Publication VI

Timo Similä (2007). Majorize-minimize algorithm for multiresponse sparse regression, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2007*, Vol. II, pp. 553–556.

VI

MAJORIZE-MINIMIZE ALGORITHM FOR MULTIRESPONSE SPARSE REGRESSION

Timo Similä

Helsinki University of Technology
Laboratory of Computer and Information Science
P. O. Box 5400, FI-02015 HUT, Finland

ABSTRACT

Multiresponse sparse regression is the problem of estimating many response variables using a common subset of input variables. Our model is linear, so row sparsity of the coefficient matrix implies subset selection. This is formulated as the problem of minimizing the residual sum of squares, where the row norms of the coefficient matrix are penalized. The proposed approach differs from existing ones in that any penalty function that is increasing, differentiable, and concave can be used. A convergent majorize-minimize algorithm is adopted for minimization. We also propose an active set strategy for tracking the nonzero rows of the coefficient matrix when the minimization is performed for a sequence of descending values of the penalty parameter. Numerical experiments are given to illustrate the active set strategy and analyze penalization with different degrees of concavity.

***Index Terms*—** MM algorithm, variable selection, row sparse matrices, simultaneous sparse approximation.

1. INTRODUCTION

Suppose that we have q response variables and m input variables from which we have n observations. Consider the model

$$\underset{(n\times q)}{\boldsymbol{Y}} = \underset{(n\times m)}{\boldsymbol{X}}\,\underset{(m\times q)}{\boldsymbol{W}} + \underset{(n\times q)}{\boldsymbol{E}}\,. \tag{1}$$

The columns of $\boldsymbol{X}$ are zero mean and constant norm input variables, $\boldsymbol{W}$ is a row sparse coefficient matrix, and $\boldsymbol{E}$ is an unknown noise matrix. The row sparsity of $\boldsymbol{W}$ has the effect that some of the input variables do not contribute to the response variables at all. Multiresponse sparse regression (MRSR) is the problem of identifying and estimating the nonzero rows of $\boldsymbol{W}$ given the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$.

A traditional paradigm for solving the MRSR problem is to apply such measures as prediction error, tests of statistical significance, or information criteria to rank different combinations of input variables. Some stepwise algorithm is used to find promising combinations. Theoretical results show that certain greedy algorithms succeed when the input variables are weakly correlated [1]. On the other hand, empirical evidence shows that they fail in the presence of higher correlations [2]. There are better and less greedy algorithms for stepwise subset selection [2], but it is not always clear, what is exactly the objective that they optimize.

Recently, relaxation techniques for the MRSR problem have emerged in the signal processing and statistical communities, apparently through independent research efforts [3], [4], [5], [6], [7]. These techniques penalize the model fitting in a way that the estimate for $\boldsymbol{W}$ becomes row sparse. Particularly, when the relaxation problem is convex, efficient methods exist for finding a global solution.

In this article, MRSR is formulated as a relaxation problem to minimize the objective function[1]

$$E_\mu^\lambda(\boldsymbol{W}) = \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 + \lambda\sum_{i=1}^{m} p_\mu(\|\boldsymbol{w}_i\|), \tag{2}$$

where $\lambda \geq 0$ is used to balance between model fitting and penalization. The subscript $\mu \geq 0$ measures perturbation from the desired objective $E_0^\lambda(\boldsymbol{W})$, in a way to be made precise in Section 2.2. Until then, one can fix $\mu = 0$. The penalty function $p_0(s)$ is increasing on $s \geq 0$.

The objective $E_0^\lambda(\boldsymbol{W})$ is introduced in [8], where the single response $q = 1$ case is studied and penalty functions resulting in estimates that satisfy the properties of unbiasedness, sparsity, and continuity are proposed. In [9], a majorize-minimize (MM) algorithm is developed for optimizing the perturbed objective $E_\mu^\lambda(\boldsymbol{W})$ with $\mu > 0$. The perturbation enables analyzing the convergence of the algorithm and it also solves some other deficiencies. This article extends the works [8] and [9] to the multiple response $q > 1$ case. The global convergence of the extended algorithm follows from general results of the MM theory [9], [10]. The main contribution of this article is a stable active set strategy for tracking the nonzero rows of the solution $\boldsymbol{W}(\lambda)$ when $E_0^\lambda(\boldsymbol{W})$ is minimized for a sequence of descending values of λ. This lessens computational burden, since only the active rows need to be optimized while the solution remains the same.

Existing studies of the $q > 1$ case are regarded to the objective (2) as follows. The choice $p_0(s) = s$ makes $E_0^\lambda(\boldsymbol{W})$ a

[1] Throughout the article, $\|\cdot\|$ denotes the L_2 vector norm and the Frobenius matrix norm is $\|\cdot\|_{\mathrm{F}} = \|\mathrm{vec}(\cdot)\|$. A bolded lower case letter refers to a column vector whose elements are taken from a single row of a matrix. For example, we have $\boldsymbol{W}^T = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m]$.

convex function and the resulting minimization problem has been studied extensively [4], [5], [7]. Here the main focus is on strictly concave penalty functions, which are needed to avoid unnecessary modeling bias when the true unknown value $\|\boldsymbol{w}_i\|$ is large [8]. The Regularized M-FOCUSS algorithm [4] uses $p_0(s) = z^{-1}s^z$, which is strictly concave on $s \geq 0$ for $z \in (0, 1)$. However, the algorithm lacks rigorous proof of convergence and it appears to be unable to change the status of a row of $\boldsymbol{W}$ from zero to nonzero in the process of iteration.

2. OPTIMIZATION

2.1. First order optimality conditions

We start with the first order optimality conditions for minimizing the desired objective $E_0^\lambda(\boldsymbol{W})$ under the assumption that $p_0(s)$ is differentiable on $s \geq 0$. In this case, the mapping $p_0(\|\boldsymbol{w}_i\|)$ is differentiable on the whole domain $\mathbb{R}^q$, possibly excluding the point $\boldsymbol{w}_i = \boldsymbol{0}$. Since nondifferentiability at this point turns out to be the key factor for row sparsity, we cannot rely on gradient-based calculus. Instead, we note that $E_0^\lambda(\boldsymbol{W})$ is directionally differentiable everywhere.

The derivative of $E_0^\lambda(\boldsymbol{W})$ at $\boldsymbol{W}$ in a direction $\boldsymbol{V}$ is

$$E_0^{\lambda\prime}(\boldsymbol{W})(\boldsymbol{V}) = \sum_{i\in\mathcal{I}} \boldsymbol{v}_i^T\left(\boldsymbol{g}_i + \lambda\frac{p_0'(\|\boldsymbol{w}_i\|)}{\|\boldsymbol{w}_i\|}\boldsymbol{w}_i\right) + \sum_{i\notin\mathcal{I}}(\boldsymbol{v}_i^T\boldsymbol{g}_i + \lambda p_0'(0)\|\boldsymbol{v}_i\|), \tag{3}$$

where $\boldsymbol{G} = -\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})$ denotes the gradient of the loss function $\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2$ and $\mathcal{I} = \{i : \|\boldsymbol{w}_i\| > 0\}$.

Proposition 1 *$E_0^{\lambda\prime}(\boldsymbol{W})(\boldsymbol{V}) \geq 0$ holds for all $\boldsymbol{V}$ if and only if the following conditions are satisfied*

$$-\boldsymbol{g}_i = \lambda\frac{p_0'(\|\boldsymbol{w}_i\|)}{\|\boldsymbol{w}_i\|}\boldsymbol{w}_i,\ i \in \mathcal{I} \tag{4}$$

$$\|\boldsymbol{g}_i\| \leq \lambda p_0'(0) \quad ,\ i \notin \mathcal{I}. \tag{5}$$

Proof Let (4)–(5) hold. Then we clearly have $E_0^{\lambda\prime}(\boldsymbol{W})(\boldsymbol{V}) = \sum_{i\notin\mathcal{I}}(\boldsymbol{v}_i^T\boldsymbol{g}_i + \lambda p_0'(0)\|\boldsymbol{v}_i\|) \geq 0\ \forall \boldsymbol{V}$ because of the Cauchy-Schwarz inequality $\boldsymbol{v}_i^T\boldsymbol{g}_i \geq -\|\boldsymbol{v}_i\|\|\boldsymbol{g}_i\|$.

Let $E_0^{\lambda\prime}(\boldsymbol{W})(\boldsymbol{V}) \geq 0$ hold $\forall \boldsymbol{V}$. (I) Assume $\exists i \in \mathcal{I}$ such that $\boldsymbol{h}_i := \boldsymbol{g}_i + \lambda\frac{p_0'(\|\boldsymbol{w}_i\|)}{\|\boldsymbol{w}_i\|}\boldsymbol{w}_i \neq \boldsymbol{0}$. Set $\boldsymbol{v}_j = \boldsymbol{0}$ for $j \neq i$ and $\boldsymbol{v}_i = -\boldsymbol{h}_i$. (II) Assume $\exists i \notin \mathcal{I}$ such that $\|\boldsymbol{g}_i\| > \lambda p_0'(0)$. Set $\boldsymbol{v}_j = \boldsymbol{0}$ for $j \neq i$. If $\boldsymbol{g}_i \neq \boldsymbol{0}$ set $\boldsymbol{v}_i = -\boldsymbol{g}_i$ and, otherwise, take any $\boldsymbol{v}_i \neq \boldsymbol{0}$. Both (I) and (II) lead to the contradiction $E_0^{\lambda\prime}(\boldsymbol{W})(\boldsymbol{V}) < 0$, so (4)–(5) must hold. □

If $\boldsymbol{W}$ minimizes the directionally differentiable function $E_0^\lambda(\boldsymbol{W})$, then $E_0^{\lambda\prime}(\boldsymbol{W})(\boldsymbol{V}) \geq 0$ holds for all $\boldsymbol{V}$ [11]. Thus, (4)–(5) are necessary for optimality. In the case that $p_0(\|\boldsymbol{w}_i\|)$ is convex, it can be shown that (4)–(5) are also sufficient. Observe that the penalty function encourages row sparsity only when $p_0'(0) > 0$ applies. The same property has been derived for the single response $q = 1$ case in [8].

2.2. MM algorithm

A general MM algorithm reduces the objective function monotonically by minimizing a succession of approximations, each of which majorizes the objective in a certain sense. In essence, the MM algorithm replaces a difficult optimization problem by a sequence of easier subproblems. Our desired objective $E_0^\lambda(\boldsymbol{W})$ is difficult to minimize, because it may be nonconvex and thereby admit multiple local minima, and because the function $p_0(\|\boldsymbol{w}_i\|)$ is nondifferentiable at $\boldsymbol{w}_i = \boldsymbol{0}$ under the sparsity assumption $p_0'(0) > 0$. The latter difficulty also prevents the use of a differentiable majorizer at this point, which spoils the idea of the MM algorithm with easy subproblems.

We adopt the approach taken in [9] and consider a perturbation $\mu > 0$, which makes $E_\mu^\lambda(\boldsymbol{W})$ differentiable everywhere, but maintains it close to $E_0^\lambda(\boldsymbol{W})$. The smaller the value μ, the more similar the two functions are. Then we use an MM algorithm to minimize the new perturbed objective. It is worth noting that minimizing $E_\mu^\lambda(\boldsymbol{W})$ directly with a gradient-based method might not be easy, since it is very close to the nondifferentiable function $E_0^\lambda(\boldsymbol{W})$ when μ is small.

To begin with the algorithm, we define $p_\mu(s)$ as follows

$$p_\mu(s) = p_0(s) - \mu\int_0^s \frac{p_0'(t)}{\mu + t}dt, \quad \mu > 0 \tag{6}$$

and construct its quadratic majorizer in the next proposition, see Eq. (3.6) and Proposition 3.2 in [9].

Proposition 2 *Suppose that $p_0(s)$ is differentiable, increasing, and concave on $s \geq 0$ such that $p_0'(0) \in (0, \infty)$. Then for all $s, s^{[k]} \geq 0$ and $\mu > 0$ the function*

$$q_\mu(s; s^{[k]}) = p_\mu(s^{[k]}) + \frac{(s^2 - s^{[k]2})p_0'(s^{[k]})}{2(\mu + s^{[k]})} \tag{7}$$

majorizes $p_\mu(s)$ at the point $s^{[k]}$ and satisfies the condition $q_\mu(s; s^{[k]}) \geq p_\mu(s)$ for all $s \geq 0$ with equality when $s = s^{[k]}$.

The function $E_\mu^\lambda(\boldsymbol{W})$ is clearly majorized by the function

$$Q_\mu^\lambda(\boldsymbol{W}; \boldsymbol{W}^{[k]}) = \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 + \lambda\sum_{i=1}^m q_\mu(\|\boldsymbol{w}_i\|; \|\boldsymbol{w}_i^{[k]}\|) \tag{8}$$

and the following *majorization property* holds

$$Q_\mu^\lambda(\boldsymbol{W}; \boldsymbol{W}^{[k]}) \geq E_\mu^\lambda(\boldsymbol{W}) \text{ for all } \boldsymbol{W} \text{ with equality when } \boldsymbol{W} \in \{\boldsymbol{W} : \|\boldsymbol{w}_i\| = \|\boldsymbol{w}_i^{[k]}\|,\ i = 1, \ldots, m\}. \tag{9}$$

Consider any iteration map, which decreases the majorizer such that $Q_\mu^\lambda(\boldsymbol{W}^{[k+1]}; \boldsymbol{W}^{[k]}) < Q_\mu^\lambda(\boldsymbol{W}^{[k]}; \boldsymbol{W}^{[k]})$ holds. This leads, together with (9), to the *monotonicity property*

$$\begin{aligned} E_\mu^\lambda(\boldsymbol{W}^{[k+1]}) &\leq Q_\mu^\lambda(\boldsymbol{W}^{[k+1]}; \boldsymbol{W}^{[k]}) \\ &< Q_\mu^\lambda(\boldsymbol{W}^{[k]}; \boldsymbol{W}^{[k]}) = E_\mu^\lambda(\boldsymbol{W}^{[k]}). \end{aligned} \tag{10}$$

The majorizer (8) is quadratic and strictly convex, so it is easy to satisfy (10) by taking the next iterate

$$\boldsymbol{W}^{[k+1]} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{\Omega}_\mu^{[k]})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{11}$$

to minimize the majorizer. Here $\boldsymbol{\Omega}_\mu^{[k]}$ is a diagonal matrix with the ith diagonal element $p_0'(\|\boldsymbol{w}_i^{[k]}\|)/(\mu + \|\boldsymbol{w}_i^{[k]}\|)$.

The convergence properties of MM algorithms are well known in general [10]. In the present context, it suffices to note that the majorizer (8) is strictly convex and the iteration map (11) is continuous, so all fixed points of the algorithm are stationary points of $E_\mu^\lambda(\boldsymbol{W})$. A fixed point is guaranteed to exist, since $E_\mu^\lambda(\boldsymbol{W})$ is lower compact. In addition, if $\boldsymbol{W}_\mu$ minimizes $E_\mu^\lambda(\boldsymbol{W})$, then any limit point of $\{\boldsymbol{W}_\mu\}$ as $\mu \to 0$ minimizes $E_0^\lambda(\boldsymbol{W})$ [9]. In the particular case that we have $p_0(s) = s$ and the columns of $\boldsymbol{X}$ are linearly independent, $E_\mu^\lambda(\boldsymbol{W})$ is strictly convex and the algorithm converges to the unique minimizer of $E_\mu^\lambda(\boldsymbol{W})$. The use of strictly concave penalty functions is likely to introduce local minima, which may also be fixed points of the algorithm.

2.3. Active set strategy

The parameter λ is free and its value must be fixed by cross-validation or related methods in practical problems. Thus it is necessary to perform the minimization for several values $\lambda^{[0]} > \lambda^{[1]} > \cdots \geq 0$. Under the sparsity assumption $p_0'(0) > 0$, some rows of the minimizer $\boldsymbol{W}(\lambda^{[t]})$ are zero for a large enough $\lambda^{[t]}$. In the next step $\lambda^{[t+1]}$, it is tempting to use $\boldsymbol{W}(\lambda^{[t]})$ as a starting point and optimize only the rows that are likely to be nonzero in $\boldsymbol{W}(\lambda^{[t+1]})$. Next we present an active set strategy for this purpose, but before going further, a word of caution is in order. If (2) is nonconvex, the path of global minimizers may be noncontinuous as a function of λ. Then it is possible to stick to a locally optimal path for some time.

Taking the norm of both sides of (4), we find that $\|\boldsymbol{g}_i(\lambda)\|$ equals $\lambda p_0'(\|\boldsymbol{w}_i(\lambda)\|)$ for $i \in \mathcal{I}(\lambda)$. Combining this to (5), we note that $\|\boldsymbol{g}_i(\lambda)\|$ is at most $\lambda p_0'(\|\boldsymbol{w}_i(\lambda)\|)\ \forall i$. Thereby, it is natural to consider those curves $\|\boldsymbol{g}_i(\lambda)\|$ active that are close to their upper bounds $\lambda p_0'(\|\boldsymbol{w}_i(\lambda)\|)$. Particularly, we define

$$\mathcal{A}(\lambda^{[t+1]}) = \{i : \|\boldsymbol{g}_i(\lambda^{[t]})\| \geq (\lambda^{[t]} - \delta) p_0'(\|\boldsymbol{w}_i(\lambda^{[t]})\|)\}. \quad (12)$$

Note that $\mathcal{I}(\lambda^{[t]}) \subseteq \mathcal{A}(\lambda^{[t+1]})$ holds for all $\delta \geq 0$. A large value of δ has the effect of including extra indices to the set $\mathcal{A}(\lambda^{[t+1]})$. This makes the active set strategy more stable, because the desired property $\mathcal{I}(\lambda^{[t+1]}) \subseteq \mathcal{A}(\lambda^{[t+1]})$ becomes more probable. In general, a suitable value of δ depends on the step length $\lambda^{[t+1]} - \lambda^{[t]}$. The choice $\delta = 0.1\lambda^{[0]}$ is a safe rule of thumb for most purposes, and it is used in [7].

We may define $\boldsymbol{W}(\lambda^{[0]}) = \boldsymbol{0}$, $\boldsymbol{G}(\lambda^{[0]}) = -\boldsymbol{X}^T\boldsymbol{Y}$,

$$\lambda^{[0]} = \max_{1 \leq i \leq m} \{\|\boldsymbol{g}_i(\lambda^{[0]})\| / p_0'(\|\boldsymbol{w}_i(\lambda^{[0]})\|)\}, \quad (13)$$

$$\mathcal{A}(\lambda^{[0]}) = \{i : \|\boldsymbol{g}_i(\lambda^{[0]})\| / p_0'(\|\boldsymbol{w}_i(\lambda^{[0]})\|) = \lambda^{[0]}\}, \quad (14)$$

because $\boldsymbol{W}(\lambda) = \boldsymbol{0}$ holds for $\lambda \geq \lambda^{[0]}$ according to (4)–(5). In the subsequent steps, the MM algorithm (11) updates only the rows that belong to the active set (12). This means that the inverse operation in (11) concerns an $|\mathcal{A}(\lambda^{[t+1]})| \times |\mathcal{A}(\lambda^{[t+1]})|$ matrix in the step $\lambda^{[t+1]}$, so computational burden lightens.

3. EXPERIMENTS

Next we present numerical experiments, where a penalty function of the form $p_0(s) = c\log(1 + s/c)$ is used. As the parameter $c > 0$ decreases, the degree of concavity increases. The MM algorithm (11) is applied and the parameter μ is decreased exponentially from 10^{-5} to 10^{-10} in the process of iteration. The active set strategy (12) has the value $\delta = 0.1\lambda^{[0]}$.

The first experiment analyzes the tobacco leaf data set, which is also used, for example, in [7]. The data set has the dimensions $n = 25$, $m = 6$, and $q = 3$. All the variables are normalized to zero mean and unit variance before the analysis. The primary purpose of the experiment is illustration, but we also computed cross-validation errors (not shown). The best models are equally accurate, but the smaller values of c lead to more accurate highly penalized models. This is explained by the fact that the row norms tend to increase rapidly with a small c, which indicates weaker regularization, see Fig. 1. Note that the rows are considered active a bit before they become nonzero according to our strategy.

The second experiment uses simulated data with the dimensions $n = 50$, $m = 100$, and $q = 5$. Input data follow the distribution $\boldsymbol{x}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, where the covariances are $[\boldsymbol{\Sigma}_{\boldsymbol{X}}]_{ij} = 0.9^{|i-j|}$. Error data follow the distribution $\boldsymbol{e}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{E}})$, where $[\boldsymbol{\Sigma}_{\boldsymbol{E}}]_{ij} = 0.2^2 \cdot 0.6^{|i-j|}$. The matrix $\boldsymbol{W}$ has 20 randomly chosen nonzero rows. The jth element of the ith row (assumed nonzero) follows the distribution $w_{ij} \sim \mathrm{N}(0, \sigma_i^2)$, where $\sigma_i \sim \mathrm{Exp}(1)$. The matrix $\boldsymbol{W}$ is normalized $\boldsymbol{W} \leftarrow \boldsymbol{W}\mathrm{diag}(\boldsymbol{W}^T\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{W})^{-\frac{1}{2}}$ after sampling so that the scales of the response variables are comparable [7].

Fig. 2 shows the results, which are calculated from 500 replicates of simulation. The lowest prediction errors are more or less the same while the most accurate model has the value $c = 0.4$. Again, small values of c lead to better accuracy for large values of λ, but the situation is the opposite for small values of λ. Differences are bigger in terms of sparsity, since decreasing c strengthens parsimony. The number of correct nonzero rows is roughly the same in all cases, but the ratio of correctness is higher when c is small.

4. EXTENSIONS

The proposed approach can be extended in many ways. Firstly, the one-step Newton update [10] is applicable with a general differentiable loss function, which enriches the feasible model family. Secondly, nondifferentiable loss functions can be majorized in the same way as the penalty function is handled in the present article. This would enable various forms of robust regression and quantile regression. Thirdly, it is possible to have several penalty functions $p_0^{(i)}(s)$, which vary between the rows of the coefficient matrix. Fourthly, the proposed MM algorithm is directly applicable to any task, where blockwise sparsity [12], not just row sparsity, is required.

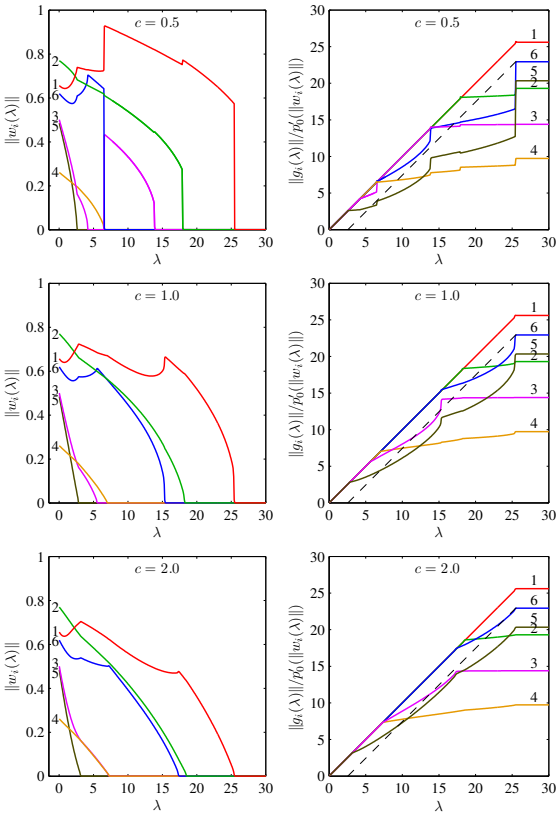

Fig. 1. Row norms (*left*) and illustrations of the active set strategy (*right*) for the tobacco leaf data set. The dashed line shows the decision boundary of the active set. A curve lying above this line indicates that the corresponding row is active.

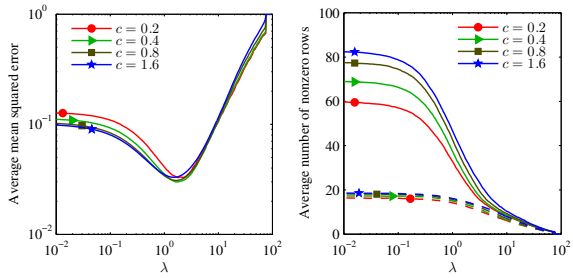

Fig. 2. Average mean squared error (*left*) and average number of nonzero rows (*right*) calculated from 500 replicates of simulated data. The dashed lines show the average proportion of correct nonzero rows.

5. ACKNOWLEDGMENTS

This work was supported in part by ComMIT graduate school and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. Nokia Foundation and Technological Foundation are also acknowledged. This publication only reflects the author's views.

6. REFERENCES

[1] J.A. Tropp, A.C. Gilbert, and M.J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, Mar. 2006.

[2] T. Similä and J. Tikka, "Common subset selection of inputs in multiresponse regression," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, July 2006, pp. 1908–1915.

[3] J.A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, Mar. 2006.

[4] S.F. Cotter, B.D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, July 2005.

[5] D. Malioutov, M. Çetin, and A.S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[6] B.A. Turlach, W.N. Venables, and S.J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, Aug. 2005.

[7] T. Similä and J. Tikka, "Input selection and shrinkage in multiresponse linear regression," Tech. Rep. A85, Helsinki Univ. Tech., Pub. in Computer and Information Science, Apr. 2006.

[8] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, Dec. 2001.

[9] D.R. Hunter and R. Li, "Variable selection using MM algorithms," *Ann. Statist.*, vol. 33, no. 4, pp. 1617–1642, Aug. 2005.

[10] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 57, no. 2, pp. 425–437, 1995.

[11] V.F. Demyanov and L.V. Vasilev, *Nondifferentiable Optimization*, Optimization Software, New York, 1985.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, Feb. 2006.