

# An efficient algorithm to estimate the sparse group structure of an high-dimensional generalized linear model

*Un algoritmo efficiente per la stima della struttura sparsa e raggruppata di un modello lineare generalizzato definito in uno spazio ad alta dimensionalità*

Luigi Augugliaro and Angelo M. Mineo

**Abstract** Massive regression is one of the new frontiers of computational statistics. In this paper we propose a generalization of the group least angle regression method based on the differential geometrical structure of a generalized linear model specified by a fixed and known group structure of the predictors. An efficient algorithm is also proposed to compute the proposed solution curve.

**Abstract** *I modelli di regressione caratterizzati da un elevato numero di variabili esplicative costituiscono una delle nuove frontiere della statistica computazionale. In questo articolo proponiamo una generalizzazione del group least angle regression method fondato sulla struttura geometrica differenziale di un modello lineare generalizzato specificato mediante struttura raggruppata e nota dei regressori. Proponiamo inoltre un efficiente algoritmo per la stima del sentiero dei coefficienti.*

**Key words:** Generalized linear models; Group lasso; dglars; Differential geometry

## 1 Introduction

This paper deals with the problem of how to find important groups of predictors in a generalized linear model (GLM) [5]. In linear regression modelling, [8] propose an extension of the lasso estimator [7], called group lasso to select a sparse predictive set of groups of predictors. A number of papers studies the theoretical properties of the group lasso estimator. In [2] the behaviour of the group lasso estimator, as a group selection method, is studied under the assumption of a fixed number of random predictors. In [6] the previous asymptotic results are extended showing that a direct generalization of the adaptive lasso [9] to the group lasso framework satis-

---

Luigi Augugliaro  
University of Palermo, e-mail: luigi.augugliaro@unipa.it

Angelo M. Mineo  
University of Palermo, e-mail: angelo.mineo@unipa.it

fies the oracle properties. They also derive the asymptotic behaviour of the group lasso when the dimension of the parameter space grows with the sample size. For a complete treatment of the group lasso estimator, the interested reader is referred to [3]. In this paper we extend the differential geometrical framework proposed in [1] to define a generalization of the group least angle regression method (glars) introduced in [8]. The rest of this paper is structured as follows. In section 2 we study how to use the differential geometrical structure of a GLM to deal with the grouping structure of the predictors and we propose an extension of the dglars that we call group dglars. Finally in section 3 we propose an efficient computational methods to estimate the solution path.

## 2 Group dglars: differential geometric foundation

Let  $\mathcal{Y}$  be a scalar random variable with probability density function  $p(y; \theta, \phi) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)]$ , where  $\theta \in \Theta \subseteq \mathbb{R}$  is called canonical parameter,  $\phi \in \Phi \subseteq \mathbb{R}^+$  is called dispersion parameter and  $a(\cdot), b(\cdot)$  and  $c(\cdot, \cdot)$  are specific given functions. We shall assume that  $\Theta$  is an open set and  $\phi$  is a fixed value so that, to simplify our notation, we can let  $\phi = 1$ . The expected value of  $\mathcal{Y}$  is related to the canonical parameter by  $E(\mathcal{Y}) = \mu = \partial_\theta b(\theta)$ . We let  $\mathcal{S} = \mu(\text{int } \Theta)$ . Since  $\mu$  is a reparametrization of the model, we denote by  $p(y; \mu)$  the probability density function of  $\mathcal{Y}$ . Let  $\mathcal{X}$  be the  $p$ -dimensional vector of random predictors. In what follows, we shall assume that we have a fixed number of groups, say  $G$ , and that such grouping structure defines a partition of the set  $\mathcal{A} = \{1, \dots, p\}$ . To simplify our notation, we shall use the multi-index notation, i.e. let  $m = \{m_1, \dots, m_{|m|}\}$  the set of indices identifying the  $m$ th group, then the corresponding block of a given  $p$ -dimensional vector is denoted using  $m$  as index, for example by  $\beta_m$  we denote the  $|m|$ -dimensional parameter vector corresponding to the  $m$ th group of predictors.

Consider  $N$  i.i.d. copies of the observation pair  $(\mathcal{Y}, \mathcal{X}^T)^T$ , under this setting a GLM is based on the assumption that the conditional expected value of  $\mathcal{Y}_i$  given  $\mathcal{X}_i = x_i$  is specified by the link function  $h(\cdot)$  as follows  $h(E(\mathcal{Y}_i | \mathcal{X}_i = x_i)) = \sum_{m \in \mathcal{A}} x_{i,m}^T \beta_m^*$ . With a slight abuse of notation we let  $h^{-1}(\sum_{m \in \mathcal{A}} x_{i,m}^T \beta_m) = \mu_i(\beta)$  and  $\mu(\beta) = (\mu_1(\beta), \dots, \mu_N(\beta))^T$ . In what follows we shall assume that our model has a sparse grouping structure, this means that there exists a subset  $\mathcal{A}^* \subset \mathcal{A}$  such that  $\beta_m^* = 0$  if and only if  $m \notin \mathcal{A}^*$ . Let  $\ell_i(\beta)$  be the log-likelihood function for the  $i$ th observation. The log-likelihood function is denoted by  $\ell(\beta) = \sum_{i=1}^N \ell_i(\beta)$  while the average observed log-likelihood function is denoted by  $\mathbb{P}_N \ell(\beta) = \sum_{i=1}^N \ell_i(\beta) / N$ . Finally, throughout this paper we shall use the convention that the indices  $i, j$  and  $l$  correspond to the quantities related to  $\mu$  while  $m, n$  and  $q$  refer to the quantities related to the groups of predictors. For example, by  $\partial_i \ell(\beta)$  we mean the derivative of the log-likelihood function with respect to  $\mu_i$  and evaluated at  $\mu(\beta)$ , i.e.  $\partial \ell(\beta) / \partial \mu_i$ , while by  $\partial_m \ell(\beta)$  we mean the block of the gradient vector corresponding to the  $m$ th group of predictors.

Let  $p(y; \mu(\beta^*))$  be the true conditional distribution function which is an element of the set  $\mathcal{M} = \{p(y; \mu(\beta)) = \prod_{i=1}^N p(y_i; \mu_i(\beta)) : \beta \in \mathbb{R}^{p+1}\}$ . Following [1],  $\mathcal{M}$  can be treated as a submanifold of  $\mathcal{S} = \{p(y; \mu) = \prod_{i=1}^N p(y_i; \mu_i) : \mu \in \widehat{\mathcal{S}}^N\}$ . The tangent space of  $\mathcal{S}$  at  $p(y; \mu(\beta))$  is defined as the linear vector space  $T_{p\{\mu(\beta)\}}\mathcal{S} = \text{span}\{\partial_i \ell(\beta)/\sqrt{N} : i = 1, \dots, N\}$  while, considering the grouped structure of the  $p$  predictors, the tangent space of  $\mathcal{M}$  at  $p(y; \mu(\beta))$  can be defined as direct sum of the  $G$  tangent spaces associated to the considered groups, i.e.  $T_{p\{\mu(\beta)\}}\mathcal{M}_m = \text{span}\{\partial_{m_j} \ell(\beta)/\sqrt{N} : m_j \in m\}$  is the tangent space associated to the  $m$ th group. An inner product can be defined by the conditional expected value of  $\mathcal{Y}$  given  $\mathcal{X}$ , i.e. the inner product between  $v_\beta = \sum_{i=1}^N v_i(\beta) \partial_i \ell(\beta)/\sqrt{N} \in T_{p\{\mu(\beta)\}}\mathcal{S}$  and  $w_\beta^m = \sum_{m_j \in m} w_{m_j}^m(\beta) \partial_{m_j} \ell(\beta)/\sqrt{N} \in T_{p\{\mu(\beta)\}}\mathcal{M}_m$ , is defined as  $\langle w_\beta^m, v_\beta \rangle_{p\{\mu(\beta)\}} = E_{\mathcal{Y}|X\beta}(w_\beta^m \cdot v_\beta)$ , where  $X$  is the design matrix. In order to complete the differential geometrical framework, it is necessary to introduce the observed response vector and to consider its relationship with the tangent spaces  $T_{p\{\mu(\beta)\}}\mathcal{M}_m$ . This can be done by the tangent residual vector, i.e.  $r_\beta = \sum_{i=1}^N r_i(\beta) \partial_i \ell(\beta)/\sqrt{N} \in T_{p\{\mu(\beta)\}}\mathcal{S}$ , where  $r_i(\beta) = y_i - \mu_i(\beta)$  are fixed values.

## 2.1 Group dgLARS: formal description of group dgLARS

Let  $\beta$  be a fixed point of  $\mathbb{R}^{p+1}$  and  $r_\beta^m = \sum_{m_h \in m} r_{m_h}^m(\beta) \partial_{m_h} \ell(\beta)/\sqrt{N}$  be the projection of  $r_\beta$  onto  $T_{p\{\mu(\beta)\}}\mathcal{M}_m$ . The following identities extend the results given in [4] to GLMs with grouped predictors:

$$\mathbb{P}_N \partial_{m_j} \ell(\beta) = \langle \partial_{m_j} \ell(\beta)/\sqrt{N}, r_\beta \rangle_{p\{\mu(\beta)\}} = \sum_{m_h \in m} \mathbb{P}_N I_{m_j m_h}(\beta) r_{m_h}^m(\beta),$$

where  $\mathbb{P}_N \partial_{m_j} \ell(\beta) = \sum_{i=1}^N \partial_{m_j} \ell_i(\beta)/N$  and  $\mathbb{P}_N I_{m_j m_h}(\beta) = \sum_{i=1}^N E_{\mathcal{Y}|X_i^T \beta} \{\partial_{m_j} \ell_i(\beta) \cdot \partial_{m_h} \ell_i(\beta)\}/N$ . Using the previous identities we can conclude that the coefficients of  $r_\beta^m$  are given by the following expression

$$r_\beta^m(\beta) = \{\mathbb{P}_N I_{mm}(\beta)\}^{-1} \mathbb{P}_N \partial_m \ell(\beta), \quad (1)$$

where  $\mathbb{P}_N I_{mm}(\beta)$  is the matrix corresponding to that diagonal block of the mean conditional Fisher information matrix that corresponds to the  $m$ th group. Combining (1) with  $\|r_\beta^m\|_{p\{\mu(\beta)\}}^2 = \langle r_\beta, r_\beta^m \rangle_{p\{\mu(\beta)\}} = \cos\{\rho_m(\beta)\} \|r_\beta\|_{p\{\mu(\beta)\}} \|r_\beta^m\|_{p\{\mu(\beta)\}}$ , we have

$$\|r_\beta^m\|_{p\{\mu(\beta)\}} = \cos\{\rho_m(\beta)\} \|r_\beta\|_{p\{\mu(\beta)\}} = \quad (2)$$

$$= \sqrt{\mathbb{P}_N \partial_m \ell(\beta)^T \{\mathbb{P}_N I_{mm}(\beta)\}^{-1} \mathbb{P}_N \partial_m \ell(\beta)} = \sqrt{\frac{\omega_m(\beta)}{N}}, \quad (3)$$

where  $\omega_m(\beta)$  is the Rao's score test statistic for the  $m$ th group and  $\rho_m(\beta)$  is the angle between  $r_\beta$  and  $r_\beta^m$  at  $p(y; \mu(\beta))$ . Expression (2) generalizes the univariate differential geometric characterization of the signed Rao's score test statistic given in [1] and shows that  $\sqrt{\omega_m(\beta)/N}$  is based on two essential elements on which any variable selection method should be based, namely, an invariant measure that can be used to locally rank the  $G$  groups of predictors, i.e.  $\cos\{\rho_m(\beta)\}$ , and a global measure of goodness of fit, i.e.  $\|r_\beta\|_{p\{\mu(\beta)\}}$ .

Since  $\|r_\beta\|_{p\{\mu(\beta)\}}$  is equal for any group, using expression (2) and (3), we say that the  $m$ th and the  $n$ th group of predictors satisfy the generalized group equiangularity condition when

$$\sqrt{\frac{\omega_m(\beta)}{N}} = \sqrt{\frac{\omega_n(\beta)}{N}}. \quad (4)$$

As suggested by [8], when the groups have different sizes, condition (4) can be substituted with its weighted version obtained using  $\omega_q(\beta)/|q|$  instead of  $\omega_q(\beta)$ . Using the Cholesky square root of  $\mathbb{P}_N I_{mm}(\beta)$ , denoted by  $\{\mathbb{P}_N I_{mm}(\beta)\}^{1/2}$ , condition (4) can be rewritten by using the Euclidean norm of the transformation

$$\zeta_m(\beta) = \{\mathbb{P}_N I_{mm}(\beta)\}^{-1/2} \mathbb{P}_N \partial_m \ell(\beta) / \sqrt{|m|}. \quad (5)$$

The group dglars method computes the solution curve, denoted by  $\hat{\beta}(\gamma)$ , finding a sequence of  $G$  transition points, denoted by  $0 \leq \gamma^{(G)} \leq \dots \leq \gamma^{(1)}$ , such that for any  $\gamma \in (\gamma^{(q+1)}; \gamma^{(q)})$  the active set  $\mathcal{A}$  contains  $q$  groups and the following conditions are satisfied:

$$\|\zeta_m(\hat{\beta}(\gamma))\| = \frac{\gamma}{N}, \quad \forall m \in \mathcal{A}, \quad (6)$$

$$\|\zeta_n(\hat{\beta}(\gamma))\| < \frac{\gamma}{N}, \quad \forall n \in \mathcal{A}^c. \quad (7)$$

When  $\gamma = \gamma^{(j)}$ , with  $j = 2, \dots, G$ , the condition

$$\exists n \in \mathcal{A}^c : \|\zeta_n(\hat{\beta}(\gamma^{(j)}))\| = \|\zeta_m(\hat{\beta}(\gamma^{(j)}))\|, \quad \forall m \in \mathcal{A}, \quad (8)$$

is satisfied and a new group is included in  $\mathcal{A}$ . Assume that  $\mathcal{A} = \{a, \dots, q\}$  and define  $\zeta_{\mathcal{A}}(\hat{\beta}_{\mathcal{A}}(\gamma)) = (\zeta_a^T(\hat{\beta}_{\mathcal{A}}(\gamma)), \dots, \zeta_q^T(\hat{\beta}_{\mathcal{A}}(\gamma)))^T$  and  $v_{\mathcal{A}} = (0, v_b^T, \dots, v_q^T)^T$ , where  $a$  identifies the intercept term and  $v_m = \zeta_m(\hat{\beta}_{\mathcal{A}}(\gamma^{(q)})) / \|\zeta_m(\hat{\beta}_{\mathcal{A}}(\gamma^{(q)}))\|$ . Then, as a consequence of conditions (6) and (7), for any  $\gamma \in (\gamma^{(q+1)}; \gamma^{(q)})$  the group dglars estimator is the Z-estimator implicitly defined by the following system of estimating equations

$$\zeta_{\mathcal{A}}(\hat{\beta}_{\mathcal{A}}(\gamma)) - \frac{\gamma}{N} v_{\mathcal{A}} = 0, \quad \forall \gamma \in (\gamma^{(q+1)}; \gamma^{(q)}), \quad (9)$$

while  $\hat{\beta}_{\mathcal{A}^c}(\gamma) = 0$ . The previous system shows that in our setting the intercept is not penalized but the method can be easily extended to the case in which is penalized or to GLMs defined with no intercept term. The definition of the group dglars method implicitly assumes that we are working in a setting in which it is possible to compute

the MLE of the model defined by all the  $G$  groups. It is easy to see that the proposed method can also be defined in a high-dimensional setting, i.e.  $p \gg N$ , simply by using a value larger than  $\gamma^{(G)}$  as the last value of the sequence of transition points.

Using the differential geometrical structure of a GLM with grouped predictors the proposed group dglars method can be described as follows: let  $\hat{\beta}_a = \arg\max_{\beta_a} \ell(\beta_a)$  be the starting point of the proposed solution curve, i.e. the MLE of the intercept term. Then  $\|r_{\hat{\beta}_a}^m\|_{p\{\mu(\hat{\beta}_a)\}}$  can be used to locally rank the  $G$  groups. Suppose that the group of predictors, identified by the set  $b$ , is the first to be included in the active set, thus  $\mathcal{A} = \{a, b\}$ . Then the solution curve  $\hat{\beta}(\gamma) = (\hat{\beta}_a(\gamma), \hat{\beta}_b^T(\gamma))^T$  is such that  $r_{\hat{\beta}(\gamma)}$  is always orthogonal to  $\partial_a \ell(\hat{\beta}(\gamma))$ , while the direction is defined by  $r_{\hat{\beta}(\gamma)}^b$ . The curve continues as defined above until  $\gamma^{(2)}$ , for which exists a new group of predictors, say  $c$ , that satisfies condition (4). Then  $c$  is included in  $\mathcal{A}$  and the curve  $\hat{\beta}(\gamma) = (\hat{\beta}_a(\gamma), \hat{\beta}_b^T(\gamma), \hat{\beta}_c^T(\gamma))^T$  is such that  $\|\zeta_b(\hat{\beta}(\gamma))\| = \|\zeta_c(\hat{\beta}(\gamma))\|$ . The proposed method proceeds as previously described until the last value of the transition points.

### 3 Computational Aspects

The predictor-corrector method proposed in [1] can be extended to compute the solution curve implicitly defined by the group dglars method. The basic idea underlying the predictor-corrector algorithm is to trace a curve that is implicitly defined by a system of non-linear equations, by generating a sequence of points satisfying a chosen tolerance criterion. Differentiating system (9) with respect to  $\gamma$  it is easy to see that, for any  $\gamma \in (\gamma^{(q+1)}, \gamma^{(q)}]$ , we can approximate the value of the non-zero elements of the solution curve at  $\gamma - \Delta\gamma$  by

$$\tilde{\beta}_{\mathcal{A}}(\gamma - \Delta\gamma) = \hat{\beta}_{\mathcal{A}}(\gamma) - \frac{\Delta\gamma}{N} \left\{ \frac{\partial \zeta_{\mathcal{A}}(\hat{\beta}_{\mathcal{A}}(\gamma))}{\partial \beta_{\mathcal{A}}} \right\}^{-1} v_{\mathcal{A}},$$

where  $\partial \zeta_{\mathcal{A}}(\hat{\beta}_{\mathcal{A}}(\gamma)) / \partial \beta_{\mathcal{A}}$  is the Jacobian matrix of  $\zeta_{\mathcal{A}}(\hat{\beta}_{\mathcal{A}}(\gamma))$  and  $\Delta\gamma \in (0; \gamma - \gamma^{(q+1)}]$ . We use  $\tilde{\beta}(\gamma - \Delta\gamma)$  for the predictor step of the proposed predictor-corrector algorithm. In literature there are different methods to compute the step size  $\Delta\gamma$ . In our algorithm we approximate the step size that will change  $\mathcal{A}$ . In according with condition (8), a change in the active set occurs when

$$\exists n \subset \mathcal{A}^c : \|\zeta_n(\hat{\beta}_{\mathcal{A}}(\gamma - \Delta\gamma))\| = \|\zeta_m(\hat{\beta}_{\mathcal{A}}(\gamma - \Delta\gamma))\|, \quad \forall m \subset \mathcal{A}. \quad (10)$$

Using the approximation  $\|\zeta_n(\hat{\beta}_{\mathcal{A}}(\gamma - \Delta\gamma))\| \approx \|\zeta_n(\hat{\beta}_{\mathcal{A}}(\gamma))\| - d_\gamma \|\zeta_n(\hat{\beta}_{\mathcal{A}}(\gamma))\| \Delta\gamma$ , and observing that  $\|\zeta_m(\hat{\beta}_{\mathcal{A}}(\gamma - \Delta\gamma))\| = \gamma - \Delta\gamma$ , for any  $m \subset \mathcal{A}$ , condition (10) can be approximated as

$$\exists n \subset \mathcal{A}^c : \left| \|\zeta_n(\hat{\beta}_{\mathcal{A}^c}(\gamma))\| - \frac{d\|\zeta_n(\hat{\beta}_{\mathcal{A}^c}(\gamma))\|}{d\gamma} \Delta\gamma \right| \approx \gamma - \Delta\gamma.$$

This leads to the following approximation of the step size

$$\Delta\gamma \approx \min_{n \subset \mathcal{A}^c}^+ \left\{ \frac{\gamma - \|\zeta_n(\hat{\beta}_{\mathcal{A}^c}(\gamma))\|}{1 - d\|\zeta_n(\hat{\beta}_{\mathcal{A}^c}(\gamma))\|/d\gamma}, \frac{\gamma + \|\zeta_n(\hat{\beta}_{\mathcal{A}^c}(\gamma))\|}{1 + d\|\zeta_n(\hat{\beta}_{\mathcal{A}^c}(\gamma))\|/d\gamma} \right\}.$$

## 4 Conclusion

In this paper we have extended the differential geometrical framework used to define the dglars to a GLM with grouped predictors. The proposed approach allows us to explore the sparse structure of a GLM specified by predictors with a group structure. This implies that the proposed geometrical setting can easily be extended to other models, such as the model based on the quasi-likelihood function or the Cox's proportional hazard model. We have shown the role of the Rao score test statistic in the equiangularity condition and we have also proposed a new efficient algorithm that can be used to compute the solution curve.

## References

1. Augugliaro, L., Mineo, A. M., Wit, E. C.: dglars: a differential geometric approach to sparse generalized linear models. *J. R. Statist. Soc. B* **75**, 471–498 (2013)
2. Bach, F. R.: Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9**, 1179–1225 (2008)
3. Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high dimensional models. *Statist. Sci.* **27**, 481–499 (2012)
4. Kass, R. E., Vos, P. W.: *Geometrical Foundations of Asymptotic Inference*. John Wiley & Sons, New York (1997)
5. McCullagh, P., Nelder, J.: *Generalized Linear Models*. Chapman & Hall, 2nd ed, London (1989)
6. Nardi, Y., Rinaldo, A.: On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Statist.* **2**, 605–633 (2008)
7. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–288 (1996)
8. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* **68**, 49–67 (2006)
9. Zou, H.: The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429 (2006)