

Statistical Applications in Genetics and Molecular Biology

Volume 9, Issue 1

2010

Article 30

Lasso Logistic Regression, GSoft and the Cyclic Coordinate Descent Algorithm: Application to Gene Expression Data

Manuel Garcia-Magariños*

Anestis Antoniadis†

Ricardo Cao‡

Wenceslao González-Manteiga**

*Universidade de Santiago de Compostela, manuel.garcia.magarinos@usc.es

†Université Joseph Fourier, antonia@imag.fr

‡Universidade da Coruña, rcao@udc.es

**Universidade de Santiago de Compostela, wenceslao.gonzalez@usc.es

Lasso Logistic Regression, GSoft and the Cyclic Coordinate Descent Algorithm: Application to Gene Expression Data*

Manuel Garcia-Magariños, Anestis Antoniadis, Ricardo Cao, and Wenceslao González-Manteiga

Abstract

Statistical methods generating sparse models are of great value in the gene expression field, where the number of covariates (genes) under study moves about the thousands while the sample sizes seldom reach a hundred of individuals. For phenotype classification, we propose different lasso logistic regression approaches with specific penalizations for each gene. These methods are based on a generalized soft-threshold (GSoft) estimator. We also show that a recent algorithm for convex optimization, namely, the cyclic coordinate descent (CCD) algorithm, provides with a way to solve the optimization problem significantly faster than with other competing methods. Viewing GSoft as an iterative thresholding procedure allows us to get the asymptotic properties of the resulting estimates in a straightforward manner. Results are obtained for simulated and real data. The leukemia and colon datasets are commonly used to evaluate new statistical approaches, so they come in useful to establish comparisons with similar methods. Furthermore, biological meaning is extracted from the leukemia results, and compared with previous studies. In summary, the approaches presented here give rise to sparse, interpretable models that are competitive with similar methods developed in the field.

KEYWORDS: penalized regression, logistic regression, lasso, GSoft, CCD algorithm, optimization, gene expression

*The authors would like to thank the associate editor and two anonymous referees for their thorough remarks and suggestions. Research supported by Grant MTM2008-00166 (EU ERDF support included) for Garcia-Magariños and Cao. Garcia-Magariños is also supported by a FPU grant of the Spanish Ministry of Education and Science. Financial support from the IPA research network P5/24 of the Belgian government is gratefully acknowledged by Antoniadis. Research for González-Manteiga is supported by Grant MTM2008-03010. Garcia-Magariños also wants to thank Antoniadis for his hospitality during his three-month stay with him at the Université Joseph Fourier in Grenoble.

1 Introduction

Advent of high-dimensional data in several fields (genetics, text categorization, combinatorial chemistry, . . .) is an outstanding challenge for statistics. Gene expression data is the paradigm of high-dimensionality, usually comprising thousands (p) of covariates (genes) for only a few dozens (n) of samples (individuals). Feature selection in regression and classification is then fundamental to get interpretable, understandable models, which might be of use to the field. First approaches to this problem (Guyon and Elisseeff, 2003, Hall, 1999, Lee et al., 2003, Weston et al., 2003) were based on filtering to select a subset of covariates related with the outcome, usually a binary response. Nevertheless, common methods developed nowadays search for variable selection and classification carried out in the same step. Sparse models are needed to account for high-dimensionality (the $p \gg n$ problem) and strong correlations between covariates.

Penalized regression methods have received much attention over the past few years, as a proper way to get sparse models in those fields with large datasets. The different approaches deal with several issues (e.g. high correlations) in many ways. Lasso (Tibshirani, 1996) was originally proposed for linear regression models, and subsequently adapted to the logistic case (Roth, 2004, Shevade and Keerthi, 2003). Lasso applies a l_1 penalization that, as opposed to ridge regression (Hoerl and Kennard, 1970), gives rise to sparse models, ruling out the influence of most of the covariates on the response. Consistency properties of lasso for the linear regression case have been well studied (Knight and Fu, 2000, Lv and Fan, 2009, Meinshausen and Bühlmann, 2006, Zhang and Huang, 2008, Zhang, 2009). An evolution of lasso that allows for specific penalizations in the l_1 penalty (adaptive lasso) is developed in Zou (2006). Lasso has been also adapted to work with categorical variables (Antoniadis and Fan, 2001, Bakin, 1999, Meier et al., 2008, Yuan and Lin, 2006) and multinomial responses (Krishnapuram et al., 2005). Other penalized regression methods include bridge estimators (Frank and Friedman, 1993), which replace the l_1 penalization with l_q penalization, for $0 < q < 1$, and the elastic net (Zou and Hastie, 2005), that uses a linear combination of l_1 and l_2 penalties. The elastic net was proposed as a solution to some of the limitations of the lasso, namely the random selection in blocks of high correlated covariates. Consistency studies about bridge and elastic net can be found in Huang et al. (2006) and De Mol et al. (2009), respectively. Application of both approaches to high-dimensional genetic data is carried out in Liu et al. (2007). Optimization of the lasso log-likelihood function is

also an important subject (Lee et al., 2006, Schmidt et al., 2007), as a result of the non-differentiability problems of the l_1 penalty around zero.

In this paper, we adopt an adaptive lasso logistic regression approach based on the generalized soft-threshold estimator (GSoft) (Klinger, 2002). A theoretical connection between existence of solution in GSoft and convergence of the cyclic coordinate descent (CCD) algorithm (Zhang and Oles, 2001) is established, allowing to get the asymptotic properties of the resulting estimates. Different vectors $\mathbf{\Gamma}$ are used for the specific penalization of each covariate (gene) and some consistency results (Huang et al., 2008b) are shown for each one. Extensive comparisons with similar approaches are carried out using simulated and real microarray data.

The rest of this paper is organized as follows: a short introduction about the CCD algorithm, GSoft and some of its asymptotic properties is given in Section 2, together with the theoretical connection between both and the three different $\mathbf{\Gamma}$ choices for the specific penalizations. Some consistency results for each one are added. Results of simulated and real data are shown in Section 3. Simulations include approximations of the variance-covariance matrix for the estimated coefficients. Real data includes leukemia (Golub et al., 1999) and colon (Alon et al., 1999) datasets. Finally Section 4 is devoted to conclusions, and the Appendix contains the proof of Theorem 2.

2 Methods

Our aim is to learn a binary gene expression classifier $y_i = f(\mathbf{x}_i)$ from a set $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ of independent and identically distributed observations. For every observation i , the vector

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p \quad (1)$$

comprises gene expression measurements. The $n \times p$ design matrix is then $X = (\mathbf{x}_j, j \in \{1, \dots, p\})$ where the \mathbf{x}_j 's represent the expression measurements of gene j along the entire sample. The vector of binary responses

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad (2)$$

informs about membership (+1) or nonmembership (-1) of the sample to the category. The logistic regression model assumes that

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})}. \quad (3)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients. Adopting a generalized linear model framework, the associated linear predictor $\boldsymbol{\eta}$ is defined as

$$\boldsymbol{\eta} = X\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}'_1\boldsymbol{\beta} \\ \vdots \\ \mathbf{x}'_n\boldsymbol{\beta} \end{pmatrix} \text{ where } X = \begin{pmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \quad (4)$$

The decision of whether to assign the i observation to the category or not is usually accomplished by comparing the probability estimate with a threshold (e.g. 0.5). Consequently, minus the log-likelihood function is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^n \ln [1 + \exp(-y_i\mathbf{x}'_i\boldsymbol{\beta})] \quad (5)$$

The lasso like logistic estimator $\hat{\boldsymbol{\beta}}$ with specific penalizations for each covariate is then given by the minimizer of the function

$$L_1(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^p \gamma_j |\beta_j| \quad (6)$$

where λ is a common nonnegative penalty parameter and the vector $\boldsymbol{\Gamma} = (\gamma_1, \dots, \gamma_p)$, with nonnegative entries, penalizes each coefficient. The standard lasso regularization (Tibshirani, 1996) takes $\gamma_j = 1 \forall j$. Minimization of these objective functions makes use of their derivatives. We refer to the gradient of $L(\boldsymbol{\beta})$ as the score vector whose components are defined by:

$$s_j(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} \quad (7)$$

The negative Hessian with respect to the linear predictor $\boldsymbol{\eta}$ is defined as

$$H(\eta) = -\frac{\partial^2 L(\eta)}{\partial \eta \partial \eta'} \quad (8)$$

The basic requirement for the weights γ_j is that their value should be large enough to get $\hat{\beta}_j = 0$ if the true value β_j is zero, and small otherwise. Obtaining a sparse, interpretable model is of crucial importance in those areas where the number of variables is usually larger than the sample size ($p \gg n$ problem). The choice of the $\mathbf{\Gamma}$ vector is therefore essential to get an accurate estimator $\hat{\beta}$.

2.1 Cyclic coordinate descent (CCD) algorithm

The choice of a proper algorithm to solve the minimization of (6) is a main issue, as it needs to deal with the problem of non-differentiability of the absolute value function around zero. Furthermore, efficiency of the algorithm is fundamental, given the high dimension of the problems at hand.

Several algorithms have been developed to obtain the optimum for the objective function. In Goldstein and Osher (2008) a “Split-Bregman” method is applied to solve L_1 -regularized problems, while in Wright et al. (2008) an algorithmic framework is proposed for minimizing the sum of a smooth convex function with a nonsmooth nonconvex one. A similar algorithm is used in Kim et al. (2008) to obtain the solution for the SCAD (smoothly clipped absolute deviation) estimator in high dimensions. Two new approaches are developed in Schmidt et al., together with a comparative study. An efficient algorithm is carried out in Lee et al. (2006), using LARS (Efron et al., 2004) in each iteration. A local linear approximation (LLA) algorithm was recently proposed by Zou and Li (2008), while Wang and Leng (2007) developed a method of least squares approximation (LSA) for lasso estimation, making use of the LARS algorithm.

Finding the estimate of β is a convex optimization problem. The cyclic coordinate descent algorithm is based on the CLG algorithm of Zhang and Oles (2001). An exhaustive description of the algorithm is beyond the scope of this paper. Interested readers are referred to the detailed description in Genkin et al. (2005). The basis of all cyclic coordinate descent algorithms is to optimize with respect to only one variable at the time while all others are held constant. When this one-dimensional optimization problem has been solved, optimization is performed with respect to the next variable, and so on. When the procedure has gone through all variables it starts all over with the first one again, and the iterations proceed in this manner until some

pre-defined convergence criterion is met. The one-dimensional optimization problem is to find β_j^{new} , the value for the j -th parameter that maximizes the penalized log-likelihood assuming that all other β_j 's are held constant. In the end, the update equation for β_j becomes

$$\beta_j^{new} = \begin{cases} \beta_j - \Delta_j & \text{if } \Delta v_j < -\Delta_j \\ \beta_j + \Delta v_j & \text{if } -\Delta_j \leq \Delta v_j < \Delta_j \\ \beta_j + \Delta_j & \text{if } \Delta_j < \Delta v_j \end{cases} \quad (9)$$

where the interval $(\beta_j - \Delta_j, \beta_j + \Delta_j)$ is an iteratively adapted trust region for the suggested update Δv_j . The width of this interval is determined based on its previous value and the previous update for β_j . The suggested update is given by

$$\Delta v_j = -\frac{s_j(\boldsymbol{\beta}) - \lambda\gamma_j \text{sign}(\beta_j)}{Q(\beta_j, \Delta_j)} \quad (10)$$

The essential idea in CCD is $Q(\beta_j, \Delta_j)$ to be an upper bound on the second derivative of $L_1(\boldsymbol{\beta})$ in the interval around β_j :

$$\frac{\partial^2 L_1(\boldsymbol{\beta})}{\partial \beta_j^2} = \sum_{i=1}^n \frac{x_{ij}^2 \exp(-y_i \mathbf{x}_i' \boldsymbol{\beta})}{[1 + \exp(-y_i \mathbf{x}_i' \boldsymbol{\beta})]^2} \quad (11)$$

The function $Q(\beta_j, \Delta_j)$ is given by the expression:

$$Q(\beta_j, \Delta_j) = \sum_{i=1}^n x_{ij}^2 F(y_i \mathbf{x}_i' \boldsymbol{\beta}, \Delta_j x_{ij}) \quad (12)$$

with the function F being defined by

$$F(B, \delta) = \begin{cases} 0.25 & \text{if } |B| \leq |\delta| \\ [2 + \exp(|B| - |\delta|) + \exp(|\delta| - |B|)]^{-1} & \text{otherwise} \end{cases} \quad (13)$$

A proof of Q being an upper bound in the aforementioned interval is straightforward. Advantages of CCD can be summarized in efficiency of the algorithm, stability and ease of implementation. Efficiency is due to several factors: CCD works following a cycling procedure along the coefficients. From a certain iteration, CCD only visits the active set, reducing considerably its

computational demands. The CCD algorithm is implemented in the R package *glmnet*, which is used here in order to estimate the model parameters. This approach is explained in Friedman et al. (2008), where it is proved to be faster than its competitors. The *penalty.factor* argument in *glmnet* allows to implement the Γ vectors containing the specific weights for each covariate.

2.2 GSoft

The generalized soft–threshold estimator or GSoft (Klinger, 2002) is claimed to be a compromise between approximately linear estimators and variable selection strategies for high dimensional problems. Our interest in GSoft lies in the fact that once a solution β exists, a bunch of asymptotic properties can be derived. The next theorem from Klinger establishes necessary and sufficient conditions for the existence of such a solution.

Theorem 1. *The following set of conditions is necessary and sufficient for the existence of an optimum $\hat{\beta}$ of $L_1(\beta)$*

(a)

$$\begin{cases} |s_j(\beta)| \leq \lambda\gamma_j & \text{if } \beta_j = 0 \\ s_j(\beta) = \lambda\gamma_j & \text{if } \beta_j > 0 \\ s_j(\beta) = -\lambda\gamma_j & \text{if } \beta_j < 0 \end{cases} \quad (14)$$

(b)

$$X'_\lambda H(\eta) X_\lambda \text{ is positive definite,} \quad (15)$$

where X_λ retains only those columns (covariates) \mathbf{x}_j of X fulfilling $|s_j(\beta)| = \lambda\gamma_j$, that is, $X_\lambda = (\mathbf{x}_j, |s_j(\beta)| = \lambda\gamma_j)$.

2.2.1 Approximation of the covariance matrix for the estimated coefficients.

Approximations to the variance–covariance matrix of $\hat{\beta}$ have to deal with the non–differentiability problem of the penalization term around $|\beta_j| = 0$. This problem is solved in Klinger (2002) by taking a differentiable approximation to the absolute value function, obtained by smoothing it around zero.

Such approximation is constructed from the well–known sandwich form developed in Huber (1967)

$$V_\delta(\hat{\beta}) = \left\{ H(\hat{\beta}) + \lambda \Gamma G(\hat{\beta}, \delta) \right\}^{-1} \text{Var} \left\{ s(\hat{\beta}) \right\} \left\{ H(\hat{\beta}) + \lambda \Gamma G(\hat{\beta}, \delta) \right\}^{-1} \quad (16)$$

GSoft solves the problems of regularity for the true zero coefficients by developing the following estimator

$$\hat{V}(\hat{\beta}_j) = \left\{ H(\hat{\beta}) + \lambda \Gamma G^* \left(\hat{\beta}, \hat{\sigma} \right) \right\}^{-1} \hat{F}(\hat{\beta}) \left\{ H(\hat{\beta}) + \lambda \Gamma G^* \left(\hat{\beta}, \hat{\sigma} \right) \right\}^{-1} \quad (17)$$

where

$$\begin{aligned} G^* \left(\hat{\beta}, \sigma \right) &= \text{diag} \left\{ \frac{2}{\sigma_1} \phi(\hat{\beta}_1/\sigma_1), \dots, \frac{2}{\sigma_p} \phi(\hat{\beta}_p/\sigma_p) \right\} \\ (\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2) &= \text{diag} \left[H(\hat{\beta})^{-1} \hat{F}(\hat{\beta}) H(\hat{\beta})^{-1} \right] \end{aligned} \quad (18)$$

ϕ is the density function of the standard normal distribution

Anyhow, the main point to get a well established approach to the real variance covariance matrix is to use an accurate estimator \hat{F} of the Fisher matrix given by

$$F(\eta) = -E \left\{ \frac{\partial L(\eta)}{\partial \eta \partial \eta'} \right\} \quad (19)$$

Therefore the theoretical framework to carry out this approximation has been established since Huber, and Klinger contributed with the adjustment to sparse models. Trying to improve this, firstly we made use of the approach carried out in Antoniadis et al. (2009). Nevertheless, after some tests we realized that such a choice really underestimates the true variance-covariance values. Our solution consists of rescaling this matrix multiplying it by a factor equal to the number of variables, p , in the model. So

$$\hat{F}(\hat{\beta}) = I(\hat{\beta}) = \frac{p[\partial^2 L(\hat{\beta})/\partial \beta_i \partial \beta_j]}{n} \quad (20)$$

Goodness of fit for this estimator is discussed in Section 3.

2.3 Connection GSoft - CCD algorithm

The main aim of this article is to establish a theoretical connection between the convergence of the CCD algorithm and the existence of an optimum for the objective function with GSoft. This theoretical connection is established by the next theorem (its proof is given in the Appendix).

Theorem 2. *The following two statements are equivalent:*

- (1) *The CCD algorithm for the lasso case converges.*
- (2) *An optimum for the objective function under the terms of the theorem in Klinger (2002) exists.*

Both GSoft and CCD are optimization methods for penalized likelihood estimation. It is however a well established fact (De Leeuw, 1994, Tseng, 2001, Tseng and Yun, 2009, Friedman et al., 2008) that CCD, as an iterative optimization procedure, is generally convergent and significantly faster than other competing methods. However, deriving asymptotic properties of the resulting estimates is not an easy task. Viewing GSoft as an iterative thresholding procedure allows to get the asymptotic properties of the resulting estimates in a straightforward manner, using either Klinger's results or Meinshausen and Bühlmann's results (Meinshausen and Bühlmann, 2006).

2.3.1 Choice of Γ

As we mentioned above, we use a global threshold λ together with a vector of specific thresholds $\Gamma = (\gamma_1, \dots, \gamma_p)$ corresponding to the coefficients β_1, \dots, β_p of each variable in the model. In this study, we will evaluate the performance of three different choices for the Γ vector:

1. $\gamma_j = \sqrt{\text{var}(\mathbf{x}_j)}$. This is one of the choices carried out in Klinger (2002). As a consequence, we will refer to it as γ -Klinger. Adjusting the thresholds like this is equivalent to standardization.
2. $\gamma_j = 1/|\beta_j^{\text{ridge}}|$. Ridge logistic regression was performed on data with a small global threshold λ_0 , obtaining coefficients $\beta_j^{\text{ridge}} \neq 0, \forall j = 1, \dots, p$. This choice is related to penalize according to the importance of the variable in ridge, and it is based on a special case of the adaptive lasso (Zou, 2006). This choice will be denoted as γ -ridge.
3. $\gamma_j = 1/|\beta_j^{\text{lasso}}|$. Lasso logistic regression was performed on data with a small global threshold λ_0 and without using specific thresholds γ . Obviously, some coefficients β_j^{lasso} will take zero values. In this case, these variables are excluded from the final model, which is equivalent to take $\gamma_j = \infty$. It will be called γ -lasso.

2.3.2 Consistency results

Variable selection consistency results in lasso can be found in the recent related literature. Oracle property (Fan and Li, 2001) for the adaptive lasso in linear

regression models is proved in Huang et al. (2008a). Consistency results shown here are based on the subsequent adaptation of these results to the logistic case, carried out in Huang et al. (2008b), for the γ -lasso, there called iterated lasso.

Under bound conditions for the true coefficients β and the covariates \mathbf{x}_j , and imposing restrictions on the number of nonzero coefficients and the value of λ , it is proved in Huang et al. (2008b) that

$$P(\text{sign}(\hat{\beta}) = \text{sign}(\beta)) \rightarrow 1 \quad (21)$$

where the sign function is now taken in a slightly different way than in (10): $\text{sign}(\theta_1, \dots, \theta_p) = (\text{sign}(\theta_1), \dots, \text{sign}(\theta_p))$ and

$$\text{sign}(t) = \begin{cases} -1 & \text{if } t < 0 \\ 0 & \text{if } t = 0 \\ 1 & \text{if } t > 0 \end{cases} \quad (22)$$

Asymptotic convergence is also proved

$$\mathbf{T}_n(\hat{\beta}_{B_0} - \beta_{B_0}) \longrightarrow_D N(0, 1) \quad (23)$$

where $B_0 = \{j : \beta_j \neq 0\}$ is the set of indices with true nonzero coefficients and \mathbf{T}_n is a vector depending on the sample size n .

These two results, (21) and (23), together mean that the γ -lasso choice has the asymptotic oracle property. The proof can be found in Huang et al. (2008b), which also refers to the proof for the linear case in Huang et al. (2008a).

This consistency result (oracle property) for the iterated lasso can be also obtained for the γ -ridge choice of specific penalizations, applying the necessary changes in the corresponding assumptions. Many of these assumptions remain exactly the same, as they just impose conditions on the true coefficient values or the data. The crux is the so-called (Huang et al., 2008b) r_n -consistency or zero-consistency of the initial estimator (ridge regression), which means the initial estimator shrinks the zero coefficients towards zero at a certain rate. In other words, in order to use the results by Huang et al., an initial estimator that is zero-consistent is needed. This means that the estimators of zero coefficients converge to zero in probability and the estimators of non-zero coefficients do not converge to zero. Under the conditions for the design matrix X and λ small enough, the L_2 consistency of the ridge logistic estimator follows from the asymptotic results for L_2 maximum penalized

likelihood estimation (see Eggermont and LaRiccia, 2009). This L_2 consistency, although does not give rise to a sparse model, it is enough to weight the nonzero coefficients with bounded weights, while the zero ones will have weights tending asymptotically to infinity. This feature is the only one really needed in the proofs of Huang et al. (2008a) and Huang et al. (2008b) for the initial estimators to satisfy the asymptotic oracle property.

When γ -Klinger penalizations are selected, this is equivalent to standardization, as proved in Klinger. Therefore, only usual consistency lasso results (Huang et al., 2008b, Meier et al., 2008) can be proved in this case, and oracle property does not hold. An upper bound for the number of estimated nonzero coefficients in lasso is given in Huang et al. (2008b). There, it is proved that the dimension of the model selected by lasso is directly proportional to n^2 and inversely proportional to λ .

3 Results

3.1 Simulated data

Three scenarios with binary response have been simulated according to one of the examples in Hunter and Li (2005). In all of them, the response follows the model:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x}'\boldsymbol{\beta})} \quad (24)$$

This example has been adapted to two specific scenarios carried out in Wang and Leng (2007) (Simulation 1) and Zou and Li (2008) (Simulation 2), with the aim of comparing our results with those obtained there. Furthermore, a third set of simulations have been developed following Hunter and Li. In addition to this, a fourth high dimensional simulation study has been carried out, trying to emulate the main characteristics of a gene expression study ($p \gg n$, correlation structure depending on distance, small proportion of nonzero coefficients) to evaluate the ability of our approaches under these controlled conditions. We have also used the scenario in Zou and Li to obtain the results of approximation of variance as explained in the last section.

3.1.1 Simulation 1

The main aim is to compare our results with those obtained with the least squares approximation (LSA) estimator. Comparisons with the results of the

Sample size	Estimation method	MS		CM	
		mean	(SE)	mean	(SE)
200	LLR γ -Klinger	3.266	(0.025)	0.762	(0.019)
	LLR γ -ridge	2.896	(0.025)	0.812	(0.017)
	LLR γ -lasso	2.96	(0.028)	0.798	(0.018)
	LSA	3.178	(0.026)	0.798	(0.018)
	PH	3.272	(0.033)	0.716	(0.020)
400	LLR γ -Klinger	3.046	(0.011)	0.956	(0.009)
	LLR γ -ridge	2.964	(0.021)	0.860	(0.016)
	LLR γ -lasso	2.982	(0.022)	0.902	(0.013)
	LSA	3.130	(0.018)	0.888	(0.014)
	PH	3.092	(0.023)	0.846	(0.016)

Table 1: *True model detection results. Comparison between Model 1 and those in Wang and Leng is established in the same terms as there.*

Park and Hastie (PH) algorithm (see Park and Hastie, 2006) shown in Wang and Leng, are also established. Model 1 is 9-dimensional with coefficients $\beta = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)'$. The components of \mathbf{x}_i are standard normal and the correlation between each pair of variables \mathbf{x}_{j_1} and \mathbf{x}_{j_2} is fixed to $0.5^{|j_1-j_2|}$. The sizes of the training samples are $n = 200$ and $n = 400$, and 500 simulation replications have been obtained each time. The BIC criterion is used to obtain the best solution for LSA and PH, while for the choice of λ in our models, we follow a slightly different approach. As choosing the λ giving rise to the smallest error rate (ER) does not necessarily produce a sparse model, we take the largest λ having an error rate smaller than $\min_{\lambda} ER + 2 \cdot \text{sd}(ER)$. The results are shown in Table 1. From now on, lasso logistic regression will be referred with the abbreviation LLR.

The different estimators are compared in terms of model size (MS) and percentage of correct models identified (CM). Unlike Wang and Leng, here we will not use the relative model error as a comparative measure, since it puts too much weight to the model error without penalty. Besides, in problems involving large amounts of noise, detection of the variables associated with the response is much more important than precise estimation of the true coefficients. Results obtained with Model 1 are slightly better than those in Wang and Leng. Comparisons between the different choices for the Γ vector are favourable to γ -ridge and γ -lasso, as the γ -Klinger seems to be more imprecise than those two regarding detection of the correct model. This imprecision grows when sample size decreases, until reaching the standard of LSA and PH.

3.1.2 Simulation 2

Comparisons with the one-step sparse estimates developed in Zou and Li are carried out, along with the SCAD estimator and other variable selection models used there. Model 2 is 12-dimensional with $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)'$, while \mathbf{x} is obtained as in Simulation 1, but with one important difference: variables with even index are translated to binary according to their sign. The size of the training sample is $n = 200$, and 1000 replicated datasets were obtained. Choice of the optimal λ in this case is carried out in a similar way to Simulation 1, but taking the largest λ having an error rate smaller than $\min_{\lambda} ER + 0.2 \cdot sd(ER)$. The change in the factor multiplying the standard deviations allowed from Simulation 1 (from 2 to 0.2) is due to the changes in the model produced by the transformation of continuous variables into binary variables. The choice of this tuning parameter should always depend on the amount of sparsity demanded, as well as on the nature of the data (categorical/continuous, range), besides advice of experts in the field. The results are shown in Table 2.

The notation in Zou and Li is also used here: columns ‘C’ and ‘IC’ measure the average number of nonzero coefficients correctly estimated to be nonzero and the average number of zero coefficients incorrectly estimated to be nonzero, respectively. “Under-fit” and “Over-fit” show the proportion of models excluding any nonzero coefficients and including any zero coefficients throughout the 1000 replications, respectively. “Correct-fit” shows the proportion of correct models obtained.

Our methods show a worse behaviour than those in Zou and Li. After some tests (results not shown here) we realized that the reason was

Method	C	IC	Proportion of		
			Under-fit	Correct-fit	Over-fit
LLR γ -Klinger	2.84	1.68	0.16	0.14	0.70
LLR γ -ridge	2.77	0.82	0.22	0.40	0.37
LLR γ -lasso	2.71	0.71	0.29	0.40	0.31
one-step SCAD	2.95	0.82	0.051	0.565	0.384
one-step LOG	2.97	0.61	0.029	0.518	0.453
one-step $L_{0.01}$	2.97	0.61	0.028	0.516	0.456
SCAD	2.92	0.51	0.076	0.706	0.218
P-SCAD	2.92	0.5	0.079	0.707	0.214
AIC	2.98	1.56	0.021	0.216	0.763
BIC	2.95	0.22	0.053	0.800	0.147

Table 2: True model detection results. Comparison between Model 2 and those in Zou and Li is established in the same terms as there.

Method	$\rho = 0.25$		$\rho = 0.75$	
	C	I	C	I
LLR γ -Klinger	5.96	0.034	5.562	0.326
LLR γ -ridge	5.9	0.166	5.912	0.778
LLR γ -lasso	5.9	0.176	5.916	0.76
New	5.922	0	5.534	0.222
LQA	5.728	0	4.97	0.090
BIC	5.86	0	5.796	0.304
AIC	4.93	0	4.86	0.092

Table 3: *True model detection results. Comparison between Model 2 and those in Hunter and Li is established in the same terms as there.*

that they suffer a lot from the presence of binary variables. This is not a major concern, since our aim was to apply these methods to gene expression data, where all the variables are continuous. Therefore, in order to test them in a continuous environment, conditions in Hunter and Li were replicated. These conditions are the same as in Simulation 1 but the correlation between variables is now fixed to $\rho = 0.25$ and $\rho = 0.75$. The sample size was also fixed to $n = 200$. The results are shown in Table 3.

The optimal λ is chosen as in Simulation 1. The columns “C” and “I” measure the average number of coefficients correctly and incorrectly set to zero, respectively. Comparisons are made with a new algorithm proposed in Hunter and Li, a local quadratic approximation (LQA) algorithm developed in Fan and Li and the best subset variable selection using BIC and AIC scores. Competitive results are obtained with respect to the procedure in Hunter and Li. The best variable selection is obtained using BIC. The results obtained with the γ -Klinger are similar to the ones with γ -ridge and γ -lasso.

3.1.3 Simulation 3

We also performed a new simulation to emulate the conditions of a gene expression study. As a consequence, the same conditions as in Simulation 1 were recreated here, but now for a 1000-dimensional model, where just 10 coefficients (chosen at random in each simulation replication) take the nonzero values $(3, 1.5, 7, 4, 2.2, 1, 10, 2, 5, 3)'$. Apart from the training sample sizes used in Simulation 1 ($n = 200$ and $n = 400$), a new one ($n = 100$) was added. Since the presence of so much noise covariates and the correlation structure gives rise to high variability for the error rate measures, we took the largest λ having an error rate smaller than $\min_{\lambda} ER + 0.2 \cdot \text{sd}(ER)$. The results are shown in Table 4.

Sample size	Method	MS		PC		PN		Ratio
		mean	(SE)	mean	(SE)	mean	(SE)	
100	LLR γ -Klinger	8.95	(3.02)	0.373	(0.093)	0.005	(0.003)	0.72
	LLR γ -ridge	7.93	(6.28)	0.331	(0.120)	0.005	(0.005)	0.68
	LLR γ -lasso	5.72	(3.94)	0.315	(0.096)	0.003	(0.003)	1.22
200	LLR γ -Klinger	17.76	(3.35)	0.674	(0.101)	0.011	(0.003)	0.61
	LLR γ -ridge	15.15	(9.62)	0.613	(0.118)	0.009	(0.009)	0.68
	LLR γ -lasso	12.02	(6.67)	0.636	(0.108)	0.006	(0.006)	1.13
400	LLR γ -Klinger	25.53	(4.22)	0.883	(0.074)	0.016	(0.004)	0.52
	LLR γ -ridge	20.59	(11.78)	0.833	(0.091)	0.012	(0.011)	0.68
	LLR γ -lasso	13.77	(6.48)	0.852	(0.085)	0.005	0.006	1.62

Table 4: True model detection results. Comparison is established in terms of detection of noise and true nonzero coefficients for a high dimensional simulation study.

Comparisons were established in terms of model size (MS), understood as the number of nonzero coefficients estimated, proportion of nonzero coefficients correctly estimated as nonzero (PC), and proportion of noise in the models (PN), understood as the proportion of zero coefficients estimated nonzero. The right column refers to the ratio of correct nonzero coefficients to incorrect nonzero coefficients for the model selected.

The results show a clear trend towards increasing model sizes when sample size grows. It is essential to recall the fact that model sizes in Table 4 strongly depend on the penalization (λ), and therefore on the factor multiplying the standard deviation selected above (0.2). The larger this factor, the smaller the model size. Considering similar model sizes, γ -ridge and, above all, γ -lasso, show higher proportions of nonzero coefficients correctly identified (PC). Many of the incorrect nonzero coefficients in the estimated models correspond with covariates correlated with the true nonzero ones (results not shown here). Figure 1 shows the changes in model size (MS), number of true nonzero (NC) and true zero (NN) detected coefficients, and Ratio, all of them averaged, when the λ penalization (or the number of standard deviations) varies, for the γ -lasso approach and different sample sizes. The PC and PN values which appear in Table 4 are, respectively, the NC and NN values expressed as proportions.

3.1.4 Approximation of variance

Covariance matrix estimation for the estimated coefficients has been obtained according to the previously explained approach. Model 2 has been used, without the translation to binary (for simplicity). In Figure 2, the behaviour of variance estimation for $\beta_1 = 3$, $\beta_2 = 1.5$ and $\beta_3 = 0$, respectively, is shown in comparison with the true variance, as a function of λ . The estimation,

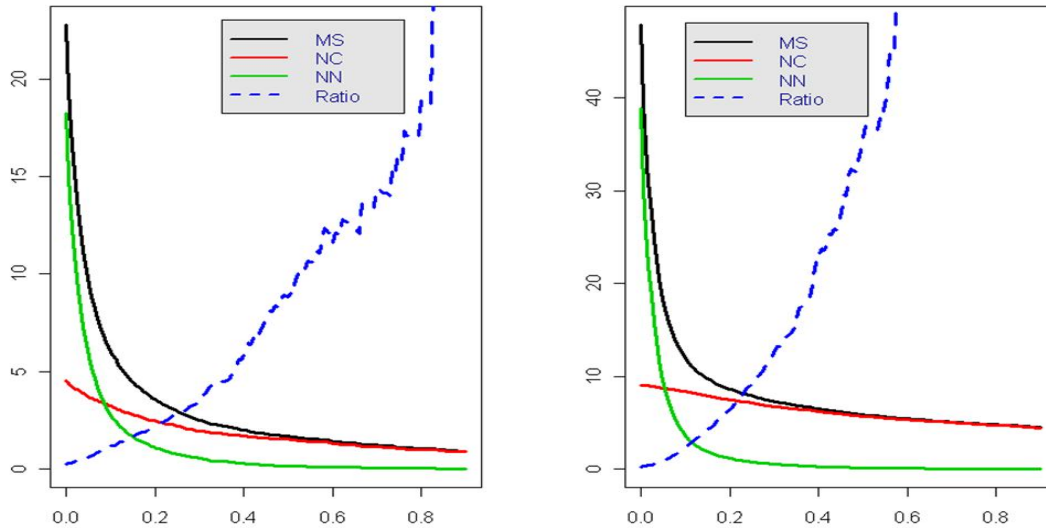


Figure 1: Changes in model detection for the γ -lasso approach and different sample sizes: $n = 100$ (left) and $n = 400$ (right), when the penalization term λ varies. As indicated in the legends, the black line denotes the model size (MS), as the average number of detected nonzero coefficients. This MS is the result of adding the number of true nonzero (NC , red line) and the number of true zero (NN , green line) detected as nonzero (both averaged). The Ratio (blue line) is the result of dividing NC by NN .

obtained as the median along 1000 replications, fits almost perfectly to the variance except for small deviations when λ is close to zero (maximum likelihood estimator), as the true variance increases enormously.

3.2 Real data

The leukemia dataset (Golub et al., 1999) has been used on countless occasions through the gene expression literature. It comprises gene expression data for 72 bone marrow and peripheral blood samples (47 cases of acute lymphoblastic leukemia (ALL) and 25 cases of acute myeloid leukemia (AML)) in 7129 genes. Initially (Golub et al.) the total sample was divided into a training sample (38 bone marrow samples) and a test sample (34 bone marrow and peripheral blood samples).

The colon dataset was analyzed initially by Alon et al. (1999). As leukemia, it is another commonly used dataset in genomic studies. A number of 62 observations (40 tumors and 22 controls) were measured in 2000 human genes. Absolute measurements from Affymetrix high-density oligonucleotide

arrays were taken for each sample in each gene in both datasets. Here, we have worked with data in two different ways. On one side, we have carried out preprocessing steps (P) following Subsection 3.1.2 of Dudoit et al. (2002), (i) thresholding of the measurements, (ii) filtering of genes, (iii) base 10 logarithmic transformation. On the other hand, we have also tried our models over the raw data (RD). With preprocessing, leukemia and colon datasets reduce their dimensionality to 3571 and 1225 genes, respectively.

As a result of combining these two ways to deal with data with the three different choices for γ , we have six different procedures. Table 5 shows the results for the leukemia dataset. To obtain accurate and precise measures for the error and its standard deviation, we randomly split 50 times the set of 72 samples into a training set of 38 samples and a test set of 34 samples. We also record the number of genes with non-zero coefficient for the optimal lambda, in terms of cross-validation (CV) error.

Table 6 shows the results for the colon dataset. The 62-sample has been randomly splitted 50 times into a training subsample of 50 observations and a test subsample of 12 observations. Different ways of splitting for the two real datasets are explained from their different nature. An unbalanced train-test data split is needed in the colon dataset to detect the existing associations (see Krishnapuram et al., 2005).

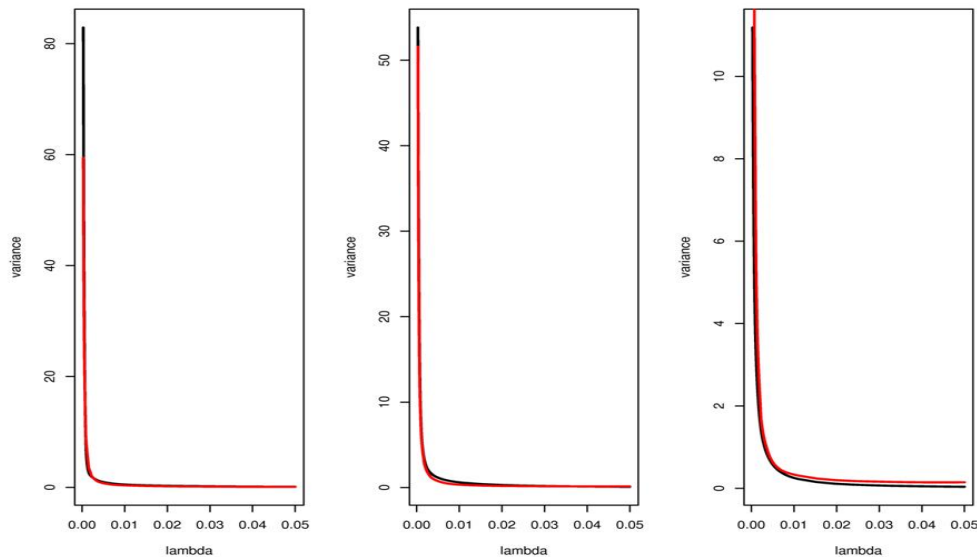


Figure 2: Variance estimation (in red) for the estimated values of β_1 , β_2 and β_3 in Simulation 2, according to the estimator (17) with \hat{F} taken as in (20). True variance (in black) was approximated by means of recursive simulation-estimation. The variance is displayed as a function of the penalty parameter λ .

Leukemia	Test error	SD	Genes
RD- γ Klinger	0.062	(0.044)	67 (of 7129)
RD- γ Zou	0.064	(0.039)	11 (of 7129)
RD- γ Lasso	0.102	(0.055)	6 (of 7129)
P- γ Klinger	0.079	(0.032)	16 (of 3571)
P- γ Zou	0.067	(0.030)	5 (of 3571)
P- γ Lasso	0.064	(0.028)	5 (of 3571)

Table 5: *Test error and sparsity results for the leukemia dataset.*

Colon	Test error	SD	Genes
RD- γ Klinger	0.195	(0.130)	10 (of 2000)
RD- γ Zou	0.147	(0.116)	17 (of 2000)
RD- γ Lasso	0.200	(0.128)	9 (of 2000)
P- γ Klinger	0.152	(0.096)	11 (of 1225)
P- γ Zou	0.182	(0.111)	15 (of 1225)
P- γ Lasso	0.215	(0.133)	10 (of 1225)

Table 6: *Test error and sparsity results for the colon dataset.*

Leukemia and colon datasets have been often used in the literature to test the performance of different methods. Nevertheless, it is difficult to find a fair comparison between methods, since each author uses a different way to obtain an error measure. Some of them only focus on a leave-one-out cross-validation rate (too optimistic); others center on the same data subdivision carried out by Golub et al. Finally, the fairest way to know the real performance of each method is to randomly split the total sample N times into two disjoint samples, training and test. Table 7 compares our best test error results with those from other methods, using similar training-testing data divisions.

Comparisons with the following methods have been established. In Boulesteix et al. (2003), a CART-based method is developed to discover the emerging patterns within the set of variables. BagBoosting (Dettling, 2004) is a combination of bagging and boosting, two ensemble learning algorithms, applied to stumps, decision trees with only one split and two terminal nodes. Different algorithms are presented in Dettling and Bühlmann (2004). *Pelora* is a penalized logistic regression method. *Forsela* is similar to *Pelora*, but making a search of single genes instead of groups, *Wilma* (Dettling and Bühlmann, 2002) shares some characteristics with *Pelora*, but suffers from a few limitations (Dettling and Bühlmann, 2004). Nguyen and Rocke (2002) use dimension reduction through partial least squares (PLS) and principal component analysis (PCA), classifying with discriminant analysis. Our error results are only slightly worse than the others for the leukemia dataset, and among

the best for colon. In any case, all the error rates are quite similar. Many of the methods we compare with stand out for grouping genes (Boulesteix et al., Nguyen and Roche, *Pelora* and *Wilma* in Dettling and Bühlmann, 2004) in one way or another. Gene preselection is carried out by means of preexisting methods in Boulesteix et al. and Dettling (2004). Our logistic lasso methods neither makes use of grouping or gene preselection nor it is necessary to select a lot of different parameters, as in Boulesteix et al., apart from the penalty λ . Moreover, its sparsity (see Tables 5 and 6) and interpretability are merits not fulfilled by the other methods.

Gene expression data is seen as the paradigm of the case $n \ll p$, as Affymetrix or oligonucleotide arrays map large parts of the human genome while only tens or hundreds of individuals are sampled. This situation makes most of traditional statistical methods inapplicable, so new variable selection approaches had to be developed to deal with this *curse of dimensionality* problem. Lasso selects a group of $p' \leq n$ genes with high importance in the classification of samples, and assigns a zero coefficient to the remaining genes. The use of the CCD algorithm to solve the optimization problem is highly desirable, as it provides with the global solution of GSoft in the fastest and most efficient way.

Dataset	Method	Test error
Leukemia	Our best	0.062
	CART-Fisher (*)	0.024–0.050
	BagBoosting (**)	0.0408
	<i>Pelora</i> (**)	0.0569
	<i>Wilma</i> (**)	0.0262
	<i>Forsela</i> (**)	0.0415
	PLS (***)	0.033–0.047
	PCA (***)	0.039–0.108
	Colon	Our best
Colon	CART-Fisher (*)	0.128–0.234
	BagBoosting (**)	0.161
	<i>Pelora</i> (**)	0.1571
	<i>Wilma</i> (**)	0.1648
	<i>Forsela</i> (**)	0.1381

Table 7: Test error rates obtained using different methods from the related literature for the leukemia and colon datasets. (*) In each random split, 10 observations in the test set. (**) In each random split, 2/3 of the data in the training set, 1/3 of the data in the test set. (***) In each random split, 1/2 of the data in the training set, 1/2 of the data in the test set.

From a more biological view, we have also studied which genes are more related with the ALL/AML status in leukemia. Observations of the genes with nonzero coefficients for each model have been carried out. As expected, some recurrences have been found with the six different procedures. Table 8 shows those genes showing up more frequently.

The fact that some genes are discovered with some procedures and not with others can be explained from the correlations between them. These correlations arise as a result of co-inheritance of nearby genes throughout generations. For instance, gene M19507 takes a nonzero coefficient with all but two of the procedures, and gene M92287 takes nonzero coefficients only with these two procedures. If we take a careful look to the correlation between them, we detect it as unusually high. A correlation study between all the genes with nonzero coefficient in any of the models has been carried out. With the aim of knowing the real meaning of each correlation value, a simple permutation testing was applied: a significance value for each correlation was obtained as the proportion of values, in a set of 10000 random correlations taken between pairs of genes from the entire dataset, which are higher than the correlation under study. In this way, significance of the correlation M19507–M92287 is 0.0558; the one between M84526–Y00787 is 0.048, which explains why they are partly complementary. Significances of correlations between gene Y00787 and the last eight genes in Table 8 are also very low, as they are detected specifically in those two models where Y00787 is not. In a similar way, pairwise correlations in this 8-gene group are often high. Complementarity in the detection by the different models emphasizes one of the main problems of lasso selection, also marked in Zou and Hastie: when there is a group of significant variables with high pairwise correlation lasso selects only one, and does not care which one.

A bunch of articles can be found in the gene expression literature looking for genes associated with the ALL/AML status. It is expected that there exists some kind of intersection between the sets of genes given by the different studies. The first five genes in the relation of Table 8 (M27891, M19507, M84526, Y00787 and M92287) are also discovered in Lee et al. (2003), being M27891 the one showing the strongest association with disease, as happens here. Three of the four genes pointed out in Guyon et al. (2002) (U82759, HG1612 and X95735) are also discovered here. On the other hand, coincidences with the list given in Thomas et al. (2001) are more limited.

Genes	RD- γ Klinger	RD- γ Zou	RD- γ Lasso	P- γ Klinger	P- γ Zou	P- γ Lasso
M27891	X	X	X	X	X	X
M19507	X	X	X		X	
M84526	X			X	X	X
Y00787		X	X		X	X
M92287				X		X
U05255		X	X			
M17733		X	X			
M63138	X		X			
M96326	X	X				
L07633	X			X		
U82759	X			X		
HG1612	X			X		
M13690	X			X		
M23197	X			X		
X95735	X			X		
Y07604	X			X		
X85116	X			X		

Table 8: Genes with nonzero estimated coefficients with the different procedures for the leukemia dataset. Only the seventeen genes detected in more than one model are shown.

4 Conclusion

We study lasso logistic regression by means of a generalized soft-threshold (GSoft) estimator. An equivalence between the existence of a solution in GSoft and convergence of the CCD algorithm to the same solution is given. An approximation of the covariance matrix for the estimated coefficients, $\hat{\beta}$, based on the GSoft approach produces very accurate results. The CCD algorithm is fast, stable and efficient, and allows different kinds of implementations. Efficiency of the optimization algorithm is a main issue nowadays, as the datasets used in many fields (text categorization, image processing,...) have extraordinary high dimensions.

We tried different options for the vector Γ of specific penalizations in GSoft. Some of them are based in the variability shown by each covariate, while others depend on previous application of penalized regression approaches to the data. Their consistency properties follow from previous results in the recent literature.

Finally, we applied these methods to simulated data and gene expression data. The same simulations carried out in other studies were used here, in order to provide honest and fair comparisons. A high dimensional simulation study was also performed, in order to emulate the characteristics of

gene expression studies. The best model is selected from the error rate results, correcting using the standard error. The factor multiplying the standard error should be guided by the authors' prior knowledge about the problem. Common gene expression datasets, like leukemia or colon, allow to know the ability of these methods to detect genes related with the disease or trait under study. The penalized regression approaches performed in this work give rise to sparse models, where only a very small percentage of covariates (genes) have a positive weight in classification/prediction.

Appendix. Proof of theorem 2

The log-likelihood functions in logistic regression and in lasso logistic regression with specific penalizations are given in (5) and (6), respectively. The first partial derivatives or score functions are:

$$s_j(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{-y_i x_{ij}}{1 + \exp(y_i \mathbf{x}'_i \boldsymbol{\beta})} \quad (25)$$

The definition of the Δv_j for the lasso case in Genkin et al., applied on a penalized regression problem with specific penalizations for each variable, is given in (10). For ease of notation, we will use here S instead of $\text{sign}(\beta_j)$. We will base the entire proof in the steps and the notations used in Figures 4 and 5 in Genkin et al. Many of the terms used there will be repeated here. To clarify the notation, we will use β_j for the true value of the coefficients and $\beta_j^{(I)}$ for the value of the j th coefficient in the I th iteration of the CCD algorithm.

We begin by proving the equivalence for the case $\beta_j = 0$, and then we will move to the more general case of $\beta_j > 0$ (similar proof for $\beta_j < 0$).

Case $\beta_j = 0$

(1) \Rightarrow (2)

We assume that the CCD algorithm, as explained in Genkin et al., converges. Therefore, from a certain iteration, I , we have $\beta_j^{(I)} = 0$ and $\Delta v_j^{(I)} = 0$. The CCD algorithm tries then to improve the objective function by searching in the positive and the negative direction, so:

$$\left\{ \begin{array}{l} S = 1 \text{ and } \Delta v_j^{(I+1)} \leq 0 \Leftrightarrow s_j(\boldsymbol{\beta}) - \lambda \gamma_j \leq 0 \\ S = -1 \text{ and } \Delta v_j^{(I+1)} \geq 0 \Leftrightarrow s_j(\boldsymbol{\beta}) + \lambda \gamma_j \geq 0 \end{array} \right\} \Leftrightarrow \quad (26)$$

$$\left\{ \begin{array}{l} s_j(\boldsymbol{\beta}) \leq \lambda\gamma_j \\ -s_j(\boldsymbol{\beta}) \leq \lambda\gamma_j \end{array} \right\} \Leftrightarrow \quad (27)$$

$$\Leftrightarrow |s_j(\boldsymbol{\beta})| \leq \lambda\gamma_j \quad (28)$$

(2) \Rightarrow (1)

We assume now that the necessary and sufficient conditions for convergence in the GSoft theorem are fulfilled. This implies

$$|s_j(\beta)| \leq \lambda\gamma_j \quad (29)$$

We need to bear in mind also that the initial value for β_j in the CCD algorithm is $\beta_j^{(0)} = 0$. In this situation and from the definitions of the CCD algorithm for the lasso case, we have that

- if we try $S = 1$ (positive direction) then $\Delta v_j^{(0)} \leq 0$ and positive direction failed.
- if we try $S = -1$ (negative direction) then $\Delta v_j^{(0)} \geq 0$ and negative direction failed.

Therefore, following the steps of the CCD algorithm for the lasso case, this means we take $\Delta v_j^{(0)} = 0$, as both directions failed, and then

$$\Delta\beta_j = \min(\max(0, -\Delta_j), \Delta_j) = \min(0, \Delta_j) = 0 \quad (30)$$

and the CCD algorithm converges.

Case $\beta_j > 0$ (the proof is similar for $\beta_j < 0$)

(1) \Rightarrow (2)

Let us suppose that $s_j(\boldsymbol{\beta}) \neq \lambda\gamma_j$ and we will show that this gives rise to a contradiction. As the true β_j is positive and the CCD algorithm converges, from any iteration, I , we will have $\beta_j^J > 0$ for all iteration $J > I$, so $S = 1$ and $\Delta v_j^J \neq 0$, following the definition in (10). Thus, for any positive constant k ,

$$\Delta\beta_j^{(J)} = \min(\max(\Delta v_j^{(J)}, -\Delta_j^{(J)}), \Delta_j^{(J)}) \neq 0 \Rightarrow \quad (31)$$

$$\Rightarrow \Delta_j^{(J+1)} = \max\left(2|\Delta\beta_j^{(J)}|, \frac{\Delta_j^{(J)}}{2}\right) > k > 0 \quad (32)$$

and this happens for every iteration $J > I$, which enters in contradiction with the convergence of the CCD algorithm to β_j .

(2) \Rightarrow (1)

We assume now that necessary and sufficient conditions for convergence in the GSoft theorem are fulfilled. Let us suppose that the CCD algorithm converges to a different “solution” $\bar{\beta} \neq \beta$ with $\bar{\beta}_j \neq \beta_j$.

In such a case, as the conditions in (a) in the GSoft theorem determine a unique solution, it has to be $s_j(\bar{\beta}) \neq \lambda\gamma_j$. Then $\Delta v_j^{(J)} \neq 0$, for all $J > I$ with $I \in \mathbb{N}$ and therefore $\Delta \bar{\beta}_j$ does not converge to 0, which means the CCD algorithm does not converge either, which is a contradiction.

We have not mentioned or used anywhere in the proof the condition about the positive definite nature of the matrix $X'_\lambda H(\hat{\eta}) X_\lambda$. So we have to prove that this condition is also fulfilled when the CCD algorithm converges. We will prove this by *reductio ad absurdum*.

Let us assume that $X'_\lambda H(\hat{\eta}) X_\lambda$ is not definite positive. As X_λ is a complete matrix, this implies that $H(\hat{\eta})$ is not definite positive, and therefore

$$\left. \begin{array}{l} -H(\hat{\eta}) \text{ (Hessian) is not definite negative} \\ \frac{\partial L_1(\hat{\beta})}{\partial \beta_j} = 0 \text{ for all } j \in \{1, \dots, p\} \end{array} \right\} \quad (33)$$

and therefore the estimated linear predictor, $\hat{\eta}$, cannot be a maximum of the objective function in Klinger, which means $\hat{\beta}$ is not a minimum of the objective function in Genkin et al. and the CCD algorithm does not converge (contradiction).

References

- Alon U., Barkai N., Notterman D., Gish K., Ybarra S., Mack D. and Levine A.J. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. Proceedings of the National Academy of Sciences USA, **96**(12), (1999), 6745–6750.
- Antoniadis A. and Fan J. *Regularization of wavelet approximations*. Journal of the American Statistical Association, **96**(455), (2001), 939–967.
- Antoniadis A., Gijbels I. and Nikolova M. *Penalized likelihood regression for generalized linear models with nonquadratic penalties*. Annals of the Institute of Statistical Mathematics (in press), (2009).
- Bakin S. *Adaptive regression and model selection in data mining problems*. PhD Thesis, Australian National University, Canberra, (1999).

- Boulesteix A.L., Tutz G. and Strimmer K. *A CART-based approach to discover emerging patterns in microarray data*. *Bioinformatics*, **19**(18), (2003), 2465–2472.
- De Leeuw J. *Block-relaxation methods in statistics*. In Bock H.H., Lenski W. and Richter M.M., editors, *Information Systems and Data Analysis*, Springer-Verlag, Berlin, (1994).
- De Mol C., De Vito E. and Rosasco L. *Elastic-net regularization in learning theory*. *Journal of Complexity*, **25**(2), (2009), 201–230.
- Dettling M. *BagBoosting for tumor classification with gene expression data*. *Bioinformatics*, **20**(18), (2004), 3583–3593.
- Dettling M. and Bühlmann P. *Finding predictive gene groups from microarray data*. *Journal of Multivariate Analysis*, **90**, (2004), 106–131.
- Dettling M. and Bühlmann P. *Supervised clustering of genes*. *Genome Biology*, **3**(12), (2002), 0069.1–0069.15.
- Dudoit S., Fridlyand J. and Speed T.P. *Comparison of discrimination methods for the classification of tumors using gene expression data*. *Journal of the American Statistical Association*, **97**(457), (2002), 77–87.
- Efron B., Hastie T., Johnstone I. and Tibshirani R. *Least angle regression*. *Annals of Statistics*, **32**(2) (2004), 407–499.
- Eggermont P.P. and LaRiccia V.N. *Maximum penalized likelihood estimation*. Vol. 2 Regression, Springer Series in Statistics, New York, (2009).
- Fan J. and Li R. *Variable selection via nonconcave penalized likelihood and its oracle properties*. *Journal of the American Statistical Association*, **96**(456), (2001), 1348–1360.
- Frank I.E. and Friedman J.H. *A statistical view of some chemometrics tools*. *Technometrics*, **35**(2), (1993), 109–135.
- Friedman J., Hastie T. and Tibshirani R. *Regularization paths for generalized linear models via coordinate descent*. Technical Report, Department of Statistics, Stanford University, (2008).
- Genkin A., Lewis D.D. and Madigan D. *Sparse logistic regression for text categorization*. DIMACS Working Group on Monitoring Message Streams, Project Report, (2005).

- Goldstein T. and Osher S. *The Split Bregman method for L1 regularized problems*. UCLA CAAM Report 08–29, (2008).
- Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A. et al. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science, **286**(5439), (1999), 531–537.
- Guyon I. and Elisseeff A. *An introduction to variable and feature selection*. Journal of Machine Learning Research, **3**(Mar), (2003), 1157–1182.
- Guyon I., Weston J., Barnhill S. and Vapnik V. *Gene selection for cancer classification using support vector machines*. Machine Learning, **46**(1–3), (2002), 389–422.
- Hall M. *Correlation-based feature selection for machine learning*. PhD Thesis, Department of Computer Science, Waikato University, New Zealand, (1999).
- Hoerl A.E. and Kennard R. *Ridge regression: biased estimation for nonorthogonal problems*. Technometrics, **12**(1), (1970), 55–67.
- Huang J., Horowitz J.L. and Ma S. *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*. Technical Report, Department of Statistics and Actuarial Science, The University of Iowa, (2006).
- Huang J., Ma S. and Zhang C.H. *Adaptive lasso for sparse high-dimensional regression models*. Statistica Sinica, **18**(4), (2008), 1603–1618.
- Huang J., Ma S. and Zhang C.H. *The iterated lasso for high-dimensional logistic regression*. Technical report No. 392, The University of Iowa, (2008).
- Huber P.J. *The behavior of maximum likelihood estimates under nonstandard conditions*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, (1967).
- Hunter D.R. and Li R. *Variable selection using MM algorithms*. Annals of Statistics, **33**(4), (2005), 1617–1642.
- Kim Y., Choi H. and Oh H.S. *Smoothly clipped absolute deviation on high dimensions*. Journal of the American Statistical Association, **103**(484), (2008), 1665–1673.

- Klinger A. *Inference in high dimensional generalized linear models based on soft thresholding*. Journal of the Royal Statistical Society Series B, **63**(2), (2002), 377–392.
- Knight K. and Fu W.J. *Asymptotics for lasso-type estimators*. Annals of Statistics, **28**(5), (2000), 1356–1378.
- Krishnapuram B., Carin L., Figueiredo M.A.T. and Hartemink A.J. *Sparse multinomial logistic regression: fast algorithms and generalization bounds*. IEEE Transactions on Pattern Analysis and Machine Intelligence, **27**(6), (2005), 957–968.
- Lee S.I., Lee H., Abbeel P. and Ng A.Y. *Efficient L_1 regularized logistic regression*. Proceedings of the Twenty-first International Conference on Machine Learning (AAAI-06), (2006).
- Lee K.E., Sha N., Dougherty E.R., Vannucci M. and Mallick B.K. *Gene selection: a bayesian variable selection approach*. Bioinformatics, **19**(1), (2003), 90–97.
- Liu Z., Jiang F., Tian G., Wang S., Sato F., Meltzer S.J. and Tan M. *Sparse logistic regression with L_p penalty for biomarker identification*. Statistical Applications in Genetics and Molecular Biology, **6**(1), (2007), article 6.
- Lv J. and Fan Y. *A unified approach to model selection and sparse recovery using regularized least squares*. Annals of Statistics, **37**(6A), (2009), 3498–3528.
- Meier L., van de Geer S. and Bühlmann P. *The group lasso for logistic regression*. Journal of the Royal Statistical Society Series B, **70**(1), (2008), 53–71.
- Meinshausen N. and Bühlmann P. *High dimensional graphs and variable selection with the lasso*. Annals of Statistics, **34**(3), (2006), 1436–1462.
- Nguyen D.V. and Rocke D.M. *Tumor classification by partial least squares using microarray gene expression data*. Bioinformatics, **18**(1), (2002), 39–50.
- Park M.Y. and Hastie T. *An L_1 regularization-path algorithm for generalized linear models*. Manuscript, Department of Statistics, Stanford University, (2006).

- Roth V. *The generalized lasso*. IEEE Transactions on Neural Networks, **15**, (2004), 16–28.
- Schmidt M., Fung G. and Rosales R. *Fast optimization methods for L_1 regularization: a comparative study and two new approaches*. European Conference on Machine Learning (ECML), (2007).
- Shevade S. and Keerthi S. *A simple and efficient algorithm for gene selection using sparse logistic regression*. Bioinformatics, **19**(17), (2003), 2246–2253.
- Tarigan B. and van de Geer S. *Classifiers of support vector machine type with L_1 complexity regularization*. Bernoulli, **12**(6), (2006), 1045–1076.
- Thomas J.G., Olson J.M. and Tapscott S.J. *An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles*. Genome Research, **11**, (2001), 1227–1236.
- Tibshirani R. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society Series B, **58**(1), (1996), 267–288.
- Tseng P. *Convergence of block coordinate descent method for nondifferentiable maximization*. Journal of Optimization Theory and Applications, **109**, (2001), 473–492.
- Tseng P. and Yun S. *A coordinate descent gradient method for nonsmooth separable minimization*. Mathematical Programming, **117**, (2009), 387–423.
- Wang H. and Leng C. *Unified lasso estimation via least squares approximation*. Journal of the American Statistical Association, **102**(479), (2007), 1039–1048.
- Weston J., Elisseeff A., Scholkopf B. and Tipping M. *Use of the zero-norm with linear models and kernel methods*. Journal of Machine Learning Research, **3**(Mar), (2003), 1439–1461.
- Wright S.J., Nowak R.D. and Figueiredo M.A.T. *Sparse reconstruction by separable approximation*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2008).
- Wu T.T. and Lange K. *Coordinate descent algorithms for lasso penalized regression*. Annals of Applied Statistics, **2**(1), (2008), 224–244.

- Yuan M. and Lin Y. *Model selection and estimation in regression with grouped variables*. *Journal of the Royal Statistical Society Series B*, **68**(1), (2006), 49–67.
- Zhang C.H. and Huang J. *The sparsity and bias of the lasso selection in high-dimensional linear regression*. *Annals of Statistics*, **36**(4), (2008), 1567–1594.
- Zhang T. *Some sharp performance bounds for least squares regression with L1 regularization*. *Annals of Statistics*, **37**(5A), (2009), 2109–2144.
- Zhang T. and Oles F. *Text categorization based on regularized linear classifiers*. *Information Retrieval*, **4**(1), (2001), 5–31.
- Zou H. *The adaptive lasso and its oracle properties*. *Journal of the American Statistical Association*, **101**(476), (2006), 1418–1429.
- Zou H. and Hastie T. *Regularization and variable selection via the elastic net*. *Journal of the Royal Statistical Society Series B*, **67**(2), (2005), 301–320.
- Zou H. and Li R. *One-step sparse estimates in nonconcave penalized likelihood models*. *Annals of Statistics*, **36**(4), (2008), 1509–1533.