Journal of Computer and Communications, 2014, 2, 1-14
Published Online September 2014 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2014.211001

A Graduated Nonconvex Regularization for Sparse High Dimensional Model Estimation

Thomas F. Coleman, Yuying Li

[1]Departmentof Combinatorics and Optimization, University of Waterloo, Waterloo, Canada
[2]Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada
Email: tfcoleman@uwaterloo.ca, yuying@uwaterloo.ca

Received May 2014

Abstract

Many high dimensional data mining problems can be formulated as minimizing an empirical loss function with a penalty proportional to the number of variables required to describe a model. We propose a graduated non-convexification method to facilitate tracking of a global minimizer of this problem. We prove that under some conditions the proposed regularization problem using the continuous piecewise linear approximation is equivalent to the original l_0 regularization problem. In addition, a family of graduated nonconvex approximations are proposed to approximate its l_1 continuous approximation. Computational results are presented to illustrate the performance.

Keywords

Sparse Model Estimation, Variable Selection l_0 Regularization

1. Introduction

Sparsity is a desired property in model estimation since it often leads to better interpretability and out-of-sample predictability. In [1], a risk bound is established for the model selection by minimizing the empirical loss function penalized by the number of variables needed to describe the model. In this context, model dimension is the number of unknown variables; sparsity refers to a small number of variables selected to define the model. Thus sparse model estimation is also sometimes referred to as variable selection.

Selecting a model with a small number variables can be formulated as minimizing an empirical loss function with a penalization for the number of nonzero variables; this is referred to as l_0-regularization. Unfortunately this is a NP-hard global optimization problem, see, e.g., [2] [3]. Relaxation of cardinality regularization has long been used as a way to approach the sparse model selection problem, see, e.g., [4] [5]. Due to its computational simplicity, regularization based on the l_2 norm is popular in practice. This is referred to as ridge regression.

It has also long been recognized that l_1 regularization often leads to sparsity in solutions, see, e.g., [6]-[11] [29]. Recent compressive sensing theory, e.g., [12]-[14]), formally establishes that, under a certain restricted isometry property (RIP) on a $n \times m$ sensing matrix $\Phi, n \le m$ a sparse vector x can be reconstructed exactly from $y = \Phi x$. In addition, computational methods have been developed to obtain solutions efficiently, see, e.g., [15]-[17].

How to cite this paper: Coleman, T.F. and Li, Y.Y. (2014) A Graduated Nonconvex Regularization for Sparse High Dimensional Model Estimation. *Journal of Computer and Communications*, **2**, 1-14. http://dx.doi.org/10.4236/jcc.2014.211001

Despite its success, the l_1 regularization approach can potentially lead to model bias for model estimations, see, e.g., [18]-[21]. In [18], a smoothly clipped absolute deviation (SCAD) penalty is proposed using a nonconvex regularization to avoid the potential bias. Iterative methods are proposed in [23] [24] to approximate a local minimizer of the SCAD penalized loss. It is proposed in [18] that a proper regularization function for model estimation should be chosen with three objectives: avoid *bias* in the resulting estimator, achieve sparsity in the estimated model, and, finally, achieve *continuity or stability* in model prediction. The l_1 regularization function is able to achieve sparsity and smoothness but can lead to model bias. In addition, using l_1 regularization does not always lead to the sparsest model which fits the data to a specified accuracy. Recently computational methods have also been proposed to iteratively solve l_0 regularization problems, see, e.g., [25] [26].

The main objective of this paper is to devise a computational method to minimize the empirical loss function with a penalization for the number of nonzero variables. We approximate the counting indicator function by a continuous piecewise linear function. We show mathematically that a continuous piecewise linear approximation has appealing theoretical properties. To facilitate tracking a global minimizer, this continuous piecewise function is further approximated by a family of continuously differentiable piecewise quadratic functions which are indexed by a parameter controlling the degree of nonconvexity in the approximation. Starting from an initial convex function, a sequence of increasingly more nonconvex approximate problems are solved, using the solution to the previous problem as a starting point for the next approximate problem. In addition, each approximation can be regarded as a regularization function for model estimation; each approximation behaves similar to the l_1 function near the origin (ensuring sparsity), the l_0 function asymptotically (avoiding bias), and continuously differentiable everywhere (for smoothness). We illustrate the efficacy of the proposed method in determining a sparse model for data fitting, computational efficiency of the approach, and the effect of the parameters on the gradual non-convex approximations.

2. Continuous Approximation to L_0

Assume that $f(x)$ measures the empirical error based on the given data. We want to solve the following l_0 regularization problem

$$\min_{x\in R^n} f(x)+\mu\sum_{i=1}^{n}\Lambda(x_i) \tag{2.1}$$

where the counting indicator function $\Lambda(\cdot)$ is

$$\Lambda(x_i)=\begin{cases}1 & \text{if } x_i\neq 0\\ 0 & \text{otherwise}\end{cases}$$

and $\mu>0$ is a penalty parameter, balancing the objectives of minimization of $f(x)$ and sparsity in the solution.

Unfortunately standard optimization methods cannot be applied to the data fitting problem (2.1) since the cardinality function $\Lambda(z)$ is discontinuous and nonconvex. Convex relaxations have been proposed for (2.1), see, e.g., [5] [27]; these relaxations can be sub-optimal since (2.1) is nonconvex. In [28] the counting indicator $\Lambda(z)$ is approximated, in the context of image processing, by the piecewise quadratic $\hat{h}_\lambda(z)$ and solves

$$\min_{x\in R^n} f(x)+\mu\sum_{i=1}^{n}\hat{h}_\lambda(x_i) \tag{2.2}$$

where

$$\hat{h}_\lambda(z)=\begin{cases}\lambda z^2 & \text{if } |z|\le\dfrac{1}{\sqrt{\lambda}}\\ 1 & \text{otherwise}\end{cases}$$

and $\lambda>0$ is a small resolution parameter. We propose to approximate the discontinuous counting indicator function $\Lambda(z)$ by the following continuous l_1 penalty function $h_\lambda(z)$,

$$h_\lambda(z)=\begin{cases}\sqrt{\lambda}|z| & \text{if } |z|\le\dfrac{1}{\sqrt{\lambda}}\\ 1 & \text{otherwise}\end{cases}$$

Figure 1 illustrates $h_\lambda(z)$ and $\hat{h}_\lambda(z)$ Using $h_\lambda(z)$, the l_0 regularization problem (2.1) is approximated by

$$\min_{x \in R^n} \tilde{p}(x;\lambda) \overset{\text{def}}{=} f(x) + \mu \sum_{i=1}^{n} h_\lambda(x_i) \tag{2.3}$$

Next we establish mathematically that there exists a finite threshold parameter $\bar{\lambda}$ such that a solution to (2.3) is a solution to the original l_0 penalty problem (2.1) for all $\lambda \geq \bar{\lambda} > 0$.

Assumption 2.1. Assume $f(x)$ is twice continuously differentiable on an open set $D \subseteq R^n$ and there exists $K > 0$ such that $|\nabla f_i(x)| \leq K$ for all $x \in D$, $i = 1, \cdots, n$.

Lemma 2.1. *Suppose Assumption* 2.1 *holds. Let* $x_* \in D$ *be a local minimizer of* $\tilde{p}(x;\lambda)$ *with* $\lambda \geq \frac{K^2}{\mu^2}$, *Then, for each index* i, *either* $(x_*)_i = 0$ *or* $|(x_*)_i| \geq \frac{1}{\sqrt{\lambda}}$. *Hence either* $h_\lambda((x_*)_i) = 0$ or $h_\lambda((x_*)_i) = 1$ *for* $i = 1, \cdots, n$

Proof. For notational simplicity, in this proof, we denote $\tilde{p}(x;\lambda)$. simply as $\tilde{p}(x)$.

Suppose the contrary, *i.e.*, for some index *i*,

$0 < |(x_*)_i| < \frac{1}{\sqrt{\lambda}}$. We note that in this situation the derivative of $\tilde{p}(x)$ with respect to x_i exists at x_* because $f(x)$ is everywhere differentiable on D and since by assumption$x_*$is not a cusp point, $h_\lambda(x_i)$ is also differentiable at $x_i = (x_*)_i$. Then, by local optimality $\nabla \tilde{p}_i(x_*) = 0$, *i.e.*,

$$(\nabla f_i(x_*))_i + \mu\sqrt{\lambda} = 0 \text{ or } (\nabla f_i(x_*))_i - \mu\sqrt{\lambda} = 0$$

But $\mu\sqrt{\lambda} > K$ and $|\nabla f_i(x_*)| \leq K$, which is contradictory. □

Lemma 2.2. *Let* x_* *be a local minimizer for* $\tilde{p}(x;\bar{\lambda})$ *for some* $\bar{\lambda} > \frac{K^2}{\mu^2}$. *Then* x_* *is a local minimizer of* $\tilde{p}(x;\lambda)$ for all $\lambda \geq \bar{\lambda}$.

Proof. Suppose that x_* is not a local minimizer of $\tilde{p}(x;\lambda)$ for some $\lambda = \lambda^+ > \bar{\lambda}$. Lemma 2.1 states that the components of x_* are partitioned into two sets: *i.e.*, for each index*i*either $(x_*)_i = 0$ or $|(x_*)_i| \geq \frac{1}{\sqrt{\lambda}}$. Since $\lambda^+ > \bar{\lambda}$, the same partition applies with respect to λ^+. Therefore

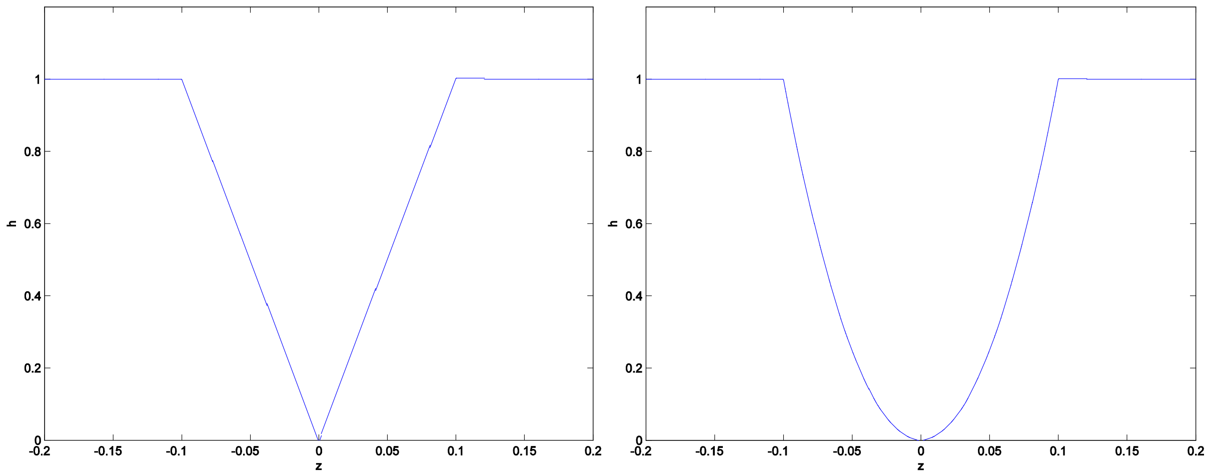

Figure 1. Quadratic (left subplot) and piecewise linear (right sunplot) approximations to the counting indicator function $\Lambda(z)$ with $\lambda = 100$.

$$\sum_{i=1}^{n} h_{\bar{\lambda}}\left((x_*)_i\right) = \sum_{i=1}^{n} h_{\lambda^+}\left((x_*)_i\right)$$

It follows from the definition for $\tilde{p}(x;\lambda)$ that x_* cannot be a minimizer of $\tilde{p}(x;\bar{\lambda})$, a contradiction. ■

Theorem 2.1. If $x_* \in D$ *is a strong local minimizer for* $\tilde{p}(x;\lambda)$ *for some* $\lambda > \frac{K^2}{\mu^2}$. *Then* x_* *is a local minimizer of* $p(x) = f(x) + \mu\sum_{i=1}^{n}\Lambda(x_i)$.

Proof. By Lemma 2.2, x_* is a local minimizer of $\tilde{p}(x;\lambda)$ for all λ sufficiently large. We show that x_* is a local minimizer of $p(x)$ by contradiction. Suppose that x_* is not a local minimizer of $p(x)$. Then there exists a sequence x_k converging to $x_*, x_k \neq x_*$ and $p(x_k) < p(x_*)$. This implies that $f(x_k) < f(x_*)$ for sufficiently large k. Since, for sufficiently large k,

$$\tilde{p}(x_k;\lambda) < p(x_k)$$

And $\tilde{p}(x_*;\lambda) = p(x_*)$, we have that $\tilde{p}(x_k;\lambda) < \tilde{p}(x_*;\lambda)$. This contradicts that x_* is a strong local minimizer of $\tilde{p}(x;\lambda)$ *for some* $\lambda > \frac{K^2}{\mu^2}$. □

If one uses $\hat{h}_\lambda(z)$ to approximate $\Lambda(z)$, a minimizer x_* to the approximation problem (2.2) will generally not be a minimizer to the l_0 regularization problem (2.1) unless either $(x_*)_i = 0$ or $|(x_*)_i| \geq \frac{1}{\sqrt{\lambda}}$, which typically does not hold for any $\lambda > 0$. Theorem 2.1 indicate that our proposed approximation $h_\lambda(z)$ is superior to $\hat{h}_\lambda(z)$ in solving the l_0 regularization problem (2.1).

3. Graduated Non-Convexification

We now address a couple of additional challenges. Firstly, $h_\lambda(z)$ is not differentiable everywhere, which is also the case for $\hat{h}_\lambda(z)$. Secondly, $\sum_{i=1}^{n} h_\lambda(x_i)$ is not convex; thus problem (2.3) has many local minimizers. Assume $\mathcal{A} \subset \mathcal{N} = \{1,2,\cdots,n\}$. Consider the following

$$\min_{x\in R^n} f(x) \text{ subject to } \mathrm{x}_i = 0, i \in \mathcal{A} \tag{3.1}$$

Then any local minimizer of (3.1) is a local minimizer of (2.1) for a fixed $\mu > 0$.

For a given $\mu > 0$, computing a *global* minimizer of the l_0 regularization problem (2.1), or the proposed approximation (2.3), is NP-hard. However, the quality of the estimated model depends on being able to find, as much as possible, a sufficiently good approximation to the global minimizer of (2.1). Next we develop a computational method to produce a good approximation to the global minimizer of the piecewise linear minimization problem (2.3).

Assume that the empirical loss function $f(x)$ is convex.

Hence the nonconvexity comes from the counting indicator function. In [28], a graduated non-convexification process is proposed, in the context of image processing, in an attempt to find the global minimizer of the piecewise quadratic approximation (2.2) as follows. The continuous function $\hat{h}_\lambda(z)$ is approximated using a family of continuously differentiable piecewise quadratic functions $\hat{g}_\lambda(z)$, where

$$\hat{g}_\lambda(z;\ \rho) = \begin{cases} \lambda z^2 & \textit{if } |z| \leq \kappa \\ 1 - \frac{\rho}{2}\left(|z| - \gamma\right)^2 & \textit{if } \kappa \leq |z| \leq \gamma \\ 1 & \text{otherwise} \end{cases} \tag{3.2}$$

$$\gamma = \sqrt{\frac{2}{\rho} + \frac{1}{\lambda}}, \quad \kappa = \frac{1}{\lambda\gamma}. \tag{3.3}$$

Here $\rho > 0$ is a parameter indexing the family of approximations to $\hat{h}_\lambda(z)$. The function $\hat{g}_\lambda(z;\rho)$ is a piece wise quadratic, with the concave quadratic

$$1-\frac{\rho}{2}\left(|z|-\gamma\right)^2$$

when $z\in(\kappa,\gamma)$. Note that, for any $\rho>0$, we have $\gamma>\frac{1}{\sqrt{\lambda}}$. Thus $\kappa<\frac{1}{\sqrt{\lambda}}<\gamma$. In addition the function $\hat{g}_\lambda\left(x_i;\rho\right)$ is continuously differentiable and symmetric with respect to z. Left subplot in **Figure 2** graphically illustrates the function $\hat{g}_\lambda\left(z;\rho\right)$. Substituting $\hat{g}_\lambda\left(z;\rho\right)$ for $\hat{h}_\lambda\left(z\right)$, we have

$$\min_{x\in R^n}\left(f\left(x\right)+\mu\sum_{i=1}^{n}\hat{g}_\lambda\left(x_i;\rho\right)\right) \tag{3.4}$$

Let $\left\{\rho_k\right\}$ be a given monotonically increasing sequence which converges to $+\infty$. In [28], the global minimizer of the piecewise quadratic minimization (2.2) is tracked by solving a sequence of problems (3.4) indexed by a monotonically increasing sequence $\left\{\rho_k\right\}$, using the solution of the $\left(k-1\right)$th problem as the starting point for the *k*th problem. Initially, the minimizer for the empirical error function, which corresponds to (3.4) with $\mu=0$, is computed. As $\rho_k\to 0$, $\kappa_k\to 0$, and $\gamma_k\to+\infty$. Thus problem (3.4) approaches the minimization of the empirical error function min $f\left(x\right)$ as $\rho_k\to 0$. As ρ_k increases, the second order derivative of the quadratic function in $\left[\kappa,\gamma\right]$ equals $-\rho$, which becomes increasingly more negative, gradually introducing non-convexity. In addition, as $\rho_k\to+\infty,\gamma_k,\kappa_k\to\frac{1}{\sqrt{\lambda}}$, and the function $\hat{g}_\lambda\left(z;\rho_k\right)$ approaches $\hat{h}_\lambda\left(z\right)$. Geometrically, the convex empirical error function $f\left(x\right)$ is gradually deformed to the concave function, with thecomputed solution sequence following a path from the minimizer of $f\left(x\right)$ to, ideally, the global minimizer of $f\left(x\right)+\mu\sum_{i=1}^{n}\hat{h}_\lambda\left(x_i\right)$.

Similarly we can design a family of approximations to track the global minimizer of $f\left(x\right)+\mu\sum_{i=1}^{n}h_\lambda\left(x_i\right)$ in (2.3). For sparsity, we want to retain the l_1 segment $\sqrt{\lambda}|z|$ in a neighborhood of $z=0$. Outside this neighborhood, we require the approximation to be continuously differentiable. Finally, similar to $\hat{g}_\lambda\left(z;\rho_k\right)$, we ensure that this approximation gradually approaches the piecewise linear $h_\lambda\left(z\right)$.

Based on $\hat{g}_\lambda\left(z;\rho\right)$ in (3.2), we construct a family of approximations for $h_\lambda\left(z\right)$ with the desired properties as follows. Let γ and κ be defined in (3.3). First we select two break points ξ and η so that

$$0<\xi<\eta<\kappa \quad\text{and}\quad \eta-\xi=\kappa-\eta.$$

In addition, ξ and η monotonically increase to $\frac{1}{\sqrt{\lambda}}$ as ρ converges to $+\infty$; this property also holds

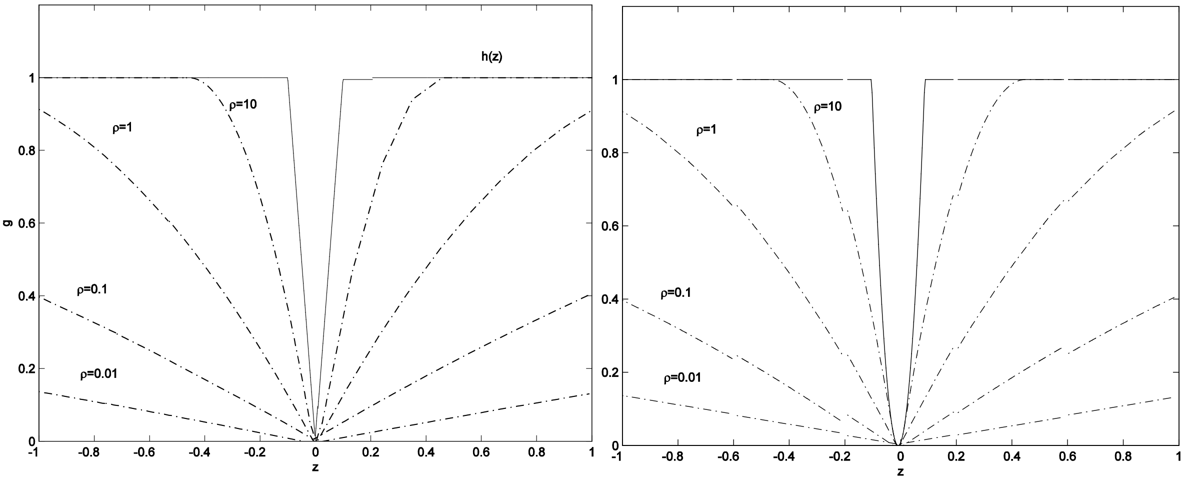

Figure 2. Graduated nonconvex approximation $\hat{g}_\lambda\left(z;\rho\right)$ (left subplot) and $g_\lambda\left(z;\rho\right)$ (right subplot).

for κ. Then we construct the unique quadratic spline $S(z)=\{s_1(z),s_2(z)\}$ on $[\xi,\eta]$:

$$S(z)=\begin{cases} s_1(z) & \text{when } \xi\le z\le\eta \\ s_2(z) & \text{when } \eta\le z\le\kappa \end{cases}$$

which satisfies the following boundary conditions: the function values and the derivative values at ξ and η are given by $\frac{1}{\gamma}z$ and $1-\frac{\rho}{2}(|z|-\gamma)^2$ respectively, *i.e.*,

$$\begin{aligned} &S(\xi)=\frac{1}{\gamma}\xi,\ \ S'(\xi)=\frac{1}{\gamma}, \\ &S(\kappa)=1-\frac{\rho}{2}(\kappa-\gamma)^2,\ \ S'(\kappa)=-\rho(\kappa-\gamma). \end{aligned} \tag{3.5}$$

We now approximate the nondifferentiable piecewise linear function $h_\lambda(z)$ when $z\in[0,+\infty)$ by the following continuously differentiable function $g(z;\rho)$ below:

$$g_\lambda(z;\rho)=\begin{cases} \frac{1}{\gamma}|z| & \text{if } |z|\le\xi \\ s_1(|z|) & \text{if } \zeta\le|z|\le\eta \\ s_2(|z|) & \text{if } \eta\le|z|\le\kappa \\ 1-\frac{\rho}{2}(|z|-\gamma)^2 & \text{if } \kappa\le|z|\le\gamma \\ 1 & \text{otherwise} \end{cases} \tag{3.6}$$

where γ and κ are defined in (3.3). Subplot (b) in **Figure 2** graphically illustrates the function $g_\lambda(z;\rho_k)$.

By construction, $g_\lambda(z;\rho_k)$ is an even function. In addition, it is continuously differentiable on $(0,+\infty)$. We now establish the monotonicity property for $g_\lambda(z;\rho_k)$ when $z\in[0,+\infty)$.

Lemma 3.1. Let $S(z)$ be the quadratic spline in $[\xi,\kappa]$ with the breakpoints $0<\xi<\eta<\kappa$ and the boundary conditions (3.5) restated below

$$\begin{aligned} &S(\xi)=\frac{1}{\gamma}\xi,\ \ S'(\xi)=\frac{1}{\gamma}, \\ &S(\kappa)=1-\frac{\rho}{2}(\kappa-\gamma)^2,\ \ S'(\kappa)=-\rho(\kappa-\gamma). \end{aligned}$$

where γ and κare defined in (3.3). Assume $\rho>0$ and $\eta-\xi=\kappa-\eta$. Then $S(z)$ is strictly monotonically increasing on $[\xi,\kappa]$ and $g_\lambda(z;\rho_k)$ is strictly monotonically increasing on $[0,\gamma]$.

Proof. Assume that $s_1(z)=a_1+b_1(z-\xi)+c_1(z-\xi)^2$ and $s_2(z)=a_2+b_2(z-\xi)+c_2(z-\xi)^2$ are the quadratics. For the spline $S(z)$ on $[\xi,\kappa]$ with the boundary conditions (3.5), we have

$$s_1'(z)=b_1+2c_1(z-\xi),$$

and $s_2'(z)=b_2+2c_2(z-\xi)$. The boundary conditions (3.5) imply that

$$\begin{aligned} &a_1=s_1(\xi)=\frac{1}{\gamma}\xi, b_1=s_1'(\xi)=\frac{1}{\gamma}, \\ &a_2=s_2(\kappa)=1-\frac{\rho}{2}(\kappa-\gamma)^2, b_2=s_2'(\kappa)=\rho(\gamma-\kappa). \end{aligned} \tag{3.7}$$

The derivative and function continuity of the spline at $z=\eta$ yields

$$\begin{aligned} &c_1(\eta-\xi)-c_2(\eta-\kappa)=\frac{b_2-b_1}{2}, \\ &c_1(\eta-\xi)^2-c_2(\eta-\kappa)^2=a_2-a_1+b_2(\eta-\kappa)-b_1(\eta-\xi). \end{aligned}$$

This leads to

$$c_1 = \frac{a_2 - a_1 + b_2(\eta-\kappa) - b_1(\eta-\xi) + \frac{b_2-b_1}{2}(\kappa-\eta)}{(\eta-\xi)^2 + (\kappa-\eta)(\eta-\xi)}.$$

Since $\hat{g}_\lambda(z;\rho)$ is continuous at $z=\kappa$, $a_2 = s_2(\kappa) = \hat{g}_\lambda(\kappa;\rho) = \lambda\kappa^2$. Using (3.3),

$$a_2 = \lambda\kappa^2 = (\lambda\kappa)\kappa = \frac{1}{\gamma}\kappa.$$

Let $\Delta = \kappa-\eta = \eta-\xi$,. From $a_1 = \frac{1}{\gamma}\xi$, using (3.7), and $\kappa-\xi = 2\Delta$,

$$a_2 - a_1 = \frac{1}{\gamma}(\kappa-\xi) = b_1(\kappa-\xi) = 2b_1\Delta$$

Since $\hat{g}_\lambda(z;\rho)$ is continuously differentiable at $z=\kappa$, using (3.7), we have

$$b_2 = s_2'(\kappa) = 2\lambda\kappa = \frac{2}{\gamma} = 2b_1.$$

Hence

$$b_2(\eta-\kappa) - b_1(\eta-\xi) = -3b_1\Delta.$$

Using above, $a_2 - a_1 = 2b_1\Delta$, and $\kappa-\eta=\Delta$, we obtain

$$a_2 - a_1 + b_2(\eta-\kappa) - b_1(\eta-\xi) + \frac{b_2-b_1}{2}(\kappa-\eta) = -\frac{1}{2}b_1\Delta.$$

From the value of c_1 and the above, we have

$$2c_1(\eta-\xi) = -\frac{1}{2}b_1$$

Hence

$$s_1'(\eta) = b_1 + 2c_1(\eta-\xi) = \frac{1}{2}b_1 > 0.$$

Since $s_1'(z)$ and $s_2'(z)$ are linear in $[\xi,\eta]$ and $[\eta,\kappa]$ respectively, from $s_1'(\xi)>0, s_1'(\eta) = s_2'(\eta) > 0$ and $s_2'(\kappa)>0$, we conclude that $s_1'(z)$ and $s_2'(z)$ are strictly monotonically increasing in $[\xi,\eta]$ and $[\eta,\kappa]$. Since $\frac{1}{\gamma}z$ and $1-\frac{\rho}{2}(z-\gamma)^2$ are strictly monotonically increasing in $[0,\xi]$ and $[\kappa,\gamma]$ respectively, we conclude that $g(z;\rho)$ is monotonically increasing in $[0,\gamma]$. □

Figure 2 compares the approximation $\hat{g}_\lambda(z;\rho)$ with the approximation $g_\lambda(z;\rho)$ for a few values of ρ. Left and right subplots visually look very similar. However, the main difference can be seen near $z=0$.

Replacing $h_\lambda(\cdot)$ by $g_\lambda(z;\rho)$, we obtain:

$$\min_{x\in R^n}\left(f(x) + \mu\sum_{i=1}^n g_\lambda(x_i;\rho)\right) \tag{3.8}$$

Each approximation $\sum_{i=1}^n g_\lambda(x_i;\rho)$ can be considered as a regularization for the empirical function minimization.

This penalty function corresponds to the l_1 penalty around $x_i = 0$ for $1\le i\le n$; it behaves like the counting indicator function when $|x_i|$ is very large. The size of each region depends on the parameter ρ. When $\rho\to+\infty$, the l_1 penalty is used in $\left[0,\frac{1}{\sqrt{\lambda}}\right]$ and the counting indicator function is used in $\left[\frac{1}{\sqrt{\lambda}},+\infty\right)$. For any given $\rho>0$, the penalty function is continuously differentiable everywhere except at the origin. The function $g_\lambda(z;\rho)$ is piecewise quadratic with a concave quadratic piece for $z\in[\kappa,\gamma]$.

Based on (3.8), a gradual nonconvexification process can similarly be applied to track the global minimizer of (2.3), and (2.1) if λ is sufficiently large. Assume that the empirical function $f(x)$ is convex. For notational simplicity, let

$$P_k(x) = f(x) + \mu \sum_{i=1}^{n} g_\lambda(x_i; \rho_k) \tag{3.9}$$

Starting from the minimizer of $f(x)$, a sequence of approximations to the penalized empirical error minimization problem (3.8), $\{\mathcal{P}_k, k = 1, 2, \cdots\}$, is solved by approximating the indicator function. As ρ_k converges to zero, ξ, η and κ all converge to zero and γ converges to $+\infty$. Thus $g_\lambda(z;\rho)$ approaches the quadratic segment $1 - \frac{\rho}{2}(|z| - \gamma)^2$ with the (negative) curvature converging to zero. Hence the optimization problem (3.8) approaches the empirical error function minimization as ρ_k approaches zero. As ρ_k increases, the curvature of the quadratic function defining $g_\lambda(z;\rho_k)$ for $z \in [\kappa_k, \gamma_k]$ becomes more negative, introducing a graduated nonconvexity. The negative curvature interacts with the positive curvature of the empirical error function in an attempt to reach the optimal subset solutions via minimizing $P_k(x)$. In addition, as $\rho_k \to +\infty$, γ_k, $\kappa_k \to \frac{1}{\sqrt{\lambda}}$ and the functions $g_\lambda(z;\rho_k)$ approach $h_\lambda(z)$.

Figure 3 graphically illustrates how this graduated nonconvexification process tracks the global minimizer of the minimization problem (2.3) with an one-dimensional function $f(z) = \frac{1}{2}(z-2)^2$. In the top-left subplot of **Figure 3**, we see the original nonconvex function $f(z) + \mu h_\lambda(z)$, a convex approximation (corresponding to $\rho_k = 0.001$), and its minimizer. Increasing ρ to 1, we see the next approximation to the original function in the top-right subplot. With the minimizer of the first approximation function as a starting point, the minimizer of the new approximation, which is very close to the global minimizer, is computed. In the bottom two subplots

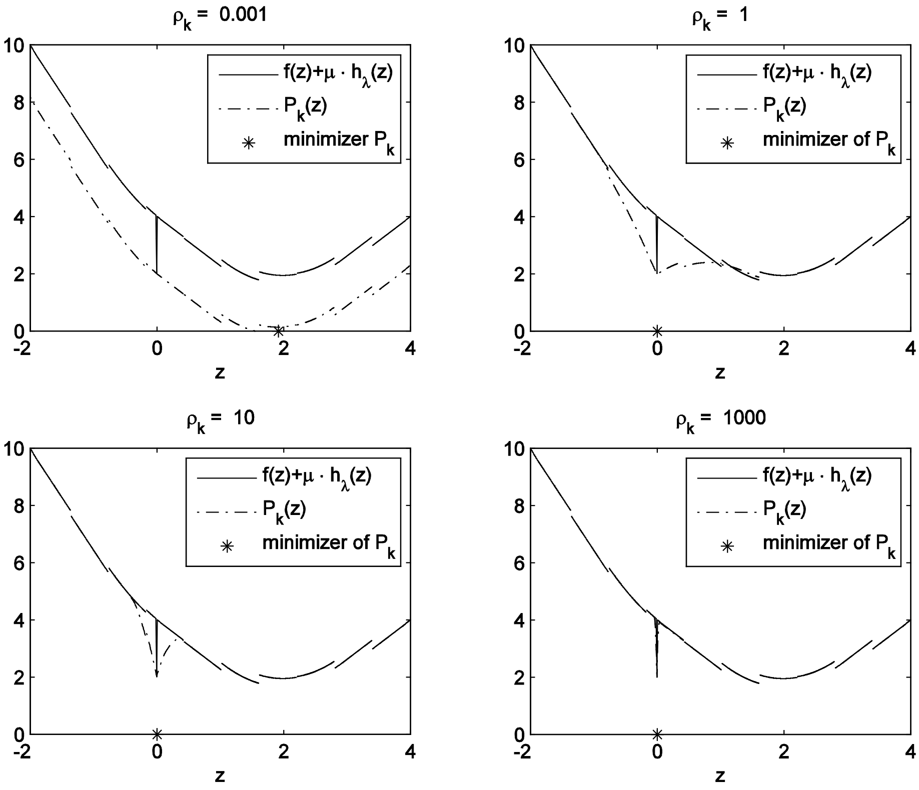

Figure 3. Tracking the global minimizer of $f(z) + \mu h_\lambda(z)$: Graduated nonconvex approximations.

($\rho_k = 10$ and 1000 respectively), we see how the approximating functions $P_k(z)$ approach the original function as ρ_k increases. From this illustration we see that the proposed process first considers large scale features of the function $f(z) + \mu h_\lambda(z)$ and gradually focuses in on features of a smaller scale. This graduated nonconvexification process can be terminated when the approximation $g_\lambda(z;\rho_k)$ to $h_\lambda(z)$ is sufficiently accurate at the computed solution.

Next Theorem 3.1 shows that when, for all i either $\left(x_i^*\right)_k \le \xi_k$ or $\left(x_i^*\right)_k \ge \gamma_k$, $\sum_{i=1}^n g_\lambda\left(\left(x_i^*\right)_k;\rho_k\right)$ accurately approximates $\sum_{i=1}^n h_\lambda\left(\left(x_i^*\right)_k\right)$ at the computed solution and remains so for any larger ρ.

Theorem 3.1. Assume that $\kappa(\rho)$ and $\gamma(\rho)$ are defined in (3.3), $0 < \xi < \eta < \kappa$, and $\xi(\rho)$ and $\eta(\rho)$ are monotonically increasing functions of ρ. Then the following holds:

- If the first order necessary condition for (3.8) is satisfied at x^* with $\bar{\rho} > 0$ and either $x^* < \xi$ or $x^* > \gamma$, then the first order optimality condition for (3.8) is satisfied at x^* for any $\rho \ge \bar{\rho}$.
- If the second order necessary condition for (3.8) is satisfied at x^* with $\bar{\rho} > 0$ and either $x^* < \xi$ or $x^* > \gamma$, then the second order optimality condition for (3.8) is satisfied at x^* for any $\rho \ge \bar{\rho}$.

Proof. Assume that $\bar{\rho} > 0$ and either $x_i^* < \xi$ or $x_i^* > \gamma$ for $1 \le i \le n$. By definition (3.3), $\gamma(\rho) < \gamma(\bar{\rho})$ and $\kappa(\rho) < \kappa(\bar{\rho})$ for $\rho \ge \bar{\rho}$.From the monotonicity assumption of ξ and η, for $\rho \ge \bar{\rho}$, $\xi(\rho) \ge \xi(\bar{\rho})$ and $\eta(\rho) \ge \eta(\bar{\rho})$. Therefore,

$$\sum_{x_i^* \ne 0}^{n} \nabla g_\lambda\left(x_i^*;\rho\right) = \sum_{x_i^* \ne 0}^{n} \nabla g_\lambda\left(x_i^*;\bar{\rho}\right), \text{ and } \sum_{x_i^* \ne 0}^{n} \nabla^2 g_\lambda\left(x_i^*;\rho\right) = \sum_{x_i^* \ne 0}^{n} \nabla^2 g_\lambda\left(x_i^*;\bar{\rho}\right).$$

Hence the first order and the second order necessary conditions for (3.7) hold at x^* with $\rho > 0$, when these conditions for (3.7) hold at x^* with $\bar{\rho} > 0$. □

Applying Theorem 3.1, the following can be used as a stopping condition for the graduated nonconvexification process:

$$\text{either}\left(x_i^*\right)_k < \xi_k \text{ or } \left(x_i^*\right)_k > \gamma_k, \quad \forall i \tag{3.10}$$

We also note that $g_\lambda(\cdot)$ only can be erroneous in approximating $h_\lambda(\cdot)$ in $[\xi,\eta]$. Thus when $\gamma - \xi$ is sufficiently small, we can also regard this approximation as sufficiently accurate. Thus, the computation can also terminate if the region of the inaccuracy $\gamma_k - \xi_k$ becomes sufficiently small.

The proposed graduated non-convex approximation process is summarized in **Figure 4**.

4. Computational Results

In this section we illustrate the performance of the proposed computational methods for mode estimation. We assume that the variables are constrained to be nonnegative; we note that the illustrated properties of the proposed sequence of l_1 approximations to l_0 regularization on sparsity will be similar when the model parameters are unconstrained.

To illustrate, we generate random sparse model selection problems based on least squares data fitting problems below:

GNC1 Algorithm. Let $\lambda > 0$ be a large constant and $\{\rho_k\}$ be a monotonically increasing sequence which converges to $+\infty$. Lettolbe a positive stopping tolerance.

1) Compute a minimizer for minimization problem (3.8) with the penalty parameter $\mu = 0$. Let $k = 1$.

2) Compute a solution to (3.8) with $\rho = \rho_k$ using the computed solution of (3.8) with $\rho = \rho_{k-1}$ as a starting point.

3) If either (3.10) or the inequality $\gamma_k - \xi_k < \text{tol}$ holds, terminate. Otherwise, $k \leftarrow k+1$ and go to the step 1).

Figure 4. A graduated nonconvexification method for the sparse model selection.

$$\min_{x\geq 0}\frac{1}{2}\|Ax-b\|_2^2+\mu\sum_{i=1}^{n}\Lambda\left(x_i\right) \tag{4.1}$$

Each random problem instance is generated by a random matrix $A=UDV$ where U is an m-by-m random orthogonal matrix, V is an n-by-n random orthogonal matrix, and D is a random diagonal matrix with a condition number equal to a specified constant. The m-by-1 vector b is set to equal to $Ax^*+\varepsilon$ where $x^*\geq 0$ is a random vector with K^* nonzero components (randomly selected); the nonzero components equal $100(1+\text{rand})$. Here rand is a random sample from the uniform distribution in $[0,1]$. The vector ε is a vector of m independent standard normal with a standard deviation equal to 10^{-2}, unless stated otherwise.

Using the graduated nonconvexification algorithm, we compute a solution to (4.1) by solving a sequence of approximations below:

$$\min_{x\geq 0}\frac{1}{2}\|Ax-b\|_2^2+\mu\sum_{i=1}^{n}g_\lambda\left(x_i;\rho_k\right) \tag{4.2}$$

where $\{\rho_k\}$ is a sequence of monotonically increasing positive numbers. Unless stated otherwise, we set $\frac{1}{\sqrt{\lambda}}=0.05$, $\rho_k=10\rho_{k-1}$, and $\rho_0=10^{-5}$. In addition, we let the spline breakpoints be as follows:

$$\xi=\left(1-0.5\min\left(\frac{1}{\rho},0.5\right)\right)\kappa,$$

$$\eta=\left(1-0.25\min\left(\frac{1}{\rho},0.5\right)\right)\kappa.$$

We refer to the computational algorithm in **Figure 4** with the specified parameter setting above as GNC1. We use the trust region algorithm for bound constrained problems proposed in [30] to solve (4.2). We illustrate various properties of the computed model $\hat{x}$. Specifically we report empirical error, sparsity, and errors compared to the true model x^* used to generate data. In measuring sparsity, we regard a component $\hat{x}_i$ as zero if $|\hat{x}_i|\leq 10^{-6}$. In addition, we assess the approximation error of using the continuous approximation $h(\cdot)$ or $\hat{h}(\cdot)$ to the indicator function at the computed solution $\hat{x}$. We also evaluate the computational cost of the graduated nonconvexification process and report the total number of optimization iterations required to obtain the estimate $\hat{x}$ and the average number of optimization iterations required for minimizing each $\mathcal{P}_k$.

Specifically, based on 100 random problem instances, we report the following attributes based on the average values from 100 random problem instances:

- Empirical error: $\|A\hat{x}-b\|_2$
- Average relative distance to the true model: $\frac{\|x-x^*\|_2}{\|x^*\|_2}$
- Sparsity: the number of zeroes in the computed solution $\hat{x}$, *i.e.*, $\sum_{i=1}^{n}\Lambda\left(|\hat{x}_i|\leq 10^{-6}\right)$.
- Accuracy in approximating the counting indicator function $\Lambda(z)$ by $h_\lambda(z)$ or $\hat{h}_\lambda(z)$: recall that the approximation is accurate if $|z|\leq\frac{1}{\sqrt{\lambda}}$ implies that $z=0$. Thus we measure this accuracy by $\frac{\sum_{i=1}^{n}\Lambda\left(|\hat{x}_i|\leq 10^{-6}\right)}{\sum_{i=1}^{n}\Lambda\left(|\hat{x}_i|\leq\frac{1}{\sqrt{\lambda}}\right)}$; value 1 means 100% accuracy.
- Average number of optimization iterations over all $\rho_k:\text{itn}_k$.
- Total number of optimization iterations for the entire gradual non-convexification process: $\sum_k\text{itn}_k$.

Table 1 reports the performance assessment for GNC1based on the piecewise linear approximation $h_\lambda(z)$ to the counting function $\Lambda(z)$ in the specified parameter setting. The number of observations*m*equals 200 and the dimension*n* of x^* equals 100. Two subpanels correspond to $\sum_{i=1}^{n}\Lambda\left(x_i^*\right)$ the number of nonzeros in x^*, equals to 50 and 90 respectively. The column under $\mu=0$ provides the properties of the least squares solution for the given data; we see that the average least squares error $\|Ax-b\|_2$ is approximately 0.1. As the penalty parameter μ increases, the average data fitting error increases but sparsity in the estimate $\hat{x}$ also increases from 0 to about 70.

In addition observe from **Table 1** that there is little difference in the fitting error when $\mu \in [0, 10^{-4}]$ but the sparsity in the solution estimate $\hat{x}$ is drastically different, from no zero components to more than 35 when the true solution x^* has about 50 zero components, *i.e.*, $K^* = 50$. From **Table 1**, we also observe that the relative average distance to the true parameter x^* is smaller than that of the least squares solution, when the penalty parameter is $\mu \in (0, 10^{-4}]$ for which estimate $\hat{x}$ is fairly sparse. This indicates that sparsity regularization is indeed important in accurate model estimation. We also note that the error in using $h(\cdot)$ to approximate the counting indicator function at the computed estimate $\hat{x}$ is very small with a minimum correctness of 80% for penalty parameter $\mu = 10^{-6}$; complete accuracy of 100% is achieved when $\mu \ge 10^{-3}$.

One potential concern of the proposed graduated non-convex method is the computational cost since a potentially large sequence (corresponding to the parameter sequence $\{\rho_k\}$ of optimization problems need to be solved. However, since for each minimization of $\mathcal{P}_k$, the solution from $k-1$ is used as a starting point, the total number of optimization iterations required is quite reasonable.

For example, when $K^* = 50$, on average a total of 77.31 iterations are required to compute for the entire GNC1 process. Since the average number of iterations for minimizing P_k is 4.83, this implies that the GNC1 process here terminates after about 16 steps on average. It can be observed that the total number of iterations required increases with the penalty parameter μ.

We also investigate performance for the underdetermined data fitting problems; the number of observations is less than the number of unknown variables (specifically $m = 100$ and $n = 200$). **Table 2** illustrates that the relative distance to the true model x^* is significantly larger in the underdetermined case than in the over determined case; the corresponding entries in **Table 1** are included in **Table 2** for comparison. This is reasonable since there is less information available to infer the true model.

Table 1. GNC1 performance statistics from 100 random problems with the number of observations *m* greater than the dimension *n*: $\rho_{k+1} = 10\rho_k$ with ρ_0 = 1e-005 and $\frac{1}{\sqrt{\lambda}} = 0.05$.

	$\mu = 0$	1e-006	0.0001	0.01	1
Number of nonzeros in the generator $K^* = 50$ $m = 200$ n = 100					
$\|A\hat{x}-b\|_2$	0.098	0.108	0.111	0.257	2.380
$\frac{\|\hat{x}-x^*\|_2}{\|x^*\|_2}$	0.199	0.172	0.188	0.469	0.686
$\sum_{i=1}^{n}\left(\Lambda\left(\lvert\hat{x}_i\rvert\le 10^{-6}\right)\right)$	0.000	20.610	35.020	53.480	68.700
$\frac{\sum_i \lvert\Lambda\left(\hat{x}_i\rvert\le 10^{-6}\right)}{\sum_i \lvert\Lambda\left(\hat{x}_i\rvert\le 1/\sqrt{\lambda}\right)}$	0.000	0.812	0.942	1.000	1.000
itn_k	0.000	2.123	4.832	7.947	7.281
$\sum_k itn_k$	0.000	33.970	77.320	127.150	116.490
Number of nonzeros in the generator $K^* = 90$ $m = 200$ $n = 100$					
$\|A\hat{x}-b\|_2$	0.099	0.102	0.103	0.309	2.831
$\frac{\|\hat{x}-x^*\|_2}{\|x^*\|_2}$	0.160	0.150	0.177	0.485	0.694
$\sum_{i=1}^{n}\Lambda\left(\lvert\hat{x}_i\rvert\le 10^{-6}\right)$	0.000	5.260	9.820	31.140	53.660
$\frac{\sum_i \Lambda\left(\lvert\hat{x}_i\rvert\le 10^{-6}\right)}{\sum_i \Lambda\left(\lvert\hat{x}_i\rvert\le 1/\sqrt{\lambda}\right)}$	0.000	0.805	0.948	1.000	1.000
itn_k	0.000	2.023	3.094	7.353	8.368
$\sum_k itn_k$	0.000	32.370	49.510	117.650	133.890

Table 2. Distance to the true model (GNC1): overdetermined vs underdetermined, $\rho_{k+1} = 10\rho_k$ with ρ_0 = 1e-005 and $\frac{1}{\sqrt{\lambda}}$ = 0.05 .

Distance to the True Model: $\frac{\|\hat{x} - x^*\|_2}{\|x^*\|_2}$					
	$\mu = 0$	1e-006	0.0001	0.01	1
(m,n)		Number of nonzeros in the generator K^* = 50			
(200,100)	0.199	0.172	0.188	0.469	0.686
(100 200)	0.720	0.714	0.715	0.781	0.857
(m,n)		Number of nonzeros in the generator K^* = 90			
(200,100)	0.160	0.150	0.177	0.485	0.694
(100 200)	0.722	0.718	0.721	0.784	0.861

5. Concluding Remarks

In high dimensional data fitting problems, the objective is to minimize an empirical loss function based on data while achieving sparsity in the model parameters at the same time. Achieving sparsity can be crucial in obtaining robust out-of-sample performance and attaining meaningful understanding of the causal relationship in data. In addition, sparsity may be an explicit goal in practical applications such as tracking market indices with a small number of assets.

The combination of minimizing empirical loss and maximizing sparsity naturally leads to a minimization problem regularized by a penalty, which is proportional to the number of variables describing the model to be estimated, this problem with many local minimizers.

The empirical loss function is often convex. Assuming this property, we propose a graduated non-convex algorithm to minimize a convex empirical loss function regularized with a penalty proportional to the number of nonzero variables.

The proposed algorithm is based on approximating the counting indicator function by a continuous piecewise linear function. We show mathematically that the continuous piecewise linear approximation to the counting indicator function has more attractive properties than the continuous piecewise quadratic approximation. Specifically, there exists a sufficiently large parameter λ so that the regularized optimization problem using this approximation is equivalent to the original optimization problem regularized by the number of nonzero in the model parameters. This property does not hold for the continuous piecewise quadratic approximation.

A graduated nonconvexification process is proposed by introducing a family of approximations to the continuous piecewise linear function. This family of approximations is indexed by a nonnegative parameter ρ . In addition, these approximations can be regarded as penalty functions themselves with the following properties. Firstly there is a region around the origin in which the penalty function is the l_1 penalty. In another region, the penalty function equals counting indicator functions. As ρ increases, both regions increases and the region with l_1 penalty converges to $\left[0, \frac{1}{\sqrt{\lambda}}\right]$ and the region with the counting indicator function as penalty converges to $\left[\frac{1}{\sqrt{\lambda}}, +\infty\right)$. In addition, for any $\rho > 0$, the penalty function is continuously differentiable everywhere except at the origin. Each penalty function is an even function with a monotonicity property. When the parameter ρ is small, a small negative curvature is added to the convex empirical loss function to form a regularized objective function. As ρ increases, more negative curvature is added but the region in which the penalty is nonconvex shrinks.

We investigate performance of the proposed graduated nonconvexification algorithm based on randomly generated least squares problems with different sparsity levels in the true solution x^* We observe that the data fitting error increases as the penalty parameter μ increases. Simultaneously sparsity in the computed solution increases as the penalty increases. In addition, sparse solutions (with sparsity close to the true solutions) can be computed without much compromise in the magnitude of the data fitting error. Indeed, the solution with the smallest relative distance to the true solution is obtained by sparse solutions with the sparsity close to that of the true solutions.

Our results also indicate that the computational costs required by the GNC process is relatively moderate, since the computed solution in the *k*th step in the graduate nonconvexification process is often a good starting point for the (*k*+1)th step.

References

[1] Barron, A., Birge, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields*, **113**, 301-302. http://dx.doi.org/10.1007/s004400050210

[2] Natarajan, B.K. (1995) Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, **24**, 227-234. http://dx.doi.org/10.1137/S0097539792240406

[3] Davis, G., Mallat, S. and Avellaneda, M. (1997) Greedy Adaptive Approximation. *Constructive Approximation*, **13**, 57-98. http://dx.doi.org/10.1007/BF02678430

[4] Tropp, J.A. (2006) Just Relax: Convex Programming Methods for Identifying Sparse Signals in Noise. *IEEE Transactions on Information Theory*, **52**, 1030-1051.

[5] Tropp, J.A. (2004) Just Relax: Convex Programming Methods for Subset Selection and Sparse Approximation. Tech. Rep. ICES Report 0404, The University of Texas at Austin.

[6] Taylor, H.L., Banks, S.C. and McCoy, J.F. (1979) Deconvolution with the l1 Norm. *Geophysics*, **44**, 39-52. http://dx.doi.org/10.1190/1.1440921

[7] Levy, S. and Fullagar, P.K. (1981) Reconstruction of a Sparse Spike Train from a Portion of Its Spectrum and Application to High-Resolution Deconvolution. *Geophysics*, **46**, 1235-1243. http://dx.doi.org/10.1190/1.1441261

[8] Wu, J.K. (1994) Two Problems of Computer Mechanics Program System. In: *Proceedings of Finite Element Analysis and CAD*, Peking University Press, Beijing, 9-15.

[9] Santosa, F. and Symes, W.W. (1986) Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*, **7**, 1307-1330. http://dx.doi.org/10.1137/0907087

[10] Li, Y. and Santosa, F. (1996) A Computational Algorithm for Minimizing Total Variation in Image Restoration. *IEEE Transactions on Image Processing*, **5**, 987-995. http://dx.doi.org/10.1109/83.503914

[11] Coleman, T.F., Li, Y. and Mariano, A. (2001) Segmentation of Pulmonary Nodule Image Using 1-Norm Minimization. *Computational Optimization and Application*, **19**, 243-272.

[12] Donoho, D.L., Elad, M. and Temlyakov, V.N. (2006) Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. *IEEE Transactions on Information Theory*, **52**, 6-17. http://dx.doi.org/10.1109/TIT.2005.860430

[13] Candès, E.J. (2006) Compressive Sampling. In: *Proceedings of the International Congress of Mathematics*, Madrid, Spain.

[14] Candés, E.J., Romberg, J. and Tao, T. (2006) Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, **52**, 489-509. http://dx.doi.org/10.1109/TIT.2005.862083

[15] Kim, S., Koh, K., Boyd, S. and Gorinvesky, D. (2007) An Interior Point Method for Large-Scale l_1 Regularized Least Squares. *IEEE Journal on Selected Topics in Signal Processing*, **1**, 606-617. http://dx.doi.org/10.1109/JSTSP.2007.910971

[16] Wright, S.J., Nowak, R.D. and Figueiredo, M.A.T. (2009) Sparse Reconstruction by Separable Approximation. *IEEE Transactions on Signal Processing*, **57**, 2470-2493. http://dx.doi.org/10.1109/TSP.2009.2016892

[17] Daubechies, I., Devore, R., Fornasier, M. and Güntürk, C.S. (2009) Iteratively Re-Weighted Least Squares Minimization for Sparse Recovery. Communications on Pure and Applied Mathematics LXIII (2010) 0001-0038, 2470-2493.

[18] 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. http://dx.doi.org/10.1109/TSP.2009.2016892

[19] Fan, J. and Li, R. (2006) Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. In: Sanz-Sole, M., Soria, J., Carona, J.L. and Verdera, J., Eds., *Proceedings of the International Congress of Mathematics*, 595-622.

[20] Lv, J. and Fan, Y. (2009) A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares. *The Annals of Statistics*, **37**, 3498-3528. http://dx.doi.org/10.1214/09-AOS683

[21] Fan, J., Lv, J. and Qi, L. (2011) Sparse High Dimensional Models in Economics. *Annual Review of Economics*, **3**, 291-317.

[22] Fan, J. (1997) Comments on “Wavelets in Statistics: A Review”. *Journal of the Italian Statistical Society*, **6**, 131-138. http://dx.doi.org/10.1007/BF03178906

[23] Hunter, D.R. and Li, R. (2005) Variable Selection Using mm Algorithms. *The Annals of Statistics*, **33**, 1617-1642. http://dx.doi.org/10.1214/009053605000000200

[24] Zhou, H. and Li, R. (2008) One-Step Sparse Estimates in Nonconvex Penalized Likelihood Models (with Discussion). *The Annals of Statistics*, **36**, 1509-1533. http://dx.doi.org/10.1214/009053607000000802

[25] Cand´es, E.J., Wakin, M.B. and Boyd, S.P. (2008) Enhancing Sparsity by Reweighted l1 Minimization. *J. FourierAna Appl*, **14**, 877-905. http://dx.doi.org/10.1007/s00041-008-9045-x

[26] Chartrand, R. and Yin, W. (2008) Iteratively Reweighted Algorithms for Compressive Sensing. In: *IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 3869-3872.

[27] Zhang, T. (2010) Analysis of Multi-Stage Convex Relaxation for Sparse Regularization. *Journal of Machine Learning Research*, **10**, 1081-1107.

[28] Blake, A. and Zisserman, A. (1987) Visual Reconstruction. Cambridge.

[29] Coleman, T.F., Henninger, J. and Li, Y. (2006) Minimizing Tracking Error While Restricting the Number of Assets. *Journal of Risk*, **8**, 35-56.

[30] Coleman, T.F. and Li, Y. (1996) An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM Journal on Optimization*, **6**, 418-445. http://dx.doi.org/10.1137/0806023