



HAL
open science

The Iso-lambda Descent Algorithm for the LASSO

Manuel Loth, Philippe Preux

► **To cite this version:**

Manuel Loth, Philippe Preux. The Iso-lambda Descent Algorithm for the LASSO. 17th International Conference on Neural Information Processing, Nov 2010, Sidney, Australia. inria-00508257v1

HAL Id: inria-00508257

<https://inria.hal.science/inria-00508257v1>

Submitted on 2 Aug 2010 (v1), last revised 6 Sep 2010 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Iso-lambda Descent Algorithm for the LASSO

Manuel Loth and Philippe Preux

INRIA Lille - Nord Europe,
Université de Lille

Abstract. Following the introduction by Tibshirani of the LASSO technique for feature selection in regression, two algorithms were proposed by Osborne et al. for solving the associated problem. One is an homotopy method that gained popularity as the LASSO modification of the LARS algorithm. The other is a finite-step descent method that follows a path on the constraint polytope, and seems to have been largely ignored. One of the reason may be that it solves the constrained formulation of the LASSO, as opposed to the more practical regularized formulation. We give here an adaptation of this algorithm that solves the regularized problem, has a simpler formulation, and outperforms state-of-the-art algorithms in terms of speed.

Key words: Lasso, algorithm, descent, regularization

1 Introduction

The Least Absolute Selection and Shrinkage Operator (LASSO) was proposed in [1] as an efficient and feasible way to produce sparse linear models in regression. Let us begin by recalling its definition and fixing notations.

Let \mathbf{v}^\top denotes the transpose of a vector, $\mathbf{u}^\top \mathbf{v}$ denotes the scalar product, $\|\mathbf{v}\|_2^2 = \sum_{i=1}^{\dim(\mathbf{v})} v_i^2 = \mathbf{v}^\top \mathbf{v}$ the squared L_2 norm, and $\|\mathbf{v}\|_1 = \sum_{i=1}^{\dim(\mathbf{v})} |v_i| = \text{sign}(\mathbf{v})^\top \mathbf{v}$ the L_1 norm, and $\text{sign}(\mathbf{v}) = (\text{sign}(v_1), \dots)^\top$.

Definition 1. *Given*

- n samples $(x_i, y_i)_{1 \leq i \leq n}$ from correlated variables $X \in \mathcal{X}$ and $Y \in \mathbb{R}$,
- a set of features $\mathcal{D} \subset \mathbb{R}^{\mathcal{X}}$,
- a constraint $t \in \mathbb{R}_+$,

the LASSO associates to the samples a model $Y = f(X)$ defined as a linear combination of features from \mathcal{D} that minimizes the squared residual subject to an L_1 -norm constraint on the linear coefficient vector:

$$\min_{\sigma \in 2^{\mathcal{D}}, \beta \in \mathbb{R}^{|\sigma|}} \|\mathbf{y} - \mathbf{X}_\sigma^\top \beta\|_2^2 \quad \text{s.t. } \|\beta\|_1 \leq t, \quad (1)$$

where σ , referred to as the active set, is a finite, ordered subset of features: $\sigma = \{\phi_1, \dots, \phi_k\} \subset \mathcal{D}$, and

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X}_\sigma = \begin{bmatrix} -\phi_1(\mathbf{x}) - \\ \vdots \\ -\phi_k(\mathbf{x}) - \end{bmatrix} = \begin{bmatrix} \phi_1(x_1) \dots \phi_1(x_n) \\ \vdots \\ \phi_k(x_1) \dots \phi_k(x_n) \end{bmatrix}.$$

It should first be noted that the constraint is active only if it disqualifies all ordinary least-square (OLS) solutions that strictly minimize the squared residual. Let t_{OLS} be the supremum of active constraints, and let us assume in the following that $t < t_{\text{OLS}}$. In this case, the inequality constraint can be turned into an equality, by simple convexity arguments.

Along with defining the operator, two ways to solve the associated problem were proposed in [1]. The first one is an “exterior-point” active-set method that starts from the OLS solution and follows a path defined by changes of active set, on which the L_1 norm of the coefficient decreases up to reaching the constraint polytope. The second one consists in casting the problem into a quadratic program (QP), and applying standard QP methods. This can be done by adding to \mathcal{D} the negative counterpart of all features, thus allowing an additional nonnegativity constraint to all coefficients, which makes the L_1 norm constraint strictly linear.

In [2], a very different approach was introduced (“shooting” algorithm), that proceeds by means of a cyclical coordinate descent (CCD). Despite its simplicity and efficiency, it lacked popularity, partly because of the unnecessary use of the OLS solution as a starting point. However, it gained credit by its presentation in [3] in a more general setting.

In [4] were proposed two algorithms more related to QP, but fitted to the specificities of the problem. One is an interior-point active-set homotopy method that starts from an empty set and follows a path defined by changes of the active set on which the coefficient’s L_1 norm increases up to reaching the constraint. Its major interest, beside its efficiency, is that the followed path corresponds to all successive LASSO solutions for a constraint going from 0 to t (regularization path). It is best known as the LASSO modification of the LARS algorithm presented in [5], in a more general framework, with more details and clarity, and in a *regularized* formulation. The second algorithm, that we reformulate and transpose in this paper, is a *surface* active-set descent method: all points in the path have an L_1 norm *equal* to the constraint t , and the squared residual decreases up to reaching an optimal solution.

In section 2, we give a quick yet rather detailed exposition of this method, from which we derive in section 3 an algorithm that solves the regularized formulation of the LASSO. Section 4 presents experimental results in which the new algorithm outperforms the homotopy and CCD in terms of running time. We conclude in section 5 by mentioning additional advantages of this algorithm.

2 Iso- $\|\beta\|_1$ Descent

The unnamed algorithm introduced in [4], that we may refer to as the iso- $\|\beta\|_1$ descent method, is based on the following two facts:

- if the active set as well as the sign of the coefficients (*signed active set*) are known or assumed, the computation reduces to finding the minimizer of the squared residual on the corresponding hyperplane,
- the optimality of the signed active set can then easily be tested.

Let us explicit these two steps, and the algorithm that naturally arises from them.

2.1 Minimizer on a signed active set

If the constraint is tightened by imposing a given signed active set $(\sigma, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \text{sign}(\boldsymbol{\beta})$, the problem reduces to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{|\sigma|}} \|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \begin{cases} \boldsymbol{\theta}^\top \boldsymbol{\beta} & = t \\ \text{sign}(\boldsymbol{\beta}) & = \boldsymbol{\theta} \end{cases} . \quad (2)$$

Minimizing a convex function subject to a linear constraint is a simple QP: the minimizer is the only point on the constraint plane where the gradient of the minimized function is normal to the plane, i.e. such that any vector normal to the plane, say $\boldsymbol{\theta}$, is equal to the gradient (or its negative half) up to a (Lagrange) multiplier λ :

$$1 \implies -\frac{1}{2} \nabla_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta}\|_2^2 = \lambda \boldsymbol{\theta} \quad (3)$$

$$\implies \mathbf{X}^\top (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta}) = \lambda \boldsymbol{\theta} \quad (4)$$

$$\implies \mathbf{X} \mathbf{X}^\top \boldsymbol{\beta} = \mathbf{X} \mathbf{y} - \lambda \boldsymbol{\theta} \quad (5)$$

$$\implies \boldsymbol{\beta} = \underbrace{(\mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{X} \mathbf{y}}_{\boldsymbol{\beta}^*} - \lambda \underbrace{(\mathbf{X} \mathbf{X}^\top)^{-1} \boldsymbol{\theta}}_{\Delta \boldsymbol{\beta}} \quad (6)$$

$$1 \iff \begin{cases} \boldsymbol{\beta} = \boldsymbol{\beta}^* - \lambda \Delta \boldsymbol{\beta} \\ \boldsymbol{\theta}^\top \boldsymbol{\beta} = t \end{cases} \quad (7)$$

$$\iff \begin{cases} \boldsymbol{\beta} = \boldsymbol{\beta}^* - \lambda \Delta \boldsymbol{\beta} \\ \boldsymbol{\theta}^\top (\boldsymbol{\beta}^* - \lambda \Delta \boldsymbol{\beta}) = t \end{cases} \quad (8)$$

$$\iff \begin{cases} \boldsymbol{\beta} = \boldsymbol{\beta}^* - \lambda \Delta \boldsymbol{\beta} \\ \lambda = \frac{\boldsymbol{\theta}^\top \boldsymbol{\beta}^* - t}{\boldsymbol{\theta}^\top \Delta \boldsymbol{\beta}} \end{cases} \quad (9)$$

Note that $\boldsymbol{\beta}^*$ is the partial least-square solution (PLS, ordinary least-square on selected feature set), $\Delta \boldsymbol{\beta}$ can be seen as a regularization direction, and λ as a regularization parameter.

It then remains for this solution to satisfy the sign constraint ($\text{sign}(\boldsymbol{\beta}) = \boldsymbol{\theta}$) that can also be seen as an initial assumption of which the validity must be checked. If it is not satisfied, this implies that $(\sigma, \boldsymbol{\theta})$ is not optimal: the squared residual being convex in $\boldsymbol{\beta}$, given any point $\boldsymbol{\beta}_0$ in the simplex (the subspace of the plane $\boldsymbol{\theta}^\top \boldsymbol{\beta}$ defined by the sign constraint), it monotonically decreases on a

line from β_0 to β , which will intercept the frontier of the simplex at which a coefficient (the same for any starting point β_0) is zeroed, thus the corresponding reduced active set contains a better minimizer. The computation can then be started again on the reduced set. An example is given in Fig. 2.1 that illustrates the sign disagreement and the different variables involved.

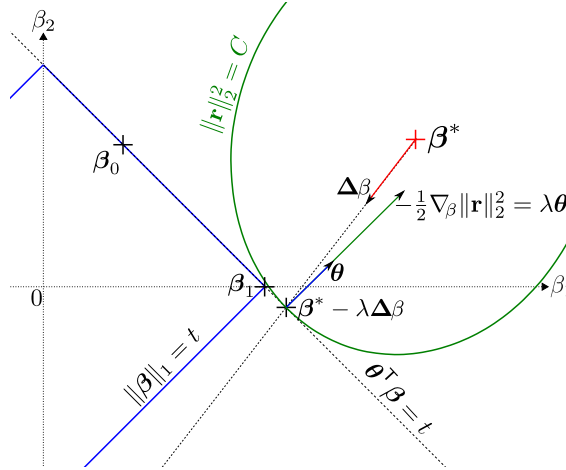


Fig. 1. From a signed active set $(\{\phi_1, \phi_2\}, (+, +))$, the regularized least-square is disagreeing with the sign hypothesis, indicating a better solution β_1 on reduced active set $(\{\phi_1\}, (+))$. \mathbf{r} is the residual $\mathbf{y} - \mathbf{X}^T \beta$

2.2 Karush-Kuhn-Tucker conditions

If there is no sign disagreement, or when there is none after a sequence of optimization/shrinkage, the optimality of the active set can be tested via the necessary and sufficient Karush-Kuhn-Tucker conditions that reduce to the following ([6]):

$$\begin{cases} \forall \phi \in \sigma, & \phi(\mathbf{x})^T (\mathbf{y} - \mathbf{X}_\sigma^T \beta) = \text{sign}(\beta_\phi) \lambda \\ \forall \phi \in \mathcal{D} \setminus \sigma, & |\phi(\mathbf{x})^T (\mathbf{y} - \mathbf{X}_\sigma^T \beta)| \leq \lambda \end{cases} \quad (10)$$

The first condition was handled by the previous computation, and the optimality test lies in the second one: no inactive feature should have a correlation to the residual higher than that of active features. This can be intuitively understood: the correlation being the derivative of the squared residual w.r.t. the coefficient of a feature, if an inactive feature is *over-correlated*, it is possible to reduce the coefficient budget allocated to the active features by a sufficiently small amount and re-allocate it to the latter with greater benefice. A formal proof can be written following this sketch, that also shows that the better minimizer of the augmented active set involves a decreased value of λ . Thus, given a signed active

set and a sign-compliant solution to the associated (local) QP, the set augmented with an over-correlated feature contains a better solution (lower residual for equal L_1 -norm). The sign imposed to the coefficient of the new active feature is that of its correlation, following 3.

2.3 Algorithm

This leads to the Algorithm 1, that shrinks the active set whenever the local QP solution is not sign-compliant, otherwise expands it with the most over-correlated feature if any, otherwise has converged to a LASSO solution. It is not mandatory to include the *most* over-correlated feature, but intuitively and empirically, this lessens the expected/average number of steps.

The algorithm has the descent property, since the residual decreases at each change of the active set, as shown previously, while the coefficient's L_1 norm is constant. It converges in at most $2^{|\mathcal{D}|}$ steps which is finite if \mathcal{D} is. The number of steps is however consistently observed to be $O(\min(n, |\mathcal{D}|))$ in experiments.

Algorithm 1 iso- $\|\beta\|_1$ Descent

Input: $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{y} \in \mathbb{R}^n$, $\mathcal{D} \subset \mathbb{R}^{\mathcal{X}}$, $t \in \mathbb{R}^+$, ($\sigma \subset \mathcal{D}$, $\beta \in \mathbb{R}^{|\sigma|}$) s.t. $\|\beta\|_1 = t$

Output: $(\sigma, \beta) \in \arg \min_{\sigma \subset \mathcal{D}, \beta \in \mathbb{R}^{|\sigma|}} \|\mathbf{y} - \mathbf{X}_\sigma^\top \beta\|_2^2$ s.t. $\|\beta\|_1 = t$

define $(x, i) = \min, \arg \min_t f(t)$: $x = \min_t f(t)$; $i \in \arg \min_t f(t)$

$\theta \leftarrow \text{sign}(\beta)$

loop

$$\beta^* = (\mathbf{X}_\sigma \mathbf{X}_\sigma^\top)^{-1} \mathbf{X}_\sigma \mathbf{y}$$

$$\Delta \beta = (\mathbf{X}_\sigma \mathbf{X}_\sigma^\top)^{-1} \theta$$

$$\lambda = \frac{\theta \cdot \beta^* - t}{\theta \cdot \Delta \beta}$$

$$\beta' = \beta^* - \lambda \Delta \beta$$

$$(\gamma, i) = \min, \arg \min_{i \in \{1, \dots, |\sigma|\}, \text{sign}(\beta'_i) \neq \text{sign}(\theta_i)} \frac{\beta_i}{\beta'_i - \beta_i}$$

$$\beta \leftarrow \beta + \min(\gamma, 1)(\beta' - \beta)$$

if $\gamma \leq 1$ **then**

$\sigma \leftarrow \sigma \setminus \{\phi_i\}$; update θ and β accordingly

else

$$(\phi, c) = \arg \max_{\phi \in \mathcal{D}, c = \phi(\mathbf{x})(\mathbf{y} - \mathbf{X}_\sigma^\top \beta)} |c|$$

if $|c| > \lambda$ **then**

$$\sigma \leftarrow \sigma \cup \{\phi\}; \theta \leftarrow (\theta, \text{sign}(c))^\top; \beta \leftarrow (\beta, 0)^\top$$

else

return (σ, β)

3 Iso- λ Descent

The Lagrange multiplier that appears in the resolution of the constrained least-square leads to an alternative *regularized* formulation of the LASSO: for any

constraint t ,

$$\min_{\substack{\sigma \in 2^{\mathcal{D}}, \\ \boldsymbol{\beta} \in \mathbb{R}^{|\sigma|}}} \|\mathbf{y} - \mathbf{X}_{\sigma}^{\top} \boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t \iff \min_{\substack{\sigma \in 2^{\mathcal{D}}, \\ \boldsymbol{\beta} \in \mathbb{R}^{|\sigma|}}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{\sigma}^{\top} \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

where λ is the Lagrange multiplier associated with the minimizer(s) of the constrained formulation. It can be shown that parameters t and λ are strictly decreasing functions of each other in this equivalence relation.

For none of the formulations can the parameter easily be tuned beforehand, but λ presents the advantage that its bounds are known in advance: the (resp. greatest and lowest) parameters such that $\mathbf{0}$ is the unique solution are $t_0 = 0$ and $\lambda_0 = \max_{\phi \in \mathcal{D}} |\phi(\mathbf{x})^{\top} \mathbf{y}|$, but t_{OLS} cannot be computed independently of the OLS solution itself, whereas $\lambda_{\text{OLS}} = 0$. Moreover, λ is more informative about the difference of residual between the regularized and ordinary least-square solutions.

In [7], the authors of the iso- $\|\boldsymbol{\beta}\|_1$ descent method mention the possibility of solving the regularized formulation by multiple runs of their algorithm inside a grid or Newton-Raphson search to find the corresponding value of t .

A much simpler possibility appears perhaps more clearly in our exposition of the algorithm, that consists in using a fix value of λ – the given regularization parameter – rather than computing at each step the value maintaining a constant L_1 -norm. This results in an even simpler algorithm described in Algorithm 2.

It remains to prove that the descent property is preserved by this modification. This can be done by noting that the L_1 norm of the coefficient is increasing at each step, as a consequence of the fact that the Lagrange multiplier is decreasing in the iso- $\|\boldsymbol{\beta}\|_1$ descent. Hence the coefficient change when augmenting the active set can be decomposed in two steps, the first one preserving the L_1 norm as previously, and the second one along the regularization direction toward the PLS. Both follow descent directions. When a sign disagreement occurs, the first step is ended at the first coefficient zeroing and the second is made on the reduced active set.

4 Experiments

We reproduced the “speed trials” experiments described in [3] with the following three algorithms: the cyclical coordinate descent algorithm described in that publication (CCD), the homotopy (LARS) method, and the iso- λ descent. The detailed settings and source code can be found at <http://chercheurs.lille.inria.fr/~loth/LASSO-speed>

The contrast between the results of the homotopy method in our experiments and those in [3] is explained by the fact that we gave it a simplified formulation and a proper implementation in C, closely resembling those of the iso- λ descent. This similarity, together with a lower complexity and comparable number of steps for going from a value of λ to another, indicate a consistently lower run time for the latter, which is confirmed by the experiments. The coordinate descent method shows slightly better results only for $|\mathcal{D}| < n$ and uncorrelated features.

Algorithm 2 iso- λ Descent

Input: $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{y} \in \mathbb{R}^n$, $\mathcal{D} \subset \mathbb{R}^{\mathcal{X}}$, $\lambda \in \mathbb{R}^+$, starting point $(\sigma \subset \mathcal{D}, \boldsymbol{\beta} \in \mathbb{R}^{|\sigma|})$
Output: $(\sigma, \boldsymbol{\beta}) \in \arg \min_{\sigma \subset \mathcal{D}, \boldsymbol{\beta} \in \mathbb{R}^{|\sigma|}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{\sigma}^{\top} \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$
define $(x, i) = \min, \arg \min_t f(t) : x = \min_t f(t); i \in \arg \min_t f(t)$
 $\boldsymbol{\theta} \leftarrow \text{sign}(\boldsymbol{\beta})$
loop
 $\boldsymbol{\beta}' = (\mathbf{X}_{\sigma} \mathbf{X}_{\sigma}^{\top})^{-1} (\mathbf{X}_{\sigma} \mathbf{y} - \lambda \boldsymbol{\theta})$
 $(\gamma, i) = \min, \arg \min_{i \in \{1, \dots, |\sigma|\}, \text{sign}(\beta'_i) \neq \text{sign}(\theta_i)} \frac{\beta'_i}{\beta'_i - \beta_i}$
 $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \min(\gamma, 1) (\boldsymbol{\beta}' - \boldsymbol{\beta})$
if $\gamma \leq 1$ **then**
 $\sigma \leftarrow \sigma \setminus \{\phi_i\}$; update $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ accordingly
else
 $(\phi, c) = \arg \max_{\phi \in \mathcal{D}, c = \phi(\mathbf{x}) (\mathbf{y} - \mathbf{X}_{\sigma}^{\top} \boldsymbol{\beta})} |c|$
if $|c| > \lambda$ **then**
 $\sigma \leftarrow \sigma \cup \{\phi\}$; $\boldsymbol{\theta} \leftarrow (\boldsymbol{\theta}, \text{sign}(c))^{\top}$; $\boldsymbol{\beta} \leftarrow (\boldsymbol{\beta}, 0)^{\top}$
else
return $(\sigma, \boldsymbol{\beta})$

Table 1. Speed trials experiments with the same settings as in [3]. The running times, in seconds, are averaged over 10 runs. All methods and trials were implemented in C in similar fashions.

n	$ \mathcal{D} $	Method	Population correlation between features					
			0	0.1	0.2	0.5	0.9	0.95
100	1000	homotopy	0.13	0.12	0.13	0.14	0.14	0.14
		iso- λ descent	0.09	0.09	0.09	0.10	0.10	0.10
		CCD	0.21	0.21	0.24	0.46	1.21	2.64
	5000	homotopy	0.63	0.65	0.62	0.67	0.62	0.60
		iso- λ descent	0.53	0.56	0.54	0.59	0.54	0.54
		CCD	1.39	1.39	1.54	2.32	7.39	8.53
	20000	homotopy	2.39	3.06	2.69	3.22	3.30	3.36
		iso- λ descent	2.06	2.56	2.25	2.68	2.76	2.85
		CCD	5.38	7.22	6.06	11.14	37.24	47.07
1000	100	homotopy	0.22	0.22	0.22	0.22	0.20	0.16
		iso- λ descent	0.19	0.19	0.19	0.19	0.17	0.13
		CCD	0.18	0.23	0.30	0.65	1.62	2.13
5000		homotopy	0.71	0.70	0.70	0.70	0.70	1.05
		iso- λ descent	0.67	0.67	0.67	0.66	0.65	0.97
		CCD	0.61	0.75	0.91	1.44	4.31	8.07

5 Conclusion

The LASSO can thus be computed by three simple and efficient algorithms. Regardless of the running times exhibited in previous section, each offers specific advantages.

The cyclical coordinate descent is a very simple algorithm that is less subject to implementation hazards and conditioning problems. Also, since it does not involve the active Gram matrix, it supports sample reweighting and can thus be used to compute the elastic net operator. On the other hand, its convergence being asymptotic, a stopping criterion/parameter is needed.

The homotopy method gives the exact full regularization path (RP), while keeping a computational cost of the same order as the two others. However, this is rarely needed, and solutions for a predefined set of regularization parameters are often sufficient.

In this respect, the iso- λ descent method offers an efficient way to compute such a sequence, using one solution as a hot start for the next one, to which it converges in a small number of steps, very close to the number of steps on the RP. The path followed between the two is itself close to the RP, as it shares the property of simultaneously increasing the coefficient L_1 norm and decreasing the squared residual. Another nice property of this algorithm is its ability to handle continuous feature sets, e.g. the set of all Gaussian functions over $\mathcal{X} = \mathbb{R}^p$ parameterized by their covariance matrix. This will be investigated, together with expositions of the proofs mentioned here, in a publication to come.

References

1. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1) (1996) 267–288
2. Fu, W.J.: Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7**(3) (September 1998) 397
3. Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. *Annals of Applied Statistics* **1**(2) (2007) 302–332
4. Osborne, M., Presnell, B., Turlach, B.: A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20**(3) (2000) 389
5. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. *The Annals of Statistics* **32**(2) (April 2004) 407–499
6. Nash, S.G., Sofer, A.: *Linear and nonlinear programming*. McGraw-Hill (New York) (1996)
7. Osborne, M., Presnell, B., Turlach, B.A.: On the LASSO and Its Dual. *Journal of Computational and Graphical Statistics* **9**(2) (2000) 319 – 337