MOX-Report No. 21/2022

A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation

Cappozzo, A.; Ieva, F.; Fiorito, G.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation

Andrea Cappozzo * Francesca Ieva* Giovanni Fiorito †

Abstract

Recent evidence highlights the usefulness of DNA methylation (DNAm) biomarkers as surrogates for exposure to risk factors for non-communicable diseases in epidemiological studies and randomized trials. DNAm variability has been demonstrated to be tightly related to lifestyle behavior and exposure to environmental risk factors, ultimately providing an unbiased proxy of an individual state of health. At present, the creation of DNAm surrogates relies on univariate penalized regression models, with elastic-net regularizer being the gold standard when accomplishing the task. Nonetheless, more advanced modeling procedures are required in the presence of multivariate outcomes with a structured dependence pattern among the study samples. In this work we propose a general framework for mixed-effects multitask learning in presence of high-dimensional predictors to develop a multivariate DNAm biomarker from a multi-center study. A penalized estimation scheme based on an expectation-maximization (EM) algorithm is devised, in which any penalty criteria for fixed-effects models can be conveniently incorporated in the fitting process. We apply the proposed methodology to create novel DNAm surrogate biomarkers for multiple correlated risk factors for cardiovascular diseases and comorbidities. We show that the proposed approach, modeling multiple outcomes together, outperforms state-of-the-art alternatives, both in predictive power and bio-molecular interpretation of the results.

1 Introduction

DNA methylation (DNAm) is an epigenetic process that regulates gene expression, typically occurring in cytosine within CpG sites (CpGs) in the DNA sequence (Singal and Ginder, 1999). The development of surrogate scores based on blood DNA methylation has received thriving attention in recent years: impressive epidemiological evidence has been established between DNAm and long-term exposure to lifestyle and environmental risk factors (Zhong et al., 2016;

*MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano, andrea.cappozzo@polimi.it, francesca.ieva@polimi.it

†Department of Biomedical Sciences, Università di Sassari, gfiorito@uniss.it

Guida et al., 2015; Fiorito et al., 2018). To this extent, multi-CpG DNAm biomarkers have been devised to predict patient-specific state of health indicators; and relevant examples include epigenetic clocks to measure "biological age" (Lu et al., 2019), smoking habits (Guida et al., 2015) and proxies for inflammatory proteins (Stevenson et al., 2020). Remarkably, DNAm based scores have been demonstrated to outperform surveyed exposure measurements when predicting diseases (Zhang et al., 2016; Conole et al., 2020). A possible explanation for this somewhat counter-intuitive behavior being that DNA methylation intrinsically accounts for biases in self-reported exposure (e.g., underestimation of smoked cigarettes) as well as individual responses to risk factors (e.g., the same amount of tobacco may produce different effects in dissimilar patients).

From a modeling perspective, state-of-the-art methods for DNAm biomarkers creation generally rely on standard univariate penalized regression models, with elastic-net (Zou and Hastie, 2005) being the routinely employed technique when accomplishing the task. Indeed, the associated learning problem entirely falls within the "p bigger than N" framework: DNA methylation levels are measured at approximately half million CpG sites for each sample, with the dimension of the latter generally not exceeding the order of thousands in most studies. The afore-described procedure is shown to be widely effective in building DNAm biomarkers, with three very recent contributions including a surrogate score for cumulative lead exposure (Colicino et al., 2021), DNAm surrogate for alcohol consumption, obesity indexes, and blood measured inflammatory proteins (Hillary and Marioni, 2021) and the identification of CpG sites associated with clinical severity of COVID-19 disease (Castro de Moura et al., 2021). Nonetheless, elastic-net penalties may be too restrictive when dealing with complex learning problems involving multivariate responses and distinctive dependence patterns across statistical units.

The aforesaid first layer of complexity is encountered when a multi-dimensional DNAm biomarker needs to be created, to jointly model multiple risk factors and to coherently account for the correlation structure among the response variables. Such a multivariate problem, also known as multi-task regression in the machine learning literature (Caruana, 1997), can be fruitfully untangled only if dedicated care is devoted in choosing the most appropriate penalty required for the analysis. For instance, one may opt for the incorporation of $\ell 1/\ell 2$ type of regularizers (Obozinski et al., 2010, 2011; Li et al., 2015), that extend the lasso (Tibshirani, 1996), group-lasso (Yuan and Lin, 2006) and sparse group-lasso (Simon et al., 2013; Laria et al., 2019) to the multiple response framework. Another option could contemplate the inclusion, within the estimation procedure, of prior information related to the association structure among CpG sites: this is effectively achieved by means of graph-based penalties (Li and Li, 2010; Kim et al., 2013; Cheng et al., 2014; Dirmeier et al., 2018). Furthermore, tree-based regularization methods have also been recently introduced in the literature, to account for hierarchical structure over the responses in a single study (Kim and Xing, 2012) as well as when multiple data sources are at our disposal (Zhao and Zucknick, 2020). For a thorough and up-to-date survey on the analysis of high-dimensional omics data via structured regularization we refer the interested

reader to Vinga (2021), while the monograph of Hastie et al. (2015) provides a general introduction to statistical learning with sparsity.

A second layer of complexity is introduced when DNA samples and related blood measured biomarkers are collected in a study comprising multiple cohorts. In such a situation, an unknown degree of heterogeneity may be included in the data, with patients coming from the same cohort sharing some degree of commonality. Observations in the dataset are thus no longer independent and the cohort-wise covariance structure needs to be properly estimated. Linear Mixed-Effects Models (LMM) provide a convenient solution to this problem by adding a random component to the model specification (see, e.g., Pinheiro and Bates, 2006; Gałecki and Burzykowski, 2013; Demidenko, 2013, for an introduction on the topic). Whilst being able to capture unobserved heterogeneity, standard mixed models, very much like their fixed counterpart, cannot directly handle situations in which the number of predictors exceeds the sample size. In order to overcome this issue Schelldorfer et al. (2011) introduced a procedure for estimating high-dimensional LMM via an ℓ_1-penalization . More recently, Rohart et al. (2014) devised a general-purpose ECM algorithm (Meng and Rubin, 1993) for solving the same issue, but achieving greater flexibility as the proposed framework can be combined with any penalty structure previously developed for linear fixed-effects models.

A Multivariate Mixed-Effects Model (MLMM) is an LMM in which multiple characteristics (response variables) are measured for the statistical units comprising the study. Despite being quite a long-established methodology (Reinsel, 1984; Shah et al., 1997), its further development has not received much attention in the recent literature. Relevant exceptions include the computational strategies for handling missing values proposed in Schafer and Yucel (2002), and the estimation theory based on hierarchical likelihood developed in Chipperfield and Steel (2012). On this account, to the best of our knowledge, a unified approach for penalized MLMM estimation is still missing in the literature and it could thus be a relevant contribution to the statistics and machine learning fields.

Motivated by the problem of creating a DNAm biomarker for hypertension and hyperlipidemia from a multi-center study, we propose in this article a general framework for high-dimensional multitask learning with random effects. Leveraging from the algorithm developed in Rohart et al. (2014) for the univariate response case, the learning mechanism is effectively constructed to accommodate custom penalty types, building upon existing routines developed for regression with fixed-effects only.

The remainder of the paper is structured as follows. Section 2 describes the EPIC Italy dataset, which gave the motivation for the development of the methodology proposed in this manuscript. In Section 3 we introduce the penalized mixed-effects model for multitask learning, covering its formulation, inference and model selection. Section 4 outlines the results of the novel method applied to the EPIC Italy data for creating DNAm surrogates for cardiovascular risk factors and comorbidities, comparing it with state-of-the-art alternatives. Section 5 presents a simulation study on synthetic data for two different scenarios. Section 6 concludes the paper with a discussion and directions for future

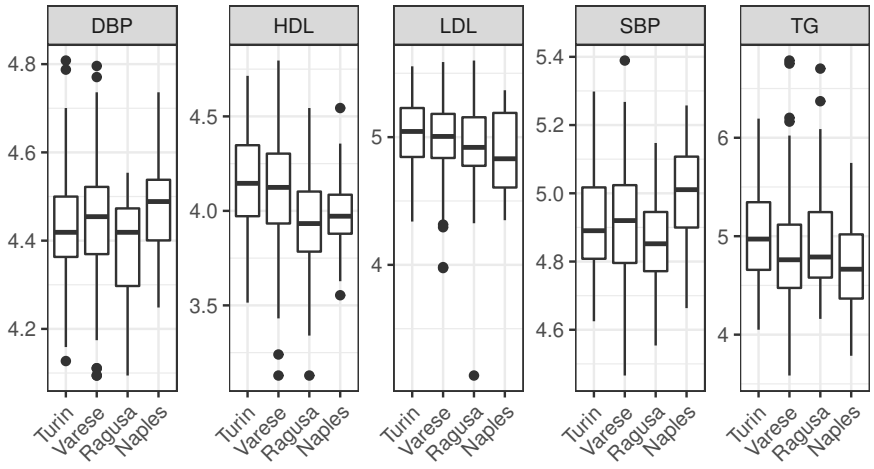

Figure 1: Boxplots of log-transformed Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Systolic Blood Pressure (SBP) and Triglycerides (TG) for different Center, Italy EPIC dataset.

research. The `R` package `emlmm` implementing the proposed method accompanies the article and it is freely available at `https://github.com/AndreaCappozzo/emlmm`.

2 EPIC Italy data and study design

The considered dataset belongs to the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC) study, one of the largest cohort study in the world, with participants recruited across 10 European countries and followed for almost 15 years (Riboli et al., 2002). For each participant, lifestyle and personal history questionnaires were recorded, together with anthropomorphic measures and blood samples for DNA extraction. The EPIC Italy dataset is comprised of four geographical sub-cohorts identified by the center of recruitment: the provinces of Ragusa and Varese and the cities of Turin and Naples. The latter center became associated with EPIC in later times through the Progetto ATENA study (Panico et al., 1992). By profiting from the information recorded in the aforementioned sub-cohorts, we aim at creating a multi-dimensional DNAm biomarker for cardiovascular risk factors and comorbidities. To this extent, we consider a multivariate response comprised of $r = 5$ measures, namely Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL) and Triglycerides (TG). These characteristics were chosen as they represent the major risk factors for cardiovascular diseases (Wu et al., 2015). In building a DNAm biomarker, the response variables are regressed on DNA methylation

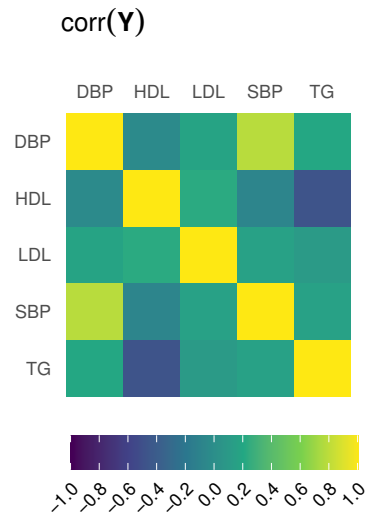

Figure 2: Sample correlation matrix of log-transformed Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Systolic Blood Pressure (SBP) and Triglycerides (TG), Italy EPIC dataset.

values for each CpG site, adjusted for sex and age. A total of $N = 574$ individuals in the $J = 4$ cohorts showcase non-missing values for every response variable: they comprise the sample onto which all subsequent analyses will be performed. An epigenome-wide association study (EWAS, Campagna et al., 2021) was performed as a pre-screening procedure. Whilst variable screening in ultra-high feature space is itself an ongoing research field (see, e.g., Fan and Lv, 2008; Fan et al., 2009; Zhong et al., 2021, and references therein), we decided to rely on the EWAS technique as it is the standard approach employed in epigenomics (Fazzari and Greally, 2010). In so doing, out of the whole set of CpG sites, 13449 DNA methylation features have been retained for subsequent modeling. Together with sex and age, this amounts to a total of $p = 13451$ predictors and a 5-dimensional response. Furthermore, as previously mentioned, the considered samples belong to four different centers distributed across Italy, with data for 128, 334, 68 and 44 units respectively collected in Turin, Varese, Ragusa and Naples provinces. The boxplots in Figure 1 emphasize the differences in the five response variables by center. To capture the center-wise variability and to maintain generalizability of the devised DNAm biomarker outside the Italy EPIC cohorts, a partial pooling random-intercept model must be adopted. That is, a $q = 1$ random effect component is included in the model specification. Furthermore, the biomarkers comprising the response vector showcase some degree of relations, as displayed by the sample correlation matrix of Figure 2, so much so that it is sensible to regress them jointly to take advantage of their association structure in the model formulation. This challenging learning task requires

an ad-hoc formulation for a multivariate mixed-effects framework applicable to high-dimensional predictors.

3 Penalized mixed-effects models for multitask learning

In this section, a novel approach for multivariate mixed-effects modeling based on penalized estimation is proposed.

3.1 Model definition

The multivariate linear mixed-effects model (Shah et al., 1997) expresses the $n_j \times r$ response matrix $\boldsymbol{Y}_j$ for the j-th group as:

$$\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{B} + \boldsymbol{Z}_j\boldsymbol{\Lambda}_j + \boldsymbol{E}_j \tag{1}$$

where, for each of the n_j units in group j and $\sum_{j=1}^{J} n_j = N$, r response variables have been measured. The remainder terms define the following quantities:

- $\boldsymbol{B}$ is the $p \times r$ matrix of fixed-effects (including the intercept)
- $\boldsymbol{\Lambda}_j$ is the $q \times r$ matrix of random-effects
- $\boldsymbol{X}_j$ is the $n_j \times p$ fixed-effects design matrix
- $\boldsymbol{Z}_j$ is the $n_j \times q$ random-effects design matrix
- $\boldsymbol{E}_j$ is the $n_j \times r$ within-group error matrix
- $j = 1, \ldots, J$, with J total number of groups.

By employing the vec operator, we assume that:

$$\operatorname{vec}(\boldsymbol{\Lambda}_j) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}),$$

where $\boldsymbol{\Psi}$ is a $qr \times qr$ positive semidefinite matrix, incorporating variations and covariations between the r responses and the q random effects. We further assume that the error term is distributed as follows:

$$\operatorname{vec}(\boldsymbol{E}_j) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}), \tag{2}$$

where $\boldsymbol{\Sigma}$ is a $r \times r$ covariance matrix, capturing dependence among responses, and $\boldsymbol{I}_{n_j}$ is the identity matrix of dimension $n_j \times n_j$. Formulation in (2) explicitly induces independence between the row vectors of $\boldsymbol{E}_j$. Therefore, the entire model can be rewritten in vec form:

$$\operatorname{vec}(\boldsymbol{Y}_j) \sim N\left((\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}), (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right).$$

Given a sample of $N = \sum_{j=1}^{J} n_j$, the log-likelihood of model (1) reads:

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} -\frac{n_j}{2}\log 2\pi - \frac{1}{2}\log|\left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right)\boldsymbol{\Psi}\left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}| + \\ - \frac{1}{2}\left(\text{vec}\left(\boldsymbol{Y}_j\right) - \left(\boldsymbol{I}_r \otimes \boldsymbol{X}_j\right)\text{vec}\left(\boldsymbol{B}\right)\right)' \left(\left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right)\boldsymbol{\Psi}\left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right)^{-1} \left(\text{vec}\left(\boldsymbol{Y}_j\right) - \left(\boldsymbol{I}_r \otimes \boldsymbol{X}_j\right)\text{vec}\left(\boldsymbol{B}\right)\right) \tag{3}$$

Where $\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\}$ is the set of parameters to be estimated. When the framework outlined in (1) is employed for DNAm biomarker creation, the number of regressors p is most certainly much larger than the sample size N. We are thus not directly interested in maximizing (3), but rather a penalized version of it, generically defined as follows:

$$\ell_{pen}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - p(\boldsymbol{B}; \lambda), \tag{4}$$

with $p(\boldsymbol{B}; \lambda)$ being a penalty term employed to regularize the fixed-effects $\boldsymbol{B}$ as a function of the complexity parameter $\lambda \geq 0$. Notice that, depending on the chosen penalty, more than one complexity parameter could be involved in the definition of $p(\boldsymbol{B}; \lambda)$ (see Section 3.3 for further details).

A general-purpose algorithm for maximizing (4) can be devised, as described in the next subsection.

3.2 Model estimation

Direct maximization of (4) is unfeasible, as the terms $\text{vec}\left(\boldsymbol{\Lambda}_j\right)$, $j = 1, \ldots, J$, are unknown. We therefore devise an EM algorithm (Dempster et al., 1977) in which the E-step computes the conditional expectations for the unobserved quantities, while a *complete penalized log-likelihood* is maximized in the M-step.

3.2.1 E-step

The E-step requires the computation of $\mathbb{E}(\text{vec}\left(\boldsymbol{\Lambda}_j\right)|\boldsymbol{Y}_j; \boldsymbol{\theta})$ and $\mathbb{E}(\text{vec}\left(\boldsymbol{\Lambda}_j\right)\text{vec}\left(\boldsymbol{\Lambda}_j\right)'|\boldsymbol{Y}_j; \boldsymbol{\theta})$. This is achieved by noticing that the conditional density $p(\text{vec}\left(\boldsymbol{\Lambda}_j\right)|\boldsymbol{Y}_j; \boldsymbol{\theta})$ is Normal. Updating formulae for the quantities of interest are thus derived as follows:

$$\hat{\boldsymbol{\Gamma}}_j = \mathbb{V}(\text{vec}\left(\boldsymbol{\Lambda}_i\right)|\boldsymbol{Y}_j; \boldsymbol{\theta}) = \left[\left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right)'\left(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right)^{-1}\left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right) + \boldsymbol{\Psi}^{-1}\right]^{-1}, \tag{5}$$

$$\widehat{\text{vec}\left(\boldsymbol{\Lambda}_j\right)} = \mathbb{E}(\text{vec}\left(\boldsymbol{\Lambda}_j\right)|\boldsymbol{Y}_j; \boldsymbol{\theta}) = \hat{\boldsymbol{\Gamma}}_j\left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right)'\left(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right)^{-1}\left(\text{vec}\left(\boldsymbol{Y}_j\right) - \left(\boldsymbol{I}_r \otimes \boldsymbol{X}_j\right)\text{vec}\left(\boldsymbol{B}\right)\right). \tag{6}$$

Consequently, the second moment $\hat{\boldsymbol{R}}_j = \mathbb{E}(\text{vec}\left(\boldsymbol{\Lambda}_j\right)\text{vec}\left(\boldsymbol{\Lambda}_j\right)'|\boldsymbol{Y}_j; \boldsymbol{\theta})$ reads:

$$\hat{\boldsymbol{R}}_j = \hat{\boldsymbol{\Gamma}}_j + \widehat{\text{vec}\left(\boldsymbol{\Lambda}_j\right)}\widehat{\text{vec}\left(\boldsymbol{\Lambda}_j\right)}'. \tag{7}$$

At the t-th iteration of the EM algorithm, the E-step requires the computation of (5)-(7) conditioning on the parameter values estimated at iteration $t-1$. Notice that we can directly define the conditional density of $\boldsymbol{Y}_j|\boldsymbol{\Lambda}_j$ by means of the matrix normal distribution

$$\boldsymbol{Y}_j|\boldsymbol{\Lambda}_j \sim m\mathcal{N}\left(\boldsymbol{X}_j\boldsymbol{B}+\boldsymbol{Z}_j\boldsymbol{\Lambda}_j, \boldsymbol{I}_{n_j}, \boldsymbol{\Sigma}\right), \tag{8}$$

where $\boldsymbol{X}_j\boldsymbol{B}+\boldsymbol{Z}_j\boldsymbol{\Lambda}_j$ is the $n_j \times r$ mean matrix, and $\boldsymbol{I}_{n_j}$, $\boldsymbol{\Sigma}$ respectively identify the row and column covariance matrices (Dawid, 1981). Such a representation will be useful in specifying the update for $\boldsymbol{B}$ in the devised M-step: details are provided in the next subsection.

3.2.2 M-step

In the M-step we maximize the *complete penalized log-likelihood*:

$$\begin{aligned}\ell_{C\,pen}(\boldsymbol{\theta}) &= \sum_{j=1}^{J} \log(p(\operatorname{vec}(\boldsymbol{Y}_j) \,|\, \operatorname{vec}(\boldsymbol{\Lambda}_j); \boldsymbol{B}, \boldsymbol{\Sigma})) + \log(p(\operatorname{vec}(\boldsymbol{\Lambda}_j); \boldsymbol{\Psi})) - p(\boldsymbol{B};\lambda) = \\ &= \sum_{j=1}^{J} -\frac{n_j}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}\otimes\boldsymbol{I}_{n_j}| - \frac{1}{2}\mathbb{E}(\boldsymbol{e}_j^{'}\left(\boldsymbol{\Sigma}\otimes\boldsymbol{I}_{n_j}\right)^{-1}\boldsymbol{e}_j|\boldsymbol{Y}_j,\boldsymbol{\theta}) + \\ &- \frac{n_j}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\mathbb{E}\left(\operatorname{vec}(\boldsymbol{\Lambda}_j)^{'}\boldsymbol{\Psi}^{-1}\operatorname{vec}(\boldsymbol{\Lambda}_j)\,|\boldsymbol{Y}_j,\boldsymbol{\theta}\right) - p(\boldsymbol{B};\lambda),\end{aligned} \tag{9}$$

where $\boldsymbol{e}_j = \operatorname{vec}(\boldsymbol{Y}_j) - (\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}) - (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\operatorname{vec}(\boldsymbol{\Lambda}_j)$ and the maximization is performed with respect to $\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\}$.

The updating formula for $\boldsymbol{B}$ clearly depends on the considered $p(\boldsymbol{B};\lambda)$ penalty. All the same, it is convenient to work with the matrix-variate representation defined in (8). In so doing, the objective function to be maximized wrt $\boldsymbol{B}$ reads:

$$Q_{\boldsymbol{B}}(\boldsymbol{B}) = -\frac{1}{2}\sum_{j=1}^{J}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{Y}}_j - \boldsymbol{X}_j\boldsymbol{B}\right)^{'}\left(\tilde{\boldsymbol{Y}}_j - \boldsymbol{X}_j\boldsymbol{B}\right)\right) - p(\boldsymbol{B};\lambda), \tag{10}$$

where $\tilde{\boldsymbol{Y}}_j = \boldsymbol{Y}_j - \boldsymbol{Z}_j\hat{\boldsymbol{\Lambda}}_j$ and $\hat{\boldsymbol{\Lambda}}_j$ is $\widehat{\operatorname{vec}(\boldsymbol{\Lambda}_j)}$, previously computed in the E-step, rearranged in matrix form. Start by noticing that, when no penalty is considered, maximization of (10) agrees with the generalized least squares (GLS) estimator assuming $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ known (Shah et al., 1997). By exploiting properties of the trace operator, we can rewrite (10) defining the following minimization problem:

$$\text{minimize}_{\boldsymbol{B}\in\mathbb{R}^{p\times r}}\ \frac{1}{2}\sum_{j=1}^{J}\left|\left|\boldsymbol{\Sigma}^{-1/2}\left(\tilde{\boldsymbol{Y}}_j - \boldsymbol{X}_j\boldsymbol{B}\right)^{'}\right|\right|_F^2 + p(\boldsymbol{B};\lambda) \tag{11}$$

where $||\cdot||_F^2$ denotes the squared Frobenius norm and $\boldsymbol{\Sigma}^{-1/2}$ is the symmetric positive definite square root of $\boldsymbol{\Sigma}^{-1}$, such that $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}$. The representation in (11) allows to employ standard routines for multivariate penalized fixed-effect models for estimating $\boldsymbol{B}$. In details, we start by computing:

$$\tilde{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \frac{1}{2} \sum_{j=1}^{J} \left|\left| \boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{Y}}_j - \boldsymbol{X}_j\boldsymbol{B} \right|\right|_F^2 + p(\boldsymbol{B};\lambda). \tag{12}$$

Notice that (12) is a fixed-effects penalized regression problem in which the response variable is $\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{Y}}_j$, $j = 1, \ldots, J$. The final update for (10) is obtained by post multiplying $\tilde{\boldsymbol{B}}$ by $\boldsymbol{\Sigma}^{1/2}$; that is, at each iteration of the EM-algorithm, we firstly compute $\tilde{\boldsymbol{B}}$ via fixed-effects routines for penalized estimation and then we set:

$$\hat{\boldsymbol{B}} = \tilde{\boldsymbol{B}}\boldsymbol{\Sigma}^{1/2}, \tag{13}$$

where $\hat{\boldsymbol{B}}$ maximizes (10). This procedure stems from the rationale outlined in Rohart et al. (2014), where, contrarily to their original solution, in our context the updating steps are made more complex by the multidimensional nature of $\boldsymbol{Y}$. The devised updating scheme allows to easily incorporate any $p(\boldsymbol{B};\lambda)$ that has been previously defined for the fixed-effects framework, and whose estimating routines are available. A list of possible penalties is proposed in Section 3.3.

Updating formulae for the covariance matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ agree with those of the unpenalized framework, namely

$$\hat{\boldsymbol{\Psi}} = \frac{1}{J} \sum_{j=1}^{J} \hat{\boldsymbol{R}}_j, \tag{14}$$

and for the (h,k)-th element of matrix $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\Sigma}}_{(h,k)} = \frac{1}{N} \sum_{j=1}^{J} \left[\mathbb{E}\left(\boldsymbol{E}_{jh}|\boldsymbol{Y}_j\right)' \mathbb{E}\left(\boldsymbol{E}_{jk}|\boldsymbol{Y}_j\right) \right] + \mathrm{tr}\left[\mathrm{cov}(\boldsymbol{E}_{jh}, \boldsymbol{E}_{jk}|\boldsymbol{Y}_j)\right], \quad h,k = 1,\ldots,r, \tag{15}$$

where $\boldsymbol{E}_{jh}$ denotes the h-th column of matrix $\boldsymbol{E}_j = \boldsymbol{Y}_j - \boldsymbol{Z}_j\hat{\boldsymbol{\Lambda}}_j - \boldsymbol{X}_j\boldsymbol{B}$, $h = 1,\ldots,r$.

3.3 On the choice of $p(\boldsymbol{B};\lambda)$

The EM algorithm devised in the previous section defines a general-purpose optimization strategy for penalized mixed-effects multitask learning. Nonetheless, in practice, a functional form for $p(\boldsymbol{B};\lambda)$ must be chosen when performing the analysis. While any penalty type can in principle be defined, three notable examples, commonly used in this context, are the elastic net penalty (Zou and Hastie, 2005), the group-lasso penalty for multivariate regression (Hastie et al., 2015) and the netReg routines for Network-regularized linear models (Dirmeier et al., 2018): each of them is briefly described in the next subsections, highlighting pros and cons wrt a mixed-effects multitask learning setting.

3.3.1 Elastic-net penalty

The first penalty type we consider is the renowned convex combination of lasso and ridge regularizers, whose magnitude of the former over the latter is controlled by the mixing parameter α, $0 \leq \alpha \leq 1$. In details, the penalty expression reads:

$$p(\boldsymbol{B}; \lambda, \alpha) = \lambda \left[(1-\alpha) \sum_{c=1}^{r} \sum_{l=2}^{p} b_{lc}^2 + \alpha \sum_{c=1}^{r} \sum_{l=2}^{p} |b_{lc}| \right], \tag{16}$$

where b_{lc} denotes the element in the l-th row and c-th column of matrix $\boldsymbol{B}$. Notice that the first row of $\boldsymbol{B}$ contains the r intercepts and it is thus not penalized. The penalty in (16) does not take into account the multivariate nature of the problem in (4), as the shrinkage is applied directly to vec($\boldsymbol{B}$). This behavior allows for capturing a wide variety of sparsity patterns that may be present in $\boldsymbol{B}$, but does not impose any specific structure that may be desirable in a multivariate context (see next subsection). Algorithmically, penalty (16) can be enforced employing standard and widely available routines for univariate penalized estimation, like the `glmnet` software (Tay et al., 2021). The only computational detail that shall be examined is how to prevent the default shrinkage of the r intercepts: the `penalty.factor` argument of the `glmnet` function effectively serves the purpose.

3.3.2 Group-lasso penalty

This type of penalty imposes a group structure on the coefficients, forcing the same subset of predictors to be preserved across all r components of the response matrix. This feature is particularly desirable when building multivariate DNAm biomarkers, since it automatically identifies the CpG sites that are *jointly* related to the considered risk factors. Such a penalty is defined as follows:

$$p(\boldsymbol{B}; \lambda, \alpha) = \lambda \left[(1-\alpha) \sum_{c=1}^{r} \sum_{l=2}^{p} b_{lc}^2 + \alpha \sum_{l=2}^{p} ||\boldsymbol{b}_{l\cdot}||_2 \right], \tag{17}$$

where $\boldsymbol{b}_{l\cdot}$ identifies the l-th row of the matrix $\boldsymbol{B}$, such that each $\boldsymbol{b}_{l\cdot}$, $l = 2, \ldots, p$ is an r-dimensional vector. Likewise Section 3.3.1, summations over rows in (17) start at 2 since we do not penalize the vector of intercepts. This penalty behaves like the lasso, but on the whole group of predictors for each of the r variables: they are either all zero, or else none are zero, but are shrunk by an amount depending on λ. Similarly to (16), the mixing parameter α controls the weight associated to ridge and group-lasso regularizers. The `glmnet` software, with `family = "mgaussian"` is again at our disposal for efficiently incorporating (17) in the framework outlined in the present paper.

3.3.3 Network-Regularized penalty

The last penalty we consider allows for the inclusion of biological graph-prior knowledge in the estimation by accounting for the contribution of two non-

negative adjacency matrices $\boldsymbol{G}_X \in \mathbb{R}_+^{(p-1)\times(p-1)}$ and $\boldsymbol{G}_Y \in \mathbb{R}_+^{r\times r}$, respectively related to $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this case, $p(\boldsymbol{B};\lambda)$ assumes the following functional form:

$$p(\boldsymbol{B};\lambda,\lambda_X,\lambda_Y) = \lambda||\boldsymbol{B}_0||_1 + \lambda_X \operatorname{tr}\left(\boldsymbol{B}_0^{'}(\boldsymbol{D}_{G_X} - \boldsymbol{G}_X)\boldsymbol{B}_0\right) + \lambda_Y \operatorname{tr}\left(\boldsymbol{B}_0(\boldsymbol{D}_{G_Y} - \boldsymbol{G}_Y)\boldsymbol{B}_0^{'}\right) \tag{18}$$

where $\boldsymbol{B}_0$ is the $(p-1)\times r$ matrix of coefficients without the intercepts and $\boldsymbol{D}_{G_X}$, $\boldsymbol{D}_{G_Y}$ indicate the degree matrices of $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$, respectively (Chung and Graham, 1997). $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$ encode a biological similarity, forcing rows and columns of $\boldsymbol{B}_0$ to be similar. Such a penalty is particularly useful when the interaction among features and/or responses is, at least partially, known, such that it can be profited from within the learning mechanism (Cheng et al., 2014). The `netReg R` package provides a convenient implementation of (18) (Dirmeier et al., 2018).

3.4 Further aspects

Hereafter, we discuss some practical considerations related to the presented methodology.

- **Initialization:** we start the algorithm with an M-step, setting $\hat{\theta}^{(0)} = \{\hat{\boldsymbol{B}}^{(0)}, \hat{\boldsymbol{\Sigma}}^{(0)}, \hat{\boldsymbol{\Psi}}^{(0)}\}$. In details, both $\hat{\boldsymbol{\Sigma}}^{(0)}$ and $\hat{\boldsymbol{\Psi}}^{(0)}$ are initialized with identity matrices of dimension $r\times r$ and $qr\times qr$ respectively, while $\hat{\boldsymbol{B}}^{(0)}$ is estimated from a penalized linear model (without the random-effects) employing the chosen penalty function with the associated hyper-parameters.

- **Convergence:** the EM algorithm is considered to have converged once the relative difference in the objective function for two subsequent iterations is smaller than ε, for a given $\varepsilon > 0$:

$$\frac{|\ell_{pen}(\hat{\boldsymbol{\theta}}^{(t+1)}) - \ell_{pen}(\hat{\boldsymbol{\theta}}^{(t)})|}{|\ell_{pen}(\hat{\boldsymbol{\theta}}^{(t)})|} < \varepsilon,$$

  where $\hat{\boldsymbol{\theta}}^{(t)} = \{\hat{\boldsymbol{B}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)}, \hat{\boldsymbol{\Psi}}^{(t)}\}$ is the set of estimated values at the end of the t-th iteration. In our analyses, ε is set equal to 10^{-6}. The procedure described in Section 3.2 falls within the class of Expectation Conditional Maximization (ECM) algorithms, whose convergence properties have been proved in Meng and Rubin (1993) and in Section 5.2.3 of McLachlan and Krishnan (2008).

- **Model selection:** a standard 10-fold cross validation (CV) strategy is employed for selecting the tuning factors. Alternatively, as suggested in Rohart et al. (2014), one could employ a modified version of the Bayesian Information Criterion (BIC, Schwarz, 1978):

$$BIC = 2\ell(\hat{\boldsymbol{\theta}}) - d_0 \log(N), \tag{19}$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the log-likelihood evaluated at $\hat{\boldsymbol{\theta}}$, obtained maximizing (4), and d_0 is the number of non-zero parameters resulting from the penalized estimation. Another option would be to rely on an interval search algorithm, like the efficient parameter selection via global optimization (Frohlich and Zell, 2005): an implementation is available in the `c060 R` package (Sill et al., 2014).

- **Scalability:** the devised methodology provides a framework for incorporating any penalty in a high-dimensional mixed-effects multitask learning framework. To this extent, the data dimensionality our procedure can cope with very much depends on the scalability associated to the chosen shrinkage term. Typically nevertheless, penalized likelihood approaches fail to be directly applied to ultrahigh-dimensional problems (Fan et al., 2009), and pre-processing procedures such as variable screening are thus required prior to modeling. The epigenetic application that motivated the procedure naturally called for an EWAS pre-screening strategy (see Section 2), but clearly other dimensionality reduction techniques could be considered when dealing with massive datasets. The interested reader is referred to Jordan (2013) for a thought-provoking investigation on the topic.

- **Implementation:** routines for fitting the penalized mixed-effects multitask learning method have been implemented in `R` (R Core Team, 2021), and the source code is freely available at `https://github.com/AndreaCappozzo/emlmm` in the form of an `R` package. The three penalties described in Section 3.3 are included in the software, and can be selected via the `penalty_type` argument of the `ecm_mlmm_penalized` function. As described in Section 3.3, the M-step heavily relies on previously developed fast and stable subroutines, while the E-step and the objective function evaluation have been implemented in `c++` to reduce the overall computing time.

- **Response-specific random-effects:** model in (1) assumes that each and every response requires a random-effects component. Whilst in principle reasonable, it may happen in specific applications that only a subset of the r characteristics in $\boldsymbol{Y}$ enjoys group-dependent heterogeneity. The occurrence of such a scenario can be unveiled by looking at the r diagonal elements of dimension q in $\hat{\boldsymbol{\Psi}}$: a response may be considered group-independent when the magnitude of the associated elements in $\text{diag}(\hat{\boldsymbol{\Psi}})$ is significantly lower than the remaining ones. Doing this way, the impact a given random-effect has on the different characteristics is retrieved as a by-product of the modeling procedure.

4 DNAm biomarkers creation from EPIC dataset

The methodology described in the previous section is employed to build a 5-dimensional DNAm biomarker of hypertension and hyperlipidemia. As men-

Table 1: Root Mean Squared Error (RMSE) and active number of CpG sites for different penalized regression models, EPIC Italy test set. Bold numbers indicate lowest RMSE for each of the $r = 5$ dimension of the response matrix.

	Framework		Root Mean Squared Error					Active #
Model	Penalty type	Response	DBP	HDL	LDL	SBP	TG	CpG sites
Random-effects	Group-lasso	Multivariate	**0.1024**	**0.2065**	0.2887	**0.1187**	**0.3958**	1824
Random-effects	Elastic-net	Multivariate	0.1089	0.2103	0.2844	0.1263	0.4138	1468
Fixed-effects	Group-lasso	Multivariate	0.1098	0.2141	0.2838	0.126	0.4036	874
Fixed-effects	Elastic-net	Multivariate	0.1162	0.2298	0.2988	0.1329	0.4227	441
Fixed-effects	Elastic-net	Univariate	0.1043	0.2106	**0.2781**	0.1226	0.4002	1933

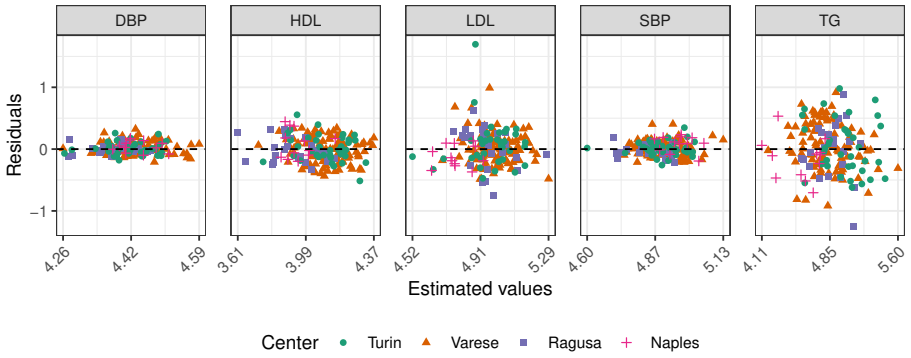

Figure 3: Residual plots of log-transformed Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Systolic Blood Pressure (SBP) and Triglycerides (TG) for group-lasso mixed-effects model for multitask learning. Residuals are colored by Center, Italy EPIC test set.

tioned in the introduction, DNAm surrogates possess extensive advantages over their blood-measured counterparts, since they directly account for genetic susceptibility and subject specific response to risk factors. Furthermore, once the DNAm biomarkers have been created (i.e., model parameters have been estimated), their values can immediately be predicted for patients not directly involved in the study, even coming from an external cohort with available DNAm data. This is particularly interesting when the risk factor or exposure has not been directly measured in the external cohort. In addition to the epidemiological usefulness of DNAm surrogates, further understanding of the biomolecular mechanisms associated with complex phenotypes can be acquired through a pathway enrichment analysis (Reimand et al., 2019). The latter allows to identify molecular pathways overrepresented among the regressors involved in the surrogate construction (i.e., the CpG sites whose associated parameters are not shrunk to 0).

To reconstruct the process of DNAm surrogates creation and validation, the EPIC Italy data is randomly split into two sets: 70% ($N_{tr} = 401$) of it is em-

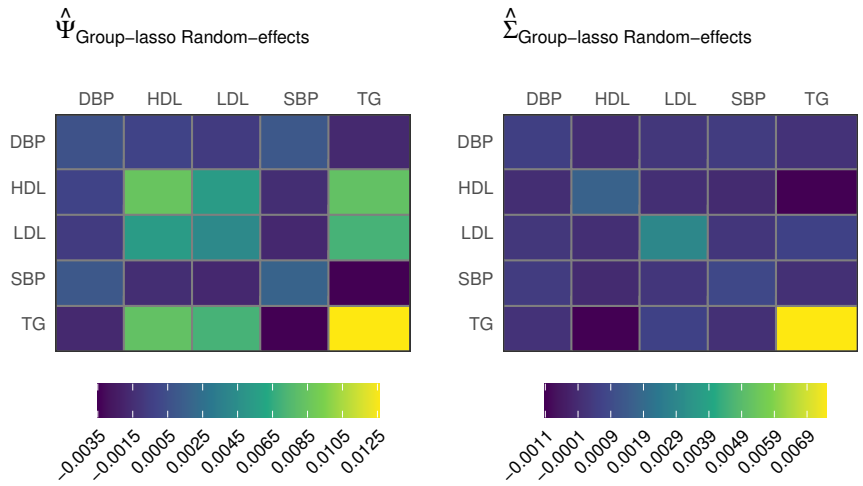

Figure 4: Estimated covariance matrix $\hat{\mathbf{\Psi}}$ of the random-effects and covariance matrix $\hat{\mathbf{\Sigma}}$ of the error term for group-lasso mixed-effects model for multitask learning, Italy EPIC dataset.

ployed for model fitting, while the remaining 30% ($N_{te} = 173$) acts as test set for assessing predictive performance. Several estimation strategies are contemplated varying penalty and modeling type. For each model, the penalty factor λ was tuned via 10-fold CV on the training set, while the mixing parameter α was kept fixed and equal to 0.5. As mentioned in Section 2, the design matrix comprises of $p = 13451$ variables and redundancies are likely to occur as the feature space is constituted by the union of CpG sites pre-screened by univariate epigenome-wide analyses. Results are summarized in Table 1, where the Root Mean Squared Error (RMSE) and the number of active CpG sites are reported. The first two rows are related to the novel penalized MLMM methodology with a random-effects design matrix that includes a $q = 1$ random intercept, coupled with elastic-net (Section 3.3.1) and group-lasso (Section 3.3.2) penalties, respectively. The corresponding fixed-effects counterparts are reported in the third and fourth rows, while univariate elastic-net metrics, obtained fitting $r = 5$ separate models, one for each response, are detailed in the last row of Table 1. Notice that our proposal outperforms the state-of-the-art approach (univariate elastic-net) for 4 out of 5 dimensions of the response variable. The reason being that our method takes advantage of the borrowing information asset typical of multivariate models (the correlation between SBP and DBP is equal to 0.77 in the training set), whilst allowing for center-wise difference to be captured by the random intercept. Furthermore, thanks to the group-lasso penalty, our penalized MLMM approach directly identifies the CpG sites that are jointly related to hypertension and hyperlipidemia, with a total number of features that

is lower with respect to univariate elastic-nets. Figure 4 displays the estimated covariance matrix $\hat{\boldsymbol{\Psi}}$ of the random-effects (left panel) and covariance matrix $\hat{\boldsymbol{\Sigma}}$ (right panel) for the penalized MLMM model with group-lasso penalty. By taking the ratio between the diagonal elements of $\hat{\boldsymbol{\Psi}}$ and the sum of the diagonals of $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Sigma}}$ it is possible to compute, for each component of the response matrix $\boldsymbol{Y}$, the analogue of the Percentage of Variation due to Random Effects (PVRE) index. For the EPIC Italy dataset, the estimated PVRE amounts to 56.07%, 83.35%, 58.4%, 67.66% and 62.59% for DBP, HDL, LDL, SBP and TG, respectively. Notice that DBP and LDL possess lower PVREs than the other biomarkers.

The employment of the group-lasso penalty within a mixed-effects multi-task learning framework is also supported by biological reasons. In fact, it is more likely that multiple correlated phenotypes affect (or are affected by, depending on the causal relationship between DNAm and the exposure variable) the same set of CpG sites. This mechanism is known as pleiotropic effect (Tyler et al., 2013; Richard et al., 2017). In addition, the incorporation of a random intercept in the model is further motivated by the intrinsic rationale of DNAm biomarkers creation: the compelling necessity of developing study-invariant DNAm biomarkers, whilst still being able to capture the center effect, can be properly achieved by modeling the latter with a random component. The network regularized penalty has not been included in the comparison as the incorporation of prior knowledge through graph-based regularizers does not seem to be suited for this context, with predictive metrics being much worse for both random-effects and fixed-effects models. Figure 3 reports the residuals vs fitted plots for the model in the first row of Table 1: each dimension displays a satisfactory diagnostic pattern, also supported by normality checks on the residuals.

In addition to the higher prediction performance and epidemiological rationale of our approach compared to the univariate elastic-nets, we investigated the biological reliability of the selected features (CpG sites). The univariate elastic-nets extracted 492, 325, 469, 481, 489 CpG sites for diastolic blood pressure, systolic blood pressure, HDL cholesterol, LDL cholesterol, and triglycerides, respectively. The total number of unique CpGs was 1933. However, despite the high degree of correlation among the multivariate outcomes, no CpGs were in common in the five sets, and only a minor percentage of CpGs was shared among two or more responses, as it is represented in Figure 5. Instead, as previously described , our MLMM procedure regularized with a group-lasso penalty extracts features that are associated with the five outcomes at the same time, a biological mechanism known as pleiotropy (Atchley and Hall, 1991a,b), increasing the biological reliability of our findings. On this wise, we extracted all the CpGs previously associated with blood pressure, HDL cholesterol, LDL cholesterol, and triglycerides from the EWAS catalogue (Battram et al., 2021), and we investigated the overlap with the CpGs extracted in the EPIC Italy dataset by the two approaches (univariate elastic-nets and penalized MLMM with group-lasso penalty). The EWAS catalogue collects the results from epigenome-wide

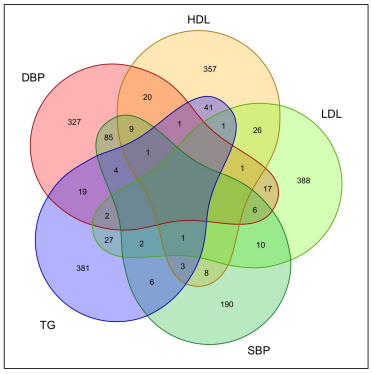

Figure 5: Venn diagram highlighting the number of CpG sites in common among those extracted by univariate elastic-nets for Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Systolic Blood Pressure (SBP) and Triglycerides (TG) biomarkers.

association studies (EWAS) satisfying stringent inclusion parameters (Battram et al., 2021). For the comparison, we selected the CpGs associated with (at least one) of the outcomes considered in this study, with p-value of the association test lower than 10^{-8}, which is considered the optimal threshold of significance to avoid false positives in EWAS studies.

Despite the low number of CpG sites selected by penalized MLMM compared with univariate elastic-nets, the former systematically identifies a higher number of CpG sites in common with the EWAS catalogue. In fact, out of 52 CpGs associated with systolic blood pressure, 8 (15.3%) are retained by our approach, whereas only 6 (11.5%) were identified as relevant by the univariate elastic-net. Interestingly, only one out of six common CpGs resulted from the systolic blood pressure specific analysis. Further, out of 32 CpGs associated with diastolic blood pressure, 7 (21.9%) were in the MLMM list, whereas 5 (15.6%) were in the univariate elastic-nets set. Of those, two out of five came out from the diastolic blood pressure specific analysis. Similarly, out of 12 CpGs associated with triglycerides, 5 (41.7%) were in the MLMM list and 4 (33.3%) in the univariate elastic-net set. No CpGs associated with LDL and HDL cholesterol were found in the EWAS catalogue. These results further support the higher reliability and reproducibility in independent datasets of a multitask learning framework compared to the current state-of-the-art methods.

All in all, the proposed approach exhibits promising results when it comes to multivariate DNAm biomarker creation, outperforming the current employed procedure, both in terms of predictive power and epidemiological interpretation.

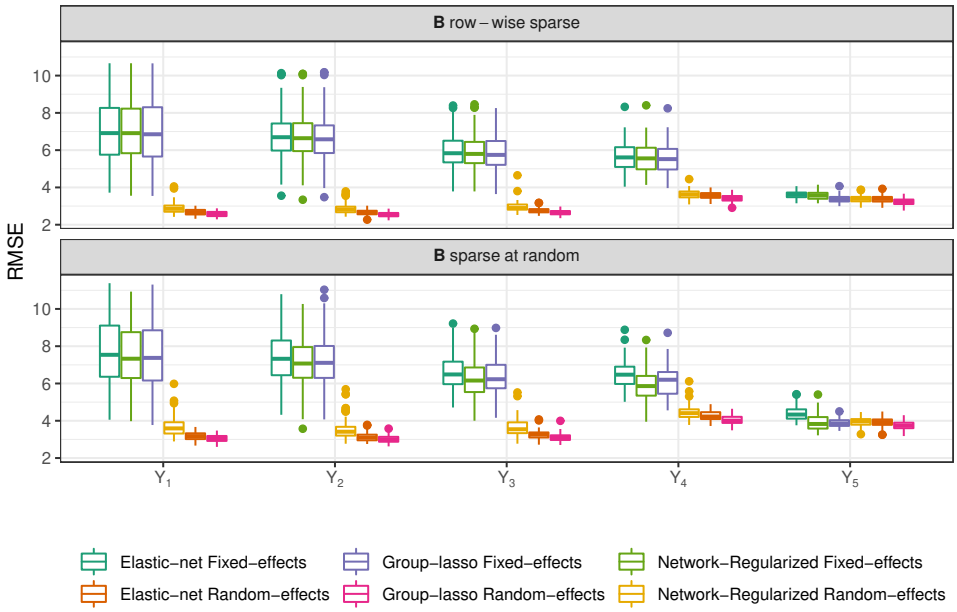

Figure 6: Boxplots of the Root Mean Squared Error (RMSE) for $MC = 100$ repetitions of the simulated experiment. RMSE is computed on 200 test points for different methods and two scenarios varying sparsity pattern for $\boldsymbol{B}$.

5 Simulation study

In this section, we evaluate the model introduced in Section 3 on synthetic data. The aim of the analyses reported hereafter is twofold. On the one hand, we would like to validate the predictive power of the proposed procedure against its fixed-effects counterpart when the random-effects vary across dimensions in the multivariate response. On the other hand, we assess the estimated model parameters and the recovery of the underlying sparsity structure for different values of the shrinkage factor λ.

5.1 Experimental setup

We generate $N = 600$ data points according to model (1) with the following parameters:

$$\boldsymbol{\Psi} = \begin{bmatrix} 50.00 & -1.59 & -0.60 & -0.22 & 2.38 \\ -1.59 & 40.00 & -0.96 & -0.91 & 0.37 \\ -0.60 & -0.96 & 30.00 & -0.43 & 0.50 \\ -0.22 & -0.91 & -0.43 & 20.00 & 0.80 \\ 2.38 & 0.37 & 0.50 & 0.80 & 0.16 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 3.56 & -2.17 & 1.15 & 2.52 & 0 \\ -2.17 & 9.04 & 0.20 & 0.59 & 0 \\ 1.15 & 0.20 & 4.35 & 2.85 & 0.02 \\ 2.52 & 0.59 & 2.85 & 5.34 & 0.03 \\ 0 & 0 & 0.02 & 0.03 & 5.03 \end{bmatrix},$$

implying that $r = 5$ and $q = 1$. Notice that $\boldsymbol{\Psi}$ is purposely constructed for the random-effects to differently affect the five dimensional response: while the first component showcases high variance (first entry in the main diagonal) the last one is very small and close to 0. The data generating process further assumes ten equally-sized subpopulations, resulting in $J = 10$. The matrix of fixed-effects $\boldsymbol{B}$ is of dimension 101×5, with distinct sparsity pattern according to two scenarios:

- $\boldsymbol{B}$ *row-wise sparse*: $\boldsymbol{B}$ has entries equal to 0.5 for the first 21 rows, while all the other entries are equal to 0,
- $\boldsymbol{B}$ *sparse at random*: $\boldsymbol{B}$ is equal to 0.5 for approximately 70% of its entries, while all the others are equal to 0.

A graphical representation of the resulting structures can be found in the top panels of Figure 10. Lastly, $\boldsymbol{Z}_j$ is an all-one column vector $\forall j = 1, \ldots, 10$, while $\boldsymbol{X}_j$ has the first column equal to 1, meaning that the intercept is included in $\boldsymbol{X}_j$ in our model specification, and the remaining 100 dimensions are generated according to a standard normal random vector.

Taking a cue from the Monte Carlo simulations of Li and Li (2010), for each replication of our experiment the learning framework is structured as follows: we equally divide the $N = 600$ units in a training set, an independent validation set and an independent test set, retrieving a sample size of 200 for each. Six different models, varying λ within a grid, are fitted on the training data:

- *Elastic-net Fixed-effects:* a penalized multitask learning model with elastic-net regularization. The considered penalty is described in Section 3.3.1,
- *Group-lasso Fixed-effects:* a penalized multitask learning model with group-lasso regularization. The considered penalty is described in Section 3.3.2,
- *Network-Regularized Fixed-effects:* graph-regularized multitask learning model with edge-based regularization. The considered penalty is described in Section 3.3.3,
- *Elastic-net Random-effects:* the penalized mixed-effects multitask learning model introduced in the paper with elastic-net regularization (Section 3.3.1),
- *Group-lasso Random-effects:* the penalized mixed-effects multitask learning model introduced in the paper, with group-lasso regularization (Section 3.3.2),
- *Network-Regularized Random-effects:* the penalized mixed-effects multitask learning model introduced in the paper, with edge-based regularization (Section 3.3.3).

The mixing parameter α was set equal to 0.5 for methods with elastic-net and group-lasso regularizers, while for the Network-Regularized penalty we employ

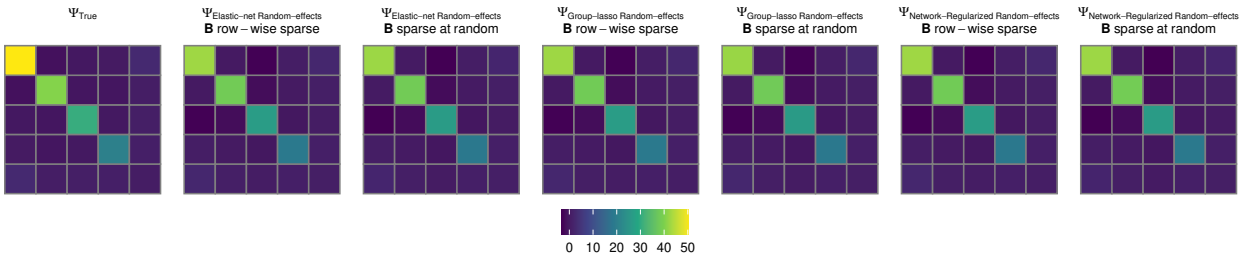

Figure 7: True covariance matrix of the random-effects and estimated $\boldsymbol{\Psi}$, averaged over $MC = 100$ replication of the simulated experiment, for different methods and scenarios.

5-fold CV to tune λ_X and λ_Y on the training set. For the latter penalty, the adjacency matrices $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$ are computed via a thresholding procedure on the correlation matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, with a threshold equal to 0.1 (Langfelder and Horvath, 2008). Subsequently, the validation dataset is used to select the best shrinkage parameter λ minimizing the RMSE for every model. Lastly, the predictive performance is assessed on the test set. The devised simulated experiment is replicated $MC = 100$ times: results are reported in the next subsection.

5.2 Simulation results

Figure 6 displays boxplots of the Root Mean Squared Error, computed for each component of the 5-dimensional response on the test set. For both scenarios ($\boldsymbol{B}$ *row-wise sparse* and $\boldsymbol{B}$ *sparse at random*) we observe that the component-wise predictive performance is heavily affected by the magnitude of the related diagonal entry in the $\boldsymbol{\Psi}$ matrix. When the grouping effect is negligible (fifth dimension Y_5), all methods showcase comparable predictive performance under both scenarios. Contrarily, the RMSE deteriorates for fixed-effects models in those response components for which the grouping impact is more relevant. The same does not happen for the mixed-effects counterparts, as the random intercept effectively captures baseline differences across groups. Interestingly, the penalty type does not seem to influence the RMSE metric, with our proposal displaying excellent results irrespective of the chosen shrinkage functional.

The same holds true when we look at the estimated covariance matrix of the random-effects: Figure 7 and 8 respectively report the average estimated $\boldsymbol{\Psi}$ and boxplots of the Frobenius distance between $\boldsymbol{\Psi}$ and $\hat{\boldsymbol{\Psi}}$ for different methods and scenarios. Particularly, the different sparsity patterns of $\boldsymbol{B}$ do not alter the recovery of the underlying random component, which is effectively attained by the three considered penalties. This also happens for the analogue of the Percentage of Variation due to Random Effects (PVRE) metric, displayed in Figure 9, in which it clearly emerges how the grouping impact differently affects the five response components.

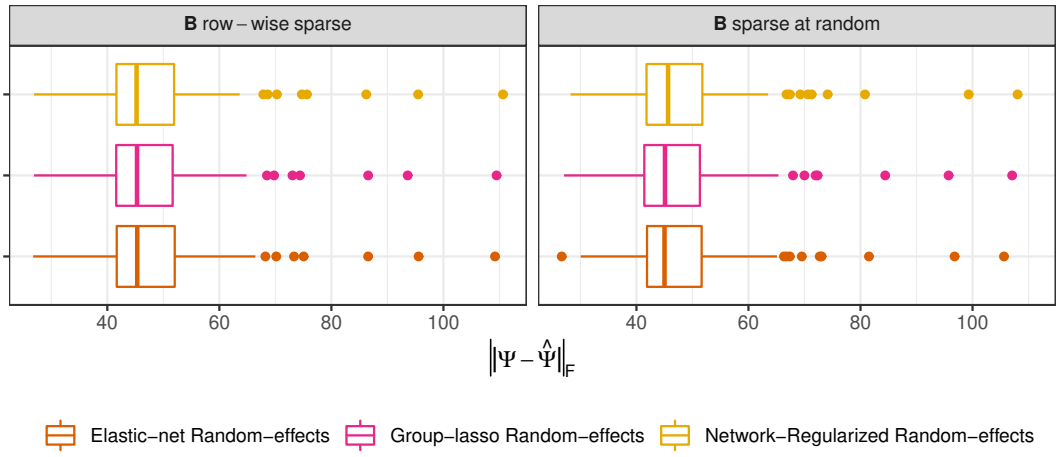

Figure 8: Boxplots of the Frobenius distance between true and estimated covariance matrix of the random-effects $\mathbf{\Psi}$ for $MC = 100$ repetitions of the simulated experiment.

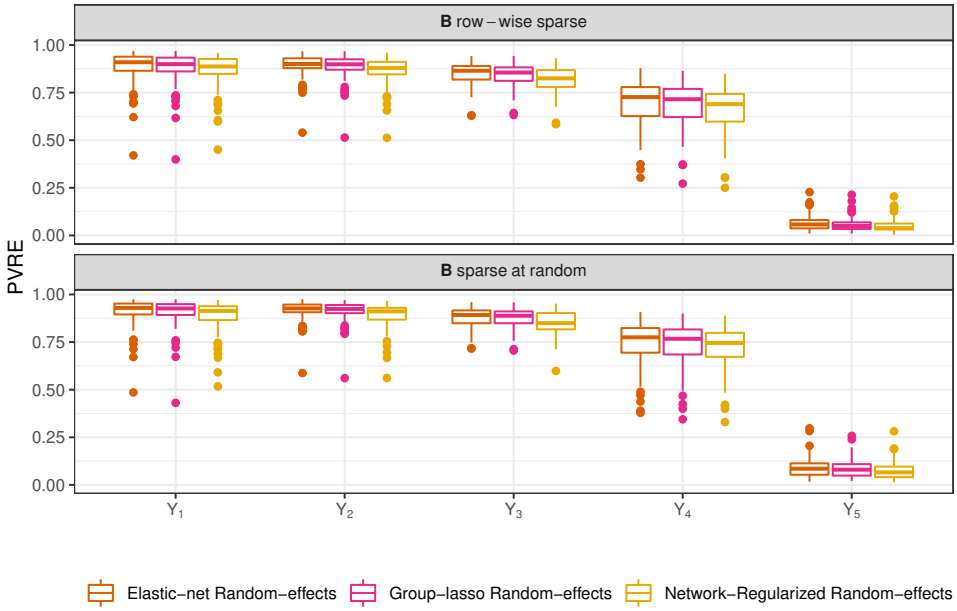

Figure 9: Boxplots of the Percentage of Variation due to Random Effects (PVRE) for $MC = 100$ repetitions of the simulated experiment.

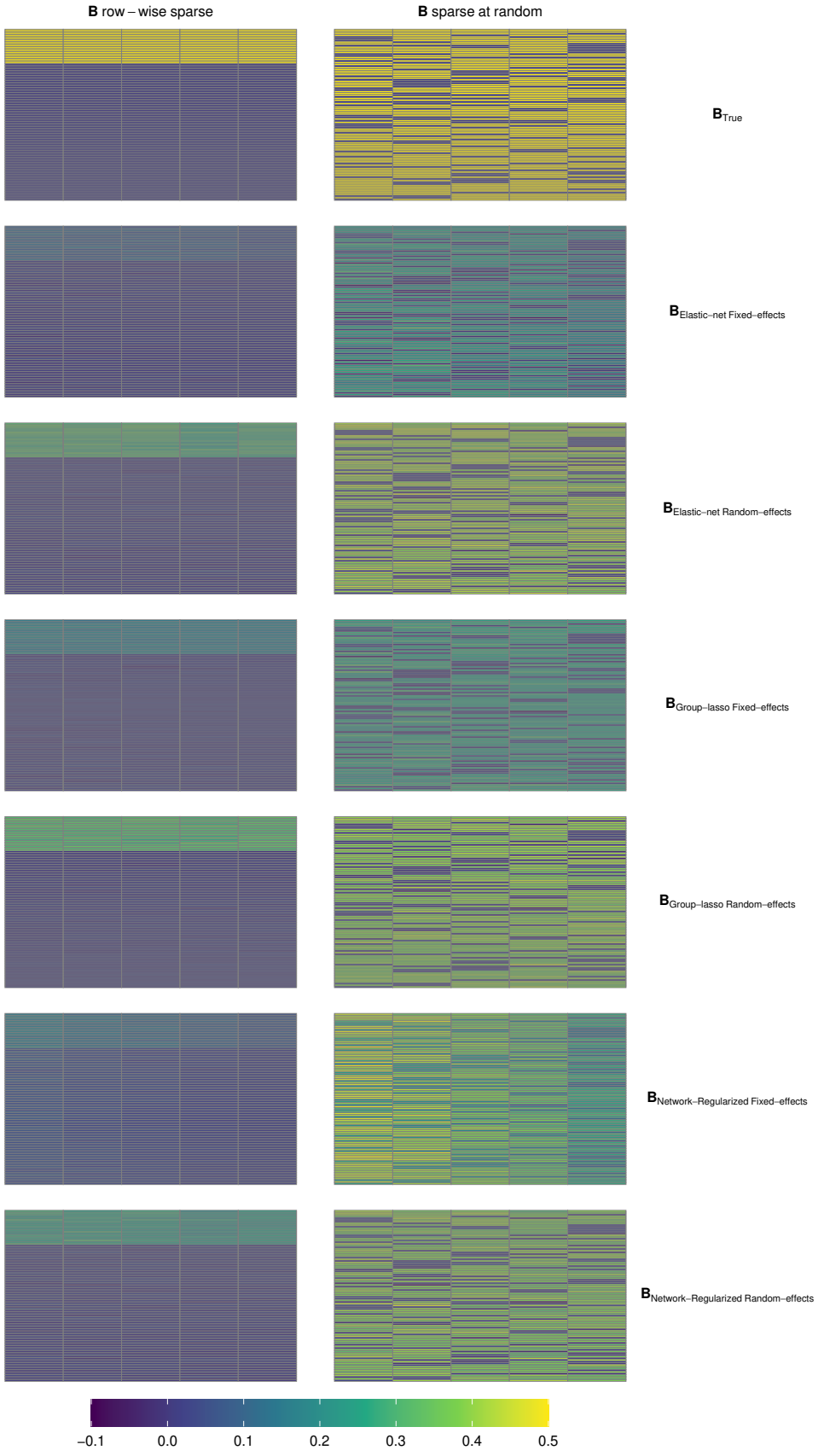

Figure 10: True and estimated matrices of fixed-effects $\boldsymbol{B}$, averaged over $MC = 100$ replication of the simulated experiment, for different methods (row-wise) and scenarios (column-wise).

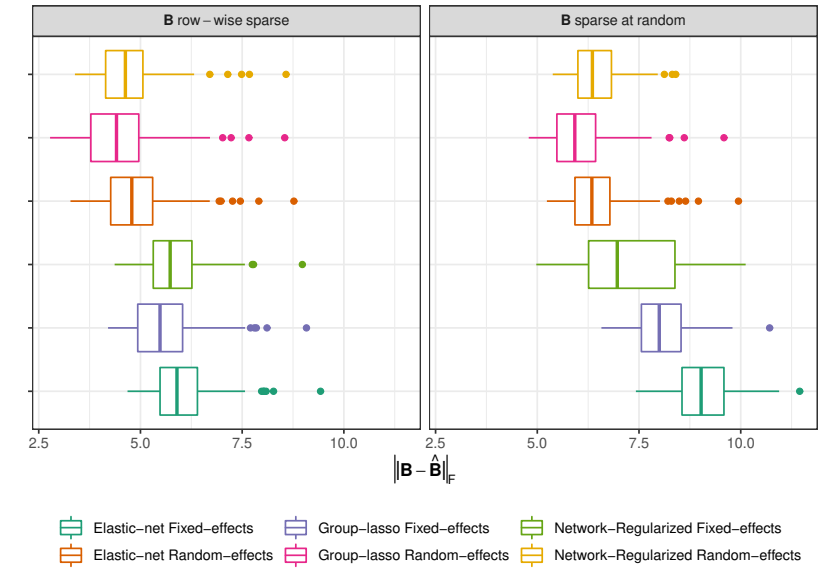

Figure 11: Boxplots of the Frobenius distance between true and estimated matrices of fixed-effects $\boldsymbol{B}$ for $MC = 100$ repetitions of the simulated experiment.

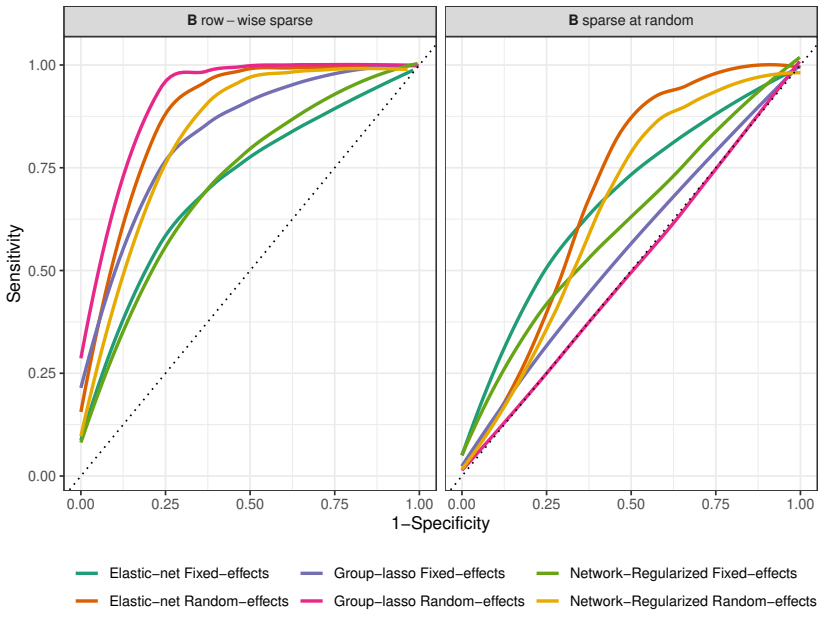

Figure 12: Receiver operating characteristic (ROC) curves averaged over $MC = 100$ replication of the simulated experiment, for different methods and scenarios. The ROCs are calculated as a function of the sparsity parameter λ.

As it may be expected, estimates for the fixed-effects matrix $\boldsymbol{B}$ differ in the two scenarios and markedly depend on the chosen penalty: Figure 10 and 11 report the average estimated $\boldsymbol{B}$ and boxplots of the Frobenius distance between $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$, respectively. First off, it is immediately noticed that methods without a random intercept showcase poorer performances than our proposals, regardless of the selected penalty and under both scenarios. Secondly, $\boldsymbol{B}$ *sparse at random* structure results in an on average higher Frobenius distance between the true and estimated $\boldsymbol{B}$ with respect to the $\boldsymbol{B}$ *row-wise sparse* case. Intuitively, the former scenario is more challenging than the latter and, while all penalty types can potentially accommodate a row-wise sparse $\boldsymbol{B}$, group-lasso regularizers only force entire rows of $\boldsymbol{B}$ to be shrunk to 0. It thus may seem surprising that the $||\boldsymbol{B} - \hat{\boldsymbol{B}}||_F$ metric displays lowest median values for the *Group-lasso Random-effects* model for both scenarios.

An explanation for this somewhat counterintuitive behavior is unraveled by looking at the recovery of the true underlying sparsity structure for $\boldsymbol{B}$ varying shrinkage factor λ. Figure 12 displays the receiver operating characteristic (ROC) curves of the different procedures for the two scenarios. In details, *Sensitivity* computes the proportion of zero entries in $\boldsymbol{B}$ correctly estimated as such over the total number of zeros in $\boldsymbol{B}$. Conversely, the *Specificity* of a solution amounts to the proportion of correctly recovered non-zero in $\boldsymbol{B}$ over its total number of non-zeros entries. By increasing the shrinkage factor λ penalized models gain in Sensitivity but lose Specificity, as it is demonstrated in the ROC curves of Figure 12. For the $\boldsymbol{B}$ *row-wise sparse* scenario (left panel) we notice that the *Group-lasso Random-effects* procedure outperforms all the other methods. While such an outcome seems logical given the sparsity structure sought by a group-lasso regularizer, it is interesting to observe that both *Elastic-net Random-effects* and *Network-Regularized Random-effects* ROC curves dominate the one associated to the *Group-lasso Fixed-effects* method; highlighting that a penalized mixed-effects modeling strategy, in presence of grouped data, not only increases the predictive accuracy but also improves the recovery of the sparse structure in the fixed-effects matrix. The zero entries of the $\boldsymbol{B}$ *sparse at random* structure are more difficult to be recovered, and all methods display lower values of both sensitivity and specificity (right panel of Figure 12). *Elastic-net Random-effects* and *Network-Regularized Random-effects* models perform better than their fixed-effects counterparts when higher values of λ are considered, while they tend to underestimate the number of zeros for moderate values of the shrinkage factor. As expected, methods coupled with a group-lasso penalty perform poorly under this scenario. Nonetheless, even though the sparsity pattern of $\boldsymbol{B}$ is not well-recovered the distance, in terms of Frobenius norm, between the estimated and the true matrix of fixed-effects does not suffer from the ill-posed penalty type, as featured in the boxplots of Figure 11.

6 Discussion and further work

In the present paper we have proposed a novel framework for mixed-effects multitask learning suitable for high-dimensional data. The ubiquitous presence in modern applications of "p bigger than N" problems asks for the development of ad-hoc statistical tools able to cope with such scenarios. By resorting to penalized likelihood estimation, we have devised a general purpose EM algorithm capable of accommodating any penalty type that has been previously defined for fixed-effects models. We have examined three functional forms for the penalty term, discussing pros and cons of each and providing convenient routines for model fitting. The proposal has been accompanied by some considerations on distinguishing features, like how to quantify response specific random-effects, and other more general issues concerning initialization, convergence and model selection.

The work has been motivated by the problem of developing a multivariate DNAm biomarker of cardiovascular and high blood pressure comorbidities from a multi-center sample. The EPIC Italy dataset has been analyzed using Diastolic Blood Pressure, Systolic Blood Pressure, High Density Lipoprotein, Low Density Lipoprotein and Triglycerides as response variables, regressing them on 13449 CpG sites and accounting for between-center heterogeneity. Our modeling framework, coupled with a group-lasso penalty, has demonstrated to outperform the state-of-the-art alternative, both in terms of predictive power and biomedical interpretation. Remarkably, the number of CpG sites deemed as relevant in the multi-dimensional surrogate creation was found to be lower than those identified by separately fitting penalized models for each risk factor. Decreasing the amount of relevant CpG sites is crucial to reduce sequencing costs for future studies, with the final aim of querying only a limited number of targeted genomic regions. Such a result may thereupon favor the adoption of our methodological approach for building DNAm surrogates.

A direction for future research concerns promoting the application of the proposed procedure in creating additional multi-dimensional DNAm biomarkers, conveniently embedding mixed-effects and customized penalty types. In addition, having assumed random intercepts for each and every component in a low-dimensional response framework was only motivated by the application at hand, and it may not be valid in general. Thus, a two-fold methodological development naturally arises: a first one concerning the definition of response-specific random-effects in multitask learning and another accounting for the inclusion of custom penalties when dealing with high-dimensional response variables. Furthermore, the latter may also possess a mixed-type structure, with components simultaneously be nominal, ordinal, discrete and/or continuous. Some proposals are currently under study and they will be the object of future work.

References

Atchley WR, Hall BK (1991a) A model for development and evolution of complex morphological structures. Biological Reviews of the Cambridge Philosophical Society 66(2):101–157

Atchley WR, Hall BK (1991b) A model for development and evolution of complex morphological structures. Biological Reviews 66(2):101–157

Battram T, Yousefi P, Crawford G, Prince C, Babaei MS, Sharp G, Hatcher C, Vega-Salas MJ, Khodabakhsh S, Whitehurst O, Langdon R, Mahoney L, Elliott HR, Mancano G, Lee MA, Watkins SH, Lay AC, Hemani G, Gaunt TR, Relton CL, Staley JR, Suderman M (2021) The EWAS Catalog: a database of epigenome-wide association studies. OSF Preprints p 4

Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, Jokubaitis VG, Lea RA (2021) Epigenome-wide association studies: current knowledge, strategies and recommendations. Clinical Epigenetics 13(1):214

Caruana R (1997) Multitask learning. Machine learning 28(1):41–75

Castro de Moura M, Davalos V, Planas-Serra L, Alvarez-Errico D, Arribas C, Ruiz M, Aguilera-Albesa S, Troya J, Valencia-Ramos J, Vélez-Santamaria V, Rodríguez-Palmero A, Villar-Garcia J, Horcajada JP, Albu S, Casasnovas C, Rull A, Reverte L, Dietl B, Dalmau D, Arranz MJ, Llucià-Carol L, Planas AM, Pérez-Tur J, Fernandez-Cadenas I, Villares P, Tenorio J, Colobran R, Martin-Nalda A, Soler-Palacin P, Vidal F, Pujol A, Esteller M (2021) Epigenome-wide association study of COVID-19 severity with respiratory failure. EBioMedicine 66:103339

Cheng W, Zhang X, Guo Z, Shi Y, Wang W (2014) Graph-regularized dual Lasso for robust eQTL mapping. Bioinformatics 30(12):139–148

Chipperfield JO, Steel DG (2012) Multivariate random effect models with complete and incomplete data. Journal of Multivariate Analysis 109:146–155

Chung FRK, Graham FC (1997) Spectral graph theory. 92, American Mathematical Soc.

Colicino E, Just A, Kioumourtzoglou MA, Vokonas P, Cardenas A, Sparrow D, Weisskopf M, Nie LH, Hu H, Schwartz JD, Wright RO, Baccarelli AA (2021) Blood DNA methylation biomarkers of cumulative lead exposure in adults. Journal of Exposure Science & Environmental Epidemiology 31(1):108–116

Conole ELS, Stevenson AJ, Green C, Harris SE, Maniega SM, Valdés-Hernández MdC, Harris MA, Bastin ME, Wardlaw JM, Deary IJ, Miron VE, Whalley HC, Marioni RE, Cox SR (2020) An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. medRxiv p 2020.10.08.20205245

Dawid AP (1981) Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application. Biometrika 68(1):265

Demidenko E (2013) Mixed models: theory and applications with R. John Wiley & Sons

Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39(1):1–22

Dirmeier S, Fuchs C, Mueller NS, Theis FJ (2018) NetReg: Network-regularized linear models for biological association studies. Bioinformatics 34(5):896–898

Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(5):849–911

Fan J, Samworth R, Wu Y (2009) Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. Journal of Machine Learning Research 10(2009):2013–2038

Fazzari MJ, Greally JM (2010) Introduction to Epigenomics and Epigenome-Wide Analysis, Humana Press, Totowa, NJ, pp 243–265

Fiorito G, Vlaanderen J, Polidoro S, Gulliver J, Galassi C, Ranzi A, Krogh V, Grioni S, Agnoli C, Sacerdote C, Panico S, Tsai MY, Probst-Hensch N, Hoek G, Herceg Z, Vermeulen R, Ghantous A, Vineis P, Naccarati A (2018) Oxidative stress and inflammation mediate the effect of air pollution on cardio- and cerebrovascular disease: A prospective study in nonsmokers. Environmental and Molecular Mutagenesis 59(3):234–246

Frohlich H, Zell A (2005) Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., IEEE, vol 3, pp 1431–1436

Gałecki A, Burzykowski T (2013) Linear Mixed-Effects Models Using R. Springer Texts in Statistics, Springer New York, New York, NY

Guida F, Sandanger TM, Castagné R, Campanella G, Polidoro S, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, Severi G, Kyrtopoulos SA, Georgiadis P, Vermeulen RC, Lund E, Vineis P, Chadeau-Hyam M (2015) Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Human Molecular Genetics 24(8):2349–2359

Hastie T, Tibshirani R, Wainwright M (2015) Statistical Learning with Sparsity. Chapman and Hall/CRC

Hillary RF, Marioni RE (2021) MethylDetectR: a software for methylation-based health profiling. Wellcome Open Research 5:283

Jordan MI (2013) On statistics, computation and scalability. Bernoulli 19(4):1378–1390

Kim S, Xing EP (2012) Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. The Annals of Applied Statistics 6(3):1095–1117

Kim S, Pan W, Shen X (2013) Network-Based Penalized Regression With Application to Genomic Data. Biometrics 69(3):582–593

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9(1):559

Laria JC, Carmen Aguilera-Morillo M, Lillo RE (2019) An Iterative Sparse-Group Lasso. Journal of Computational and Graphical Statistics 28(3):722–731

Li C, Li H (2010) Variable selection and regression analysis for graph-structured covariates with an application to genomics. The Annals of Applied Statistics 4(3):1498–1516

Li Y, Nan B, Zhu J (2015) Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. Biometrics 71(2):354–363

Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Assimes TL, Ferrucci L, Horvath S (2019) DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging 11(2):303–327

McLachlan GJ, Krishnan T (2008) The EM Algorithm and Extensions, 2E, Wiley Series in Probability and Statistics, vol 54. John Wiley & Sons, Inc., Hoboken, NJ, USA

Meng XL, Rubin DB (1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. Biometrika 80(2):267

Obozinski G, Taskar B, Jordan MI (2010) Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing 20(2):231–252

Obozinski G, Wainwright MJ, Jordan MI (2011) Support union recovery in high-dimensional multivariate regression. Annals of Statistics 39(1):1–47

Panico S, Dello Iacovo R, Celentano E, Galasso R, Muti P, Salvatore M, Mancini M (1992) Progetto ATENA, A study on the etiology of major chronic diseases in women: Design, rationale and objectives. European Journal of Epidemiology 8(4):601–608

Pinheiro J, Bates D (2006) Mixed-effects models in S and S-PLUS. Springer science & business media

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, Merico D, Bader GD (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols 14(2):482–517

Reinsel G (1984) Estimation and Prediction in a Multivariate Random Effects Generalized Linear Model. Journal of the American Statistical Association 79(386):406–414

Riboli E, Hunt K, Slimani N, Ferrari P, Norat T, Fahey M, Charrondière U, Hémon B, Casagrande C, Vignat J, Overvad K, Tjønneland A, Clavel-Chapelon F, Thiébaut A, Wahrendorf J, Boeing H, Trichopoulos D, Trichopoulou A, Vineis P, Palli D, Bueno-de Mesquita H, Peeters P, Lund E, Engeset D, González C, Barricarte A, Berglund G, Hallmans G, Day N, Key T, Kaaks R, Saracci R (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutrition 5(6b):1113–1124

Richard MA, Huan T, Ligthart S, Gondalia R, Jhun MA, Brody JA, Irvin MR, Marioni R, Shen J, Tsai PC, Montasser ME, Jia Y, Syme C, Salfati EL, Boerwinkle E, Guan W, Mosley TH, Bressler J, Morrison AC, Liu C, Mendelson MM, Uitterlinden AG, van Meurs JB, Franco OH, Zhang G, Li Y, Stewart JD, Bis JC, Psaty BM, Chen YDI, Kardia SL, Zhao W, Turner ST, Absher D, Aslibekyan S, Starr JM, McRae AF, Hou L, Just AC, Schwartz JD, Vokonas PS, Menni C, Spector TD, Shuldiner A, Damcott CM, Rotter JI, Palmas W, Liu Y, Paus T, Horvath S, O'Connell JR, Guo X, Pausova Z, Assimes TL, Sotoodehnia N, Smith JA, Arnett DK, Deary IJ, Baccarelli AA, Bell JT, Whitsel E, Dehghan A, Levy D, Fornage M, Heijmans BT, 't Hoen PA, van Meurs J, Isaacs A, Jansen R, Franke L, Boomsma DI, Pool R, van Dongen J, Hottenga JJ, van Greevenbroek MM, Stehouwer CD, van der Kallen CJ, Schalkwijk CG, Wijmenga C, Zhernakova A, Tigchelaar EF, Slagboom PE, Beekman M, Deelen J, van Heemst D, Veldink JH, van den Berg LH, van Duijn CM, Hofman A, Uitterlinden AG, Jhamai PM, Verbiest M, Suchiman HED, Verkerk M, van der Breggen R, van Rooij J, Lakenberg N, Mei H, van Iterson M, van Galen M, Bot J, van 't Hof P, Deelen P, Nooren I, Moed M, Vermaat M, Zhernakova DV, Luijk R, Bonder MJ, van Dijk F, Arindrarto W, Kielbasa SM, Swertz MA, van Zwet EW (2017) DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. The American Journal of Human Genetics 101(6):888–902

Rohart F, San Cristobal M, Laurent B (2014) Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. Computational Statistics & Data Analysis 80:209–222

Schafer JL, Yucel RM (2002) Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. Journal of Computational and Graphical Statistics 11(2):437–457

Schelldorfer J, Bühlmann P, De Geer SV (2011) Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. Scandinavian Journal of Statistics 38(2):197–214

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464

Shah A, Laird N, Schoenfeld D (1997) A Random-Effects Model for Multiple Characteristics with Possibly Missing Data. Journal of the American Statistical Association 92(438):775–779

Sill M, Hielscher T, Becker N, Zucknick M (2014) c060 : Extended Inference with Lasso and Elastic-Net Regularized Cox and Generalized Linear Models. Journal of Statistical Software 62(5)

Simon N, Friedman J, Hastie T, Tibshirani R (2013) A Sparse-Group Lasso. Journal of Computational and Graphical Statistics 22(2):231–245

Singal R, Ginder GD (1999) DNA Methylation. Blood 93(12):4059–4070

Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, Walker RM, Evans KL, Boutin TS, Hayward C, McRae AF, McColl BW, Spires-Jones TL, McIntosh AM, Deary IJ, Marioni RE (2020) Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. Clinical Epigenetics 12(1):113

Tay JK, Narasimhan B, Hastie T (2021) Elastic Net Regularization Paths for All Generalized Linear Models

Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1):267–288

Tyler AL, Crawford DC, Pendergrass SA (2013) Detecting and characterizing pleiotropy: new methods for uncovering the connection between the complexity of genomic architecture and multiple phenotypes. In: Biocomputing 2014, WORLD SCIENTIFIC, pp 183–187

Vinga S (2021) Structured sparsity regularization for analyzing high-dimensional omics data. Briefings in Bioinformatics 22(1):77–87

Wu CY, Hu HY, Chou YJ, Huang N, Chou YC, Li CP (2015) High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults. Medicine 94(47):e2160

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1):49–67

Zhang Y, Elgizouli M, Schöttker B, Holleczek B, Nieters A, Brenner H (2016) Smoking-associated DNA methylation markers predict lung cancer incidence. Clinical Epigenetics 8(1):1–12

Zhao Z, Zucknick M (2020) Structured penalized regression for drug sensitivity prediction. Journal of the Royal Statistical Society: Series C (Applied Statistics) 69(3):525–545

Zhong J, Agha G, Baccarelli AA (2016) The Role of DNA Methylation in Cardiovascular Risk and Disease. Circulation Research 118(1):119–131

Zhong W, Wang J, Chen X (2021) Censored mean variance sure independence screening for ultrahigh dimensional survival data. Computational Statistics & Data Analysis 159:107206

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(5):768–768

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

20/2022 Clementi, L.; Gregorio, C; Savarè, L.; Ieva, F; Santambrogio, M.D.; Sangalli, L.M.
A Functional Data Analysis Approach to Left Ventricular Remodeling Assessment

18/2022 Bennati, L; Vergara, C; Giambruno, V; Fumagalli, I; Corno, A.F; Quarteroni, A; Puppini, G; L
An image-based computational fluid dynamics study of mitral regurgitation in presence of prolapse

19/2022 Lupo Pasini, M.; Perotto, S.
Hierarchical model reduction driven by machine learning for parametric advection-diffusion-reaction problems in the presence of noisy data

17/2022 Regazzoni, F.
Stabilization of staggered time discretization schemes for 0D-3D fluid-structure interaction problems

14/2022 Zappon, E.; Manzoni, A.; Quarteroni A.
Efficient and certified solution of parametrized one-way coupled problems through DEIM-based data projection across non-conforming interfaces

15/2022 G. Ciaramella, T. Vanzan
Spectral coarse spaces for the substructured parallel Schwarz method

16/2022 G. Ciaramella, T. Vanzan
Substructured Two-grid and Multi-grid Domain Decomposition Methods

13/2022 Grasselli, M.; Parolini, N.; Poiatti, A.; Verani, M.
Non-isothermal non-Newtonian fluids: the stationary case

12/2022 Antonietti, P.F.; Dassi, F.; Manuzzi, E.
Machine Learning based refinement strategies for polyhedral grids with applications to Virtual Element and polyhedral Discontinuous Galerkin methods

09/2022 Corti, M.; Zingaro, A.; Dede', L.; Quarteroni, A.
Impact of Atrial Fibrillation on Left Atrium Haemodynamics: A Computational Fluid Dynamics Study