A MULTIDIMENSIONAL SHRINKAGE-THRESHOLDING OPERATOR

Arnau Tibau Puig[1],[2], *Ami Wiesel*[1] *and Alfred O. Hero III*[1]

(1) University of Michigan - Department of Electrical Engineering
1301 Beal Avenue Ann Arbor, MI 48109-2122
(2) Supélec - Département SSE
3, rue Joliot-Curie, 91192 Gif-sur-Yvette

ABSTRACT

The scalar shrinkage-thresholding operator (SSTO) is a key ingredient of many modern statistical signal processing algorithms including: sparse inverse problem solutions, wavelet denoising, and JPEG2000 image compression. In these applications, it is customary to select the threshold of the operator by solving a scalar sparsity penalized quadratic optimization. In this work, we present a natural multidimensional extension of the scalar shrinkage thresholding operator. Similarly to the scalar case, the threshold is determined by the minimization of a convex quadratic form plus an euclidean penalty, however, here the optimization is performed over a domain of dimension $N \geq 1$. The solution to this convex optimization problem is called the multidimensional shrinkage threshold operator (MSTO). The MSTO reduces to the standard SSTO in the special case of $N = 1$. In the general case of $N > 1$ the optimal MSTO threshold can be found by a simple convex line search. We present three illustrative applications of the MSTO in the context of non-linear regression: l_2-penalized linear regression, Group LASSO linear regression and Group LASSO logistic regression.

Index Terms— Multidimensional Shrinkage-Thresholding Operator, Iterative Group Shrinkage-Thresholding, Group LASSO regression

1. INTRODUCTION

The scalar shrinkage-threshold operator is central to modern signal processing algorithms such as Iterative Thresholding [1] for image deblurring [2], wavelet-based deconvolution [3] or sparse approximation [4].

In this paper, we introduce a multidimensional generalization of the scalar shrinkage thresholding operator. We define this operator as the minimization of a convex quadratic form plus an Euclidean norm penalty. We analyze this non-differentiable optimization problem and discuss its properties. In particular, in analogy to the scalar shrinkage operator, we show that this generalization yields a multidimensional Shrinkage Thresholding Operator (MSTO) which takes a vector as an input and shrinks it or thresholds it depending on its Euclidean norm. For this purpose, we reformulate the problem as a constrained quadratic problem with a conic constraint. This principle leads to a theoretical result that transforms this multidimensional optimization problem into a simple line search which can be efficiently implemented. We show by simulations that evaluating the MSTO using line search is competitive with state-of-the-art convex solvers.

In the second part of the paper, we discuss applications of the MSTO to statistical regression. First, we consider the Euclidean-norm penalized least squares and discuss its relation to ridge regression and robust regression [5]. Next, we address group LASSO linear regression [6]. In the special case of a block-orthogonal design matrix, we show that the problem can be reduced to evaluating the MSTO for each block. For other Group LASSO problems, we propose two iterative applications of the MSTO. In the first approach, we use Block Coordinate Descent to solve the linear regression problem with an arbitrary design matrix. The second approach tackles more complicated cost functions such as the logistic regression objective. Due to its similarity to the well-known class of Iterative Thresholding Algorithms [1], we name the latter Iterative Group Shrinkage-Thresholding (IGST). In both cases, the MSTO enables one to solve large scale Group LASSO problems using a small number of simple line searches.

This paper is organized as follows. In Section 2 we define the MSTO and introduce our main theoretical result. In Section 3 we illustrate how to apply the MSTO to solve four different regression problems. We present simulation results in Section 5.

The following notation is used. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. The superscripts $'$ and $\dagger$ denote the transpose and the matrix pseudoinverse operators, respectively. $\boldsymbol{\theta}_{(S)}$ denotes the subvector constructed from the indices S. For a set of indices S, $\bar{S}$ denotes its complementary set. $\lambda_i(\mathbf{X})$, $\lambda_{max}(\mathbf{X})$ and $\lambda_{min}(\mathbf{X})$ refer to the i-th, maximum and minimum eigenvalue of a matrix $\mathbf{X}$ and $\mathcal{R}(\mathbf{X})$ denotes its range. $\mathbf{I}$ is the identity matrix. $\mathbf{X}_{S,T}$ denotes the submatrix obtained from taking the columns and rows indexed by S and T respectively and $[\mathbf{X}]$ represents the subspace spanned by the columns of $\mathbf{X}$. We define $\text{diag}(\mathbf{x})$ as the operator that returns a diagonal

matrix with the vector $\mathbf{x}$ in its diagonal. Also, we define $(x)_+$ as the operator that yields 0 is $x \leq 0$ and x otherwise. We denote the multivariate normal distribution of mean μ and covariance matrix Σ by $\mathcal{N}(\mu, \Sigma)$. Finally, $x \sim \mathcal{D}$ means that x is a random variable with distribution $\mathcal{D}$.

2. MULTIDIMENSIONAL SHRINKAGE-THRESHOLDING OPERATOR (MSTO)

The scalar shrinkage-thresholding operator is usually defined as:

$$\begin{aligned} \Phi_{\lambda,a}(g) &:= \arg\min_x \tfrac{1}{2}ax^2 + gx + \lambda|x| \\ &= -\tfrac{1}{a}(|g| - \lambda)_+ \operatorname{sign}(g), \end{aligned} \tag{1}$$

where $a, \lambda > 0$ and $g \in \mathbb{R}$. A natural multidimensional generalization of this operator is:

$$\mathcal{T}_{\lambda,\mathbf{H}}(\mathbf{g}) := \arg\min_{\mathbf{x}} \tfrac{1}{2}\mathbf{x}'\mathbf{H}\mathbf{x} + \mathbf{g}'\mathbf{x} + \lambda\|\mathbf{x}\|_2, \tag{2}$$

where $\mathbf{H} \succeq 0$, $\lambda > 0$ and $\mathbf{g} \in \mathbb{R}^N$. The next theorem shows that this operator behaves in fact as a *Multidimensional Shrinkage Thresholding Operator* (MSTO) which takes $\mathbf{g}$ and *thresholds* or *shrinks* it to a value that depends on its norm.

Theorem 2.1. *Let $\mathbf{H} \succeq 0$, $\mathbf{g} \in \mathcal{R}(\mathbf{H})$ and $\lambda > 0$. The optimal value of the N-dimensional, non-differentiable problem:*

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}'\mathbf{H}\mathbf{x} + \mathbf{g}'\mathbf{x} + \lambda\|x\|_2 \tag{3}$$

is equal to the optimal value of the convex one-dimensional problem:

$$\min_{\eta \geq 0} \quad \eta\left(1 - \tfrac{1}{2}\mathbf{g}'\mathbf{B}^{\dagger}(\eta)\,\mathbf{g}\right), \tag{4}$$

where $\mathbf{B}(\eta) := \eta\mathbf{H} + \frac{\lambda^2}{2}\mathbf{I}$. Furthermore, the solution of (4) satisfies:

$$\eta = \begin{cases} 0 & \text{if } \|\mathbf{g}\|_2 \leq \lambda \\ \eta > 0 : \left\|\frac{\lambda}{2}\mathbf{B}^{\dagger}(\eta)\,\mathbf{g}\right\|_2^2 = 1 & \text{if } \|\mathbf{g}\|_2 > \lambda \end{cases} \tag{5}$$

and the solutions of (3) and (4) are related by:

$$\mathbf{x} = \begin{cases} -\eta\mathbf{B}^{\dagger}(\eta)\,\mathbf{g} & \text{if } \|\mathbf{g}\|_2 > \lambda \\ \mathbf{0} & \text{if } \|\mathbf{g}\|_2 \leq \lambda \end{cases}, \tag{6}$$

where $\eta \geq 0$ satisfies (5).

Proof. Since $\mathbf{H} \succeq 0$ and $\|\cdot\|_2$ is a norm, it follows that $\mathbf{x}'\mathbf{H}\mathbf{x} + \mathbf{g}'\mathbf{x}$ and $\|\mathbf{x}\|_2$ are convex functions of $\mathbf{x}$. Then (3) is equivalent to the following quadratic program with a second order conic constraint:

$$\begin{aligned} \min_{x,t} \quad & \tfrac{1}{2}\mathbf{x}'\mathbf{H}\mathbf{x} + \mathbf{g}'\mathbf{x} + t \\ \text{s.t.} \quad & \|\lambda\mathbf{x}\|_2 - t \leq 0. \end{aligned} \tag{7}$$

Slater's condition for generalized inequalities is verified and strong duality holds. Then the dual can be written as ([7], Section 5.9.1):

$$\max_{\|\mathbf{u}\|_2 - \mu \leq 0} \min_{x,t} \frac{1}{2}\mathbf{x}'\mathbf{H}\mathbf{x} + \mathbf{g}'\mathbf{x} + t - \mathbf{u}'(\lambda\mathbf{x}) - \mu t. \tag{8}$$

The inner minimization is unbounded in t unless $\mu = 1$ and in $\mathbf{x}$ unless $\lambda\mathbf{u} \in \mathcal{R}(\mathbf{H})$. Otherwise, its optimum satisfies:

$$\mathbf{x} = -\mathbf{H}^{\dagger}(\mathbf{g} - \lambda\mathbf{u}). \tag{9}$$

Plugging (9) in (8), and using the fact that a non differentiable dual conic constraint $\|\mathbf{u}\|_2 - \mu \leq 0$ with $\mu = 1$ is equivalent to a standard quadratic constraint $\|\mathbf{u}\|_2^2 \leq 1$, we obtain the following dual concave maximization:

$$\max_{\|\mathbf{u}\|_2^2 \leq 1, \mathbf{u} \in \mathcal{R}(\mathbf{H})} -\frac{1}{2}(\mathbf{g} - \lambda\mathbf{u})'\mathbf{H}^{\dagger}(\mathbf{g} - \lambda\mathbf{u}). \tag{10}$$

The standard lagrange dual of this problem is:

$$\min_{\eta \geq 0} \max_{\mathbf{u} \in \mathcal{R}(\mathbf{H})} -\frac{1}{2}(\mathbf{g} - \lambda\mathbf{u})'\mathbf{H}^{\dagger}(\mathbf{g} - \lambda\mathbf{u}) - \eta(\mathbf{u}'\mathbf{u} - 1). \tag{11}$$

Since $\mathbf{H} \succeq 0$ and $\mathbf{H}^{\dagger}\mathbf{g} \in \mathcal{R}\left(\mathbf{H}^{\dagger}\right)$, the inner maximization is a simple quadratic problem in $\mathbf{u}$ with solution:

$$\mathbf{u} = \frac{\lambda}{2}\mathbf{B}^{\dagger}(\eta)\,\mathbf{g}, \tag{12}$$

where $\mathbf{B}(\eta) = \left(\eta\mathbf{H} + \frac{\lambda^2}{2}\mathbf{I}\right)$. This leads to the following one-dimensional minimization over the lagrange multiplier η:

$$\min_{\eta \geq 0} \eta\left(1 - \frac{1}{2}\mathbf{g}'\mathbf{B}^{\dagger}(\eta)\,\mathbf{g}\right), \tag{13}$$

which proves the equivalence between (3) and (4). The eigenvalues of $\mathbf{B}^{\dagger}(\eta)$ are real and can be characterized as:

$$\lambda_i\left(\mathbf{B}^{\dagger}(\eta)\right) = \frac{1}{\eta\lambda_i(\mathbf{H}) + \frac{\lambda^2}{2}}. \tag{14}$$

Since $\eta \geq 0$, $\lambda_i(\mathbf{H}) \geq 0$ and $\lambda > 0$, it holds that $0 < \lambda_i\left(\mathbf{B}^{\dagger}(\eta)\right) \leq \frac{2}{\lambda^2}$. Therefore, if $\|\mathbf{g}\|_2 \leq \lambda$ then $\frac{1}{2}\mathbf{g}'\mathbf{B}^{\dagger}(\eta)\,\mathbf{g} \leq 1$ and it follows that $\eta\left(1 - \frac{1}{2}\mathbf{g}'\mathbf{B}^{\dagger}(\eta)\,\mathbf{g}\right) \geq 0$. Then the minimum in (13) is attained by chosing $\eta = 0$. On the other hand, if $\|\mathbf{g}\|_2 > \lambda$, by complementary slackness $\mathbf{u}$ has to verify $\|\mathbf{u}\|_2^2 = 1$. Using (12), this leads to the following equation in $\eta \geq 0$:

$$\left\|\frac{\lambda}{2}\mathbf{B}^{\dagger}(\eta)\,\mathbf{g}\right\|_2^2 = 1, \tag{15}$$

which has no solution at $\eta = 0$. Combining the two solutions of (13) yields (5). Plugging (12) in (9) yields (6). □

Remark 2.2. This result shows that evaluating the MSTO in a space of arbitrarily large dimension reduces to solving the one-dimensional convex problem (13) or, equivalently, equation (5), only if the norm of the input vector is above the threshold λ. Otherwise, the evaluation is immediate and no computation is necessary, the output being set to the zero vector.

Furthermore, in the special case of $\mathbf{H} = k\mathbf{I}$ for some $k > 0$, using Theorem 2.1 leads to a closed form expression for the MSTO. In this case, equation (5) has the positive solution $\eta = \frac{\lambda}{2k}\left(\|\mathbf{g}\|_2 - \lambda\right)$ and the MSTO reduces to:

$$\mathcal{T}_{\lambda,k\mathbf{I}}\left(\mathbf{g}\right) = -\frac{1}{k}\left(\|\mathbf{g}\|_2 - \lambda\right)_+ \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \tag{16}$$

which is analogous to (1) if we define the multidimensional sign function as $\text{sign}\left(\mathbf{x}\right) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. This expression coincides with the *vectorial soft-threshold* defined in [8]. If $\mathbf{H} \neq k\mathbf{I}$, the unique positive solution of the non-linear equation in (5) can be efficiently found using any standard line-search method. In particular, in our simulations we use a Newton Raphson implementation inspired on [9].

3. APPLICATIONS

Here we illustrate the MSTO by considering a few applications in statistical signal processing.

3.1. Linear regression with l_2 norm penalty

Given a vector of n observations $\mathbf{y}$ and an $n \times N$ design matrix $\mathbf{X}$, we consider the following class of problems:

$$\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^p + \lambda\|\boldsymbol{\theta}\|_2^q. \tag{17}$$

Depending on p and q, this problem specializes to ridge regression ($p = 2$, $q = 2$), robust least-squares (RLS) ($p = 1$, $q = 1$) [Theorem 3.2, [5]] or l_2-penalized least squares ($p = 2$, $q = 1$). To our knowledge, the latter has not been treated in the literature. The following corollary characterizes the solution to this problem.

Corollary 3.1. *The solution to the l_2-penalized least squares*

$$\min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2 \tag{18}$$

is:

1. *If $\|\mathbf{X}'\mathbf{y}\|_2 \leq \frac{\lambda}{2}$, then $\boldsymbol{\theta} = \mathbf{0}$, i.e., thresholding.*
2. *If $\|\mathbf{X}'\mathbf{y}\|_2 > \frac{\lambda}{2}$, then $\boldsymbol{\theta}$ is given by the classical shrinkage least squares*

$$\boldsymbol{\theta} = \left(\mathbf{X}'\mathbf{X} + \epsilon\mathbf{I}\right)^{\dagger}\mathbf{X}'\mathbf{y}, \tag{19}$$

with shrinkage parameter $\epsilon = \frac{\lambda^2}{4\eta}$ where $\eta > 0$ is chosen to satisfy the complementary slackness condition $\|\lambda\left(2\eta\mathbf{X}\mathbf{X}' + \frac{\lambda^2}{2}\mathbf{I}\right)^{\dagger}\mathbf{X}'\mathbf{y}\|_2^2 = 1$.

In the special case where $\mathbf{X}$ is orthogonal ($2\mathbf{X}'\mathbf{X} = k\mathbf{I}$) then (18) has a closed form solution with $\epsilon = \frac{\lambda k}{2(k\|\mathbf{y}\|_2 - \lambda)}$.

The proof of this Corollary follows immediately from Theorem 2.1.

3.2. Group LASSO Linear Regression with block-orthogonal design

Given $\mathbf{X}$, $\mathbf{y}$ as in the previous section and p disjoint groups of indices $G_i \subseteq \{1, ..., N\}$ satisfying $\cup_i G_i = \{1, ..., N\}$, the Group LASSO linear regression problem [6] is defined as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in R^N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^{p} \lambda_i \|\boldsymbol{\theta}_{(G_i)}\|_2, \tag{20}$$

where λ_i are fixed penalty parameters which we assume known.

When the design matrix $\mathbf{X}$ is block-orthogonal (i.e. $\mathbf{X}'_{G_i,G_j}\mathbf{X}_{G_i,G_j} = \mathbf{0}$ for $i \neq j$) and letting $\mathbf{H} = 2\mathbf{X}'\mathbf{X}$, $\mathbf{g} = -2\mathbf{X}'\mathbf{y}$, we can use Theorem 2.1 to obtain the following solution to (20) for each group G_i:

$$\hat{\boldsymbol{\theta}}_{(G_i)} = \mathcal{T}_{\lambda_i,\mathbf{H}_{G_i,G_i}}\left(\mathbf{g}_{(G_i)}\right). \tag{21}$$

Therefore computing the optimal $\boldsymbol{\theta}$ reduces to p evaluations of the MSTO.

3.3. Block Coordinate Descent for Linear Regression

For an arbitrary design matrix $\mathbf{X}$, problem (20) can be solved using a Block Coordinate Descent (BCD) algorithm. The main idea of the BCD method is to iteratively solve (20) for each block G_i, letting the parameters corresponding to the other blocks remain fixed. Defining $\mathbf{H} = 2\mathbf{X}'\mathbf{X}$, $\mathbf{g} = -2\mathbf{X}'\mathbf{y}$ and using the MSTO operator (2) we can obtain the following update rule for each group G_i at iteration t:

$$\boldsymbol{\theta}^t_{(G_i)} \leftarrow \mathcal{T}_{\lambda,\mathbf{H}_{G_i,G_i}}\left(2\boldsymbol{\theta}^{t-1}_{(\bar{G}_i)}\mathbf{H}_{\bar{G}_i,G_i} + \mathbf{g}_{(G_i)}\right), \tag{22}$$

where $\bar{G}_i$ is the complementary set of indices with respect to G_i. This sequence is guaranteed to converge to the global solution to (20) [10].

3.4. Iterative Group Shrinkage Thresholding (IGTS) for Logistic Regression

Defining the same quantities $\mathbf{y}$, $\mathbf{X}$, $\mathbf{H}$, $\mathbf{g}$ and $\{G_i\}_{i=1}^p$ as in the section above, we next consider the Group LASSO logistic regression problem [11]:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in R^N} \sum_{i=1}^{n} \log\left(1 + e^{-y_i\mathbf{x}_i'\boldsymbol{\theta}}\right) + \sum_{i=1}^{p} \lambda_i \|\boldsymbol{\theta}_{(G_i)}\|_2, \tag{23}$$

where $\mathbf{x}_i$ are the rows of the design matrix $\mathbf{X}$ and λ_i are fixed penalty parameters. Our algorithm emulates the well-known Iterative Shrinkage Thresholding (IST) class of algorithms for l_1 penalized regression [1] or classification [12]. For each group G_i, and choosing $\mathbf{H}$ block-diagonal with respect to the groups G_i, the update rule of the IGTS algorithm at iteration t is:

$$\boldsymbol{\theta}^t_{(G_i)} \leftarrow \mathcal{T}_{\lambda,\mathbf{H}_{G_i,G_i}}\left(\left[\nabla f\left(\boldsymbol{\theta}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{t-1}} - 2\boldsymbol{\theta}^{t-1}\mathbf{H}\right]_{(G_i)}\right), \tag{24}$$

where $\nabla f\left(\boldsymbol{\theta}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{t-1}}$ is the gradient of the logistic cost evaluated at $\boldsymbol{\theta}^{t-1}$. Optimization transfer theory guarantees that the sequence defined by the update rule (24) converges to the global solution to (23) if $\mathbf{H} \succeq \frac{1}{4}\mathbf{X}'\mathbf{X}$ [12].

4. SIMULATIONS

Here we demonstrate the computational advantages of evaluating the MSTO using Theorem 2.1. Due to space limitations we can only present simulations for the l_2-penalized linear regression problem presented in section 3.1. We generate the random symmetric matrices $\mathbf{H} \succeq 0$ from the random matrices $\mathbf{X_i}$ defined in Table 1 and taking $\mathbf{H}_i = \mathbf{X}_i'\mathbf{X}_i$. We also

Diagonal	Arbitrary Well-conditioned	Arbitrary Ill-conditioned
$\mathbf{X}_1 \sim \text{diag}\left(\mathcal{N}\left(N\mathbf{1}, \mathbf{I}\right)\right)$	$\mathbf{X}_2 \sim \mathcal{N}\left(N\mathbf{I}, \mathbf{I}\right)$	$\mathbf{X}_3 \sim \mathcal{N}\left(\mathbf{I}, \mathbf{I}\right)$

Fig. 1. Generation of random $\mathbf{X}_i$ to obtain $\mathbf{H}_i = \mathbf{X}_i'\mathbf{X}_i \succeq 0$.

generate $\mathbf{g} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$.

We consider two experiments. For each experiment, we evaluate the MSTO using two different approaches: (1) *MSTO N-R* (using Theorem 2.1 and Newton Raphson to solve (5)) and (2) *Sedumi* (a state-of-the-art SOCP solver to solve (3)). In the first experiment we calculate the average elapsed computation times for evaluating the MSTO over 10 realizations when choosing $\lambda = 10^{-2}$ and dimensions ranging from $N = 5$ to $N = 300$. In the second one, we fix $N = 50$ and consider different λ's in the range $\left[10^{-6}, ..., 100\right]$.

Figure 2 depicts the results of the first experiment, showing a consistently better performance than *sedumi* for the ill-conditioned matrix $\mathbf{H}_3$. For $\mathbf{H}_1$ and $\mathbf{H}_2$, evaluating the MSTO using Newton Rapshon is faster for dimensions up to $N = 150$ and in the same order of *sedumi* elsewhere. On the other hand, Figure 3 shows that our method dominates for $N = 50$ over a wide range of values of the regularization parameter λ.

We conclude that evaluating the MSTO using Theorem 2.1 and Newton Raphson is specially advantageous when $\mathbf{H}$ in (3) is bad-conditionned, which is usual in small-sample regression (i.e. when the $n \times N$ design matrix $\mathbf{X}$ is such that $N \gg n$). Also, Figure 3 leads us to think that our result will be also particularly useful when we know a priori that the regression result is likely to be set to $\mathbf{0}$ as the MSTO does not require line search when $\lambda > \|\mathbf{g}\|_2$.

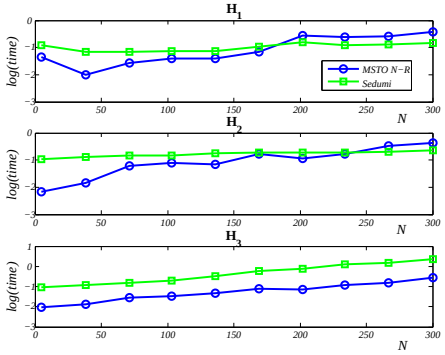

Fig. 2. Comparison of Sedumi and MSTO with Newton-Raphson implementation for solving (3). We consider three situations: $\mathbf{H}_1$, a well conditioned diagonal matrix (top), $\mathbf{H}_2$, a well conditioned matrix (middle) and $\mathbf{H}_3$, an ill-conditioned matrix (bottom).

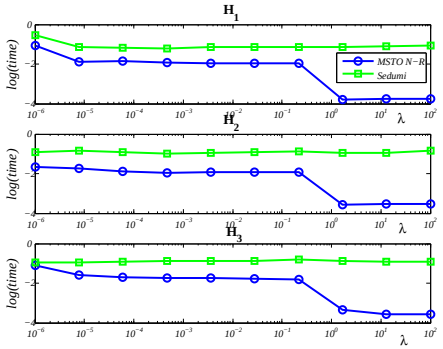

Fig. 3. Same as in Figure 2, now for λ in the range $\left[10^{-6}, ..., 100\right]$ and fixed $N = 50$.

5. ACKNOWLEDGEMENT

The authors would like to thank Amir Beck for his Matlab code solving an equation similar to (5). This work was partially supported by a grant from the National Science Foundation CCR-0325571 and a Marie Curie Fellowship to the second author.

6. REFERENCES

[1] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math*, vol. 57, no. 11, pp. 1413–1457, 2004.

[2] A. Beck and M. Teboulle, "Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J. Imaging Sciences*, vol. To appear., 2008.

[3] RD Nowak and MAT Figueiredo, "Fast wavelet-based image deconvolution using the EM algorithm," in *Conference Record of the 35th Asilomar Conference*, 2001, vol. 1.

[4] KK Herrity, AC Gilbert, and JA Tropp, "Sparse Approximation Via Iterative Thresholding," in *IEEE ICASSP Proceedings*, 2006, vol. 3.

[5] L. El Ghaoui and H. Lebret, "Robust solutions to least squares problems with uncertain data," *SIAM Journal Matrix Analysis and Applications*, October 1997.

[6] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, pp. 49–67, 2006.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press.

[8] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2479–2493, July 2009.

[9] A. Beck, "The matrix-restricted total least-squares problem," *Signal Process.*, vol. 87, no. 10, pp. 2303–2312, 2007.

[10] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

[11] L. Meier, S. van de Geer, and P. Buhlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B*, vol. 70, February 2008.

[12] B. Krishnapuram, L. Carin, M. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, 2005.