University of Groningen

A differential-geometric approach to generalized linear models with grouped predictors

Augugliaro, Luigi; Mineo, Angelo M.; Wit, Ernst C.

Published in:
Biometrika

DOI:
[10.1093/biomet/asw023](https://doi.org/10.1093/biomet/asw023)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Augugliaro, L., Mineo, A. M., & Wit, E. C. (2016). A differential-geometric approach to generalized linear models with grouped predictors. *Biometrika*, 103(3), 563-577. <https://doi.org/10.1093/biomet/asw023>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

A differential-geometric approach to generalized linear models with grouped predictors

BY LUIGI AUGUGLIARO, ANGELO M. MINEO

*Department of Economics, Business and Statistics, University of Palermo, Building 13,
Viale delle Scienze, 90128 Palermo, Italy*

luigi.augugliaro@unipa.it angelo.mineo@unipa.it

AND ERNST C. WIT

*Johann Bernoulli Institute of Mathematics and Computing Science, University of Groningen,
9747 AG Groningen, The Netherlands*

e.c.wit@rug.nl

SUMMARY

We propose an extension of the differential-geometric least angle regression method to perform sparse group inference in a generalized linear model. An efficient algorithm is proposed to compute the solution curve. The proposed group differential-geometric least angle regression method has important properties that distinguish it from the group lasso. First, its solution curve is based on the invariance properties of a generalized linear model. Second, it adds groups of variables based on a group equiangularity condition, which is shown to be related to score statistics. An adaptive version, which includes weights based on the Kullback–Leibler divergence, improves its variable selection features and is shown to have oracle properties when the number of predictors is fixed.

Some key words: Differential-geometric least angle regression; Differential geometry; Generalized linear model; Group lasso; Score statistic.

1. INTRODUCTION

This paper deals with finding important groups of predictors in a generalized linear model (McCullagh & Nelder, 1989). To perform sparse inference in a linear regression model specified by groups of predictors, Yuan & Lin (2006) develop an extension of the lasso estimator (Tibshirani, 1996), called the group lasso estimator. Bach (2008) investigates its behaviour as a group selection method, under the assumption of a fixed number of random predictors, generalizing the irrepresentability condition (Zou, 2006; Zhao & Yu, 2006) and showing that the group lasso is a consistent group selection method. Nardi & Rinaldo (2008) extend asymptotic results obtained by Bach (2008) showing that a direct generalization of the adaptive lasso (Zou, 2006) to the group lasso framework has oracle properties (Fan & Li, 2001), and they derive the asymptotic behaviour of the group lasso when the dimension of the parameter space grows with the sample size. Wei & Huang (2010) study the asymptotic properties of the group lasso and its adaptive version under generalized sparsity and sparse Riesz conditions (Zhang & Huang, 2008). Huang & Zhang (2010) study the conditions under which the group lasso performs better than

the classical lasso. For a more complete treatment of the group lasso estimator, see [Huang et al. \(2012\)](#).

In this paper we propose a new path estimator called the group differential-geometric least angle regression estimator, generalizing the approach proposed in [Augugliaro et al. \(2013\)](#). Although the description of our proposal is closely related to the method developed in [Efron et al. \(2004\)](#) to compute the path of solutions for the lasso problem, we shall show that it does not solve the group lasso problem for generalized linear models; the difference comes from the use of score statistics to define a new Z -estimator.

2. GROUP DIFFERENTIAL-GEOMETRIC LEAST ANGLE REGRESSION

2.1. Notation

Let Y be a scalar random variable with exponential family probability density function, i.e., $p(y; \theta, \phi) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)]$, where $\theta \in \Theta \subseteq \mathbb{R}$ is the canonical parameter, $\phi \in \Phi \subseteq \mathbb{R}^+$ is the dispersion parameter and $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are given functions. We assume that Θ is an open set. To simplify notation, we assume that the dispersion parameter $\phi = 1$. We can also treat scenarios where ϕ is unknown, but this complicates the exposition. The expected value of Y , denoted by μ , is related to the canonical parameter by $\mu = \partial b(\theta)/\partial \theta$. Since μ is a reparameterization of the model, we denote by $p(y; \mu)$ the probability density function of Y . Let $X = (X_0, \dots, X_p)^\top$ be a vector of predictors where $X_0 = 1$ and the remaining predictors are random. We assume that we have a fixed number of groups of random predictors, say K . In our setting we use the partition $\mathcal{A} = \{g_0, \dots, g_K\}$ of the set $\{0, \dots, p\}$, where $g_0 = \{0\}$ and g_m is the set of indices identifying the predictors belonging to the m th group; $|g_m|$ denotes the cardinality of the set g_m . In this paper any $(p+1)$ -dimensional vector, say $v = (v_0, \dots, v_p)^\top$, indexed by an element of partition $\mathcal{A}$ denotes the subvector of v associated with the corresponding group, i.e., $v_{g_m} = (v_h)_{h \in g_m}$. Furthermore, a $(p+1)$ -dimensional vector indexed by a subpartition $\mathcal{A}_1 \subseteq \mathcal{A}$ denotes the subvector $v_{\mathcal{A}_1} = (v_{g_m}^\top)_{g_m \in \mathcal{A}_1}^\top$. The Euclidean norm of a vector v is denoted by $\|v\|$. Similar notation is also used to identify the submatrices of a given $(p+1) \times (p+1)$ matrix, say $M = (m_{hk})$. For example, given the two subpartitions $\mathcal{A}_1$ and $\mathcal{A}_2$ of $\mathcal{A}$, by $M_{\mathcal{A}_1 \mathcal{A}_2}$ we denote the submatrix of M with elements m_{hk} , where $h \in g_m \in \mathcal{A}_1$ and $k \in g_n \in \mathcal{A}_2$.

Consider N independent and identically distributed copies of the observation pair $(Y, X^\top)^\top$. A generalized linear model is based on the assumption that the conditional expected value of Y_i given $X_i = x_i$ can be specified as

$$E_{\beta^*}(Y_i | X_i = x_i) = \mu(\eta_i^*) = \mu(x_i^\top \beta^*) \quad (i = 1, \dots, N),$$

where the inverse of the function $\mu(\cdot)$ is called the link function, the linear combination $\eta_i^* = x_i^\top \beta^*$ is called the i th linear predictor and β^* is the true parameter vector. The conditional expected value is indexed by β^* in order to emphasize that it is computed using the true probability density function. We assume that our model has a sparse grouping structure, so there exists a subpartition $\mathcal{A}^* \subset \mathcal{A}$ such that $\|\beta_{g_m}^*\| = 0$ if and only if $g_m \notin \mathcal{A}^*$. With a slight abuse of notation we let $\mu(x_i^\top \beta) = \mu_i(\beta)$ and $\mu(\beta) = \{\mu_1(\beta), \dots, \mu_N(\beta)\}^\top$. Let $\ell_i(\beta) = \log p\{y_i; \mu_i(\beta)\}$ be the loglikelihood function for the i th observation and let $\mathbb{P}_N \ell(\beta) = \sum_{i=1}^N \ell_i(\beta)/N$. Finally, we denote by $\mathbb{P}_N \partial_h \ell(\beta) = N^{-1} \sum_{i=1}^N \partial \ell_i(\beta)/\partial \beta_h$ the h th element of the gradient of $\mathbb{P}_N \ell(\beta)$ and by $\mathbb{P}_N I_{hk}(\beta) = N^{-1} \sum_{i=1}^N E_\beta[\{\partial \ell_i(\beta)/\partial \beta_h\} \{\partial \ell_i(\beta)/\partial \beta_k\} | X_i = x_i]$ the generic element of the conditional Fisher information matrix $\mathbb{P}_N I(\beta)$.

2.2. Differential-geometric least angle regression

Using the terminology introduced in §2.1, in this section each group contains only one predictor, i.e., $K = p$, $g_m = \{m\}$ and $\mathcal{A} = \{0, \dots, p\}$. Furthermore, to be coherent with the asymptotic results in §3 we slightly modify the original description of the differential-geometric least angle regression method; in this section we scale score test statistics by a factor of $N^{1/2}$ while the tuning parameter γ is scaled by N . It is easy to see that the tuning parameter of the original differential-geometric least angle regression method is scaled by a factor of $N^{1/2}$. Let $r_{g_m}^u(\beta) = \{\mathbb{P}_N I_{g_m g_m}(\beta)\}^{-1/2} \mathbb{P}_N \partial_{g_m} \ell(\beta)$ be the scaled score statistic for the m th predictor.

Differential-geometric least angle regression is a method for constructing a path of solutions, indexed by a positive parameter γ , where the nonzero estimates of each solution can be formally defined as follows. For any dataset there exists with probability one a finite set of transition points, denoted by $0 \leq \gamma^{(K)} \leq \dots \leq \gamma^{(1)}$, such that for any $\gamma^{(k+1)} < \gamma < \gamma^{(k)}$ the active set is defined as $\hat{\mathcal{A}} = \{g_m \in \mathcal{A} : |\hat{\beta}_{g_m}(\gamma)| > 0\}$ and the corresponding vector of nonzero estimates, denoted by $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma)$, is defined as the solution of the system of nonlinear equations

$$r_{\hat{\mathcal{A}}}^u\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma)\} - N^{-1} \gamma \text{sign}[r_{\hat{\mathcal{A}}}^u\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}] = 0, \quad (1)$$

satisfying the further condition that $|r_{g_n}^u\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma)\}| < \gamma/N$, for any $g_n \notin \hat{\mathcal{A}}$. We denote the estimator of the true active set $\mathcal{A}^*$ by $\hat{\mathcal{A}}$. System (1) comes from the differential-geometric characterization of the score test statistic, i.e., $r_{g_m}^u(\beta)$ is proportional to a generalization of the Euclidean notion of cosine of the angle between the m th column of the design matrix and the residual vector $r(\beta) = y - \mu(\beta)$, so that it can be used to extend the equiangularity condition of Efron et al. (2004) to generalized linear models. Formally, we say that the m th and n th predictors satisfy the generalized equiangularity condition if $|r_{g_m}^u(\beta)| = |r_{g_n}^u(\beta)|$. From (1) it is easy to see that any pair of predictors with indices in $\hat{\mathcal{A}}$ satisfies the equiangularity condition. At each transition point a new index is included in the active set; in other words, at a transition point there exists a predictor with parameter estimate zero satisfying the equiangularity condition.

The solution path can be constructed, in the following way. Let $\hat{\mathcal{A}} = \{g_0\}$. The maximum likelihood estimate of a generalized linear model specified with only the intercept term, denoted by $\hat{\beta}_{g_0}$, is used as the starting point of the solution curve, and then $|r_{g_m}^u(\hat{\beta}_{g_0})|$ are used to rank the p predictors. If $g_i = \arg \max_{g_m} |r_{g_m}^u(\hat{\beta}_{g_0})|$, then the index of the i th predictor is included in $\hat{\mathcal{A}}$ and $\gamma^{(1)}$ is equal to $N|r_{g_i}^u(\hat{\beta}_{g_0})|$. The first segment of the solution curve, denoted by $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma) = \{\hat{\beta}_{g_0}(\gamma), \hat{\beta}_{g_i}(\gamma)\}^T$, is implicitly defined by the equations

$$r_{g_0}^u(\beta_{\hat{\mathcal{A}}}) = 0, \quad r_{g_i}^u(\beta_{\hat{\mathcal{A}}}) - N^{-1} \gamma \text{sign}\{r_{g_i}^u(\hat{\beta}_{g_0})\} = 0.$$

This method traces the first segment of the solution curve, reducing γ until we find the transition point $\gamma^{(2)}$ corresponding to the inclusion of a new index in the active set, i.e., there exists a predictor, say the j th, satisfying the condition

$$|r_{g_j}^u\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})\}| = |r_{g_i}^u\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})\}| = \gamma^{(2)}/N,$$

where $g_j \notin \hat{\mathcal{A}}$. At this point g_j is included in $\hat{\mathcal{A}}$ and the new segment of the solution curve $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma) = \{\hat{\beta}_{g_0}(\gamma), \hat{\beta}_{g_i}(\gamma), \hat{\beta}_{g_j}(\gamma)\}^T$ is implicitly defined by the system obtained by adding the nonlinear equation $r_{g_j}^u(\beta_{\hat{\mathcal{A}}}) - N^{-1} \gamma \text{sign}[r_{g_j}^u\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})\}] = 0$ to the previous system. The solution curve is traced as previously described until $\gamma = 0$. When we are in a high-dimensional

setting, $p \gg N$, this method works by fixing the last value of the sequence of γ to a small positive value, to avoid overfitting of the model.

2.3. Group differential-geometric least angle regression method

To extend the method proposed in [Augugliaro et al. \(2013\)](#) to a generalized linear model with grouped predictors, the description given in § 2.2 shows that we need to generalize the equiangularity condition defined by the score test statistic and the nonlinear system (1).

The first element can be obtained using the differential-geometric structure of a generalized linear model with grouped predictors, which generalizes the group equiangularity condition proposed in [Yuan & Lin \(2006\)](#) for a linear regression model.

THEOREM 1. *If the differential-geometric representation of a generalized linear model with grouped predictors is used to generalize the Euclidean notion of angle, then the square root of the score test statistic for the m th group, denoted by $\omega_m^{1/2}(\beta)$, is proportional to the cosine of the angle between the tangent residual vector and its orthogonal projection onto the tangent space associated with the m th group of predictors.*

The proof and the necessary differential-geometric notions are reported in the Supplementary Material. Using the geometric characterization of $\omega_m^{1/2}(\beta)$ given in Theorem 1, we say that the m th and the n th groups of predictors satisfy the generalized group equiangularity condition when $\omega_m^{1/2}(\beta) = \omega_n^{1/2}(\beta)$. Following [Yuan & Lin \(2006\)](#), when the groups have different sizes we replace $\omega_q(\beta)$ with $\omega_q(\beta)/|g_q|$.

To define the system of nonlinear equations to obtain the solution curve, observe that

$$\omega_m(\beta)/N = \mathbb{P}_N \partial_{g_m} \ell(\beta)^\top \{\mathbb{P}_N I_{g_m g_m}(\beta)\}^{-1} \mathbb{P}_N \partial_{g_m} \ell(\beta).$$

Denoting by $\{\mathbb{P}_N I_{g_m g_m}(\beta)\}^{1/2}$ the Cholesky square root matrix of $\mathbb{P}_N I_{g_m g_m}(\beta)$, the weighted equiangularity condition can be expressed in terms of the Euclidean norm of the transformation

$$\zeta_{g_m}(\beta) = \{\mathbb{P}_N I_{g_m g_m}(\beta)\}^{-1/2} \mathbb{P}_N \partial_{g_m} \ell(\beta) / |g_m|^{1/2}. \quad (2)$$

Using (2), the proposed estimator can be formally defined as follows. For any dataset there exists a finite set of transition points, say $0 \leq \gamma^{(K)} \leq \dots \leq \gamma^{(1)}$, such that for any $\gamma^{(k+1)} < \gamma < \gamma^{(k)}$ the active set is defined as $\hat{\mathcal{A}} = \{g_m \in \mathcal{A} : \|\hat{\beta}_{g_m}(\gamma)\| > 0\}$ and the corresponding group differential-geometric least angle regression estimator is the Z-estimator defined by the system

$$\zeta_{\hat{\mathcal{A}}} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma)\} - N^{-1} \gamma v_{\hat{\mathcal{A}}} = 0, \quad (3)$$

with $v_{g_m} = \zeta_{g_m} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\} / \|\zeta_{g_m} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\|$, satisfying the further condition that $\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma)\}\| < \gamma/N$ for any $g_n \notin \hat{\mathcal{A}}$. As a consequence of the definition of the proposed estimator, we have that for any $\gamma^{(k+1)} < \gamma < \gamma^{(k)}$ the following conditions are satisfied for any $g_m \in \hat{\mathcal{A}}$ and any $g_n \notin \hat{\mathcal{A}}$:

$$\|\zeta_{g_m} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma)\}\| = \gamma/N, \quad \|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma)\}\| < \gamma/N. \quad (4)$$

At $\gamma^{(k+1)}$ there exists a group, say the n th, with $\|\hat{\beta}_{g_n}(\gamma^{(k+1)})\| = 0$, such that

$$\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k+1)})\}\| = \gamma^{(k+1)}/N.$$

Algorithm 1. Group differential-geometric least angle regression.

Start with the estimate of the intercept-only model.
 Include in $\hat{\mathcal{A}}$ the set of indices with largest $\|\zeta_{g_m}(\hat{\beta}_{g_0})\|$.
 Repeat the following steps until γ reaches a fixed value.
 Trace the path of solutions, reducing γ and solving system (3) until
 a new transition point is identified.
 Include in $\hat{\mathcal{A}}$ the set of indices of the group of predictors satisfying
 the weighted group equiangularity condition.

Algorithm 1 shows how the entire path of solutions can be constructed, generalizing the description given in § 2.2, i.e., let $\hat{\beta}_{g_0}$ be the maximum likelihood estimate of a generalized linear model specified only with the intercept term; then $\|\zeta_{g_m}(\hat{\beta}_{g_0})\|$ are used to rank the K groups. If $g_i = \arg \max_{g_m} \|\zeta_{g_m}(\hat{\beta}_{g_0})\|$, then g_i is included in $\hat{\mathcal{A}}$ and the first directional vector is $v_{g_i} = \zeta_{g_i}(\hat{\beta}_{g_0}) / \|\zeta_{g_i}(\hat{\beta}_{g_0})\|$. The first segment of the solution curve is implicitly defined by the system of nonlinear equations (3). This segment is traced, reducing γ until we find the transition point $\gamma^{(2)}$ corresponding to the inclusion of a new set of indices in $\hat{\mathcal{A}}$, i.e., at $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})$ there exists a group, say the j th, such that the weighted group equiangularity condition is satisfied:

$$\|\zeta_{g_j}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})\}\| = \|\zeta_{g_i}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})\}\| = \gamma^{(2)} / N.$$

Then, g_j is included in $\hat{\mathcal{A}}$ and $v_{g_j} = \zeta_{g_j}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})\} / \|\zeta_{g_j}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})\}\|$. The new segment of the solution curve is implicitly defined by the new nonlinear system (3). As described before, our method goes on until $\gamma = 0$ or γ reaches a fixed positive value when we are in a high-dimensional setting.

For a linear regression model, the proposed solution curve is equal to the solution curve obtained by using the groupwise prediction penalty function proposed in Bühlmann & van de Geer (2011, § 4.5.1). However, this is not true for the generalized linear models, since those authors do not use the differential-geometric structure of the exponential family to generalize the penalty function. When each group contains only one predictor, this method is that proposed in Augugliaro et al. (2013).

The differential-geometric least angle regression method is closely related to the implicit function theorem, which gives the condition for the existence and uniqueness of the continuous segments of the solution curve. In our case, this condition is that the Jacobian matrix of the vector-valued function $\zeta_{\hat{\mathcal{A}}}(\beta_{\hat{\mathcal{A}}})$ is nonsingular at least for any β belonging to the corresponding segment. In other words, zero is not a critical value of the function $H(\beta_{\hat{\mathcal{A}}}; \gamma) = \zeta_{\hat{\mathcal{A}}}(\beta_{\hat{\mathcal{A}}}) - \gamma v_{\hat{\mathcal{A}}} / N$. When we work with a generalized linear model specified using the variance-stabilizing transformation as link function, the Fisher information matrix is constant, which implies the nonsingularity of the Jacobian matrix for any β belonging to the parameter space. In the other cases, since $H(\beta_{\hat{\mathcal{A}}}; \gamma)$ is a C^2 function, Sard's theorem (Sard, 1942) tells us that its set of critical values has Lebesgue $|\hat{\mathcal{A}}|$ -dimension measure zero, i.e., any segment of the solution curve exists and is unique with probability one. For an extensive description of the role of Sard's theorem in numerical homotopy methods, the reader is referred to Allgower & Georg (2003, Ch. 11).

2.4. A comparison with the group lasso estimator

A comparison between the proposed estimator and the group lasso estimator can be done by comparing conditions (4) with the corresponding conditions for the group lasso estimator. For

a linear regression model, [Yuan & Lin \(2006, Proposition 1\)](#) show that for a given value of γ , different from the transition points, the following conditions are satisfied by the group lasso estimator:

$$\|\mathbb{P}_N \partial_{g_m} \ell(\beta)\| = \frac{\gamma}{N} |g_m|^{1/2}, \quad \|\mathbb{P}_N \partial_{g_n} \ell(\beta)\| < \frac{\gamma}{N} |g_n|^{1/2}, \quad (5)$$

for any $g_m \in \hat{\mathcal{A}}$ and any $g_n \notin \hat{\mathcal{A}}$. [Bühlmann & van de Geer \(2011, Lemma 4.2\)](#) show that conditions (5) are valid also for a generalized linear model with grouped predictors subject to a lasso constraint. Conditions (5) say that for a generalized linear model the behaviour of the group lasso estimator depends on the Euclidean norm of the blocks of the gradient vector. For linear regression models this behaviour is studied by [Simon & Tibshirani \(2012\)](#), who suggest that a more natural approach for the group lasso estimator is obtained by standardizing the blocks of the design matrix corresponding to the groups of predictors. This corresponds to a reparameterization of the linear regression model where the blocks of the conditional Fisher information matrix are identity matrices.

In a generalized linear model the standardization of the blocks of the design matrix is not sufficient to improve the behaviour of the group lasso estimator, since it does not remove the dependence of $\mathbb{P}_N I_{g_m g_m}(\beta)$ on the parameter vector. Instead, our estimator uses the transformation (2) to define the conditions determining its behaviour as a variable selection method. This means that the proposed method takes into account the conditional Fisher information and overcomes the parameterization problem of the group lasso method. Using Theorem 3 in [Augugliaro et al. \(2013\)](#), it is straightforward to see that conditions (5) are equal to (4) if and only if we are working with a generalized linear model specified using the variance-stabilizing transformation as a link function and, at the same time, the blocks of the design matrix are orthonormalized. The invariant nature of transformation (2) implies that the solution curve of our method is invariant with respect to any reparameterization within groups. This is an important advantage over the group lasso.

3. ASYMPTOTIC RESULTS

3.1. Adaptive group differential-geometric least angle regression

The behaviour of the lasso is related to the covariance structure of the p predictors. [Zou \(2006\)](#) and [Zhao & Yu \(2006\)](#) introduced the irrerepresentability condition, which shows that under conditions on the correlations between the predictors the lasso may always include incorrect predictors even if the number of observations goes to infinity. [Bach \(2008\)](#) showed that this is also the case for the group lasso. [Yuan & Lin \(2007, Theorem 3\)](#) show that least angle regression may also be inconsistent in variable selection. Theorem 2 generalizes this result to group differential-geometric least angle regression, showing that the behaviour of this method is related to the covariance structure among groups of false and true predictors and to the nonlinearity of the function $\mu(\cdot)$. However, we show how this can be overcome in an adaptive version of group differential-geometric least angle regression.

THEOREM 2. *Let $\tilde{\beta}_{g_0} = \arg \max_{\beta_{g_0}} E[\log p\{Y; \mu(\beta_{g_0})\}]$, let $\tilde{\beta} = (\tilde{\beta}_{g_0}, 0^T)^T$ be a $(p+1)$ -dimensional vector and $\tilde{\beta}_{\mathcal{A}^*}$ a value between $\beta_{\mathcal{A}^*}^*$ and $\tilde{\beta}_{\mathcal{A}^*}$. If*

$$\begin{aligned} \max_{g_n \notin \mathcal{A}^*} |g_n|^{-1} \left\| \Sigma_{g_n g_n}^{-1/2} E\{X_{g_n} \partial \mu(X_{\mathcal{A}^*}^T \tilde{\beta}_{\mathcal{A}^*}) / \partial \eta X_{\mathcal{A}^*}^T\} (\beta_{\mathcal{A}^*}^* - \tilde{\beta}_{\mathcal{A}^*}) \right\|^2 \\ \geq \max_{g_q \in \mathcal{A}^*} |g_q|^{-1} \left\| \Sigma_{g_q g_q}^{-1/2} E\{X_{g_q} \partial \mu(X_{\mathcal{A}^*}^T \tilde{\beta}_{\mathcal{A}^*}) / \partial \eta X_{\mathcal{A}^*}^T\} (\beta_{\mathcal{A}^*}^* - \tilde{\beta}_{\mathcal{A}^*}) \right\|^2, \quad (6) \end{aligned}$$

where $\tilde{\beta}_{A^*}$ is a value between $\beta_{A^*}^*$ and $\tilde{\beta}_{A^*}$, then the first group of predictors included in the active set is wrong with probability tending to one.

The proof of this theorem is given in the Supplementary Material. Several authors have proposed to use a two-stage method to overcome the irrepresentability condition for the lasso estimator. An important example is the adaptive lasso (Zou, 2006). The main idea underlying this estimator is to use a $N^{1/2}$ -consistent estimator to remove the irrepresentability condition and to recover variable selection consistency. The extension of the adaptive lasso to the group lasso is proposed in Nardi & Rinaldo (2008) and Wei & Huang (2010). Here, we extend the approach proposed in Zou (2006) by using a weighted version of the differential-geometric representation of a generalized linear model. In the following we replace γ with γ_N in order to emphasize its dependence on the sample size N .

Let $\hat{\beta}$ be an initial estimate based on a $N^{1/2}$ -consistent estimator, such as the maximum likelihood estimator. Then, using the relationship between the maximum likelihood estimator and our estimator, it can be shown that $\hat{\beta}$ is the last point of the group differential-geometric least angle regression solution curve. Consider the straight line $\hat{\beta}_{g_m}(\tau) = \tau \hat{\beta}_{g_m}$ and the corresponding tangent vector $w_m = N^{1/2} \hat{\beta}_{g_m}^T \mathbb{P}_N \partial_{g_m} \ell(\hat{\beta})$. Let $\beta_{g_m}^*$ be the true $(p+1)$ -dimensional parameter vector in which the parameters corresponding to the m th group of predictors are equal to zero and let $\mathcal{KL}(\beta^*, \beta_{g_m}^*)$ be the Kullback–Leibler divergence between $p\{y; \mu(\beta^*)\}$ and $p\{y; \mu(\beta_{g_m}^*)\}$. Under the regularity assumptions reported in the Supplementary Material, $\|w_m\|_{p\{\mu(\beta)\}}^2 = \hat{\beta}_{g_m}^T \{\mathbb{P}_N I_{g_m g_m}(\hat{\beta})\} \hat{\beta}_{g_m}$ converges in probability to $\beta_{g_m}^{*T} \mathcal{I}_{g_m g_m}(\beta^*) \beta_{g_m}^* \approx 2\mathcal{KL}(\beta^*, \beta_{g_m}^*)$. Then, we can define the weights

$$\hat{v}_m = \hat{\beta}_{g_m}^T \{\mathbb{P}_N I_{g_m g_m}(\hat{\beta})\} \hat{\beta}_{g_m} / (2|g_m|), \quad (7)$$

which can be seen as an estimate of the amount of information lost when we incorrectly set $\beta_{g_m}^*$ equal to zero. The factor $|g_m|$ is used in order to deal with different group sizes. The invariant nature of the tangent vectors on which the weights (7) are based implies that $\hat{v}_m$ are invariant to reparameterization of the β_{g_m} . We stress that the weights proposed in Zou (2006), Nardi & Rinaldo (2008), Wei & Huang (2010) or Bühlmann & van de Geer (2011, § 4.6) do not have this property. The adaptive group differential-geometric least angle regression is defined using the adaptive version of conditions (4), namely

$$\|\zeta_{g_m}\{\hat{\beta}(\gamma_N)\}\| = \gamma_N / (N\hat{v}_m), \quad \|\zeta_{g_n}\{\hat{\beta}(\gamma_N)\}\| < \gamma_N / (N\hat{v}_n),$$

for any $g_m \in \hat{A}$ and for any $g_n \notin \hat{A}$. Theorem 3 shows that this adaptive method satisfies the oracle properties of Fan & Li (2001) when the number of predictors p is fixed.

THEOREM 3. *If p is fixed, $\gamma_N \rightarrow \infty$ and $\gamma_N = o(N^{1/2})$, the adaptive group differential-geometric least angle regression estimator satisfies:*

- (i) *consistency in variable selection, $\lim_{N \rightarrow \infty} pr\{\hat{A}(\gamma_N) = A^*\} = 1$;*
- (ii) *asymptotic normality, $N^{1/2}\{\hat{\beta}_{A^*}(\gamma_N) - \beta_{A^*}^*\} \rightarrow N\{0; \mathcal{I}_{A^* A^*}^{-1}(\beta_{A^*}^*)\}$ in distribution as $N \rightarrow \infty$.*

The proof of this theorem and the regularity conditions are given in the Supplementary Material. In a linear regression model with no group structure, i.e., $g_m = \{m\}$, and under the assumption that $\mathbb{P}_N x_m = 1$, the weights (7) are equal to $\hat{v}_m = \hat{\beta}_m^2$, corresponding to what is proposed in

Fig. 1. Results of the simulation study. Panel (a) shows the behaviour of the proportion of true models identified by the proposed adaptive estimator. Circles refer to results with $c = 40$ and triangles to results with $c = 60$. Panels (b) and (c) show the quantile-quantile plots of the simulated distributions of the two adaptive estimators with respect to the asymptotic normal distribution postulated in Theorem 3.

Zou (2006) when $\gamma = 2$. When Zou (2006) studied the behaviour of the adaptive lasso for different values of the parameter γ , it is interesting that $\gamma = 2$ seems always to be the best choice.

3.2. Simulation study

We simulate a Poisson regression model with 25 groups of predictors. The size of each group is equal to 2 and each predictor has a standard normal distribution. Only the first group is used to simulate the response variable and the corresponding nonzero parameters are set equal to 1. To study the asymptotic properties reported in Theorem 3 we let $\gamma_N = c \log N$ with $c = (40, 60)$ and $N = (100, 150, \dots, 500)$. The results are based on 5000 simulation runs. In any simulation run the weights (7) are computed using the maximum likelihood estimator of the parameter vector.

Fig. 1(a) shows the proportion of the true models identified by the adaptive group differential-geometric least angle estimator as a function of the sample size. This is closely related to the parameter c , i.e., the proportion is always an increasing function of the sample size but the convergence to its asymptotic value is improved when we increase c . When $c = 60$, under our setting, a sample size larger than 200 is sufficient to obtain $\hat{\text{pr}}\{\hat{\mathcal{A}}(\gamma_N) = \mathcal{A}^*\}$ close to 1. Figure 1 (b) and (c) report the asymptotic behaviour of the proposed adaptive estimator of the first nonzero parameter and the second nonzero parameter, respectively, using the results from the simulation study with $c = 60$ and $N = 500$. Since the asymptotic covariance matrix postulated in part (ii) of Theorem 3 cannot be computed in closed form, we estimated it by $\{\mathbb{P}_N I_{\mathcal{A}^* \mathcal{A}^*}(\beta_{\mathcal{A}^*}^*)\}^{-1}$ with $N = 10^7$. As we can see in Figs. 1(b) and (c), the simulated distributions of the two estimators are close to the normal distribution postulated in Theorem 3.

4. COMPUTATIONAL ASPECTS

To compute the solution curve we must find an efficient numerical method to approximate the segments of the solution curve and a method to identify the transition points.

Suppose that we have traced the first $k - 1$ segments of the solution curve and that we know the transition points $\gamma^{(k+1)}$ and $\gamma^{(k)}$. Under this setting, the first subproblem can be formalized as follows: defining a discrete set of nodes, say $\gamma^{(k+1)} = \gamma_J^{(k)} < \dots < \gamma_1^{(k)} = \gamma^{(k)}$, we need to compute the corresponding points lying on the solution curve. Then, the k th segment is obtained by interpolating the computed points.

Suppose that we have computed $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_{j-1}^{(k)})$. Differentiating system (3) with respect to γ , we can approximate $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)})$ using Euler's predictor formula (Allgower & Georg, 2003), i.e.,

$$\tilde{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)}) = \hat{\beta}_{\hat{\mathcal{A}}}(\gamma_{j-1}^{(k)}) - \frac{(\gamma_{j-1}^{(k)} - \gamma_j^{(k)})}{N} \left[\frac{\partial \zeta_{\hat{\mathcal{A}}} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_{j-1}^{(k)})\}}{\partial \beta_{\hat{\mathcal{A}}}} \right]^{-1} v_{\hat{\mathcal{A}}}, \quad (8)$$

where $\partial \zeta_{\hat{\mathcal{A}}} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_{j-1}^{(k)})\} / \partial \beta_{\hat{\mathcal{A}}}$ is the Jacobian matrix of the vector-valued function $\zeta_{\hat{\mathcal{A}}}(\beta_{\hat{\mathcal{A}}})$ at the point $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_{j-1}^{(k)})$. To compute $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)})$ we use approximation (8) to solve the nonlinear system

$$\zeta_{\hat{\mathcal{A}}}(\beta_{\hat{\mathcal{A}}}) - N^{-1} \gamma_j^{(k)} v_{\hat{\mathcal{A}}} = 0, \quad (9)$$

using a Newton–Raphson algorithm. The entire algorithm is known as the predictor-corrector algorithm, used to numerically trace a curve implicitly defined by a system of nonlinear equations (Allgower & Georg, 2003). The predictor-corrector schema has a trade-off between accuracy of the approximation of the k th segment and the computational burden of the global algorithm. In other words, to improve the approximation of the solution curve we need to reduce the step size $(\gamma_{j-1}^{(k)} - \gamma_j^{(k)})$, but this increases the computational complexity of the algorithm, which anyway can be significantly reduced in many practical applications using the values of the solution curve at the transition points, i.e., $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})$.

This brings us to the problem of how to approximate the step size $\Delta \gamma^{(k)} = \gamma^{(k)} - \gamma^{(k+1)}$, a method for which was proposed in Park & Hastie (2007) for ℓ_1 -penalized generalized linear models. In our case, to approximate $\Delta \gamma^{(k)}$ we consider that in $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)} - \Delta \gamma^{(k)})$ there is a change in the active set, i.e., there exists a nonactive group, say the n th, such that

$$\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)} - \Delta \gamma^{(k)})\}\| = \|\zeta_{g_m} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)} - \Delta \gamma^{(k)})\}\| = (\gamma^{(k)} - \Delta \gamma^{(k)})/N, \quad (10)$$

for any $g_m \in \hat{\mathcal{A}}$. Using a first-order approximation

$$\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)} - \Delta \gamma^{(k)})\}\| \approx \|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\| - \frac{d\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\|}{d\gamma} \Delta \gamma^{(k)},$$

condition (10) can be locally approximated by

$$\left| \|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\| - \frac{d\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\|}{d\gamma} \Delta \gamma^{(k)} \right| \approx \frac{\gamma^{(k)} - \Delta \gamma^{(k)}}{N},$$

which leads to

$$\Delta \gamma^{(k)} \approx \min_{g_n \notin \hat{\mathcal{A}}}^+ \left[\frac{\gamma^{(k)} - N\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\|}{1 - Nd\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\|/d\gamma}; \frac{\gamma^{(k)} + N\|\zeta_n \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\|}{1 + Nd\|\zeta_{g_n} \{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}\|/d\gamma} \right], \quad (11)$$

where $\min^+(\cdot)$ indicates that the minimum is taken over only positive components. The proposed algorithm can be described as follows. As we have seen in §2.3, the $\|\zeta_{g_m}(\hat{\beta}_{g_0})\|$ are used to locally rank the K groups. If $g_i = \arg \max_{g_m} \|\zeta_{g_m}(\hat{\beta}_{g_0})\|$, then g_i is included in $\hat{\mathcal{A}}$, $v_{g_i} = \zeta_{g_i}(\hat{\beta}_{g_0}) / \|\zeta_{g_i}(\hat{\beta}_{g_0})\|$ and $\gamma_1^{(1)} = N\|\zeta_{g_i}(\hat{\beta}_{g_0})\|$. Using (11) we approximate the transition point $\gamma^{(2)}$ by $\gamma_2^{(1)} = \gamma_1^{(1)} - \Delta \gamma^{(1)}$ and by (8) we compute the corresponding Euler's predictor $\tilde{\beta}_{\hat{\mathcal{A}}}(\gamma_2^{(1)})$ which will be used as the starting point to compute $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_2^{(1)})$. If there exists a nonactive

group, say the j th, such that $\|\zeta_{g_j}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_2^{(1)})\}\| = \|\zeta_{g_i}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_2^{(1)})\}\| = \gamma_2^{(1)}/N$, then g_j is included in $\hat{\mathcal{A}}$ and $\gamma^{(2)} = \gamma_2^{(1)}$; otherwise the previous steps are repeated using $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_2^{(1)})$ as the new starting point. This procedure is repeated until $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(2)})$ is computed. When the approximation (11) exceeds the corresponding transition point, the step size is reduced. The proposed algorithm computes the entire solution curve, repeating the previous procedure until the last value of the solution curve is reached. The proposed algorithm is reported in Algorithm 2 and implemented in the R package (R Development Core Team, 2016) `gdglars`. In this package the user can always choose whether to use the approximation (11) or a small constant as the step size. In this way the user can always improve the approximation to the solution curve when necessary.

We emphasize that the k th segment of the solution curve can equivalently be defined as the solution of a system of ordinary differential equations with initial condition specified by $\hat{\beta}_{\hat{\mathcal{A}}}(\gamma^{(k)})$, so that the algorithm proposed in Wu (2011), and extended in Wu (2012), Zhou & Wu (2014) and Xiao et al. (2015), can be easily adapted to obtain our estimator. This algorithm has high flexibility since we can use, for example, the algorithms based on the Runge–Kutta formulae (Dormand & Prince, 1980) to trace any segment of the solution curve.

Algorithm 2. Pseudo-code of the proposed algorithm.

```

Set  $\hat{\beta}_{g_0} = \arg \max_{\beta_{g_0}} \mathbb{P}_N \ell(\beta_{g_0})$ 
Set  $\hat{\mathcal{A}} = \{g_0\}$ 
Set  $\gamma_1^{(1)} = N \max_{g_m} \|\zeta_{g_m}(\hat{\beta}_{g_0})\|$ 
For  $k = 1$  to  $k = K$ 
  Set  $j = 1$ 
  Set  $g_m = \arg \max_{g_n \notin \hat{\mathcal{A}}} \|\zeta_{g_n}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_1^{(k)})\}\|$ 
  Set  $v_{g_m} = \zeta_{g_m}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_1^{(k)})\} / \|\zeta_{g_m}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_1^{(k)})\}\|$ 
  Include  $g_m$  in  $\hat{\mathcal{A}}$ 
  Repeat
    Use (11) to compute  $\Delta\gamma^{(k)}$ 
     $j \leftarrow j + 1$ 
    Repeat
       $\gamma_j^{(k)} \leftarrow \gamma_{j-1}^{(k)} - \Delta\gamma^{(k)}$ 
      Use (8) to compute  $\tilde{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)})$ 
      Use  $\tilde{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)})$  as a starting point to solve the system (9)
      If  $\max_{g_n \notin \hat{\mathcal{A}}} \|\zeta_{g_n}\{\tilde{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)})\}\| > \gamma_j^{(k)}/N$ 
         $\Delta\gamma^{(k)} \leftarrow c\Delta\gamma^{(k)}$ , where  $c$  is a contractor factor
      Else exit from Repeat
    If  $\max_{g_n \notin \hat{\mathcal{A}}} \|\zeta_{g_n}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)})\}\| = \gamma_j^{(k)}/N$ 
      Set  $\gamma_1^{(k+1)} = N \max_{g_n \notin \hat{\mathcal{A}}} \|\zeta_{g_n}\{\hat{\beta}_{\hat{\mathcal{A}}}(\gamma_j^{(k)})\}\|$ 
      Exit from Repeat
  
```

5. SIMULATIONS

5.1. Time comparison

We compare the run times of the predictor-corrector algorithm described in § 4 with the block coordinate gradient descent algorithm proposed in Meier et al. (2008) and available in the R package *grplasso*. Algorithm 2 has been implemented in FORTRAN and is called in R using wrapper functions. Our simulation study is based on a logistic regression model with the following scenarios:

- (a) $K = 2, 4, 6, 8, 10$ and $|g_m| = 2, 4$ for any group of predictors. The random vector X is sampled from a multivariate normal distribution with zero mean and $\text{cov}(X_i, X_j) = 0.9^{|i-j|}$. The binary response is simulated using only the first two groups and the corresponding nonzero elements of the parameter vector are fixed to 0.1.
- (b) We first simulate a K -dimensional latent variable, denoted by W , from a multivariate normal distribution with zero mean and $\text{cov}(W_i, W_j) = 0.5^{|i-j|}$. As in scenario (a) we set $K = 2, 4, 6, 8, 10$. We then obtain two different subscenarios transforming W_i into two categorical random variables using the tertiles and quartiles of the standard normal distribution, respectively. In any subscenario each element of the parameter vector is fixed to 0.2, i.e., we do not have a sparse group structure.

In order to evaluate the effect of the sample size, we set $N = 200, 400, 600$. The two algorithms use different criteria to compute the sequence of tuning parameters: our proposed algorithm uses expression (11) to compute the step size, while the block coordinate gradient descent algorithm uses a fixed grid. To make the run times comparable, we first trace the solution curve of our method and then compute the group lasso solution curve using a sequence of equispaced values of the tuning parameter: this sequence has the same number of values as used for our method. The last value of the two sequences is set to zero. For each scenario the average processor time used to trace the path of solutions is computed over 100 simulation runs.

The complete results, reported in the Supplementary Material, show that, in the scenarios considered, the proposed algorithm is always faster than that of Meier et al. (2008). Increasing the sample size increases the run times, and this effect seems to be larger with highly correlated predictors, as in scenario (a). Computing time also increases when we double the size of each group, but this effect seems to be less pronounced for logistic regression models with sparse design matrices, as in scenario (b).

5.2. Group selection behaviour

We compare the group selection behaviour of the proposed method to that of the group lasso. The simulation scheme is based on a logistic regression model with the following scenarios:

- (a) $K = 20, 60, 100$ and $|g_m| = 2$, for any group of predictors. The i th predictor of the m th group is simulated according to the following model: $X_i = W_m + \varepsilon_i$, where ε_i has a standard normal distribution and the K -dimensional random vector W is sampled from a multivariate normal distribution with zero mean and covariance matrix Σ with elements $\text{cor}(X_i, X_j) = \rho_w$, for any $i, j \in g_m$, and $\text{cor}(X_i, X_h) = \rho_b$, for $i \in g_m$ and $h \in g_n$. We set: $\rho_w = 0.6$ and $\rho_b = 0.5, 0.3, 0.1$; $\rho_w = 0.4$ and $\rho_b = 0.3, 0.1$; $\rho_w = 0.2$ and $\rho_b = 0.1$, respectively. This scheme ensures that Σ is positive definite; only the first three groups are used to generate the binary response variable, while the sample size is 100. Each nonzero element of the parameter vector β^* is sampled from a normal distribution with mean 1 and standard deviation 0.1.

- (b) $K = 20, 40, 60$ and the size $|g_m| = 2, 5$. The p -dimensional random vector X is sampled from a normal distribution with zero mean. A Toeplitz structure is used for Σ . As in scenario (a), the first three groups are used to simulate the response variable, the sample size is 100 and each nonzero element of β^* is sampled from a normal distribution with mean 1 and standard deviation 0.1.
- (c) This scenario is similar to that of [Yuan & Lin \(2006\)](#) and [Meier et al. \(2008\)](#). We simulate a 9-dimensional latent random variable, denoted by W , from a multivariate normal distribution with mean 0 and $\text{cov}(W_i, W_j) = \rho^{|i-j|}$. We set $\rho = (0.5, 0.3, 0.1)$. Then W_i is transformed into a three-valued categorical random variable, denoted by X_i , using the tertiles of the standard normal distribution. The treatment contrast is used as an encoding scheme. The binary response variable is simulated according to the true models: $X_1X_2 + X_3X_4$; $X_1X_2 + X_3X_4 + X_5X_6$. The sample size is 500 and each nonzero coefficient is sampled from a normal distribution with mean 4 and standard deviation 0.1. As in [Meier et al. \(2008\)](#), the design matrix used for the group logistic lasso and for our method is obtained from a model with all the two-way interaction terms.

For the tuning parameter of the logistic group lasso we used a grid with 100 equally spaced points ranging from $\lambda_{\max}$ to 0.05. The value of $\lambda_{\max}$ is computed as suggested in [Meier et al. \(2008\)](#). A ten-fold crossvalidation deviance is used to select the optimal tuning parameter for the two methods. For each scenario we simulated 200 training sets, used to estimate the solution curves, and 200 test sets, to evaluate the prediction behaviour of the methods; see the Supplementary Material. To compare the two methods, we use the mean number of groups included in the final model, the mean false discovery rate, the mean false positive rate and the mean false negative rate. The median deviance is used to evaluate the prediction behaviour. The results are summarized in Table 1.

The behaviour of group differential-geometric least angle regression is closely related to the covariance structure of the groups of predictors. For scenario (a), the difference between the mean number of groups of predictors selected by the group lasso and by our method increases as the number of groups K increases. Our method tends to select sparser models than group lasso and this is even more marked in a high-dimensional setting, $K = 100$. However, our method does not have a higher mean false discovery rate or mean false positive rate; these rates are significantly lower than for the group lasso. The mean deviance values show that the proposed method can give the prediction behaviour of the group lasso, but with a smaller number of groups. The results for the two adaptive methods show that the weights used in the adaptive group lasso give only a slight improvement on its global behaviour; see also [Wei & Huang \(2010\)](#). The adaptive version of our method significantly reduces the dimension of the selected model. This is associated with a lower false discovery rate and a slightly higher false negative rate. The same results are confirmed for the two scenarios (b) and (c), when we consider the adaptive versions of the two methods.

6. APPLICATION

Alzheimer's disease, the most common form of dementia in elderly persons, has a strong genetic component. Accumulation and aggregation of beta-amyloid peptide is hypothesized to be its primary cause. [Shi et al. \(2005\)](#) focused their attention on neprilysin, also known as neutral endopeptidase, a putative beta-amyloid peptide-degrading enzyme, which is thought to play an important role in Alzheimer's disease pathogenesis because its decreased expression and/or activity may result in cerebral beta-amyloid peptide accumulation. Their study is based on a sample of 257 Alzheimer's disease patients and 242 age-matched controls. We use the proposed method to investigate the association of eight single nucleotide polymorphisms with Alzheimer's disease.

Table 1. Results from the simulation study with setting $\rho_w = 0.6$ and $\rho_b = 0.5$ for scenario (a), with $|g_m| = 5$ and $\rho = 0.6$ for scenario (b), and with $\rho = 0.5$ for scenario (c). The number of simulation runs is 200. Median absolute deviation is reported in parentheses for the median deviance while standard deviation is reported for the other measures

Scenario		Method	Size	FDR ($\times 10^2$)	FPR ($\times 10^2$)	FNR ($\times 10^2$)	Dev ($\times 10$)
(a)	$K = 100$	gdglars	10.9 (2.8)	72.3 (9.3)	8.1 (3.3)	5.6 (13.3)	6.0 (1.4)
		group lasso	17.5 (4.0)	83.3 (4.6)	15.3 (4.1)	5.8 (12.9)	5.8 (1.2)
		adp-gdglars	4.2 (1.5)	35.9 (24.7)	1.8 (1.6)	18.8 (22.1)	5.9 (2.4)
		adp-group lasso	17.2 (3.9)	82.7 (4.8)	14.9 (4.1)	5.7 (12.6)	7.3 (2.6)
	$K = 60$	gdglars	10.1 (2.6)	68.9 (8.8)	12.8 (4.7)	2.2 (8.1)	5.9 (1.1)
		group lasso	15.6 (3.6)	79.9 (5.1)	22.1 (6.4)	2.4 (7.2)	5.9 (1.0)
		adp-gdglars	3.9 (1.4)	29.4 (22.9)	2.4 (2.2)	16.7 (20.9)	6.0 (2.1)
		adp-group lasso	14.5 (3.6)	78.7 (6.2)	20.3 (6.4)	3.7 (10.4)	2.1 (2.1)
	$K = 20$	gdglars	7.7 (2.2)	58.4 (12.0)	27.7 (13.1)	0.7 (4.7)	5.0 (1.0)
		group lasso	10.5 (2.3)	70.2 (7.3)	43.6 (14.3)	0.5 (4.1)	5.1 (1.0)
		adp-gdglars	3.1 (1.3)	15.4 (18.9)	3.9 (5.4)	17.0 (22.7)	6.4 (3.7)
		adp-group lasso	9.9 (2.5)	67.7 (9.4)	40.9 (14.9)	0.5 (4.1)	5.3 (1.5)
(b)	$K = 60$	gdglars	9.1 (2.3)	64.3 (11.2)	11.2 (3.9)	0.3 (0.1)	6.9 (1.1)
		group lasso	17.2 (4.9)	80.6 (8.2)	24.9 (8.7)	0.1 (0.2)	7.2 (1.2)
		adp-gdglars	3.1 (0.4)	2.9 (8.4)	0.2 (0.6)	1.2 (6.1)	5.8 (1.6)
		adp-group lasso	17.1 (4.9)	80.4 (8.2)	24.8 (8.6)	0.2 (0.1)	8.4 (2.4)
	$K = 40$	gdglars	8.5 (2.0)	61.8 (10.6)	14.6 (5.4)	0.2 (0.4)	6.7 (1.1)
		group lasso	15.1 (3.7)	78.6 (7.4)	32.8 (10.0)	0.1 (0.1)	6.9 (1.1)
		adp-gdglars	3.1 (0.4)	3.6 (10.0)	0.4 (1.2)	0.7 (4.7)	5.7 (1.5)
		adp-group lasso	15.1 (3.7)	78.4 (7.4)	32.6 (9.9)	0.1 (0.1)	8.0 (2.2)
	$K = 20$	gdglars	7.5 (1.9)	56.9 (13.1)	25.6 (11.3)	0.1 (0.1)	6.5 (1.0)
		group lasso	11.3 (2.7)	71.6 (7.6)	48.9 (15.7)	0.1 (0.1)	6.7 (1.0)
		adp-gdglars	3.1 (0.3)	2.3 (7.9)	0.6 (1.9)	0.7 (4.7)	5.7 (1.3)
		adp-group lasso	11.2 (2.5)	71.6 (7.6)	48.1 (14.7)	0.1 (0.1)	1.5 (1.5)
(c)	Model 1	gdglars	9.5 (3.0)	53.1 (15.9)	3.5 (2.0)	33.2 (1.2)	1.8 (0.3)
		group lasso	10.9 (3.7)	58.4 (15.2)	4.4 (2.4)	33.3 (0.1)	1.8 (0.3)
		adp-gdglars	4.0 (0.9)	4.0 (10.8)	0.2 (0.5)	36.9 (9.0)	2.2 (1.1)
		adp-group lasso	10.9 (3.7)	58.3 (15.1)	4.3 (2.4)	33.3 (0.1)	1.8 (0.3)
	Model 2	gdglars	9.0 (1.7)	31.2 (13.1)	2.0 (1.1)	33.7 (2.7)	0.9 (0.3)
		group lasso	11.0 (2.1)	43.0 (12.7)	3.3 (1.4)	33.4 (2.2)	0.8 (0.2)
		adp-gdglars	5.8 (1.0)	5.8 (10.1)	0.3 (0.5)	39.6 (7.7)	0.9 (0.3)
		adp-group lasso	10.9 (2.1)	42.9 (12.7)	3.2 (1.4)	33.6 (2.6)	0.8 (0.2)

gdglars, group differential-geometric least angle regression; adp-gdglars, adaptive group differential-geometric least angle regression; adp-group lasso, adaptive group lasso; Size, average number of groups included in the final model; FDR, average false discovery rate; FPR, average false positive rate; FNR, average false negative rate; Dev, median deviance.

Each is modelled by a factor with three levels and Alzheimer's disease status is modelled by a factor with two levels representing disease and control. Then, the dataset is summarized by a $3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 2$ contingency table, which can be studied by a log-linear model with first-order interaction effects. The minimum value of γ was fixed to 0.5 and the Bayesian information criterion was used to identify the best model.

Figure 2 shows the estimated paths of the square root of the score statistic. At $\gamma^{(1)} = 12.37$ the factor c.401A/G is included in the model. At $\gamma^{(2)} = 11.72$ the path corresponding to the group $-204G \rightarrow C$ intersects the path of the first group included in $\hat{A}$, i.e., at $\gamma^{(2)}$ the first generalized group equiangularity condition is satisfied. For any $\gamma \in (\gamma^{(3)}; \gamma^{(2)}]$ the group

Fig. 2. Paths of the square root of the score statistic for the Alzheimer's disease dataset. The vertical dashed line identifies the optimal value of the tuning parameter selected using the Bayesian information criterion, while the vertical dotted lines identify the transition points. Paths corresponding to the significant effects are black while grey dashed lines correspond to effects that are not included in the final model.

generalized equiangularity condition is always satisfied and this is shown in Fig. 2 by a solid black line. At $\gamma^{(3)} = 11.19$ a new group satisfies the group equiangularity condition and it is included in the model. The method proceeds until the minimum value of γ is achieved.

Similar to the results in Shi et al. (2005), our method shows that $-204G \rightarrow C$ is significantly associated with Alzheimer's disease. Figure 2 also shows that the interaction between $IVS17-294C \rightarrow T$ and $IVS22+36C \rightarrow A$ is related to Alzheimer's disease. This result is also observed in Shi et al. (2005), but they report that the corresponding single nucleotide polymorphisms seemed unlikely to influence neprilysin expression level or activity, because both of them are located in introns and are 36 bp or approximately 300 bp away from the splice sites. The authors observe that both single nucleotide polymorphisms are in linkage disequilibrium with other Alzheimer's disease risk variants within or near the single nucleotide polymorphism gene, such as $c.401A \rightarrow G$, which causes $Gln \rightarrow Arg$ mutation.

7. DISCUSSION

The proposed geometrical setting can easily be extended to other models, such as those based on quasilielihood or Cox's proportional hazard model. Our method suffers from the lack of selection consistency of least angle regression (Yuan & Lin, 2007), so we have extended the geometrical derivation of the weights proposed in Zou (2006) by means of the Kullback–Leibler divergence and defined an adaptive version of our method, which has oracle selection properties. The proposed methods compare favourably with the group lasso and with its adaptive extension. Although the asymptotic results given in § 3 do not cover the high-dimensional setting, $p \gg N$, the results of the simulation studies suggest that a zero-consistent estimator can be used to define the adaptive version of the proposed method, even when $p \gg N$.

ACKNOWLEDGEMENT

The authors thank the associate editor and the referees for comments that have improved the paper.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes the proofs of the theorems and the complete simulation results.

REFERENCES

- ALLGOWER, E. L. & GEORG, K. (2003). *Introduction to Numerical Continuation Methods*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.
- AUGUGLIARO, L., MINEO, A. M. & WIT, E. C. (2013). dglars: a differential geometric approach to sparse generalized linear models. *J. R. Statist. Soc. B* **75**, 471–98.
- BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9**, 1179–225.
- BÜHLMANN, P. & VAN DE GEER, S. (2011) *Statistics for High-Dimensional Data*. Berlin: Springer.
- DORMAND, J. R. & PRINCE, P. J. (1980). A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.* **6**, 19–26.
- EFRON, B., HASTIE, T. J., JOHNSTONE, I. M. & TIBSHIRANI, R. J. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407–99.
- FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.
- HUANG, J., BREHENY, P. & MA, S. (2012). A selective review of group selection in high dimensional models. *Statist. Sci.* **27**, 481–99.
- HUANG, J. & ZHANG, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38**, 1978–2004.
- MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall, 2nd ed.
- MEIER, L., VAN DE GEER, S. & BÜHLMANN, P. (2008). The group lasso for logistic regression. *J. R. Statist. Soc. B* **70**, 53–71.
- NARDI, Y. & RINALDO, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Statist.* **2**, 605–33.
- PARK, M. Y. & HASTIE, T. J. (2007). l_1 -regularization path algorithm for generalized linear models. *J. R. Statist. Soc. B* **69**, 659–77.
- R DEVELOPMENT CORE TEAM. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- SARD, A. (1942). The measure of the critical values of differentiable maps. *Bull. Am. Math. Soc.* **48**, 883–90.
- SHI, J., ZHANG, S., TANG, M., MA, C., ZHAO, J., LI, T., LIU, X., SUN, Y., GUO, Y., HAN, H. ET AL. (2005). Mutation screening and association study of the neprilysin gene in sporadic Alzheimer's disease in Chinese persons. *J. Gerontol. A: Biol. Sci. Med. Sci.* **60**, 301–6.
- SIMON, N. & TIBSHIRANI, R. J. (2012). Standardization and the group lasso penalty. *Statist. Sinica* **22**, 983–1001.
- TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–88.
- WEI, F. & HUANG, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16**, 1369–84.
- WU, Y. (2011). An ordinary differential equation based solution path algorithm. *J. Nonparam. Statist.* **23**, 185–99.
- WU, Y. (2012). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Statist. Sci.* **22**, 271–94.
- XIAO, W., WU, Y. & ZHOU, H. (2015). Convexlar: An extension of least angle regression. *J. Comp. Graph. Statist.* **24**, 603–26.
- YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* **68**, 49–67.
- YUAN, M. & LIN, Y. (2007). On the non-negative garrotte estimator. *J. R. Statist. Soc. B* **69**, 143–61.
- ZHANG, C. & HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–94.
- ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.
- ZHOU, H. & WU, Y. (2014). A generic path algorithm for regularized statistical estimation. *J. Am. Statist. Assoc.* **109**, 686–99.
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[Received November 2013. Revised May 2016]