Statistics
in Medicine

Research Article

Received 12 June 2012, Accepted 29 March 2013 Published online 6 May 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5834

Variable selection in monotone single-index models via the adaptive LASSO

Jared C. Foster,*† Jeremy M. G. Taylor and Bin Nan

We consider the problem of variable selection for monotone single-index models. A single-index model assumes that the expectation of the outcome is an unknown function of a linear combination of covariates. Assuming monotonicity of the unknown function is often reasonable and allows for more straightforward inference. We present an adaptive LASSO penalized least squares approach to estimating the index parameter and the unknown function in these models for continuous outcome. Monotone function estimates are achieved using the pooled adjacent violators algorithm, followed by kernel regression. In the iterative estimation process, a linear approximation to the unknown function is used, therefore reducing the situation to that of linear regression and allowing for the use of standard LASSO algorithms, such as coordinate descent. Results of a simulation study indicate that the proposed methods perform well under a variety of circumstances and that an assumption of monotonicity, when appropriate, noticeably improves performance. The proposed methods are applied to data from a randomized clinical trial for the treatment of a critical illness in the intensive care unit. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: adaptive LASSO; isotonic regression; kernel estimator; single-index models; variable selection

1. Introduction

Linear regression is a simple and commonly used technique for assessing relationships of the form $y = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + \epsilon$ between an outcome of interest, y, and a set of covariates, $x_1, \ldots, x_p$; however, in many cases, a more general model may be desirable. As noted by Härdle *et al.* [1], one particularly useful and more general variation of the linear regression formulation is the single-index model

$$y_i = \eta\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right) + \epsilon_i, \tag{1}$$

where $\boldsymbol{x}_i'$s are subject-specific covariate vectors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $y_i \in \mathbb{R}$, η is an unknown function, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. errors with mean zero and variance σ^2, and ϵ_i's and $\boldsymbol{x}_i$'s are independent. To ensure identifiability, no intercept is included, and β_1 is assumed to be equal to 1. These models are able to capture important features in high-dimensional data while avoiding the difficulties associated with high-dimensionality, as dimensionality is reduced from many covariates to a univariate index [2]. Single-index models have applications to a number of fields, including discrete choice analysis in econometrics and dose–response models in biometrics [1].

There is a rich literature on estimation of $\boldsymbol{\beta}$ and η, including [1–6], among many others. Additionally, variable selection for single-index models was considered by Kong and Xia [7], who proposed the separated cross-validation method, and Liang *et al.* [8], who applied the smoothly clipped absolute deviation (SCAD) approach to partially linear single-index models. However, little consideration has been given to such problems for monotone single-index models, where η is required to be nondecreasing (or nonincreasing). In the case of linear models, a great many authors, including [9–11] and [12], have considered variable selection via penalized least squares, which allows for simultaneous selection of variables and estimation of regression parameters. Several penalty functions, including the SCAD [10], the adaptive

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.
**Correspondence to: Jared C. Foster, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
†E-mail: jaredcf@umich.edu

LASSO [11] and the adaptive elastic net [12], have been shown to possess favorable theoretical properties, including the *oracle* properties, that is, consistency of selection and asymptotic normality, with the asymptotic covariance matrix being the same as that which would be obtained if the true underlying model were known. Hence, for large samples, oracle procedures perform as well as if the true underlying model were known in advance. Furthermore, Liang *et al.* [8] established the oracle properties for the SCAD for partially linear single-index models. Given the desirable properties of the SCAD, adaptive LASSO and adaptive elastic net approaches, it is natural to consider the extension of these methods to monotone single-index models. Unlike the adaptive LASSO and adaptive elastic net, which present a convex optimization problem, the SCAD optimization problem is nonconvex and thus more computationally demanding [13]. In addition, the adaptive elastic net and SCAD methods require the selection of two tuning parameters, whereas the adaptive LASSO requires the selection of a single tuning parameter. Therefore, for convenience, computational efficiency, and because covariates in our example data are not highly correlated (a condition under which the adaptive elastic net is especially good), we consider adaptive LASSO penalized least squares estimation of $\boldsymbol{\beta}$ in monotone single-index models.

The assumption of monotonicity and the desire to select a subset of the covariates are motivated in part by the randomized clinical trial data considered in [14]. A monotonicity assumption is often reasonable, and such an assumption may improve prediction and reduction in model complexity while also allowing for more straightforward inference. Foster *et al.* [14] consider methods for subgroup identification in randomized clinical trial data. In such cases, should a subgroup be identified, it is desirable that this subgroup be easily described and depend on only a small number of covariates. Application of the methods proposed in this paper results in estimates $\hat{\eta}$ and $\hat{\boldsymbol{\beta}}$, such that $\hat{y}_i = \hat{\eta}\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i\right)$, where $\hat{\eta}$ is monotone and $\hat{\boldsymbol{\beta}}$ generally includes a number of zero values. Using this model, one can classify individuals with $\hat{y}$'s beyond some predefined threshold, c, as being in the subgroup. Then, because of the monotonicity of $\hat{\eta}$, the predefined threshold can be converted into an equivalent threshold, c', on $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}$, and the impact of the chosen covariates on subgroup membership can be easily described. Without the assumption of monotonicity, the subgroup may be a collection of several disjoint subregions of the covariate space, making each covariate's impact on subgroup membership more difficult to ascertain.

The remaining sections of this article are as follows. In Section 2, we consider penalized least squares estimation for monotone single-index models, briefly discuss asymptotics and discuss a method to obtain standard error estimates for $\boldsymbol{\beta}$. In Section 3, we present the results of a simulation study implemented to assess the performance of the adaptive LASSO penalized single-index models. In Section 4, we briefly discuss the application of this method to the randomized clinical trial data, and in Section 5, we give concluding remarks.

2. Estimation for monotone single index models

Our estimation procedure iterates between estimation of $\boldsymbol{\beta}$ and η until convergence. Given some η, the penalized least squares estimator of $\boldsymbol{\beta}$ can be found by minimizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \eta(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)\right)^2 + \lambda_n \sum_{j=2}^{p} w_j|\beta_j|, \tag{2}$$

where w_j, $j = 2,\ldots,p$ are known weights and covariates $\boldsymbol{x}_i$ are standardized to have mean zero and variance 1. Because of the identifiability constraint specified in model (1), β_1 is not penalized. Following [11], we choose $w_j = |\hat{\beta}_{\mathrm{init},j}|^{-\gamma}$ for $\gamma > 0$, where $\hat{\boldsymbol{\beta}}_{\mathrm{init}}$ is a n^{α}-consistent estimator of $\boldsymbol{\beta}$, where $0 < \alpha \leqslant \frac{1}{2}$. We use linear ordinary least squares estimates for $\hat{\boldsymbol{\beta}}_{\mathrm{init}}$, as under the assumptions of Theorem 2.1 in [15], these are shown to be $\sqrt{n}$-consistent up to a multiplicative scalar. Once obtained, $\hat{\boldsymbol{\beta}}_{\mathrm{init}}$ is rescaled by $\hat{\beta}_{1,\mathrm{init}}$. Alternatively, weights could be defined using the unpenalized single-index model estimates of $\boldsymbol{\beta}$.

For a given $\boldsymbol{\beta}$, without considering the monotonicity constraint, η can be estimated at some point t using the Nadaraya–Watson kernel-weighted average:

$$\hat{\eta}(t;\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},h) = \frac{\sum_j y_j K\left(\frac{t-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j}{h}\right)}{\sum_j K\left(\frac{t-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j}{h}\right)}, \tag{3}$$

where $\boldsymbol{X}$ is the covariate matrix, K is a fixed kernel function and h is a bandwidth. Note that, when $\boldsymbol{\beta}$ is known, $\hat{\eta}$ is determined by h, so a value of h must be chosen. We consider kernel functions that are symmetric probability densities. For numerical stability, we hold (3) fixed for all t outside the range of the $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$'s. That is, $\hat{\eta}(t) = \hat{\eta}(\min_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i))$ if $t < \min_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$ and $\hat{\eta}(\max_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i))$ if $t > \max_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$.

By combining (2) and (3), the adaptive LASSO estimator for $\boldsymbol{\beta}$ is obtained by minimizing

$$\hat{Q}(\boldsymbol{\beta}, h) = \sum_i{}' \left(y_i - \hat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}, h)\right)^2 + \lambda_n \sum_{j=2}^{p} w_j |\beta_j| \tag{4}$$

with respect to $\boldsymbol{\beta}$, where $\sum_i'$ denotes summation over i such that the denominator in the kernel estimator is not too close to zero. Details can be found in [1]. With the inclusion of the penalty term in (4), $\hat{\boldsymbol{\beta}}$ becomes a function of λ_n, so in addition to h, a value of λ_n must be chosen if $\hat{\boldsymbol{\beta}}$ is to be obtained. Throughout this paper, we refer to the method of estimating $\boldsymbol{\beta}$ and η without a monotonicity constraint, using objective function (4), as the *unconstrained* approach.

2.1. *A smooth monotone function estimate for η with fixed $\boldsymbol{\beta}$*

There are a variety of ways to obtain smooth monotone regression function estimates, including quadratic B-splines [16], I-splines [17], empirical distribution tilting [18], the scatterplot smoothing approach of Friedman and Tibshirani [19] and the kernel-based approach of Mukerjee [20] and Mammen [21]. We consider the kernel-based method of the latter two papers, which we briefly describe later.

Assume $\boldsymbol{\beta}$ is known. The proposed monotone estimator $\hat{\eta}_m$ requires two steps:

Isotonization. This step involves the application of the pooled adjacent violator algorithm (PAVA) [22]. Using $(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, y_i)$ ordered by increasing $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i$ as data, PAVA produces monotone estimates $\hat{m}_1, \ldots, \hat{m}_n$, which are averages of y_j's near i (unless y's are already monotone, in which case $\hat{m}_i = y_i$), and which are not necessarily smooth [19].

Smoothing. Apply the kernel estimator (3) with y_i replaced by $\hat{m}_i$ for all i to estimate η. That is, $\hat{\eta}_m(t) = \hat{\eta}(t; \boldsymbol{\beta}, \widehat{\boldsymbol{m}}, \boldsymbol{X}, h)$.

Because $\hat{m}_1, \ldots, \hat{m}_n$ are monotone, the resulting function estimate is monotone in t. It is worth noting that this may not necessarily be the case for other smoothing methods, such as local linear regression.

As previously mentioned, a bandwidth is needed to estimate η and can be found using cross-validation; however, our algorithm requires estimation of both η and its derivative η', so care must be taken. In particular, to ensure good algorithmic convergence, it is crucial that $\hat{\eta}'$ be smooth, but to obtain a smooth estimate of η', it is often necessary to oversmooth η. Thus, we restrict the range of potential bandwidths in our cross-validation. Specifically, h is restricted to be between $0.1^* sd(\boldsymbol{X}\boldsymbol{\beta})$ and $sd(\boldsymbol{X}\boldsymbol{\beta})$, as values in this range were found to perform well in our simulations.

2.2. *Estimation for $\boldsymbol{\beta}$ with fixed η*

The shooting algorithm proposed by Fu [23] has been shown to perform well in solving LASSO penalized least squares problems for linear models [24]. Therefore, we consider the application of this algorithm to LASSO problems for the single-index model. One way to achieve this is to employ a linear approximation via Taylor series expansion of $\eta(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$ about $\boldsymbol{\beta}_0^T\boldsymbol{x}_i$, where $\boldsymbol{\beta}_0$ is known. We define the linear approximation as follows:

$$\eta\left(\boldsymbol{\beta}^T\boldsymbol{x}_i\right) \approx \eta\left(\boldsymbol{\beta}_0^T\boldsymbol{x}_i\right) + \eta'\left(\boldsymbol{\beta}_0^T\boldsymbol{x}_i\right)\left[\boldsymbol{\beta}^T\boldsymbol{x}_i - \boldsymbol{\beta}_0^T\boldsymbol{x}_i\right]. \tag{5}$$

Let

$$y_i^* = y_i - \eta\left(\boldsymbol{\beta}_0^T\boldsymbol{x}_i\right) + \eta'\left(\boldsymbol{\beta}_0^T\boldsymbol{x}_i\right)\boldsymbol{\beta}_0^T\boldsymbol{x}_i$$

and

$$\boldsymbol{x}_i^* = \eta'\left(\boldsymbol{\beta}_0^T\boldsymbol{x}_i\right)\boldsymbol{x}_i.$$

Then, we have

$$y_i - \eta\left(\boldsymbol{\beta}^T\boldsymbol{x}_i\right) \approx y_i - \eta\left(\boldsymbol{\beta}_0^T\boldsymbol{x}_i\right) - \eta'\left(\boldsymbol{\beta}_0^T\boldsymbol{x}_i\right)\left[\boldsymbol{\beta}^T\boldsymbol{x}_i - \boldsymbol{\beta}_0^T\boldsymbol{x}_i\right] = y_i^* - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i^*,$$

and (4) can be approximated by

$$\hat{Q}_{\text{lin}}(\boldsymbol{\beta}) = \sum_i \left(y_i^* - \boldsymbol{\beta}^T \boldsymbol{x}_i^*\right)^2 + \lambda_n \sum_{j=2}^{p} w_j |\beta_j|, \tag{6}$$

which is a LASSO penalized least squares problem for the linear model and can thus be solved using the shooting algorithm.

Note that (5) involves an estimate of η'. This estimate is obtained as follows. Sort the observations by increasing $\boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{x}_i$, and define new data $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) :, \ i = 1, \ldots, n-1\}$, where $\tilde{y}_i = \frac{\eta(\boldsymbol{\beta}_0^T \boldsymbol{x}_{i+1}) - \eta(\boldsymbol{\beta}_0^T \boldsymbol{x}_i)}{\boldsymbol{\beta}_0^T \boldsymbol{x}_{i+1} - \boldsymbol{\beta}_0^T \boldsymbol{x}_i}$ and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \frac{\boldsymbol{x}_{i+1} - \boldsymbol{x}_i}{2}$. This new data should 'look like' data coming from the model $\tilde{y}_i = \eta'\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i\right) + \tilde{\epsilon}_i$, so $\eta'(t)$ can be estimated using (3), but with $\{(\boldsymbol{x}_i, y_i) : \ i = 1, \ldots, n\}$ replaced by $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) : \ i = 1, \ldots, n-1\}$, that is, $\hat{\eta}'(t) = \hat{\eta}(t; \boldsymbol{\beta}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}, \tilde{h})$, where $\tilde{h}$ is a new bandwidth for the derivative estimate. To select $\tilde{h}$, cross-validation can again be used.

2.3. Algorithm

The algorithm to obtain final estimates of η and $\boldsymbol{\beta}$ iterates between the steps in Sections 2.1 and 2.2 until convergence. After k iterations, let $\hat{\boldsymbol{\beta}}^{(k)}$, $\hat{\eta}_m^{(k)}$ and $\hat{m}_1^{(k)}, \ldots, \hat{m}_n^{(k)}$ denote the current estimates of $\boldsymbol{\beta}$ and η and the current PAVA estimates, respectively. For a given λ_n, the 'final' estimates of $\boldsymbol{\beta}$ and η are obtained as follows:

1. Using data $\left\{(\hat{\boldsymbol{\beta}}^{(k)T} \boldsymbol{x}_i, y_i) : \ i = 1, \ldots, n\right\}$, apply PAVA to obtain new monotone data $\left\{\left(\hat{\boldsymbol{\beta}}^{(k)T} \boldsymbol{x}_i, \hat{m}_i^{(k+1)}\right) : i = 1, \ldots, n\right\}$, and define the monotone function estimate $\hat{\eta}_m^{(k+1)}(t)$ using (3). Select h via a grid search on values $\left\{0.1^* sd\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}\right), 0.2^* sd\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}\right), \ldots, \ sd\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}\right)\right\}$ using leave-one-out cross-validation. For computational convenience, fix h after a small number, say a, of iterations (i.e., when $k = a$).
2. Using data $\left\{\left(\hat{\boldsymbol{\beta}}^{(k)T} \boldsymbol{x}_i, \hat{m}_i^{(k+1)}\right) : i = 1, \ldots, n\right\}$, obtain the derivative data $\left\{\left(\hat{\boldsymbol{\beta}}^{(k)T} \tilde{\boldsymbol{x}}_i, \tilde{y}_i^{(k+1)}\right) : i = 1, \ldots, n-1\right\}$, and define the derivative $\hat{\eta}^{(k+1)'}(t)$ using (3). Select $\tilde{h}$ from the grid $\left\{0.1^* sd\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}\right), 0.2^* sd\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}\right), \ldots, sd\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}\right)\right\}$ using leave-one-out cross-validation. As with h, $\tilde{h}$ is fixed after a iterations.
3. Let the general notation $z^{(k^l)}$ indicate the lth update to $z^{(k)}$. Using approximation (5), obtain data $\left\{\left(\boldsymbol{x}_i^{*(k^1)}, y_i^{*(k^1)}\right) : \ i = 1, \ldots, n\right\}$ and minimize $\hat{Q}_{lin}^{(k^1)}(\boldsymbol{\beta})$ from (6), giving $\hat{\boldsymbol{\beta}}^{(k^1)}$. Repeat this step $m-1$ more times, for a total of m iterations, each time updating the linear approximation (5), so that $\hat{\boldsymbol{\beta}}^{(k^m)} \equiv \hat{\boldsymbol{\beta}}^{(k+1)}$ comes from data $\left\{\left(\boldsymbol{x}_i^{*(k^m)}, y_i^{*(k^m)}\right) : \ i = 1, \ldots, n\right\} \equiv \left\{\left(\boldsymbol{x}_i^{*(k+1)}, y_i^{*(k+1)}\right) : \ i = 1, \ldots, n\right\}$.
4. Cycle through steps 1–3 until $\left\|\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}\right\|$ becomes smaller than a prespecified precision level. The final estimate of η is then obtained by implementing step 1 once more using the converged $\boldsymbol{\beta}$ estimate.

The identifiability constraint is imposed by rescaling $\hat{\boldsymbol{\beta}}^{(k)}$ by $\hat{\beta}_1^{(k)}$ each time step 3 is completed, so it is desirable that β_1 be nonzero to avoid potential numerical problems. To help ensure this in practice, one could first fit a linear model and choose the largest (or most significant) β_j estimate to be that which is subsequently unpenalized and forced to be 1. If in the final model another coefficient is larger, then one could re-run the analysis with that coefficient being the one that is unpenalized and forced to be 1. As suggested by one of the reviewers, one could also consider a sensitivity analysis in which multiple models were fit, each time forcing a different coefficient to be 1.

A value of λ_n must be chosen before $\boldsymbol{\beta}$ can be estimated. Suppose that $\hat{\boldsymbol{\beta}}(\lambda_n)$ and $\hat{\eta}_m(t; \lambda_n)$ are the estimates of $\boldsymbol{\beta}$ and $\eta(t)$, given tuning parameter λ_n. To choose a value of λ_n, we use the Bayes information criterion (BIC) measure of [8]. Specifically, we choose the value of λ_n that minimizes

$$BIC(\lambda_n) = \log\left\{\frac{1}{n}\sum_i^{\nu}\left(y_i - \hat{\eta}_m(\hat{\boldsymbol{\beta}}(\lambda_n)^{\mathrm{T}}\boldsymbol{x}_i;\lambda_n)\right)^2\right\} + \frac{\log(n)}{n}DF_{\lambda_n},$$

where DF_{λ_n} is one less than the number of nonzero values in $\hat{\boldsymbol{\beta}}(\lambda_n)$, because $\hat{\beta}_1$ is forced to be nonzero. To find the optimal λ_n, a grid search is employed.

In the remaining sections, the monotone-constrained method described earlier is referred to as the *constrained* approach.

2.4. Asymptotics

Using the results of Härdle *et al.* [1] and arguments similar to those of Zou [11], it is possible to establish the *oracle* properties for the unconstrained approach. We provide an outline of such an argument here.

Suppose the regularity conditions of Härdle *et al.* [1] hold. Then, by their main theorem, we can rewrite the sum of squares portion of (4) as a sum of three terms, $\tilde{S}$, T and R, where $\tilde{S}$ and T depend only on $\boldsymbol{\beta}$ and h, respectively, and the remainder term R is negligible. Thus, as $\tilde{S}$ is the only term that depends on $\boldsymbol{\beta}$, (4) can be reduced to $\tilde{S}(\boldsymbol{\beta})+\lambda_n\sum_{j=2}^{p} w_j|\beta_j| = n\left\{\boldsymbol{W}_0^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-\frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\right\}^{\mathrm{T}}\left\{\boldsymbol{W}_0^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-\frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\right\}+\lambda_n\sum_{j=2}^{p} w_j|\beta_j|$, where $\boldsymbol{W}_0$ is a $p\times p$ matrix, $\boldsymbol{\beta}_0$ is the true index parameter and $\boldsymbol{Z}$ is an asymptotically normal $N(\boldsymbol{0},\boldsymbol{I})$ p-vector. Now, suppose that $\frac{\lambda_n}{\sqrt{n}}\to 0$, and $\lambda_n n^{(\gamma-1)/2}\to\infty$, where $\gamma\in(0,\frac{3}{5}]$, and let $\boldsymbol{\beta}=\boldsymbol{\beta}_0+\frac{\boldsymbol{u}}{\sqrt{n}}$, where $\|\boldsymbol{u}\|\leqslant C$. From here, by following arguments very similar to Zou [11], the *oracle* properties can be established.

It seems that the *oracle* properties will also hold for $\boldsymbol{\beta}$ estimates from the constrained approach under certain conditions. Specifically, under the conditions of Theorem 2 in [21], we have $\hat{\eta}_m(t) = \hat{\eta}(t) + O_p(n^{-8/15})$, for all t, where $\hat{\eta}_m$ is our monotone estimator of η, and $\hat{\eta}$ is the Nadaraya–Watson kernel-weighted average. Thus, it is possible to reduce the penalized sum of squares for the constrained approach to (4) plus a negligible remainder term. The *oracle* properties for the constrained approach would hold by the same reasoning used for the unconstrained approach.

In practice, it is difficult to verify that the conditions needed for the theory hold. Because a data-driven method (BIC) is used to select the tuning parameter, λ_n, we cannot guarantee the required rate of convergence. Thus, the assumptions $\frac{\lambda_n}{\sqrt{n}}\to 0$ and $\lambda_n n^{(\gamma-1)/2}\to\infty$ may not hold.

2.5. Bootstrap standard errors

Standard errors for our $\boldsymbol{\beta}$ estimates can be obtained via the bootstrap. In particular, for a given data set, we employ the adaptive LASSO-based residual bootstrap approach discussed by Chatterjee and Lahiri [25] to obtain many, say M, bootstrap data sets. A penalized single-index model is then fit on each of these bootstrap data sets, giving M sets of estimates. The estimated standard errors are then obtained by taking the standard deviations of the M estimates for each β_j.

For a given data set, we obtain a residual bootstrap data set as follows. Suppose $\hat{\boldsymbol{\beta}}$ and $\hat{\eta}$ are final estimates of $\boldsymbol{\beta}$ and η for a particular data set. Let $e_i = y_i - \hat{\eta}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x_i})$, $i=1,\ldots,n$, be the residuals for this data set. A residual bootstrap data set is then obtained by replacing y_i with $\hat{\eta}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x_i}) + e_i^*$, $i=1,\ldots,n$, where $\{e_1^*,\ldots,e_n^*\}$ is a random sample (drawn with replacement) from the centered residuals, $e_i - \frac{1}{n}\sum_{i=1}^{n} e_i$, $i=1,\ldots,n$. The covariate matrix remains the same across the bootstrap data sets.

On the basis of additional simulations (results not given), creating residual bootstrap data sets using permuted (sampled without replacement) residuals gives nearly identical results to those shown in Table II. As noted by one reviewer, in practice, the interpretation for the standard errors can be awkward, particulary in cases where a number of covariates are highly correlated. In such cases, one might expect the distribution of these estimates to be a mixture of a continuous distribution and a point mass at zero. Thus, the estimates are a product of both selection and estimation, which can make interpretation difficult. This may be due to the known shortcomings of the adaptive LASSO for highly correlated predictors. If one believes that a number of covariates may be highly correlated, an alternative approach, such as the adaptive elastic net, may perform better and may lead to bootstrap standard error estimates based on a smaller number of zeros.

We generally suggest that one reselect λ_n with each bootstrap data set; however, our simulations (results not shown) suggest that holding λ_n fixed across bootstrap data sets gives standard error

estimates that are nearly identical to those found by reselecting λ_n for each bootstrap data set. Thus, it may be reasonable to consider fixed-λ_n bootstrap standard errors if reselecting λ_n for each bootstrap data set is too computationally burdensome.

3. Simulations

A simulation study was performed using R software to evaluate the performance of the proposed methods. To comply with the conditions in Section 2.4, a value of $\frac{3}{5}$ was chosen for γ for adaptive LASSO. Additionally, for each example, a large test set ($n = 10,000$) was generated, and final $\boldsymbol{\beta}$ estimates from each of the simulated data sets were used to calculate the mean squared error (MSE) for this large test set. To evaluate the performance of all methods considered, we recorded the number of correct and incorrect zero values in $\hat{\boldsymbol{\beta}}$, as well as the total proportion of $\hat{\beta}_j$'s correctly estimated as zero or nonzero for each data set. The average of these proportions across all simulated data sets is referred to in Table I as the *relative frequency correct*. We also computed the false discovery rate (FDR), which is the percentage of

Table I. Simulation results: variable selection performance.

		Avg. no. $\hat{\boldsymbol{\beta}} = 0$*			
Method	Rel. freq Correct	Correct	Incorrect	FDR	Mean test MSE (×100)
Case (i)[1]					
Cons.	0.92	6.22	0.00	0.21	4.93
Uncons.	0.85	5.47	0.01	0.34	5.63
Cons. oracle	1.00	7.00	0.00	0.00	4.74
Uncons. oracle	1.00	7.00	0.00	0.00	4.96
Case (ii)[2] **					
Cons.	0.89	6.09	0.17	0.24	10.87
Uncons.	0.75	4.68	0.16	0.45	12.84
Cons. oracle	1.00	7.00	0.00	0.00	10.07
Uncons. oracle	1.00	7.00	0.00	0.00	10.56
Case (iii)[3]					
Cons.	0.86	6.25	0.65	0.24	4.64
Uncons.	0.73	4.83	0.56	0.47	5.51
Cons. oracle	1.00	7.00	0.00	0.00	4.45
Uncons. oracle	1.00	7.00	0.00	0.00	4.65
Case (iv)[4]					
Cons.	0.88	6.15	0.32	0.24	4.79
Uncons.	0.82	5.45	0.27	0.36	5.38
Cons. oracle	1.00	7.00	0.00	0.00	4.51
Uncons. oracle	1.00	7.00	0.00	0.00	4.78
Case (v)[5]					
Cons.	0.94	53.70	0.10	0.53	5.64
Uncons.	0.87	49.36	0.13	0.74	7.16
Cons. oracle	1.00	57.00	0.00	0.00	4.74
Uncons. oracle	1.00	57.00	0.00	0.00	4.96

'Oracle' indicates true zero $\boldsymbol{\beta}$ values known. η is estimated in all methods.
*Average number of variables dropped in final model.
**Required 101 simulated data sets because of numerical problems.
[1] $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0)^{\mathrm{T}}$, $\mathrm{Corr}(x_{ij}, x_{ik}) = 0,\ j \neq k,\ \sigma = 0.20$.
[2] Same as case (i), but $\sigma = 0.3$.
[3] Same as case (i), but $\beta_7 = -0.2$.
[4] Same as case (i), but $\mathrm{Corr}(x_{ij}, x_{ik}) = 0.5,\ j \neq k$.
[5] Same as case (i), but $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0, \mathbf{0}_{1\times 50})^{\mathrm{T}}$.

Table II. Performance of standard error estimates.

	$\hat{\beta}_2$		$\hat{\beta}_7$	
Method	SD	SE (SE_{sd})	SD	SE (SE_{sd})
Cons.	0.25	0.20 (0.08)	0.20	0.18 (0.06)
Uncons.	0.28	0.34 (0.41)	0.28	0.27 (0.26)

Required 102 simulated data sets because of numerical problems.

nonzero $\hat{\boldsymbol{\beta}}$ values that should have been zero. For each data set, the optimal tuning parameter value λ_n was chosen from the grid $\{0, 0.01, \ldots, 0.25\}$ using BIC.

3.1. *Examples*

For all simulations, 100 data sets of size 100 were generated from the model

$$y_i = \left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i\right)^3 + \epsilon_i,$$

where $\boldsymbol{x}_i$'s were Unif$[-\frac{1}{2}, \frac{1}{2}]$ and error terms were normal with mean zero and variance σ^2. We considered five different cases:

(i) $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{x}_i$'s independent, and ϵ's independent with $\sigma = 0.20$;
(ii) Same as case (i), but with $\sigma = 0.30$;
(iii) Same as case (i), but with $\boldsymbol{\beta}$ changed to $(1, 0.8, 0, 0, 0, 0, -0.2, 0, 0, 0)^{\mathrm{T}}$;
(iv) Same as case (i), but with $\mathrm{Corr}(x_{ij}, x_{ik}) = 0.5,\ j \neq k$;
(v) Same as case (i), but with an additional 50 noise covariates, so that $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0, \mathbf{0}_{1\times 50})^{\mathrm{T}}$.

From Table I, we can see that, in all cases, the constrained approach shows noticeably better reduction in model complexity and smaller FDR than the unconstrained approach. Additionally, the constrained approach has mean test MSEs that are smaller and closer to the corresponding oracle test MSEs than the unconstrained approach. Reduction in model complexity for the constrained approach appears to be reasonably insensitive to the changes in simulation settings considered previously; however, the unconstrained approach appears to suffer in this regard, especially when true parameter values are decreased or error standard deviation is increased.

Additional simulations were implemented to evaluate the performance of the proposed methods under alternative monotonic functions, η (results not shown). In particular, we considered a linear function and two spline functions; one resembling the cubic function from the aforementioned examples, but with two knots chosen to create a wider 'flat' section around the origin and one that is constant to the left of the origin and quadratic to the right. As expected, both methods performed well in the linear case. In the cubic spline case, reduction in model complexity was good, but mean test MSE became noticeably larger, and in the case of the constant spline with the quadratic knot, mean test MSE was good, but reduction in model complexity was noticeably worse. Thus, as one might expect, the proposed methods are less useful in cases where η contains large sections that are nearly flat, or exactly constant.

To evaluate the performance of our standard error estimates, residual bootstrap standard errors (based on 100 bootstrap data sets) were calculated for case (i). Let SD denote the standard deviation of the 100 β_j estimates, $j = 1, \ldots, p$. Additionally, let SE and SE_{sd} denote the mean and the standard deviation of the 100 estimated SEs respectively. Looking at Table II, we can see that the standard error estimates appear to perform reasonably well, although they sometimes slightly underestimate or overestimate the true values.

To demonstrate the ability of penalized monotone single-index models to capture nonlinear relationships, we computed $\hat{\eta}_m$ values across a fine grid of input values and averaged these $\hat{\eta}_m$'s across the 100 data sets in case (i). These average values can be found in Figure 1, along with the true function η and 90% empirical pointwise confidence bands for $\hat{\eta}_m$. As we can see, the monotone function estimate $\hat{\eta}_m$ appears to closely follow the true function.

Because we are interested in using the proposed methods to identify subgroups, we also compared 'enhancement' classification between the two methods for case (i). For this comparison, we consider a subject to be enhanced if $\eta(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) > 0$. On average, 88% of subjects identified as enhanced by the

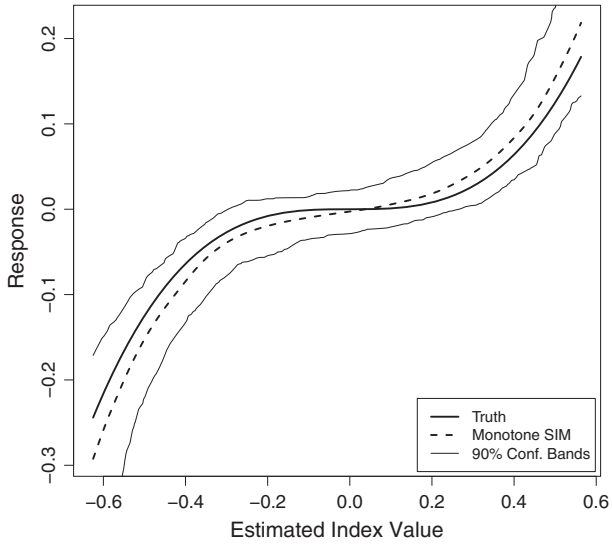

Figure 1. Average $\hat{\eta}_m$ values from 100 simulations.

constrained approach were truly enhanced, whereas for the unconstrained approach, only 73% were correctly identified on average. Thus, the constrained approach may be advantageous for applications to subgroup identification.

The two methods require approximately the same amount of time to complete a single iteration of our algorithm for a given value of λ. However, for some data sets, the constrained approach requires more iterations to achieve the same degree of convergence as the unconstrained approach. For example, for case (i) of our simulations, the median run time for a data set for the constrained approach was approximately 64% longer than that for the unconstrained approach.

4. Example data

In this example, we apply the proposed methods to the Eli Lilly data in [14], which come from a randomized, double-blinded clinical trial in patients with a critical illness in the intensive care unit conducted over a decade ago. We consider 1019 individuals; of whom, 512 received the experimental treatment in addition to the standard of care. The remaining patients received placebo with the standard of care. The intervention is a drug that is intended to improve survival in patients with a critical illness, and the endpoint was survival at 28 days post-randomization to treatment/placebo. We consider 58 covariates analyzed by Foster *et al.* [14], which include demographic, laboratory, medical history and questionnaire data. Of these, 9 are binary, 22 are regarded as continuous and 27 are dummy variables coming from subdivision of 12 categorical variables.

In [14], a random forest was used to obtain two predicted probabilities, $\hat{P}_{1i}$ and $\hat{P}_{0i}$, for each individual, where P_{1i} is the probability of survival at 28 days post-randomization for subject i if that individual had received treatment and P_{0i} is that if subject i had received placebo. The estimation of these probabilities was motivated by the fact that the methods of Foster *et al.* [14] were designed to identify subgroups of enhanced treatment effect in randomized clinical trial data. Therefore, a new outcome representing the treatment effect for person i, $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$, $i = 1, \ldots, n$, was subsequently defined, because individuals in such a subgroup should ideally have values of P_{1i}, which are much larger than P_{0i}. Then, a single regression tree was fit using Z as the outcome and the covariates as predictors. This tree identified subgroups of enhanced treatment effect, which depended on age at admission, baseline creatinine clearance, baseline interleukin 6 and hypertension (yes, no or unknown). This method was referred to by Foster *et al.* [14] as 'Virtual Twins'.

Using Z as the outcome and the 58 covariates as predictors, we fit penalized single-index models with and without monotonicity constraints. All covariates were standardized in this analysis because of large differences in scale, and age at admission was chosen to be the first column of $\boldsymbol{X}$, as its corresponding initial estimate was the largest and most significant value of $\hat{\boldsymbol{\beta}}_{\text{init}}$. It should be noted that this analysis was

also performed with baseline creatinine clearance as the first column (results now shown), and the same six additional covariates were chosen, along with one other. The relative magnitude of the coefficients in this analysis were similar for most variables. Results from these models (with age at admission as first column of $\boldsymbol{X}$) can be found in Table III. Estimates for the constrained and unconstrained approaches were fairly similar, although an additional covariate, baseline index of independence in activities of daily living (ADL) [26], was included by the constrained approach.

In addition to $\boldsymbol{\beta}$ estimates, we computed bootstrap standard errors by using 300 bootstrap samples. Because less important covariates will tend to be removed from the model in most bootstrap samples, resulting in many zero bootstrap estimates, we expect such covariates to have very small bootstrap standard errors.

The six covariates selected by both methods were age at admission, baseline central lab platelet count, baseline creatinine clearance, baseline interleukin 6 (log scale), number of baseline organ failures and pre-infusion acute physiology and chronic health evaluation II (APACHE II) score, of which age at admission, creatinine clearance and interleukin 6 were also selected by the Virtual Twins method. Plots of the data (from the constrained approach) and the final η estimates can be found in Figure 2. We can see that both estimates of η are reasonably close, with the constrained estimate being noticeably more smooth. From Figure 2, we can see that the predicted region of enhanced treatment effect consists of $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}$ values, which are larger than approximately -2, with the degree of enhancement increasing as $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}$ becomes larger. The constrained and unconstrained approaches identified 847 and 864 subjects as being enhanced, respectively, and of the 864 identified by the unconstrained approach, 845 were also identified

Table III. Estimates for Eli Lilly data.

Variable	Unconstrained estimate	SE	Constrained estimate	SE
Age	1.00	—	1.00	—
ADL[1]	—	—	−0.13	0.04
Platelet count	−0.12	0.03	−0.19	0.08
Creat. clear.	−0.70	0.17	−0.81	0.21
Interleukin 6	0.60	0.11	0.70	0.13
# Organ fail.	0.14	0.06	0.23	0.11
APACHE II[2]	0.24	0.09	0.33	0.14

[1]Baseline index of independence in activities of daily living (ADL).
[2]Pre-infusion acute physiology and chronic health evaluation II score.

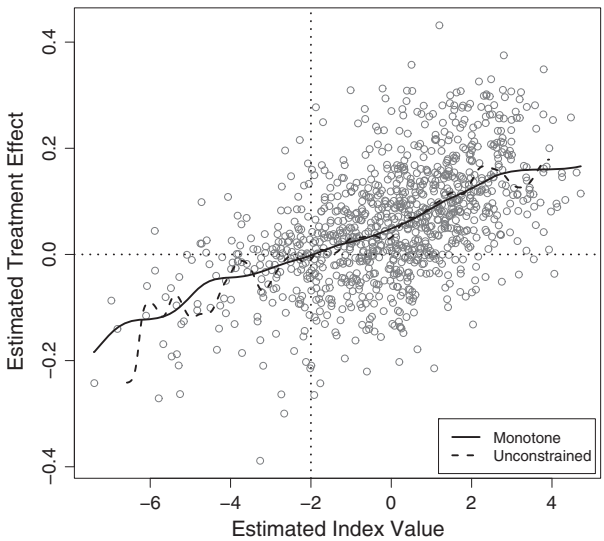

Figure 2. Estimates of function $\hat{\eta}(\cdot)$ from Eli Lilly data. Index values in the plotted data are $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}$, where $\hat{\boldsymbol{\beta}}$ comes from the constrained approach, and treatment effect estimates are the Z values from Virtual Twins procedure. Those points to the right of the vertical dotted line would be considered 'enhanced' on the basis of this analysis.

by the constrained approach. Furthermore, for the constrained model, older individuals and those with higher baseline IL-6 respond very well to treatment, and patients with lower baseline creatinine clearance show a greater treatment differential. The findings from this analysis are reasonably consistent with the original conclusions from this trial, which suggested that patients who had higher risk factors for mortality responded better to the treatment.

As both fits suggest a relationship that is close to linear, an adaptive LASSO penalized linear model was also fit (results not shown), once using the default tuning parameter selection settings (10-fold cross-validation using squared error loss) in the R *glmnet* package, and once using BIC to select the tuning parameter. The model resulting from the default tuning parameter selection settings contained 24 covariates, whereas the model selected using BIC contained seven covariates. Although BIC is known to give smaller models than cross-validation, this dramatic difference in model complexity was mildly surprising to us. On the basis of the results of the linear model (using BIC), it appears that the single-index models may not have added much compared with a linear model in this case.

5. Discussion

We proposed the use of adaptive LASSO variable selection for monotone single-index models and showed that it performs well in a variety of situations. The constrained approach noticeably outperformed the unconstrained and has the advantage of more straightforward interpretation. A linear approximation to η via Taylor series was also proposed, thus allowing for the use of standard LASSO algorithms, such as coordinate descent, which have been shown to perform well. In addition, we suggested the use of residual bootstrap standard errors for $\boldsymbol{\beta}$ estimates and showed that they perform reasonably well in simulations.

We argue that the unconstrained adaptive LASSO penalized single-index model estimates possess the *oracle* properties when η is estimated using the Nadaraya–Watson formula. Additionally, we briefly argue that, following the results of Mammen [21], the *oracle* properties may also hold for the constrained approach, and it would be interesting to investigate this more formally. Furthermore, the proof outlined in Section 2.4 assumes that $\boldsymbol{\beta}$ is in a $\sqrt{n}$-neighborhood of the true value, which is likely true given that the initial estimator of $\boldsymbol{\beta}$ is in a $\sqrt{n}$-neighborhood of $\boldsymbol{\beta}_0$.

Our method of obtaining a monotone function estimate is very similar to that of Friedman and Tibshirani [19]. They suggested that it may be possible to improve the estimation of the monotone penalized single-index model if one considers 'one-step' monotone function estimates, such as those suggested by He and Shi [16] and Ramsay [17]. This is worthy of further investigation.

The adaptive LASSO penalty was chosen for convenience; however, one may wish to consider other penalty functions. For instance, as noted by a reviewer, the adaptive elastic net can often outperform the adaptive LASSO approach, particularly when covariates are highly correlated. Note that the linear approximation to the function η does not involve the penalty function. Thus, the proposed method and algorithm could easily be modified if one wished to use a different penalty function, such as the SCAD or adaptive elastic net.

Acknowledgements

This research was partially supported by a grant from Eli Lilly, grants CA083654 and AG036802 from the National Institutes of Health and grant DMS-1007590 from the National Science Foundation.

References

1. Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *The Annals of Statistics* 1993; **21**(1):157–178.
2. Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 2002; **97**(460):1042–1054.
3. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 1993; **58**(1–2):71 –120.
4. Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. *Journal of the American Statistical Association* 1997; **92**(438):477–489.
5. Xia Y, Tong H, Li WK, Zhu LX. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B* 2002; **64**(3):363–410. http://ideas.repec.org/a/bla/jorssb/v64y2002i3p363-410.html.
6. Xia Y, Härdle WK. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis* 2006; **97**(5):1162–1184. http://EconPapers.repec.org/RePEc:eee:jmvana:v:97:y:2006:i:5:p:1162-1184.

7. Kong E, Xia Y. Variable selection for the single-index model. *Biometrika* 2007; **94**(1):217–229.
8. Liang H, Liu X, Li R, Tsai CL. Estimation and testing for partially linear single-index models. *Annals of Statistics* 2010; **38**(6):3811–3836.
9. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996; **58**(1):267–288.
10. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**:1348–1360.
11. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 2006; **101**:1418–1429.
12. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* 2009; **37**(4):1733–1751.
13. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer: New York, NY, 2009.
14. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 2011:2867–2880. http://dx.doi.org/10.1002/sim.4322.
15. Li KC, Duan N. Regression analysis under link violation. *The Annals of Statistics* 1989; **17**(3):1009–1052.
16. He X, Shi P. Monotone b-spline smoothing. *Journal of the American Statistical Association* 1998; **93**(442):643–650.
17. Ramsay JO. Monotone regression splines in action. *Statistical Science* 1988; **3**(4):425–441.
18. Hall P, Huang LS. Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics* 2001; **29**(3):624–647.
19. Friedman J, Tibshirani R. The monotone smoothing of scatterplots. *Technometrics* 1984; **26**(3):243–250.
20. Mukerjee H. Monotone nonparametric regression. *The Annals of Statistics* 1988; **16**(2):741–750.
21. Mammen E. Estimating a smooth monotone regression function. *The Annals of Statistics* 1991; **19**(2):724–740.
22. Barlow RE, Bartholomew RJ, Bremner JM, Brunk HD. *Statistical Inference Under Order Restrictions*. John Wiley and Sons: New York, 1972.
23. Fu WJ. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* 1998; **7**(3):397–416.
24. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. *The Annals of Applied Statistics* 2007; **1**(2):302–332.
25. Chatterjee A, Lahiri SN. Bootstrapping lasso estimators. *Journal of the American Statistical Association* 2011; **106**(494):608–625. DOI: 10.1198/jasa.2011.tm10159. http://pubs.amstat.org/doi/abs/10.1198/jasa.2011.tm10159.
26. Katz S, Akpom CA. Index of ADL. *Medical Care* 1976; **14**(5):116–118.