Sparse estimation for case-control studies with multiple disease subtypes

Nadim Ballout, Cédric Garcia, Vivian Viallon

Sparse estimation for case-control studies with multiple disease subtypes.

Journal:	*Biostatistics*
Manuscript ID	BIOSTS-19034.R2
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	BALLOUT, Nadim; IFSTTAR Lyon-Bron; Universite Claude Bernard Lyon 1 Garcia, Cedric; IFSTTAR Departement Amenagement Mobilites et Environnement Viallon, Vivian; International Agency for Research on Cancer
Keywords:	Conditional logistic regression, Multinomial logistic regression, Lasso, Sparsity, Structured sparsity
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
MainManuscript.zip SupplementaryMaterials.zip	

Biostatistics (2019), **0**, 0, *pp.* 1–24
doi:10.1093/biostatistics/Ballout-Garcia-Viallon-MainManuscript˙rev

Sparse estimation for case-control studies with multiple disease subtypes.

Nadim Ballout*,[1], Cedric Garcia[2],Vivian Viallon[3]

[1]*IFSTTAR, UMRESTTE, Université Claude Bernard Lyon 1, Lyon, FRANCE.*

[2]*IFSTTAR, AME, DEST, Marne la vallée, FRANCE.*

[3]*IARC, WHO, Lyon, FRANCE.*

nadim.ballout@ifsttar.fr

Summary

The analysis of case-control studies with several disease subtypes is increasingly common, e.g. in cancer epidemiology. For matched designs, a natural strategy is based on a stratified conditional logistic regression model. Then, to account for the potential homogeneity among disease subtypes, we adapt the ideas of data shared lasso, which has been recently proposed for the estimation of stratified regression models. For unmatched designs, we compare two standard methods based on L_1-norm penalized multinomial logistic regression. We describe formal connections between these two approaches, from which practical guidance can be derived. We show that one of these approaches, which is based on a symmetric formulation of the multinomial logistic regression model, actually reduces to a data shared lasso version of the other. Consequently, the relative performance of the two approaches critically depends on the level of homogeneity that exists among disease subtypes: more precisely, when homogeneity is moderate to high, the non-symmetric formulation with controls as the reference is not recommended. Empirical results obtained from synthetic data are presented, which confirm the benefit of properly accounting for potential homogeneity under both matched and unmatched designs, in terms of estimation and prediction accuracy, variable selection and identification of heterogeneities. We also present preliminary results from the analysis of a case-control study nested within the EPIC cohort, where the objective is to identify metabolites associated with the occurrence of subtypes of breast cancer.

Key words: Conditional logistic regression; Multinomial logistic regression; Lasso; Sparsity; Structured sparsity.

*To whom correspondence should be addressed.

1. Introduction

The rise of -omics and other high-dimensional data in medical science gives researchers access to numerous features that may predict outcomes of interest, like cancer development. However, this relatively cheap source of information comes at a price: the curse of dimensionality makes multivariate modeling of such data impossible without further assumptions. In other words, some prior piece of information has to be properly accounted for to reduce dimensionality and accurately estimate high-dimensional multivariate models. The prior information about the sparsity of the parameter vector is one common assumption for the parametric regression models. The use of L_1-norm regularized approaches, such as the Lasso (Tibshirani, 1996), has been shown to yield optimal sparse estimates when the true vector is sparse, under technical assumptions on the design matrix (Wainwright, 2009; Bach, 2010; Bickel, Ritov and Tsybakov, 2009). As a result, L_1-penalized logistic models are now standard tools when studying risk factors of a disease in a high-dimensional setting (Park and Hastie, 2007; Wu *and others*, 2009) .

For many diseases that were primarily considered as one single disease (breast cancer, colorectal cancer), several subtypes have now been recognized. They can either be histological, as for breast cancer, or anatomical, as for colorectal cancer. Even if commonalities may exist among these subtypes, they have their own specificities regarding both prognosis and etiology. For example, the cancer epidemiology community is now increasingly concerned with the identification of subtype specific risk factors for various cancer sites. One illustrating example is presented in Section 5, where the objective is the identification of metabolites associated with breast cancer subtypes, based on a matched case-control study nested in the EPIC (European Prospective Investigation into Cancer and nutrition) cohort study.

Formally, let $K-1$ denote the number of case/disease subtypes, for some $K > 1$. In matched case-control studies, and assuming for simplicity a 1:1 matching, each case has his own control. Then, the overall sample can naturally be divided into $K-1$ subsamples. Each subsample can be analyzed separately using, e.g., a conditional logistic regression model. On the other hand, for unmatched studies with multiple subtypes, controls are "shared" for all case subtypes, and the sample can not be split according to disease subtype. The analysis of such data typically relies on a multinomial logistic regression model (McCullagh and Nelder, 1989; Begg and Gray, 1984).

Under both matched and unmatched settings, the inference boils down to the estimation of $K-1$ parameter vectors. But, as mentioned above, commonalities are generally expected among disease subtypes. More precisely, some risk factors are likely to be shared by some subtypes, and these shared risk

factors may have the same level of association across various subtypes. Then, the $K-1$ parameter vectors are expected to show some level of homogeneity, in the sense that some zeros are likely to be in the same positions, and that some non-zero values are likely identical across subtypes. Properly accounting for this particular structured sparsity (Bach *and others*, 2012) is key to reduce the complexity of the inference task and improve estimation efficiency (Viallon *and others*, 2016). Recently, data shared lasso has been introduced as a way to account for the expected homogeneity among the $K-1$ parameter vectors to be estimated under stratified regression models (Gross and Tibshirani, 2016; Ollier and Viallon, 2017).

In this article, we will show how the ideas of data shared lasso can be applied to analyze both matched and unmatched case-control studies with multiple disease subtypes. In Section 2, we consider stratified sparse conditional logistic models under matched designs, for which data shared lasso is naturally appealing. Section 3 is devoted to the unmatched setting and sparse multinomial logistic regression models, for which the link with data shared lasso is less obvious at first sight. Two formulations of sparse multinomial logistic regression models exist in the literature (Krishnapuram *and others*, 2005; Friedman, Hastie and Tibshirani, 2010), without clear guidance on how to chose between them. We will formally establish that one of these two formulations corresponds to a data shared lasso version of the other. In Section 4, we present results from a simulation study. Under both the matched and unmatched settings, our results illustrate the superiority of data shared lasso compared to its competitors when homogeneity exists among the parameter vectors to be estimated, in terms of prediction and estimation accuracy, as well as support recovery (i.e., the ability to identify the position of the non-zero entries of these vectors) and identification of heterogeneities among these vectors. Section 5 is devoted to our illustrative example. Concluding remarks are given in Section 6.

2. Matched case-control studies with multiple subtypes of cases and stratified conditional logistic models

Conditional logistic regression is a standard tool for the analysis of matched case-control studies when a single type of disease is considered (Pearce, 2016; Rothman, Greenland and Lash, 2008). Here, we show how the ideas of data shared lasso can be applied to handle the situation where $K-1$ disease subtypes are present, for some given integer $K > 1$.

2.1 *Setting*

Consider a matched case-control study where information about subtype is available for each case. We denote the number of subtypes by $K-1$, for some given integer $K > 1$. For simplicity, we further assume a 1:1 matched case-control design, and we denote by $m \geqslant 1$ the total number of pairs of individuals. Because each case has his own control, the total sample can be divided into $K-1$ subsamples. For any $k \in \{1, .., K-1\}$, the k-th subsample $\mathcal{M}_k$ is made of the m_k pairs composed by each case of subtype k and his matched control.

For any $\ell \in \{1, \ldots, m_k\}$, we let $\mathbf{x}^{(k)}_{\ell,case}$ and $\mathbf{x}^{(k)}_{\ell,control}$ denote the vectors of covariates (of length p) for the case and the control, respectively, in the ℓ-th matched pair of $\mathcal{M}_k$. We then have $Y^{(k)}_{\ell,case} = 1$ and $Y^{(k)}_{\ell,control} = 0$, which represent the disease indicators for the two individuals composing this matched pair. The association between covariates and disease subtype k can be studied by applying a conditional logistic regression model restricted to observations in $\mathcal{M}_k$. Under this model, we assume the existence of a vector $\boldsymbol{\delta}^*_k \in \mathbb{R}^p$ of true values of parameters such that the probability that the case is the one observed in pair ℓ, given that a case is observed in pair ℓ, is (Greenland, 2000)

$$\Pr(Y^{(k)}_{\ell,case} = 1 | Y^{(k)}_{\ell,case} + Y^{(k)}_{\ell,control} = 1, \mathbf{x}^{(k)}_{\ell,case}, \mathbf{x}^{(k)}_{\ell,control}) = \frac{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case})}{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case}) + \exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,control})}. \quad (2.1)$$

Introduce $\mathbf{1}_{m_k} = (1, \ldots, 1)^T \in \mathbb{R}^{m_k}$ and let $\boldsymbol{\Delta}^{(k)}$ denote the $m_k \times p$ matrix whose ℓ-th row equals $(\mathbf{x}^{(k)}_{\ell,control} - \mathbf{x}^{(k)}_{\ell,case})$, for $\ell \in \{1, .., m_k\}$. Vector $\boldsymbol{\delta}^*_k$ can be estimated by maximizing the log conditional likelihood $L^{(cond)}_k$ restricted to pairs in $\mathcal{M}_k$, which is defined for any vector $\boldsymbol{\delta}_k \in \mathbb{R}^p$ as

$$\begin{aligned} L^{(cond)}_k(\boldsymbol{\delta}_k) &= -\sum_{\ell=1}^{m_k} \log[1 + \exp\{\boldsymbol{\delta}^T_k (\mathbf{x}^{(k)}_{\ell,control} - \mathbf{x}^{(k)}_{\ell,case})\}] \\ &= -[\log\{\mathbf{1}_{m_k} + \exp(\boldsymbol{\Delta}^{(k)} \boldsymbol{\delta}_k)\}]^T \mathbf{1}_{m_k}. \end{aligned} \quad (2.2)$$

Equivalently, vectors $\boldsymbol{\delta}^*_1, \ldots, \boldsymbol{\delta}^*_{K-1}$ can be estimated simultaneously by maximizing the following global criterion over $\boldsymbol{\delta} = (\boldsymbol{\delta}^T_1, \ldots, \boldsymbol{\delta}^T_{K-1})^T$,

$$L^{(cond)}(\boldsymbol{\Delta}_{In}, \boldsymbol{\delta}) = \sum_{k=1}^{K-1} L^{(cond)}_k(\boldsymbol{\delta}_k) = -[\log\{\mathbf{1}_m + \exp(\boldsymbol{\Delta}_{In}\boldsymbol{\delta})\}]^T \mathbf{1}_m, \quad (2.3)$$

with

$$\boldsymbol{\Delta}_{In} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} & \ldots & \mathbf{0}_{m_1,p} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{m_{K-1},p} & \ldots & \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}.$$

For future use, observe that function $L^{(cond)}$ is defined for any pair $(\boldsymbol{\Delta}, \boldsymbol{\delta})$ with $\boldsymbol{\Delta} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{\delta} \in \mathbb{R}^d$, for any integer $d \geqslant 1$. Moreover, estimation of the $K-1$ vectors $\boldsymbol{\delta}^*_1, \ldots, \boldsymbol{\delta}^*_{K-1}$, is performed simultaneously

but still independently when maximizing the above criterion. Coupling the estimation of the $K-1$ vectors, that is making the estimation of each vector to depend on each other, is deemed necessary to allow the estimates to share the same similarities as $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$ when such similarities exist. This can be achieved by using appropriate penalties, such as the one employed in data shared lasso presented below.

2.2 *Data shared lasso*

Data shared lasso was introduced by Gross and Tibshirani (2016) and Ollier and Viallon (2017) in the context of stratified regression models, as a way to account for the expected homogeneities among the parameter vectors to be estimated. The key to the approach is a reparametrization of the model. More precisely, instead of the original parametrization based on $\delta_{k,j}^*$, for $k \in \{1, ..., K-1\}$ and $j \in \{1, ..., p\}$, data shared lasso is based on the following over-parametrized decomposition

$$\delta_{k,j}^* = \mu_j^* + \gamma_{k,j}^*. \tag{2.4}$$

Here μ_j^* can be seen as the "global" parameter for covariate j and is common to all subtypes, while $\gamma_{k,j}^*$ captures the variation of the parameter for subtype k around this global parameter. As will be shown in Section 2.3, data shared lasso can be seen as a generalization of several more standard L_1-penalized approaches based on other parametrizations of the model, which correspond to particular constraints in decomposition (2.4).

Even if decomposition (2.4) is over-parametrized, estimates of μ_j^* and $\gamma_{k,j}^*$ for $k \in \{1, ..., K-1\}$ and $j \in \{1, ..., p\}$ can be derived by maximizing the following L_1-penalized criterion over $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and the $\boldsymbol{\gamma}_k$'s, with $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,p})$,

$$\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1). \tag{2.5}$$

As usual, appropriate values of the tuning parameter λ can be obtained in practice by cross-validation (Bühlmann and Geer, 2011) or through the maximization of BIC-like criteria (Schwarz, 1978). We will refer to this approach as CondLogist_DataSharedLasso. The L_1-norm penalty $\|\boldsymbol{\mu}\|_1$ encourages sparsity of the vector of global parameters, while the $\|\boldsymbol{\gamma}_k\|_1$'s encourage homogeneity among vectors $\widehat{\boldsymbol{\delta}}_k$ defined as $\widehat{\boldsymbol{\delta}}_k = \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\gamma}}_k$, for $k \in \{1, ..., K-1\}$. Moreover, Gross and Tibshirani (2016) and Ollier and Viallon (2017) showed that optimal parameters especially satisfy

$$\widehat{\mu}_j = \underset{m}{\text{argmin}}\{|m| + \sum_{k=1}^{K-1} |\widehat{\delta}_{k,j} - m|\} = \text{median}(\widehat{\delta}_{1,j}, \ldots, \widehat{\delta}_{K-1,j}, 0).$$

In words, the estimated global parameter for covariate j corresponds to a shrunk version of the median of the estimated parameters for covariate j across disease subtypes. As a result, estimates $(\widehat{\boldsymbol{\delta}}_1, \ldots, \widehat{\boldsymbol{\delta}}_{K-1})$ produced by CondLogist_DataSharedLasso, are encouraged to be close to their shrunk median $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_p)$ in the L_1-norm sense, hence similar.

We shall stress that the penalty $\sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1$ can be generalized to $\sum_{k=1}^{K-1} \tau_k \|\boldsymbol{\gamma}_k\|_1$, for some $(\tau_k)_{k \geqslant 1}$, e.g., to penalize more heavily terms $\|\boldsymbol{\gamma}_k\|_1$ associated with larger sample size m_k. For simplicity, we focus on the case $\tau_k = 1$ here, and refer to Gross and Tibshirani (2016) and Ollier and Viallon (2017) for more details on the general case.

2.3 *Implementation and relationship with more standard strategies*

A first nice property of data shared lasso is that it can be written as a simple lasso, which makes it readily implementable. In particular, the data shared lasso criterion can be rewritten as

$$\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1) = -[\log\{\mathbf{1}_m + \exp(\boldsymbol{\Delta}_{DS}\boldsymbol{\Gamma})\}]^T \mathbf{1}_m - \lambda\|\boldsymbol{\Gamma}\|_1 \qquad (2.6)$$

with $\boldsymbol{\Gamma} = (\boldsymbol{\mu}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_{K-1}^T)^T \in \mathbb{R}^{K \times p}$ and

$$\boldsymbol{\Delta}_{DS} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} & \boldsymbol{\Delta}^{(1)} & \mathbf{0}_{m_1,p} & \ldots & \mathbf{0}_{m_1,p} \\ \boldsymbol{\Delta}^{(2)} & \mathbf{0}_{m_2,p} & \boldsymbol{\Delta}^{(2)} & \ldots & \mathbf{0}_{m_2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Delta}^{(K-1)} & \mathbf{0}_{m_{K-1},p} & \mathbf{0}_{m_{K-1},p} & \ldots & \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}.$$

Here, $\boldsymbol{\Delta}^{(k)}$ still denotes the $m_k \times p$ matrix whose ℓ-th row equals $(\mathbf{x}_{\ell,control}^{(k)} - \mathbf{x}_{\ell,case}^{(k)})$. Criterion (2.6) corresponds to an L_1-penalized version of the log-likelihood (2.2) or (2.3), with design matrix $\boldsymbol{\Delta}_{DS}$ instead of $\boldsymbol{\Delta}^{(k)}$ or $\boldsymbol{\Delta}_{In}$. In other words, any solver for the L_1-penalized conditional logistic regression model can be used to implement CondLogist_DataSharedLasso. For instance, the cLogitLasso (Avalos *and others*, 2015) and cLogitL1 (Reid and Tibshirani, 2014) packages are available for R users.

In addition, this new writing of the data shared lasso criterion highlights its connection with three more standard approaches based on other reparametrizations of the model, and which correspond to particular constraints in decomposition (2.4). These standard approaches consist in maximizing a criterion similar to (2.6) above with $\boldsymbol{\Delta}_{DS}$ replaced, in turn, by $\boldsymbol{\Delta}_{In}$ given above, and $\boldsymbol{\Delta}_{Po}$ and $\boldsymbol{\Delta}_{Re}^{(1)}$ given by

$$\boldsymbol{\Delta}_{Po} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} \\ \boldsymbol{\Delta}^{(2)} \\ \vdots \\ \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}, \quad \boldsymbol{\Delta}_{Re}^{(1)} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} & \mathbf{0}_{m_1,p} & \ldots & \mathbf{0}_{m_1,p} \\ \boldsymbol{\Delta}^{(2)} & \boldsymbol{\Delta}^{(2)} & \ldots & \mathbf{0}_{m_2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Delta}^{(K-1)} & \mathbf{0}_{m_{K-1},p} & \ldots & \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}.$$

First, consider the constraint $\mu_j^* = 0$ for all $j \in \{1, ..., p\}$ in decomposition (2.4). In this case, the reparametrization is simply a change of notation compared with the original parametrization: $\delta_{k,j}^* = \gamma_{k,j}^*$.

The constraint $\mu_j^* = 0$ for all $j \in \{1, ..., p\}$ can be imposed in criterion (2.6) simply by eliminating the first block of p columns in $\mathbf{\Delta}_{DS}$, that is by replacing $\mathbf{\Delta}_{DS}$ by $\mathbf{\Delta}_{In}$. As detailed in Appendix A.1 of the Supplementary Materials, under this additional contraint, CondLogist_DataSharedLasso reduces to the simple approach which consists in running one L_1-penalized conditional logistic regression (that is an L_1-penalized version of criterion (2.2)) on each subsample $\mathcal{M}_k$ independently, and so we refer to this approach as CondLogist_IndepLasso. Second, consider the constraint $\gamma_{k,j}^* = 0$ for all $k \in \{1, ..., K-1\}$ and $j \in \{1, ..., p\}$. In this case, $\delta_{k,j}^* = \mu_j^*$ for all k: working under this constraint corresponds to assuming that vectors $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_K^*$ are all equal to a common vector, $\boldsymbol{\mu}^*$. This vector $\boldsymbol{\mu}^* \in \mathbb{R}^p$ can again be estimated by maximizing the same criterion as (2.6), this time after eliminating the $K-1$ last blocks of p columns in $\mathbf{\Delta}_{DS}$, that is after replacing $\mathbf{\Delta}_{DS}$ by $\mathbf{\Delta}_{Po}$. This corresponds to pooling all the subsamples together, and we will refer to this approach as CondLogist_PooledLasso. Finally, consider the constraint $\gamma_{1,j}^* = 0$ for all $j \in \{1, ..., p\}$. In this case, we have $\mu_j^* = \delta_{1,j}^*$ and $\gamma_{k,j}^* = \delta_{k,j}^* - \delta_{1,j}^*$ for all $j \in \{1, ..., p\}$ and $k > 1$. The $(K-1) \times p$ parameters $\mu_j^* (= \delta_{1,j}^*)$ and $\gamma_{k,j}^*$ for $j \in \{1, ..., p\}$ and $k \geqslant 2$, can be estimated by maximizing the same criterion as (2.6), after eliminating the second block of p columns in $\mathbf{\Delta}_{DS}$, that is after replacing $\mathbf{\Delta}_{DS}$ by $\mathbf{\Delta}_{Re}^{(1)}$. This corresponds to working under the decomposition $\boldsymbol{\delta}_k^* = \boldsymbol{\delta}_1^* + \boldsymbol{\gamma}_k^*$ for $k \geqslant 2$. In other words, this corresponds to considering the first subtype as the reference subtype, while parameter $\gamma_{k,j}^*$, for $j \in \{1, ..., p\}$ and $k \geqslant 2$, captures how the association of covariate j and subtype k differs from that of covariate j and subtype 1. We will refer to this approach as CondLogist_RefLasso. Of course, any subtype r can be considered as the reference, not necessarily the first one.

Each of these three more standard approaches, CondLogist_IndepLasso, CondLogist_PooledLasso and CondLogist_RefLasso, can therefore be regarded as one particular constrained version of CondLogist_DataSharedLasso, where the additional constraint makes decomposition (2.4) identifiable. However, the flexibility of the over-parametrization on which CondLogist_DataSharedLasso relies makes the approach generally better than the other three, as we now explain. First, the parametrization used in CondLogist_PooledLasso is not flexible enough to account for subtype specificities, and then results in biased estimates unless all vectors $\boldsymbol{\delta}_k^*$ are equal. On the other hand, the parametrizations used in CondLogist_IndepLasso and CondLogist_RefLasso are flexible enough to avoid such a bias. But, as detailed in Ollier and Viallon (2017), these parametrizations are still suboptimal, because they generally involve unnecessarily large numbers of non-zero true parameters. As a matter of fact, the optimal parametrization of the form (2.4) is such that $\|\boldsymbol{\mu}^*\|_0 + \sum_k \|\boldsymbol{\delta}_k^* - \boldsymbol{\mu}^*\|_0$ is minimized, with $\|\cdot\|_0$ standing for the

L_0 pseudo-norm. The optimal choice for μ_j^* is therefore $\delta^*_{r_j,j}$ for any $r_j \in \{1, ..., K-1\}$ such that $\delta^*_{r_j,j}$ is the mode of the collection of values $(\delta^*_{1,j}, \ldots, \delta^*_{K-1,j}, 0)$. In other words, the optimal parametrization of the form (2.4) relies on optimal covariate-specific references. The corresponding optimal version of CondLogist_RefLasso, applied with such optimal covariate-specific references, can of course not be implemented in practice because these optimal covariate-specific references are unknown. But, in the setting of stratified linear models, the data shared lasso strategy was shown to target the same parametrization as this optimal version of CondLogist_RefLasso (Ollier and Viallon, 2017). It was further shown to perform as well as this optimal version of CondLogist_RefLasso, and to outperform the three more standard approaches, both theoretically and empirically (Ollier and Viallon, 2017). Results from our simulation study presented in Section 4 will confirm those described in Ollier and Viallon (2017) under linear regression models. In particular, the strategy based on the data shared lasso penalty usually better accounts for homogeneity than the other three appraoches, which translates into better estimation and prediction accuracy, overall support recovery and identification of heterogeneities.

3. Unmatched case-control studies with multiple subtypes of cases and sparse multinomial logistic models

We now turn our attention to the unmatched setting. When $K-1$ subtypes of cases are present for some given integer $K > 1$, the outcome Y can be modeled as a categorical variable, taking values in $\{1, ..., K\}$. Hereafter, we will assume that $Y = K$ for controls, while $Y = k$ for cases of subtype k, for any $k \in \{1, ..., K-1\}$. When no natural order exists among the categories of Y, the multinomial logistic regression model is a natural extension of the standard logistic regression model. Below, we will recall some basics about the multinomial logistic regression model. We will first introduce the L_1-penalized approach based on the symmetric formulation of the model, as implemented in the popular `glmnet` R package (Friedman, Hastie and Tibshirani, 2010). We will then show that it corresponds to a data shared lasso version of the more standard formulation, which relies on the initial choice of a reference category. For ease of notation, we will mostly focus on models with no intercept. Our presentation would mainly be the same if intercepts were considered, except that intercept terms are generally not penalized, and L_1-norms $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ would be replaced by $\sum_{j=2}^p |\beta_j|$ if β_1 corresponds to the intercept. See the last paragraph in Section 3.1 for additional details.

3.1 *The multinomial logistic regression model*

For any collection of vectors $(\mathbf{u}_1, \ldots, \mathbf{u}_K) \in \mathbb{R}^{p\times K}$, any $k \in \{1, ..., K\}$, and any $\mathbf{x}_0 \in \mathbb{R}^p$ for some $p \geqslant 1$, introduce the function $p_k(\mathbf{x}_0; \mathbf{u}_1, \ldots, \mathbf{u}_K) = \exp(\mathbf{x}_0^T \mathbf{u}_k)/\{\sum_{\ell=1}^K \exp(\mathbf{x}_0^T \mathbf{u}_\ell)\}$. In its symmetric formulation, the multinomial logistic regression model assumes the existence of K vectors $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*) \in \mathbb{R}^{p\times K}$ of true values of parameters such that

$$\Pr(Y = k|\mathbf{x} = \mathbf{x}_0) = \frac{\exp(\mathbf{x}_0^T \boldsymbol{\beta}_k^*)}{\sum_{\ell=1}^K \exp(\mathbf{x}_0^T \boldsymbol{\beta}_\ell^*)} = p_k(\mathbf{x}_0; \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*), \tag{3.7}$$

for any value $\mathbf{x}_0 \in \mathbb{R}^p$ of the covariate vector. Because $\sum_{k=1}^K \Pr(Y = k|\mathbf{x} = \mathbf{x}_0) = 1$ for any $\mathbf{x}_0 \in \mathbb{R}^p$, this formulation is over-parametrized and vectors $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$ in Equation (3.7) are defined up to a constant only. Indeed, for any $\boldsymbol{\nu} \in \mathbb{R}^p$ and any $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p\times K}$, $p_k(\mathbf{x}_0; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = p_k(\mathbf{x}_0; \boldsymbol{\beta}_1 + \boldsymbol{\nu}, \ldots, \boldsymbol{\beta}_K + \boldsymbol{\nu})$. In other words, if model (3.7) holds with vectors $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)$, then it holds with vectors $(\boldsymbol{\beta}_1^* + \boldsymbol{\nu}, \ldots, \boldsymbol{\beta}_K^* + \boldsymbol{\nu})$ for any $\boldsymbol{\nu} \in \mathbb{R}^p$ as well. For future use, note that it especially holds with vectors $(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_K^*, \ldots, \boldsymbol{\beta}_{K-1}^* - \boldsymbol{\beta}_K^*, \mathbf{0}_p)$, which corresponds to the particular choice $\boldsymbol{\nu} = -\boldsymbol{\beta}_K^*$. Because of this lack of identifiability, standard maximum likelihood estimation based on this parametrization can not be used to derive estimates of $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$, and constrained or penalized versions of the likelihood have to be used instead. In particular, the `glmnet` R package (Friedman, Hastie and Tibshirani, 2010) produces estimates defined as maximizers of the L_1-penalized version of the log-likelihood

$$L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) - \lambda \sum_{k=1}^K \|\boldsymbol{\beta}_k\|_1 = \frac{1}{n} \sum_{i=1}^n \log\{p_{y_i}(\mathbf{x}_i; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)\} - \lambda \sum_{k=1}^K \|\boldsymbol{\beta}_k\|_1 \tag{3.8}$$

for an appropriate value of the regularization parameter λ. We will refer to this approach as MultinomLogist_SymLasso. It works under the implicit assumption that (at least) one of the infinitely many collections of vectors $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$ satisfying (3.7) is sparse, and looks for the "sparsest", or more precisely, the one with lowest $\sum_k \|\boldsymbol{\beta}_k^*\|_1$. In particular, Friedman, Hastie and Tibshirani (2010) show that maximizers $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K$ of criterion (3.8) are such that

$$\text{median}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}) = 0, \quad \text{for all } j \in \{1, ..., p\}. \tag{3.9}$$

Equation (3.9) establishes that the L_1-norm penalization solves the lack of identifiability for each covariate by targeting a collection of vectors $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K$ such that, for each covariate, the median of its parameters across the K categories is null. As mentioned above, when intercepts are considered, they are generally not penalized, in which case the lack of identifiability remains for them. In `glmnet`, this is resolved by mean centering, which corresponds to imposing the constraint $\sum_{k=1}^K \hat{\beta}_{k,1} = 0$ (Friedman, Hastie and Tibshirani, 2010), with $\hat{\beta}_{k,1}$ standing for the intercept estimate for the k-th category.

3.2 *Relationship with data shared lasso*

Now, let us turn our attention to the "standard" formulation of the multinomial logistic regression model, which resolves the lack of identifiability of the symmetric one by first selecting a reference category, typically K. Then, this formulation assumes the existence of $K-1$ parameter vectors, say $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$, such that $\Pr(Y = k|\mathbf{x} = \mathbf{x}_0) = p_k(\mathbf{x}_0, \boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*, \mathbf{0}_p)$. The two formulations – symmetric and standard – are strictly equivalent. Indeed, and as mentioned above, for any $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model, vectors $\boldsymbol{\delta}_1^*, \ldots \boldsymbol{\delta}_{K-1}^*$ defined as $\boldsymbol{\delta}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_K^*$ for $k \in \{1, ..., K-1\}$ satisfy the standard one. When the dimension p of the covariates is large, the expected sparsity within vectors $(\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*)$ can be accounted for by looking for estimates maximizing an L_1-penalized log-likelihood (Krishnapuram *and others*, 2005)

$$\frac{1}{n}\sum_{i=1}^{n} \log\{p_{y_i}(\mathbf{x}_i; \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{K-1}, 0)\} - \lambda \sum_{k=1}^{K-1} \|\boldsymbol{\delta}_k\|_1.$$

We will refer to this approach as MultinomLogist_StdLasso. The ideas of data shared lasso can further be applied to account for the homogeneity among vectors $(\boldsymbol{\delta}_k^*)$'s when the subtypes are expected to share commonalities. Considering as in Section 2 the decomposition $\boldsymbol{\delta}_k = \boldsymbol{\mu} + \boldsymbol{\gamma}_k$ for $k \in \{1, ..., K-1\}$, the method we will refer to as MultinomLogist_StdDataSharedLasso then simply consists in maximizing the criterion

$$\frac{1}{n}\sum_{i=1}^{n} \log\{p_{y_i}(\mathbf{x}_i; \boldsymbol{\mu} + \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\mu} + \boldsymbol{\gamma}_{K-1}, \mathbf{0}_p)\} - \lambda\Big(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1\Big).$$

Interestingly, this criterion is exactly the same as the one in Equation (3.8), after using the change of variable $\boldsymbol{\mu} = -\boldsymbol{\beta}_K$ and $\boldsymbol{\gamma}_k = \boldsymbol{\beta}_k$ for all $k < K$; see Appendix A.2 in the Supplementary Materials for the detailed derivation of this result. This equality formally establishes that working under the symmetric formulation (3.7) with an L_1-norm penalty, as in the `glmnet` R package, exactly corresponds to working under the more standard formulation with a data shared lasso penalty to encourage homogeneity among vectors $(\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*)$. More precisely, the estimates $(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$ and $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_{K-1})$ produced by MultinomLogist_SymLasso and MultinomLogist_StdDataSharedLasso, respectively, are such that $\widehat{\boldsymbol{\mu}} = -\widehat{\boldsymbol{\beta}}_K$ and $\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\gamma}}_k$ for all $k \in \{1, ..., K-1\}$.

This equivalence between MultinomLogist_SymLasso and MultinomLogist_StdDataSharedLasso further allows the derivation of guidance on whether to use MultinomLogist_SymLasso or MultinomLogist_StdLasso in practice: by by-passing the arbitrary choice of the reference category, MultinomLogist_SymLasso will typically target a sparser parametrization than MultinomLogist_StdLasso if disease subtypes share commonalities, and is then expected to produce better estimates. MultinomLo-

gist_StdDataSharedLasso can be seen as a way to compensate any suboptimal choice of the reference category in MultinomLogist_StdLasso. Although different at first sight, MultinomLogist_SymLasso and MultinomLogist_StdDataSharedLasso produce the same estimates and we will simply refer to any of them as MultinomLogist_SymLasso in the rest of our article.

4. Simulation study

4.1 *Evaluation criteria*

To compare the performance of the considered approaches (under both the matched and unmatched settings), several criteria are evaluated. Given estimates $\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}$ of true vectors of parameters $\boldsymbol{\delta}^*_1, \ldots, \boldsymbol{\delta}^*_{K-1}$ (under the unmatched setting, they correspond to vectors involved in the standard formulation with controls as the reference category), a first common criterion when evaluating L_1-penalized approaches is the accuracy with respect to support recovery, which measures the ability to correctly identify patterns of null, positive and negative entries in the vector of parameters to be estimated. In our context, we consider the following criterion:

$$\text{Sgn\_Accuracy} = \frac{\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left(\mathbb{1}[\text{sgn}(\delta^*_{k,j}) = \text{sgn}(\hat{\delta}_{k,j})] - \mathbb{1}[\text{sgn}(\delta^*_{k,j}) \times \text{sgn}(\hat{\delta}_{k,j}) = -1]\right)}{(K-1)p},$$

where $\text{sgn}(x) = +1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$ and $\text{sgn}(x) = 0$ if $x = 0$. This criterion is a slight modification of the standard accuracy (Metz, 1978; Viallon *and others*, 2016), where the term $-\mathbb{1}[\text{sgn}(\delta^*_{k,j}) \times \text{sgn}(\hat{\delta}_{k,j}) = -1]$ is included to penalize approaches that tend to produce positive [resp. negative] estimate while the true value is negative [resp. positive], since this is particularly unwanted in practice. Good approaches are expected to have a high Sgn_Accuracy.

In our framework, vectors $\boldsymbol{\delta}^*_1, ..., \boldsymbol{\delta}^*_{K-1}$ are not only expected to be sparse. They may also have some zeros in the same positions, and some non-zero entries may be equal for different subtypes. Estimates $\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}$ should share the same structure to be able to identify heterogeneities. For any $j \in \{1, ..., p\}$, a good approach should then be able to produce estimates $\hat{\delta}_{1,j}, \ldots, \hat{\delta}_{K-1,j}$ such that, for any $(k_1 \neq k_2) \in \{1, ..., K-1\}^2$, $\hat{\delta}_{k_1,j} = \hat{\delta}_{k_2,j}$ if and only if $\delta^*_{k_1,j} = \delta^*_{k_2,j}$. One standard criterion to evaluate this capacity is the Rand Index (Rand, 1971), which is defined in our context as

$$\text{RandIndex} = \frac{\sum_{j=1}^{p}\sum_{k1=1}^{K-2}\sum_{k2>k1}^{K-1}\left(\mathbf{1}[\delta^*_{k1,j} = \delta^*_{k2,j}, \hat{\delta}_{k1,j} = \hat{\delta}_{k2,j}] + \mathbf{1}[\delta^*_{k1,j} \neq \delta^*_{k2,j}, \hat{\delta}_{k1,j} \neq \hat{\delta}_{k2,j}]\right)}{p(K-2)!}.$$

Again, good approaches are expected to have a high RandIndex.

We also evaluate the approaches with respect to estimation error and prediction accuracy. As for the estimation error, we used the following criterion, which should be as low as possible

$$\text{Est\_Error} = \sum_{k=1}^{K-1} \frac{\|\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*\|_2^2}{\|\boldsymbol{\delta}_k^*\|_2^2}.$$

As for the prediction accuracy, we computed an AUC-like criterion, which was adapted to our matched and unmatched settings. Under both settings, our AUC compares predicted probabilities with observed outcomes on an independent test sample of size $n^{(test)} = 10,000$. In the matched setting, our AUC is defined as the weighted average of the AUCs computed in each subsample $\mathcal{M}_k^{(test)}$. In the unmatched setting, we adapted the one class versus all other classes approach (Provost and Domingos, 2000; Fawcett, 2006); see Appendix A.3 in the Supplementary Material for details on this adaptation. In either setting, good approaches are expected to have a high AUC.

4.2 *The matched setting*

We performed a simulation study to assess the performance of data shared lasso in the context of matched case-control studies. We compared CondLogist_DataSharedLasso with CondLogist_IndepLasso, CondLogist_PooledLasso, and CondLogist_RefLasso. For the latter, the first subtype was selected as the reference. We set the number of covariates to $p = 100$, and the number of disease subtypes to $(K-1) = 6$. We further set the number of pairs of observations in each subsample to $m_1 = 200$, $m_2 = 100$ and $m_k = 50$ for $k = 3, \ldots, 6$, so that the total number of observations was $n = 1000$. In this "high"-dimensional setting, we implemented a cross-validation technique in the spirit of the one-step lasso (Bühlmann and Meier, 2008) to select the optimal regularization parameters and obtain the final parameter estimates.

Here, we briefly describe the simulation designs we considered. Additional details are provided in Appendix A.4 of the Supplementary Materials. Four configurations corresponding to four levels of homogeneity among vectors $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_6^*$ were considered: full homogeneity (the 6 vectors are equal), low heterogeneity ($\boldsymbol{\delta}_2^*, \ldots, \boldsymbol{\delta}_6^*$ are equal, and $\boldsymbol{\delta}_1^*$ is different from them), moderate heterogeneity ($\boldsymbol{\delta}_4^*, \boldsymbol{\delta}_5^*$, and $\boldsymbol{\delta}_6^*$ are equal, while the three other vectors are different from them, and from each other) and full heterogeneity (the 6 vectors have nothing in common). We shall stress that under the low and moderate heterogeneity configurations, the first subtype is the worst choice for the reference used in CondLogist_RefLasso, in the sense that $\|\boldsymbol{\delta}_r^*\|_0 + \sum_{k \neq r} \|\boldsymbol{\delta}_k^* - \boldsymbol{\delta}_r^*\|_0$ is maximized for $r = 1$. The comparison between the performance of CondLogist_RefLasso and CondLogist_DataSharedLasso will allow the assessment of the impact of a

suboptimal choice for the reference when applying CondLogist_RefLasso.

To illustrate the relative performance of the approaches as a function of signal strength, we made it vary through a parameter $\delta \in \{0.1, 0.25, 0.5, 0.75\}$, which determines the magnitude of the non-zero true parameters of our generating model, and is related to the log-odds-ratio for an increase of one standard-deviation of the corresponding covariates; see Appendix A.4 of the Supplementary Materials for more details. Under each of the four configurations, and for each of the four signal strengths, we generated 200 samples under model (2.1). Figure 1 presents the criteria averaged over these 200 replicates, along with the 95% confidence intervals, for each of the four methods we compared here, that is CondLogist_DataSharedLasso, CondLogist_IndepLasso, CondLogist_PooledLasso, and CondLogist_RefLasso. Boxplots showing the distribution of the criteria over the 200 replicates for each method, under each configuration and for each signal strength, are presented on Figure 1 of the Supplementary Materials. First consider the case of full homogeneity. Because all the vectors $\boldsymbol{\delta}_k^*$ are equal, the optimal strategy is of course CondLogist_PooledLasso, which is based on a parametrization with p_0 non-zero parameters (where p_0 is the number of non-zero parameters in each vector $\boldsymbol{\delta}_k^*$; we have $p_0 = 10$ in our simulation study). On the other hand, because CondLogist_IndepLasso is based on a parametrization with $(K-1)p_0$ non-zero parameters, it performs poorly compared to CondLogist_PooledLasso in this configuration, in terms of the four criteria we considered: because it is unable to account for homogeneity, estimates produced by CondLogist_IndepLasso are fully heterogeneous (its RandIndex is very low, as expected), hence with a large variance, and performs poorly in terms of estimation and prediction accuracy, and also support recovery (because it is unable to borrow strength from the various subtypes). On the other hand, both CondLogist_DataSharedLasso and CondLogist_RefLasso account for homogeneity, and perform nearly as well as CondLogist_PooledLasso in terms of each of our criteria under this configuration of full homogeneity. It is noteworthy that in this particular case, any subtype is an optimal reference in CondLogist_RefLasso ($\|\boldsymbol{\delta}_r^*\|_0 + \sum_{k \neq r} \|\boldsymbol{\delta}_k^* - \boldsymbol{\delta}_r^*\|_0 = p_0$ for any r), which explains why CondLogist_DataSharedLasso and CondLogist_RefLasso perform similarly in this case. Next, in the case of low heterogeneity, CondLogist_PooledLasso produces biased estimates and is not optimal since vectors $\boldsymbol{\delta}_k^*$ are not all equal anymore. Interestingly, CondLogist_RefLasso does not outperform CondLogist_PooledLasso in this case, and these two approaches actually produce very similar estimates under this configuration. This is due to the particular choice for the reference subtype in CondLogist_RefLasso: when $\boldsymbol{\delta}_2^* = \ldots = \boldsymbol{\delta}_6^*$, and $\boldsymbol{\delta}_1^*$ is different from them, the penalty term $\sum_k \|\boldsymbol{\delta}_k - \boldsymbol{\delta}_1\|_1$ generally prevents the approach to identify these heterogeneities. As a matter of fact, any other choice for the reference would have led to better

performance for CondLogist_RefLasso. As mentioned above, CondLogist_DataSharedLasso bypasses the arbitrary choice of the reference, and mimics the optimal version of CondLogist_RefLasso applied with an optimal, possibly covariate-specific, reference. Under this low heterogeneity configuration, CondLogist_DataSharedLasso allows the identification of heterogeneities (its RandIndex is higher than that of CondLogist_RefLasso and CondLogist_PooledLasso), and substantially outperforms the other approaches with respect to all criteria. As the level of heterogeneity increases, the complexity of the estimation task increases, and the performance of CondLogist_DataSharedLasso tends to that of CondLogist_IndepLasso. But, as long as some level of homogeneity is present (moderate heterogeneity configuration), CondLogist_DataSharedLasso outperforms the other approaches. Under the full heterogeneity configuration, CondLogist_DataSharedLasso still performs on average as well as CondLogist_IndepLasso, which is the optimal strategy in this case, while CondLogist_PooledLasso, and to a lesser extent CondLogist_RefLasso, perform worse.

Overall, our results illustrate that the performance of CondLogist_IndepLasso does not depend on the level of heterogeneity, in terms of support recovery, prediction accuracy and estimation accuracy. In the total absence of homogeneity, this performance is optimal. But, as the level of homogeneity increases, methods that account for homogeneity can target better (*i.e.*, sparser) parametrizations, and yield substantial improvements in terms of estimation performance. Among the four approaches we compared here, CondLogist_DataSharedLasso appears as the best approach to account for homogeneity when it is present. In addition, it performs as well as CondLogist_IndepLasso on average when no homogeneity is present at all.

4.3 *The unmatched setting*

We further performed a simulation study in the unmatched setting to illustrate the relative performance of MultinomLogist_StdLasso and MultinomLogist_SymLasso (the later being the same as MultinomLogist_StdDataSharedLasso), depending on the level of homogeneity among vectors $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$ of the standard formulation. We again set $K-1=6$ disease subtypes, and considered four configurations: full homogeneity, low heterogeneity, moderate heterogeneity and full heterogeneity. To save computational time, a low-dimensional setting with $n=1000$ and $p=20$ was considered here. To generate the data, we adapted the framework described in Section 4.2 to the unmatched setting. We used intercept terms, $(\delta_{1,0}, \ldots, \delta_{K-1,0})$, chosen in such a way that $\Pr(Y=K)=0.5$ and $\Pr(Y=k)$ ranged from 0.05 to 0.2 for

$k \in \{1, ..., K-1\}$. In this low-dimensional setting, regularization parameters were selected as minimizers of the BIC after adapting the Lasso-OLS hybrid ideas to our context (Efron *and others*, 2004), in the same way as in Viallon *and others* (2016).

Figure 2 presents the criteria averaged over 200 replicates, along with their 95% confidence intervals. Boxplots summarizing the full distribution of the criteria over the 200 replicates are presented in Figure 2 of the the Supplementary Materials. Overall, the conclusions drawn from the comparison between MultinomLogist_SymLasso and MultinomLogist_StdLasso in this unmatched setting are consistent with those drawn when comparing CondLogist_DataSharedLasso with CondLogist_IndepLasso in the matched setting. More precisely, the two methods perform similarly in case of full heterogeneity, but the performance of MultinomLogist_SymLasso improves as the level of homogeneity increases, while that of MultinomLogist_StdLasso remains roughly unchanged. In particular, MultinomLogist_SymLasso substantially outperforms MultinomLogist_StdLasso with respect to all criteria in the case of full homogeneity. This was expected since the number of non-zero parameters to be estimated under the standard formulation is $(K-1)p_0$ (where p_0 is the number of non-zero parameters in each $\boldsymbol{\delta}_k^*$; this was set to $p_0 = 10$ in our simulation study), while MultinomLogist_SymLasso (or equivalently MultinomLogist_StdDataSharedLasso) is able to target a parametrization with only p_0 non-zero parameters in the case of full homogeneity; see Appendix A.6 of the Supplementary Materials, for more details. Just as in the matched setting, our results confirm that using data shared lasso (or, equivalently, the symmetric formulation in this unmatched setting) allows the homogeneity to be accounted for when present, which translates into better estimation and prediction accuracy, support recovery and identification of heterogeneties.

5. Application

5.1 *Data description*

The European Prospective Investigation into Cancer and Nutrition (EPIC) study is an ongoing multicenter prospective study aiming to investigate prospectively the etiology of cancer in relation to diet, lifestyle and environmental factors. Its design has been previously described in detail (Riboli *and others*, 2002). From 1992 to 2000, a total of 521,324 participants were recruited across 10 European countries. Among these participants, 246,000 women, aged from 35 to 70 years, provided a blood sample at inclusion. Here, we present preliminary results from the analysis of a case-control study nested in EPIC, whose main objective was to assess the association between metabolites and the risk of subtypes of breast

cancer for women older than 50: 1415 cases of breast cancer were included, along with 1415 matched controls (using incidence density sampling). We shall stress that the methods presented in Section 2 can be applied when the case-control study is nested within a cohort, as is the case here. This is because the analysis of the k-th disease subtype still relies on a conditional logistic regression model with parameter $\boldsymbol{\delta}_k^*$, which measures the level of association between the covariates and disease subtype k.

For these 2830 individuals, plasma samples collected at inclusion in the study were analyzed by mass spectrometry (AbsoluteIDQ p180 Kit) allowing the measurement of the concentrations of 127 metabolites. These concentrations were log-transformed to reduce skewness. We considered six histological subtypes of breast cancer, based on the presence/absence of hormone receptors: HER2-enriched (100 pairs of observations), triple negative (134 pairs), Luminal A PR+ (164 pairs), Luminal A PR- (820 pairs), Luminal B PR+ (58 pairs) and Luminal B PR- (139 pairs).

5.2 *Results*

Figure 3 provides a graphical representation of the log odds-ratio estimates $\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_6$ produced by each of the four methods for the 6 subtypes of breast cancer. For 79 out of the 127 measured metabolites, all methods produced a zero estimate for all subtypes. These "constantly" null estimates are not reported on Figure 3 to improve legibility. Also, the remaining 48 metabolites were anonymized as the biological interpretation of the results is out of the scope of this preliminary analysis. When analyzing such data, most practitioners would start by pooling all subtypes together (that is, ignoring subtypes) to identify metabolites associated with breast cancer as a whole. In this application, CondLogist_PooledLasso does identify several metabolites associated with breast cancer, which naturally raises the question of whether these identified metabolites (and maybe other ones as well) may be more specifically associated with particular subtypes. The independent analyses of each subtype, as implemented in CondLogist_IndepLasso, identifies many metabolites associated with the Luminal A PR- subtype, and fewer metabolites for the other subtypes. In particular, no metabolite is identified for the Luminal B PR- and the HER2-enriched subtypes. Moreover, very few metabolites were found to be associated with more than one subtype: to name a few exceptions, M96 appeared to be associated with both Luminal A PR+ and Luminal A PR-, and M28 with Luminal A PR+ and Triple Negative. Clearly, this heterogeneity across the subtypes can be the result of a combination of: (i) true heterogeneities, (ii) lack of power for some subtypes (many metabolites are identified in the case of Luminal A PR-, which is the most frequent subtype, while no

metabolite is identified for Luminal B PR- or the HER2-enriched which are the two least frequent subtypes), and (*iii*) sample variability combined with correlations among the metabolites. Indeed, if two metabolites are strongly correlated, CondLogist_IndepLasso will typically identify one or the other on two different samples even if these samples are drawn from the same population (that is, in the absence of true heterogeneity between the two samples). In other words, and just as in subgroup analyses (Wang *and others*, 2007), it is hazardous to claim and interpret heterogeneities on the basis of the independent analyses of subtypes. Because heterogeneities are penalized when using CondLogist_DataSharedLasso (and, in a less optimal way, when using CondLogist_RefLasso), heterogeneities identified by CondLogist_DataSharedLasso are supported by the data, and are more likely true ones. In the present application, CondLogist_DataSharedLasso produces estimates that are quite similar to those produced by CondLogist_PooledLasso, suggesting that the data does not support departure from homogeneity in the levels of association between most metabolites and breast cancer across subtypes. A few heterogeneities are identified though, suggesting that some metabolites might be more specifically associated with the Luminal A PR+ subtype (M18, M27, M42, M43, M63 but also M111 whose association with other subtypes exist, but is stronger with Luminal A PR+), or Luminal A PR- (M96). The comparison with the results produced by CondLogist_RefLasso is also instructive, in particular the estimates produced for M18 and M63. Because Luminal A PR- was chosen as the reference when applying CondLogist_RefLasso, it is here unable to identify any heterogeneity for this particular subtype, which is consistent with the results of our simulation study under the low heterogeneity configuration.

6. Discussion

In this article, we considered the analysis of high-dimensional case-control studies, when several disease subtypes exist, under both unmatched and matched settings. In the latter case, our analysis further covers matched case-control studies nested within a cohort. We have shown that estimation and prediction accuracy, support recovery and the ability to identify heterogeneities across subtypes, could all be substantially improved when commonalities exist among subtypes, provided methods that properly account for these commonalities, e.g. those based on the data shared lasso penalty, are used. Our findings are in line with the empirical and theoretical results of Ollier and Viallon (2017) in the case of stratified linear regression models, as well as the empirical results of Ballout and Viallon (2017) for stratified binary graphical models.

Under matched designs, the original parametrization relies on $K-1$ vectors $\boldsymbol{\delta}_k^*$, which represent the log odds-ratios that compare each of the $K-1$ disease subtypes with controls. Based on an over-parametrized reparametrization, CondLogist_DataSharedLasso is able to target a sparser parametrization when commonalities exist among subtypes, which can yield substantial improvements in terms of estimation efficiency. In the absence of commonalities, it still performs as well as the standard, independent analysis of each subtype. Under unmatched designs, the standard formulation of multinomial logistic regression models relies on the same parametrization, involving $K-1$ vectors $\boldsymbol{\delta}_k^*$ that compare each disease subtype with controls. We formally established that applying the ideas of data shared lasso along with this parametrization was actually equivalent to applying a standard lasso on the symmetric formulation of the model. This symmetric formulation relies on an over-parametrized parametrization with K vectors $\boldsymbol{\beta}_k^*$, and takes advantage of the fact that controls do not necessarily have to be considered as the reference category in unmatched settings. Again, the resulting parametrization can be much sparser than the standard one, and yields generally better estimation efficiency, especially when the level of homogeneity among subtypes is high.

The methods we presented to account for potential commonalities are simple to implement under both designs. Under matched designs, CondLogist_DataSharedLasso is as easy to implement as CondLogist_RefLasso or CondLogist_IndepLasso. Under unmatched designs, MultinomLogist_SymLasso (which is equivalent to MultinomLogist_StdDataSharedLasso) is implemented in the `glmnet` R package. Given the simplicity of their implementation and the possibly substantial gain in terms of estimation performance, we strongly encourage the use of these approaches when analyzing case-control studies with several disease subtypes.

As pointed out in our application to the EPIC data, the methods that account for potential commonalities are especially useful to claim and interpret heterogeneities across subtypes, contrary to methods that do not account for them. An interesting extension would concern the derivation of valid p-values or confidence intervals for the nonzero parameters identified by CondLogist_DataSharedLasso or MultinomLogist_SymLasso, in particular those corresponding to heterogeneities across subtypes. Given the connection of data shared lasso with the lasso (see, e.g., Equation (2.6) under matched designs), this post-selection inference could be derived by extending strategies proposed for lasso estimates (Lee *and others*, 2016). In other respects, when the identification of heterogeneities is of primary interest, study design is an important step to ensure balanced sample sizes across subtypes (which was not the case in our application to the EPIC data).

The estimation of several parameter vectors considered here is closely related to multi-task learning (Evgeniou and Pontil, 2004), for which a number of other structured sparsity inducing norms have been proposed in the literature, including the group lasso and generalized fused lasso (Lounici *and others*, 2011; Viallon *and others*, 2016). We shall first mention that the group lasso is not well suited for the identification of heterogeneities. On the other hand, the generalized fused lasso has shown good properties in the context of stratified regression models, both under generalized linear models (Viallon *and others*, 2016), survival models (Sennhenn-Reulen and Kneib, 2016) and binary graphical models (Ballout and Viallon, 2017). Its extension to conditional logistic regression models or multinomial logistic models constitutes another interesting lead for future work.

7. Software

Software in the form of R codes is available on `Github`. The link to the codes in the matched setting is `https://github.com/NadimBLT/SL1CLR`. The link to the codes in the unmatched setting is `https://github.com/NadimBLT/L1MLR`.

8. Supplementary Materials

Supplementary materials is available online at `http://biostatistics.oxfordjournals.org`.

Acknowledgments

This work was partially supported by the French National Cancer Institute (L'Institut National du Cancer; INCA) (grant number 2015-166; PI: S. Rinaldi). The authors are grateful to the Principal Investigators of each of the EPIC centres for sharing the data of our illustrative example.

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

References

Avalos, M., Pouyes, H., Grandvalet, Y., Orriols, L. and Lagarde, E. (2015). Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple

algorithm. *BMC bioinformatics* **16**(6), S1.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics* **4**, 384–414.

Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statistical Science* **27**(4), 450–468.

Ballout, N. and Viallon, V. (2017). Structure estimation of binary graphical models on stratified data: application to the description of injury tables for victims of road accidents. *arXiv preprint arXiv:1709.10298*.

Begg, C. B. and Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* **71**(1), 11–18.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 1705–1732.

Bühlmann, P. and Geer, S. (2011). *Statistics for High-Dimensional Data: Method, Theory and Applications*. Springer Series in Statistics.

Bühlmann, P. and Meier, L. (2008). Discussion of "one-step sparse estimates in nonconcave penalized likelihood models" by h. zou and r. li. *Ann. Statist* **36**, 1534–1541.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics* **32**, 407–499.

Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. pp. 109–117.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters* **27**(8), 861 – 874. ROC Analysis in Pattern Recognition.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Soft.* **33**(1), 1–22.

Greenland, S. (2000). Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics* **1**(1), 113–122.

Gross, S. M. and Tibshirani, R. (2016). Data shared lasso: A novel tool to discover uplift. *Computational statistics & data analysis* **101**, 226–235.

Krishnapuram, B., Carin, L., Figueiredo, M. A. and Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence* **27**(6), 957–968.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference with the lasso. *Ann. Statist.* **44**, 907–927.

Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 2164–2204.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, Volume 37. CRC press.

Metz, C. E. (1978). Basic principles of roc analysis. *Seminars in Nuclear Medicine* **8**(4), 283 – 298.

Ollier, E. and Viallon, V. (2017). Regression modeling on stratified data with the lasso. *Biometrika* **104**(1), 84–96.

Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69**(4), 659–677.

Pearce, N. (2016). Analysis of matched case-control studies. *BMJ* **352**, i969.

Provost, F. and Domingos, P. (2000). Well-trained pets: Improving probability estimation trees. CeDER Working Paper #IS-00-04, Stern School of Business, New York University, NY 10012.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**(336), 846–850.

Reid, S. and Tibshirani, R. (2014). Regularization paths for conditional logistic regression: the clogitl1 package. *Journal of statistical software* **58**(12).

Riboli, E., Hunt, K., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondiere, U., Hemon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A. *and others*. (2002). European prospective investigation into cancer and nutrition (epic): study populations and data collection. *Public health nutrition* **5**(6b), 1113–1124.

ROTHMAN, K. J., GREENLAND, S. AND LASH, T. L. (2008). *Modern epidemiology, 3rd edition*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The annals of statistics* **6**(2), 461–464.

SENNHENN-REULEN, H. AND KNEIB, T. (2016). Structured fusion lasso penalized multi-state models. *Statistics in medicine* **35**(25), 4637–4659.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

VIALLON, V., LAMBERT-LACROIX, S., HOEFLING, H. AND PICARD, F. (2016). On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing* **26**(1), 285–301.

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on* **55**(5), 2183–2202.

WANG, R., LAGAKOS, S. W., WARE, J. H., HUNTER, D. J. AND DRAZEN, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357**(21), 2189–2194.

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. AND LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**(6), 714–721.

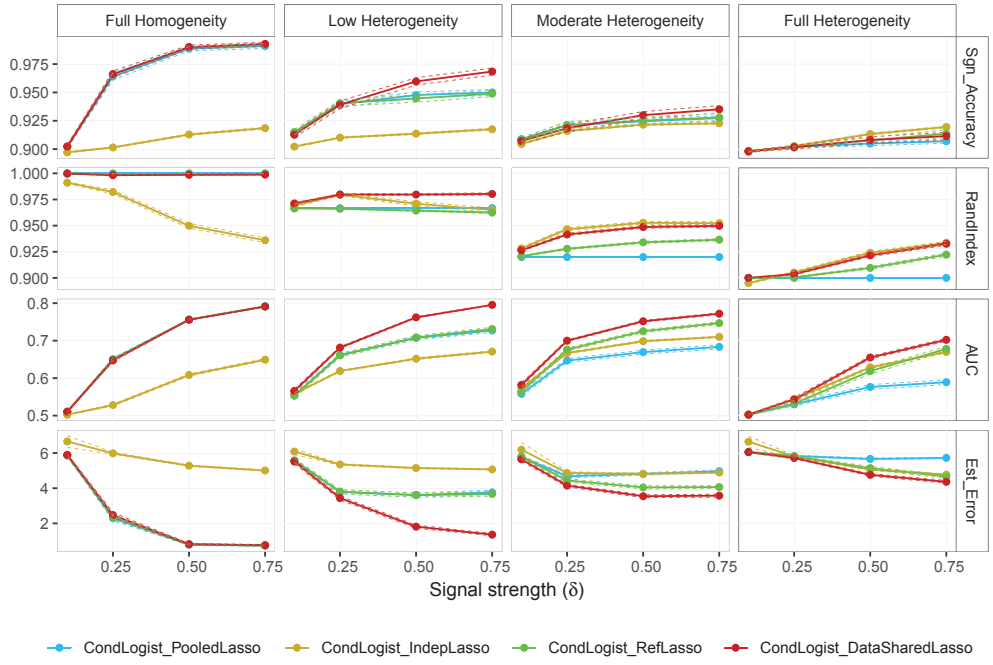

Fig. 1. Results of the simulation study in the matched setting. Solid lines correspond to averages over the 200 replicates, while 95% confidence intervals appear as dotted lines.

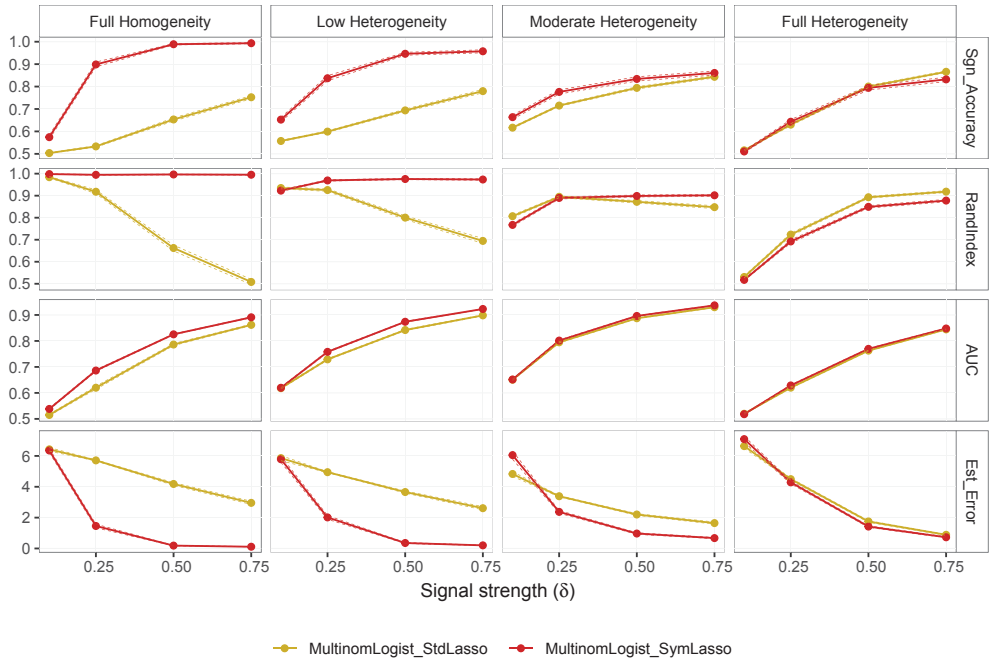

Fig. 2. Results of the simulation study in the unmatched setting. Solid lines correspond to averages over the 200 replicates, while 95% confidence intervals appear as dotted lines

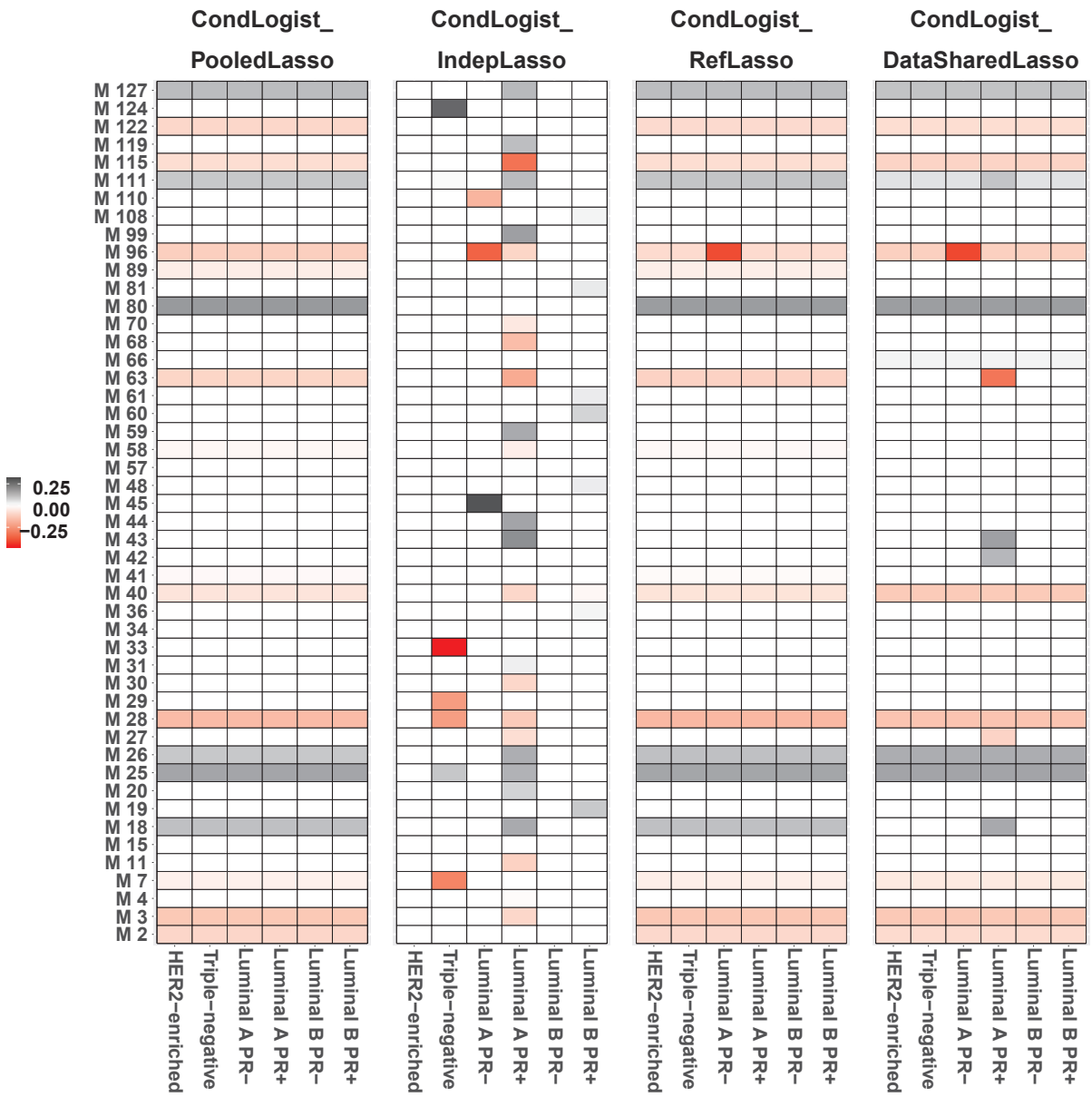

Fig. 3. Preliminary results from the analysis of the matched case-control study nested in EPIC. Six breast cancer histological subtypes are considered: HER-enriched, Triple Negative, Luminal A PR-, Luminal A PR+, Luminal B PR- and Luminal B PR+. Results obtained after the application of four different methods (CondLogist_PooledLasso, CondLogist_IndepLasso, CondLogist_RefLasso, and CondLogist_DataSharedLasso) are presented. For CondLogist_RefLasso, the Luminal A PR+ subtype was selected as the reference. For each method, estimates of $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_6^*$ are combined in a matrix, with 6 columns (one for each subtype) and 48 rows (out of the 127 original metabolites, the 79 metabolites for which the four methods produced a zero estimate for all 6 subtypes were eliminated from the plot). In each of the four matrices, each entry represents the estimated level of association between one metabolite and one particular breast cancer subtype. White entries correspond to null associations, grey entries indicate positive associations, while red entries indicate negative association; see the scale on the left of the figure. For example, CondLogist_IndepLasso identifies a strongly inverse association between metabolite M33 and Triple-Negative breast cancer.

Biostatistics (2019), **0**, 0, *pp.* 1–9
doi:10.1093/biostatistics/Ballout-Garcia-Viallon-SupplementaryMaterials˙Rev

Sparse estimation for case-control studies with multiple disease subtypes Supplementary Materials

Nadim Ballout*,[1], Cedric Garcia[2],Vivian Viallon[3]

[1]*IFSTTAR, UMRESTTE, Université Claude Bernard Lyon 1, Lyon, FRANCE.*

[2]*IFSTTAR, AME, DEST, Marne la vallée, FRANCE.*

[3]*IARC, WHO, Lyon, FRANCE.*

nadim.ballout@ifsttar.fr

APPENDIX

A. Additional technical details

A.1 *Details on the "standard" L_1-penalized approaches presented in the matched design*

Here, we provide additional details on the link between CondLogist_DataSharedLasso and the three more standard approaches presented in the matched design (CondLogist_IndepLasso, CondLogist_PooledLasso and CondLogist_RefLasso).

CondLogist_DataSharedLasso. First recall that estimates produced by CondLogist_DataSharedLasso are defined as $\hat{\boldsymbol{\delta}}_k = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\gamma}}_k$ with

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \underset{(\boldsymbol{\mu}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{K-1}) \in \mathbb{R}^{p \times K}}{\operatorname{argmin}} \sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1)$$

CondLogist_IndepLasso. This approach simply consists in working with the original parametrization and performing one L_1-penalized conditional logistic regression on each subsample independently. Then,

*To whom correspondence should be addressed.

the most natural way of defining estimates returned by CondLogist_IndepLasso is

$$(\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}) = \underset{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_{K-1})\in\mathbb{R}^{p\times(K-1)}}{\operatorname{argmin}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\delta}_k) - \lambda\|\boldsymbol{\delta}_k\|_1 \big\} \Big]$$

But, CondLogist_IndepLasso can also be seen as a special case of CondLogist_DataSharedLasso since we also have $(\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}) = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1})$, with

$$(\mathbf{0}_p, \hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \underset{\substack{(\boldsymbol{\mu},\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{K-1})\in\mathbb{R}^{p\times K} \\ \boldsymbol{\mu}=\mathbf{0}_p}}{\operatorname{argmin}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\mu}+\boldsymbol{\gamma}_k) \big\} - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

$$i.e., \quad (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \underset{(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{K-1})\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\gamma}_k) - \lambda\|\boldsymbol{\gamma}_k\|_1 \big\} \Big]$$

While estimated simultaneously, the $(K-1)$ parameter vectors are still estimated independently. This approach cannot take advantage of any potential similarity among $(\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_K^*)$, and typically produces estimates with suboptimal properties when such similarity exists.

CondLogist_PooledLasso. This approach works under the (strong) assumption that all disease subtypes share the same parameter vector: $\boldsymbol{\delta}_1^* = \ldots = \boldsymbol{\delta}_{K-1}^* = \boldsymbol{\delta}^*$. Then, the most natural way of defining estimates returned by CondLogist_PooledLasso is

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}\in\mathbb{R}^p}{\operatorname{argmin}} \Big[\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\delta}) - \lambda\|\boldsymbol{\delta}\|_1 \Big]$$

Again, there is a link between CondLogist_PooledLasso and CondLogist_DataSharedLasso since $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}$, with

$$(\hat{\boldsymbol{\mu}}, \mathbf{0}_p, \ldots, \mathbf{0}_p) = \underset{\substack{(\boldsymbol{\mu},\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{K-1})\in\mathbb{R}^{p\times K} \\ \boldsymbol{\gamma}_1=\ldots=\boldsymbol{\gamma}_{K-1}=\mathbf{0}_p}}{\operatorname{argmin}} \Big[\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu}+\boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

$$i.e., \quad \hat{\boldsymbol{\mu}} = \underset{\hat{\boldsymbol{\mu}}\in\mathbb{R}^p}{\operatorname{argmin}} \Big[\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu}) - \lambda\|\boldsymbol{\mu}\|_1 \Big]$$

Because all $\boldsymbol{\delta}_k^*$'s are assumed to be equal, this approach obviously produces biased estimates when differences exist among the $\boldsymbol{\delta}_k^*$'s.

CondLogist_RefLasso. For simplicity, assume that the first disease subtype is chosen as the reference. Then, CondLogist_RefLasso works under the following reparametrization: $\boldsymbol{\delta}_k^* = \boldsymbol{\delta}_1^* + \boldsymbol{\gamma}_k^*$ for all $k \geqslant 2$. The most natural way of defining estimates returned by CondLogist_RefLasso is

$$(\hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \underset{(\boldsymbol{\delta}_1,\boldsymbol{\gamma}_2\ldots,\boldsymbol{\gamma}_{K-1})\in\mathbb{R}^{p\times(K-1)}}{\operatorname{argmin}} \Big[L_1^{(cond)}(\boldsymbol{\delta}_1) + \sum_{k=2}^{K-1} L_k^{(cond)}(\boldsymbol{\delta}_1+\boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\delta}_1\|_1 + \sum_{k=2}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

But, we also have $(\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\gamma}}_{K-1})$, with

$$(\hat{\boldsymbol{\mu}}, \mathbf{0}_p, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \underset{\substack{(\boldsymbol{\mu}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{K-1}) \in \mathbb{R}^{p \times K} \\ \boldsymbol{\gamma}_1 = \mathbf{0}_p}}{\operatorname{argmin}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) \big\} - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

$$i.e., \quad (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \underset{(\boldsymbol{\mu}, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_{K-1}) \in \mathbb{R}^{p \times (K-1)}}{\operatorname{argmin}} \Big[L_1^{(cond)}(\boldsymbol{\mu}) + \sum_{k=2}^{K-1} L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=2}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

More generally, when used with the r-th subtype as the reference, this approach encourages similarity between $\boldsymbol{\delta}_r^*$ and the remaining $K-2$ vectors. Consequently, its performance depends on the choice of the reference subtype. In particular, it will perform poorly if the true parameter vector pertaining to the chosen reference subtype is actually very different from the $K-2$ other ones. As explained in the main text, the optimal reference subtype is generally covariate-specific. CondLogist_DataSharedLasso bypasses the arbitrary choice of the reference subtype. Moreover, in the setting of stratified linear regression models, the data shared lasso strategy was shown to perform as well as the optimal (and non-implementable) strategy based on an a priori selection of optimal covariate-specific references.

A.2 *Equivalence between MultinomLogist_SymLasso and MultinomLogist_StdDataSharedLasso*

First observe that the contribution of an individual with covariate vector $\mathbf{x}_0$ to the likelihood of the symmetric formulation of the model (see Equation (3.8) of the Main Manuscript) is

$$\begin{aligned}
\prod_{k=1}^{K} \{p_k(\mathbf{x}_0; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)\}^{\mathbf{I}(Y=k)} &= \prod_{k=1}^{K} \left\{ \frac{\exp(\mathbf{x}_0^T \boldsymbol{\beta}_k)}{\sum_{\ell=1}^{K} \exp(\mathbf{x}_0^T \boldsymbol{\beta}_\ell)} \right\}^{\mathbf{I}(Y=k)} \\
&\overset{(*)}{=} \prod_{k=1}^{K-1} \left\{ \frac{\exp(\mathbf{x}_0^T \boldsymbol{\gamma}_k)}{\exp(-\boldsymbol{\mu}^T \mathbf{x}_0) + \sum_{\ell=1}^{K-1} \exp(\mathbf{x}_0^T \boldsymbol{\gamma}_\ell)} \right\}^{\mathbf{I}(Y=k)} \times \left\{ \frac{\exp(-\boldsymbol{\mu}^T \mathbf{x}_0)}{\exp(-\boldsymbol{\mu}^T \mathbf{x}_0) + \sum_{\ell=1}^{K-1} \exp(\mathbf{x}_0^T \boldsymbol{\gamma}_\ell)} \right\}^{\mathbf{I}(Y=K)} \\
&= \prod_{k=1}^{K-1} \left\{ \frac{\exp(\mathbf{x}_0^T (\boldsymbol{\mu} + \boldsymbol{\gamma}_k))}{1 + \sum_{\ell=1}^{K-1} \exp(\mathbf{x}_0^T (\boldsymbol{\mu} + \boldsymbol{\gamma}_\ell))} \right\}^{\mathbf{I}(Y=k)} \times \left\{ \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\mathbf{x}_0^T (\boldsymbol{\mu} + \boldsymbol{\gamma}_\ell))} \right\}^{\mathbf{I}(Y=K)} \\
&= \prod_{k=1}^{K} \{p_k(\mathbf{x}_0; \boldsymbol{\mu} + \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\mu} + \boldsymbol{\gamma}_{K-1}, \mathbf{0}_p)\}^{\mathbf{I}(Y=k)}
\end{aligned}$$

where we used the change of variable $\boldsymbol{\mu} = -\boldsymbol{\beta}_K$ and $\boldsymbol{\gamma}_k = \boldsymbol{\beta}_k$ for all $k < K$ to obtain the equality $(*)$. Using the same change of variable, we get $\sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_1 = \|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1$. Putting this all together, the L_1-penalized criterion (3.9) of the Main Manuscript equals, up to a change of variable,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) - \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_1 &= \frac{1}{n} \sum_{i=1}^{n} \log\{p_{y_i}(\mathbf{x}_i; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)\} - \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_1 \\
&= \frac{1}{n} \sum_{i=1}^{n} \log\{p_{y_i}(\mathbf{x}_i; \boldsymbol{\mu} + \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\mu} + \boldsymbol{\gamma}_{K-1}, \mathbf{0}_p)\} - \lambda\Big(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1\Big).
\end{aligned}$$

A.3 *Additional details on the AUC criteria*

In the unmatched setting, the AUC was computed as an adaptation of the one class versus all other classes approach (Provost and Domingos, 2000; Fawcett, 2006). More precisely, first remind that we generate data such that, e.g., $\boldsymbol{\delta}_4^* = \boldsymbol{\delta}_5^* = \boldsymbol{\delta}_6^*$ (under the configurations described as full homogeneity, low heterogeneity and moderate heterogeneity). Then, the three classes 4, 5, 6 are undistinguishable under these configurations. More generally, the set of classes $\{1, ..., K\}$ can be partitioned into $G = \{g_1, ..., g_I\}$, where $I \leqslant K$ and $g_i \subset \{1, ..., K\}$, for all $i = \{1, \ldots, I\}$, such that $(\exists i \in \{1, \ldots, I\} : (k_1, k_2) \in g_i) \Leftrightarrow (\boldsymbol{\delta}_{k_1}^* = \boldsymbol{\delta}_{k_2}^*)$; we set $\boldsymbol{\delta}_K^* = \mathbf{0}_p$ for the class corresponding to controls. For the sake of completeness, these partitions are as follows under the four configurations we considered

- Full homogeneity: $G = \{g_1, g_2\}$, $|G| = 2$, $g_1 = \{1, ..., 6\}$ and $g_2 = \{7\}$
- Low heterogeneity: $G = \{g_1, g_2, g_3\}$, $|G| = 3$, $g_1 = \{1\}$, $g_2 = \{2, ..., 6\}$ and $g_2 = \{7\}$
- Moderate heterogeneity: $G = \{g_1, g_2, g_3, g_4, g_5\}$, $|G| = 5$, $g_1 = \{1\}$, $g_2 = \{2\}$, $g_3 = \{3\}$, $g_4 = \{4, ..., 6\}$ and $g_5 = \{7\}$
- Full heterogeneity: $G = \{g_1, g_2, ..., g_7\}$, $|G| = 7$, for each $k \in \{1, ..., k\}$, $g_k = k$

Given such a partition G of $\{1, \ldots, K\}$, for any $g \subset G$, let $y_i^{(g)} = 1$ if $y_i \in g$ and 0 otherwise. Then set $n_g = \sum_{i=1}^n \mathbb{1}[y_i \in g]$, $T_0^{(g)} = \{i \in \{1, ..., n\} : y_i^{(g)} = 0\}$, $T_1^{(g)} = \{i \in \{1, ..., n\} : y_i^{(g)} = 1\}$, and

$$\hat{p}_i^{(g)} = \sum_{k \in g} \hat{p}_i^{(k)} = \sum_{k \in g} \frac{\exp\left(x_i^T \hat{\boldsymbol{\delta}}_k\right)}{\sum_{\ell=1}^K \exp\left(x_i^T \hat{\boldsymbol{\delta}}_\ell\right)}.$$

The AUC was then simply computed as

$$AUC = \sum_{g \subset G} \frac{n_g}{n} \frac{1}{|T_1^{(g)}| \times |T_0^{(g)}|} \sum_{i_1 \in T_1^{(g)}} \sum_{i_2 \in T_0^{(g)}} \mathbb{1}[\hat{p}_{i_1}^{(g)} > \hat{p}_{i_2}^{(g)}].$$

It is a weighted average of the one class versus all class AUC (Provost and Domingos, 2000; Fawcett, 2006), with classes replaced by the groups of classes in G.

A.4 *Additional details on the simulation study*

For any real numbers $a < b$, we denote the uniform distribution on $[a, b]$ by $\mathcal{U}_{[a,b]}$. For any $p \in [0, 1]$, we further denote the Bernoulli distribution with parameter p by $B(p)$. Parameters $\delta_{k,j}^*$ were generated as follows. One subset $J_1 \subset \{1, ..., p\}$ was first randomly selected, with $|J_1| = 10$. For $j \notin J_1$, we set $\delta_{k,j}^* = 0$ for all $k \in \{1, ..., K-1\}$. For $j \in J_1$, four configurations were considered, allowing the level

of homogeneity among $(\delta^*_{1,j}, \ldots, \delta^*_{K-1,j})$ to vary. In the first configuration (full homogeneity), we set $\delta^*_{k,j} = (2\iota_j - 1)\delta$, for some $\delta > 0$ and with $\iota_j \sim B(1/2)$. In the second configuration (low heterogeneity), for $j \in J_1$, we set $\delta^*_{k,j} = (2\iota_j - 1)\delta$ for $k \geqslant 2$ and $\delta^*_{1,j} = (2\iota_{k_j,j} - 1)\delta(1 + U_{k_j,j})$, with each $\iota_{k,j} \sim B(1/2)$ and $U_{k_j,j} \sim \mathcal{U}_{[\sqrt{K}/2, 2\sqrt{K}]}$. Here, the limits $[\sqrt{K}/2, 2\sqrt{K}]$ were motivated by the non-asymptotic results obtained by Ollier and Viallon (2017) under stratified linear regression models (see the comment right after Theorem 1). In the third configuration (moderate heterogeneity), we set $\delta^*_{k,j} = (2\iota_j - 1)\delta$ for $k \notin \{1, 2, 3\}$ and $\delta^*_{k,j} = (2\iota_{k,j} - 1)\delta(1 + U_{k,j})$ for $k \in \{4, 5, 6\}$, with again $\iota_j \sim B(1/2)$, $\iota_{k,j} \sim B(1/2)$ and $U_{k,j} \sim \mathcal{U}_{[\sqrt{K}/2, 2\sqrt{K}]}$. Finally, in the fourth configuration (full heterogeneity), we set $\delta^*_{k,j} = (2\iota_{k,j} - 1)\delta(1 + U_{k,j})$ for $k \in \{1, ..., K-1\}$ with again $\iota_{k,j} \sim B(1/2)$ and $U_{k,j} \sim \mathcal{U}_{[\sqrt{K}/2, 2\sqrt{K}]}$. In each configuration, parameter δ varied in $\{0.1, 0.25, 0.5, 0.75\}$ to study the impact of signal strength on the performance of the approaches; these δ values correspond to log-odds-ratio for an increment of one standard deviation.

For each observation, covariates were generated under a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$, where $\mathbf{\Sigma}_{i,j} = 0.3^{|i-j|}$. Pairs of observations were then created and randomly assigned to one stratum $\mathcal{M}_k$ in such a way that $m_1 = 200$, $m_2 = 100$ and $m_k = 50$ for $k = 3, \ldots, 6$. Within each pair ℓ of each stratum $\mathcal{M}_k$, the response variable $Y^{(k)}_{\ell,1}$ was then generated under model (2.1), that is, $Y^{(k)}_{\ell,1}$ was drawn from a Bernoulli distribution with parameter equal to

$$\frac{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,1})}{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,1}) + \exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,2}))}.$$

Then, $Y^{(k)}_{\ell,2}$ was set to $1 - Y^{(k)}_{\ell,1}$. Denoting by *case* [resp. *control*] the index of the case [resp. control] in the ℓ-th generated pair, we then have

$$\Pr(Y^{(k)}_{\ell,case} = 1 | Y^{(k)}_{\ell,case} + Y^{(k)}_{\ell,control} = 1, \mathbf{x}^{(k)}_{\ell,case}, \mathbf{x}^{(k)}_{\ell,control}) = \frac{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case})}{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case}) + \exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,control})},$$

and our data are indeed generated under model (2.1).

A.5 *Additional results from the simulation study*

Figure 1 illustrates the distribution of the criteria whose averages are presented on Figure 1 (matched setting) in the main text.

Figure 2 illustrates the distribution of the criteria whose averages are presented on Figure 2 (unmatched setting) in the main text.

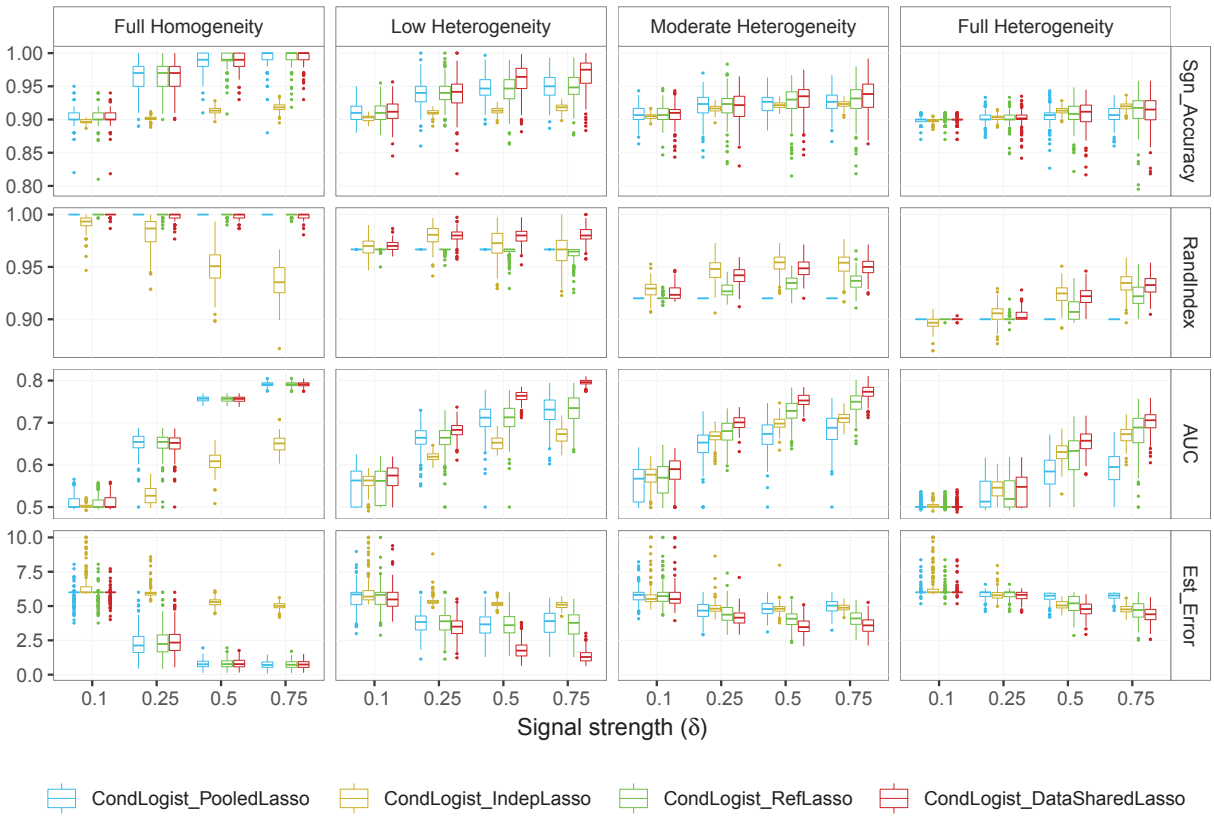

Fig. 1. Boxplots showing the distributions of the criteria for each of the four methods compared in the matched setting, over the 200 replicates of each considered configuration and signal strength.

A.6 *The influence of the reference category when using MultinomLogist_StdLasso: a toy example*

Consider a multinomial logistic regression model with $K-1$ disease subtypes, in the particular case of full homogeneity, that is when covariates associated with disease have the same level of association with all disease subtypes. For any $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model, we would then have (i) $\boldsymbol{\beta}_1^* = \cdots = \boldsymbol{\beta}_{K-1}^*$, and (ii) $\boldsymbol{\beta}_K^* \neq \boldsymbol{\beta}_1^*$ (assuming as in the main text that $Y = K$ indicates controls, while $Y = k$ for $k \in \{1, ..., K-1\}$ indicates disease subtype k). Let p_0 denote the number of covariates associated with the disease, for some $1 \leqslant p_0 \leqslant p$. If the standard formulation of the model is used with controls as the reference, parameters to be estimated will be $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$, with $\boldsymbol{\delta}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_K^*$ for $k = 1, \ldots, K-1$. Then, these $K-1$ vectors are all equal (since $\boldsymbol{\beta}_1^* = \ldots = \boldsymbol{\beta}_{K-1}^*$), and they all have p_0 non-zero components. The total number of non-zero parameters (hereafter referred to as model complexity) is therefore $(K-1)p_0$. Now, consider again the standard formulation, but this time using the first disease subtype as the reference. The parameters to be estimated would then be $\tilde{\boldsymbol{\delta}}_2^*, \ldots, \tilde{\boldsymbol{\delta}}_K^*$, with $\tilde{\boldsymbol{\delta}}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*$ for $k = 2, \ldots, K$. Because $\boldsymbol{\beta}_1^* = \ldots = \boldsymbol{\beta}_{K-1}^*$, the $K-2$ vectors $(\tilde{\boldsymbol{\delta}}_k^*)_{k=2,\ldots,K-1}$ are all null, while $\tilde{\boldsymbol{\delta}}_K^* = -\boldsymbol{\delta}_1^*$ has p_0 non-zero components. The model complexity is therefore p_0, which is much lower

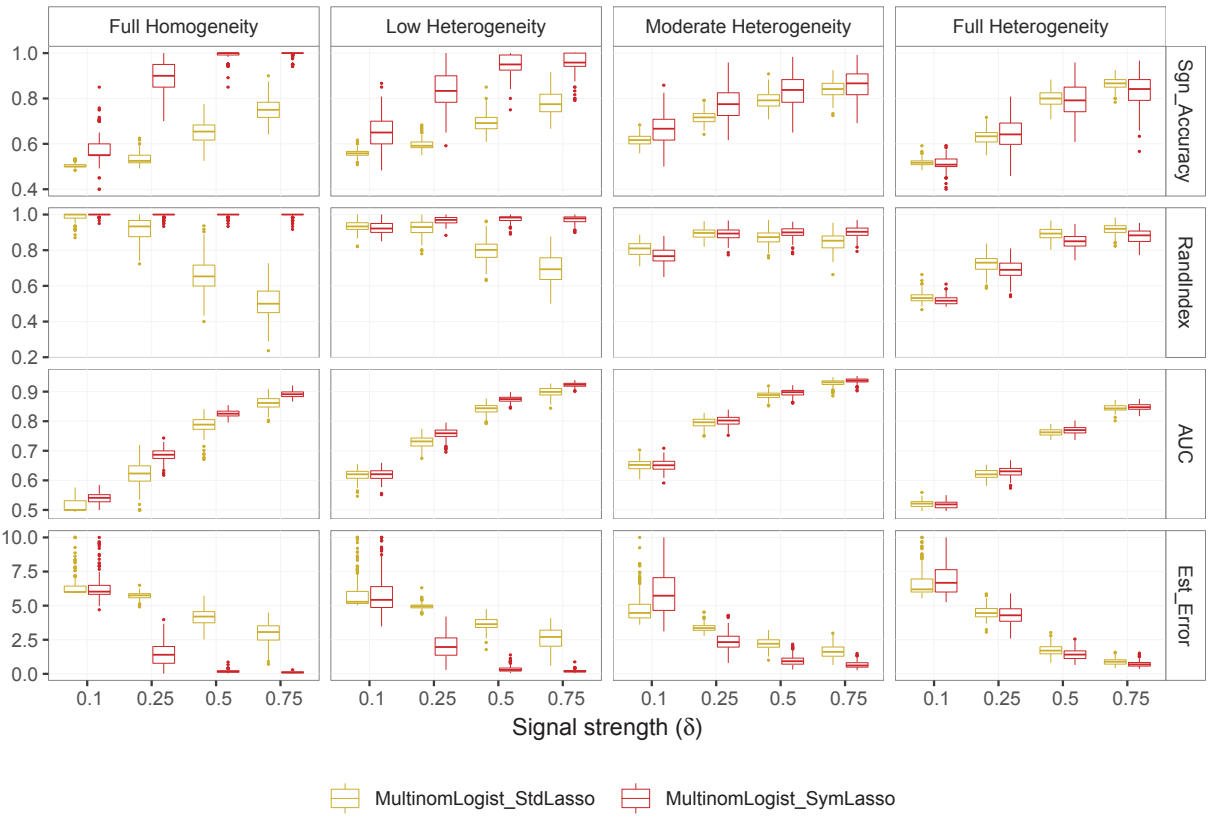

Fig. 2. Boxplots showing the distributions of the criteria for the two methods compared in the unmatched setting, over the 200 replicates of each considered configuration and signal strength.

than $(K-1)p_0$. In this particular case, using subtype 1 (or any other subtype) as the reference category yields a parametrization which is much sparser than the parametrization we consider when using controls as the reference. Therefore, when applying MultinomLogist_StdLasso, controls represent the worst choice for the reference category in this situation (we recall that the choice of the reference category would not have any influence on the quality of the estimation in an unpenalized framework). Still considering this toy example, applying MultinomLogist_StdDataSharedLasso corrects any suboptimal choice for the reference category. For example, whether the reference category is set to controls or to the first disease subtype, the model complexity is p_0 when applying MultinomLogist_StdDataSharedLasso, as illustrated in Figure 3 below. Finally, MultinomLogist_SymLasso would target the sparsest collection of vectors $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model. In the toy example considered here, this would be $(\mathbf{0}_p, \ldots, \mathbf{0}_p, \tilde{\boldsymbol{\delta}}_K^*)$, as illustrated in Figure 3. The corresponding model complexity is again p_0.

Figure 3 below gives a graphical representation of our toy example. It especially illustrates how model complexity (denoted by C) is affected by the choice of the reference category when working under the standard formulation of the multinomial logistic regression model, and how the decomposition targeted

by data shared lasso corrects any sub-optimal choice for this reference category. For simplicity, Figure 3 represents the case where $p_0 = p$, and where all covariates share the same level of association with the disease. In this case, a typical collection of vectors $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model is such that, for all $j \in \{1, ..., p\}$, $\beta_{1,j}^* = \beta_1$ and $\beta_{K,j}^* = \beta_2$ for some $\beta_1, \beta_2 \neq 0$ such that $\beta_1 - \beta_2 \neq 0$.

References

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters* **27**(8), 861 – 874. ROC Analysis in Pattern Recognition.

Ollier, E. and Viallon, V. (2017). Regression modeling on stratified data with the lasso. *Biometrika* **104**(1), 84–96.

Provost, F. and Domingos, P. (2000). Well-trained pets: Improving probability estimation trees. CeDER Working Paper #IS-00-04, Stern School of Business, New York University, NY 10012.

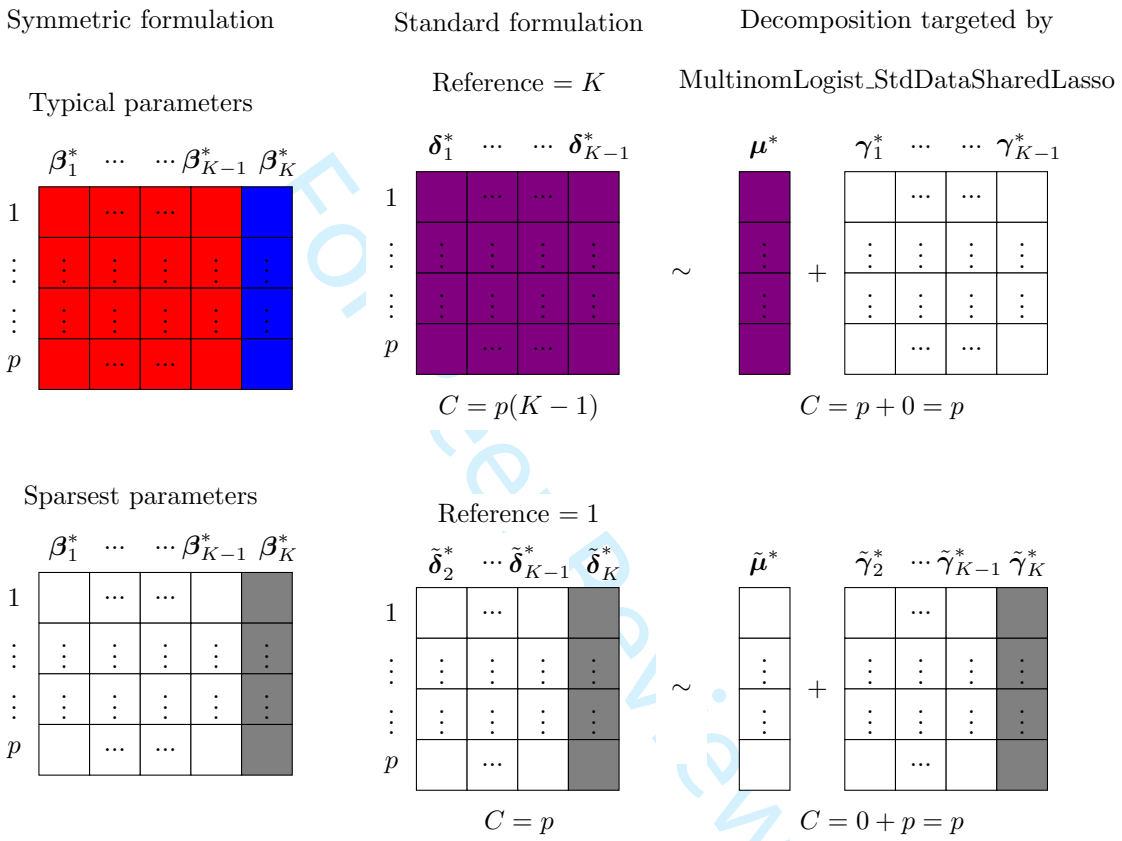

Fig. 3. Graphical representation of our toy example. In each matrix, red entries correspond to the "common" value β_1, blue entries correspond to the value β_2, purple entries to the value $(\beta_1 - \beta_2)$, gray entries to the value $-(\beta_1 - \beta_2)$ and white entries to the value 0. If MultinomLogist_StdLasso is applied after selecting the K-th category as the reference, model complexity is $C = (K-1)p$. The choice of the K-th categoty as the reference is clearly sub-optimal since selecting any other category as the reference, e.g. the first one, leads to a model complexity $C = p$. On the other hand, irrespective of the initial choice of the reference category, the complexity of the decomposition targeted by MultinomLogist_StdDataSharedLasso is optimal and equals p.