beta'_j \right)^2 \right\} + \lambda_0 \cdot \sum_{j=1}^{p} w(\beta_j), \quad (6)$$

where

$$w(\beta_j) = \tanh \left( a \cdot \beta_j^2 \right) \quad \text{and} \quad \beta'_j = \beta_j \cdot w(\beta_j),$$

for $a > 0$ and $j = 1, \ldots, p$. In matrix form, (6) can be rewritten as

$$\min_{\beta} n \log \| y - XW\beta \|^2 + \lambda_0 \cdot \text{tr}(W), \quad (7)$$
where \( W = \text{diag}(w_j) \) with \( w_j = w(\beta_j) \). Again, \( \lambda_0 \) is fixed according to an information criterion. We recommend having \( a = 50 \) and its choice will be further discussed.

I call this method “subtle uprooting” because variables are uprooted or virtually unselected in a subtle way. The method essentially involves two steps: a smooth surrogate function is first applied for approximating cardinality, followed by an additional technical maneuver for enhancing sparsity while retaining differentiability of the objective function.

### 2.1 The Surrogate Function \( \omega(\cdot) \)

First, we replace the cardinality in (3) with a continuous or smooth surrogate penalty function \( w(\cdot) \). This makes the discrete optimization problem become continuous. To preserve \( \lambda_0 \), a suitable \( w(\cdot) : \Bar{\mathbb{R}} \rightarrow [0, 1] \) must be an even function ranging from 0 to 1 and satisfying that \( w(0) = 0 \) and \( w(\beta) = 1 \) when \( \beta \) is away from 0. Here, \( \Bar{\mathbb{R}} \) denotes the set of extended real numbers that adjoins \( +\infty \) and \( -\infty \) to \( \mathbb{R} \). The \([0, 1]\) range requirement essentially makes \( w(\cdot) \) nonconvex, but this is necessary in order for \( \sum w(\beta) \) to function similarly as cardinality does, namely,

\[
\sum_{j=1}^{p} w(\beta_j) \approx \sum_{j=1}^{p} 1(\beta_j \neq 0) = \text{card}(\beta). \tag{8}
\]

Several choices of \( w(\cdot) \) are listed below:

1. Truncated \( \ell_r \) penalty given by

\[
w(\beta; a, r) = \begin{cases} (|\beta|/a)^r & \text{if } |\beta| \leq a, \\ 1 & \text{otherwise}, \end{cases}
\]

for positive \( r > 0 \) and \( a > 0 \). Typically, \( r = 0.5, 1, 2 \) can be considered. The magnitude of \( a \) controls the sharpness of the approximation. The corresponding penalty functions are continuous but not smooth.

2. Several nonconvex penalties are studied in the literature, including SCAD (Fan and Li 2001) and MCP (Zhang 2010). Both being even functions, the penalty function in SCAD is given by, for \( \beta > 0 \),

\[
w_{a,b}(\beta) = \begin{cases} a\beta & \text{if } \beta \leq a, \\ \frac{2ab\beta - a^2 - \beta^2}{2(b-1)} & \text{if } a < \beta \leq ab, \\ \frac{a^2(b + 1)}{2} & \text{if } \beta > ab, \end{cases}
\]

with first derivative \( \dot{w}_{a,b}(\beta) = a[I(\beta \leq a) + \frac{(a-b)\beta}{a(b-1)}I(\beta > a)] \), for \( a \geq 0 \) and \( b > 2 \). It corresponds to a quadratic spline function with knots at \( a \) and \( ab \). For \( \beta > 0 \), the MCP penalty is given by

\[
w_{a,b}(\beta) = \begin{cases} a\beta - \frac{\beta^2}{2b} & \text{if } \beta \leq ab, \\ \frac{b a^2}{2} & \text{if } \beta > ab, \end{cases}
\]

with first derivative \( \dot{w}_{a,b}(\beta) = (a - \beta/b) I(\beta \leq a b) \) for \( a \geq 0 \) and \( b > 1 \). Both SCAD and MCP penalties are smooth in \( \beta \geq 0 \) with singularity at \( \beta = 0 \). Since
they converge to some constants when \( \beta \) gets large, they can be modified to satisfy the condition that \( \lim_{\beta \to \infty} w(\beta) = 1 \). To do so, set \( b := (2 - a^2)/a^2 \) in SCAD or \( b := 2/a^2 \) in MCP. But the range of \( a \) needs further modification due to the constraints on \( b \).

3. The hyperbolic tangent penalty \( w(\beta) = \tanh(a \cdot |\beta|^r) \) for some constant \( r \geq 0 \) and steepness factor \( a \geq 0 \). When \( r = 1 \), \( w(\beta) \) is nonsmooth and singular at 0. When \( r = 2 \), \( w(\beta) \) is smooth. Figure 1(a) plots the hyperbolic tangent penalty for different choices of \( a \) when \( r = 2 \), showing a sigmoid shape on \( \mathbb{R}_+ \). It can be seen that a larger \( a \) provides a sharper discrimination between 0 and values that are close to 0.

4. The “weight elimination” penalty (Weigend, Rumelhart, and Huberman 1991) in artificial neural network (ANN) \( w(\beta) = \{1 + (a \cdot |\beta|)^{-r}\}^{-1} \) with \( r = 2 \) provides another smooth approximation. The constant \( a \geq 0 \) adjusts the precision of the approximations. Its shape is similar to the hyperbolic tangent penalty, yet with slower progression in its approximation to \( I(\beta \neq 0) \) as \( \beta \) increases for same \( a \).
With a surrogate function $w(\cdot)$, we seek to solve

$$
\min_{\beta} n \cdot \log \| y - X\beta \|^2 + \lambda_0 \cdot \sum_{j=1}^{p} w(\beta_j). \tag{12}
$$

This optimization problem is equivalent to minimizing $\| y - X\beta \|^2$ or $(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})$ subject to $\sum_{j} w(\beta_j) = t_0$ for some $t_0 \geq 0$, where $\hat{\beta} = (\hat{\beta}_j)$ is the MLE or least-square estimate (LSE) of $\beta$. A two-dimensional illustration is given in Figure 1(b) for the tanh penalty function with $r = 2$. Its contour plot (if $t_0$ is adjustable) starts with circles for small $t_0$ values and then becomes sharpened diamonds as $t_0$ increases.

To gain further insight, consider the simplified case with orthogonal design $X^T X = I$. Equation (12) reduces to a componentwise optimization problem

$$
\min_{\beta} \sum_{j} \left\{ (\beta_j - \hat{\beta}_j)^2 + \Lambda_0 w(\beta_j) \right\} \tag{13}
$$

for some $\Lambda_0 \geq 0$. Although there are no simple formulas for the connections among $\{\lambda_0, t_0, \Lambda_0\}$, implying that the values of $t_0$ and $\Lambda_0$, albeit fixed, cannot be explicitly determined, the association in these optimization problems can be analytically useful. Let $\tilde{\beta}$ denote the “subtle uprooting” estimates. It can be seen from (13) that, if $\dot{w}(\beta_j) = 0$, which occurs to a large $\beta_j$ with $w(\beta_j) = 1$, then $\tilde{\beta}_j = \hat{\beta}_j$. To force sparsity, various regularization methods resort to estimates of thresholding forms, which entails a nonsmooth penalty function $w(\beta)$ with singularity at 0.

In this article, we would like to advocate the use of smooth penalty functions, as given by the hyperbolic tangent function with $r = 2$

$$
w(\beta) = \tanh(a\beta^2) = \frac{\exp(2a\beta^2) - 1}{\exp(2a\beta^2) + 1}. \tag{14}
$$

The primary reason is that the proposed method could be viewed as a natural extension of MLE. Since most likelihood or log-likelihood functions are smooth, we do not want to alter this nature. Furthermore, the smoothness property allows us to capitalize on many well-developed theories and methods in both optimization and statistical estimation. As shown in Figure 1(b), sparsity could be forced if $t_0$ (or $\lambda_0$) is treated as an adjustable tuning parameter; but that would substantially intensify the computation burden. Our approach here is to have a fixed $\lambda_0$ as suggested by the information criteria and achieve sparsity of $\tilde{\beta}$ in a different way.

While other smooth functions could be used as well, the derivatives of tanh($\cdot$) are easily available and the tanh($\cdot$) function has connection with the logistic or expit function that is widely used in statistics. Some properties of the hyperbolic tangent penalty $w(\beta)$ are summarized in Proposition 1. The proof is omitted.

**Proposition 1.** Denote $\pi(t)$ the expit function such that $\pi(t) = \expit(t) = \{1 + \exp(-t)\}^{-1}$. Let $\pi = \pi(2a\beta^2)$ and $w = w(\beta) = \tanh(a\beta^2)$. We have

(i) $w(\beta) = 2\pi - 1$. 
The first two derivative of \( w(\beta) \) are given by

\[
\dot{w}(\beta) = 2a\beta(1 - w^2) = 8a\beta \pi (1 - \pi)
\]

and

\[
\ddot{w}(\beta) = 2a(1 - w^2)(1 - 4a\beta^2 w) = 8a\pi(1 - \pi)(1 + (1 - 2\pi)\beta).
\]

By Taylor’s expansion, \( w(\beta) = a\beta^2 + O(\beta^6) \) around 0.

According to (iii), \( w(\beta) \) behaves similarly to the ridge \( \ell_2 \) penalty around 0. This means that sparsity may not be enforced, that is, solution to (12) may not supply 0-valued estimates.

### 2.2 Uprooting

This problem motivates us to enhance the shrinkage around the neighborhood of 0 by having a nonconvex penalty with singularity at 0 as SCAD or MCP does. To do so, a step called “uprooting” is proposed on the basis of the fact that \( \beta = \beta \cdot 1(\beta \neq 0) \approx \beta w(\beta) \). Denote \( \beta' = \beta w(\beta) \). As shown in Figure 1(c), \( \beta' \) and \( \beta \) have a one-to-one correspondence and \( \beta' = \beta \) everywhere except a small neighborhood of 0 where the shrinkage in \( |\beta| \) is effective.

We replace \( \beta_j \) in Model (1) with \( \beta'_j \) so that \( \mu_i \) becomes

\[
\mu_i = \beta_1 w_1 x_{i1} + \cdots + \beta_p w_p x_{ip} = x_i^T W \beta = x_i^T \beta',
\]

where \( \beta' = W \beta = (\beta'_j) \). This can be simply viewed as a reparameterization of the model. The regression coefficient for \( X_j \) is no longer \( \beta_j \), but \( \beta'_j \). The associated optimization problem now becomes

\[
\min_{\beta} n \log \| y - XW \beta \|^2 + \lambda_0 \cdot \sum_{j=1}^p w(\beta_j),
\]

(15)

where the optimization is done with respect to decision variables \( \beta \), instead of \( \beta' \).

To see why this step would help enforce sparsity, let us consider what happens if (15) is optimized with respect to \( \beta' \):

\[
\min_{\beta} n \log \| y - XW \beta \|^2 + \lambda_0 \cdot \sum_{j=1}^p w(\beta_j).
\]

(16)

Compared to (12), the only change is the use of reparameterized \( \beta' \) in (16). In other words, (16) would reduce to (12) if \( w(\beta') \), instead of \( w(\beta) \), is used as penalty for \( \beta' \) in (16). Since \( |\beta_j| \geq |\beta'_j| \) and the tanh function is monotonically increasing in \( \mathbb{R}_+ \), it follows that \( w(\beta) \geq w(\beta') \). Namely, a larger penalty is applied, especially for values that are close to 0.

**Proposition 2.** Let \( w = w(\beta) \), \( \dot{w} = \dot{w}(\beta) \), and \( \ddot{w} = \ddot{w}(\beta) \). Concerning \( w(\beta) \) as a penalty function of \( \beta' \), we have

(i) \[
\frac{d w(\beta)}{d \beta'} = \frac{\dot{w}}{w + \beta \dot{w}},
\]

and
\( \frac{d^2 w(\beta)}{d \beta'^2} = \frac{\dot{w} \ddot{w} - 2 \dot{w}^2}{(w + \beta \dot{w})^3} \)

when \( \beta' \neq 0 \); the first two derivatives do not exist at \( \beta' = 0 \).

**Proof.** Since it is difficult to explicitly express \( \beta \) in \( \beta' \), we apply the chain rule and differentiation of the inverse function:

\[
\frac{dw(\beta)}{d \beta'} = \frac{dw(\beta)}{d \beta} \cdot \frac{d \beta}{d \beta'} = \frac{dw(\beta)}{d \beta} \cdot \left( \frac{d \beta'}{d \beta} \right)^{-1},
\]

where \( \frac{d \beta'}{d \beta} = w + \beta \dot{w} \). Similarly for the second derivative,

\[
\frac{d^2 w(\beta)}{d \beta'^2} = \left( \frac{d}{d \beta} \frac{d w}{d \beta'} \right) \cdot \frac{d \beta}{d \beta'} = \frac{d}{d \beta} \left( \frac{\dot{w}}{w + \beta \dot{w}} \right) \cdot \frac{1}{w + \beta \dot{w}}.
\]

The validity of the above formulas requires \( d \beta'/d \beta \) be nonzero. As a result, both derivatives are expressed as a function of \( \beta \) instead and do not exist at \( \beta' = 0 \). The rest of the proof is omitted. \( \square \)

Proposition 2 indicates that \( w(\beta) \), as a function of \( \beta' \), has singularity at \( \beta' = 0 \) after the reparameterization. Figure 1(d) plots \( w(\beta) \) versus \( \beta' \). Compared to Figure 1(a), the curve now shows a cusp at \( \beta' = 0 \) and a much steeper change rate around 0, a pattern being similar to the nonconvex SCAD or MCP penalties. Noting that \( w + \beta \dot{w} > 0 \) in the denominator of the second derivative \( d^2 w(\beta)/d \beta'^2 \), the convexity of \( w(\beta) \) on \( \mathbb{R}_+ \) (where \( |\beta| \) belongs) can be assessed with a detailed look at the numerator, that is,

\[
\dot{w} \ddot{w} - 2 \dot{w}^2 = 8 a \pi (1 - \pi) [\{1 + (1 - 2 \pi) \beta\}(2 \pi - 1) - 16 a \beta \pi (1 - \pi)],
\]

which is mostly negative. In particular, this is true whenever \( \beta > (2 \pi - 1)^{-1} \). Note that \( \pi \) has range \( 1 \geq \pi \geq 1/2 \). This is also true when \( \beta \to 0_+ \) since, using Proposition 1(iii), the sign of the second derivative is determined by the dominating term \(-8a\beta + O(\beta^2)\).

It is important to emphasize that the decision variables involved in the optimization problem (15) are \( \beta \), instead of \( \beta' \). The objective function remains smooth for \( \beta \). This allows us to use well-established optimization algorithms and theoretical results from statistical estimation. Illustration on the resultant \( \hat{\beta} \) will be made by simulation in Section 5. As we will see, some estimates in \( \hat{\beta}' \) are shrunk to very small values and can be virtually taken as 0 after rounding.

2.2.1 Connection to Nonnegative Garrotte. The uprooting step described above can be motivated from a connection with the nonnegative garrotte (NG) method of Breiman (1995). Nonnegative garrotte can be viewed as a sign-constrained regularization method (see, e.g., Meinshausen 2013) based on the decomposition \( \beta = \text{sgn}(\beta) \cdot |\beta| \). The sign of each component \( \beta \) is assumed to be correctly specified by an initial estimator, say, LSE \( \hat{\beta} \) or any other consistent estimator (Yuan and Lin 2007). Thus, what remains to be done is
estimation of $|\beta|$. The NG estimator is given as the solution to
\[
\min_{\mathbf{w}} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \hat{\beta}_j w_j x_{ij} \right\}^2, \quad \text{subject to } \sum_{j=1}^{p} w_j \leq t \text{ and } w_j \geq 0.
\]
If $\text{sgn}(\beta)$ is wrongly specified by the initial estimator, then NG would have no chance to make correction. The uprooting method is based on a different decomposition $\beta = \beta \cdot 1 \ (\beta \neq 0)$. The indicator $1 \ (\beta \neq 0)$ indicates the binary selection status for a corresponding predictor, which can be approximated by some measure of the likelihood for the variable to be selected. The problem can then be formulated as
\[
\min_{\beta, \mathbf{w}} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \beta_j w_j x_{ij} \right\}^2, \quad \text{subject to } \sum_{j=1}^{p} w_j \leq t \text{ and } 0 \leq w_j \leq 1,
\]
where $w_j$ can be considered as some measure of the likelihood for $j$th variance to be selected. Furthermore, we restate the condition $0 \leq w_j \leq 1$ is restated by relating it to $\beta_j$. With the tanh function, $w_j = w(\beta_j) = \tanh(\alpha \beta_j^2)$. By doing so, $\beta$ becomes the only decision variables in the above optimization problem.

### 3. OPTIMIZATION VIA MODIFIED BFGS

Solving (6) is essentially an unconstrained smooth nonconvex optimization problem. While many optimization methods are applicable, we adopted the modified Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm proposed by Li and Fukushima (2001), who extended the popular BFGS quasi-Newton method for nonconvex optimization problems. Denote the objective function in subtle uprooting as
\[
Q(\beta) = n \log \| \mathbf{y} - \mathbf{XW} \beta \|^2 + \lambda_0 \text{tr}(\mathbf{W}). \tag{17}
\]
Given a current estimate $\beta_k$ at the $k$th step, let $\mathbf{g}_k$ and $\mathbf{H}_k$ denote the gradient vector $\mathbf{g}$ and Hessian matrix $\mathbf{H}$ of $Q(\beta)$, evaluated at $\beta_k$. In Newton’s method, the direction $\mathbf{d}_k$ is determined by $\mathbf{H}_k \mathbf{d}_k + \mathbf{g}_k = 0$.

The gradient vector of $Q(\beta)$ is
\[
\mathbf{g} = -\frac{2n}{\| \mathbf{y} - \mathbf{XW} \beta \|^2} \mathbf{VX}^T (\mathbf{y} - \mathbf{XW} \beta) + \lambda_0 \hat{\mathbf{w}}, \tag{18}
\]
where $\hat{\mathbf{w}} = \{\hat{w}(\beta_j)\} \in \mathbb{R}^p$ and $\mathbf{V} = \text{diag}(w(\beta_j) + \beta_j \hat{w}_j(\beta_j))$. It also can be numerically approximated via finite difference. The form of its Hessian matrix $\mathbf{H}$ is more involved. Quasi-Newton methods seek approximation of $\mathbf{H}$ at each step, and, in particular, the BFGS algorithm computes an approximate Hessian $\mathbf{B}_{k+1}$ by adding a rank-two update matrix to its predecessor $\mathbf{B}_k$.

Nevertheless, owing to the nonconvex nature, $\mathbf{H}_k$ may not be positive definite. As a result, there is no guarantee that the BFGS updated matrix $\mathbf{B}_{k+1} \succ 0$ given $\mathbf{B}_k \succ 0$ and $\mathbf{d}_k$ necessarily be a descent direction. The modified BFGS algorithm of Li and Fukushima (2001), as presented in Algorithm 1, is intended to address these issues. We refer interested readers to Li and Fukushima (2001) for a detailed description.
Algorithm 1 Modified BFGS for Solving Subtle Uprooting

set constants \(0 < c_0 < 0.5\) and \(0 < s_0 < 1\);
set number of maximum runs \(K\) and tolerance \(\varepsilon > 0\);
initialize \(\beta_0, B_0 \succ 0\), and \(k = 0\);
while \(k \leq K\) and \(\| g_k \| > \varepsilon\) do
  obtain \(d_k\) by solving \(B_k d_k + g_k = 0\), done via Cholesky decomposition of \(B_k\);
  set \(i = 0\)
  repeat
    \(\beta_k^i := \beta_k + s_k d_k\);
    set \(\Delta g_k = g_{k+1} - g_k\) and \(\Delta \beta_k = \beta_{k+1} - \beta_k = s_k d_k\);
    compute \(r_k = 1 + \max\left\{ -\frac{\Delta g_k^T \Delta \beta_k}{\| \Delta \beta_k \|^2}, 0 \right\}\) and \(u_k = \Delta g_k + r_k \| \Delta g_k \| \Delta \beta_k\);
    update \(B_{k+1} := B_k - \frac{B_k \Delta \beta_k \Delta \beta_k^T B_k}{\Delta \beta_k^T B_k \Delta \beta_k} \frac{u_k u_k^T}{u_k^T \Delta \beta_k} \); \(k := k + 1\);
end while
return \(\beta_k\).

The next proposition explores the asymptotic properties of the modified BFGS algorithm.

**Proposition 3.** Assume matrix \(X\) is of full column rank \(p\). Let \(\{\beta_k : k = 0, 1, \ldots\}\) denote the sequence of estimates generated by Algorithm 1.

(i) \(\liminf_{k \to \infty} \| g_k \| = 0\).

(ii) Assume that \(\{\beta_k : k = 0, 1, \ldots\}\) converges to a local minimum point \(\beta^*\), where \(g(\beta^*) = 0\) and \(H(\beta^*) \succ 0\). Then \(\lim_{k \to \infty} \frac{\beta_{k+1} - \beta^*}{\| \beta_{k+1} - \beta^* \|} = 0\).

Proposition 3(i) establishes its global convergence. Given any initial value \(\beta_0\), the algorithm converges to a critical point of \(Q(\beta)\). Proposition 3(ii) shows that, if the algorithm converges to a local minimum point, then it converges at least at a superlinear rate. The proof is omitted as the results follow immediately from Theorems 5.1 and 5.2 of Li and Fukushima (2001) by checking the conditions.

### 4. SELECTION CONSISTENCY

In this section, we show that the subtle uprooting estimator \(\tilde{\beta}'\) is consistent in selecting variables under mild regularity conditions. All our discussions are restricted to the conventional setting where the total number of predictors \(p \ll n\) is fixed. Denote the true model as \(y_i = x_i^T \beta_0 + \epsilon_i\), where \(E(x_i) = 0\) and \(\text{cov}(x_i) = E(x_i x_i^T) = \Sigma_x > 0\); \(\epsilon_i \sim N(0, \sigma^2)\) is independent of \(x_i\). We make the following regularity assumptions, which are rather standard in studying the asymptotics of LSE or MLE in Gaussian linear models.
(D1) The parameter set $B$ is compact with bounded $\beta'$.

(D2) $X'X$ is nonsingular with probability one for all $n > p$.

(D3) Assume that $X'X/n \xrightarrow{p} \Sigma_x$, $X'e/n \xrightarrow{p} 0$, and $e'n \xrightarrow{p} 0$.

The subtle uprooting estimator $\tilde{\beta}'$ is the minimizer of

$$Q_n(\beta') = \log \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' \beta')^2 \right\} + \frac{\ln(n)}{n} \text{pen}(\beta')$$

over all $\beta' \in B$, where the penalty function $\text{pen}(\beta') = \sum_{j=1}^{p} w(\beta_j)$ with $\beta_j' = \beta_j w(\beta_j)$. The first term $\log(\sum_{i=1}^{n} (y_i - x_i' \beta')^2/n)$ in (19) is, up to some constant, the profile likelihood of $\beta'$.

To have a form of the sum of iid terms for the objective function in (19), we consider $\tilde{\beta}'$ as the solver of $\min_{\beta' \in B} \exp(Q_n(\beta'))$ with

$$\exp(Q_n(\beta')) = \frac{1}{n} \sum_{i=1}^{n} n^{\sum_j w(\beta_j)/n} (y_i - x_i' \beta')^2.$$

Therefore, $\tilde{\beta}'$ is an $M$-estimator (see, e.g., chap. 5 of van der Vaart 1998) with base function $m_{\beta'} = m(\beta', n) = n^{\sum_j w(\beta_j)/n} (y_i - x_i' \beta')^2$.

With standard notations in empirical processes, we have $\exp(Q(\beta')) = \mathbb{P}_m m_{\beta'}$ or simply $\mathbb{P}_m$. The corresponding population version $\mathbb{P}_m m_{\beta'}$ or $\mathbb{P}_m$ is given by

$$\mathbb{P}_m m_{\beta'} = n^{\sum_j w(\beta_j)/n} E (y - x' \beta')^2 = n^{\sum_j w(\beta_j)/n} \left\{ (\beta_0' - \beta')' \Sigma (\beta_0' - \beta') + \sigma^2 \right\},$$

where the expectation is taken with respect to the joint distribution of $(y_i, x_i)$. Denote its limit as $\mathbb{P}_0 m_{\beta'}$ or simply $\mathbb{P}_0 m$, that is,

$$\mathbb{P}_0 m_{\beta'} = \lim_{n \to \infty} \mathbb{P}_m m_{\beta'} = (\beta_0' - \beta')' \Sigma (\beta_0' - \beta') + \sigma^2,$$

which is convex in $\beta'$ and uniquely minimized at $\beta' = \beta_0'$ regardless of $\sigma^2$.

Theorem 1 shows that $\tilde{\beta}'$ is a consistent estimator of $\beta_0'$ and provides a rate of convergence. The proof can be found in the supplementary materials. Our arguments are essentially based on $M$-estimation. However, unlike most $M$-estimators, the base function $m_{\beta'}$ involves $n$ and, owing to the penalty part, $\mathbb{P}_m$ is not differentiable in $\beta'$ whenever $\beta_j' = 0$. Thus, some modifications are made to accommodate these peculiarities.

**Theorem 1.** Under conditions (D1)–(D3), we have (i) $\tilde{\beta}' \xrightarrow{p} \beta_0'$ and (ii) $n^{3/8} \| \tilde{\beta}' - \beta_0' \| = O_p(1)$.

The $O(n^{-3/8})$ convergence rate may not be the best rate of $\tilde{\beta}'$, but it allows us to establish the selection consistency. To proceed, we define $\ominus(\cdot)$ as an indexing operator such that $\ominus(\beta)$ extracts the indices of nonzero components in $\beta$. Let $M$ denote the linear model space under consideration. With slight abuse of notations, we let $m$ denote a candidate model in $M$ and also the index set of predictors with nonzero coefficients. Thus, the cardinality of $m$, $|m|$, gives the number of regression coefficients in model $m$. Let $m_0$ denote the true
model or the index set $\emptyset (\beta'_0)$ associated with $\beta'_0$. Any candidate model $m$ in $\mathcal{M}$ can be categorized into one of three classes: the underfitting class $\mathcal{M}_-$, correct selection $\mathcal{M}_0$, and the overfitting class $\mathcal{M}_+$, which are, respectively, defined as follows:

$$\mathcal{M}_- = \{m : m \not\supset m_0\},$$
$$\mathcal{M}_0 = \{m_0\}, \text{ and}$$
$$\mathcal{M}_+ = \{m : m \supset m_0 \text{ and } m \neq m_0\}.$$  

Note that the underfitting class $\mathcal{M}_-$ contains scenarios whenever one true predictor is not included in the model.

For a model $m$, let $X_m$ denote the corresponding design matrix after removing those predictor columns with 0 coefficients from $X$. Denote the column space of $X_m$ as $C_m$ and the space orthogonal to $C_m$ as $C_m^\perp$. Let $P_{C_m}$ denote the projection matrix for subspace $C_m$. The following theorem indicates that the global minimizer of $Q_n(\beta')$ must have the same sparsity as the true $\beta'_0$ (w.p.a.1) as $n \to \infty$.

**Theorem 2.** Assume conditions (D1)–(D3) hold. Furthermore, for any $m \in \mathcal{M}_-$, we assume that $\beta'_0 X^T P_{C_m} X \beta'_0 / n \xrightarrow{p} c > 0$. Then $\lim_{n \to \infty} \Pr\{\emptyset (\hat{\beta}') \in \mathcal{M}_0\} = 1$.

It is particularly worth noting that all the above asymptotic results hold for any constant $a > 0$. The practical implication is that there is no need to select $a$ with large $n$.

## 5. NUMERICAL RESULTS

In this section, simulated experiments and real data examples are used to assess and illustrate the subtle uprooting method. To make comparisons with other methods, we have used the same simulation settings and datasets as in Yuan and Lin (2007) and others.

All computations were conducted in R (R Core Team 2013). For practical implementation, we have used simulated annealing as a convenient preprocessing step in solving subtle uprooting, although this method fits best to discrete search space. With the estimates obtained from simulated annealing, modified BFGS is applied next to continue with the optimization till convergence. Optionally, the multistart strategy can also be used in searching for the global minimum. In this approach, solutions from a number of runs with different starting points are compared and the one with the smallest $Q(\beta)$ value is selected as the final estimate. In all the numerical examples presented in this section, the BIC penalty $\lambda_0 = \ln(n)$ was used unless otherwise indicated and the parameter estimates were rounded to four significant digits for sparsity.

### 5.1 SIMULATED EXAMPLE 1

We first consider Example 1 of Yuan and Lin (2007). The data were generated from

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

where $(\beta_1, \beta_2, \beta_3) = (1, 1, 0)$; $\{X_1, X_2, \epsilon\} \sim \mathcal{N}(0, 1)$ independently; conditional on $X_1$ and $X_2$, a noisy variable $X_3$ was generated from $\mathcal{N}(\rho \cdot (X_1 + X_2), 1 - 2\rho^2)$. Thus,
marginally $X_3 \sim \mathcal{N}(0, 1)$ as well. The value of $\rho$ controls the correlation among the predictors. Two choices of $\rho$, 0.35 and 0.65, were considered. According to the path consistency definition of Yuan and Lin (2007), the LASSO, LARS, and elastic net methods (Zou and Hastie 2005) all fail when $\rho > 0.5$.

We examined the subtle uprooting method in different ways. Figure 2(a) and 2(b) plots the proportion of correct selections over 100 simulation runs versus the sample size $n$, with different choices of $a$, while Figure 2(c) and 2(d) plots the proportion of correct selections versus $a$, with different sample sizes. Table 1 tabulated the average and SD of the resultant $\beta$ estimates over 100 runs in several selected scenarios, as well as the proportion of correct selections. It can be seen that subtle uprooting provides reasonably very accurate variable selection, even with highly correlated variables ($\rho = 0.65$). While varying with the choices of $a$ and $n$, its performance shows substantial reliability and robustness. It is not surprising that its performance generally improves with increased sample size with fixed $a$, except the case when $a = 1$ and $\rho = 0.35$. Comparing Figure 2(a) with 2(b), it is interesting to see that the performance of the method seems to improve with larger values of $a$ for less correlated data, but improve with smaller values of $a$ for highly correlated data.

To make comparisons, Figure 3 plots the proportion of correct selections over 100 runs versus $n$ for four competitive methods: best subset regression with BIC, LASSO, SCAD, and MCP. The R packages, bestglm, lars, and ncvreg (for both SCAD and MCP),
Table 1. Illustration of subtle uprooting with simulated Example 1. Rounding with four significant digits after decimal points was applied to have sparsity

| Sample size (n) | Parameter | $a = 1$ | 10 | 20 | 50 | 100 |
|----------------|-----------|---------|----|----|----|-----|
| $\beta_1$     | 0.35      | 0.969 (0.175) | 0.992 (0.162) | 0.996 (0.161) | 0.996 (0.164) | 0.990 (0.163) |
| $\beta_2$     | 1.008 (0.210) | 1.032 (0.197) | 1.034 (0.199) | 1.035 (0.199) | 1.033 (0.202) |
| $\beta_3$     | 0.071 (0.153) | 0.040 (0.130) | 0.041 (0.139) | 0.051 (0.152) | 0.057 (0.157) |
| Correct selection | 0.74     | 0.88    | 0.91  | 0.88 | 0.86 |
| $\beta_1$     | 300       | 1.000 (0.059) | 1.005 (0.059) | 1.005 (0.059) | 1.004 (0.058) | 1.003 (0.058) |
| $\beta_2$     | 0.995 (0.057) | 1.000 (0.057) | 0.999 (0.057) | 0.998 (0.057) | 0.998 (0.058) |
| $\beta_3$     | 0.027 (0.047) | 0.013 (0.033) | 0.010 (0.029) | 0.004 (0.020) | 0.002 (0.018) |
| Correct selection | 0.63     | 0.84    | 0.88  | 0.95 | 0.99 |
| $\beta_1$     | 0.65      | 0.878 (0.482) | 0.973 (0.459) | 0.989 (0.456) | 0.987 (0.477) | 1.013 (0.429) |
| $\beta_2$     | 0.884 (0.439) | 0.973 (0.450) | 0.976 (0.468) | 0.983 (0.442) | 1.007 (0.410) |
| $\beta_3$     | 0.312 (0.586) | 0.266 (0.573) | 0.313 (0.604) | 0.316 (0.590) | 0.323 (0.568) |
| Correct selection | 0.68     | 0.77    | 0.69  | 0.65 | 0.59 |
| $\beta_1$     | 300       | 0.989 (0.071) | 0.999 (0.062) | 1.001 (0.064) | 1.001 (0.063) | 1.002 (0.079) |
| $\beta_2$     | 0.985 (0.075) | 0.996 (0.061) | 0.997 (0.063) | 0.997 (0.067) | 0.998 (0.080) |
| $\beta_3$     | 0.033 (0.086) | 0.008 (0.045) | 0.007 (0.048) | 0.013 (0.065) | 0.042 (0.108) |
| Correct selection | 0.81     | 0.97    | 0.98  | 0.96 | 0.84 |

were used for computation, respectively. These results can be compared to Fig. 1 in Yuan and Lin (2007). However, note that the proportion of correct selection in Yuan and Lin (2007) is defined as the proportion of consistent paths, that is, frequency when the whole regularization path contains the true variable selection. Even with a consistent path, the final variable selection still depends on the choice of the regularization parameter. It can be seen that the best subset selection with BIC provides the best performance for this doable problem in its capacity, followed by SCAD and MCP. Despite the variations, their performances do not seem to be affected too much by increased correlation among variables. LASSO has the lowest proportion of correct selections and its performance deteriorates substantially when predictors are more correlated.

Subtle uprooting shows a performance in between best subset and SCAD or MCP, varying with different values of $a$. Intuitively, a relatively larger $a$ would lead to a better approximation to the best subset selection, while a smaller $a$ value would lead to a smoother objective function. In any case, the optimization problem remains nonconvex. This nonconvex nature also plays a critical role and global optimization techniques can be helpful. In the supplementary materials, additional simulations were conducted by implementing the multistart strategy. It can be seen that, with a reasonable large $a$, subtle uprooting essentially approximates the best subset selection, both outperforming other methods in this setting.

5.2 Simulation Example 2

We next consider three models used in Tibshirani (1996) and other authors including Fan and Li (2001) and Yuan and Lin (2007). Data were simulated from the linear model $y = X\beta + \sigma \varepsilon$, where $p = 8$ and $\sigma = 3$. The correlation between $X_j$ and $X_{j'}$ is $\rho |j - j'|$ with
Figure 3. Simulation results from Example 1: comparison with other methods: (a) $\rho = 0.35$ and (b) $\rho = 0.65$. Four methods are considered: best subset selection (with BIC), LASSO, SCAD, and MCP.

$\rho = 0.5$. The three models differ in $\beta$, as specified below:

Model A $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$;
Model B $\beta_j \equiv 0.85$, for $j = 1, \ldots, p$;
Model C $\beta = (5, 0.5, 0.5, 0.5, 0, 0, 0, 0)^T$.

Three sample sizes: $n = 20$, 50, and 100 were considered. For each sample size, 200 simulation runs were made. For each simulated dataset, subtle uprooting was applied with four choices of $\alpha$: 1, 10, 50, and 100.

We recorded the model error, model size, false positive selections, and false negative selections, all averaged over the 200 simulation runs. Given an estimated $\hat{\beta}$, the model error (ME) is

$$\text{ME}(\hat{\beta}) = (\hat{\beta} - \beta)^T V(\hat{\beta} - \beta),$$

where $V = E(X^T X)$ is the population covariance matrix of $X$.

Table 2 summarizes the results. Since our simulation settings are exactly the same as in Yuan and Lin (2007), the results for Models A, B, and C in Table 2 can be compared to
Table 2. Simulation Example 2 with $\sigma = 3$ (weaker signals): subtle uprooting with different $a$ values and comparison with four other selection methods—all subsets selection, LASSO, SCAD, and MCP. Tabulated quantities include average model error (ME), model size (Size), false positive selections (FP), and false negative selections (FN), all averaged over 200 runs. The true model sizes for Model A, B, and C are 3, 8, and 4, respectively.

| Model | $n$ | $a$ | ME  | Size | FP  | FN  | Method     | ME  | Size | FP  | FN  |
|-------|-----|-----|-----|------|-----|-----|------------|-----|------|-----|-----|
| A     | 20  | 1   | 5.765 | 3.69 | 1.33 | 0.64 | best subset | 5.653 | 3.24 | 0.96 | 0.72 |
|       | 10  |     | 6.276 | 4.68 | 2.14 | 0.46 | LASSO      | 4.421 | 4.84 | 2.12 | 0.29 |
|       | 50  |     | 6.443 | 5.89 | 3.17 | 0.29 | SCAD       | 6.312 | 4.11 | 1.76 | 0.65 |
|       | 100 |     | 6.577 | 6.35 | 3.58 | 0.23 | MCP        | 6.679 | 3.80 | 1.58 | 0.78 |
|       | 50  | 1   | 1.179 | 3.45 | 0.62 | 0.17 | best subset | 1.210 | 3.08 | 0.31 | 0.23 |
|       | 10  |     | 1.384 | 3.47 | 0.67 | 0.20 | LASSO      | 1.382 | 5.05 | 2.08 | 0.03 |
|       | 50  |     | 1.531 | 4.45 | 1.56 | 0.12 | SCAD       | 1.515 | 4.23 | 1.38 | 0.15 |
|       | 100 |     | 1.565 | 4.74 | 1.85 | 0.11 | MCP        | 1.622 | 3.88 | 1.12 | 0.25 |
|       | 100 | 1   | 0.467 | 3.62 | 0.64 | 0.02 | best subset | 0.443 | 3.13 | 0.18 | 0.05 |
|       | 10  |     | 0.499 | 3.29 | 0.34 | 0.05 | LASSO      | 0.631 | 5.25 | 2.25 | 0.00 |
|       | 50  |     | 0.562 | 3.83 | 0.84 | 0.01 | SCAD       | 0.568 | 3.98 | 1.01 | 0.04 |
|       | 100 |     | 0.617 | 4.17 | 1.19 | 0.02 | MCP        | 0.576 | 3.92 | 0.96 | 0.04 |
| B     | 20  | 1   | 6.710 | 3.55 | 0.00 | 4.45 | best subset | 6.711 | 3.10 | 0.00 | 4.91 |
|       | 10  |     | 6.949 | 4.51 | 0.00 | 3.50 | LASSO      | 4.976 | 4.95 | 0.00 | 3.05 |
|       | 50  |     | 7.025 | 5.83 | 0.00 | 2.18 | SCAD       | 7.527 | 4.03 | 0.00 | 3.98 |
|       | 100 |     | 6.939 | 6.43 | 0.00 | 1.58 | MCP        | 7.335 | 3.77 | 0.00 | 4.24 |
|       | 50  | 1   | 3.101 | 4.50 | 0.00 | 3.51 | best subset | 3.526 | 3.82 | 0.00 | 4.18 |
|       | 10  |     | 3.017 | 4.45 | 0.00 | 3.55 | LASSO      | 1.732 | 7.29 | 0.00 | 0.71 |
|       | 50  |     | 2.435 | 5.50 | 0.00 | 2.51 | SCAD       | 2.277 | 6.52 | 0.00 | 1.49 |
|       | 100 |     | 2.307 | 5.77 | 0.00 | 2.24 | MCP        | 2.191 | 6.46 | 0.00 | 1.54 |
|       | 100 | 1   | 1.749 | 5.70 | 0.00 | 2.31 | best subset | 2.008 | 4.99 | 0.00 | 3.01 |
|       | 10  |     | 1.779 | 5.44 | 0.00 | 2.57 | LASSO      | 0.862 | 7.84 | 0.00 | 0.17 |
|       | 50  |     | 1.450 | 6.11 | 0.00 | 1.89 | SCAD       | 0.974 | 7.44 | 0.00 | 0.57 |
|       | 100 |     | 1.308 | 6.48 | 0.00 | 1.52 | MCP        | 0.957 | 7.45 | 0.00 | 0.56 |
| C     | 20  | 1   | 3.599 | 2.74 | 0.81 | 2.07 | best subset | 3.467 | 2.26 | 0.50 | 2.24 |
|       | 10  |     | 5.234 | 3.92 | 1.56 | 1.65 | LASSO      | 3.035 | 3.34 | 1.01 | 1.67 |
|       | 50  |     | 5.759 | 5.43 | 2.43 | 1.01 | SCAD       | 3.060 | 2.36 | 0.64 | 2.28 |
|       | 100 |     | 5.711 | 5.78 | 2.68 | 0.90 | MCP        | 3.232 | 2.12 | 0.50 | 2.38 |
|       | 50  | 1   | 1.075 | 2.61 | 0.45 | 1.84 | best subset | 1.222 | 2.13 | 0.24 | 2.11 |
|       | 10  |     | 1.328 | 2.62 | 0.50 | 1.88 | LASSO      | 1.226 | 4.20 | 1.20 | 1.01 |
|       | 50  |     | 1.451 | 3.53 | 1.11 | 1.58 | SCAD       | 1.380 | 2.81 | 0.73 | 1.92 |
|       | 100 |     | 1.475 | 3.95 | 1.31 | 1.36 | MCP        | 1.368 | 2.61 | 0.65 | 2.04 |
|       | 100 | 1   | 0.640 | 2.89 | 0.44 | 1.56 | best subset | 0.737 | 2.25 | 0.15 | 1.90 |
|       | 10  |     | 0.764 | 2.42 | 0.23 | 1.82 | LASSO      | 0.598 | 4.95 | 1.55 | 0.60 |
|       | 50  |     | 0.749 | 3.06 | 0.61 | 1.55 | SCAD       | 0.812 | 3.87 | 1.19 | 1.33 |
|       | 100 |     | 0.792 | 3.61 | 1.01 | 1.40 | MCP        | 0.825 | 3.45 | 0.96 | 1.51 |

those presented in Tables 1, 2, and 3 of Yuan and Lin (2007), respectively. Model A is also considered in Table 1 of Fan and Li (2001), but they used $n = 40$. Additional simulation results can be found in supplementary materials, where subtle uprooting implemented with multistart strategy was further investigated by considering an increased sample size $n = 200$ and different dependence levels $\rho = 0.2, 0.5, \text{ and } 0.8$.

It can be seen that the results from subtle uprooting, regardless of the choices of $a$, are generally comparable to the results obtained from other methods. Keep in mind that subtle
Table 3. Simulation Example 2 with $\sigma = 1$ (stronger signals): subtle uprooting with different $a$ values and comparison with four other selection methods—all subsets selection, LASSO, SCAD, and MCP. Tabulated quantities include average model error (ME), model size (Size), false positive selections (FP), and false negative selections (FN), all averaged over 200 runs. The true model sizes for Model A, B, and C are 3, 8, and 4, respectively.

| Model | n     | $a$    | ME    | Size | FP | FN | Method      | ME    | Size | FP | FN |
|-------|-------|--------|-------|------|----|----|-------------|-------|------|----|----|
| A     | 20    | 1      | 0.449 | 4.53 | 1.53 | 0.00 | best.subset | 0.457 | 3.73 | 0.74 | 0.01 |
|       |       | 10     | 0.485 | 4.02 | 1.03 | 0.01 | LASSO       | 0.532 | 5.33 | 2.33 | 0.00 |
|       |       | 50     | 0.610 | 4.55 | 1.55 | 0.00 | SCAD        | 0.473 | 3.77 | 0.81 | 0.05 |
|       |       | 100    | 0.654 | 5.06 | 2.06 | 0.00 | MCP         | 0.481 | 3.71 | 0.75 | 0.04 |
|       | 50    | 1      | 0.124 | 4.40 | 1.40 | 0.00 | best.subset | 0.100 | 3.30 | 0.30 | 0.00 |
|       |       | 10     | 0.102 | 3.70 | 0.70 | 0.00 | LASSO       | 0.148 | 5.23 | 2.23 | 0.00 |
|       |       | 50     | 0.121 | 3.60 | 0.60 | 0.00 | SCAD        | 0.102 | 3.59 | 0.59 | 0.00 |
|       |       | 100    | 0.141 | 3.98 | 0.98 | 0.00 | MCP         | 0.103 | 3.51 | 0.51 | 0.00 |
|       | 100   | 1      | 0.053 | 4.24 | 1.24 | 0.00 | best.subset | 0.040 | 3.18 | 0.18 | 0.00 |
|       |       | 10     | 0.043 | 3.66 | 0.66 | 0.00 | LASSO       | 0.066 | 5.29 | 2.29 | 0.00 |
|       |       | 50     | 0.045 | 3.34 | 0.34 | 0.00 | SCAD        | 0.043 | 3.51 | 0.51 | 0.00 |
|       |       | 100    | 0.052 | 3.51 | 0.51 | 0.00 | MCP         | 0.043 | 3.40 | 0.40 | 0.00 |
| B     | 20    | 1      | 1.449 | 6.59 | 0.00 | 1.42 | best.subset | 1.371 | 6.34 | 0.00 | 1.66 |
|       |       | 10     | 1.217 | 6.70 | 0.00 | 1.31 | LASSO       | 0.810 | 7.84 | 0.00 | 1.16 |
|       |       | 50     | 0.877 | 7.37 | 0.00 | 0.63 | SCAD        | 1.078 | 7.26 | 0.00 | 0.75 |
|       |       | 100    | 0.791 | 7.61 | 0.00 | 0.40 | MCP         | 1.115 | 7.22 | 0.00 | 0.79 |
|       | 50    | 1      | 0.301 | 7.83 | 0.00 | 0.18 | best.subset | 0.228 | 7.90 | 0.00 | 0.10 |
|       |       | 10     | 0.215 | 7.97 | 0.00 | 0.04 | LASSO       | 0.195 | 8.00 | 0.00 | 0.00 |
|       |       | 50     | 0.196 | 7.99 | 0.00 | 0.02 | SCAD        | 0.201 | 8.00 | 0.00 | 0.01 |
|       |       | 100    | 0.196 | 7.99 | 0.00 | 0.02 | MCP         | 0.200 | 7.99 | 0.00 | 0.01 |
|       | 100   | 1      | 0.118 | 7.96 | 0.00 | 0.04 | best.subset | 0.092 | 8.00 | 0.00 | 0.00 |
|       |       | 10     | 0.092 | 8.00 | 0.00 | 0.00 | LASSO       | 0.092 | 8.00 | 0.00 | 0.00 |
|       |       | 50     | 0.092 | 8.00 | 0.00 | 0.00 | SCAD        | 0.092 | 8.00 | 0.00 | 0.00 |
|       |       | 100    | 0.092 | 8.00 | 0.00 | 0.00 | MCP         | 0.092 | 8.00 | 0.00 | 0.00 |
| C     | 20    | 1      | 0.546 | 4.40 | 1.31 | 0.91 | best.subset | 0.649 | 3.42 | 0.74 | 1.32 |
|       |       | 10     | 0.666 | 3.85 | 1.02 | 1.18 | LASSO       | 0.525 | 5.01 | 1.43 | 0.42 |
|       |       | 50     | 0.705 | 4.38 | 1.36 | 0.98 | SCAD        | 0.769 | 3.84 | 1.09 | 1.25 |
|       |       | 100    | 0.707 | 4.89 | 1.67 | 0.78 | MCP         | 0.770 | 3.55 | 0.97 | 1.43 |
|       | 50    | 1      | 0.158 | 4.77 | 1.02 | 0.25 | best.subset | 0.201 | 3.55 | 0.24 | 0.69 |
|       |       | 10     | 0.183 | 4.07 | 0.53 | 0.47 | LASSO       | 0.167 | 5.74 | 1.78 | 0.04 |
|       |       | 50     | 0.191 | 4.02 | 0.50 | 0.48 | SCAD        | 0.202 | 5.06 | 1.42 | 0.36 |
|       |       | 100    | 0.178 | 4.42 | 0.75 | 0.34 | MCP         | 0.210 | 4.71 | 1.19 | 0.48 |
|       | 100   | 1      | 0.069 | 4.91 | 0.98 | 0.07 | best.subset | 0.063 | 4.05 | 0.15 | 0.11 |
|       |       | 10     | 0.064 | 4.51 | 0.54 | 0.03 | LASSO       | 0.076 | 5.87 | 1.87 | 0.00 |
|       |       | 50     | 0.060 | 4.18 | 0.24 | 0.07 | SCAD        | 0.070 | 5.07 | 1.13 | 0.06 |
|       |       | 100    | 0.063 | 4.35 | 0.39 | 0.05 | MCP         | 0.073 | 4.97 | 1.02 | 0.06 |

Subtle uprooting is intended for an approximation of all subsets selection with BIC and BIC works best with large sample and strong signals. However, the simulation settings considered here all have relatively weak signals with a large error variance $\sigma = 3$, especially for Model B. As a result, all subsets selection did not win out as much as in Example 1. Subtle uprooting also shows mixed performances when compared to others. In Table 3, we used $\sigma = 1$ for stronger signals. Now it can be seen that subtle uprooting outperforms other methods most of the time in terms of ME for some $a$. When fixing $a = 50$, the results remain either
superior or comparable. Despite the variations, subtle uprooting demonstrates considerable robustness with respect to the choices of \( a \). As theoretically shown in Section 4, subtle uprooting is selection consistent for any constant \( a > 0 \). For these reasons, we recommend simply fixing \( a = 50 \).

### 5.3 Simulation Example 3

This section compares the computation time of subtle uprooting with other selection methods. Data were generated from the linear model \( y = X\beta + \sigma \varepsilon \), where \( \sigma = 3 \) and \( \beta = (\beta_j) \in \mathbb{R}^p \) with \( \beta_j = 1 \) for \( j \leq p/2 \) and 0 otherwise for \( j = 1, \ldots, p \). Same as in Section 5.2, \( X \) were generated from a multivariate normal distribution, where \( \text{cor}(X_j, X_{j'}) = \rho |j - j'| \) with \( \rho = 0.5 \). To see how computing time varies with \( n \) and \( p \), the combinations of \( n \in \{200, 2000\} \) and \( p \in \{10, 50, 100\} \) are considered.

We considered four choices of \( a \in \{1, 10, 50, 100\} \) for subtle uprooting and made comparison with four other methods mentioned before. Table 4 presents the computation time in seconds for performing variable selection with each setting, averaged from three

---

**Table 4. Simulation Example 3 on computation time.** Subtle uprooting with different choices of \( a \) is compared to four methods, which are best subset selection, LASSO, SCAD, and MCP. The recorded CPU time (in seconds) are averaged over three repetitions.

| \( n \) | \( p \) | \( a \) | \( \rho = 0.25 \) | \( 0.50 \) | \( 0.75 \) | **Method** | \( \rho = 0.25 \) | \( 0.50 \) | \( 0.75 \) |
|------|------|------|-------|-------|-------|-----|-------|-------|-------|
| 200  | 10   | 1    | 0.023 | 0.033 | 0.027 | best subset | 0.017 | 0.013 | 0.013 |
|      |      | 10   | 0.020 | 0.020 | 0.023 | LASSO       | 0.070 | 0.057 | 0.057 |
|      |      | 50   | 0.037 | 0.027 | 0.020 | SCAD        | 0.083 | 0.057 | 0.060 |
|      |      | 100  | 0.020 | 0.027 | 0.027 | MCP         | 0.097 | 0.090 | 0.067 |
| 50   | 1    | 0.327 | 0.313 | 0.323 |       | best subset |       |       |       |
|      |      | 10   | 0.307 | 0.320 | 0.460 | LASSO       | 0.230 | 0.223 | 0.217 |
|      |      | 50   | 0.483 | 0.410 | 0.460 | SCAD        | 0.370 | 0.443 | 0.500 |
|      |      | 100  | 0.470 | 0.503 | 0.480 | MCP         | 0.433 | 0.583 | 0.467 |
| 100  | 1    | 1.357 | 1.190 | 1.917 |       | best subset |       |       |       |
|      |      | 10   | 1.263 | 1.357 | 1.850 | LASSO       | 0.683 | 0.587 | 0.613 |
|      |      | 50   | 1.773 | 1.843 | 2.143 | SCAD        | 1.527 | 1.533 | 1.697 |
|      |      | 100  | 1.867 | 1.980 | 2.383 | MCP         | 1.973 | 2.117 | 1.560 |
| 2000 | 10   | 0.060 | 0.047 | 0.067 |       | best subset | 0.020 | 0.017 | 0.023 |
|      |      | 10   | 0.063 | 0.053 | 0.057 | LASSO       | 0.113 | 0.097 | 0.100 |
|      |      | 50   | 0.057 | 0.047 | 0.050 | SCAD        | 0.430 | 0.297 | 0.343 |
|      |      | 100  | 0.063 | 0.047 | 0.057 | MCP         | 0.387 | 0.520 | 0.317 |
| 50   | 1    | 1.380 | 1.443 | 2.507 |       | best subset |       |       |       |
|      |      | 10   | 0.780 | 0.847 | 1.263 | LASSO       | 0.500 | 0.520 | 0.500 |
|      |      | 50   | 0.630 | 1.243 | 1.733 | SCAD        | 2.303 | 2.500 | 2.990 |
|      |      | 100  | 1.160 | 1.587 | 1.750 | MCP         | 2.380 | 4.267 | 3.820 |
| 100  | 1    | 8.393 | 8.760 | 13.300 |       | best subset |       |       |       |
|      |      | 10   | 2.767 | 3.737 | 5.093 | LASSO       | 1.247 | 1.303 | 1.387 |
|      |      | 50   | 3.490 | 4.057 | 6.557 | SCAD        | 3.700 | 6.937 | 9.130 |
|      |      | 100  | 5.820 | 7.020 | 7.910 | MCP         | 6.767 | 10.577 | 12.197 |
Figure 4. Subtle uprooting estimates $\hat{\beta}'$ for the prostate data example with $a = 1, 5, 10, 15, \ldots, 100$.

repetitions for every method. Note that the best subset method is not available when $p = 50$ or $p = 100$, although it is the fastest when $p = 10$.

Subtle uprooting offers comparable speed and works well for high-dimensional cases. Its computational time varies with the choice of $a$; yet, with $a = 50$, it is faster than SCAD and MCP most of the time. It is worth mentioning that our implementation of subtle uprooting is solely R-based while other packages were written in either C or FORTRAN. Thus, the comparison may not fully demonstrate the advantage of subtle uprooting. In principle, subtle uprooting solves a similar problem to either SCAD or MCP, yet with only one fixed $\lambda$. Supposedly, the computation time is expected to increase with larger correlation for all methods; but this does not seem to the case always.

5.4 REAL EXAMPLE: PROSTATE CANCER DATA

For further illustration, we revisit the prostate cancer dataset that was previously analyzed by Tibshirani (1996) and many others including Yuan and Lin (2007). The dataset contains 97 observations and 9 variables. The objective is to regress log(the level of prostate specific antigen) ($\text{lpsa}$) on eight medical measures that consist of log (cancer volume) ($\text{lcavol}$), log(prostate weight) ($\text{lweight}$), age, log(benign prostatic hyperplasia amount) ($\text{lbph}$), seminal vesicle invasion ($\text{svi}$), log(capsular penetration) ($\text{lcp}$), Gleason score ($\text{gleason}$), and percentage Gleason scores 4 or 5 ($\text{pgg45}$). All previous variable selection results clearly indicate that $\text{lcavol}$ is most predictive while $\text{gleason}$ seems unimportant. To illustrate how various variable selection methods perform with correlated predictors, Yuan and Lin (2007) replaced $\text{gleason}$ with an artificial predictor, computed as $(2 \times \text{lcavol} + \text{gleason})$. A good variable selection method is anticipated to recognize the feigned multicolinearity and continue excluding the modified $\text{gleason}$.

Figure 4 plots the estimated coefficients $\hat{\beta}'$ from subtle uprooting with different values of $a = \{1, 5, 10, 15, 20, \ldots, 100\}$. We can see that $\hat{\beta}'$ is very stable when $a \geq 25$. Table 5 tabulates the detailed results involved in subtle uprooting when $a = 50$, as well as results
Table 5. Variable selection with prostate cancer data. In subtle uprooting, $a = 50$ and $\lambda_0 = \ln(n)$ were used. First, a global $\hat{\beta}$ were obtained with simulated annealing and then used as starting values for modified BFGS, which yielded a local $\tilde{\beta}$. Finally, the subtle uprooting estimate $\tilde{\beta}'$ is computed as $\tilde{\beta}' = \tilde{\beta} w(\tilde{\beta})$

|                | LSE $\hat{\beta}$ | Global $\hat{\beta}$ | LSE $\tilde{\beta}$ | $w(\tilde{\beta})$ | $\tilde{\beta}'$ | Selection | Best | Subset | LASSO | NG | Net | SCAD | MCP |
|----------------|------------------|------------------|------------------|------------------|------------------|-----------|------|-------|------|----|-----|------|-----|
| lcavol         | 0.476            | 0.495            | 0.540            | 1.000            | 0.540            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
| lweight        | 0.231            | 0.209            | 0.245            | 0.995            | 0.243            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
| age            | −0.137           | −0.047           | 0.000            | 0.000            | 0.000            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
| lbph           | 0.122            | 0.120            | 0.000            | 0.000            | 0.000            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
| svi            | 0.273            | 0.184            | 0.237            | 0.993            | 0.235            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
| lcp            | −0.128           | −0.038           | 0.000            | 0.000            | 0.000            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
| pgg45          | 0.109            | −0.018           | 0.000            | 0.000            | 0.000            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
| gleason        | 0.117            | 0.014            | 0.000            | 0.000            | 0.000            | ✓         | ✓    | ✓     | ✓    | ✓  | ✓   | ✓    | ✓   |
from six other variable selection methods. Except LASSO and elastic net, all methods are able to exclude the modified gleason variable. NG, best subset, and subtle uprooting provide a more parsimonious model choice than the one found by MCP and SCAD.

6. DISCUSSION

This article makes two noteworthy contributions. First, we point out that variable selection can be reformulated into one single optimization problem, by approximating cardinality with a smooth surrogate function in information criteria. Second, we propose a special technical maneuver that helps maintain smoothness of the objective function and enhance sparsity of the coefficient estimates. There are several interesting avenues for future research. First of all, the method can be extended in a few ways. It is straightforward to apply subtle uprooting to generalized linear models and other parametric or semiparametric models. Subtle uprooting can be most useful when the likelihood function itself is not log-concave. Extensions to structured sparsity as in grouped LASSO can also be considered. Second, we have used $\lambda = \log(n)$ in spirit of BIC. Note that BIC is no longer a consistent selection criterion when $p \gg n$, in which scenario, the use of the modified BIC criterion by Chen and Chen (2008) may facilitate a natural extension. With large $n$ and $p$, the limited-memory BFGS (L-BFGS; Liu and Nocedal 1989) algorithm could be helpful, although some similar modification is needed to handle nonconvexity. Third, further studies are ongoing to better understand the theoretical and empirical properties of this method in terms of both variable selection and parameter estimation.

SUPPLEMENTARY MATERIALS

The supplementary materials contain additional simulation results and proofs. The supplementary materials contain two sections. In the first section, additional simulated experiments were carried out to further address three issues in subtle uprooting: global optimization with multi-start strategy, scenarios with stronger signals, and robustness with respect to the choice of $\alpha$. The second section contains detailed proofs for Theorem 4.1 and 4.2 in Section 4.

ACKNOWLEDGMENTS

The author thanks the editor, an associate editor, and two referees for helpful comments and suggestions that greatly improved the article.

[Received November 2012. Revised August 2014.]

REFERENCES

Akaike, H. (1974), “A New Look at Model Identification,” IEEE Transactions on Automatic Control, 19, 716–723. [1092]
Breheny, P., and Huang, J. (2011), “Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications to Biological Feature Selection,” Annals of Applied Statistics, 5, 232–253. [1093]
Breiman, L. (1995), “Better Subset Regression Using the Nonnegative Garrote,” *Technometrics*, 37, 373–384. [1099]

Chen, J., and Chen, Z. (2008), “Extended Bayesian Information Criterion for Model Selection With Large Model Spaces,” *Biometrika*, 95, 759–771. [1112]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least Angle Regression,” *The Annals of Statistics*, 32, 407–489. [1093]

Fan, J., and Li, R. (2001), “Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties,” *Journal of the American Statistical Association*, 96, 1348–1360. [1093,1094,1095,1105,1107]

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007), “Pathwise Coordinate Optimization,” *Annals of Applied Statistics*, 1, 302–332. [1093]

Fu, W. (1998), “Penalized Regressions: The Bridge vs. the LASSO,” *Journal of Computational and Graphical Statistics*, 7, 397–416. [1093]

Furnival, G. M., and Wilson, R. W. (1974), “Regression by Leaps and Bounds,” *Technometrics*, 16, 499–511. [1093]

Höler, A. E., and Kennard, R. W. (1970), “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, 42, 80–86. [1093]

Li, D.-H., and Fukushima, M. (2001), “A Modified BFGS Method and its Global Convergence in Nonconvex Minimization,” *Journal of Computational and Applied Mathematics*, 129, 15–35. [1100,1101]

Liu, D. C., and Nocedal, J. (1989), “On the Limited Memory BFGS Method for Large Scale Optimization,” *Mathematical Programming*, 45, 503–528. [1112]

Mazumder, R., Friedman, H., and Hastie, T. (2011), “SparseNet: Coordinate Descent With Nonconvex Penalties,” *Journal of the American Statistical Association*, 106, 1125–1138. [1093]

Meinshausen, N. (2013), “Sign-Constrained Least Squares Estimation for High-Dimensional Regression,” *Electronic Journal of Statistics*, 7, 1607–1631. [1099]

Natarajan, B. K. (1995), “Sparse Approximate Solutions to Linear Systems,” *SIAM Journal on Computing*, 24, 227–234. [1093]

Osborne, M., Presnell, B., and Turlach, B. (2000), “On the Lasso and its Dual,” *Journal of Computational and Graphical Statistics*, 9, 319–337. [1093]

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at [http://www.r-project.org](http://www.r-project.org) [1103]

Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464. [1092]

Tibshirani, R. J. (1996), “Regression Shrinkage and Selection Via the LASSO,” *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1093,1105,1110]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, New York: Cambridge University Press. [1102]

Weigend, A. S., Rumelhart, D. E., and Huberman, B. A. (1991), “Generalization by Weight-Elimination With Application to Forecasting,” in *Advances in Neural Information Processing Systems 3* (Denver 1990), eds. R. P. Lippmann, J. E. Moody, and D. S. Touretzky, San Mateo, CA: Morgan Kaufmann, pp. 875–882. [1096]

Yuan, M., and Lin, Y. (2007), “On the Non-Negative Garrotte Estimator,” *Journal of the Royal Statistical Association*, Series B, 69, 143–161. [1099,1103,1105,1106,1110]

Zhang, C.-H. (2010), “Nearly Unbiased Variable Selection Under Minimax Concave Penalty,” *The Annals of Statistics*, 38, 894–942. [1093,1095]

Zou, H., and Hastie, T. (2005), “Regression Shrinkage and Selection Via the Elastic Net With Application to Microarrays,” *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1104]