Regression modeling on stratified data with the lasso

Edouard Ollier, Vivian Viallon

Regression modeling on stratified data with the lasso

E. Ollier[⋆] **and V. Viallon**[‡]

[⋆] *Ecole Normale Supérieure de Lyon 46 Allée d'Italie, 69364 Lyon Cedex 07, France*
[‡]*Université de Lyon, F-69622, Lyon, France; Université Lyon 1, UMRESTTE, F-69373 Lyon; IFSTTAR, UMRESTTE, F-69675 Bron.*
vivian.viallon@univ-lyon1.fr edouard.ollier@ens-lyon.fr

Abstract

We consider the estimation of regression models on strata defined using a categorical covariate, in order to identify interactions between this categorical covariate and the other predictors. A basic approach requires the choice of a reference stratum. We show that the performance of a penalized version of this approach depends on this arbitrary choice. We propose a refined approach that bypasses this arbitrary choice, at almost no additional computational cost. Regarding model selection consistency, our proposal mimics the strategy based on an optimal and covariate-specific choice for the reference stratum. Results from an empirical study confirm that our proposal generally outperforms the basic approach in the identification and description of the interactions. An illustration is provided on gene expression data.

1. Introduction

We consider the estimation of regression models when the population under study is stratified, as is standard in epidemiology and clinical research. For instance, when studying relapse after a primary breast cancer, it is now common to analyze various histological subtypes at once (Voduc et al., 2010; Rosner et al., 2013). In order to accurately estimate the risk of relapse according to cancer subtype, risk factors that interact with cancer subtype need to be identified, and the corresponding interactions need to be precisely described. In pharmacokinetics, it is often of interest to describe how parameters related to absorption and clearance depend on dosage and type of adjuvant. In Ollier et al. (2016), for example, non-linear mixed effect models are estimated on strata defined according to the treatment dosage and the type of adjuvant. In Section 4, we study the association between the expression of the epidermal growth factor receptor gene, and those of 44 other transcription factors at eight time points of a differentiation process. We use a data set described in Kouno et al. (2013) where expression of these factors has been measured by profiling, for each time point, 120 different single cells. In this application, our aim is to describe how the association between the epidermal growth factor receptor gene and the other 44 transcription factors varies over these eight strata.

In all these examples, the general objective is to study the relationship between a response variable $y \in \mathbb{R}$ and a vector of $p \geq 1$ predictors $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ over $K \geq 1$ strata defined according to a categorical covariate Z of primary interest, such as cancer subtype. A key objective is to determine how Z modifies the effect of x on y, that is to identify and describe interactions between predictors x and the categorical effect modifier Z (Gertheiss & Tutz, 2012). Identifying and taking account of these interactions is important even when the assessment of categorical effect

modification is not the main objective. In particular, estimating one model per stratum or one model on all the strata pooled together generally leads to overfitting or underfitting, respectively.

For simplicity, we mostly focus on linear regression models. Denote by $\beta_k^* = (\beta_{k,1}^*, \ldots, \beta_{k,p}^*) \in \mathbb{R}^p$, for any $k \in [K] = \{1, \ldots, K\}$, the parameter vector describing the association between y and x in the model corresponding to stratum $Z = k$. For any $j \in [p]$, denote by $d_j \in [K]$ the number of distinct values in the set $\{\beta_{1,j}^*, \ldots, \beta_{K,j}^*\}$. Further consider the partition $\{\mathcal{K}_j^{(1)}, \ldots, \mathcal{K}_j^{(d_j)}\}$ of $[K]$ such that $\beta_{k_1,j}^* = \beta_{k_2,j}^*$ if and only if $(\beta_{k_1,j}^*, \beta_{k_2,j}^*) \in \mathcal{K}_j^{(d)}$ for some $d \in [d_j]$. The full description of how Z modifies the effect of x_j on y relies on the identification of this partition. The total number of partitions of $[K]$ is the Kth Bell number B_K, with, for instance, $B_5 = 52$. Standard statistical test procedures would require the comparison of $(B_K)^p$ models for the identification of the p partitions, so they are not well-suited to fully describe how Z modifies the effect of x on y.

Penalized approaches have been advocated in this context, which can be seen as a special case of multi-task learning (Evgeniou & Pontil, 2004). In particular, a version of the adaptive generalized fused lasso (Tibshirani et al., 2005; Viallon et al., 2016) has been shown to enjoy an asymptotic oracle property if Kp does not grow with the sample size n (Gertheiss & Tutz, 2012; Oelker et al., 2014). If Kp is fixed then this method identifies partitions $\{\mathcal{K}_j^{(1)}, \ldots, \mathcal{K}_j^{(d_j)}\}$ for all $j \in [p]$ with probability tending to one as $n \to \infty$, under mild assumptions. However, theoretical results in a non-asymptotic framework are still lacking for this method. Results obtained by Sharpnack et al. (2012) for generalized fused lasso estimates in the Gaussian means setting are not straightforward to extend to our context but they suggest that this method might only be able to identify the partitions under very particular settings. See Section 2.6 for more details.

An alternative strategy, which is standard in epidemiology, consists in first selecting a reference stratum, say the first one, then coding the other strata by indicators $\mathbb{1}(Z = k)$, and finally including them into the regression model, along with their interactions with the predictors. This corresponds to considering the decompositions $\beta_k^* = \beta_1^* + \delta_k^*$, for all $k \in [K]$, with $\delta_1^* = 0_p$, the null vector in $\mathbb{R}^p$. Then, the lasso (Tibshirani, 1996) can be used to identify null components in vectors β_1^* and δ_k^*, for $k \neq 1$. Of course, this strategy can generally identify only one element of each partition $\{\mathcal{K}_j^{(1)}, \ldots, \mathcal{K}_j^{(d_j)}\}$. However, a natural question is whether it is sparsistent, that is whether it can identify the sets $S_1^* = \{j \in [p] : \beta_{1,j}^* \neq 0\}$ and $T_1^* = \{(k, j) \in [K] \times [p] : \beta_{k,j}^* \neq \beta_{1,j}^*\}$ with high probability. If so, it could partially describe the effect of Z on the association between y and x, while results of Sharpnack et al. (2012) suggest that the method based on the fused lasso generally fails to fully describe it.

In this article, we establish that the sparsistency of this basic approach depends on the choice of the reference stratum. We then present a refined approach which bypasses this arbitrary choice. Under linear regression models, we show that our proposal is sparsistent under conditions similar to those ensuring the sparsistency of the approach based on an optimal and covariate-specific choice of the reference stratum. In addition, our proposal can be implemented with available packages under a variety of models, at approximately the same computational cost as that of the basic approach.

2. Methods

2.1 Notations and setting

Methods are first presented in the linear regression model for ease of notation. Extensions to generalized linear models (McCullagh & Nelder, 1989) are briefly presented in Section 2.4.

For any positive integer $m \geq 1$, define $[m] = \{1, \ldots, m\}$. Let 0_m and 1_m be the vectors of size m with components all equal to 0 and 1 respectively, and let I_m be for the $(m \times m)$ identity matrix. For any vector $x = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$, let $\text{supp}(x) = \{j \in [m] : x_j \neq 0\}$ denote its support. We further set $\|x\|_q = (\sum_{j\in[m]} |x_j|^q)^{1/q}$ for any real number $q \in (0, \infty)$, $\|x\|_\infty = \max_j |x_j|$ and $\|x\|_0 = |\text{supp}(x)|$, where $|E|$ is the cardinality of the set E. For any set $E \subseteq [m]$, let x_E denote the vector of $\mathbb{R}^{|E|}$ with components $(x_j)_{j\in E}$. For any real matrix M, let M_j be its jth column, and let M_E be the sub-matrix made of columns $(M_j)_{j\in E}$. Further denote the smallest singular value of M by $\Lambda_{\min}(M)$. Finally, $\mathbb{1}(\cdot)$ is the indicator function.

Denote the number of levels of variable Z, that is the number of strata, by $K \geq 1$. Let n_k be the number of observations in stratum $k \in [K]$, and denote the total number of observations by $n = \sum_{k\in[K]} n_k$. For $k \in [K]$, further denote the response vector in stratum k by $y^{(k)} \in \mathbb{R}^{n_k}$. Similarly, let $X^{(k)}$ be the $(n_k \times p)$ design matrix in stratum k. For all $k \in [K]$, we assume that $y^{(k)} = X^{(k)}\beta_k^* + \varepsilon^{(k)}$, with noise vector $\varepsilon^{(k)} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$. Vectors $\beta_k^* \in \mathbb{R}^p$ include the Kp parameters to be estimated.

2.2 Basic approach

A basic approach consists in picking a reference stratum for any j, say ℓ_j. Most often in practice, ℓ_j is chosen so that it does not depend on j; here we consider the most general version. Introduce $\ell = (\ell_1, \ldots, \ell_p) \in [K]^p$, $\mu_\ell^* = (\beta_{\ell_1,1}^*, \ldots, \beta_{\ell_p,p}^*)$ and $\delta_k^* = \beta_k^* - \mu_\ell^*$. The basic approach relies on the decomposition $\beta_k^* = \mu_\ell^* + \delta_k^*$, for all $k \in [K]$, with $\delta_{\ell_j,j}^* = 0$ (Gertheiss & Tutz, 2012). Following the lasso (Tibshirani, 1996), estimates of μ_ℓ^* and δ_k^*, $k \in [K]$, can be defined as

$$\underset{\substack{\mu,\delta_1,\ldots,\delta_K \\ \delta_{\ell_j,j}=0 \text{ for all } j\in[p]}}{\text{argmin}} \left\{ \sum_{k=1}^K \frac{\|y^{(k)} - X^{(k)}(\mu + \delta_k)\|_2^2}{2n} + \lambda_1\|\mu\|_1 + \sum_{k=1}^K \lambda_{2,k}\|\delta_k\|_1 \right\}, \tag{1}$$

for appropriate non-negative λ_1 and $\lambda_{2,k}$. As will be seen in Sections 2.5 and 2.6, conditions ensuring the sparsistency of this approach depend on the arbitrary choice of the vector of reference strata ℓ. As a matter of fact, the underlying model dimension is $\|\mu_\ell^*\|_0 + \sum_{k\neq\ell} \|\beta_k^* - \mu_\ell^*\|_0$, which depends on ℓ. The lowest dimension, and then the best possible performance, is attained if $\mu_{\ell_j}^*$ is a mode of the collection of values $(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*)$, $\text{mode}(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*)$, for all $j \in [p]$. For any $j \in [p]$, such an optimal and covariate-specific reference stratum will be denoted by ℓ_j^* below. Because $\ell^* = (\ell_1^*, \ldots, \ell_p^*)$ is generally unknown, the corresponding optimal version of the basic approach cannot be implemented in practice.

2.3 Our proposal

Our proposal aims at bypassing the arbitrary choice of the vector of reference strata ℓ, while mimicking the optimal version of the basic approach. We consider an overparametrization involving $(K + 1)p$ parameters, $\beta_k^* = \mu^* + \gamma_k^*$ for any $k \in [K]$. It generalizes the decomposition used in the basic approach in the sense that no coefficient is constrained to be zero here. For nonnegative values of λ_1 and $\lambda_{2,k}$'s, our proposal returns

$$(\widehat{\mu}, \widehat{\gamma}_1 \ldots, \widehat{\gamma}_K) \in \underset{\mu,\gamma_1,\ldots,\gamma_K}{\text{argmin}} \left\{ \sum_{k=1}^K \frac{\|y^{(k)} - X^{(k)}(\mu + \gamma_k)\|_2^2}{2n} + \lambda_1\|\mu\|_1 + \sum_{k=1}^K \lambda_{2,k}\|\gamma_k\|_1 \right\}. \tag{2}$$

Working with a large enough value for $\lambda_{2,r}$, for some $r \in [K]$, is equivalent to constraining $\widehat{\gamma}_r = 0_p$, and then $\widehat{\mu} = \widehat{\beta}_r$, and reduces to the basic approach with $\ell = (r, \ldots, r)$. More generally, setting $\tau = (\tau_1, \ldots, \tau_K)$ with $\tau_k = \lambda_{2,k}/\lambda_1$, and defining the shrunk and τ-weighted version of the median of $(b_1, \ldots, b_K)$ as $\mathrm{WSmedian}(b_1, \ldots, b_K; \tau) = \mathrm{argmin}_b(|b| + \sum_{k\in[K]} \tau_k |b_k - b|)$, it is easy to see that $\widehat{\mu}_j \in \mathrm{WSmedian}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}; \tau)$. In other words, for any particular value of the $\lambda_{2,k}/\lambda_1$ ratios, our approach encourages solutions $(\widehat{\beta}_1 = \widehat{\mu} + \widehat{\gamma}_1, \ldots, \widehat{\beta}_K = \widehat{\mu} + \widehat{\gamma}_K)$ with a sparse vector $\widehat{\mu}$ and sparse vectors of differences $\widehat{\gamma}_k = \widehat{\beta}_k - \widehat{\mu}$, and with the overall effect of the jth covariate $\widehat{\mu}_j$ defined as $\mathrm{WSmedian}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}; \tau)$.

Moreover, working with a large enough value for λ_1 is equivalent to constraining $\widehat{\mu} = 0_p$, and our approach then reduces to K independent lasso's. In contrast, working with large enough $\lambda_{2,k}$ values is equivalent to constraining $\widehat{\beta}_k = \widehat{\mu}$ for all k, and our approach then reduces to one lasso run on all the strata pooled together.

2.4 Rewriting as a lasso on a transformation of the original data

Our proposal reduces to the lasso on a simple transformation of the original data, just as the basic approach does. Set $\mathscr{Y} = ({y^{(1)}}^T, \ldots, {y^{(K)}}^T)^T$ the vector containing the n observations of the response variable. For any $k \in [K]$, introduce $P_\ell^{(k)} = \{j \in [p] : k \neq \ell_j\}$, with ℓ_j still denoting the reference stratum chosen for covariate j in the basic approach. Note that $\sum_k |P_\ell^{(k)}| = (K-1)p$ and set $\tilde{X}_\ell^{(k)} = X^{(k)}_{P_\ell^{(k)}}$. Now introduce

$$\mathcal{X}_\ell = \begin{pmatrix} X^{(1)} & \tilde{X}_\ell^{(1)}/\tau_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ X^{(K)} & 0 & \ldots & \tilde{X}_\ell^{(K)}/\tau_K \end{pmatrix}, \qquad \mathcal{X}_0 = \begin{pmatrix} X^{(1)} & X^{(1)}/\tau_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ X^{(K)} & 0 & \ldots & X^{(K)}/\tau_K \end{pmatrix},$$

Criteria to be minimized in (1) and (2) reduce to the lasso

$$\frac{1}{2n}\|\mathscr{Y} - \mathcal{X}\theta\|_2^2 + \lambda_1 \|\theta\|_1, \tag{3}$$

with $\mathcal{X}$ set to either $\mathcal{X}_\ell$, for the basic approach, or $\mathcal{X}_0$, for our proposal, and θ a vector of $\mathbb{R}^{Kp}$ or $\mathbb{R}^{(K+1)p}$ as appropriate. Therefore, our proposal comes at almost no additional computational cost compared to the basic approach.

This rewriting as a lasso extends to generalized linear models, Cox models, etc., and makes our proposal directly implementable using the glmnet package of Friedman et al. (2010), for instance. Considering logistic models, set $\mathcal{L}_{\mathrm{logistic}}(y, z) = \sum_{i\in[n]} y_i z_i - \log(1 + e^{z_i})$ for any $y \in \{0, 1\}^n$ and $z \in \mathbb{R}^n$. The criterion of our proposal writes as the logistic lasso, $-\mathcal{L}_{\mathrm{logistic}}(\mathscr{Y}, \mathcal{X}_0\theta) + \lambda_1\|\theta\|_1$. In addition, the glmnet package can benefit from the sparse structure of $\mathcal{X}_0$.

2.5 Sparsistency

From now on we assume that $\bar{\beta}_j^* = \mathrm{mode}(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*)$ is uniquely defined for all $j \in [p]$, for ease of notation. For any vector of reference strata $\ell = (\ell_1, \ldots, \ell_p) \in [K]^p$, introduce the sets $S_\ell = \{j \in [p] : \beta^*_{\ell_j,j} \neq 0\}$ and $T_\ell = \{(k, j) \in [K] \times [p] : \beta^*_{k,j} \neq \beta^*_{\ell_j,j}\}$. Further define $\theta_\ell^* = (\mu_\ell^{*T}, \tau_2\gamma_{\ell,2}^{*T}, \ldots, \tau_K\gamma_{\ell,K}^{*T})^T \in \mathbb{R}^{Kp}$, with $\mu^*_{\ell,j} = \beta^*_{\ell,j}$ and $\gamma^*_{\ell,k} = (\beta_k^* - \mu_\ell^*)_{P_k^{(\ell)}}$. The

basic approach is sparsistent if it identifies these two sets, or, equivalently, the set $J_\ell = \mathrm{supp}(\theta^*_\ell)$, with high probability. Now, consider an optimal vector of reference strata $\ell^* = (\ell^*_1, \ldots, \ell^*_p)$ such that $\beta^*_{\ell^*_j, j} = \bar{\beta}^*_j$ for all $j \in [p]$. Define $\theta^*_0 = (\mu^{*T}_{\ell^*}, \tau_1 \gamma^{*T}_{0,1}, \ldots, \tau_K \gamma^{*T}_{0,K})^T \in \mathbb{R}^{(K+1)p}$ where $\gamma^*_{0,k} = \beta^*_k - \mu^*_{\ell^*}$. We will say our proposal is sparsistent if it identifies S_{ℓ^*} and T_{ℓ^*}, or, equivalently, the set $J_0 = \mathrm{supp}(\theta^*_0)$, with high probability.

For the lasso to be sparsistent, a sufficient and almost necessary condition on the design matrix is the irrepresentability condition (Zhao & Yu, 2006; Wainwright, 2009). Consider the general formulation (3) of the lasso and denote by θ^* the true value of the parameter vector to be estimated. Defining $J^* = \mathrm{supp}(\theta^*)$, the matrix $\mathcal{X}$ fulfills the irrepresentability condition, with respect to J^*, if $\Lambda_{\min}(\mathcal{X}^T_{J^*}\mathcal{X}_{J^*}) \geq C_{\min}$ for some $C_{\min} > 0$ and $\max_{j \notin J^*} \|(\mathcal{X}^T_{J^*}\mathcal{X}_{J^*})^{-1}\mathcal{X}^T_{J^*}\mathcal{X}_j\|_1 < 1$. Using the vector of reference strata ℓ in the basic approach, the irrepresentability condition of matrix $\mathcal{X}_\ell$ writes as $(IC)_\ell$, while in our approach, the irrepresentability condition of matrix $\mathcal{X}_0$ writes as $(IC)_0$:

$$(IC)_\ell \qquad \Lambda_{\min}(\mathcal{X}^T_{\ell J_\ell}\mathcal{X}_{\ell J_\ell}) \geq C_\ell > 0 \text{ and } c_\ell = \max_{j \notin J_\ell} \|(\mathcal{X}^T_{\ell J_\ell}\mathcal{X}_{\ell J_\ell})^{-1}\mathcal{X}^T_{\ell J_\ell}\mathcal{X}_{\ell j}\|_1 < 1,$$

$$(IC)_0 \qquad \Lambda_{\min}(\mathcal{X}^T_{0J_0}\mathcal{X}_{0J_0}) \geq C_0 > 0 \text{ and } c_0 = \max_{j \notin J_0} \|(\mathcal{X}^T_{0J_0}\mathcal{X}_{0J_0})^{-1}\mathcal{X}^T_{0J_0}\mathcal{X}_{0j}\|_1 < 1.$$

The comparison of $(IC)_{\ell^*}$ and $(IC)_0$ is of particular interest. First, because $\theta^*_{\ell^* J_{\ell^*}} = \theta^*_{0J_0}$, we have $\mathcal{X}_{\ell^* J_{\ell^*}} = \mathcal{X}_{0J_0}$ and $\Lambda_{\min}(\mathcal{X}^T_{\ell^* J_{\ell^*}}\mathcal{X}_{\ell^* J_{\ell^*}}) = \Lambda_{\min}(\mathcal{X}^T_{0J_0}\mathcal{X}_{0J_0})$. Second, the maxima in the definitions of c_0 and c_{ℓ^*} are taken over J_0^c and $J_{\ell^*}^c$, respectively, with $|J_0^c| = p + |J_{\ell^*}^c|$. Indeed, columns corresponding to $\gamma^*_{0,\ell^*_j,j}$, for $j \in [p]$, are present in matrix $\mathcal{X}_0$ while they are absent from $\mathcal{X}_{\ell^*}$. Moreover, the corresponding indexes belong to J_0^c since $\gamma^*_{0,\ell^*_j,j} = 0$. Therefore, the only difference between $(IC)_0$ and $(IC)_{\ell^*}$ comes from the fact that $c_0 \geq c_{\ell^*}$, and $(IC)_0$ is only slightly stronger than $(IC)_{\ell^*}$. Focusing on the case of balanced strata and orthogonal designs in each stratum, Lemma 2 in Section 2.6 states that these two conditions are identical in this particular case. It further explicitly relates them to the ratios $\tau_k = \lambda_{2,k}/\lambda_1$ and to the maximum level of heterogeneity that is allowed among the collections of values $(\beta^*_{1,j} \ldots, \beta^*_{K,j})$, for all $j \in [p]$. In addition, they are shown to be generally weaker than $(IC)_\ell$ for the choice $\ell = (r, \ldots, r)$, for some $r \in [K]$.

We can now state Theorem 1, according to which our proposal identifies S_{ℓ^*} and T_{ℓ^*} under nearly the same assumptions as those required by the optimal version of the basic approach if the ℓ^*_j's were given in advance. More precisely, besides the fact that $(IC)_0$ is generally a little stronger that $(IC)_{\ell^*}$, the only difference lies in the terms $(\lambda^{(1)}_1, \beta^{(1)}_{\min})$ and $(\lambda^{(0)}_1, \beta^{(0)}_{\min})$, where $K+1$ replaces K. Our result is a consequence of Theorem 1 in Wainwright (2009).

Theorem 1 *Assume that the noise variables $(\varepsilon^{(k)}_i)_{i \in [n_k], k \in [K]}$ are independent and identically distributed centered sub-Gaussian variables with parameter $\sigma > 0$. Further assume that $n_k^{-1/2}\|X^{(k)}_j\|_2 \leq 1$ for all $(k, j) \in [K] \times [p]$. Introduce $\tau_0 > 0$ and set $\tau_k = \tau_0(n_k/n)^{1/2}$ for all $k \in [K]$. If $(IC)_{\ell^*}$ holds then set $\gamma_1 = 1 - c_{\ell^*}$. Further set $\gamma_0 = 1 - c_0$ if $(IC)_0$ holds. For $\eta \in \{0, 1\}$, introduce*

$$\lambda^{(\eta)}_1 > \frac{2}{\gamma_\eta \min(1, \tau_0)} \left\{ \frac{2\sigma^2 \log((K+\eta)p))}{n} \right\}^{1/2}, \quad \beta^{(\eta)}_{\min} = \lambda^{(\eta)}_1 \left\{ \frac{(|S_{\ell^*}| + |T_{\ell^*}|)^{1/2}}{C_{\ell^*}} + 4\frac{\sigma}{C^{1/2}_{\ell^*}} \right\}.$$

If $(IC)_{\ell^*}$ holds then solutions $\widehat{\theta}_{\ell^*}$ of (3) with $\mathcal{X} = \mathcal{X}_{\ell^*}$ and $\lambda_1 = \lambda^{(0)}_1$ as above are such that, with probability at least $1 - 4\exp(-a_0 n \lambda_1)$ for some constant $a_0 > 0$: (i) $\widehat{\theta}_{\ell^*}$ is uniquely defined,

(ii) $\widehat{\theta}_{\ell^* J_{\ell^*}^c} = 0_{|J_{\ell^*}^c|}$, *and (iii)* $\|\widehat{\theta}_{\ell^* J_{\ell^*}} - \theta^*_{\ell^* J_{\ell^*}}\|_\infty \leq \beta^{(0)}_{\min}$. *If, in addition,* $|\bar{\beta}^*_j| > \beta^{(0)}_{\min}$ *for all* $j \in S_{\ell^*}$ *and* $|\beta^*_{k,j} - \bar{\beta}^*_j| > \beta^{(0)}_{\min}/\tau_k$ *for all* $(k,j) \in T_{\ell^*}$, *then* J_{ℓ^*}, *hence both* S_{ℓ^*} *and* T_{ℓ^*}, *are perfectly identified with probability at least* $1 - 4\exp(-a_0 n \lambda_1^2)$.

If $(IC)_0$ *holds then solutions* $\widehat{\theta}_0$ *of (3) with* $\mathcal{X} = \mathcal{X}_0$ *and* $\lambda_1 = \lambda_1^{(1)}$ *as above are such that, with probability at least* $1 - 4\exp(-a_1 n \lambda_1)$ *for some constant* $a_1 > 0$*:* (i) $\widehat{\theta}_0$ *is uniquely defined,* (ii) $\widehat{\theta}_{0 J_0^c} = 0_{|J_0^c|}$, *and* (iii) $\|\widehat{\theta}_{0 J_0} - \theta^*_{\ell^* J_{\ell^*}}\|_\infty \leq \beta^{(1)}_{\min}$. *If, in addition,* $|\bar{\beta}^*_j| > \beta^{(1)}_{\min}$ *for all* $j \in S_{\ell^*}$ *and* $|\beta^*_{k,j} - \bar{\beta}^*_j| > \beta^{(1)}_{\min}/\tau_k$ *for all* $(k,j) \in T_{\ell^*}$, *then* J_0, *hence both* S_{ℓ^*} *and* T_{ℓ^*}, *are perfectly identified with probability at least* $1 - 4\exp(-a_1 n \lambda_1^2)$.

This result especially confirms that it is harder to identify T_{ℓ^*} than S_{ℓ^*}, in the sense that heterogeneities have to be at least $|\beta^*_{k,j} - \bar{\beta}^*_j| > (n/n_k)^{1/2}\beta^{(\eta)}_{\min}/\tau_0$ for $(k,j) \in T_{\ell^*}$, while $|\bar{\beta}^*_j|$ has only to be greater than $\beta^{(\eta)}_{\min}$ for $j \in S_{\ell^*}$.

2.6 The orthogonal and balanced case

It is instructive to inspect in more detail the simple setting where $n_k = n/K$ and $({X^{(k)}}^T X^{(k)})/n_k = I_{n_k}$ for all $k \in [K]$. This orthogonality assumption does not make matrices $\mathcal{X}_\ell$ and $\mathcal{X}_0$ orthogonal and is therefore not sufficient for the irrepresentability condition.

For any vector of reference strata ℓ and all $j \in [p]$, define $K^*_{\ell,j} = \{k \in [K] : \beta^*_{k,j} = \beta^*_{\ell_j,j}\}$. Set $\mathcal{D}_{\ell,0} = \max_{j \notin S_\ell} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{\ell_j,j}\}|$ if $S_\ell \neq [K]$ and 0 otherwise and $\mathcal{D}_{\ell,1} = \max_{j \in S_\ell} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{\ell_j,j}\}|$ if $S_\ell \neq \emptyset$ and $-\infty$ otherwise. For all $k \in [K]$, we set $\tau_k = \tau_0 K^{-1/2}$, for some $\tau_0 > 0$.

Lemma 2 *The matrix* $\mathcal{X}_\ell$ *fulfills the irrepresentability condition if and only if*

$$(sIC)_\ell \qquad 0 \leq \frac{K^{1/2}}{K - 2\mathcal{D}_{\ell,1}} < \tau_0 < \frac{K^{1/2}}{\mathcal{D}_{\ell,0}}.$$

The matrix $\mathcal{X}_0$ *fulfills the irrepresentability condition if and only if*

$$(sIC)_0 \qquad 0 \leq \frac{K^{1/2}}{K - 2\mathcal{D}_{\ell^*,1}} < \tau_0 < \frac{K^{1/2}}{\mathcal{D}_{\ell^*,0}}.$$

Conditions $(sIC)_0$ and $(sIC)_{\ell^*}$ in Lemma 2 are identical. In the orthogonal and balanced case, the two sets of assumptions required by our proposal and the optimal version of the basic approach to identify the sets S_{ℓ^*} and T_{ℓ^*} are therefore identical, except for the terms $(\lambda_1^{(1)}, \beta^{(1)}_{\min})$ and $(\lambda_1^{(0)}, \beta^{(0)}_{\min})$ where $K+1$ replaces K, as in Theorem 1 above.

In addition, $(sIC)_{\ell^*}$, or equivalently $(sIC)_0$, imposes that $2\mathcal{D}_{\ell^*,1} + \mathcal{D}_{\ell^*,0} < K$. In particular, if $\mathcal{D}_{\ell^*,1} = \mathcal{D}_{\ell^*,0} = \mathcal{D}_{\ell^*}$, this implies that $\mathcal{D}_{\ell^*} < K/3$. The irrepresentability conditions $(sIC)_{\ell^*}$ and $(sIC)_0$ are therefore directly related to the maximum level of heterogeneity among the values $(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$. Similarly, $(IC)_\ell$ imposes $2\mathcal{D}_{\ell,1} + \mathcal{D}_{\ell,0} < K$, which is generally a stronger constraint. For simplicity, consider the situation where $\ell = (r, \ldots, r)$ for some $r \in [K]$, which is a common choice in practice. Without loss of generality, set $r = 1$. Then $(IC)_\ell$ entails that, for each $j \in [p]$, the effect of the jth covariate on most strata is $\beta^*_{1,j}$, while $(sIC)_{\ell^*}$ and $(sIC)_0$ only entail that, for each $j \in [p]$, the effect of the jth covariate on most strata is $\mathrm{mode}(0, \beta^*_{1,j}, \ldots, \beta^*_{K,j})$.

Moreover, if $\ell = (1, \ldots, 1)$ and $\{1\} \notin \cap_{j\in[p]} K^*_{\ell^*,j}$ then we have $T_\ell \neq T_{\ell^*}$ and, possibly, $S_\ell \neq S_{\ell^*}$: the identification of T_ℓ and S_ℓ is both less interesting and less likely and our proposal should be preferred over the basic approach.

To recap, our non-asymptotic analysis shows that the partial description of the categorical effect modification due to Z through decompositions of the type $\beta^*_k = \mu^* + \gamma^*_k$ is guaranteed only when the level of heterogeneity is not too high, that is when the number of nonzero components in vectors $(\gamma^*_{1,j}, \ldots, \gamma^*_{K,j})$, for $j \in [p]$, is not too high. In particular, the lowest level of heterogeneity is attained for the choice $\mu^*_j = \mathrm{mode}(0, \beta^*_{1,j}, \ldots, \beta^*_{K,j})$. Our proposal is able to target this optimal decomposition and is sparsistent under nearly the same assumptions as those required for the optimal version of the basic approach.

2.7 Connection with the generalized fused lasso

The fact that the level of heterogeneity must be low to ensure the sparsistency of our proposal as well as the sparsistency of the optimal version of the basic approach has connections with other results in the literature. Consider for instance the approach mentioned in Section 1, which was proposed by Gertheiss & Tutz (2012); see also Oelker et al. (2014) and Viallon et al. (2016). This is based on a fusion penalty that encourages similarities among solutions $\widehat{\beta}_k \in \mathbb{R}^p$, $k \in [K]$. More precisely, for appropriate non-negative λ_1 and λ_2's, it returns estimators $\widehat{\beta}_k$ defined as minimizers of the criterion

$$\sum_{k\in[K]} \frac{\|y^{(k)} - X^{(k)}\beta_k\|_2^2}{2n} + \lambda_1 \sum_{k\in[K]} \|\beta_k\|_1 + \lambda_2 \sum_{k_1<k_2} \|\beta_{k_1} - \beta_{k_2}\|_1. \tag{4}$$

This criterion is that of the generalized fused lasso, where the graph used in the penalty is made of p cliques of size K (Viallon et al., 2016). The jth clique corresponds to the jth predictor and connects all the components in $(\beta_{1,j}, \ldots, \beta_{K,j})$: the $K(K-1)/2$ differences $|\beta_{k_1,j} - \beta_{k_2,j}|$, for $k_1 < k_2$, appear in the penalty term.

Both the basic approach and our proposal are related to generalized fused lasso estimates too. In particular, consider criterion (1) with the optimal choice ℓ^* for the vector of reference strata. It can be seen as a version of a generalized fused lasso, with a graph made of p star-graphs of size K instead of p cliques: for each $j \in [p]$, only the $K-1$ differences $|\beta_{k,j} - \beta_{\ell^*_j,j}|$ appear in the penalty term, for $k \neq \ell^*_j$ and $\ell^*_j \in [K]$ fixed.

Non-asymptotic analyses of the sparsistency of generalized fused lasso estimates are scarce in the literature. Sharpnack et al. (2012) study them in the normal means setting, which can be seen as a special case of the stratified linear regression considered here. Sharpnack et al. (2012) establish that generalized fused lasso estimates are sparsistent only if the graph used in the fused penalty is in good agreement with the true structure of the vector of parameters; see also Qian & Jia (2016). Although it is not straightforward to extend to our case, these results suggest that estimates derived from (4) can only be sparsistent if the level of heterogeneity is not too high.

Our results precisely quantify the maximum level of heterogeneity above which a version of generalized fused lasso estimates, based on star-graphs, can attain sparsistency in stratified regression, in the balanced and orthogonal case. Because star-graphs are less connected than cliques, it is likely that sparsistency for clique-based estimates, such as those minimizing criterion (4), requires an even lower maximum level of heterogeneity. That being said, sparsistency for clique-based estimates refers to the full identification of the partitions $\{\mathcal{K}^{(1)}_j, \ldots, \mathcal{K}^{(d_j)}_j\}$, for $j \in [p]$. For the

basic approach, its optimal version and our proposal, sparsistency refers to the identification of one element of this partition only, say $\mathcal{K}_j^{(1)}$, and its complementary $[K] \setminus \mathcal{K}_j^{(1)}$.

In the asymptotic regime, assuming that Kp is fixed and $n_k/n \to \rho_k$ for some $\rho_k \in (0,1)$, for all $k \in [K]$, oracle properties have been derived for adaptive versions of clique-based estimates under mild assumptions (Gertheiss & Tutz, 2012): in particular, no assumption regarding the level of heterogeneity is required to ensure perfect recovery of the full partition $\{\mathcal{K}_j^{(1)}, \ldots, \mathcal{K}_j^{(d_j)}\}$, for all $j \in [p]$. Similar results are easily derived for adaptive versions of the basic approach for instance. In view of (3), we can apply Theorem 2 of Zou (2006) to show that an adaptive version of the basic approach enjoys an oracle property too, without having to assume any irrepresentability condition. Here again, no assumption regarding the maximum level of heterogeneity is required, but the identification of only one element of the partition is guaranteed for all j.

To recap, clique-based estimates are optimal and should be preferred over our proposal or the basic approach in the asymptotic regime, assuming that Kp is fixed and $n_k/n \to \rho_k$ for some $\rho_k \in (0,1)$, for all $k \in [K]$. In a non-asymptotic setting, results for clique-based estimates are still lacking, while conditions ensuring the sparsistency of our proposal and the basic approach are established in the present article. In the following simulation study, we especially compare the clique-based strategy and our proposal on finite samples.

3. Simulation Study

Theorem 1 states that our proposal and the optimal version of the basic approach perform similarly with regard to the identification of S_{ℓ^*} and T_{ℓ^*} for appropriate values of λ_1 and τ_0, under technical assumptions on the design matrices. The main objective of this simulation study is to assess the empirical performance of our proposal under general designs, and for λ_1 and τ_0 selected by 5-fold cross-validation. Comparisons are made with the basic approach with the choice $\ell = (1, \ldots, 1)$ as well as an optimal choice ℓ^*. Clique-based estimates are also considered.

We set $K = 20$ and take $n_k \in \{10, 50, 100\}$ and $p \in \{20, 100, 500\}$. For each $k \in [K]$, rows of the design matrix $X^{(k)}$ are drawn from an $\mathcal{N}(0_p, \Sigma)$ distribution, with Σ the $(p \times p)$ Toeplitz matrix with element (i,j) equal to $0.5^{|i-j|}$. We then randomly select a subset $P_0 \subset [p]$ of size 20 and set $\beta^*_{k,j} = 0$ for all $j \notin P_0$ and $k \in [K]$. As for the values $(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$ for $j \in P_0$, we consider four levels of heterogeneity d_H. More precisely, for any given $d_H \in \{1, 3, 6, 9\}$, we set $\beta^*_{k,j} = 1$ for $k > d_H$, and $\beta^*_{k,j} = 1 + \delta^*_{k,j}$ for $k \leq d_H$ for 10 indexes j randomly selected in P_0. For the other 10 indexes in P_0, we set $\beta^*_{k,j} = 1$ for $k \leq d_H$, and $\beta^*_{k,j} = 1 + \delta^*_{k,j}$ for $k > d_H$. We further consider two cases for the $\delta^*_{k,j}$ values: they are either constantly set to $K^{1/2}$ or drawn from the uniform distribution on $[K^{1/2}/2, 2K^{1/2}]$ and then multiplied by ± 1 (with probability $1/2$). When the $\delta^*_{k,j}$'s are constant, the collection of values $(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$ for each covariate $j \in P_0$ is made of two groups of distinct values, of sizes $K - d_H$ and d_H. This situation should favor clique-based estimates. When $\delta^*_{k,j}$ is random, the collection of values $(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$ for each covariate $j \in P_0$ is made of $d_H + 1$ groups, one of size $K - d_H$, and the other d_H of size 1. This situation should favor our proposal and the optimal version of the basic approach. Observations of the response variable are then generated according to $y^{(k)} = X^{(k)}\beta^*_k + \varepsilon^{(k)}$, with each component of $\varepsilon^{(k)}$ drawn from an $\mathcal{N}(0, \sigma^2)$ distribution. The variance σ^2 is set to $\sum_{k \in [K]} \|X^{(k)}\beta^*_k\|_2^2/n$, giving overall signal-to-noise ratio equal to 1. For each particular choice of $n_k \in \{10, 50, 100\}$, $p \in \{20, 100, 500\}$ and $d_H \in \{1, 3, 6, 9\}$, and for both the random and constant choice for the $\delta^*_{k,j}$'s, we generate 50

replicates of data $(X^{(k)}, y^{(k)})$, $k \in [K]$. Our results correspond to averages over these 50 replicates; see Figure 1.

In all the configurations, $\beta^*_{20,j} = \text{mode}(0, \beta^*_1, \ldots, \beta^*_{20,j})$ for all $j \in [p]$. We then set $\ell^*_j = 20$ for all $j \in [p]$ for the optimal version of the basic approach. On the other hand, $\beta^*_{1,j} \neq \text{mode}(0, \beta^*_1, \ldots, \beta^*_{20,j})$ for all $j \in P_0$. Under this setting, which is of course extreme, the comparison between the results obtained using either ℓ or ℓ^* for the reference strata allows a precise description of the impact of the reference stratum on the performance of the basic approach. Top panels of Figure 1 present results regarding the identification of the sets $T^*_{P_0} = \{(k,j) \in [K] \times P_0 : \beta^*_{k,j} \neq \beta^*_{\ell^*_j,j}\}$ for the optimal version of the basic approach, our proposal and clique-based estimates, and that of the set $T^*_{1,P_0} = \{(k,j) \in [K] \times P_0 : \beta^*_{k,j} \neq \beta^*_{1,j}\}$ for the basic approach. Here, we only consider covariates in P_0 because they are those that are the most differently accounted for by the four approaches we compare.

In the constant $\delta^*_{k,j}$ case, our empirical results clearly illustrate our theoretical ones. First, our proposal performs nearly as well as the optimal version of the basic approach. Second, the lower d_H, the better they perform, as expected since $d_H = \mathcal{D}_1$ and $\mathcal{D}_0 = 0$ here. Third, it is more difficult to recover T^*_{1,P_0} than $T^*_{P_0}$, which is also expected since $\mathcal{D}^{(1)}_1 = K - d_H > \mathcal{D}_1$. For the same reason however, as d_H increases, T^*_{1,P_0} is easier to recover. A nice symmetry appears between the performance of our proposal and the optimal version of the basic approach on the one hand and the basic approach using $\ell = (1, \ldots, 1)$, on the other hand. In addition, the recovery of the sets $T^*_{P_0}$ and T^*_{1,P_0} is only marginally affected by p. Results in the random $\delta^*_{k,j}$ case are mostly consistent with those in the constant case, except for the recovery of T^*_{1,P_0} which is mostly due to the fact that $T^*_{1,P_0} = [K-1] \times P_0$ irrespective of d_H in this random case. Finally, the clique-based strategy performs similarly to, or a little worse than, our proposal with respect to this criterion.

The bottom panels of Figure 1 present results regarding $\log(\sum_{k \in [K]} \|X^{(k)}(\beta^*_k - \widehat{\beta}_k)\|_2^2/n)$ (Dalalyan et al., 2017), which is a measure of the prediction error. Overall, our proposal and the optimal version of the basic approach perform similarly. They both outperform the basic approach, especially in the random $\delta^*_{k,j}$ case. The clique-based strategy outperforms our proposal in the constant $\delta^*_{k,j}$ case if d_H is high enough, but only when the n_k/p ratio is not too small. In the random $\delta^*_{k,j}$ case, our proposal clearly outperforms the clique-based strategy when $p = 500$, and more generally as d_H increases and/or the n_k/p ratio decreases. These results suggest that the clique-based strategy might not be able to fully account for, nor benefit from, the true structure in a high-dimensional setting. They also suggest that our proposal is better suited for this high-dimensional setting, as long as heterogeneity is not too high.

4. Application on single-cell data

We analyse data describing mielocytic leukemia cells undergoing differentiation to macrophage. Expression levels of 45 transcription factors are measured at $K = 8$ distinct time points of this differentiation process ($H0, H1, H6, H12, H24, H48, H72$ and $H96$). Each time point defines a stratum where data on $n_k = 120$ single cells are available. This data set is described in Kouno et al. (2013). In this application, the main objective is to determine how associations among the 45 transcription factors vary over time. Kouno et al. (2013) focus on marginal associations and use univariate analyses while graphical models, which describe conditional associations, might be better suited. Their inference can be reduced to the identification and description of the neighborhood of each covariate (Meinshausen Bühlmann, 2006). Here, as a first step, we study how the neighbor-

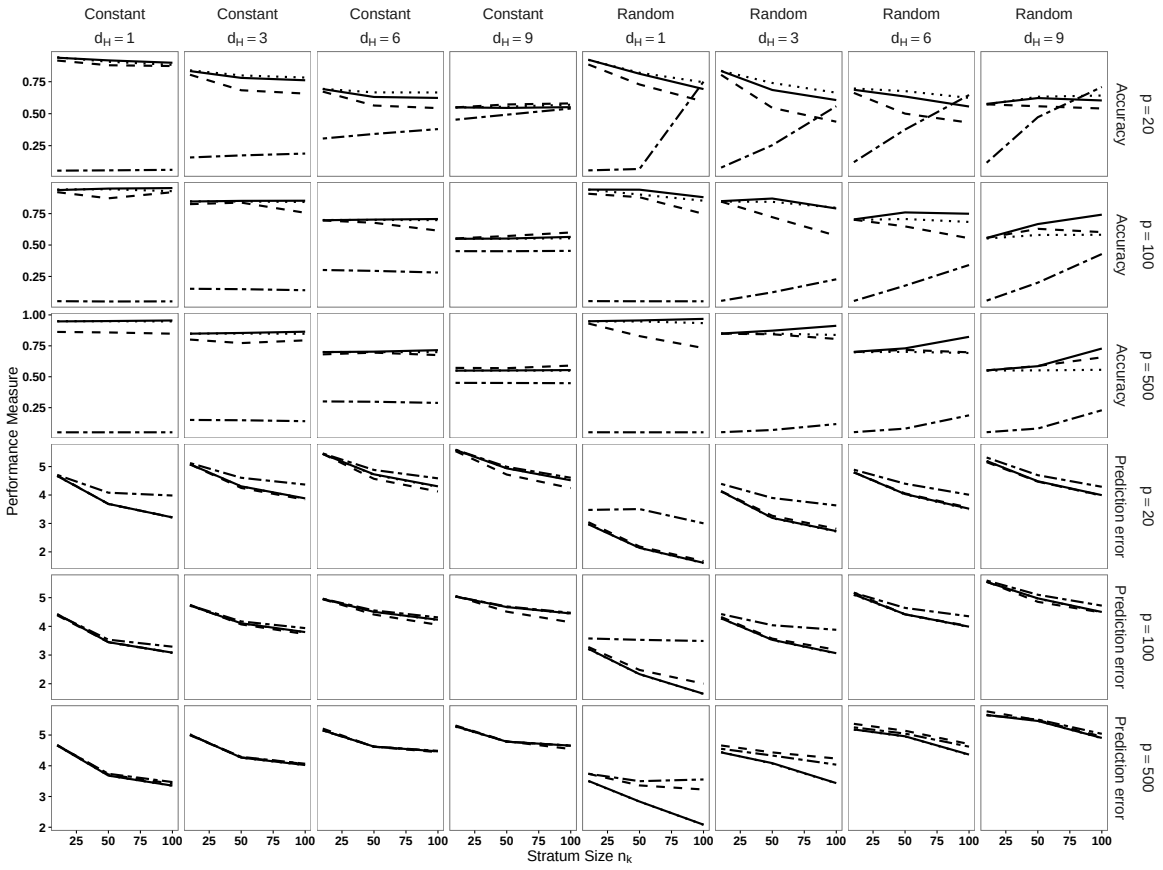

Figure 1: Results from the simulation study. The top three panels show the accuracy regarding the recovery of the set $T^*_{1,P_0} = \{(k,j) \in [K] \times P_0 : \beta^*_{k,j} \neq \beta^*_{1,j}\}$ for the basic approach and the set $T^*_{P_0} = \{(k,j) \in [K] \times P_0 : \beta^*_{k,j} \neq \beta^*_{\ell^*_j,j}\}$ for the other approaches. The higher, the better. The bottom three panels illustrate the prediction error; the lower, the better. Results are presented for both the constant and random $\delta^*_{k,j}$ cases. All results correspond to averages over 50 replicates in each configuration. Solid line: our proposal. Dotted line: optimal version of the basic approach. Dash-dot line: basic approach. Dashed line: clique-based approach.

hood of one particular transcription factor, EGR2, varies over time. Towards this end, we consider stratified linear regression models that relates EGR2 to the other $p = 44$ factors on the $K = 8$ strata. Expression levels of EGR2 are centered within each stratum, and no intercept is included in the models. Then, parameters of interest are vectors $\beta^*_1, \ldots, \beta^*_8$, where $\beta^*_k \in \mathbb{R}^{44}$ describes the association between EGR2 and the $p = 44$ transcription factors at the kth time point. We compare estimates returned by our proposal and four competitors. The basic approach is considered with two distinct choices for the reference stratum: we set it to either $H0$ or $H96$ for each covariate. We further consider the clique-based strategy. Finally, given the ordinal nature of the strata in this

particular example, the variant based on chain graphs (Gertheiss & Tutz, 2012) can be seen as the reference method. We include it as well, even if our main objective in this illustrative application is to compare the other four approaches, which do not account for this additional information. For each approach, regularization parameters are selected by 5-fold cross-validation.

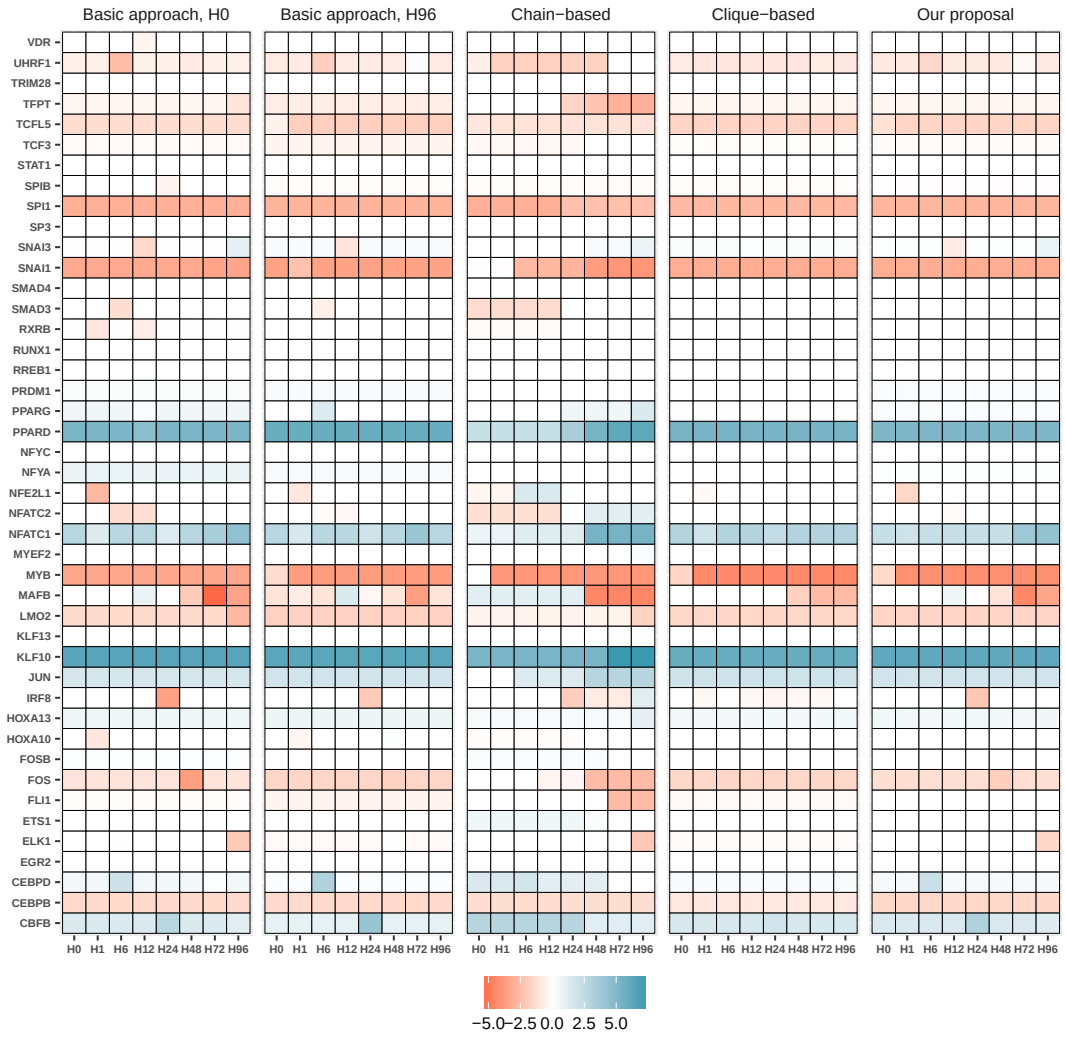

Figure 2: Estimation of the $K = 8$ parameter vectors in the linear regression models describing the association between EGR2 and the $p = 44$ other transcription factors, at times $H0, H1, H6, H12, H24, H48, H72$ et $H96$. Each column corresponds to the estimation obtained according to one of the five considered approaches : the basic approach with the reference stratum set to either $H0$ or $H96$ for every covariate, two versions of the generalized fused lasso estimates, one based on cliques and one based on chain graphs, and our proposal.

Results are presented in Figure 2. For each method, the estimates correspond to a 44×8 matrix of the form $(\widehat{\beta}_1, \ldots, \widehat{\beta}_K)$ with $K = 8$ and $\widehat{\beta}_k \in \mathbb{R}^{44}$ for any $k \in [K]$. Therefore, for any $j \in [p]$, the jth row of each matrix corresponds to estimates $(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$ of the effect of the jth covariate over the K time points, returned by the corresponding approach. This heat map representation allows an easy comparison of the pattern identified among the effects $(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$ of each covariate across the 8 strata. Our first objective is to illustrate the impact of the reference stratum when using the basic approach. Considering for instance the association between EGR2 and MYB, most approach identify the same pattern: the association is constant between $H1$ and $H96$, while it is lower, or even null, at $H0$. However, setting the reference stratum to $H0$, the basic approach does not detect any heterogeneity. These results are consistent with what is expected if the true pattern is the one identified by the other approaches: the basic approach used with $H0$ as the reference stratum is unlikely to identify the heterogeneity if it occurs at $H0$. Now consider the association with ELK1. Our proposal, the basic approach with $H0$ as the reference stratum and the strategy based on chain graphs all suggest the absence of association between EGR2 and ELK1 on the time interval $H0$ and $H72$, and a positive association at $H96$. If the reference stratum is set to $H96$, the basic approach suggests a quite different pattern, which is again expected if the true pattern is the one returned by the other approaches. These results confirm that the reference stratum is critical for the basic approach. They further suggest that our proposal is able to identify appropriate covariate-specific reference strata.

The comparison of the patterns returned by our approach and the two fused lasso strategies mostly highlights that the clique-based strategy identifies fewer heterogeneities and that strategy based on chain graphs returns smoother patterns. Again, these results were expected given the connectivity of the clique and chain graph, respectively. Prediction error was evaluated by double 5-fold cross-validation. Among the approaches that do not account for the ordering of the strata, the best prediction error is obtained with our proposal, while the worst is 1.8% higher and is obtained with the clique-based strategy. The chain graph strategy leads to an improvement of 1.8% compared to our approach.

Two main conclusions can be drawn. When data come from several strata of the population and no information is available regarding which strata are likely to share similar effects, our proposal is a competitive approach. When additional information is available, as in this particular application where strata are naturally ordered, accounting for it can be beneficial.

5. Discussion

After submitting a first version of this work, we became aware of concurrent work by Gross & Tibshirani (2016) where the authors introduce similar ideas. They apply it, in particular for the uplift problem in clinical research where the objective is to find sub-populations in a randomized trial for which an intervention is beneficial. In addition, there has been a recent line of works on penalized approaches aiming at identifying interactions, not necessarily between a categorical covariate and other predictors (Lim & Hastie , 2015; Radchenko & James, 2010). In this general context, strong hierarchy is often imposed: whenever an interaction between two variables is included in the model, the corresponding main effects are included too. However, this strong hierarchy is not desirable in our setting, where a coefficient can be nonzero in only one of the strata (Gross & Tibshirani, 2016). Moreover, when applied in our setting, these approaches can be seen as versions of the basic approach (1) based on extensions of the L_1-norm penalty. In particular, a reference stratum has to

be chosen as a first step, and the performance of these approaches would depend on this arbitrary choice. For instance, Radchenko & James (2010) establish assumptions under which their approach is sparsistent. When applied with the reference strata ℓ, their main condition on the design matrix is very similar to $(IC)_\ell$, which is generally stronger than $(IC)_{\ell^*}$ and $(IC)_0$. Then, these approaches could benefit from the ideas we developed in this article.

Our proposal is based on an overparametrization, which naturally raises the question of identifiability. We refer to Gross & Tibshirani (2016) for some discussion. We shall add that there is no identifiability issue under the conditions of Theorem 1. If these conditions do not hold, and in particular if $(IC)_{\ell^*}$ is not fulfilled, even the optimal version of the basic approach is not sparsistent, and the identifiability issue related to the overparametrization is secondary.

Prediction bounds for our proposal can be derived under the conditions presented in this work. But weaker conditions might be sufficient, following recent work studying the lasso for correlated designs (Dalalyan et al., 2017). Another extension might concern the derivation of valid p-values or confidence intervals for the nonzero parameters identified by our proposal. Given its connection with the lasso, this post-selection inference might be derived by extending recent strategies proposed for lasso estimates (Lee et al., 2016).

We also plan to extend our proposal to other regression models, which is straightforward for a variety of models given its connection with the lasso. In particular, our proposal could easily be extended to stratified Cox models used in survival analysis when competing risks arise (Rosner et al., 2013), or to the conditional logistic models used in case-controls studies (Reid & Tibshirani, 2014). The extension of clique-based estimates to other models is generally more computationally burdensome, partly because there is no proximal operator for the fused penalty.

Acknowledgement

We are grateful to Adeline Samson and Philippe Rigollet and the anonymous referees and associate editor for their fruitful comments on a preliminary version of this article.

Supplementary material

Supplementary material available at *Biometrika* online includes two sections. Section 1 presents technical details: the proof of Lemma 1 and a generalized version of Lemma 1, the version of Theorem 1 in the balanced and orthogonal design along with its proof, and two corollaries describing the particular cases where $S_{\ell^*} = \emptyset$ and $T_{\ell^*} = \emptyset$. Section 2 presents additional results from our empirical study: accuracies for the recovery of other sets of interest in the settings described heres, and additional results obtained under an alternative settings which should favor the clique-based strategy.

References

DALALYAN, A. S., HEBIRI, M., & LEDERER, J. (2017). On the prediction performance of the lasso. *Bernoulli*, **1**, 552-581.

EVGENIOU, T. & PONTIL, M. (2004). Regularized multi–task learning. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 109–17.

FRIEDMAN, J. H., HASTIE, T. J., & TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, **33**, 1–22.

GERTHEISS, J. & G. TUTZ, J. (2012). Regularization and model selection with categorial effect modifiers. *Stat Sin*, **22**, 957–82.

GROSS, S. & TIBSHIRANI, R. J. (2016). Data Shared Lasso: A novel tool to discover uplift. *Comput. Stat. Data Anal.*, **101**, 226–35.

KOUNO, T., DE HOON, M., MAR, J. C., TOMARU, Y., KAWANO, M., CARNINCI, P., SUZUKI, H., HAYASHIZAKI, Y. & SHIN, J. W. (2013). Temporal dynamics and transcriptional control using single-cell gene expression analysis. *Genome Biol.*, **14**, R118.

LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–27.

LIM, M. & HASTIE, T. J. (2015). Learning interactions via hierarchical group-lasso regularization. *J. Comp. Graph. Stat.*, **24**, 627–54.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, revised ed.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–62.

OELKER, M. R., GERTHEISS, J. & TUTZ, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Stat. Model.*, **14**, 157–77.

OLLIER, E. , SAMSON, A., DELAVENNE, X. & VIALLON, V. (2016). A SAEM algorithm for fused lasso penalized non linear mixed effect models: Application to group comparison in pharmacokinetic. *Comput. Stat. Data Anal.*, **95**, 207–21.

QIAN, J. and JIA, J. (2016). On stepwise pattern recovery of the fused Lasso. *Comput. Stat. Data Anal.*, **94**, 221–37.

RADCHENKO, P. & JAMES, T. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Am. Statist. Assoc.*, **105**, 1541–53.

REID, S. & TIBSHIRANI, R. J. (2014). Regularization paths for conditional logistic regression: The clogitl1 package. *J Stat Softw*, **58**, 1–23.

ROSNER, B., GLYNN, R., TAMIMI, R., CHEN, W., COLDITZ, G., WILLETT, W. & HANKINSON, S. (2013). Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. *J. Am Epidemiol*,**178**, 296–308.

SHARPNACK, J., RINALDO, A. & SINGH, A. (2012). Sparsistency of the edge lasso over graphs. *AISTAT*.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–88.

TIBSHIRANI, R. J., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B, **67**, 91–108.

VIALLON, V., LAMBERT-LACROIX, S., HOEFLING, H. & PICARD, F. (2016). On the robustness of the generalized fused lasso to prior specifications. *Stat Comput*, **26**, 285–301.

VODUC, K. D., CHEANG, M. C., TYLDESLEY, S., GELMON, K., NIELSEN, T.O. & KENNECKE, H. (2010). Breast cancer subtypes and the risk of local and regional relapse. *J. Clin. Oncol.*, **28**, 1684–91.

WAINWRIGHT, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, **55**, 2183–202.

ZHAO, P. & YU, B. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–63.

ZOU, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Am. Statist. Assoc.*, **101**, 141829.

Appendix: Supplementary Files

5.1 Technical details

5.1.1 PROOF OF LEMMA 1

Lemma 1 is established for matrix $\mathcal{X}_0$; the proof for $\mathcal{X}_\ell$ follows from similar arguments and is omitted.

Fix $\tau_0 > 0$ and set $\tau_k = \tau_0 K^{-1/2}$ for all $k \in [K]$. Recall that $\theta_0^* = ({\mu_{\ell^*}^*}^T, \tau_1 {\gamma_{0,1}^*}^T, \ldots, \tau_K {\gamma_{0,K}^*}^T)^T$, with $\mu_{\ell^* j}^* = \beta_{\ell_j^*, j}^*$ for $j \in [p]$, and $J_0 = \{j \in [(K+1)p] : \theta_{0j}^* \neq 0\}$. For the sake of brevity, the proof is only presented in the case where $S_{\ell^*} \neq \emptyset$ and $T_{\ell^*} \neq \emptyset$, where $S_{\ell^*} = \{j \in [p] : \mu_{\ell^* j}^* \neq 0\}$ and $T_{\ell^*} = \{(k,j) : \beta_{k,j}^* \neq \mu_{\ell^* j}^*\}$. Setting $T_k^* = \{j \in [p] : (k,j) \in T_{\ell^*}\}$ and $\Sigma_{S_{\ell^*}, T_k}^{(k)} = {X_{S_{\ell^*}}^{(k)}}^T X_{T_k}^{(k)}$ we have

$$\mathcal{X}_{0J_0} = \begin{pmatrix} X_{S_{\ell^*}}^{(1)} & \frac{X_{T_1^*}^{(1)}}{\tau_1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ X_{S_{\ell^*}}^{(K)} & 0 & \ldots & \frac{X_{T_k^*}^{(K)}}{\tau_K} \end{pmatrix}, (\mathcal{X}_{0J_0}^T \mathcal{X}_{0J_0}) = \begin{pmatrix} nI_{|S_{\ell^*}|} & \frac{\Sigma_{S_{\ell^*}, T_1^*}^{(1)}}{\tau_1} & \ldots & \ldots & \ldots & \frac{\Sigma_{S_{\ell^*}, T_K^*}^{(K)}}{\tau_K} \\ \frac{{\Sigma_{S_{\ell^*}, T_1^*}^{(1)}}^T}{\tau_1} & \frac{nI_{|T_1^*|}}{\tau_0^2} & 0 & \ldots & \ldots & 0 \\ \frac{{\Sigma_{S_{\ell^*}, T_2^*}^{(2)}}^T}{\tau_2} & 0 & \frac{nI_{|T_2^*|}}{\tau_0^2} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ \frac{{\Sigma_{S_{\ell^*}, T_K^*}^{(K)}}^T}{\tau_K} & 0 & \ldots & \ldots & 0 & \frac{nI_{|T_K^*|}}{\tau_0^2} \end{pmatrix}.$$

For any $s \in [|S_{\ell^*}|]$ and $t \in [|T_k^*|]$, denote the sth element of S_{ℓ^*} by $S_{\ell^* s}$ and the tth element of T_k^* by $T_{k,t}^*$. For all $j \in S_{\ell^*}$, further denote by $N_j^* = n|K_j^*|/K$ the number of observations in strata contained in $K_j^* = \{k \in [K] : \beta_{k,j}^* = \beta_{\ell_j^*, j}^*\}$. Now introduce C, the matrix of size $|T_{\ell^*}| \times |S_{\ell^*}|$ made of K blocks C_k. Each block is of size $|T_k^*| \times |S_{\ell^*}|$, and the element (t,s) of C_k is $(C_k)_{(t,s)} = -\tau_k / N_{S_{\ell^* s}}^*$ if $S_{\ell^* s} = T_{k,t}^*$ and 0 otherwise ($s \in [|S_{\ell^*}|]$ and $t \in [|T_k^*|]$).

For $k, \ell \in [K]$ and $k \neq \ell$, further introduce B_{k_1,k_2}, the matrix of size $|T^*_{k_1}| \times |T^*_{k_2}|$ with element (t_1, t_2) equal to $\tau_{k_1}\tau_{k_2}/N^*_{S_{\ell^* s}}$ if $T^*_{k_1,t_1} = T^*_{k_2,t_2} = S_{\ell^* s}$ for some $s \in [|S_{\ell^*}|]$, and 0 otherwise. For $k \in [K]$, denote by $B_{k,k}$ the diagonal matrix of size $|T^*_k| \times |T^*_k|$ with tth diagonal term equal to $\tau_k^2(N^*_{S_{\ell^* s}} + n_k)/(N_{\ell^* s} n_k)$ if $T^*_{k,t} = S_{\ell^* s}$ for some $s \in [|S^*|]$ and τ_k^2/n_k otherwise. Finally denote by D the diagonal matrix of size $|S_{\ell^*}| \times |S_{\ell^*}|$ with jth diagonal term equal to $1/N^*_j$, and by B the matrix of size $|T_{\ell^*}| \times |T_{\ell^*}|$ made of K^2 blocks, with block (k_1, k_2) equal to B_{k_1,k_2}. By standard algebra, we have

$$(\mathcal{X}^T_{0J_0}\mathcal{X}_{0J_0})^{-1} = \begin{pmatrix} D & C^T \\ C & B \end{pmatrix}.$$

Now, for any $j \notin J_0$, the jth column $\mathcal{X}_{0j}$ of $\mathcal{X}_0$ is of the form either (A) or (B):

(A) $\mathcal{X}_{0j} = (0^T, \ldots, 0^T, {X^{(k_0)}_{j_0}}^T, 0^T, \ldots, 0^T)^T$ for some $k_0 \in [K]$ and some $j_0 \notin T^*_{k_0}$,

(B) $\mathcal{X}_{0j} = ({X^{(1)}_{j_0}}^T, \ldots, {X^{(K)}_{j_0}}^T)^T$ for some $j_0 \notin S_{\ell^*}$.

If $\mathcal{X}_{0j}$ is of form (A), then ${\mathcal{X}_{0j}}^T\mathcal{X}_{0J_0} = (\bar{d}^T, 0^T_{|T^*_1|}, \ldots, 0^T_{|T^*_K|}) \in \mathbb{R}^{|J_0|}$ with $\bar{d} \in \mathbb{R}^{|S_{\ell^*}|}$ and

$$\bar{d}_s = \begin{cases} n_{k_0}/\tau_{k_0} & \text{if } j_0 = S_{\ell^* s}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, if $\mathcal{X}_{0j}$ is of form (A), we have

$$\|{\mathcal{X}_{0j}}^T\mathcal{X}_{0J_0}({\mathcal{X}_{0J_0}}^T\mathcal{X}_{0J_0})^{-1}\|_1 \leq \max_{j \in S_{\ell^*}} \max_{k \in K^*_j} \frac{n_k}{\tau_k N^*_j}(1 + \sum_{\ell \notin K^*_j} \tau_\ell).$$

If $\mathcal{X}_{0j}$ is of form (B), then $\mathcal{X}^T_{0j}\mathcal{X}_{0J_0} = (0_{|S_{\ell^*}|}, d_1, \ldots, d_K)^T \in \mathbb{R}^{|J_0|}$, with $d_k \in \mathbb{R}^{|T^*_k|}$ and

$$d_{k,t} = \begin{cases} n_k/\tau_k & \text{if } j_0 = T^*_{k,t}, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, we have

$$\|{\mathcal{X}_{0j}}^T\mathcal{X}_{0J_0}({\mathcal{X}_{0J_0}}^T\mathcal{X}_{0J_0})^{-1}\|_1 \leq \max_{j \notin S_{\ell^*}} \sum_{\ell \notin K^*_j} \tau_\ell.$$

Moreover, under the setting considered in Lemma 1 we have $n_k = n/K$ for all $k \in [K]$ and $\tau_k = \tau_0 K^{-1/2}$. Therefore, $\max_{j \notin J_0} \|{\mathcal{X}_{0j}}^T\mathcal{X}_{0J_0}({\mathcal{X}_{0J_0}}^T\mathcal{X}_{0J_0})^{-1}\|_1 < 1$ if and only if assumption $(sIC)_0$ holds, which completes the proof of Lemma 1 for matrix $\mathcal{X}_0$.

5.1.2 Version of Theorem 1 in the setting of orthogonal designs and balanced strata

Theorem 3 below is the version of Theorem 1 in the setting considered in Lemma 1, where $n_k = n/K$ and ${X^{(k)}}^T X^{(k)})/n_k = I_{n_k}$ for all $k \in [K]$. We set $\mathcal{D}_0 = \mathcal{D}_{\ell^*,0}$ and $\mathcal{D}_1 = \mathcal{D}_{\ell^*,1}$; see Section 2.6 in the main text for the corresponding definitions.

Theorem 3 *For all $k \in [K]$, assume that the noise variables $\varepsilon_i^{(k)}$, $i \in [n_k]$, are independent and identically distributed centered sub-Gaussian variables with parameter $\sigma > 0$. Further assuming that $(sIC)_0$ holds, define*

$$\gamma = \min\left(1 - \mathcal{D}_0\tau_0 K^{-1/2}, 1 - \frac{K^{1/2} + \mathcal{D}_1\tau_0}{(K - \mathcal{D}_1)\tau_0}\right)$$

and

$$C_{\min} = \min\left(1, \tau_0^{-2}, \frac{1}{2}\left[\left(\tau_0^{-2} + 1\right) - \left\{\left(\tau_0^{-2} - 1\right)^2 + \frac{4\mathcal{D}_1}{\tau_0^2 K}\right\}\right]\right).$$

For $\eta \in \{0, 1\}$, we set

$$\lambda_1^{(\eta)} > \frac{2}{\gamma \min(1, \tau_0)}\left[\frac{2\sigma^2 \log((K + \eta)p)}{n}\right]^{1/2}, \quad \lambda_{2,k}^{(\eta)} = \tau_k \lambda_1^{(\eta)}.$$

Finally introduce $\beta_{\min}^{(\eta)} = \lambda_1^{(\eta)}[(|S_{\ell^*}| + |T_{\ell^*}|)^{1/2} C_{\min}^{-1} + 4\sigma C_{\min}^{-1/2}]$, and consider the following β-min conditions:

$$(C_{\beta_{\min}^{(\eta)}})(i) : \forall j \in S_{\ell^*},\ |\bar{\beta}_j^*| > \beta_{\min}^{(\eta)}; \quad (C_{\beta_{\min}^{(\eta)}})(ii) : \forall j \in [p], \forall k \notin K_j^*,\ |\beta_{k,j}^* - \bar{\beta}_j^*| > \frac{K^{1/2}\beta_{\min}^{(\eta)}}{\tau_0}.$$

Then, S_{ℓ^*} and T_{ℓ^*} are both recovered

- *with probability superior to $1 - 4\exp(-c_1 n \lambda_1^{(0)^2})$, for some $c_1 > 0$, by the optimal version of the basic approach run with $\lambda_1 = \lambda_1^{(0)}$ and $\lambda_{2,k} = \lambda_{2,k}^{(0)}$ under $(C_{\beta_{\min}^{(0)}})(i, ii)$ and we have $\|\widehat{\theta}_{\ell^* J_{\ell^*}} - \theta^*_{J_{\ell^*}}\|_\infty \leq \beta_{\min}^{(0)}$;*
- *with probability superior to $1 - 4\exp(-c_1 n \lambda_1^{(1)^2})$, for some $c_1 > 0$, by our approach run with $\lambda_1 = \lambda_1^{(1)}$ and $\lambda_{2,k} = \lambda_{2,k}^{(1)}$ under $(C_{\beta_{\min}^{(1)}})(i, ii)$ and we have $\|\widehat{\theta}_{0J_0} - \theta^*_{0J_0}\|_\infty \leq \beta_{\min}^{(1)}$.*

Consider the asymptotic setting with K, and possibly p, tending to infinity as $n \to \infty$. Further assume that $\mathcal{D}_{\ell^*,0} = \mathcal{D}_{\ell^*,1} = \mathcal{D}_{\ell^*}$. If $\mathcal{D}_{\ell^*} \ll K^{1/2}$ or $\mathcal{D}_{\ell^*} = cK^{1/2}$ for some $0 < c \leq 1/2$, then $\tau_0 = 1$ ensures perfect recovery for signals such that $\beta_{\min}^{(\eta)} = \mathcal{O}(n^{-1/2}[(|S_{\ell^*}| + |T_{\ell^*}|)\log((K + 1)p)]^{1/2})$, which is optimal up to log-terms. If $\mathcal{D}_{\ell^*} = cK^{1/2}$ for some $c > 1/2$, we get the same order of magnitude for $\beta_{\min}^{(\eta)}$, but with $\tau_0 = (2c)^{-1} < 1$. If $K^{1/2} \ll \mathcal{D}_{\ell^*}$, then the regime changes. For $K^{1/2} \ll \mathcal{D}_{\ell^*} \ll K$ the optimal choice is $\tau_0 = K^{1/2}/(2\mathcal{D}_{\ell^*})$ which only ensures perfect recovery for $\beta_{\min}^{(\eta)} = \mathcal{O}((nK)^{-1/2}\mathcal{D}_{\ell^*}[(|S_{\ell^*}| + |T_{\ell^*}|)\log((K + 1)p)]^{1/2})$. Finally, if $\mathcal{D}_{\ell^*} = cK$ for some $0 < c < 1/3$, then the result of Theorem S1 becomes almost meaningless: the optimal τ_0 is $\mathcal{O}(K^{-1/2})$ which only ensures perfect recovery for $\beta_{\min}^{(\eta)} = \mathcal{O}(n^{-1/2}[K(|S_{\ell^*}| + |T_{\ell^*}|)\log((K + 1)p)]^{1/2})$.

5.1.3 Proof of Theorem 3

In view of Lemma 1, and because $\mathcal{X}_{\ell^* J_{\ell^*}} = \mathcal{X}_{0J_0}$, we simply have to compute $\Lambda_{\min}((\mathcal{X}_{0J_0}^T \mathcal{X}_{0J_0}/n))$ in order to apply Theorem 1 of Wainwright (2009) and establish Theorem 3. To do so, we look for

solutions of the characteristic polynomial of the matrix $(\mathcal{X}_{0J_0}^T\mathcal{X}_{0J_0}/n)$, $p(\lambda) = \det(\mathcal{X}_{0J_0}^T\mathcal{X}_{0J_0}/n - \lambda I_{|J_0|})$. Using the block structure of matrix $(\mathcal{X}_{0J_0}^T\mathcal{X}_{0J_0}/n - \lambda I_{|J_0|})$, we get the following expression

$$p(\lambda) = (1-\lambda)^{r_1}(\tau^{-2}-\lambda)^{r_2} \prod_{j \in S^* \cap \{\cup_k T_k^*\}} \left(\lambda^2 - \lambda(\tau^{-2}+1) + \tau^{-2}\Big(1 - \frac{n - N_j^*}{n}\Big)\right),$$

with $r_1 = |S_{\ell^*} \backslash \{\cup_k T_k^*\}|$ and $r_2 = |J_0| - |S_{\ell^*}| - |S_{\ell^*} \cap \{\cup_k T_k^*\}|$. It follows that $\Lambda_{\min}((\mathcal{X}_{0J_0}^T\mathcal{X}_{0J_0}/n)) \geq \min\left(1, \tau^{-2}, \frac{1}{2}\left[(\tau^{-2}+1) - \{(\tau^{-2}-1)^2 + 4\mathcal{D}_1\tau^{-2}/K\}^{1/2}\right]\right)$. Denote by $|||M|||_\infty$ the maximum row sum matrix norm of matrix M, and by $|||M|||_2$ its spectral norm. Because $|||(\mathcal{X}_{0J_0}^T\mathcal{X}_{0J_0}/n)|||_\infty \leq |J_0|^{1/2}|||(\mathcal{X}_{0J_0}^T\mathcal{X}_{0J_0}/n)|||_2 \leq (|S_{\ell^*}| + |T_{\ell^*}|)^{1/2}/C_{\min}$, Theorem 3 now follows from Theorem 1 of Wainwright (2009).

5.1.4 Generalization of Lemma 1

Here, we do not consider the orthogonal and balanced setting anymore and present general conditions ensuring that the irrepresentability conditions $(IC)_\ell$ and $(IC)_0$ are fulfilled by the design matrices $\mathcal{X}_\ell$ and $\mathcal{X}_0$ involved in the basic approach and our proposal, respectively. For all $k \in [K]$, we assume that $\tau_k = \tau_0(n_k/n)^{1/2}$ for some $\tau_0 > 0$ and that $n_k^{-1/2}\|X_j^{(k)}\|_2 \leq 1$ for all $(j) \in [p]$. For $\mathcal{X}$ equal to either $\mathcal{X}_\ell$ or $\mathcal{X}_0$, this ensures that $n^{-1}\|\mathcal{X}_j\|_2 \leq \max(1, \tau^{-1})$, for each column $\mathcal{X}_j$ of $\mathcal{X}$.

For any given vector of reference strata $\ell \in [K]^p$ and any $j \in [p]$, set $\bar{K}_{\ell,j} = \{k \in [K] : \beta_{k,j}^* = \beta_{\ell_j,j}^*\}$ and $K_{\ell,j} = \bar{K}_{\ell,j} \setminus \{\ell_j\}$. Further set, for any $k \in [K]$, $T_{\ell,k} = \{j \in [p] : \beta_{k,j}^* \neq \beta_{\ell_j,j}^*\}$, $\Sigma_{\ell,k} = X_{T_{\ell,k}}^{(k)T}X_{T_{\ell,k}}^{(k)}$, $\Pi_{\ell,k} = X_{T_{\ell,k}}^{(k)}\Sigma_{\ell,k}^{-1}X_{T_{\ell,k}}^{(k)T}$, and $Z_{\ell,j}^{(k)} = (I_{n_k} - \Pi_{\ell,k})X_j^{(k)}$. Define $\omega_{\ell,j}^{(k)} = \Sigma_{\ell,k}^{-1}X_{T_{\ell,k}}^{(k)^T}X_j^{(k)}$ and $\Omega_\ell^{(k)} = \Sigma_{\ell,k}^{-1}X_{T_{\ell,k}}^{(k)^T}X_{S_\ell}^{(k)}$. Introduce the quantities $\widetilde{\Sigma}_\ell = \sum_{k\in[K]} X_{S_\ell}^{(k)T}(I_{n_k} - \Pi_{\ell,k})X_{S_\ell}^{(k)}$ and $\widetilde{\Omega}_{\ell,j}^{(k)} = \widetilde{\Sigma}_\ell^{-1}X_{S_\ell}^{(k)T}Z_{\ell,j}^{(k)}$. Finally set

$$\begin{aligned}
c_1(\ell) &= \max_{j \in S_\ell^c}\left\{\|\sum_{k\in[K]}\widetilde{\Omega}_{\ell,j}^{(k)}\|_1 + \sum_{k\in[K]}\tau_k\|\sum_{l\in[K]}\Omega_\ell^{(l)}\widetilde{\Omega}_{\ell,j}^{(k)}\|_1\right\} \\
c_2(\ell) &= \max_{j\in[p]}\max_{k\in K_{\ell,j}}\left\{\frac{\|\widetilde{\Omega}_{\ell,j}^{(k)}\|_1}{\tau_k} + \sum_{l\neq k}\frac{\tau_l}{\tau_k}\|\Omega_\ell^{(l)}\widetilde{\Omega}_{\ell,j}^{(k)}\|_1 + \|\omega_{\ell,j}^{(k)} + \Omega_\ell^{(k)}\widetilde{\Omega}_{\ell,j}^{(k)}\|_1\right\} \\
\bar{c}_2(\ell) &= \max_{j\in[p]}\max_{k\in\bar{K}_{\ell j}}\left\{\frac{\|\widetilde{\Omega}_{\ell,j}^{(k)}\|_1}{\tau_k} + \sum_{l\neq k}\frac{\tau_l}{\tau_k}\|\Omega_\ell^{(l)}\widetilde{\Omega}_{\ell,j}^{(k)}\|_1 + \|\omega_{\ell,j}^{(k)}\Omega_\ell^{(k)} + \widetilde{\Omega}_{\ell,j}^{(k)}\|_1\right\}.
\end{aligned}$$

Lemma 4 *Let $\ell \in [K]^p$ be a given vector of reference strata. Assume that $\Lambda_{\min}(\Sigma_{\ell,k}) > 0$ for $k \in [K]$ and $\Lambda_{\min}(\widetilde{\Sigma}_\ell) > 0$. Condition (IC_ℓ) holds if and only if $c_1(\ell) < 1$ and $c_2(\ell) < 1$.*

Assume that $\Lambda_{\min}(\Sigma_{\ell^*,k}) > 0$ for $k \in [K]$ and $\Lambda_{\min}(\widetilde{\Sigma}_{\ell^*}) > 0$. Condition $(IC)_0$ holds if and only if $c_1(\ell^*) < 1$ and $\bar{c}_2(\ell^*) < 1$.

The proof of Lemma 4 follows from the same arguments as those presented in Section 5.1.1 for the proof of Lemma 1, and is omitted. Again, the conditions ensuring that $(IC)_{\ell^*}$ and $(IC)_0$ hold are very similar. This shows that our proposal is able to mimic the optimal version of the basic approach even when the designs are not orthogonal or strata are not balanced.

5.1.5 OTHER RESULTS

Corollary 5 and Corollary 6 consider the two special cases where $T_{\ell^*} = \emptyset$ and $S_{\ell^*} = \emptyset$, respectively.

Corollary 5 *Assume that the noise variables $(\varepsilon_i^{(k)})_{i\in[n_k],k\in[K]}$ are independent and identically centered sub-Gaussian variables with parameter $\sigma > 0$. Define $\mathbb{X} = (X^{(1)^T}, \ldots, X^{(K)^T})^T$, the $n \times p$ matrix with all the strata pooled together. Set $\tau_k = \tau_0(n_k/n)^{1/2} > 0$ for all $k \in [K]$, for some $\tau_0 > 0$. For all $j \in [p]$, assume that there exists some $\beta_j^* \in \mathbb{R}$ such that $\beta_{k,j}^* = \beta_j^*$ for all $k \in [K]$ and set $S_{\ell^*} = \{j \in [p] : \beta_j^* \neq 0\}$. Further assume that $\Lambda_{\min}(\mathbb{X}_{S_{\ell^*}}^T \mathbb{X}_{S_{\ell^*}}/n) \geq C_{\min}$ for some $C_{\min} > 0$, and that $n_k^{-1/2}\|X_j^{(k)}\|_2 \leq 1$ for all $(k, j) \in [K] \times [p]$. Finally assume that the three following conditions hold:*

$$
\begin{aligned}
(\widetilde{\mathrm{A}}) \qquad & \sum_{k\in[K]} \tau_k > 1, \\
(\widetilde{\mathrm{C}}.i.1) \qquad & \tilde{c}_1 := \max_{j\notin S^*} \|(\mathbb{X}_{S_{\ell^*}}^T \mathbb{X}_{S_{\ell^*}})^{-1} \mathbb{X}_{S_{\ell^*}}^T \mathbb{X}_j\|_1 < 1, \\
(\widetilde{\mathrm{C}}.ii) \qquad & \tilde{c}_2 := \max_{j\in[p]} \max_{k\in[K]} \tau_k^{-1} \|(\mathbb{X}_{S_{\ell^*}}^T \mathbb{X}_{S_{\ell^*}})^{-1} X_{S_{\ell^*}}^{(k)^T} X_j^{(k)}\|_1 < 1.
\end{aligned}
$$

Now, set $\tilde{\gamma} = (1 - \tilde{c}_1) \wedge (1 - \tilde{c}_2)$ and

$$
\lambda_1 = \frac{2}{(1 \wedge \tau_0)\tilde{\gamma}} \left\{ \frac{2\sigma^2 \log((K+1)p)}{n} \right\}^{1/2}, \quad \beta_{\min} = \lambda_1 \left(\frac{|S_{\ell^*}|^{1/2}}{C_{\min}} + 4\sigma C_{\min}^{-1/2} \right).
$$

Then our proposal run with parameters λ_1 and $\lambda_{2,k} = \tau_k \lambda_1$ identifies S_{ℓ^*} and $T_{\ell^*} = \emptyset$ with probability at least $1 - 4\exp(-c_1 n \lambda_1^2)$ for some $c_1 > 0$, as long as $\min_{j\in S^*} |\beta_j^*| > \beta_{\min}$.

Condition $(\widetilde{\mathrm{C}}.i.1)$ is exactly the irrepresentability condition on matrix $\mathbb{X}$, while conditions $(\widetilde{\mathrm{C}}.ii)$ and $(\widetilde{\mathrm{A}})$, which are very similar, both simply require that τ_0 is high enough. Moreover, $\gamma = 1 - \tilde{c}_1$ and $1 \wedge \tau_0 = 1$ for τ high enough. Therefore, our proposal mimics the lasso run on $(\mathbb{X}, \mathscr{Y})$ provided τ_0 high enough and is optimal, up to log-terms, when the β_k^*'s are all equal.

Corollary 6 *Assume that the noise variables $(\varepsilon_i^{(k)})_{i\in[n_k],k\in[K]}$ are independent and identically centered sub-Gaussian variables with parameter $\sigma > 0$. Set $\tau_k = \tau_0(n_k/n)^{1/2} > 0$ for all $k \in [K]$, for some $\tau_0 > 0$. For all $j \in [p]$, set $K_j^* = \{k \in [K] : \beta_{k,j}^* = 0\}$ and for all $k \in [K]$, set $T_k^* = \{j \in [p] : \beta_{k,j}^* \neq 0\}$. Assume that $\min_k\{\Lambda_{\min}(X_{T_k^*}^{(k)^T} X_{T_k^*}^{(k)}/n_k)\} \geq C_{\min}$ for some $C_{\min} > 0$, and that $n_k^{-1/2}\|X_j^{(k)}\|_2 \leq 1$ for all $(k, j) \in [K] \times [p]$. Further assume that the three following conditions hold:*

$$
\begin{aligned}
(\bar{\mathrm{A}}) \qquad & \forall j \in [p], \sum_{\ell\notin K_j^*} \tau_\ell < 1 + \sum_{k\in K_j^*} \tau_k, \\
(\bar{\mathrm{C}}.i.1) \qquad & \bar{c}_1 := \max_{j\in[p]} \max_{k\in[K]} \sum_{k\in[K]} \tau_k \|(X_{T_k^*}^{(k)^T} X_{T_k^*}^{(k)})^{-1} X_{T_k^*}^{(k)^T} X_j^{(k)}\|_1 < 1, \\
(\bar{\mathrm{C}}.ii) \qquad & \bar{c}_2 := \max_{k\in[K]} \max_{j\notin T_k^*} \|(X_{T_k^*}^{(k)^T} X_{T_k^*}^{(k)})^{-1} X_{T_k^*}^{(k)^T} X_j^{(k)}\|_1 < 1.
\end{aligned}
$$

Now, set $\bar{\gamma} = (1 - \bar{c}_1) \wedge (1 - \bar{c}_2)$*, and*

$$\lambda_1 = \frac{2}{(1 \wedge \tau_0)\bar{\gamma}} \left\{ \frac{2\sigma^2 \log((K+1)p)}{n} \right\}^{1/2}, \quad \beta_{\min} = \lambda_1 \left(\frac{\tau_0 |T_{\ell^*}|^{1/2}}{C_{\min}} + 4\sigma C_{\min}^{-1/2} \right).$$

Then our proposal run with parameters λ_1 *and* $\lambda_{2,k} = \tau_k \lambda_1$ *recovers* $S_{\ell^*} = \emptyset$ *and* $T_{\ell^*} = \{(k,j) : \beta^*_{k,j} \neq 0\}$ *with probability at least* $1 - 4\exp(-c_1 n \lambda_1^2)$*, for some* $c_1 > 0$*, as long as* $\min_{(k,j) \in T^*} |\beta^*_{k,j}| > \beta_{\min}(n/n_k)^{1/2}$.

Condition $(\bar{\mathrm{C}}.ii)$ is exactly the union of the irrepresentability conditions for each matrix $X^{(k)}$, $k \in [K]$, while conditions $(\bar{\mathrm{C}}.i.1)$ and $\bar{\mathrm{A}}$ both simply require that τ_0 is small enough. For τ_0 small enough, our proposal then mimics the strategy consisting in performing K lasso on the data $(X^{(k)}, y^{(k)})$, $k = 1, \ldots, K$, independently, with a common λ_1 value for each lasso. It is optimal, up to log-terms, when $S_{\ell^*} = \emptyset$.

5.2 Additional empirical results

5.2.1 Under the designs considered in the main text

Figure 3 presents additional results regarding the recovery of the set $S^*_{1,P_0} = \{j \in P_0 : \beta^*_{1,j} \neq 0\}$ for the basic approach run with reference strata $\ell = (1, \ldots, 1)$ and the recovery of the set $S^*_{P_0} = \{j \in P_0 : \beta^*_{\ell^*_j, j} \neq 0\}$ for the other approaches. Overall, our proposal and the optimal version of the basic approach perform similarly according to this criterion too. In the constant $\delta^*_{k,j}$ case, all methods perform similarly, and their performance does not depend on either p or d_H. In the random $\delta^*_{k,j}$ case, the performance of each method decreases as p and/or d_H increases. This discrepancy with the results obtained in the constant case illustrates that it is harder for the optimal version of the basic approach, our proposal and the clique-based strategy to determine whether the overall effect $\beta^*_{\ell^*_j, j}$ of any covariate j is null when the collection of values $(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$ varies around zero; keep in mind that in the random case, we have either $\beta^*_{k,j} = 1$ or $\beta^*_{k,j} = 1 \pm \delta^*_{k,j}$ with $\delta^*_{k,j} \sim \mathcal{U}_{[K^{1/2}/2, 2K^{1/2}]}$ so that $\beta^*_{k,j}$ can be negative. Interestingly, it is harder for the basic approach to determine whether $\beta^*_{1,j}$ is null in this situation too. In addition, the basic approach is generally outperformed by the other approaches in this random $\delta^*_{k,j}$ case. The clique-based strategy performs well for $d_H \geq 6$, especially when n_k/p is not too small. But it is outperformed by our proposal, and the optimal version of the basic approach, for $d_H = 3$ and $p = 500$.

5.2.2 Under an alternative scenario

Here, we present additional empirical results obtained under a scenario which should favor the clique-based strategy. We still consider the case where $K = 20$ and $P_0 \subset [p]$, with $|P_0| = 20$ but, for each $j \in [P_0]$, we set $\beta_{1,j} = \cdots = \beta_{5,j} = -a$, $\beta_{6,j} = \cdots = \beta_{10,j} = 0$, $\beta_{11,j} = \cdots = \beta_{15,j} = a$ and $\beta_{16,j} = \cdots = \beta_{20,j} = 2a$, for some $a > 0$. In other words, for each $j \in [P_0]$, the effects of the jth covariate across the 20 strata are made of 4 groups of distinct values, which should favor the clique-based strategy. Two values of a were considered, $a = K^{1/2}$ and $a = K^{1/2}/3$. Because results were very similar, only those obtained for $a = K^{1/2}/3$ are presented here. Figure 4 presents the predictive performance of each approach. We especially observe that our proposal performs nearly as well as the clique-based strategy in general, and outperforms it for $p = 500$. It also slightly outperforms the two versions of the basic approach.

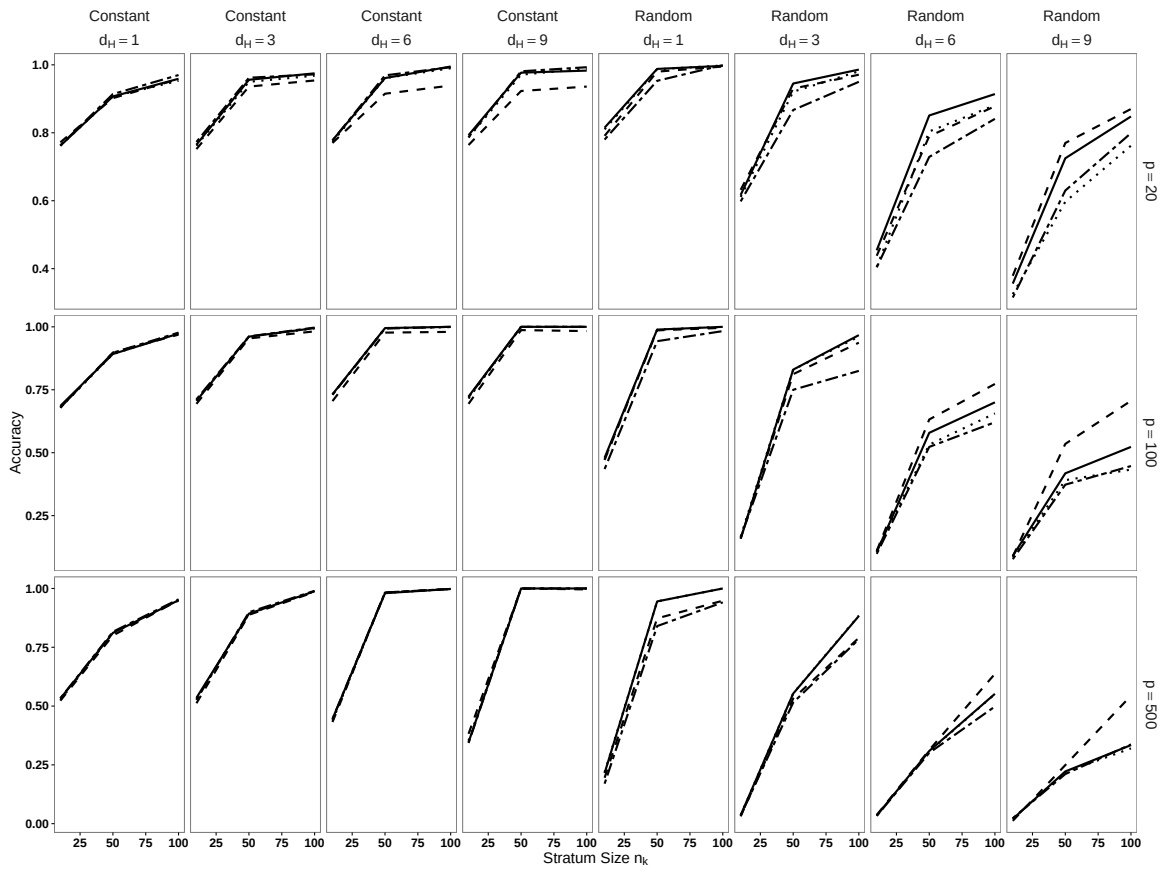

Figure 3: Accuracy regarding the recovery the set $S^*_{1,P_0} = \{j \in P_0 : \beta^*_{1,j} \neq 0\}$ for the basic approach and of the set $S^*_{P_0} = \{j \in P_0 : \beta^*_{\ell^*_j,j} \neq 0\}$ for the three other approaches. *(Left)*: Constant $\delta^*_{k,j}$ case. *(Right)*: Random $\delta^*_{k,j}$ case. Results correspond to averages over 50 replicates in each configuration. Solid line: our proposal. Dotted line: optimal version of the basic approach. Dash-dot line: basic approach. Dashed line: clique-based approach.

Figure 5 presents the estimates returned by each approach for one particular simulation in the configuration $n_k = 100$ and $p = 100$. It can especially be seen that the clique-based strategy does not fuse coefficients sensibly better than the other approaches.

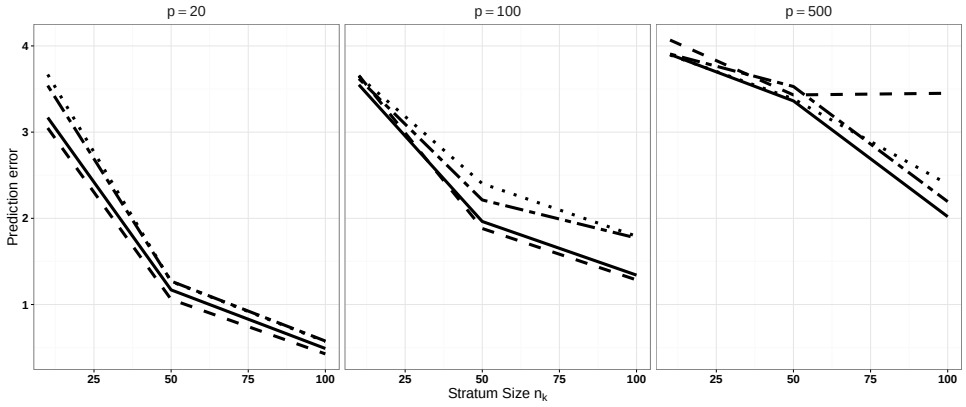

Figure 4: Prediction error (the lower, the better). Results correspond to averages over 50 replicates in each configuration. Solid line: our proposal. Dotted line: basic approach with $\ell = (20, \ldots, 20)$. Dash-dot line: basic approach with $\ell = (1, \ldots, 1)$. Dashed line: clique-based approach.

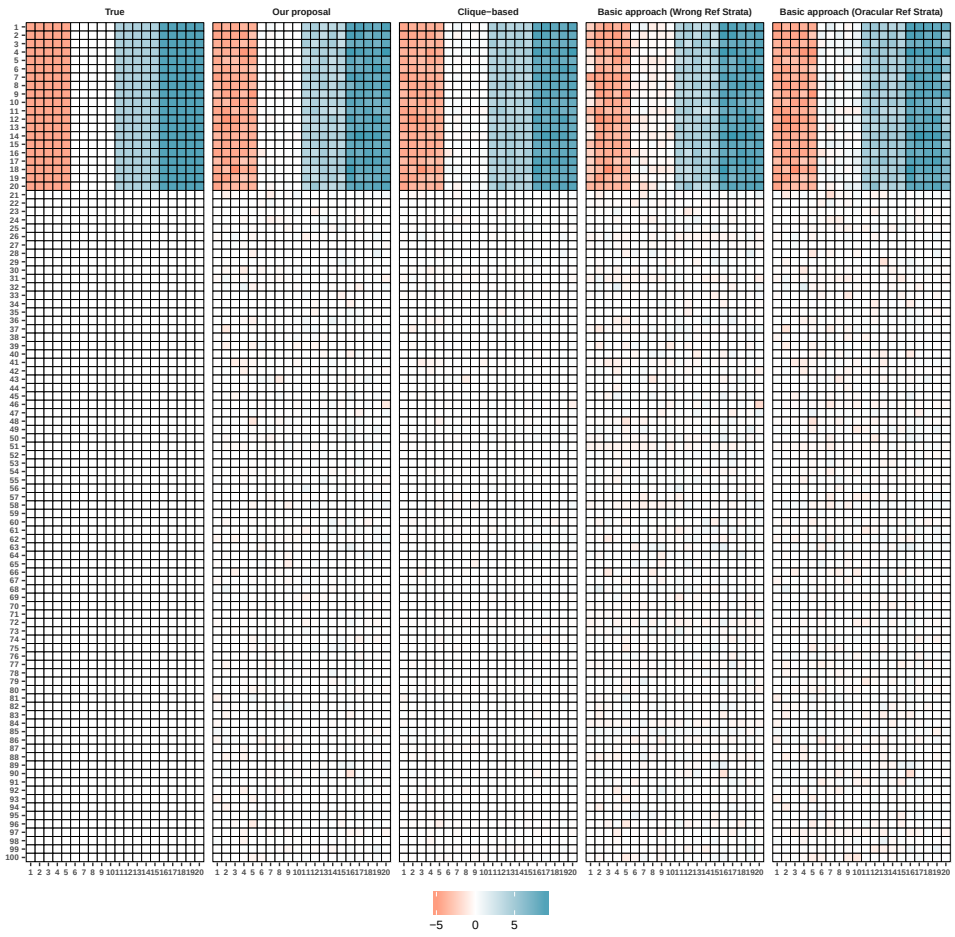

Figure 5: Estimation of the $K = 20$ parameter vectors in one particular simulation with $n_k = 100$ and $p = 100$. The first column presents the true values. Each of the four remaining columns presents the estimates obtained according to one of the four considered approaches: our proposal, the clique-based approach and the basic approach with the reference stratum set to either 1 or 20 for every covariate.