Stat Comput
DOI 10.1007/s11222-014-9497-6

On the robustness of the generalized fused lasso to prior specifications

Vivian Viallon · Sophie Lambert-Lacroix ·
Hölger Hoefling · Franck Picard

Received: 12 December 2013 / Accepted: 16 July 2014

Abstract Using networks as prior knowledge to guide model selection is a way to reach structured sparsity. In particular, the fused lasso that was originally designed to penalize differences of coefficients corresponding to successive features has been generalized to handle features whose effects are structured according to a given network. As any prior information, the network provided in the penalty may contain misleading edges that connect coefficients whose difference is not zero, and the extent to which the performance of the method depend on the suitability of the graph has never been clearly assessed. In this work we investigate the theoretical and empirical properties of the adaptive generalized fused lasso in the context of generalized linear models. In the fixed p setting, we show that, asymptotically, adding misleading edges in the graph does not prevent the adaptive generalized fused lasso from enjoying asymptotic oracle properties, while forgetting suitable edges can be more problematic. These theoretical results are complemented by an extensive simulation study that assesses the robustness of the adaptive generalized fused lasso against misspecification of the network as well as its applicability when theoretical coefficients are not exactly equal. Our contribution is also to evaluate the applicability of the generalized fused lasso for the joint modeling of multiple sparse regression functions. Illustrations are provided on two real data examples.

Keywords Lasso · Generalized linear models · Joint modeling · Model selection

Electronic supplementary material The online version of this article (doi:10.1007/s11222-014-9497-6) contains supplementary material, which is available to authorized users.

V. Viallon (✉)
Université de Lyon, 69622 Lyon, France
e-mail: vivian.viallon@univ-lyon1.fr

V. Viallon
Université Lyon 1, UMRESTTE, 69373 Lyon, France

V. Viallon
IFSTTAR, UMRESTTE, 69675 Bron, France

S. Lambert-Lacroix
UMR 5525 UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG,
38041 Grenoble, France

H. Hoefling
Novartis Pharma, Basel, Switzerland

F. Picard
Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS
5558 Univ. Lyon 1, 69622 Villeurbanne, France

1 Introduction

Network data have now become standard in many fields of statistical applications to describe interactions or relationships between observations and/or variables. In molecular biology for instance, protein-protein interaction networks describe physical interactions between proteins (Franceschini et al. 2013). In web analysis studies, networks are extracted from structured data-bases and Web contents to describe hidden structures (Han 2011). Network data have become so common that they now constitute some prior knowledge for downstream statistical analysis: two proteins of the same biological pathway are likely to share similar effects on the response to a treatment or on disease development. Consequently, statistical methods, like regression and model selection have recently focused on structured sparsity. In addition to the classical sparsity assumption (under which only a small fraction of the variables are relevant), these methods work under the assumption that two connected

Published online: 11 September 2014

covariates in the network may share similar effects on the response variable. Consequently the objective of structured sparsity is twofold: improve model selection by using some prior knowledge on the structure; increase prediction performance by effective dimensionality reduction based on the prior knowledge that several covariates may share the exact same coefficient.

Most methods proposed so far use a penalized version of the log-likelihood based on some structured sparsity-inducing penalty. The fused lasso of Tibshirani et al. (2005) is one particular example: in addition to the ℓ_1-norm penalty of the lasso (Tibshirani 1996), the fused lasso penalizes the ℓ_1-norm of the vector of successive differences. It is therefore especially adapted for smoothing, when covariates are ordered and are likely to share similar effects with their direct neighbor. It has further been generalized to handle more complex structure among feature effects, in particular networks of features (Höfling et al. 2010). The network is modeled as a graph with vertices standing for the p coefficients of the model, and with a set of edges. It is used in the penalty as prior information to penalize the absolute value of the difference of connected coefficients, leading to the generalized fused lasso. Interestingly, structure in the vector of coefficients also naturally arises when jointly estimating multiple regression models. For instance, when data are collected from distinct strata (in epidemiology, these strata can be defined by crossing age, gender, ethnicity), models defined on each stratum are expected to share similarities. Consequently, structured-sparsity and in particular the generalized fused lasso can be used in this context to enforce some structure while estimating the different models (Gertheiss and Tutz 2012). In this context, the graph to be used in the penalty is usually provided by the design of the study itself (see Sect. 2.2 below).

As any prior information, the underlying graph can be more or less informative. For instance, the clustered lasso (She 2010) was proposed when only the existence of a structure is assumed but no particular knowledge allows for its precise description. Its main step involves a penalty based on the ℓ_1-norm of the vector of all the $p(p-1)/2$ differences among the parameter values. This strategy corresponds to the generalized fused lasso with the graph set to a clique that connects all coefficients. When penalizing all differences, it is very likely that some differences are unnecessarily penalized, which raises the question of the method robustness to graph misspecification. Interestingly, any structured-sparsity approach is concerned by this robustness property, but this question has never been thoroughly investigated (Azencott et al. 2013).

In this work we focus on adaptive generalized fused lasso estimates in the context of generalized linear models and show how the framework of joint regressions can be recast as a generalized fused lasso problem. In Sect. 3, we prove that adaptive generalized fused lasso estimators enjoy asymptotic oracle properties in the fixed p setting. Our results extend those obtained in the case of clique-based strategies (She 2010; Sharma et al. 2013), and in the case of joint linear regressions (Gertheiss and Tutz 2012). In particular, we observe that only adaptive versions of the generalized fused lasso enjoy asymptotic oracle properties (i.e., are such that, as n grows to infinity, the correct support is recovered with probability tending to one and estimates of non-zero coefficients perform as well as if the true underlying model were given in advance). In a further step we investigate the empirical benefits of using an ℓ_1-based fusion penalty on support recovery and prediction as compared with other penalization strategies, under logistic models. We assess the robustness of the generalized fused lasso to graph misspecification as well as its performance when true coefficients are not exactly equal. We also illustrate the benefits of using adaptive weights and/or relaxation with generalized fused lasso estimates (Sect. 4). The performance of the generalized fused lasso in the context of joint logistic regression models is also empirically evaluated. Finally the application of the generalized fused lasso is illustrated on two data sets in Sects. 5 and 6.

2 The adaptive generalized fused lasso in generalized linear models

2.1 Models, loss functions and penalty

We consider the generalized linear models framework (McCullagh and Nelder 1989) with Y_i the *response* variable $i = 1, \ldots, n$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ a p-dimensional vector of features. We further set $\mathbf{z}_i = (1, \mathbf{x}_i^T)^T$, and we consider the *fixed design* case with $\sum_{i=1}^n x_{ij} = 0$. For generalized linear models the distribution of the response variable is given by

$$f\left(y_i, \boldsymbol{\beta}^*, \phi\right) = \exp\left(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

where ϕ is a dispersion parameter and functions $b(\cdot)$, $a(\cdot)$ and $c(\cdot,\cdot)$ are known. The linear predictor η_i is given by $\mathbf{z}_i^T\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\beta}_{\setminus 0}^*)^T \in \mathbb{R}^{p+1}$ stands for the vector of coefficients, with β_0^* the intercept parameter and $\boldsymbol{\beta}_{\setminus 0}^* = (\beta_1^*, \ldots, \beta_p^*)$. The mean $\mu_i = \mathbb{E}(Y_i)$ is related to the linear predictor via the link function g: $g(\mu_i) = \eta_i$. Here we consider the canonical link function. Estimation of the parameter vector $\boldsymbol{\beta}^*$ is usually performed by the maximum likelihood method. It consists in minimizing J, given by $J(\boldsymbol{\beta}) = -\sum_{i=1}^n \log f(y_i, \boldsymbol{\beta}, \phi)$, with respect to $\boldsymbol{\beta}$. In the simulation studies and the applications below we focus on the logistic model for which $Y_i \in \{0, 1\}$, $a(\phi) = 1$, $b(x) = \log(1 + \exp(x))$ and $c \equiv 0$. Under logistic models, the mean and the linear predictor are related by $\mu_i = 1/(1 + \exp(-\mathbf{z}_i^T\boldsymbol{\beta}^*)) = g^{-1}(\eta_i)$.

As mentioned above, we further focus on the generalized fused lasso (Höfling et al. 2010). Consider a graph $G = (V, E)$, with node set $V=\{1,\ldots p\}$ that corresponds to the coefficient indices in $\boldsymbol{\beta}_{\backslash 0}$, and edge set E that corresponds to pairs of connected coefficient indices (j, ℓ) with $j > \ell$. The graph G that is used in the penalty is fixed and represents some prior knowledge, given by an expert. The adaptive generalized fused lasso penalty consists in penalizing all coefficients along with all coefficient differences for which an edge exists in G:

$$\text{pen}_{\text{Ada}}(\boldsymbol{\beta}; G, \mathbf{w}) = \lambda_n^{(1)} \sum_{j \in V} w_j^{(1)} |\beta_j| + \lambda_n^{(2)} \sum_{(j,\ell) \in E} w_{j\ell}^{(2)} |\beta_j - \beta_\ell|.$$

In the fixed p case considered here, and following the idea of the adaptive lasso (Zou 2006), adaptive weights $w_j^{(1)}$ and $w_{j\ell}^{(2)}$ are based on initial Maximum-Likelihood estimates $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$ (under the assumptions introduced in Sect. 3, these initial estimates are $\sqrt{n}$-consistent). More precisely, for some $\gamma > 0$, we set $w_j^{(1)} = |\widetilde{\beta}_j|^{-\gamma}$ and $w_{j\ell}^{(2)} = |\widetilde{\beta}_j - \widetilde{\beta}_\ell|^{-\gamma}$. The rationale is to penalize more heavily coefficients (or differences of coefficients) when their initial estimates are small. A typical value (that we use) for γ is 1. The adaptive generalized fused lasso criterion Q is then simply defined, for given graph G and weights $\mathbf{w}$, as

$$Q(\boldsymbol{\beta}) = J(\boldsymbol{\beta}) + \text{pen}_{\text{Ada}}(\boldsymbol{\beta}; G, \mathbf{w}).$$

2.2 Application of the generalized fused penalty to joint modeling

Interestingly, in some situations the graph to be used in the penalty is provided by the design of the study itself. This is notably the case when jointly estimating multiple sparse regression models. This joint modeling framework is actually a particular instance of multi-task learning (Argyriou et al. 2008), where tasks to be learned correspond to different *strata* on which the same response variable and covariates are recorded (Huang et al. 2012; Danaher et al. 2014). This case of data collected from distinct *strata*, is very common in epidemiology for instance where each stratum can be defined by crossing gender, age and ethnicity. The design is structured according to a given vector of categorical variables $(\mathcal{C}_1, \ldots, \backslash CC_n)$, each variable taking values in $\{1, \ldots, C\}$, with $C \geq 1$ the total number of strata. Let n_c be the number of observations falling into stratum c (so that $n = \sum_c n_c$). We will assume that $n_c/n \to \kappa_c$ as $n \to \infty$, with $0 < \kappa_c < 1$ (i.e., that *stratum* sizes all tend to infinity at the same rate). Under generalized linear models, we would have $g(\mu_i) = \mathbf{z}_i^T \boldsymbol{\beta}^*_{\mathcal{C}_i}$, for $i = 1, \ldots, n$, where $\boldsymbol{\beta}^*_c$ denotes the vector of parameters for stratum c. The purpose of the analysis is to determine whether the distribution of the response varies across strata, *i.e.* to detect which coefficients of $\boldsymbol{\beta}^*_c$ do vary with c. Constructing independent (possibly sparse) models for each stratum would not take advantage of the common structure, while constructing a single model for the whole data set would mask the differences. Alternatively, the generalized fused lasso can be used to couple estimations obtained from each stratum, encouraging them to share some common structure (Gertheiss and Tutz 2012). More precisely, the following penalty can be used:

$$\sum_{c=1}^{C} \left\{ \lambda_n^{(1)} \sum_{j=1}^{p} w_j^{(1)} |\beta_{c,j}| \right\}$$
$$+ \lambda_n^{(2)} \sum_{j=0}^{p} \sum_{c_1 > c_2} w_{c_1,c_2,j}^{(2)} |\beta_{c_1,j} - \beta_{c_2,j}|,$$

where $w_j^{(1)}$ and $w_{c_1,c_2,j}^{(2)}$ are appropriate adaptive weights (when $n_c/n \to \kappa_c$ with $0 < \kappa_c < 1$, Maximum-Likelihood estimates computed on each stratum are $\sqrt{n_c}$, and hence $\sqrt{n}$, consistent under general assumptions and can therefore be used to compute these weights). Parameter $\lambda_n^{(2)}$ governs the amount of shrinkage for differences between strata: if null, this penalty resumes to C independent lasso penalties. If positive, the fused part of the penalty encourages coefficients $\beta_{c_1,j}$ and $\beta_{c_2,j}$ to be at least close to each other (that is, the jth coefficient in strata c_1 and c_2 respectively). In the general context described here, the problem reduces to an adaptive generalized fused lasso where the graph is composed by $p+1$ cliques of size C and the j-th clique connects $\beta_{1,j}$, …, $\beta_{C,j}$ all together (see the Supplementary Material for more details).

3 Theoretical results

We study the asymptotic properties of the adaptive generalized fused lasso estimator in generalized linear models. Before stating our results some notations and assumptions are needed. Let $\mathcal{A} = \{1 \leq j \leq p,\ \beta_j^* \neq 0\}$ be the *support* of $\boldsymbol{\beta}^*_{\backslash 0}$ and $p_0 = |\mathcal{A}|$ its cardinality. Further consider the set

$$\mathcal{B} = \{(j, \ell) \in E,\ \beta_j^* \neq 0 \text{ and } \beta_j^* = \beta_\ell^*\} \subset \mathcal{A} \times \mathcal{A}.$$

We denote by $\mathcal{I}(\boldsymbol{\beta})$ the empirical Fisher's matrix of size $(p+1) \times (p+1)$. For future use, observe that $\mathcal{I}(\boldsymbol{\beta}^*) = \mathbf{Z}^T \boldsymbol{D}\mathbf{Z}$, where $\boldsymbol{D}$ denotes an $n \times n$ diagonal matrix. For instance under logistic regression models, we have $D_{ii} = \mu_i(1 - \mu_i)$. For any $\delta \geq 0$, we further denote by $N_n(\delta)$ the neighborhood of $\boldsymbol{\beta}^*$ defined by

$$N_n(\delta) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{p+1} / \left\| \left[\mathcal{I}(\boldsymbol{\beta})^{-\frac{1}{2}} \right]^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right\| \leq \delta \right\}.$$

We will work under the following conditions:

- **AL1** $\mathcal{I}(\boldsymbol{\beta}^*)/n$ converges to $\mathbf{C}$ where $\mathbf{C}$ is a positive definite $(p+1) \times (p+1)$ matrix.

– **AL2** As n goes to ∞,

$$\max_{\boldsymbol{\beta}\in N_n(\delta)} \left\| \mathcal{I}(\boldsymbol{\beta})^{-\frac{1}{2}} \mathcal{I}(\boldsymbol{\beta}^*)^T \left[\mathcal{I}(\boldsymbol{\beta})^{-\frac{1}{2}}\right]^T - \mathbf{I}_{p+1} \right\| \to 0.$$

Assumptions **AL1** and **AL2** are standard when working under generalized linear models (McCullagh and Nelder 1989). Assumption **AL1** is similar to the one used to study the fused lasso in the Gaussian context (Tibshirani et al. 2005). Let us remark that under **AL1**, the minimization of criterion Q defined in Sect. 2.1 corresponds, for n large enough, to a strictly convex optimization problem, and thus is not concerned by the issue of multiple minima.

In Theorem 1 below, we generalize the first theorem of Tibshirani et al. (2005) to the case of generalized fused lasso, in generalized linear models. This result establishes the root-n consistency of non-adaptive generalized fused lasso estimates. However, it also implies that when $\lambda_n^{(m)} = O(\sqrt{n})$, for $m = 1, 2$, the support of $\boldsymbol{\beta}^*$ can not be recovered with high probability by non-adaptive fused lasso estimates, as stated in Proposition 1 below. Proofs are given in the Appendix.

Theorem 1 *Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of criterion Q defined in Sect. 2.1 with $w_j^{(1)} = 1$ and $w_{j\ell}^{(2)} = 1$ for all j, ℓ. If $\lambda_n^{(m)}/\sqrt{n} \to \lambda_0^{(m)} \geq 0$ ($m = 1, 2$), then under assumptions* **AL1-2**, *$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \to_d \arg\min(\mathcal{V})$, where $\mathcal{V}$ is the function defined, for $\mathbf{u} = (u_0, \ldots, u_p) \in \mathbb{R}^{p+1}$, as*

$$\begin{aligned}\mathcal{V}(\mathbf{u}) &= \mathbf{u}^T\mathbf{W} + \frac{1}{2}\mathbf{u}^T\mathbf{C}\mathbf{u}\\ &+\lambda_0^{(1)} \sum_{j=1}^{p} \left\{u_j sign(\beta_j^*)\mathbb{I}(\beta_j^* \neq 0) + |u_j|\mathbb{I}(\beta_j^* = 0)\right\}\\ &+\lambda_0^{(2)} \sum_{(j,\ell)\in E} \left\{(u_j - u_\ell) sign(\beta_j^* - \beta_\ell^*)\mathbb{I}(\beta_j^* \neq \beta_\ell^*)\right.\\ &\left.+|u_j - u_\ell|\mathbb{I}(\beta_j^* = \beta_\ell^*)\right\}.\end{aligned}$$

Above, $\mathbf{W}$ has an $\mathcal{N}(\mathbf{0}_{p+1}, \mathbf{C})$ distribution.

Proposition 1 *Under the assumptions of Theorem 1, and setting $\widetilde{\mathcal{A}}_n = \{1 \leq j \leq p, \widehat{\beta}_j \neq 0\}$, we have*

$$\limsup_n \mathbb{P}(\widetilde{\mathcal{A}}_n = \mathcal{A}) \leq c < 1,$$

where c is a constant depending on the true model.

Now we show that for appropriate choices of $\lambda_n^{(m)} = O(\sqrt{n})$ for $m = 1, 2$, the adaptive generalized fused lasso estimator $\widehat{\boldsymbol{\beta}}^{ad}$, defined as the minimizer of criterion Q in Sect. 2.1, enjoys asymptotic oracle properties, contrasting with its non-adaptive counterpart. Some more notations are needed before stating our result: in particular, the number s_0 of distinct non-zero values in $\boldsymbol{\beta}^*_{\setminus 0}$ "supported" by G needs to be precisely defined (s_0 can be seen as the theoretical model complexity "supported" by G). To this end, first observe that $\mathcal{A} \subseteq V$ and $\mathcal{B} = \{(j, \ell) \in E : \beta_j^*\beta_\ell^* \neq 0, \beta_j^* = \beta_\ell^*\} \subseteq E$. Then consider the graph $G_\mathcal{B} = (\mathcal{A}, \mathcal{B})$ and denote by s_0 the number of its connected components (e.g., in the particular case where G is a chain graph, s_0 is the number of segments consisting of non-zero and equal coefficients). Observe that $d_0 \leq s_0 \leq p_0$, where $p_0 = |\mathcal{A}|$ is the number of non-zero coefficients in $\boldsymbol{\beta}^*_{\setminus 0}$ and d_0 is the number of *distinct* non-zero values in $\boldsymbol{\beta}^*_{\setminus 0}$. We actually have $s_0 = p_0$ if and only if $(\beta_j^* = \beta_\ell^* \neq 0 \Rightarrow (j, \ell) \notin E)$. Moreover, two coefficients that are theoretically equal can not be fused together if they do not belong to the same connected component in $G_\mathcal{B}$. Furthermore, $s_0 = d_0$ if and only if for all (j, ℓ) such that $\beta_j^* = \beta_\ell^*$, j and ℓ belong to the same connected component of $G_\mathcal{B}$. Now denote by $\mathcal{A}_1, \ldots, \mathcal{A}_{s_0}$ the sets of vertices of each connected component of $G_\mathcal{B}$. Of course, we have $\mathcal{A} = \bigcup_{s=1}^{s_0} \mathcal{A}_s$. Further set $j_s = \min\{\mathcal{A}_s\}$ for $s = 1, \ldots, s_0$. Now we can define $\boldsymbol{\beta}^*_\mathcal{B} = (\beta_0^*, \beta_{j_1}^*, \ldots, \beta_{j_{s_0}}^*)^T$, which is composed by the intercept and the s_0 distinct non-zero values of $\boldsymbol{\beta}^*_{\setminus 0}$ supported by G; we further set $\widehat{\boldsymbol{\beta}}^{ad}_\mathcal{B}$ its estimate. Now denote by $\mathbf{X}_\mathcal{B}$ the matrix of size $n \times s_0$, whose s-th column is $X_{\mathcal{B}_s} = \sum_{j\in\mathcal{A}_s} X_j$, where X_j is the j-th column of $\mathbf{X}$. Further set $\mathbf{Z}_\mathcal{B} = (\mathbf{1}_n, \mathbf{X}_\mathcal{B})$ and denote by $\mathbf{C}_\mathcal{B}$ the $(s_0 + 1) \times (s_0 + 1)$ positive definite matrix that is defined as the limit, as $n \to \infty$, of $\mathcal{I}(\boldsymbol{\beta}^*_\mathcal{B})/n$, where $\mathcal{I}(\boldsymbol{\beta}^*_\mathcal{B}) = \mathbf{Z}_\mathcal{B}^T\boldsymbol{D}\mathbf{Z}_\mathcal{B}$. Finally introduce $\mathcal{A}_n = \{1 \leq j \leq p, \widehat{\beta}_j^{ad} \neq 0\}$ and $\mathcal{B}_n = \{(j, \ell) \in E, \widehat{\beta}_j^{ad} \neq 0$ and $\widehat{\beta}_j^{ad} = \widehat{\beta}_\ell^{ad}\}$. We have now all the ingredients to state our main result, whose proof is given in the Appendix (see Sect. 1).

Theorem 2 *If $\lambda_n^{(m)}/\sqrt{n} \to 0$ and $\lambda_n^{(m)} n^{(\gamma-1)/2} \to \infty$, $m = 1, 2$, then, under assumptions* **AL1-2**, *the adaptive generalized fused lasso estimator satisfies the following properties:*

1. *Consistency in variable selection: $\mathbb{P}[\mathcal{A}_n = \mathcal{A}] \to 1$ and $\mathbb{P}[\mathcal{B}_n = \mathcal{B}] \to 1$ as $n \to +\infty$.*
2. *Asymptotic normality:*
   $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{ad}_\mathcal{B} - \boldsymbol{\beta}^*_\mathcal{B}\right) \longrightarrow_d \mathcal{N}\left(\mathbf{0}_{s_0+1}, \mathbf{C}_\mathcal{B}{}^{-1}\right).$

Recall that weights are set to $w_j^{(1)} = |\widetilde{\beta}_j|^{-\gamma}$ and $w_{j\ell}^{(2)} = |\widetilde{\beta}_j - \widetilde{\beta}_\ell|^{-\gamma}$ for some $\gamma > 0$, where $\widetilde{\boldsymbol{\beta}}$ are initial Maximum-Likelihood estimates of $\boldsymbol{\beta}^*$. As mentioned above, under our assumptions, these initial estimates are $\sqrt{n}$-consistent. Other initial estimates could be used and $\sqrt{n}$-consistency is not a necessary condition: if initial estimates are n^ξ-consistent, with $0 < \xi < 1/2$, then the result of Theorem 2 remains valid if $\lambda_n^{(m)}/\sqrt{n} \to 0$ and $\lambda_n^{(m)} n^{\xi\gamma-1/2} \to \infty$, $m = 1, 2$.

In other respect, in the joint modeling context, and under the assumption that $n_c/n \to \kappa_c$ as $n \to \infty$, with $0 < \kappa_c < 1$, Theorem 2 implies results similar to those presented in Gertheiss and Tutz (2012) in the linear regression case.

Interestingly Theorem 2 also allows us to compare the asymptotic theoretical performance of various graph-based methods. Observing that $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{B}})$ is the information matrix of the true submodel as soon as $s_0 = d_0$, Theorem 2 states that in the fixed p scenario, the estimator $\widehat{\boldsymbol{\beta}}^{ad}_{\mathcal{B}}$ is asymptotically efficient as soon as $s_0 = d_0$, which is notably the case for clique-based methods (Sharma et al. 2013; She 2010). Moreover, because adding edges in any given graph between coefficients with theoretical different values does not modify the set $\mathcal{B}$, and so leaves the quantity s_0 unchanged, our theoretical results state that, asymptotically, adding edges in the graph can only improve the adaptive generalized fused lasso performance. However, removing edges between coefficients with theoretical equal value may modify the set $\mathcal{B}$ and increase the quantity s_0, leading to poorer asymptotic performance. These results being asymptotic, we evaluate the finite sample properties of the generalized fused lasso in the forthcoming simulation study, with an emphasis on its robustness to graph misspecifications.

4 Simulation study

We perform an extensive simulation study to compare the performance of the generalized fused lasso with other penalized strategies in two contexts: (i) when the variables lie on a graph and the support is densely connected on the graph and (ii) in the joint modeling framework. In (i), our main objective is twofold: to study the impact of a graph misspecification on the generalized fused lasso performance and also to study the robustness of the generalized fused lasso when non-null connected coefficients do not share the exact same value. In (ii), the graph being "given" by the design of the study, we will study how the performance of the generalized fused lasso vary with the level of heterogeneity across strata.

4.1 Simulation framework, implementation and evaluation

Our results being in the fixed design framework, we set N, the maximal sample size considered in a given scenario, and we generate N *i.i.d.* predictors $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1,\ldots,N$, from a $\mathcal{N}(\mathbf{0}_p, \mathbf{C}_{\text{AR}(1)})$ distribution, where $\mathbf{C}_{\text{AR}(1)} = \left(\rho^{|i-j|}\right)/16\,(i = 1\ldots p,\ j = 1\ldots p)$. Then, given $n \leq N$, $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$ and $\boldsymbol{\beta}^*$ (with $\beta^*_0 = 0$), the vector of labels is generated such that $Y_i \sim \mathcal{B}(\mu_i)$, with μ_i defined as in Sect. 2.1. Unless otherwise stated, the p_0 non-null coefficients of $\boldsymbol{\beta}^*_{\setminus 0}$ are all set to a common value β^*, that we make vary in $\{\log(1.1), \log(2), \log(4), \log(8), \log(12)\}$. Fifty replicates are considered for each configuration.

The adaptive generalized fused lasso is solved with the coordinate-wise optimization algorithm of Höfling et al. (2010) implemented in the `FusedLasso` R package, that we made available from the CRAN repository. Tuning parameters (λ_1, λ_2) are selected using the BIC, which is standard when $n \gg p$ (degrees of freedom are set to the number of distinct non-null coefficients in the estimated model). Because using shrunk estimates within the BIC can result in severe over-fitting, we also propose an extension of the relaxed lasso (Meinshausen 2007) to the generalized (adaptive) fused lasso (see Section 2 in the Supplementary Material).

Prediction accuracy is assessed using an independent test sample of N observations $(\mathbf{z}_i^{(0)}, Y_i^{(0)})$ and computing $\text{Acc.Pred} = (1/N)\sum_{i=1}^N \mathbb{I}(Y_i^{(0)} = \widehat{Y}_i^{(0)})$, with $\widehat{Y}_i^{(0)}$ equal to $\mathbb{I}(1/[1+\exp(-\mathbf{z}_i^{(0)T}\widehat{\boldsymbol{\beta}})] > 0.5)$. To assess accuracy on support recovery we define $\text{Acc.A} = (|\bar{\mathcal{A}} \cap \bar{\mathcal{A}}_n| + |\mathcal{A} \cap \mathcal{A}_n|)/p$ where $\bar{\mathcal{A}}$ (resp. $\bar{\mathcal{A}}_n$) is the set of null coefficients in $\boldsymbol{\beta}^*$ (resp. $\widehat{\boldsymbol{\beta}}$). Moreover, to evaluate the performance regarding the classification of pairs of coefficients, we focus on $\text{Acc.B} = |\{(j,\ell) \in \mathcal{B} : \widehat{\beta}_j = \widehat{\beta}_\ell\}|/|\mathcal{B}|$. Because the *lasso* does not encourage equality among non-null coefficients, its Acc.B only reflects the proportion of pairs in $\mathcal{B}$ whose elements are both put to zero by the *lasso*. Consequently, the *lasso* should show poor results regarding this criterion (especially when β^* is large). In the joint modeling framework, Acc.B assesses the capacity of the methods to detect homogeneity across strata, and could be interpreted as a type-1 error of an interaction test.

In the following simulation study we explore the properties of the "raw" generalized fused lasso (without adaptive weights nor relaxation), the adaptive generalized fused lasso, the relaxed generalized fused lasso, and of the relaxed adaptive generalized fused lasso. Comparisons are made using the relaxed adaptive *lasso* as a reference that does not account for any structure among feature effects. The sparse group lasso (Huang et al. 2012), which includes the lasso penalty and the group lasso penalty, and an extension of the elastic-net (Sun and Wang 2012) were also tested, but their performance were similar to those achieved by the lasso (see Supplementary Figs. 1 and 6).

4.2 Performance of adaptive generalized fused lasso estimates and influence of the provided graph

We set $p = 24$ and sample $\mathbf{x}_i$ as describe above with $\rho = -0.39$ (Zou 2006). Our theoretical results being asymptotic in n we consider cases where $n/p \in \{1, 5, 10, 50\}$. We also explore different degrees of sparsity with $p_0 \in \{12, 8, 3\}$. To study the robustness of selection method to a graph misspecification, we generate graphs with varying suitabilities such that equal (resp. non-equal) coefficients are connected with probability θ (resp. $1-\theta$). When θ increases fewer edges connect distinct coefficients, the most [resp. least] favorable configuration for generalized fused estimates being when $\theta = 1$ [resp. $\theta = 0$]. We mention that edges of the graph are fixed across replicates. When no prior information is available on graph G, a strategy can be to use no graph (with the lasso) or to use a graph that connects all coefficients (clique-

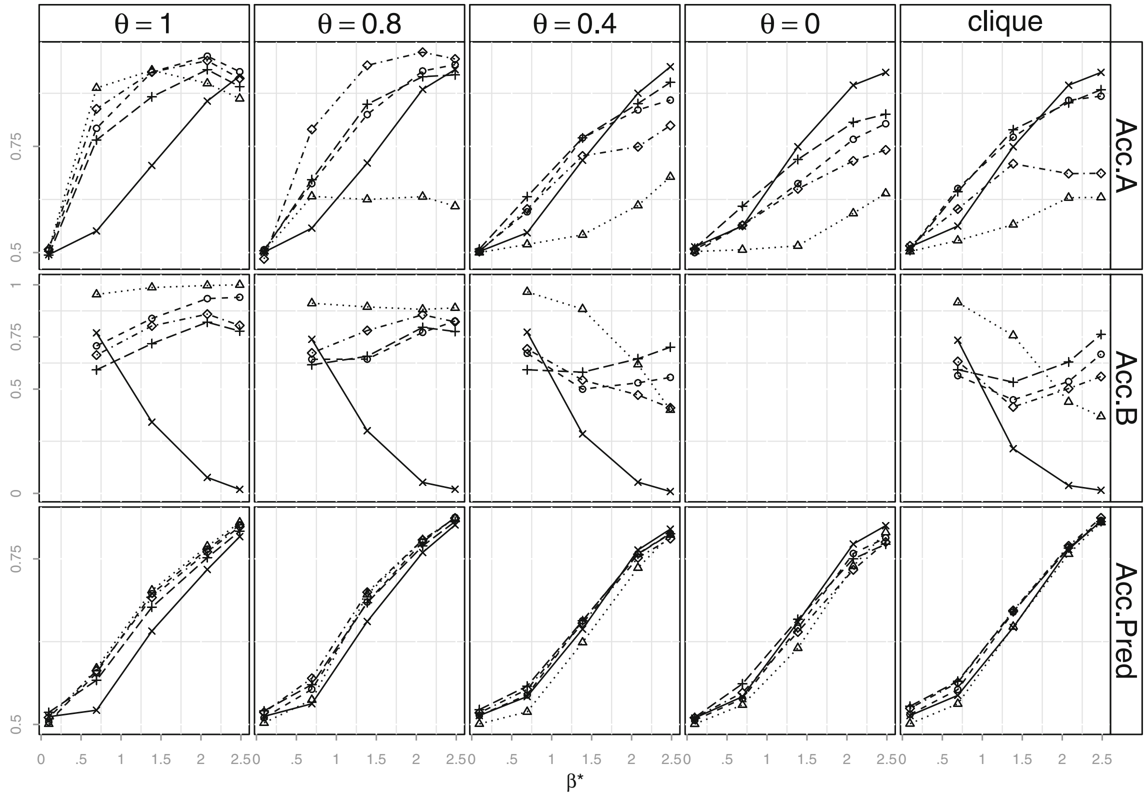

Fig. 1 Accuracies for ℓ_1-based fused penalties ($n/p = 10,\ p_0 = p/2$). Raw fused lasso (*opentriangle-dotted*), adaptive fused lasso (*opencircle-dashed*), relaxed fused lasso (*diamond-dotdash*), relaxed adaptive fused lasso (*plus-longdashed*), relaxed adaptive lasso (*times-solid*)

graph). This latter option is also considered here to compare the generalized fused lasso with clique-based methods (Sharma et al. 2013; She 2010).

In the following we choose the adaptive lasso as a reference that does not depend on any graph. In the best-case scenario (perfectly suited graph, $\theta = 1$), all graph-based method are more accurate than the lasso for support recovery, which reflects a cooperative effect that is characteristic of ℓ_1-based fused penalties (Fig. 1). Also, all fused penalties are more accurate than the lasso for prediction, without much difference among them (Fig. 1). Then the sensitivity to graph misspecification depends on the penalty: the "raw" fused lasso is very sensitive and using adaptive weights and/or relaxation increases robustness to graph misspecifications (Fig. 2). Note that the adaptive and the relaxed adaptive generalized fused lasso perform better than (or at least similarly to) the lasso except when $\theta = 0$. Averaged accuracies on support recovery show that the *relaxation* of the adaptive generalized fused lasso does not necessarily increase its accuracy (Table 1). Then when varying the n/p ratio (Supplementary Fig. 3), adaptive and/or relaxed strategies remain more accurate than the "raw" generalized fused lasso except when the graph is perfectly suited (i.e., $\theta = 1$, which is very unlikely in practice), or in situations where maximum likelihood estimates may lack in precision ($n/p \simeq 1$).

In our simulation design all non-null coefficients were set to a unique value so that the task was to detect a group of p_0 non-null parameters among a set of p. This was a way to study the effect of varying p_0 on the robustness to a graph misspecification. Then, as expected, all graph-based fused methods perform better when the graph is highly suitable and/or p_0 is high (as compared to the lasso, Fig. 2).

Next, focusing on the strategy that consists in penalizing all possible differences using the clique-graph (Sharma et al. 2013; She 2010), we observe that they are close to those of graph-based methods with low suitability ($\theta = 0$; 0.4), whatever the degree of sparsity p_0 of the true model (see Fig. 2). It appears that generalized fused lasso estimates obtained with the clique-graph never significantly improve upon the lasso for support recovery in our experiments. This is actually explained by the clique-graph strategy itself, for which the number of edges connecting non-null and equal coefficients increases with the number of non-null coefficients p_0, but so does the number of misleading edges that connect null and non-null coefficients ($p_0(p - p_0)$ is an increasing function of $p_0 \in [0, p/2]$). These results complement those of Theorem 2 above: while using a clique-graph is asymptotically optimal, this strategy is clearly sub-optimal on finite samples.

Lastly, we consider the classification of pairs of non-null (and then equal) coefficients (Acc.B; Fig. 1). When the graph

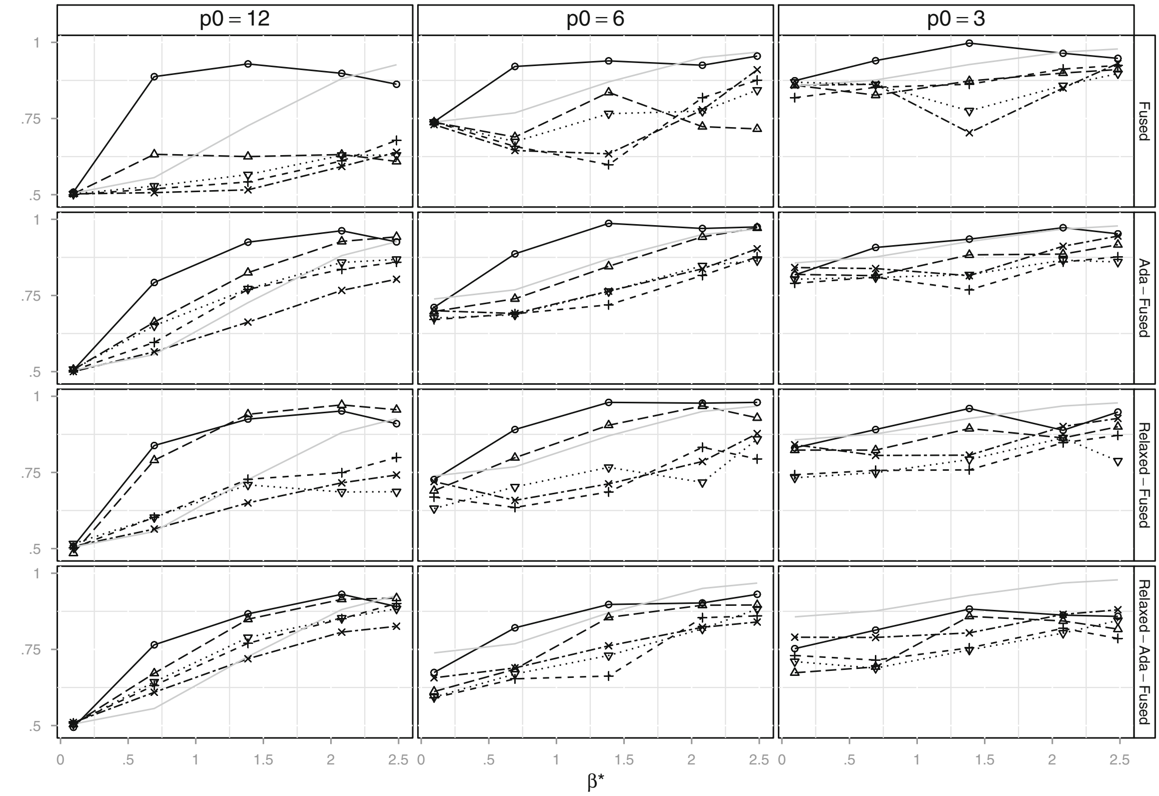

Fig. 2 Accuracy on support recovery $\mathcal{A}$ for different graphs provided to each method, with $\theta = 1$ (*opencircle-solid*), $\theta = .8$ (*opentriangle-long-dash*), $\theta = .4$ (*plus-dash*), $\theta = 0$ (*times-dot-dash*), lasso (*solid-grey*), clique (*opentriangledown-dot*). The lasso strategy corresponds to the case where no graph is provided. The clique strategy corresponds to the case where all differences are penalized

Table 1 Average Accuracies ($\times 100$) for support (Acc.A), pairs of non-zero and equal coefficients (Acc.B) and prediction (Acc.pred)

n/p		Acc.A		Acc.B		Acc.Pred	
		Non-R	R	Non-R	R	Non-R	R
1	Non-adaptive	62	57	75	45	58	58
	Adaptive	54	55	41	47	58	58
5	Non-adaptive	67	73	87	66	61	62
	Adaptive	71	69	66	59	61	61
10	Non-adaptive	71	78	82	70	62	63
	Adaptive	78	76	72	68	63	63
50	Non-adaptive	79	88	75	80	65	65
	Adaptive	89	88	87	86	65	65

Note that the relaxation strategy does not systematically improve the accuracies of the adaptive strategy
R relaxed, *non-R* non-relaxed

is completely misspecified ($\theta = 0$) Acc.B can not be computed ($\mathcal{B} = \emptyset$ in this case). Clearly, no method but the "raw" version of the generalized fused lasso is robust on graph mis-specification regarding pairs of coefficients. However, what seems to be a good performance for the "raw" fused lasso is actually linked to a poor model selection and estimation precision since the "raw" fused lasso tends to return the same value for all coefficients (Supplementary Fig. 4). To explain why the accuracy on pairs drops with the graph mis-specification, we first note that *both* the number of edges connecting non-zero equal coefficients $|\mathcal{B}|$ and Acc.B decrease as θ decreases. When the graph is misspecified there is an increased discrepancy between the true number of distinct coefficients (d_0), the number of distinct coefficients that would be possible to estimate in the asymptotic setting (s_0), and s_n the actual non-asymptotic model complexity. Then as θ decreases, $s_0 - d_0$ obviously tends to increase but so does $s_n - s_0$. It means that graph misspecifications lead to models that are too complex with respect to the theoretical one. This is still true with the clique-graph for which we do have $s_0 = d_0$, but Acc.B is still moderate which means that $s_n > s_0 = d_0$: even when all the edges that should be present

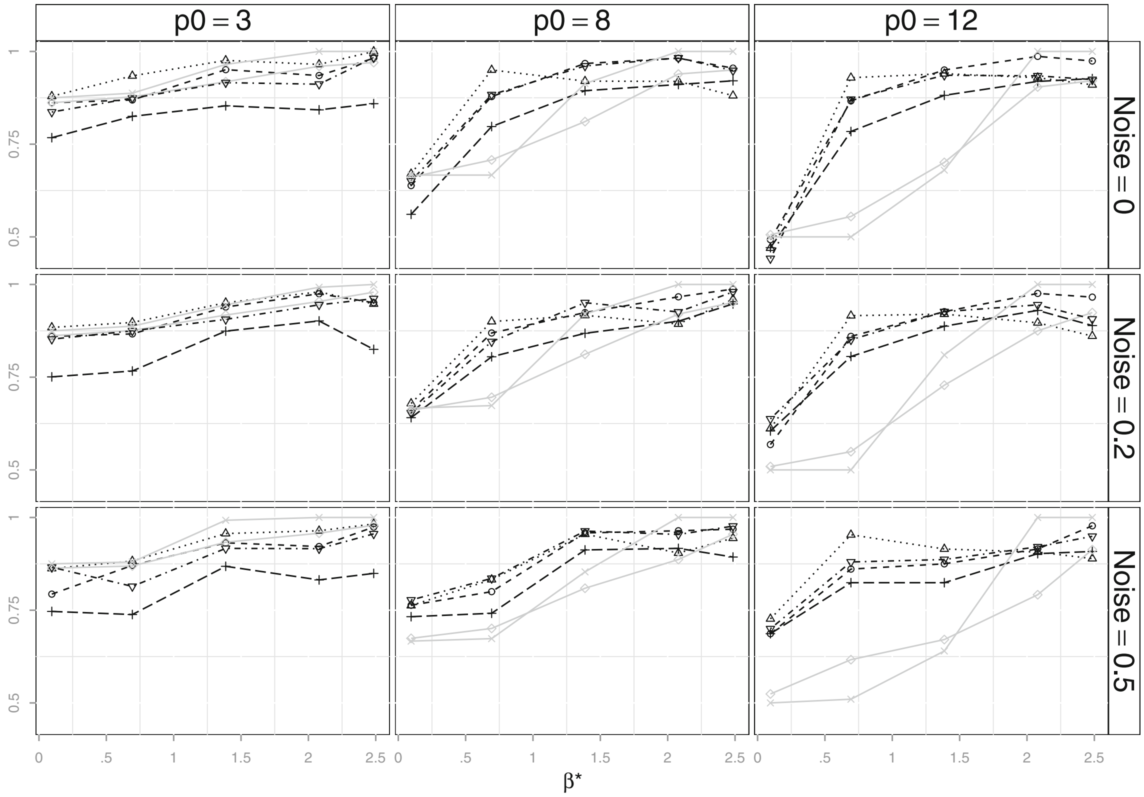

Fig. 3 Accuracy on support recovery $\mathcal{A}$ when the graph is perfect ($\theta = 1$) but non-zero (and connected) coefficients may not be exactly equal: fused-like estimates are represented in *black* (*opentriangle-dotted*: generalized fused lasso, *opencircle-dashed*: adaptive generalized fused lasso, *opentriangledown-dotdash*: relaxed generalized fused lasso, *plus-longdash*: relaxed adaptive generalized fused lasso). Results for the relaxed group lasso (*grey*, *times-solid*) and the relaxed adaptive lasso (*grey*, *diamond-solid*) are also reported

in the graph are indeed present, misleading edges prevent the method from selecting the right model: Acc.B is moderate and, in turn, so is Acc.A.

4.3 Performance of the generalized fused lasso when connected coefficients may not be exactly equal

Here, we adopt the exact same setting as in the previous paragraph, but instead of setting the p_0 non-null coefficients to the same β^* value, we set each of them to $|\beta^* + \nu|$, with $\nu \sim \mathcal{N}(0, \sigma_\nu^2)$. The variance σ_ν^2 governs the amount of "noise" (variability across the p_0 non-null coefficients), and we make it vary in $\{0, 0.2, 0.5\}$. Illustrations of the resulting $\boldsymbol{\beta}^*$ vector are provided on Supplementary Fig. 5. In this setting, the reference method is arguably the original group lasso (i.e., with the group lasso penalty only) where the penalty accounts for the true group structure: one group consisting of the null components and the other one consisting of the non-null coefficients. A relaxed version of this method was therefore included in this particular setting [using the grplasso R package (Meier et al. 2008)].

We first compete it with generalized fused lasso estimates, for which exact knowledge of the group structure is also assumed; that is we consider generalized fused lasso estimates with $\theta = 1$ (results for the relaxed adaptive lasso are also provided). Results regarding support recovery (Acc.A) are presented on Fig. 3 for $n = 240$. Interestingly, most generalized fused lasso estimates are always at least comparable to the group lasso. More importantly, they significantly outperform the group lasso in many situations, especially when the variability across the p_0 non-null coefficients is weak, p_0 is large or signal is weak (low values of β^*). Comparisons based on prediction accuracy advocate even more for the use of generalized fused lasso estimates (see Supplementary Fig. 6).

Then, considering the situation where the graph used in the generalized fused penalty is imperfect, Fig. 4 further shows that, in the case where $\sigma_\nu^2 = 0.2$, the performance of the group lasso (with exact knowledge of the group structure) in terms of support recovery is comparable to or worse than that of the generalized fused lasso in the following configurations. (i) when $p_0 = 12$ and $\theta \geq 0.8$ or $\beta^* \leq 1.5$ (irrespective of the θ value); (ii) when $p_0 = 8$ and $\theta = 1$ or $\theta = 0.8$ and $\beta^* \leq 1.5$; and (iii) when $p_0 = 3$ and $\theta = 1$. Analogous results were obtained for other values of σ_ν^2. In other words, even when the true structure is unknown, generalized fused

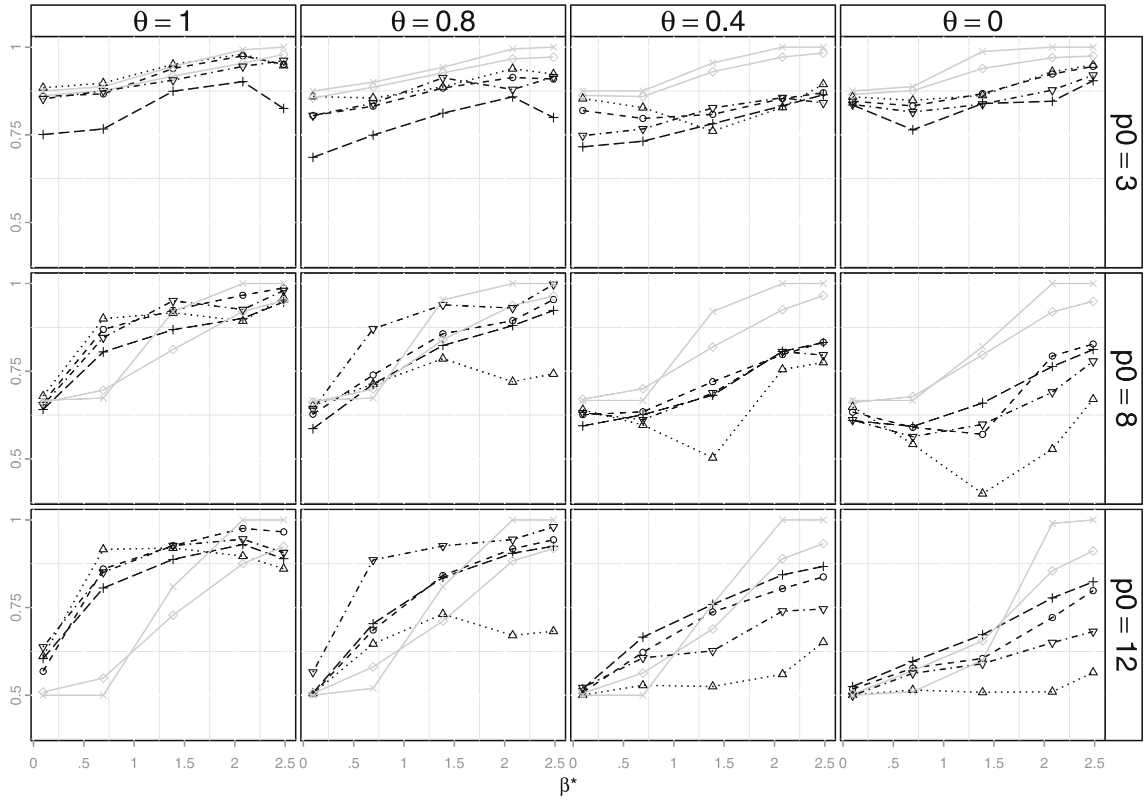

Fig. 4 Accuracy on support recovery $\mathcal{A}$ when non-zero coefficients are not exactly equal ($\sigma_v^2 = 0.2$) and the graph provided in the generalized fused lasso is imperfect: fused-like estimates are represented in *black* (*opentriangle-dotted*: generalized fused lasso, *opencircle-dashed*: adaptive generalized fused lasso, *opentriangledown-dotdash*: relaxed generalized fused lasso, *plus-longdash*: relaxed adaptive generalized fused lasso). Results for the relaxed group lasso (*grey*, *times-solid*) and the relaxed adaptive lasso (*grey*, *diamond-solid*) are also reported

lasso estimates (with $\theta < 1$ then) can achieve performance similar to those that would achieve the group lasso if the true structure were known, especially if groups of non-null coefficients are large enough. Again, results regarding prediction accuracy support even more the generalized fused lasso (Supplementary Fig. 7).

All these results clearly state the potential benefits of using generalized fused lasso estimates instead of group lasso estimates, in the setting considered in our experiments.

4.4 Simulation study in the context of joint modeling.

Our objective here is to investigate in more details the performance of the generalized fused lasso in the context of joint logistic regressions, in particular with varying signal-strengths and levels of heterogeneity across strata.

We set $C = 4$, $n_c = 200$ (so that $n = \sum_c n_c = 800$), $p = 20$, and for each stratum we sample covariates $\mathbf{x}_i$ as before, but with $\rho = 0.5$. Then we control for the level of heterogeneity across strata by using different repartitions of null and non-null coefficients for a given covariate in the various strata, while keeping $\beta^*_{0c} = 0$ for every stratum c. In Configuration 1, all strata share the same $\boldsymbol{\beta}^*_c$ with 6 non-null elements. In the other two configurations $\boldsymbol{\beta}^*_c$s *do* vary across strata while keeping the number of non-null elements equal to 6 for each stratum. In Configuration 2, 8 coefficients are null on all strata, 6 are null on all but one stratum, and 6 are *non-null* (and equal) on all but one stratum. In Configuration 3, 8 coefficients are null on all strata and 12 are non-null (and equal) on two strata (and null on the other two). See Supplementary Fig. 8 for a graphical description of these three configurations. Then we compete the generalized fused lasso with two versions of the relaxed adaptive lasso. The "*independent*" version consists in computing the relaxed adaptive lasso on each stratum independently. In the "*interaction*" version, a reference stratum is first selected and interaction terms between the remaining strata and the covariates are included as explained in Section 3.2 in the Supplementary Material (this corresponds to the generalized fused lasso with a star-graph and with $\lambda_1 = \lambda_2$). Then one relaxed adaptive lasso is computed on this whole data set. This latter approach can be regarded as the reference method in this context.

Here we mostly focus on performance for the classification of pairs, which is of primary interest in the con-

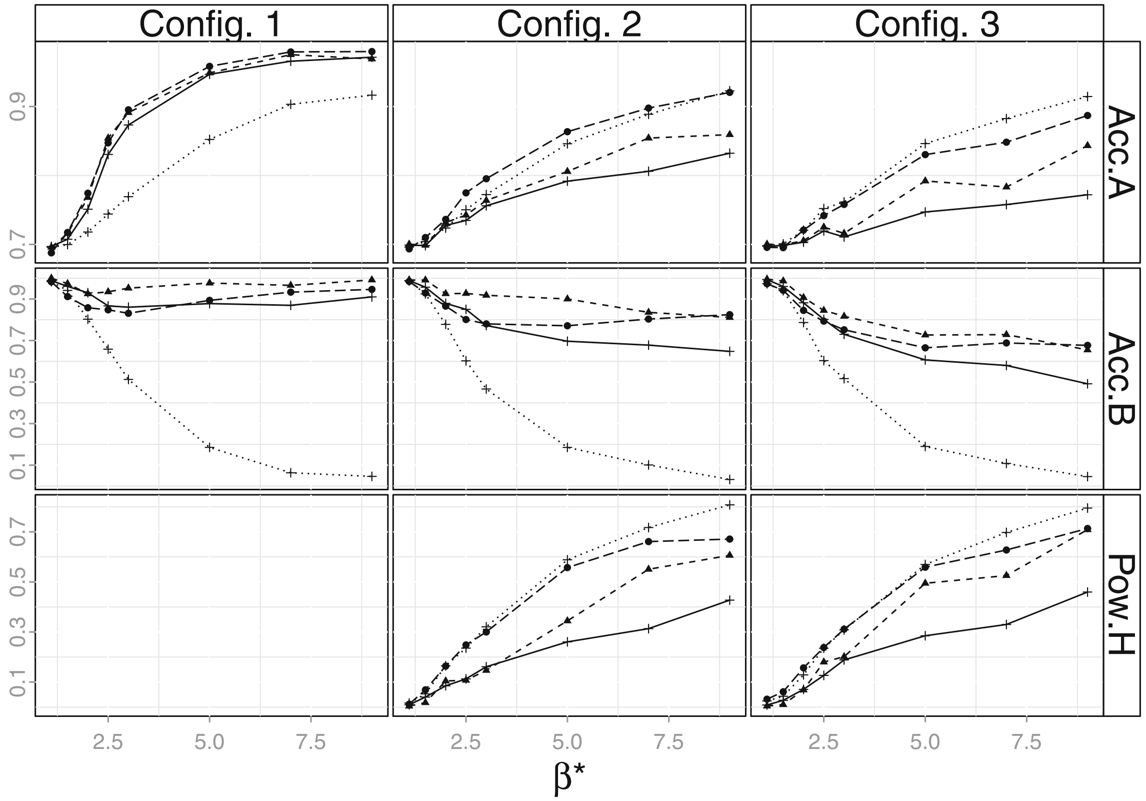

Fig. 5 Performance regarding support recovery and the classification of pairs of coefficients in the joint modeling context. Adaptive fused lasso (*blacktriangle-dashed*), relaxed adaptive fused lasso (*bullet-longdash*), independent relaxed adaptive lasso (*plus-solid*), interaction relaxed adaptive lasso (*plus-dotted*). Columns correspond to levels of heterogeneity between strata: from homogeneous (Configuration 1) to the most heterogeneous (Configuration 3)

text of joint modeling. As before, we first consider Acc.B. Because $\beta^*_{c_1,j}\beta^*_{c_2,j} \neq 0 \Rightarrow \beta^*_{c_1,j} = \beta^*_{c_2,j} \neq 0$ in our simulations, it assesses the capacity of methods to detect homogeneity across strata, and could be interpreted as a type-1 error of an interaction test, as mentioned above. We also consider the percentage of truly heterogeneous edges (i.e. edges connecting nodes that correspond to different true coefficients that are detected as heterogeneous). It can be interpreted as a power of an interaction test, and is denoted by Pow.H. Note that the "independent" relaxed adaptive lasso is expected to show good performance regarding this criterion since it does not encourage coefficients to be equal across strata: 1 - Pow.H for this method only represents the proportion of heterogeneous edges for which both coefficients are set to 0 (so that 1-Pow.H is expected to decrease, and Pow.H to increase, as β^* increases for the "*independent*"' relaxed adaptive lasso). Figure 5 shows that the relaxed adaptive generalized fused lasso outperforms the "*interaction*" relaxed adaptive lasso in terms of Pow.H and, in most cases, in terms of Acc.B as well. Moreover, it is almost "perfect" to detect heterogeneity since it shares similar performance with the "independent" relaxed adaptive lasso regarding Pow.H. The non-relaxed version is in most cases a little better (resp. worse) than the relaxed one in terms of Acc.B (resp. Pow.H). As for support recovery, (relaxed) adaptive generalized fused lasso also outperforms the "*interaction*" relaxed adaptive lasso, especially as the level of heterogeneity across strata increases (as in Configuration 2 and 3). Interestingly, even under the most heterogeneous configuration (Configuration 3), the relaxed adaptive generalized fused lasso attains performance similar to those achieved by the "*independent*" lasso in terms of Acc.A.

In other respect, Gertheiss and Tutz (2012) especially observed that imposing $\lambda_1 = \lambda_2$ lead to results comparable to those obtained with "free" λ_1 and λ_2. As shown on Supplementary Fig. 9, this was not the case on our experiments for non-adaptive versions of the generalized Fused Lasso, nor for the relaxed adaptive generalized fused lasso, especially under Configuration 3. Therefore, unless computational time is a critical issue, we would not recommend to impose $\lambda_1 = \lambda_2$ when using generalized fused lasso in the context of joint modeling (especially if non-adaptive versions are to be used).

5 Network-based prediction of cancer status based on expression data

Genomics has faced a fload of network data in the last years, ranging from protein-protein interaction data, pathway data to regulation networks (Girvan and Newman 2002; Rual et al. 2005). The molecular characterization of cancers has been at the core of many projects, especially to establish molecular subtypes of histologically similar tumors. In particular finding genomic signatures has been the *graal* for many studies to predict patient outcome, survival or relapse (Guedj et al. 2012). Such signatures can be determined using a penalized logistic regression model based on gene-expression data as covariates. Here we consider the prediction of the 5-year relapse status of 214 women with breast cancer (80 relapse in the sample) (Guedj et al. 2012). Covariates correspond to the measurement of the $p = 54,613$ gene expressions reduced to the 248 genes differentially expressed (FDR=0.05), and we use 5-fold Monte Carlo cross validations to assess prediction performance. Interestingly, the expression of different genes is structured according to some unknown regulatory network that can be inferred from the data using Gaussian Graphical models for instance (Chiquet et al. 2009). Our hypothesis here is that using this inferred network can help in the prediction of patients outcome. However, since this regulatory network is not perfectly known, our strategy is based on the hypothesis that there is no "true" regulatory network, and we explore the robustness of the generalized fused lasso to the addition/removal of edges in the penalty, as we did in the simulation study. To proceed we consider the regulatory network that is inferred on the training data by the SIMoNE package (Chiquet et al. 2009). This package is based on sparse Gaussian Graphical models, and by varying the amount of shrinkage, we were able to consider networks with increasing number of edges, and then to assess the impact of changes in the network on prediction performance and estimated model dimensions.

The first conclusion is that the gain in using fused-based strategies is massive: the AUC (Area under the Curve) jumps from ∼0.7 for the lasso to ∼0.95 for generalized fused estimates, and the empirical error rate drops from ∼0.3 to ∼0.1, which clearly indicates that the network has helped in the correct classification of samples (Fig. 6). Moreover, as previously mentioned in the simulations, fused-based methods are more performant than the lasso, but there is no significant difference among them for prediction. Very interestingly, the classification performance remain with the addition of edges, which suggests that the suitability of these new edges is about $\theta \simeq 0.5$. Then the number of non-null estimated coefficients is higher for all fused-penalties, and the estimated number of distinct non-null coefficients (a crude estimate of quantity s_0) converges towards a set of ∼10 distinct values for estimated parameters.

6 Joint modeling to analyze road-safety data

Driving under the influence of alcohol (DUI) is an established risk factor of car accidents. Interestingly, several studies also suggest that DUI increases the risk of dying in an accident, but this result remains controversial: biological evidence supporting this assumption is still lacking and the observed effect of DUI could be due to confounding variables only. We present an application of the generalized fused lasso to the joint modeling framework, where the main objective is to study the effect of alcohol consumption on the risk of death (for drivers involved in a car crash). Our dataset consists of $n = 21{,}064$ drivers involved in reported single-car personal injury crashes from 2006 to 2009 in France (Onisr 2010). Current data show 33 covariates including the characteristics of the crash, of the drivers and of crash-involved vehicles, and we focus on the vital status of the driver only. We define 4 strata based on gender and DUI: strata 1–2 (resp. 3–4) for males and females not driving (resp. driving) under the influence when the accident occurred.

We consider logistic models to relate the probability of dying in a car accident to risk factors in each stratum. Since most factors are expected to share similar effects, joint modeling is used to couple the estimations of the four models. Intercept parameters are of particular importance: they should be homogeneous across strata if neither gender nor alcohol directly modified the risk of death. Then we compare the relaxed adaptive fused lasso, to the "*independent*" and "*interaction*" versions of the relaxed adaptive lasso (with stratum 1, i.e., sober males, as the reference stratum). We also present unpenalized estimates derived from standard logistic regression models independently built on each stratum (see Supplementary Fig. 6 for complete results).

All methods agree on the absence of effect of most factors on the risk of death in a car crash, and on the intensity of the effects of other factors, such as a higher risk for older drivers and a lower risk associated with the use of a seat-belt. Then inspecting the influence of city roads emphasizes interesting differences between the relaxed adaptive fused lasso and "*independent*" relaxed adaptive lassos in the presence of highly correlated variables (see Table 2). Indeed covariates "City Roads" (crash on a road managed by a city) and "City" (crash in a city) were both kept in the analysis despite redundancy. Thus, the global effect should be the sum of the two corresponding coefficients. Based on this global effect, all methods agree on the absence of interaction with gender and DUI (the sum of the two coefficients is roughly constant over the 4 strata according to every method). Accordingly, the relaxed adaptive fused lasso (as well as the "*interactions*" lasso) return equal individual effects for "City Roads" and "City" on every stratum. However, because these two variables are highly correlated, "*independent*" las-

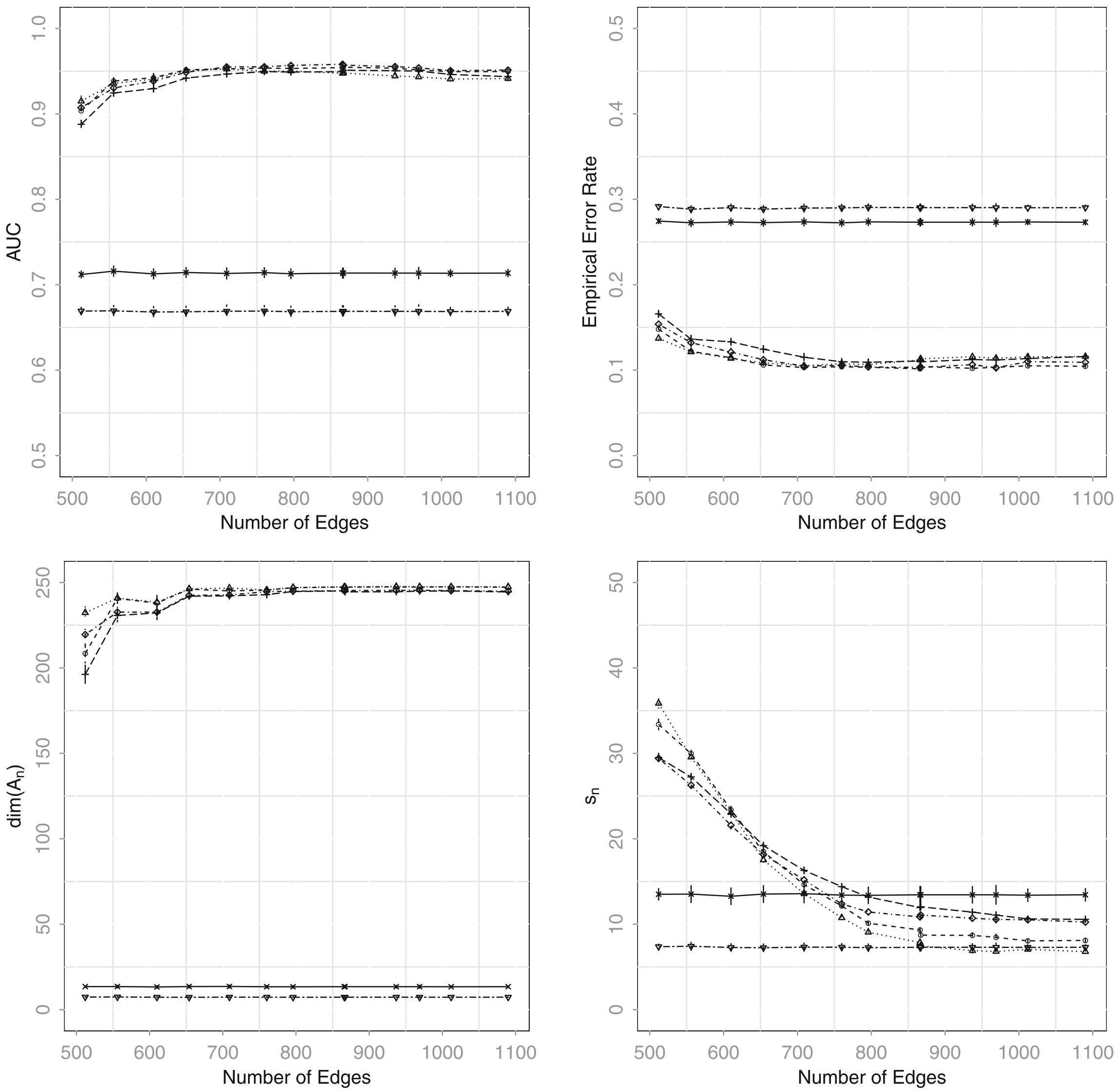

Fig. 6 Performance of penalization strategies for AUC (Area Under the Curve) and Empirical Error Rate on the breast cancer dataset with an increasing number of edges in the network provided in the penalty. Bottom panels display the estimated model size and complexities (number of non-null estimated coefficients, and number of distinct estimated coefficients) according to an increase in the number of edges in the penalty graph. Raw fused lasso (*opentriangle-dotted*), adaptive fused lasso (*opencircle-dashed*), relaxed fused lasso (*diamond-dotdash*), relaxed adaptive fused lasso (*plus*-longdashed), relaxed adaptive lasso (*times-solid*), relaxed Generalized Elastic Net (*triangledown-twodash*)

sos return widely different individual effect for “City Roads” and “City”, which does not make any sense. There are a few other differences between the three penalized methods, but when relaxed adaptive fused lasso and “*interactions*” relaxed adaptive lasso estimates disagree, the relaxed adaptive fused lasso most often agrees with “*independent*” relaxed adaptive lassos. This is consistent with our conclusions from the simulation study where these latter two methods generally performed the best in terms of support recovery.

Finally, the relaxed adaptive fused lasso agrees with the other methods on the fact that intercepts *do* vary across strata, suggesting an effect of both gender and DUI on the risk of death (see Table 2). More precisely, sober females are at a higher risk than sober males, and, to a lesser extent, females under the influence are at a higher risk than males under the influence. Moreover, irrespective on gender, drivers under the influence are at a higher risk than sober drivers. However this result should be tempered by potential confounding due to speed, which was not available here. For instance, drivers under the influence are likely to drive faster than sober drivers. The effects of speed may be partly captured by other covariates, but not entirely. Consequently residual “speed effects” could be responsible for detected heterogeneities between intercepts.

Table 2 Parameter estimates obtained on the road-safety data for the intercept terms and coefficients of variables “City” and “City roads”

Variable	Relaxed adaptive fused lasso				“*Independent*” relaxed adaptive lasso				“*Interaction*” relaxed adaptive lasso			
Intercept	−2.86	−3.89	−2.14	−2.14	−2.85	−3.86	−2.01	−2.22	−2.83	−3.67	−2.22	−2.39
City	−0.37	−0.37	−0.37	−0.37	−0.43	0.00	−0.50	−0.70	−0.39	−0.39	−0.39	−0.39
City roads	−0.21	−0.21	−0.21	−0.21	0.00	−0.54	0.00	0.00	−0.21	−0.21	−0.21	−0.21

For each method and each parameter, four values are given, corresponding to the estimate obtained on each of the four considered strata

7 Discussion

In this paper, we investigate theoretical and empirical properties of the generalized fused lasso, in various settings. From the theoretical point of view, we especially show that using adaptive weights leads to estimators enjoying asymptotic oracle properties. However, for the true underlying dimension of the problem d_0 (that is the number of distinct non-null values in $\boldsymbol{\beta}^*$) to be equal to the asymptotic dimension s_0 of the estimator, the graph G used in the penalty has to enjoy the following property: for all (j, ℓ) such that $\beta^*_j = \beta^*_\ell$, j and ℓ should belong to the same connected component of $G_\mathcal{B}$, the sub-graph of G such that $G_\mathcal{B} = (\mathcal{A}, \mathcal{B})$. In particular, our results indicate that setting G to the clique connecting all coefficients of $\boldsymbol{\beta}^*$ together (in which case all the $p(p-1)/2$ differences are penalized) is asymptotically optimal. Therefore, they confirm those obtained in the linear regression setting (She 2010), and extend them to generalized linear models. In words, it means that, asymptotically, adding misleading edges in the graph is harmless, while forgetting relevant ones can be harmful.

From the modeling point of view however, we empirically studied the robustness of generalized fused lasso estimates against graph misspecification on finite samples. On our experiments, we observed that adaptive weights and/or relaxation lead to some improved robustness. However, and overall, we demonstrated that the performance of generalized fused lasso estimates on finite samples are deeply related to the suitability of the graph in the penalty, especially for support recovery. In particular, we show that, under the designs considered in our simulations, the clique-based strategy is clearly sub-optimal for support recovery, so that misleading edges are harmful on finite samples. The graph used in the penalty constitutes a formal description of some prior knowledge on the problem that is investigated, and has to be determined with caution, especially if support recovery matters. We may stress that this graph does not describe correlations among features but similarity between their effects under the considered model. Of course, correlated features may share similar coefficients, but not necessarily. For instance in epidemiological studies, smoking and alcohol consumption are generally highly correlated. They further may share similar effects under a logistic model when studying cardiovascular diseases so that it might make sense to penalize the difference of their effects. However, when studying lung cancer, they are not expected to share similar coefficients at all, so that their difference should not be penalized. That being said, in our application of Sect. 5, using a graph based on (partial) correlation still leads to highly improved prediction accuracy compared to the Lasso for instance (of course, support recovery performance can not be assessed on real data).

We further show that generalized fused lasso can be useful even in situations where theoretical coefficients are not exactly equal. Interestingly, it can improve upon state-of-the-art competing methods, such as the group lasso in this context, even when the graph is not perfectly suited.

A particular situation where the graph is suggested by the design of the study itself is the joint modeling framework, where data come from various strata. When the main question is the detection of heterogeneities across strata, we believe that the graph made of cliques is very appealing as it encourages coefficients to be homogeneous across the strata. This strategy has some connections with the statistical tests theory where tests are generally performed under the null hypothesis (absence of heterogeneity in this case), and data need to be far enough from this assumption for the null hypothesis to be rejected. But even in the joint modeling context, other graphs may be considered: for instance, if strata correspond to various treatments, the control treatment can serve as the reference and star-graphs may be more appropriate than cliques (see Section 1 in the Supplementary Material).

We established the asymptotic oracle properties of the adaptive generalized fused lasso estimates under generalized linear models, for fixed p. These results are the first established for fused lasso estimates in the setting of generalized linear models and for the generalized fused penalty based on a graph. Even if the fixed p case is relevant, especially in the joint modeling framework, they should be extended to cover the high-dimensional case. Most published papers on the fused lasso in high-dimension focus on the chain-based fused penalty in the Gaussian sequence model. A notable exception is the work of Vaiter et al. (2013) in which results that encompass generalized fused lassos under a bounded noise assumption were recently established. The extension of such results to the random noise case and generalized linear models would be an interesting lead. Moreover, non-asymptotic oracle prediction inequalities have been recently obtained for the fused lasso and piecewise constant functions

(Dalayan et al. 2014). Extending this result to more general contexts is not straightforward and constitutes a promising research direction as well.

Acknowledgments Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24, by the MSTIC project of the Joseph-Fourier University, and by the ABS4NGS ANR project (ANR-11-BINF-0001-06). The authors thank M. Guedj for his insight on the breast cancer dataset, and J. Chiquet for discussions on the SIMoNE package. This work was performed using the computing facilities of the CC LBBE/PRABI.

Appendix

Throughout our proofs, we will make frequent use of the sign function $sign$ defined on $\mathbb{R}^*$ as $sign(u) = +1$ if $u > 0$ and $sign(u) = -1$ if $u < 0$.

Proof of Theorem 1

This proof is an adaptation of the proof given by Tibshirani et al. (2005) to account for the generalized linear model loss and the generalized fused lasso penalty. Let us define $\mathcal{V}_n(\mathbf{u}) = Q(\boldsymbol{\beta}^* + \mathbf{u}/\sqrt{n}) - Q(\boldsymbol{\beta}^*)$ with $\mathbf{u} = (u_0, \ldots, u_p)^T$, and Q defined as in Sect. 2.1. Obviously $\mathcal{V}_n(\mathbf{u})$ is minimized at $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Similarly to Tibshirani et al. (2005), we obtain:

$$\begin{aligned}\mathcal{V}_n(\mathbf{u}) = J\left(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}\right) - J(\boldsymbol{\beta}^*)\\ +\lambda_n^{(1)} \sum_{j=1}^{p} \left\{\left|\beta_j^* + \frac{u_j}{\sqrt{n}}\right| - |\beta_j^*|\right\}\\ +\lambda_n^{(2)} \sum_{(j,\ell)\in E} \left\{\left|\beta_j^* - \beta_\ell^* + \frac{(u_j - u_\ell)}{\sqrt{n}}\right| - |\beta_j^* - \beta_\ell^*|\right\}.\end{aligned}$$

For any fixed $\mathbf{u}$, the last two terms of the right-hand side converge to the last two terms in expression (1) of $\mathcal{V}(\mathbf{u})$ as n goes to ∞. As for the first two terms, a Taylor expansion yields

$$\begin{aligned}J\left(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}\right) - J(\boldsymbol{\beta}^*) = \nabla J(\boldsymbol{\beta}^*)^T \frac{\mathbf{u}}{\sqrt{n}}\\ +\frac{1}{2}\mathbf{u}^T \frac{\mathcal{I}(\boldsymbol{\beta}^*)}{n}\mathbf{u} + o_{\mathbb{P}}(1/n).\end{aligned}$$

Under **AL1**, we have $\mathbf{u}^T(\mathcal{I}(\boldsymbol{\beta}^*)/n)\mathbf{u} \to_d \mathbf{u}^T\mathbf{C}\mathbf{u}$. Moreover, **AL1** implies that the minimum eigenvalue of $\mathcal{I}(\boldsymbol{\beta}^*)$ goes to ∞ and, under **AL2**, we have $(\nabla J(\boldsymbol{\beta}^*)/\sqrt{n}) \to_d \mathbf{W}$, where $\mathbf{W}$ has an $\mathcal{N}(\mathbf{0}_{p+1}, \mathbf{C})$ distribution (Gourieroux and Monfort 1981). By Slutsky's theorem, we therefore have $\mathcal{V}_n(\mathbf{u}) \to_d \mathcal{V}(\mathbf{u})$. Since $\mathcal{V}_n$ is convex, the epi-convergence results of Geyer (1994) [see in particular Theorem 5 in Knight (1999)] can finally be used to complete the proof of Theorem 1.

Proof of Proposition 1

If $\lambda_0^{(2)} = 0$, the proof is the same as in the lasso case (Zou 2006). Now assume that $\lambda_0^{(2)} \neq 0$ (we further assume that $\lambda_0^{(1)} \neq 0$, the case where $\lambda_0^{(1)} = 0$ being slightly easier and omitted). First observe that $\mathbb{P}(\widetilde{\mathcal{A}}_n = \mathcal{A}) \leq \mathbb{P}(\sqrt{n}\hat{\beta}_j = 0\ \forall j \notin \mathcal{A})$. Moreover, in virtue of Theorem 1, we have $\limsup_n \mathbb{P}(\sqrt{n}\hat{\beta}_j = 0\ \forall j \notin \mathcal{A}) \leq \mathbb{P}(u_j^* = 0\ \forall j \notin \mathcal{A})$, with $\mathbf{u}^* = \text{argmin}(\mathcal{V})$. Therefore, we only need to show that $c = \mathbb{P}(u_j^* = 0\ \forall j \notin \mathcal{A}) < 1$.
For any $j \in \{1, \ldots, p\}$, introduce $E_j^{=}(\boldsymbol{\beta}^*) = \{\ell : (\ell, j) \in E \text{ or } (j, \ell) \in E \text{ and } \beta_j^* = \beta_\ell^*\}$ and $E_j^{\neq}(\boldsymbol{\beta}^*) = \{\ell : (\ell, j) \in E \text{ or } (j, \ell) \in E \text{ and } \beta_j^* \neq \beta_\ell^*\}$. Setting $\mathbf{W} = (W_0, \ldots, W_p)^T$ and $\mathbf{Cu}^* = ((\mathbf{Cu}^*)_0, \ldots, (\mathbf{Cu}^*)_p)^T$, we have, by the KKT conditions,

$$W_0 + (\mathbf{Cu}^*)_0 = 0, \tag{1}$$

and for all $j \in \mathcal{A}$,

$$\begin{aligned}W_j + (\mathbf{Cu}^*)_j + \lambda_0^{(1)} sign(\beta_j^*) + \lambda_0^{(2)}\Bigg\{ \sum_{\ell\in E_j^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_j^* - \beta_k^*)\\ + \sum_{\ell\in E_j^{=}(\boldsymbol{\beta}^*)} (-1)^{\mathbb{I}(j<\ell)} t_{j\ell}\Bigg\} = 0,\end{aligned}$$

and for all $j \notin \mathcal{A}$,

$$\begin{aligned}W_j + (\mathbf{Cu}^*)_j + \lambda_0^{(1)} r_j + \lambda_0^{(2)}\Bigg\{ \sum_{\ell\in E_j^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_j^* - \beta_k^*)\\ + \sum_{\ell\in E_j^{=}(\boldsymbol{\beta}^*)} (-1)^{\mathbb{I}(j<\ell)} t_{j\ell}\Bigg\} = 0.\end{aligned}$$

Above, $r_j = sign(u_j^*)$ if $u_j^* \neq 0$ and r_j is some real number in $[-1, 1]$ otherwise. Similarly, $t_{j\ell} = sign(u_j^* - u_\ell^*)$ if $u_j^* \neq u_\ell^*$ and $t_{j\ell}$ is some real number in $[-1, 1]$ otherwise.

For any index $j \in \mathcal{A}$ there is some $s = s(j)$ such that $j \in \mathcal{A}_{s(j)}$, where $\mathcal{A}_s$ still denotes the set of vertices of the s-th connected component of $G_{\mathcal{B}}$ (see the paragraph before the statement of Theorem 2 for the definitions of these objects). Then summing up the KKT conditions over $\mathcal{A}_{s(j)}$, we have

$$\begin{aligned}\sum_{k\in\mathcal{A}_{s(j)}} \Bigg\{W_k + (\mathbf{Cu}^*)_k + \lambda_0^{(1)} sign(\beta_k^*)\\ +\lambda_0^{(2)} \sum_{\ell\in E_k^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_k^* - \beta_\ell^*)\Bigg\} = 0.\end{aligned} \tag{2}$$

Similarly, setting $\widetilde{\mathcal{B}} = \{(j, \ell) \in E \cap \bar{\mathcal{A}} \times \bar{\mathcal{A}}\}$ and denoting by $G_0 = (\bar{\mathcal{A}}, \widetilde{\mathcal{B}})$, the set $\bar{\mathcal{A}}$ can be decomposed as $\bar{\mathcal{A}} = \cup_{s=1}^{s_1} \bar{\mathcal{A}}_s$, where $1 \leq s_1 \leq p - p_0$ and $\bar{\mathcal{A}}_s$ is the subset of vertices constituting the s-th connected component of G_0. Then, for any $j \notin \mathcal{A}$, there exists some $s = s(j)$ such that $j \in \bar{\mathcal{A}}_{s(j)}$ and summing up the KKT optimality conditions over $\bar{\mathcal{A}}_{s(j)}$, we have

$$\sum_{k\in\bar{\mathcal{A}}_{s(j)}}\Big\{W_k+(\mathbf{Cu}^*)_k+\lambda_0^{(1)}r_k +\lambda_0^{(2)}\sum_{\ell\in E_k^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_k^*-\beta_\ell^*)\Big\}=0, \tag{3}$$

with $|r_k|\leq 1$. If $u_j^*=0$ for all $j\notin\mathcal{A}$, equations (2) along with Eq. (1) form a system of s_0+1 equations with $p_0+1\geq s_0+1$ variables, that can be written as

$$\mathbf{W}_{\mathcal{B}}+\mathbf{M}_1\mathbf{u}^*_{\{0\}\cup\mathcal{A}}+\lambda_0^{(1)}\mathbf{r}_{\mathcal{B}}+\lambda_0^{(2)}\mathbf{t}_{\mathcal{B}}=\mathbf{0},$$

where $\mathbf{M}_1$ is the $(s_0+1)\times(p_0+1)$ matrix whose (s,j) element is $m_{s,j}=\sum_{k\in\mathcal{A}_{s-1}}C_{k,j}$ (with $C_{k,j}$ the (k,j) element of $\mathbf{C}$ and $\mathcal{A}_0=\{0\}$), and $\mathbf{W}_{\mathcal{B}}$, $\mathbf{r}_{\mathcal{B}}$ and $\mathbf{t}_{\mathcal{B}}$ are vectors in $\mathbb{R}^{s_0+1}$ whose s-th elements are $\sum_{k\in\mathcal{A}_{s-1}}W_k$, $\sum_{k\in\mathcal{A}_{s-1}}sign(\beta_k^*)$ and $\sum_{k\in\mathcal{A}_{s-1}}\sum_{\ell\in E_k^{\neq}(\boldsymbol{\beta}^*)}sign(\beta_k^*-\beta_\ell^*)$ respectively. Now, denoting by $\mathbf{M}_1^\dagger$ the pseudo-inverse of $\mathbf{M}_1$, there exists some vector $\boldsymbol{\omega}\in\mathbb{R}^{p_0+1}$ such that $\mathbf{u}^*_{\{0\}\cup\mathcal{A}}=(\mathbf{I}_{p_0+1}-\mathbf{M}_1^\dagger\mathbf{M}_1)\boldsymbol{\omega}-\mathbf{M}_1^\dagger(\mathbf{W}_{\mathcal{B}}+\lambda_0^{(1)}\mathbf{r}_{\mathcal{B}}+\lambda_0^{(2)}\mathbf{t}_{\mathcal{B}})$. Therefore, if $u_j^*=0$ for all $j\notin\mathcal{A}$, Eq. (3) form a system of s_1 equations that can be written,

$$\Big|\mathbf{W}_{\bar{\mathcal{B}}}+\lambda_0^{(2)}\mathbf{t}_{\bar{\mathcal{B}}}+\mathbf{M}_2\big\{(\mathbf{I}_{p_0}-\mathbf{M}_1^\dagger\mathbf{M}_1)\boldsymbol{\omega} -\mathbf{M}_1^\dagger(\mathbf{W}_{\mathcal{B}}+\lambda_0^{(1)}\mathbf{r}_{\mathcal{B}}+\lambda_0^{(2)}\mathbf{t}_{\mathcal{B}})\big\}\Big|\leq\lambda_0^{(1)}\mathbf{r}_{\bar{\mathcal{B}}},$$

where $\mathbf{W}_{\bar{\mathcal{B}}}$, $\mathbf{r}_{\bar{\mathcal{B}}}$ and $\mathbf{t}_{\bar{\mathcal{B}}}$ are the vectors in $\mathbb{R}^{s_1}$ whose s-th elements are given by $\sum_{k\in\bar{\mathcal{A}}_s}W_k$, $|\bar{\mathcal{A}}_s|$ and

$$\sum_{k\in\bar{\mathcal{A}}_s}\sum_{\ell\in E_k^{\neq}(\boldsymbol{\beta}^*)}sign(\beta_k^*-\beta_\ell^*)$$

respectively. We can now conclude by observing that

$$c\leq\mathbb{P}\Big(\Big|\mathbf{W}_{\bar{\mathcal{B}}}+\lambda_0^{(2)}\mathbf{t}_{\bar{\mathcal{B}}}+\mathbf{M}_2\big\{(\mathbf{I}_{p_0}-\mathbf{M}_1^\dagger\mathbf{M}_1)\boldsymbol{\omega} -\mathbf{M}_1^\dagger(\mathbf{W}_{\mathcal{B}}+\lambda_0^{(1)}\mathbf{r}_{\mathcal{B}}+\lambda_0^{(2)}\mathbf{t}_{\mathcal{B}})\big\}\Big|\leq\lambda_0^{(1)}\mathbf{r}_{\bar{\mathcal{B}}}\Big)<1.$$

Proof of Theorem 2

The following proof is a modification to the proof given by Zou (2006) to account for both the generalized linear model loss and the generalized fused penalty. Let us define $\mathbf{V}_n(\mathbf{u})=Q(\boldsymbol{\beta}^*+\mathbf{u}/\sqrt{n})-Q(\boldsymbol{\beta}^*)$ with $\mathbf{u}=(u_0,\ldots,u_p)^T$ and Q defined as Sect. 2.1. Note that $\mathbf{V}_n(\mathbf{u})$ is minimized at $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ad}-\boldsymbol{\beta}^*)$. We have

$$\begin{aligned}\mathbf{V}_n(\mathbf{u})&=\nabla J(\boldsymbol{\beta}^*)^T\frac{\mathbf{u}}{\sqrt{n}}+\frac{1}{2}\mathbf{u}^T\frac{\mathcal{I}(\boldsymbol{\beta}^*)}{n}\mathbf{u}+o_{\mathbb{P}}(1/n)\\&\quad+\frac{\lambda_n^{(1)}}{\sqrt{n}}\sum_{j=1}^p w_j^{(1)}\sqrt{n}\left\{\left|\beta_j^*+\frac{u_j}{\sqrt{n}}\right|-|\beta_j^*|\right\}\\&\quad+\frac{\lambda_n^{(2)}}{\sqrt{n}}\sum_{(j,\ell)\in E}w_{j\ell}^{(2)}\sqrt{n}\left\{\left|\beta_j^*-\beta_\ell^*+\frac{(u_j-u_\ell)}{\sqrt{n}}\right|-|\beta_j^*-\beta_\ell^*|\right\}\\&=:\nabla J(\boldsymbol{\beta}^*)^T\frac{\mathbf{u}}{\sqrt{n}}+\frac{1}{2}\mathbf{u}^T\frac{\mathcal{I}(\boldsymbol{\beta}^*)}{n}\mathbf{u}+o_{\mathbb{P}}(1/n)\\&\quad+\sum_{j=1}^p T_j^{(1)}+\sum_{(j,\ell)\in E}T_{j\ell}^{(2)}.\end{aligned}$$

We have the two following behaviors :

$$T_j^{(1)}\rightarrow_p\begin{cases}0 & \text{if } \beta_j^*\neq 0 \text{ or } (\beta_j^*=0 \text{ and } u_j=0)\\ \infty & \text{otherwise}\end{cases}$$

and

$$T_{j\ell}^{(2)}\rightarrow_p\begin{cases}0 & \text{if } \beta_j^*\neq\beta_\ell^* \text{ or } (\beta_j^*=\beta_\ell^* \text{ and } u_j=u_\ell)\\ \infty & \text{otherwise.}\end{cases}$$

Denote by $\mathbf{C}_{\mathcal{A}}$ the $(p_0+1)\times(p_0+1)$ sub-matrix of $\mathbf{C}$ constituted of rows and columns associated with indexes in $\{0\}\cup\mathcal{A}$ and by $\mathbf{W}_{\mathcal{A}}$ a random Gaussian vector $\mathcal{N}\left(\mathbf{0}_{p_0+1},\mathbf{C}_{\mathcal{A}}\right)$. Then, as in the proof of Theorem 1, $\mathbf{V}_n(\mathbf{u})\rightarrow_d\mathbf{V}(\mathbf{u})$ for every $\mathbf{u}$, with $\mathbf{V}$ defined for $\mathbf{u}=(u_0,\ldots,u_p)\in\mathbb{R}^{p+1}$, by

$$\mathbf{V}(\mathbf{u})=\begin{cases}\frac{1}{2}\mathbf{u}_{\mathcal{A}}^T\mathbf{C}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}}+\mathbf{u}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} & \text{if } u_j=0 \text{ for } j\notin\mathcal{A} \text{ and}\\ & u_j=u_\ell \text{ for } (j,\ell)\in\mathcal{B},\\ \infty & \text{otherwise.}\end{cases}$$

Recall the notations introduced just before stating Theorem 2. Any vector $\mathbf{u}\in\mathbb{R}^{p+1}$ such that $u_j=0$ for all $j\notin\mathcal{A}$ and $u_j=u_\ell$ for all $(j,\ell)\in\mathcal{B}$ has s_0+1 distinct non-zero values. Denoting by $u_0,u_{j_1},\ldots,u_{j_{s_0}}$ these values, and setting $\mathbf{u}_{\mathcal{B}}=(u_0,u_{j_1},\ldots,u_{j_{s_0}})^T$, we have

$$\mathbf{V}(\mathbf{u})=\begin{cases}\frac{1}{2}\mathbf{u}_{\mathcal{B}}^T\mathbf{C}_{\mathcal{B}}\mathbf{u}_{\mathcal{B}}+\mathbf{u}_{\mathcal{B}}^T\mathbf{W}_{\mathcal{B}} & \text{if } u_j=0 \text{ for } j\notin\mathcal{A} \text{ and}\\ & u_j=u_\ell \text{ for } (j,\ell)\in\mathcal{B},\\ \infty & \text{otherwise,}\end{cases}$$

where $\mathbf{W}_{\mathcal{B}}\sim\mathcal{N}\left(\mathbf{0}_{s_0+1},\mathbf{C}_{\mathcal{B}}\right)$. Clearly, $\mathbf{V}$ has a unique minimum for $\mathbf{u}\in\mathbb{R}^{p+1}$ such that $u_j=0$ for all $j\notin\mathcal{A}$ and $u_j=u_\ell$ for all $(j,\ell)\in\mathcal{B}$ and $\mathbf{u}_{\mathcal{B}}=-\mathbf{C}_{\mathcal{B}}^{-1}\mathbf{W}_{\mathcal{B}}$. Since $\mathbf{V}_n$ is convex we can proceed by using the epi-convergence results (Geyer 1994) to prove the asymptotic normality part (Zou 2006).

Let us now turn our attention to the variable selection consistency. Namely, we have to show that $\forall j\in\mathcal{A}$, $\mathbb{P}(j\in\mathcal{A}_n)\rightarrow 1$ and that $\forall j\notin\mathcal{A}$, $\mathbb{P}(j\in\mathcal{A}_n)\rightarrow 0$. The first claim is an easy consequence of the previous asymptotic result (Zou 2006). To prove the second claim, consider an index $j\notin\mathcal{A}$ and denote by C_j the subset of vertices constituting the connected component of G to which j belongs. Let $C_j^0=$

$\{\ell \in C_j,\ \beta^*_\ell = 0\}$; clearly, $j \in C^0_j$. Our aim is to prove that $\mathbb{P}(\ell \in \mathcal{A}_n) \to 0$, for all $\ell \in C^0_j$. Observe that the KKT conditions write, for $k = 1, \dots, p$:

$$\left[\nabla J(\widehat{\boldsymbol{\beta}}^{ad})\right]_k = \lambda^{(1)}_n w^{(1)}_k r_k + \lambda^{(2)}_n \left(\sum_{(k,\ell)\in E} w^{(2)}_{k\ell} t_{k\ell} - \sum_{(\ell,k)\in E} w^{(2)}_{k\ell} t_{\ell k} \right)$$

where $r_k = sign(\widehat{\beta}^{ad}_k)$ for $\widehat{\beta}^{ad}_k \neq 0$ and r_k is some real number in $[-1, 1]$ if $\widehat{\beta}^{ad}_k = 0$; likewise, for any $(k, \ell) \in E$, $t_{k\ell} = sign(\widehat{\beta}^{ad}_k - \widehat{\beta}^{ad}_\ell)$ for $\widehat{\beta}^{ad}_k \neq \widehat{\beta}^{ad}_\ell$ and $t_{k\ell}$ is some real number in $[-1, 1]$ if $\widehat{\beta}^{ad}_k = \widehat{\beta}^{ad}_\ell$. Introducing the set $\tilde{E} = \{(k, \ell) : (k, \ell) \in E \text{ or } (\ell, k) \in E\}$, and setting $t_{k\ell} = -t_{\ell k}$ for $(\ell, k) \in E$, we have the following more compact form for the KKT conditions:

$$\left[\nabla J(\widehat{\boldsymbol{\beta}}^{ad})\right]_k = \lambda^{(1)}_n w^{(1)}_k r_k + \lambda^{(2)}_n \sum_{(k,\ell)\in\tilde{E}} w^{(2)}_{k\ell} t_{k\ell},$$

where, in particular, $t_{k\ell} = sign(\widehat{\beta}^{ad}_k - \widehat{\beta}^{ad}_\ell)$ for any $(k, \ell) \in \tilde{E}$ such that $\widehat{\beta}^{ad}_k \neq \widehat{\beta}^{ad}_\ell$. Next, since $(\nabla J(\boldsymbol{\beta}^*)/\sqrt{n}) \to_d \mathbf{W}$ (as shown in the proof of Theorem 1), the assumption **AL1** enables us to show that $M_n(k) := [\nabla J(\widehat{\boldsymbol{\beta}}^{ad})]_k/\sqrt{n} = O_{\mathbb{P}}(1)$ as $n \to \infty$ as well.

Let us now suppose that there exist some $\ell \in C^0_j$ such that $\widehat{\beta}^{ad}_\ell \neq 0$. In this case, either the set $\mathcal{S}_{neg} = \{\ell \in C^0_j : \widehat{\beta}^{ad}_\ell < 0\}$ or the set $\mathcal{S}_{pos} = \{\ell \in C^0_j : \widehat{\beta}^{ad}_\ell > 0\}$ is not empty (or both sets are not empty). If $\mathcal{S}_{neg} \neq \varnothing$, let $b^{\min} = \min_{k\in\mathcal{S}_{neg}} \widehat{\beta}^{ad}_k$. Further denote by L the subset of $\mathcal{S}_{neg}$ of connected indices ℓ such that $\widehat{\beta}^{ad}_\ell = b^{\min}$. Since $S_{neg} \neq \varnothing$, L has at least one element. Then, summing up the KKT conditions over L, we obtain

$$\begin{aligned}\sum_{k\in L} M_n(k) = {} & \frac{\lambda^{(1)}_n}{\sqrt{n}} n^{\gamma/2} \sum_{k\in L} \frac{r_k}{|\sqrt{n}\widetilde{\beta}_k|^\gamma} \\ & + \frac{\lambda^{(2)}_n}{\sqrt{n}} \sum_{k\in L} \sum_{(k,\ell)\in\tilde{E},\ \beta^*_\ell\neq 0} \frac{t_{k\ell}}{|\widetilde{\beta}_k - \widetilde{\beta}_\ell|^\gamma} \\ & + \frac{\lambda^{(2)}_n}{\sqrt{n}} n^{\gamma/2} \sum_{k\in L} \sum_{\substack{(k,\ell)\in\tilde{E} \\ \beta^*_\ell=0\ \&\ \widehat{\beta}^{ad}_\ell > b^{\min}}} \frac{t_{k\ell}}{|\sqrt{n}(\widetilde{\beta}_k - \widetilde{\beta}_\ell)|^\gamma}.\end{aligned}$$

Since $L \subset \mathcal{S}_{neg}$, $r_k = -1$, for all $k \in L$, and by definition of L, $t_{k\ell} = -1$ for all ℓ such that $\widehat{\beta}^{ad}_\ell \neq b^{\min}$. Moreover when $\beta^*_\ell \neq 0$ then $\beta^*_\ell \neq \beta^*_k$, and $\lambda^{(2)}_n t_{k\ell}/(\sqrt{n}|\widetilde{\beta}_k - \widetilde{\beta}_\ell|^\gamma) \to_{\mathbb{P}} 0$, as n goes to ∞. Since $\lambda^{(m)}_n n^{\gamma/2}/\sqrt{n}$ $(m = 1, 2)$ tends to ∞, $\sum_{k\in L} M_n(k)$ tends to $-\infty$, which contradicts $M_n(\ell) = O_{\mathbb{P}}(1)$ for all $\ell = 1, \dots, p$. That leads to $\mathbb{P}(\mathcal{S}_{neg} = \varnothing) \to 1$. If $S_{neg} = \varnothing$, then $S_{pos} \neq \varnothing$, and similar arguments can be used (with maxima instead of minima) to get a contradiction. Putting all this together, we conclude that for all $\ell \in C^0_j$, $\mathbb{P}(\ell \in \mathcal{A}_n) \to 0$.

It remains to show the consistency for the set $\mathcal{B}_n$. As for $\mathcal{A}_n$, we need to prove that $\forall (j, \ell) \notin \mathcal{B}$, $\mathbb{P}((j, \ell) \in \mathcal{B}^c_n) \to 1$ and that $\forall (j, \ell) \in \mathcal{B}$, $\mathbb{P}((j, \ell) \in \mathcal{B}^c_n) \to 0$. Let us prove the first claim. If $(j, \ell) \notin \mathcal{B}$ either $(\beta^*_j = 0$ and/or $\beta^*_\ell = 0)$, or $(\beta^*_j \neq 0,\ \beta^*_\ell \neq 0$ and $\beta^*_j \neq \beta^*_\ell)$. In the first case, when $j \in \mathcal{A}^c$, we have proved previously that $\mathbb{P}(j \in \mathcal{A}^c_n) \to 1$, so $\mathbb{P}((j, \ell) \in \mathcal{B}^c_n) \to 1$. In the second case, if $(j,\ \ell) \in \mathcal{A}$, and $(j, \ell) \notin \mathcal{B}$, the asymptotic normality result indicates that $\widehat{\beta}^{ad}_j - \widehat{\beta}^{ad}_\ell \to_{\mathbb{P}} \beta^*_j - \beta^*_\ell \neq 0$; thus $\mathbb{P}((j, \ell) \in \mathcal{B}^c_n) \to 1$. Now let us prove the second claim, using KKT conditions as before. Let j be an index of $\mathcal{A}$ such that for some $\ell \in \mathcal{A}$ we have $(j, \ell) \in \mathcal{B}$. Then, for some $1 \leq s(j) \leq s_0$, $j \in \mathcal{A}_{s(j)}$, where $\mathcal{A}_s$ still denotes the set of vertices of the s-th connected component of $G_{\mathcal{B}}$. Suppose that there exists some $\ell \in \mathcal{A}_{s(j)}$ such that $\widehat{\beta}^{ad}_\ell \neq \widehat{\beta}^{ad}_j$. As previously we define $b^{\min} = \min_{k\in\mathcal{A}_{s(j)}} \widehat{\beta}^{ad}_k$ and L the subset of $\mathcal{A}_{s(j)}$ of connected indices ℓ such that $\widehat{\beta}^{ad}_\ell = b^{\min}$. Then, summing up the KKT conditions over L, we obtain

$$\begin{aligned}\sum_{k\in L} M_n(k) = {} & \frac{\lambda^{(1)}_n}{\sqrt{n}} \sum_{k\in L} \frac{r_k}{|\widetilde{\beta}_k|^\gamma} \\ & + \frac{\lambda^{(2)}_n}{\sqrt{n}} \sum_{k\in L} \sum_{(k,\ell)\in\tilde{E},\ \beta^*_\ell\neq\beta^*_k} \frac{t_{k\ell}}{|\widetilde{\beta}_k - \widetilde{\beta}_\ell|^\gamma} \\ & + \frac{\lambda^{(2)}_n}{\sqrt{n}} n^{\gamma/2} \sum_{k\in L} \sum_{\substack{(k,\ell)\in\tilde{E} \\ \beta^*_\ell=\beta^*_k\ \&\ \widehat{\beta}^{ad}_\ell > b^{\min}}} \frac{t_{k\ell}}{|\sqrt{n}(\widetilde{\beta}_k - \widetilde{\beta}_\ell)|^\gamma}.\end{aligned}$$

Since $L \subset \mathcal{A}$, the first sum converges to 0 in probability. Moreover, the second sum also converges to 0 in probability, while the third sum tends to $-\infty$, which contradicts $M_n(\ell) = O_{\mathbb{P}}(1)$ for all $\ell = 1, \dots, p$. We therefore conclude that $\mathbb{P}((j, \ell) \in \mathcal{B}^c_n) \to 0$, for all $(j, \ell) \in \mathcal{B}$, which completes the proof of Theorem 2.

References

Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Mach. Learn. **73**(3), 243–272 (2008)

Azencott, C.A., Grimm, D., Sugiyama, M., Kawahara, Y., Borgwardt, K.M.: Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics **29**(13), i171–179 (2013)

Chiquet, J., Smith, A., Grasseau, G., Matias, C., Ambroise, C.: SIMoNe: statistical inference for modular networks. Bioinformatics **25**(3), 417–418 (2009)

Dalalyan, A., Hebiri, M., Lederer, J.: On the Prediction Performance of the Lasso. Arxiv preprint arXiv:1402.1700 (2014)

Danaher, P., Wang, P., Witten, D.: The joint graphical lasso for inverse covariance estimation across multiple classes. J. R. Stat. Soc. Ser. B **76**(2), 373–397 (2014)

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., et al.: String v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. **41**(D1), D808–D815 (2013)

Gertheiss, J., Tutz, G.: Regularization and model selection with categorial effect modifiers. Stat. Sin. **22**, 957–982 (2012)

Geyer, C.J.: On the asymptotics of constrained M-estimation. Ann. Stat. **22**, 1993–2010 (1994)

Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA. **99**(12), 7821–7826 (2002)

Gourieroux, C., Monfort, A.: Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. J. Econom. **17**, 83–97 (1981)

Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., Bertrand, F., Bonnet, F., Martin, A.L., Feugeas, J.P., Bieche, I., Lehmann-Che, J., Lidereau, R., Birnbaum, D., Bertucci, F., de The, H., Theillet, C.: A refined molecular taxonomy of breast cancer. Oncogene **31**(9), 1196–1206 (2012)

Han, J.: Construction and analysis of web-based computer science information networks. In: Kuznetsov, S.O., Slezak, D., Hepting,D.H., Mirkin, B.G. (eds.). Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Lecture Notes in Computer Science, vol. 6743, pp. 1–2. Springer, Heidelberg (2011)

Höfling, H., Binder, H., Schumacher, M.: A coordinate-wise optimization algorithm for the Fused Lasso. Arxiv preprint arXiv:1011.6409 (2010)

Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high dimensional models. Stat. Sci. **27**(4), 481–499 (2012)

Knight, K.: Epi-convergence in distribution and stochastic equisemicontinuity. (unpublished manuscript) (1999)

McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman & Hall, New-York (1989)

Meier, L., van de Geer, S., Bühlmann, P.: The group lasso for logistic regression. J. R. Stat. Soc. Ser. B **70**(1), 53–71 (2008)

Meinshausen, N.: Relaxed lasso. Comput. Stat. Data Anal. **52**(1), 374–393 (2007)

Onisr: La sécurité routière en France, bilan de l'année 2009. Paris: La documentation Française (2010)

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M.: Towards a proteome-scale map of the human protein-protein interaction network. Nature **437**(7062), 1173–1178 (2005)

Sharma, D., Bondell, H., Zhang, H.: Consistent group identification and variable selection in regression with correlated predictors. J. Comput. Gr. Stat. **22**, 319–340 (2013)

She, Y.: Sparse regression with exact clustering. Electron. J. Stat. **4**, 1055–1096 (2010)

Sun, H., Wang, S.: Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics **28**(10), 1368–1375 (2012)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 267–288 (1996)

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B **67**, 91–108 (2005)

Vaiter, S., Peyré, G., Dossal, C., Fadili, J.: Robust sparse analysis regularization. IEEE Trans. Inf. Theory **59**(4), 2001–2016 (2013). doi:10.1109/TIT.2012.2233859.

Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)