The Annals of Applied Statistics
0, Vol. 0, No. 00, 1–26
https://doi.org/10.1214/23-AOAS1760

A GENERAL FRAMEWORK FOR PENALIZED MIXED-EFFECTS MULTITASK LEARNING WITH APPLICATIONS ON DNA METHYLATION SURROGATE BIOMARKERS CREATION

BY ANDREA CAPPOZZO[1,a], FRANCESCA IEVA[1,b] AND GIOVANNI FIORITO[2,c]

[1]*MOX Lab, Department of Mathematics, Politecnico di Milano,* [a]*andrea.cappozzo@polimi.it,* [b]*francesca.ieva@polimi.it*
[2]*Clinical Bioinformatics Unit, IRCCS Istituto Giannina Gaslini,* [c]*giovannifiorito@gaslini.org*

Recent evidence highlights the usefulness of DNA methylation (DNAm) biomarkers as surrogates for exposure to risk factors for noncommunicable 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 multicenter study. A penalized estimation scheme, based on an expectation-maximization 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 biomolecular 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)). DNAm regulates gene expression in different manners. Specifically, high DNAm has been observed in bodies of highly transcribed genes, whereas DNAm in gene promoters and first introns typically have an inverse correlation with gene expression (Anastasiadi, Esteve-Codina and Piferrer (2018), Rauluseviciute, Drabløs and Rye (2020)). Also, recent studies suggest that the relationship between genetic variation, DNAm and gene expression is complex and tissue-specific, highlighting that DNAm in non-CpG island regions regulates the transcription of distal genes (van Eijk et al. (2012)). Advanced technology allows measuring whole-genome DNAm for many samples at the same time. The most common ways for DNAm measurements consist of whole-genome bisulphite sequencing and DNAm microarray. The first commercial high-density microarray measuring genome-wide methylation was the HumanMethylation27 (27K CpGs) released by Illumina in 2009, followed by the HumanMethylation450 (450K CpGs) and, more recently, by the IlluminaMethylation850 (850K CpGs, Campagna et al. (2021)). Since then, a tremendous amount of associations between DNAm at individual CpG sites and different exposures, traits and diseases have been identified in the so-called epigenome-wide association studies (EWAS, Battram et al. (2022)). Concurrently, the development of surrogate scores, based on blood

Received April 2022; revised February 2023.
Key words and phrases. Mixed-effects models, multitask learning, EM algorithm, penalized estimation, multivariate regression, personalized medicine.

DNA methylation, has also received thriving attention in recent years: impressive epidemiological evidence has been established between DNAm and individual history of exposure to lifestyle and environmental risk factors (Zhong, Agha and Baccarelli (2016), 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 a 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 very recent contributions, including surrogate scores for short-term risk of cardiovascular events (Cappozzo et al. (2022)), cumulative lead exposure (Colicino et al. (2021)), DNAm surrogate for alcohol consumption, obesity indexes and blood measured inflammatory proteins (Hillary and Marioni (2020)), 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 afore-said first layer of complexity is encountered when a multidimensional 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 multitask 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, Taskar and Jordan (2010), Obozinski, Wainwright and Jordan (2009), Li, Nan and Zhu (2015)) that extend the lasso (Tibshirani (1996)), group-lasso (Yuan and Lin (2006)) and sparse group-lasso (Laria, Carmen Aguilera-Morillo and Lillo (2019), Simon et al. (2013)) 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, Pan and Shen (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), Zhao et al. (2022)). 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, Tibshirani and Wainwright (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 cohortwise 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). While 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, Bühlmann and van de Geer (2011) introduced a procedure for estimating high-dimensional LMM via an ℓ_1-penalization. More recently, Rohart, San Cristobal and Laurent (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, Laird and Schoenfeld (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 and 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 multicenter study, we propose in this article a general framework for high-dimensional multitask learning with random effects. Leveraging from the algorithm introduced in Rohart, San Cristobal and Laurent (2014) for the univariate response case, the estimation 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 presents a simulation study on synthetic data for three different scenarios. Section 5 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 6 concludes the paper with a discussion and directions for future 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 branch 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 geographical subcohorts identified by the center of recruitment; particularly, we will consider 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)). DNAm was measured with the HumanMethylation450 array, following standard laboratory procedures (see Fiorito et al. (2022), for a detailed description), while the preprocessing included removing CpG sites and samples with a call rate lower than 95%, BMIQ method for reducing technical variability and bias introduced by type II probes and ComBat technique for batch effect adjustment (Marabita et al. (2013)).

By profiting from the information recorded in the aforementioned subcohorts, we aim at creating a multidimensional 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 are 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 values for each CpG site, adjusted for sex and age. A total of $N = 574$ individuals in the $J = 4$ cohorts showcase nonmissing values for every response variable: they comprise the sample onto which all subsequent analyses will be performed. To reconstruct the process of DNAm surrogates creation and validation, the EPIC Italy data is randomly split into two sets: 70% ($N_{\mathrm{tr}} = 401$) of it is employed for preprocessing and model fitting, while the remaining 30% ($N_{te} = 173$) acts as test set for assessing prediction accuracy. In addition, we will consider samples from the EXPOsOMICS project (Fiorito et al. (2018)) as an external validation dataset to assess out of groups predictive performance. In details EXPOsOMICS is a case-control study on cardiovascular diseases (CVDs) nested in the EPIC Italy cohort, composed by 276 volunteers (not overlapping with the main dataset), whose center of recruitment is unknown or different from the $J = 4$ observed in the learning phase.

Coming back to the data analysis pipeline, an epigenome-wide association study (EWAS, Campagna et al. (2021)) is performed on the training set as a pre-screening procedure. In details log-transformed DBP, SBP, HDL, LDL and TG are separately regressed on each available CpG site, adjusting for sex and age. P-values are then collected and arranged in increasing order. We then screen the set of predictors retaining, for each dimension of the multivariate response, the CpG sites whose p-values are smaller than the fifth percentile of the resulting empirical distributions. The final set of covariates for the multitask learning problem is achieved by taking the union of the resulting CpG sites separately preserved for DBP, SBP, HDL, LDL and TG. In so doing, out of the whole initial set of 295,614 CpG sites, 62,128 DNA methylation features are retained for subsequent modeling. Together with sex and age, this amounts to a total of $p = 62{,}130$ predictors and a *five*-dimensional response for a training sample size of $N_{\mathrm{tr}} = 401$. While variable screening in ultra-high feature space is itself an ongoing research field (see, e.g., Fan and Lv (2008), Zhong, Wang and Chen (2021), Fan, Samworth and Wu (2009), and references therein), we decided to rely on the EWAS technique, as it is the standard approach employed in epigenomics (Fazzari and Greally (2010)).

As previously mentioned, the considered training samples belong to four different centers distributed across Italy, with data for 91, 234, 44 and 32 volunteers, 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 centerwise variability and to maintain generalizability of the devised DNAm biomarker outside the Italy EPIC cohorts, a partial pooling random-intercept model shall 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 specification for a multivariate mixed-effects framework applicable to high-dimensional predictors.

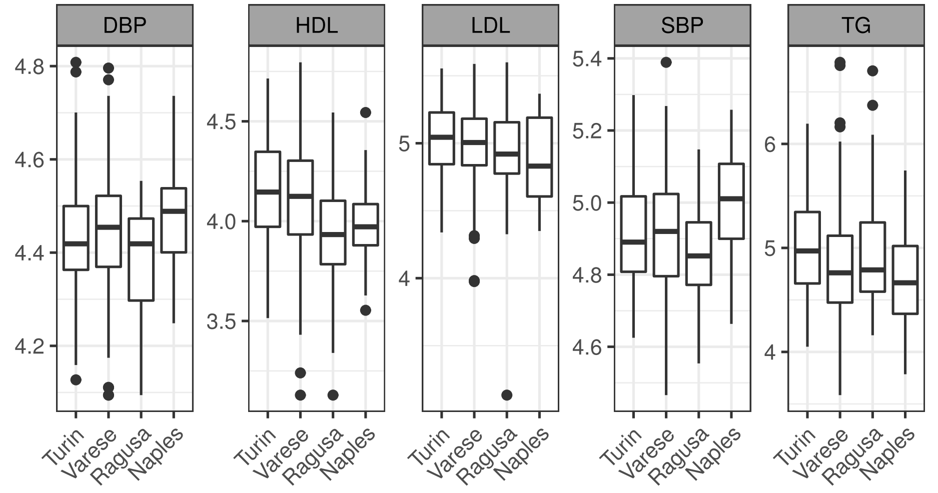

FIG. 1. *Boxplots of log-transformed diastolic blood pressure (DBP), high-density lipoprotein (HDL), low-density lipoprotein (LDL), systolic blood pressure (SBP) and triglycerides (TG) for different Center, Italy EPIC training dataset.*

3. Penalized mixed-effects model 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, Laird and Schoenfeld (1997)) expresses the $n_j \times r$ response matrix $\boldsymbol{Y}_j$ for the jth 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 samples 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.

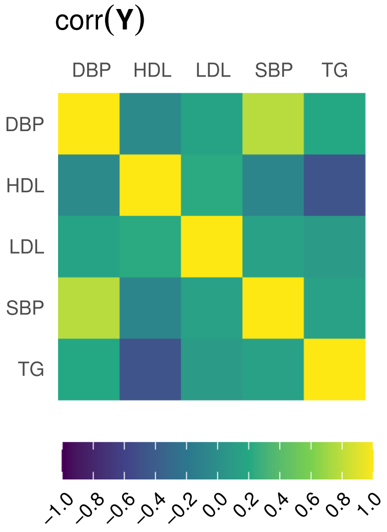

FIG. 2. *Sample correlation matrix of log-transformed diastolic blood pressure (DBP), high-density lipoprotein (HDL), low-density lipoprotein (LDL), systolic blood pressure (SBP) and triglycerides (TG), Italy EPIC training dataset.*

- $\boldsymbol{X}_j$ 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.
- $j = 1, \dots, J$, with J total number of groups.

By employing the vec operator, we assume that

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

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:

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

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,

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

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

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} & -\frac{n_j}{2}\log 2\pi - \frac{1}{2}\log\big|(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\big| \\ & - \frac{1}{2}\big(\operatorname{vec}(\boldsymbol{Y}_j) - (\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B})\big)'\big((\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\big)^{-1} \\ & \times \big(\operatorname{vec}(\boldsymbol{Y}_j) - (\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B})\big), \end{aligned} \tag{3}$$

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:

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

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 $\operatorname{vec}(\boldsymbol{\Lambda}_j)$, $j = 1, \dots, J$ are unknown. We, therefore, devise an EM algorithm (Dempster, Laird and Rubin (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 $\mathbb{E}(\operatorname{vec}(\boldsymbol{\Lambda}_j)|\boldsymbol{Y}_j; \boldsymbol{\theta})$ and $\mathbb{E}(\operatorname{vec}(\boldsymbol{\Lambda}_j)\operatorname{vec}(\boldsymbol{\Lambda}_j)'|\boldsymbol{Y}_j; \boldsymbol{\theta})$. This is achieved by noticing that the conditional density

$p(\mathrm{vec}(\boldsymbol{\Lambda}_j)|\boldsymbol{Y}_j;\boldsymbol{\theta})$ is Normal. Updating formulae for the quantities of interest are thus derived as follows:

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

$$\begin{aligned}\widehat{\mathrm{vec}(\boldsymbol{\Lambda}_j)} &= \mathbb{E}\big(\mathrm{vec}(\boldsymbol{\Lambda}_j)|\boldsymbol{Y}_j;\boldsymbol{\theta}\big)\\ &= \hat{\boldsymbol{\Gamma}}_j(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)'(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j})^{-1}\big(\mathrm{vec}(\boldsymbol{Y}_j) - (\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\,\mathrm{vec}(\boldsymbol{B})\big).\end{aligned} \tag{6}$$

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

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

At the tth 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

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

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*,

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

where $\boldsymbol{e}_j = \mathrm{vec}(\boldsymbol{Y}_j) - (\boldsymbol{I}_r\otimes\boldsymbol{X}_j)\,\mathrm{vec}(\boldsymbol{B}) - (\boldsymbol{I}_r\otimes\boldsymbol{Z}_j)\,\mathrm{vec}(\boldsymbol{\Lambda}_j)$ 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

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

where $\tilde{\boldsymbol{Y}}_j = \boldsymbol{Y}_j - \boldsymbol{Z}_j\hat{\boldsymbol{\Lambda}}_j$. $\hat{\boldsymbol{\Lambda}}_j$ is recovered by applying the inverse of the vectorization operator to $\widehat{\mathrm{vec}(\boldsymbol{\Lambda}_j)}$, previously computed in the E-step. Simply put, the $\widehat{\mathrm{vec}(\boldsymbol{\Lambda}_j)}$ vector of length qr is rearranged in a $q\times r$ matrix, obtaining $\hat{\boldsymbol{\Lambda}}_j$. 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, Laird and Schoenfeld (1997)). By exploiting properties of the trace operator, we can rewrite (10) defining the following minimization problem:

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

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-effects models for estimating $\boldsymbol{B}$. In details for solving (11), a two-step updating scheme is devised. First, we compute

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

that is, a fixed-effects penalized regression problem in which the response variable is $\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{Y}}_j,\ j = 1, \dots, J$; $\tilde{\boldsymbol{B}}$ is thus easily retrieved via fixed-effects routines for penalized estimation. Second, the solution to (11) is obtained postmultiplying $\tilde{\boldsymbol{B}}$ by $\boldsymbol{\Sigma}^{1/2}$. Therefore, at each iteration of the EM-algorithm we first compute $\tilde{\boldsymbol{B}}$, 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, San Cristobal and Laurent (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 setting, 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 $\boldsymbol{\Sigma}$

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

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

3.3. *Definition 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. 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 multivariate group-lasso penalty (Obozinski, Wainwright and Jordan (2011b)) and the netReg routines for Network-regularized linear models (Dirmeier et al. (2018)). Each of them is briefly described in the next subsections.

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, $0 \le \alpha \le 1$. 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 lth row and cth column of matrix $\boldsymbol{B}$. Notice that the first row of $\boldsymbol{B}$ contains the r intercepts, and it is thus not penalized. Algorithmically, penalty

(16) can be enforced employing standard and widely available routines for univariate penalized estimation, like the `glmnet` software (Tay, Narasimhan and Hastie (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. The latter can also be employed in our framework to force coefficients that need not be penalized to enter the model specification.

3.3.2. *Multivariate 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:

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

where $\boldsymbol{b}_{l.}$ identifies the lth row of the matrix $\boldsymbol{B}$ such that each $\boldsymbol{b}_{l.}$, $l=2,\dots,p$ 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 nonnegative 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:

$$\begin{aligned}p(\boldsymbol{B};\lambda,\lambda_X,\lambda_Y)&=\lambda\|\boldsymbol{B}_0\|_1+\lambda_X\operatorname{tr}\bigl(\boldsymbol{B}_0'(\boldsymbol{D}_{G_X}-\boldsymbol{G}_X)\boldsymbol{B}_0\bigr)\\&\quad+\lambda_Y\operatorname{tr}\bigl(\boldsymbol{B}_0(\boldsymbol{D}_{G_Y}-\boldsymbol{G}_Y)\boldsymbol{B}_0'\bigr),\end{aligned} \tag{18}$$

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. The `netReg R` package (Dirmeier et al. (2018)) provides a convenient implementation of (18).

3.3.4. *On the choice of* $p(\boldsymbol{B};\lambda)$. Leaving the flexibility attained by the methodology proposed in Section 3.1 aside, in practice, a functional form for $p(\boldsymbol{B};\lambda)$ must be chosen when performing the analysis. Hereafter, we highlight pros and cons of the proposed approaches with respect to a mixed-effects multitask learning setting.

The elastic-net penalty in (16) does not take into account the multivariate nature of the problem in (4), as the shrinkage is applied directly to $\operatorname{vec}(\boldsymbol{B})$. 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 could be desirable in a multivariate context. Differently, the multivariate group-lasso of Section 3.3.2 defines a shrinkage term that forces the same subset of predictors to be preserved across all r components of the response $\boldsymbol{Y}$. This can be seen as the generalization of the variable selection problem to the multivariate response setting, which is also known as *support union problem* or *row selection problem* in the literature (Obozinski, Wainwright and Jordan (2011b)). Lastly, the network-Regularized penalty in

3.3.3 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)).

In relation to the DNAm surrogate creation task motivating our methodological proposal, the multivariate group-lasso is definitely the most appropriate penalty, as it not only showcases better prediction performances but it is also supported by biological reasons: a thorough analysis for the EPIC dataset is reported in Section 5.

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 $\hat{\boldsymbol{\theta}}^{(0)} = \{\hat{\boldsymbol{B}}^{(0)}, \hat{\boldsymbol{\Sigma}}^{(0)}, \hat{\boldsymbol{\Psi}}^{(0)}\}$. 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 hyperparameters.
- *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 $\varepsilon > 0$,

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

  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 tth 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, San Cristobal and Laurent (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 nonzero 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 with which our procedure can cope as well as the overall computing time very much depends on the scalability and efficiency associated to the chosen shrinkage term. Typically nevertheless, penalized likelihood approaches fail to be directly applied to ultrahigh-dimensional problems (Fan, Samworth and Wu (2009)), and preprocessing procedures, such as variable screening, are thus required prior to modeling. The epigenetic application that motivated our work naturally called for an EWAS prescreening 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 (2022)), and the source code is freely available in the Supplementary Material and 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. While 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 $\operatorname{diag}(\hat{\boldsymbol{\Psi}})$ is significantly lower than the remaining ones. Doing this way, the impact random-effects have on the different characteristics is retrieved as a by-product of the modeling procedure.

4. 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.

4.1. *Experimental setup.*

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

$$\boldsymbol{\Psi} = \begin{bmatrix} 50.00 & -1.59 & -0.60 & -0.22 & 2.38 \\ -1.59 & 40.00 & -0.96 & -0.91 & 0.37 \\ -0.60 & -0.96 & 30.00 & -0.43 & 0.50 \\ -0.22 & -0.91 & -0.43 & 20.00 & 0.80 \\ 2.38 & 0.37 & 0.50 & 0.80 & 0.16 \end{bmatrix},$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.16 & 0.09 & -0.80 & 0.91 & -0.26 \\ 0.09 & 2.16 & -0.33 & 0.55 & -0.10 \\ -0.80 & -0.33 & 2.16 & -0.03 & -0.13 \\ 0.91 & 0.55 & -0.03 & 2.16 & 0.02 \\ -0.26 & -0.10 & -0.13 & 0.02 & 2.16 \end{bmatrix},$$

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. Further, the error variances (diagonal elements of $\boldsymbol{\Sigma}$) are held constant across dimensions to better highlight the impact the variability of the random-effects has on the models performance. The data-generating process assumes 10 equally-sized subpopulations, resulting in $J = 10$. The matrix of fixed-effects $\boldsymbol{B}$ is of dimension $10{,}001 \times 5$, with distinct sparsity pattern according to three scenarios:

- $\boldsymbol{B}$ *rowwise sparse*: $\boldsymbol{B}$ has entries equal to 0.5 for the first 100 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.

- ***B** with dependence structure*: $\boldsymbol{B}$ has entries whose magnitude agrees with the correlation structure between the covariates, inducing coefficients to be similar when the absolute correlation between two predictors is high.

Lastly, $\boldsymbol{Z}_j$ is an all-one column vector $\forall j = 1, \dots, 10$, 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. The remaining 10,000 dimensions are generated according to a normal random vector with independent marginals for the ***B** rowwise sparse* and ***B** sparse at random* scenarios, while the `cormat_from_triangle` function from the `faux` package (DeBruine (2021)) has been used to simulate correlated predictors in the ***B** with dependence structure* experiment.

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. Notice that, as to mimic the process of DNAm surrogates creation, the total number of variables ($p = 10{,}001$) is much larger than the sample size. Seven different models, varying λ within a grid, are fitted on the training data:

- *Univariate elastic-net fixed-effects:* Univariate elastic-net regression, obtained fitting independent models to each dimension of the multivariate response.
- *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 multivariate 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 MLMM methodology introduced in the paper with elastic-net regularization (Section 3.3.1).
- *Group-lasso random-effects:* The penalized MLMM methodology introduced in the paper with group-lasso regularization (Section 3.3.2).
- *Network-regularized random-effects:* The penalized MLMM methodology introduced in the paper with edge-based regularization (Section 3.3.3).

Such an extensive comparison can be regarded as performing a within-scenario ablation study in which we start from a complex method, and we subsequently remove the random-effects component and, finally, the borrow strength property of multivariate regression to be left with univariate elastic-net fixed-effects models. In this way we investigate the contribution of our proposal to the overall system. 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 *five*-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. The predictive performance is then evaluated on the test set. Lastly, to assess out of groups prediction, models are further validated on 100 external samples, generated according to (1), coming from five extra subpopulations not observed in the training set. The devised simulated experiment is replicated $MC = 100$ times: results are reported in the next subsection.

4.2. *Simulation results.* Figure 3 displays boxplots of the Root Mean Squared Error (RMSE) computed for each component of the *five*-dimensional response on the test set. For all scenarios we observe that the componentwise predictive performance is heavily affected by the magnitude of the related diagonal entry in the $\boldsymbol{\Psi}$ matrix. When the grouping effect

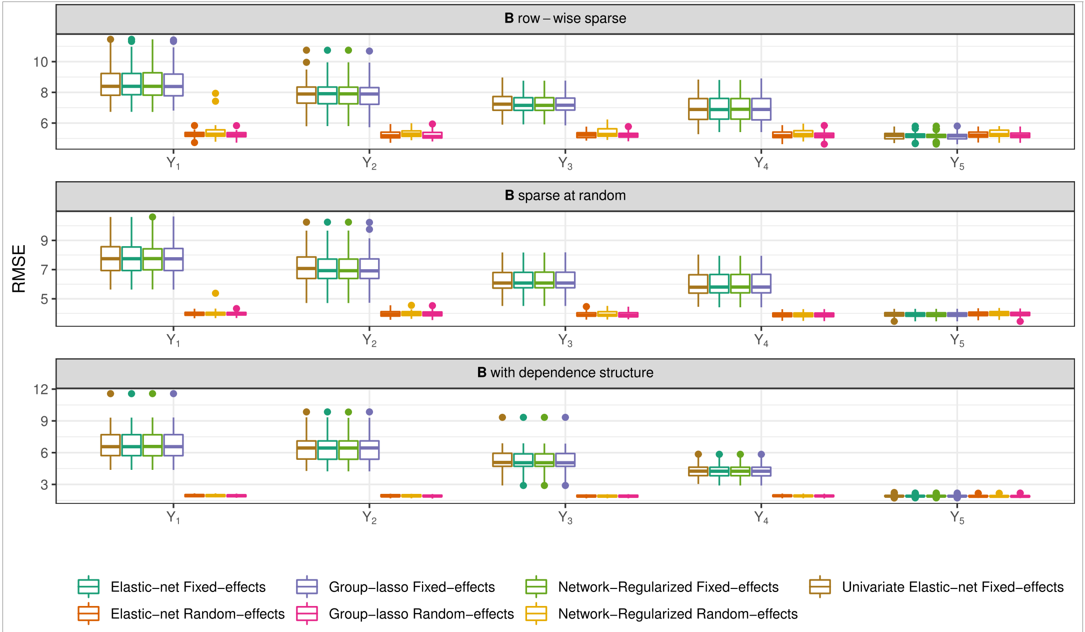

FIG. 3. *Boxplots of the Root Mean Squared Error (RMSE) for $MC = 100$ repetitions of the simulated experiment. RMSE is computed on 200 test points for different methods and three scenarios varying sparsity pattern for $\boldsymbol{B}$.*

is negligible (fifth dimension Y_5), all methods showcase comparable predictive performance. 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 for all scenarios. On the other hand, when it comes to perform out of groups prediction, the gain achieved by including random-effects decreases, and the outcome of models with fixed and mixed-effects are fairly similar. In details for the latter class of methods, the unconditional (population level) intercepts are employed when making predictions for unobserved groups. Notwithstanding, we recognize that results are no worse than those obtained with fixed-effects procedures, corroborating the generalizability of our proposals in external cohorts.

Figure 4 displays the analogue of the percentage of variation due to random effects (PVRE) metric, computed taking the ratio between the diagonal elements of $\hat{\boldsymbol{\Psi}}$ and the sum of the diagonals of $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Sigma}}$. From the plot, it clearly emerges how the grouping impact differently affects the variability in the five components of the response.

Figure 5 reports boxplots of the Frobenius distance between true and estimated matrices of fixed-effects $\boldsymbol{B}$. When looking at $\|\boldsymbol{B} - \hat{\boldsymbol{B}}\|_F$ under the three scenarios, we observe some interesting facts. First off, it is immediately noticed that the *univariate elastic-net fixed-effects* model showcases the poorest performance, in particular, for the $\boldsymbol{B}$ *with dependence structure* experiment. This is due to the fact that the different components of the response vector are related in our simulated specification, and, therefore, fitting separate regression models results in a loss of quality for the estimator. Second, we observe that for the $\boldsymbol{B}$ *rowwise sparse* scenario the *group-lasso random-effects* model is the best performing one among all the competitors, displaying the lowest median distance to $\boldsymbol{B}$. This may be expected, as such method is precisely constructed to identify a matrix of fixed-effects with a rowwise sparsity pattern. Furthermore, notice that the performance of the *group-lasso random-effects* is slightly better

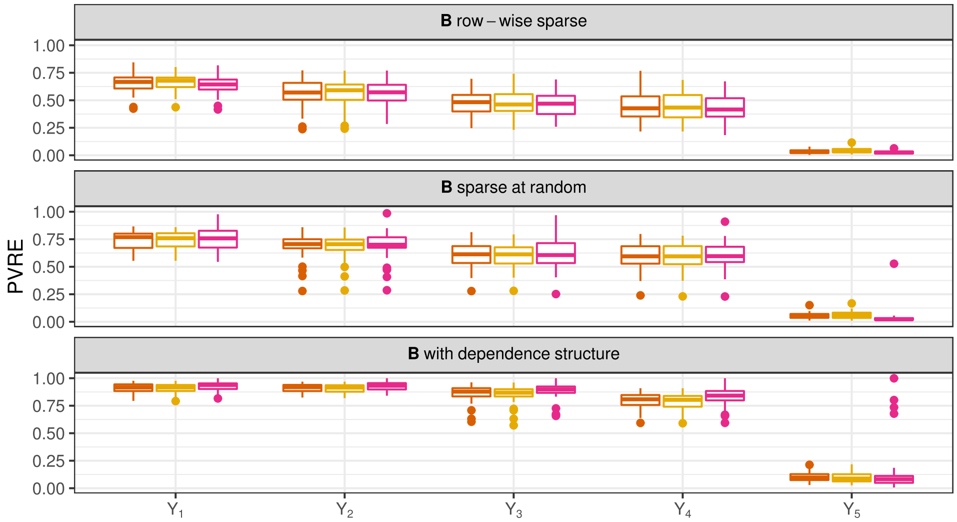

FIG. 4. *Boxplots of the Percentage of Variation due to Random Effects (PVRE) for $MC = 100$ repetitions of the simulated experiment.*

than its fixed-effects counterpart. For the remaining scenarios, the superiority of the mixed-effects procedures is not so apparent, and both fixed and random-effects models demonstrate a comparable performance. An only modest gain is showcased by the *network-regularized random-effects* method for which the inclusion of the adjacency matrices $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$ in the penalty specification helps in better recovering the $\boldsymbol{B}$ structure.

We now look at the ability of the competing procedures in identifying the true underlying sparsity patterns in the matrix of fixed-effects $\boldsymbol{B}$ under the different scenarios. In so doing we compute, for each replication of the simulated experiment, the F_1 score defined as follows:

$$F_1 = \frac{\mathtt{tp}}{\mathtt{tp} + 0.5(\mathtt{fp} + \mathtt{fn})}, \tag{20}$$

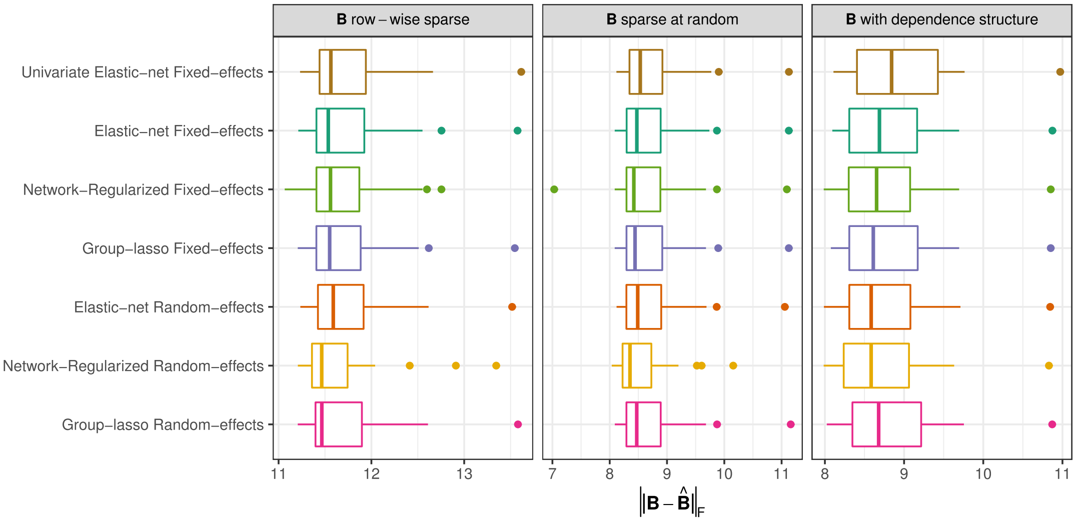

FIG. 5. *Boxplots of the Frobenius distance between true and estimated matrices of fixed-effects* $\boldsymbol{B}$ *for $MC = 100$ repetitions of the simulated experiment.*

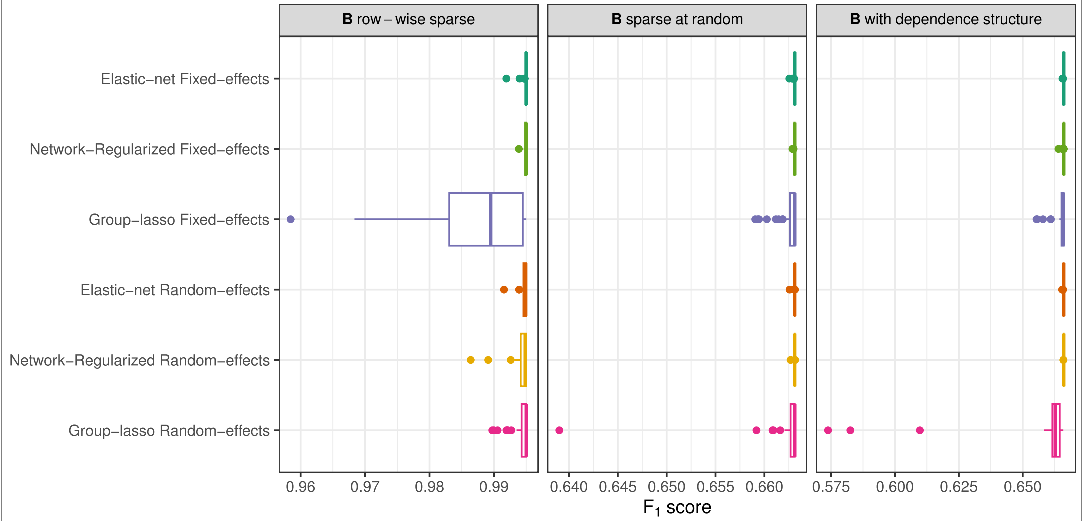

FIG. 6. *Boxplots of the F_1 score for $MC = 100$ repetitions of the simulated experiment for different methods and three scenarios varying sparsity pattern for $\boldsymbol{B}$.*

where with tp we denote the number of zero entries in $\boldsymbol{B}$ correctly estimated as such, while fp and fn represent the number of nonzero entries wrongly shrunk to 0 and the number of zero entries not shrunk to 0, respectively. Figure 6 displays boxplots of such metric for different methods and scenarios. We notice that $\boldsymbol{B}$ *rowwise sparse* structure displays much higher F_1 score, irrespective of the considered method, than the other two cases. Intuitively, the former scenario is less challenging since, while all penalty types can potentially accommodate a rowwise sparse $\boldsymbol{B}$, group-lasso regularizers only force entire rows of $\boldsymbol{B}$ to be shrunk to 0. We further observe that the F_1 score is higher for the *group-lasso random-effects* than for its fixed-effects counterpart, highlighting that a penalized mixed-effects modeling strategy, in presence of grouped data and a rowwise sparse $\boldsymbol{B}$, not only increases the predictive accuracy but also improves the recovery of the sparsity pattern in the fixed-effects matrix. The same does not happen in the remaining two scenarios for which all methods display a comparable empirical distribution of the F_1 metric across simulations. The same behavior was already observed in high-dimensional linear mixed-effects modeling for univariate responses (Schelldorfer, Bühlmann and van de Geer (2011)).

As a last worthy note, we acknowledge that, as rightly underlined by an anonymous reviewer, the present simulation study does not consider any violation in the distributional assumptions of the involved quantities. In this regard we replicated the experiment using both a multivariate skew-normal and multivariate skew-t distributions (Azzalini and Capitanio (2013)) as generative models for the error term, but we did not report the results in the paper since no dramatic changes were observed in model performances. While clearly more extreme scenarios could be considered, results in the literature have previously validated the robustness of linear mixed-effects models to violations of distributional assumptions (McCulloch and Neuhaus (2011)). For further details about the simulation study, the Supplementary Material (Cappozzo, Ieva and Fiorito (2023)) provides additional figures and a note on the overall computing times.

All in all, the good performances displayed by our proposal, particularly when coupled with a multivariate group-lasso penalty, encourage its usage in multivariate DNAm surrogates creation: promising results are reported in the next section.

5. DNAm biomarkers analysis for EPIC and EXPOsOMICS datasets. The methodology described in Section 3 is employed to build a *five*-dimensional DNAm biomarker of hypertension and hyperlipidemia. As mentioned in the Introduction, DNAm surrogates possess extensive advantages over their blood-measured counterparts since:

1. DNAm biomarkers directly account for genetic susceptibility and subject specific response to risk factors.
2. DNAm biomarkers can immediately be computed whenever DNAm values are accessible. This is particularly useful when the risk factors of interest have not been directly measured.
3. Further understanding of the biomolecular mechanisms associated with complex phenotypes can be acquired through a pathway enrichment analysis (Reimand et al. (2019)), allowing 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).

We subsequently assess how well the so-devised surrogates perform, for both internal and external cohorts, in reconstructing the blood measured biomarkers (Section 5.1) and in predicting the clinical endpoint of interest, namely, the future presence/absence of CVD events (Section 5.2). Lastly, we study from a biological perspective the CpG sites selection operated by the multivariate group-lasso penalty, comparing it with previous findings available in the literature (Section 5.3).

5.1. *DNAm surrogates creation and validation*. To construct multivariate DNAm surrogates, several penalized models are fitted to the EPIC Italy training set, varying shrinkage factors and considering both fixed and random-effects components. As mentioned in Section 2, the design matrix comprises of $p = 62{,}130$ variables. Thus, redundancies are likely to occur as the feature space is constituted by the union of CpG sites prescreened by univariate epigenome-wide analyses. After having standardized the covariates, for each model the penalty term λ is tuned via 10-fold CV, while the mixing parameter α is kept fixed and equal to 0.5. Results on the internal cohort are summarized in Table 1, where the RMSE computed on the EPIC Italy test set, the number of active CpG sites and the overall elapsed time 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 multivariate group-lasso (Section 3.3.2) and elastic-net (Section 3.3.1) 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 *four* out of *five* 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), while allowing for centerwise difference to be captured by the random intercept. Furthermore, thanks to the multivariate 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. 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 7.97%, 16.05%, 5.56%, 14.26% and 6.01% for DBP, HDL, LDL, SBP and TG, respectively, giving reason for the performance improvement showcased by the random-effects models.

The employment of the multivariate group-lasso penalty within a mixed-effects multitask 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

TABLE 1

Root Mean Squared Error (RMSE), active number of CpG sites and overall computing times 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						
Model	Penalty type	Response	DBP	HDL	LDL	SBP	TG	Active # CpG sites	Elapsed time (secs)
Random-effects	Group-lasso	Multivariate	**0.1167**	**0.2442**	**0.3236**	0.1374	**0.4745**	417	304
Random-effects	Elastic-net	Multivariate	0.1178	0.2480	0.3318	0.1379	0.4853	773	549
Fixed-effects	Group-lasso	Multivariate	0.1317	0.2602	0.3321	0.1364	0.4853	382	54
Fixed-effects	Elastic-net	Multivariate	0.1176	0.2513	0.3324	0.1362	0.4841	115	311
Fixed-effects	Elastic-net	Univariate	0.1179	0.2596	0.3383	**0.1359**	0.4996	1712	115

TABLE 2

Root Mean Squared Error (RMSE) for different penalized regression models, EXPOsOMICS validation set. Bold numbers indicate lowest RMSE for each of the $r = 5$ dimension of the response matrix

Model	Penalty type	Response	DBP	HDL	LDL	SBP	TG
Random-effects	Group-lasso	Multivariate	**0.1314**	**0.2384**	**0.2707**	0.1412	**0.4735**
Random-effects	Elastic-net	Multivariate	0.1335	0.2504	0.2821	0.1450	0.4890
Fixed-effects	Group-lasso	Multivariate	0.1409	0.2750	0.2969	0.1574	0.5142
Fixed-effects	Elastic-net	Multivariate	0.1286	0.2479	0.2859	0.1359	0.5002
Fixed-effects	Elastic-net	Univariate	0.1331	0.2733	0.3136	**0.1368**	0.5251

between DNAm and the exposure variable) the same set of CpG sites (Tyler, Crawford and Pendergrass (2013), Richard et al. (2017)).

Internal validation results, obtained for the EPIC Italy test set, highlight the benefits of mixed-effects modeling while constructing DNAm surrogates for data possessing a grouping-structure. The random intercept allows to account for centerwise variability that is induced by geographic genetical variation as well as by samples collection and storage that are likely to differ across centers. Notice that, in general, if not properly modeled, the unexplained heterogeneity in multicenter studies could be taken care of with dedicated batch-effect removal procedures (see, e.g., Johnson, Li and Rabinovic (2007)). Nonetheless, when developing DNAm biomarkers, it is of interest to devise study-invariant surrogates to be readily computed also for samples not belonging to the learning cohort. To this aim, we validate the performance of the models estimated on the EPIC training set in constructing surrogates for the external EXPOsOMICS cohort (see Section 2). In this context the grouping information (i.e., the center of recruitment) cannot be considered when performing predictions with mixed-effects models, and the unconditional (population-level) intercepts are thus utilized. RMSE between estimated and blood-measured biomarkers for the EXPOsOMICS validation cohort are reported in Table 2. Likewise for the EPIC Italy test set, the lowest RMSEs for all but SBP biomarker are retained employing a penalized random-intercept model with multivariate group-lasso penalty. Interestingly, the predictive outcomes, obtained in the EXPOsOMICS validation dataset, are comparable with the ones reported in Table 1 with slightly worse performances, as it may be expected for those dimensions of the response displaying higher PVRE indexes. All in all, the proposed approach exhibits promising results when it comes to multivariate DNAm biomarker creation, outperforming the current employed procedure in both internal and external validation cohorts.

5.2. *Association of DNAm surrogates with CVD risk.* DNAm surrogates creation is not a stand-alone regression problem, as its primary aim is to provide reliable covariates for diseases prediction models (Fernández-Sanlés et al. (2021), Odintsova et al. (2021), Hidalgo et al. (2021)). We, therefore, validate whether employing the estimated DNAm surrogates acts as a superior proxy of blood measured biomarkers in association analyses. In details within the cohort of patients in the EPIC Italy test set, we build logistic regression models to predict the probability of cardiovascular risk, using as regressors either the blood measured biomarkers or the two best performing DNAm surrogates devised in the previous section, adjusting for sex and age. The receiver operating characteristic (ROC) curves and the associated area under the curve (AUC) metrics for the considered methods are displayed in Figure 7. As expected, in light of the RMSE results reported in Table 1, we notice that classification performances are similar among the competing models. Nonetheless, the logistic curves regressed on the DNAm based surrogates seem to outperform the blood measured counterparts. Interestingly, all surrogates in Table 1 define logistic regression models whose AUCs are higher

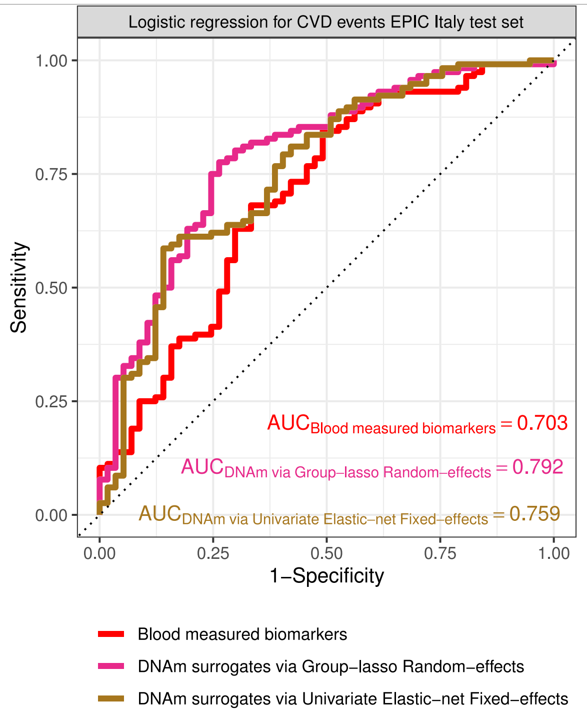

FIG. 7. *Receiver operating characteristic curves and area under the curve metrics for the association analyses of DNAm surrogates with CVD risk, Italy EPIC test set.*

than those retrieved by the blood-measured biomarkers, with the best performance attained by the penalized random-intercept model with multivariate group-lasso penalty.

We further assess the association of DNAm surrogates with CVD risk in the external EXPOsOMICS study. For this dataset values of the blood-based biomarkers are available only for a subset of volunteers; we thus construct the logistic regression models by means of the surrogates only. Such a situation is quite common in validation data and in line with the principle DNAm surrogates that were devised in the first place. Also, in this context the predictive performance of our novel proposal, coupled with a multivariate group lasso penalty, is higher with respect to state-of-the-art surrogates created via elastic-net fixed-effects models. The associated receiver operating characteristic curves as well as additional figures related to the DNAm biomarkers analysis are reported in the Supplementary Material (Cappozzo, Ieva and Fiorito (2023)).

The association analyses described in this section cast light on the applicability of the devised DNAm surrogates as an enriched and patient-specific proxy of their blood measured counterparts, in both internal and external cohorts. These favorable outcomes indicate that using models based on DNAm surrogates could be more appropriate for prediction tasks, such as CVD prevention, since they can possibly incorporate individual characteristics not directly recorded in the blood-measured biomarkers.

5.3. *CpG sites selection and gene set enrichment analyses of inflammatory pathways.* In the previous sections the newly devised random-intercept model for multitask learning has demonstrated superior predictive performance when it comes to DNAm surrogates creation and CVD prediction. Hereafter, we examine the epidemiological rationale of the multivariate group-lasso penalty, compared to univariate elastic-nets, investigating the biological reliability of the selected features (CpG sites). The univariate elastic-nets extract 178, 504, 518, 79, 497 CpG sites for diastolic blood pressure, HDL cholesterol, LDL cholesterol, systolic blood pressure and triglycerides, respectively. As reported in Table 1, the total number of unique CpGs is 1712. 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. Instead, as previously described, our MLMM procedure regularized with a multivariate group-lasso penalty extracts 417 features that are associated with the five outcomes at the same time, a biological mechanism known as pleiotropy (Atchley and Hall (1991a, 1991b)).

We are further interested in assessing whether the selected CpG sites are associated with specific biological pathways. To do so, gene set enrichment analyses (Subramanian et al. (2005)) are performed on the features retained by penalized MLMM and univariate elastic-net models. Specifically, given that the number of CpGs extracted for each method is small (hundreds of CpGs selected from an initial set of 295,614 features), an analysis based on all of the biomolecular pathways described in the canonical datasets, that is, KEGG (Yi et al. (2020)), GO (Gene Ontology Consortium (2004)) and Reactome (Fabregat et al. (2018)), would be underpowered. Therefore, to overcome this limitation and with inflammation being the main mechanism involved in the onset of the majority of chronic diseases, we focus our analyses on the 17 inflammatory pathways described in Loza et al. (2007). Enrichment analyses results are summarized in Table 3. For each list of CpGs, we test for overrepresentation of features in inflammatory pathways using the method implemented in the `missMethyl R` package (Phipson, Maksimovic and Oshlack (2016)). Considering the CpGs extracted by our multivariate approach, we find significant enrichment for CpGs in four inflammatory pathways. These results agree with previous literature suggesting that hypertension and hyperlipidaemia are associated with the dysregulation of molecular pathways regulating apoptosis, oxidative stress and the immune system (Senoner and Dichtl (2019), Dong et al. (2020)). Instead, modeling the outcomes one by one, using univariate models, leads to a less consistent

TABLE 3

Empirical p-values of the enrichment analyses computed using a permutation procedure via the `gometh` *function in the* `missMethyl R` *package. Empirical p-values lower than* 0.05, *highlighted in bold, indicate significant overrepresentation*

		Univariate Elastic-net Fixed-effects				
Inflammatory pathway	Group-lasso Random-effects	DBP	HDL	LDL	SBP	TG
Leukocyte signaling	**0.006**	0.14	0.35	0.01	1	1
ROS/Glutathione/Cytotoxic granules	**0.009**	0.06	0.15	1	**0.03**	0.15
Apoptosis Signaling	**0.01**	1	1	0.17	1	0.07
Natural Killer Cell Signaling	**0.01**	0.36	0.10	1	1	1
PI3K/AKT Signaling	0.18	1	0.26	0.03	1	0.22
Innate pathogen detection	0.26	0.12	0.30	0.31	1	1
Cytokine signaling	0.36	1	1	**0.0003**	1	0.10
Adhesion-Extravasation-Migration	1	0.18	**0.003**	0.02	0.09	0.10
Calcium Signaling	1	1	1	0.08	1	1
Complement Cascase	1	1	1	1	1	0.16
Glucocorticoid/PPAR signaling	1	1	0.10	1	0.17	0.10
G-Protein Coupled Receptor Signaling	1	1	1	1	0.05	0.27
MAPK signaling	1	1	0.14	0.49	1	**0.005**
NF-kB signaling	1	1	0.16	0.16	1	0.15
Phagocytosis-Ag presentation	1	1	1	0.26	1	1
Eicosanoid Signaling	1	1	1	1	1	1
TNF Superfamily Signaling	1	1	0.01	0.14	1	1

pattern of associations. In fact, we find only one (and always different) significant pathway per analysis.

All in all, the results reported in this section support the advantages of modeling multiple correlated outcomes not only from a prediction perspective, for both blood measured biomarkers and endpoint of interest but also considering the biological reliability of the extracted features.

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 multicenter study. 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 $62{,}128$ CpG sites and accounting for between-center heterogeneity. Our modeling framework, coupled with a multivariate 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 multidimensional 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.

The devised pipeline also possesses some limitations. The EWAS results are adjusted for clinical covariates external to the analysis, and this may thus affect the preprocessing outcome. Moreover, the level of strictness in the screening process is influenced by the chosen threshold on the p-values. On this wise two different, yet both sensible, strategies can be adopted. On the one hand, one may rely on the "Occam's razor" principle, preferring to use a stricter threshold being it the simplest and fastest option. On the other hand, one can positively include many redundant variables in the design matrix, relying on the model ability to shrink coefficients of irrelevant features to zero. Concurrently, further insights about the associations between DNA methylation and blood-measured biomarkers may be unraveled by means of sparse multiple canonical correlation analysis (Rodosthenous, Shahrezaei and Evangelou (2020), Witten, Tibshirani and Hastie (2009), Witten and Tibshirani (2009)), while other modeling approaches, such as deep-learning (Nguyen et al. (2022), Yuan et al. (2022)), Bayesian methods (Zhao et al. (2021a, 2021b)) and boosting machines (Sigrist (2022)), could be profitably adapted to build DNAm multidimensional surrogates.

A direction for future research concerns promoting the application of the proposed procedure in creating additional multidimensional DNAm biomarkers, conveniently embedding mixed-effects and customized penalty types. In this regard and of particular interest may be the definition of a shrinkage term for which the grouping in $\boldsymbol{B}$ is introduced from both responses and predictors: the former is induced by the multivariate nature of $\boldsymbol{Y}$ (i.e., the r responses), while the latter can stem from any structure present in $\boldsymbol{X}$ (e.g., CpG islands). Such a problem could be solved by extending the multivariate sparse group lasso, proposed by Li, Nan and Zhu (2015), to the mixed-effects framework.

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 being nominal, ordinal, discrete and/or continuous. Some proposals are currently under study, and they will be the object of future work.

Acknowledgments. The authors would like to thank the three anonymous reviewers, the Editor, and the Associate Editor for their thorough examination of the manuscript and their valuable, constructive comments. Their input has greatly enhanced the quality of the work. The authors also thank the EPIC Italy research group (Carlotta Sacerdote, Vittorio Krogh, Domenico Palli, Salvatore Panico, Rosario Tumino, Paolo Vineis and their collaborators) for giving access to the data used in this study.

SUPPLEMENTARY MATERIAL

Additional figures (DOI: 10.1214/23-AOAS1760SUPPA; .pdf). It contains additional figures for both the simulation study and the real data analysis reported in Sections 4 and 5.

Code (DOI: 10.1214/23-AOAS1760SUPPB; .zip). It contains the `R` package `emlmm` implementing the method proposed in the manuscript.

REFERENCES

ANASTASIADI, D., ESTEVE-CODINA, A. and PIFERRER, F. (2018). Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenet. Chromatin* **11** 37. https://doi.org/10.1186/s13072-018-0205-1

ATCHLEY, W. R. and HALL, B. K. (1991a). A model for development and evolution of complex morphological structures. *Biol. Rev. Camb. Philos. Soc.* **66** 101–157.

ATCHLEY, W. R. and HALL, B. K. (1991b). A model for development and evolution of complex morphological structures. *Biol. Rev.* **66** 101–157.

AZZALINI, A. and CAPITANIO, A. (2013). *The Skew-Normal and Related Families* **3**. Cambridge Univ. Press, Cambridge.

BATTRAM, T., YOUSEFI, P., CRAWFORD, G., PRINCE, C., BABAEI, M. S., SHARP, G., HATCHER, C., VEGA-SALAS, M. J., KHODABAKHSH, S. et al. (2022). The EWAS catalog: A database of epigenome-wide association studies. *Wellcome Open Res.* **7**.

CAMPAGNA, M. P., XAVIER, A., LECHNER-SCOTT, J., MALTBY, V., SCOTT, R. J., BUTZKUEVEN, H., JOKUBAITIS, V. G. and LEA, R. A. (2021). Epigenome-wide association studies: Current knowledge, strategies and recommendations. *Clin. Epigenet.* **13** 214. https://doi.org/10.1186/s13148-021-01200-8

CAPPOZZO, A., IEVA, F. and FIORITO, G. (2023). Supplement to “A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation.” https://doi.org/10.1214/23-AOAS1760SUPPA, https://doi.org/10.1214/23-AOAS1760SUPPB

CAPPOZZO, A., MCCRORY, C., ROBINSON, O., FRENI STERRANTINO, A., SACERDOTE, C., KROGH, V., PANICO, S., TUMINO, R., IACOVIELLO, L. et al. (2022). A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. *Clin. Epigenet.* **14** 121.

CARUANA, R. (1997). Multitask learning. *Mach. Learn.* **28** 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. et al. (2021). Epigenome-wide association study of Covid-19 severity with respiratory failure. *eBioMedicine* **66** 103339.

CHENG, W., ZHANG, X., GUO, Z., SHI, Y. and WANG, W. (2014). Graph-regularized dual Lasso for robust eQTL mapping. *Bioinformatics* **30** 139–148.

CHIPPERFIELD, J. O. and STEEL, D. G. (2012). Multivariate random effect models with complete and incomplete data. *J. Multivariate Anal.* **109** 146–155. MR2922860 https://doi.org/10.1016/j.jmva.2012.02.014

CHUNG, F. R. K. and GRAHAM, F. C. (1997). *Spectral Graph Theory* **92**. Am. Math. Soc., Providence.

COLICINO, E., JUST, A., KIOUMOURTZOGLOU, M.-A., VOKONAS, P., CARDENAS, A., SPARROW, D., WEISSKOPF, M., NIE, L. H., HU, H. et al. (2021). Blood DNA methylation biomarkers of cumulative lead exposure in adults. *J. Expo. Sci. Environ. Epidemiol.* **31** 108–116.

CONOLE, E. L. S., STEVENSON, A. J., GREEN, C., HARRIS, S. E., MANIEGA, S. M., VALDÉS-HERNÁNDEZ, M. D. C., HARRIS, M. A., BASTIN, M. E., WARDLAW, J. M. et al. (2020). An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. *MedRxiv* 2020.10.08.20205245.

GENE ONTOLOGY CONSORTIUM (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** 258D–261.

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963 https://doi.org/10.1093/biomet/68.1.265

DEBRUINE, L. (2021). faux: Simulation for Factorial Designs.

DEMIDENKO, E. (2013). *Mixed Models: Theory and Applications with R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3235905

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DIRMEIER, S., FUCHS, C., MUELLER, N. S. and THEIS, F. J. (2018). netReg: Network-regularized linear models for biological association studies. *Bioinformatics* **34** 896–898. https://doi.org/10.1093/bioinformatics/btx677

DONG, W., CHEN, H., WANG, L., CAO, X., BU, X., PENG, Y., DONG, A., YING, M., CHEN, X. et al. (2020). Exploring the shared genes of hypertension, diabetes and hyperlipidemia based on microarray. *Braz. J. Pharm. Sci.* **56** 1–12.

FABREGAT, A., JUPE, S., MATTHEWS, L., SIDIROPOULOS, K., GILLESPIE, M., GARAPATI, P., HAW, R., JASSAL, B., KORNINGER, F. et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* **46** D649–D655.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099

FAZZARI, M. J. and GREALLY, J. M. (2010). Introduction to Epigenomics and Epigenome-Wide Analysis. In *Statistical Methods in Molecular Biology* 243–265. Humana Press, Totowa, NJ.

FERNÁNDEZ-SANLÉS, A., SAYOLS-BAIXERAS, S., SUBIRANA, I., SENTÍ, M., PÉREZ-FERNÁNDEZ, S., DE CASTRO MOURA, M., ESTELLER, M., MARRUGAT, J. and ELOSUA, R. (2021). DNA methylation biomarkers of myocardial infarction and cardiovascular disease. *Clin. Epigenet.* **13** 86. https://doi.org/10.1186/s13148-021-01078-6

FIORITO, G., PEDRON, S., OCHOA-ROSALES, C., MCCRORY, C., POLIDORO, S., ZHANG, Y., DUGUÉ, P.-A., RATLIFF, S., ZHAO, W. N. et al. (2022). The role of epigenetic clocks in explaining educational inequalities in mortality: A multicohort study and meta-analysis. *J. Gerontol., Ser. A* **77** 1750–1759.

FIORITO, G., VLAANDEREN, J., POLIDORO, S., GULLIVER, J., GALASSI, C., RANZI, A., KROGH, V., GRIONI, S., AGNOLI, C. et al. (2018). Oxidative stress and inflammation mediate the effect of air pollution on cardio- and cerebrovascular disease: A prospective study in nonsmokers. *Environ. Mol. Mutagen.* **59** 234–246. https://doi.org/10.1002/em.22153

FROHLICH, H. and 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. **3** 1431–1436. IEEE, Los Alamitos.

GAŁECKI, A. and BURZYKOWSKI, T. (2013). *Linear Mixed-Effects Models Using R*. *Springer Texts in Statistics*. Springer, New York. MR3024843 https://doi.org/10.1007/978-1-4614-3900-4

GUIDA, F., SANDANGER, T. M., CASTAGNÉ, R., CAMPANELLA, G., POLIDORO, S., PALLI, D., KROGH, V., TUMINO, R., SACERDOTE, C. et al. (2015). Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum. Mol. Genet.* **24** 2349–2359. https://doi.org/10.1093/hmg/ddu751

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

HIDALGO, B. A., MINNIEFIELD, B., PATKI, A., TANNER, R., BAGHERI, M., TIWARI, H. K., ARNETT, D. K. and IRVIN, M. R. (2021). A 6-CpG validated methylation risk score model for metabolic syndrome: The HyperGEN and GOLDN studies. *PLoS ONE* **16** e0259836. https://doi.org/10.1371/journal.pone.0259836

HILLARY, R. F. and MARIONI, R. E. (2020). MethylDetectR: A software for methylation-based health profiling. *Wellcome Open Res*. **5** 283. https://doi.org/10.12688/wellcomeopenres.16458.2

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

JORDAN, M. I. (2013). On statistics, computation and scalability. *Bernoulli* **19** 1378–1390. MR3102908 https://doi.org/10.3150/12-BEJSP17

KIM, S., PAN, W. and SHEN, X. (2013). Network-based penalized regression with application to genomic data. *Biometrics* **69** 582–593. MR3106586 https://doi.org/10.1111/biom.12035

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping. *Ann. Appl. Stat.* **6** 1095–1117. MR3012522 https://doi.org/10.1214/12-AOAS549

LANGFELDER, P. and HORVATH, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform*. **9** 559.

LARIA, J. C., CARMEN AGUILERA-MORILLO, M. and LILLO, R. E. (2019). An iterative sparse-group lasso. *J. Comput. Graph. Statist.* **28** 722–731. MR4007753 https://doi.org/10.1080/10618600.2019.1573687

LI, C. and LI, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.* **4** 1498–1516. MR2758338 https://doi.org/10.1214/10-AOAS332

LI, Y., NAN, B. and ZHU, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363. MR3366240 https://doi.org/10.1111/biom.12292

LOZA, M. J., MCCALL, C. E., LI, L., ISAACS, W. B., XU, J. and CHANG, B.-L. (2007). Assembly of inflammation-related genes for pathway-focused genetic analysis. *PLoS ONE* **2** e1035.

LU, A. T., QUACH, A., WILSON, J. G., REINER, A. P., AVIV, A., RAJ, K., HOU, L., BACCARELLI, A. A., LI, Y. et al. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11** 303–327.

MARABITA, F., ALMGREN, M., LINDHOLM, M. E., RUHRMANN, S., FAGERSTRÖM-BILLAI, F., JAGODIC, M., SUNDBERG, C. J., EKSTRÖM, T. J., TESCHENDORFF, A. E. et al. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the illumina HumanMethylation450 BeadChip platform. *Epigenetics* **8** 333–346. https://doi.org/10.4161/epi.24008

MCCULLOCH, C. E. and NEUHAUS, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statist. Sci.* **26** 388–402. MR2917962 https://doi.org/10.1214/11-STS361

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2392878 https://doi.org/10.1002/9780470191613

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503 https://doi.org/10.1093/biomet/80.2.267

NGUYEN, T. M., LE, H. L., HWANG, K.-B., HONG, Y.-C. and KIM, J. H. (2022). Predicting high blood pressure using DNA methylome-based machine learning models. *Biomedicines* **10** 1406.

OBOZINSKI, G., TASKAR, B. and JORDAN, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput.* **20** 231–252. MR2610775 https://doi.org/10.1007/s11222-008-9111-x

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2009). High-dimensional support union recovery in multivariate regression. In *Advances in Neural Information Processing Systems* 21*—Proceedings of the* 2008 *Conference* 1217–1224.

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011b). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839 https://doi.org/10.1214/09-AOS776

ODINTSOVA, V. V., REBATTU, V., HAGENBEEK, F. A., POOL, R., BECK, J. J., EHLI, E. A., VAN BEIJSTERVELDT, C. E. M., LIGTHART, L., WILLEMSEN, G. et al. (2021). Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles. *Front. Psychiatr.* **12** 1–17.

PANICO, S., DELLO IACOVO, R., CELENTANO, E., GALASSO, R., MUTI, P., SALVATORE, M. and MANCINI, M. (1992). Progetto ATENA, a study on the etiology of major chronic diseases in women: Design, rationale and objectives. *Eur. J. Epidemiol.* **8** 601–608.

PHIPSON, B., MAKSIMOVIC, J. and OSHLACK, A. (2016). missMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32** 286–288.

PINHEIRO, J. and BATES, D. (2006). *Mixed-Effects Models in S and S-PLUS*. Springer, Berlin.

RAULUSEVICIUTE, I., DRABLØS, F. and RYE, M. B. (2020). DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Med. Genom.* **13** 6.

REIMAND, J., ISSERLIN, R., VOISIN, V., KUCERA, M., TANNUS-LOPES, C., ROSTAMIANFAR, A., WADI, L., MEYER, M., WONG, J. et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14** 482–517.

REINSEL, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc.* **79** 406–414. MR0755095

RIBOLI, E., HUNT, K., SLIMANI, N., FERRARI, P., NORAT, T., FAHEY, M., CHARRONDIÈRE, U., HÉMON, B., CASAGRANDE, C. et al. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): Study populations and data collection. *Public Health Nutr.* **5** 1113–1124.

RICHARD, M. A., HUAN, T., LIGTHART, S., GONDALIA, R., JHUN, M. A., BRODY, J. A., IRVIN, M. R., MARIONI, R., SHEN, J. et al. (2017). DNA methylation analysis identifies loci for blood pressure regulation. *Am. J. Hum. Genet.* **101** 888–902.

RODOSTHENOUS, T., SHAHREZAEI, V. and EVANGELOU, M. (2020). Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: A comparison study. *Bioinformatics* **36** 4616–4625. https://doi.org/10.1093/bioinformatics/btaa530

ROHART, F., SAN CRISTOBAL, M. and LAURENT, B. (2014). Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. *Comput. Statist. Data Anal.* **80** 209–222. MR3240488 https://doi.org/10.1016/j.csda.2014.06.022

SCHAFER, J. L. and YUCEL, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *J. Comput. Graph. Statist.* **11** 437–457. MR1938143 https://doi.org/10.1198/106186002760180608

SCHELLDORFER, J., BÜHLMANN, P. and VAN DE GEER, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.* **38** 197–214. MR2829596 https://doi.org/10.1111/j.1467-9469.2011.00740.x

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SENONER, T. and DICHTL, W. (2019). Oxidative stress in cardiovascular diseases: Still a therapeutic target? *Nutrients* **11**.

SHAH, A., LAIRD, N. and SCHOENFELD, D. (1997). A random-effects model for multiple characteristics with possibly missing data. *J. Amer. Statist. Assoc.* **92** 775–779. MR1467867 https://doi.org/10.2307/2965726

SIGRIST, F. (2022). Latent Gaussian model boosting. *IEEE Trans. Pattern Anal. Mach. Intell.* 1–1.

SILL, M., HIELSCHER, T., BECKER, N. and ZUCKNICK, M. (2014). c060: Extended inference with lasso and elastic-net regularized Cox and generalized linear models. *J. Stat. Softw.* **62**.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712 https://doi.org/10.1080/10618600.2012.681250

SINGAL, R. and GINDER, G. D. (1999). DNA methylation. *Blood* **93** 4059–4070.

STEVENSON, A. J., MCCARTNEY, D. L., HILLARY, R. F., CAMPBELL, A., MORRIS, S. W., BERMINGHAM, M. L., WALKER, R. M., EVANS, K. L., BOUTIN, T. S. et al. (2020). Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin. Epigenet.* **12** 113.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

TAY, J. K., NARASIMHAN, B. and HASTIE, T. (2021). Elastic net regularization paths for all generalized linear models.

R CORE TEAM (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TYLER, A. L., CRAWFORD, D. C. and PENDERGRASS, S. A. (2013). Detecting and characterizing pleiotropy: New methods for uncovering the connection between the complexity of genomic architecture and multiple phenotypes. In *Biocomputing* 2014 183–187. World Scientific, Singapore.

VAN EIJK, K. R., DE JONG, S., BOKS, M. P. M., LANGEVELD, T., COLAS, F., VELDINK, J. H., DE KOVEL, C. G. F., JANSON, E., STRENGMAN, E. et al. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13** 636.

VINGA, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. *Brief. Bioinform.* **22** 77–87. https://doi.org/10.1093/bib/bbaa122

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

WITTEN, D. M. and TIBSHIRANI, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.* **8** 28. MR2533636 https://doi.org/10.2202/1544-6115.1470

WU, C.-Y., HU, H.-Y., CHOU, Y.-J., HUANG, N., CHOU, Y.-C. and LI, C.-P. (2015). High blood pressure and all-cause and cardiovascular disease mortalities in community-dwelling older adults. *Medicine* **94** e2160.

YI, Y., FANG, Y., WU, K., LIU, Y. and ZHANG, W. (2020). Comprehensive gene and pathway analysis of cervical cancer progression. *Oncol. Lett.* **19** 3316–3332.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

YUAN, T., EDELMANN, D., FAN, Z., ALWERS, E., KATHER, J. N., BRENNER, H. and HOFFMEISTER, M. (2022). Machine learning in the identification of prognostic DNA methylation biomarkers among patients with cancer: A systematic review of epigenome-wide studies. *MedRxiv*.

ZHANG, Y., ELGIZOULI, M., SCHÖTTKER, B., HOLLECZEK, B., NIETERS, A. and BRENNER, H. (2016). Smoking-associated DNA methylation markers predict lung cancer incidence. *Clin. Epigenet.* **8** 1–12.

ZHAO, Z., BANTERLE, M., BOTTOLO, L., RICHARDSON, S., LEWIN, A. and ZUCKNICK, M. (2021a). BayesSUR: An R package for high-dimensional multivariate Bayesian variable and covariance selection in linear regression. *J. Stat. Softw.* **100**.

ZHAO, Z., BANTERLE, M., LEWIN, A. and ZUCKNICK, M. (2021b). Structured Bayesian variable selection for multiple related response variables and high-dimensional predictors. ArXiv Preprint. Available at arXiv:2101.05899, 1–33.

ZHAO, Z., WANG, S., ZUCKNICK, M. and AITTOKALLIO, T. (2022). Tissue-specific identification of multi-omics features for pan-cancer drug response prediction. *iScience* **25** 104767.

ZHAO, Z. and ZUCKNICK, M. (2020). Structured penalized regression for drug sensitivity prediction. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 525–545. MR4098960

ZHONG, J., AGHA, G. and BACCARELLI, A. A. (2016). The role of DNA methylation in cardiovascular risk and disease: Methodological aspects, study design, and data analysis for epidemiological studies. *Circ. Res.* **118** 119–131. https://doi.org/10.1161/CIRCRESAHA.115.305206

ZHONG, W., WANG, J. and CHEN, X. (2021). Censored mean variance sure independence screening for ultrahigh dimensional survival data. *Comput. Statist. Data Anal.* **159** 107206. MR4233350 https://doi.org/10.1016/j.csda.2021.107206

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 768–768.

THE ORIGINAL REFERENCE LIST

The list of entries below corresponds to the original Reference section of your article. The bibliography section on previous page was retrieved from MathSciNet applying an automated procedure.
Please check both lists and indicate those entries which lead to mistaken sources in automatically generated Reference list.

Anastasiadi, D., Esteve-Codina, A. and Piferrer, F. (2018). Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics & Chromatin* **11** 37.

Atchley, W. R. and Hall, B. K. (1991a). A model for development and evolution of complex morphological structures. *Biological Reviews of the Cambridge Philosophical Society* **66** 101–157.

Atchley, W. R. and Hall, B. K. (1991b). A model for development and evolution of complex morphological structures. *Biological Reviews* **66** 101–157.

Azzalini, A. and Capitanio, A. (2013). *The Skew-Normal and Related Families* **3**. Cambridge University Press.

Battram, T., Yousefi, P., Crawford, G., Prince, C., Babaei, M. S., Sharp, G., Hatcher, C., Vega-Salas, M. J., Khodabakhsh, S., Whitehurst, O. and Others (2022). The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome open research* **7**.

Campagna, M. P., Xavier, A., Lechner-Scott, J., Maltby, V., Scott, R. J., Butzkueven, H., Jokubaitis, V. G. and Lea, R. A. (2021). Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clinical Epigenetics* **13** 214.

Cappozzo, A., Ieva, F. and Fiorito, G. (2023). Supplement to "A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation".

Cappozzo, A., McCrory, C., Robinson, O., Freni Sterrantino, A., Sacerdote, C., Krogh, V., Panico, S., Tumino, R., Iacoviello, L., Ricceri, F., Sieri, S., Chiodini, P., McKay, G. J., McKnight, A. J., Kee, F., Young, I. S., McGuinness, B., Crimmins, E. M., Arpawong, T. E., Kenny, R. A., O'Halloran, A., Polidoro, S., Solinas, G., Vineis, P., Ieva, F. and Fiorito, G. (2022). A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. *Clinical Epigenetics* **14** 121.

Caruana, R. (1997). Multitask learning. *Machine learning* **28** 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, J. P., Albu, S., Casasnovas, C., Rull, A., Reverte, L., Dietl, B., Dalmau, D., Arranz, M. J., Llucià-Carol, L., Planas, A. M., Pérez-Tur, J., Fernandez-Cadenas, I., Villares, P., Tenorio, J., Colobran, R., Martin-Nalda, A., Soler-Palacin, P., Vidal, F., Pujol, A. and 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. and Wang, W. (2014). Graph-regularized dual Lasso for robust eQTL mapping. *Bioinformatics* **30** 139–148.

Chipperfield, J. O. and Steel, D. G. (2012). Multivariate random effect models with complete and incomplete data. *Journal of Multivariate Analysis* **109** 146–155.

Chung, F. R. K. and Graham, F. C. (1997). *Spectral graph theory* **92**. American Mathematical Soc.

Colicino, E., Just, A., Kioumourtzoglou, M.-A., Vokonas, P., Cardenas, A., Sparrow, D., Weisskopf, M., Nie, L. H., Hu, H., Schwartz, J. D., Wright, R. O. and Baccarelli, A. A. (2021). Blood DNA methylation biomarkers of cumulative lead exposure in adults. *Journal of Exposure Science & Environmental Epidemiology* **31** 108–116.

Conole, E. L. S., Stevenson, A. J., Green, C., Harris, S. E., Maniega, S. M., Valdés-Hernández, M. d. C., Harris, M. A., Bastin, M. E., Wardlaw, J. M., Deary, I. J., Miron, V. E., Whalley, H. C., Marioni, R. E. and Cox, S. R. (2020). An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. *medRxiv* 2020.10.08.20205245.

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32** 258D–261.

Dawid, A. P. (1981). Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application. *Biometrika* **68** 265.

DeBruine, L. (2021). faux: Simulation for Factorial Designs.

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*) **39** 1–22.

Dirmeier, S., Fuchs, C., Mueller, N. S. and Theis, F. J. (2018). NetReg: Network-regularized linear models for biological association studies. *Bioinformatics* **34** 896–898.

DONG, W., CHEN, H., WANG, L., CAO, X., BU, X., PENG, Y., DONG, A., YING, M., CHEN, X., ZHANG, X. and YAO, L. (2020). Exploring the shared genes of hypertension, diabetes and hyperlipidemia based on microarray. *Brazilian Journal of Pharmaceutical Sciences* **56** 1–12.

FABREGAT, A., JUPE, S., MATTHEWS, L., SIDIROPOULOS, K., GILLESPIE, M., GARAPATI, P., HAW, R., JASSAL, B., KORNINGER, F., MAY, B., MILACIC, M., ROCA, C. D., ROTHFELS, K., SEVILLA, C., SHAMOVSKY, V., SHORSER, S., VARUSAI, T., VITERI, G., WEISER, J., WU, G., STEIN, L., HERMJAKOB, H. and D'EUSTACHIO, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46** D649–D655.

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

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research* **10** 2013–2038.

FAZZARI, M. J. and GREALLY, J. M. (2010). Introduction to Epigenomics and Epigenome-Wide Analysis In *Statistical Methods in Molecular Biology* 243–265. Humana Press, Totowa, NJ.

FERNÁNDEZ-SANLÉS, A., SAYOLS-BAIXERAS, S., SUBIRANA, I., SENTÍ, M., PÉREZ-FERNÁNDEZ, S., DE CASTRO MOURA, M., ESTELLER, M., MARRUGAT, J. and ELOSUA, R. (2021). DNA methylation biomarkers of myocardial infarction and cardiovascular disease. *Clinical Epigenetics* **13** 86.

FIORITO, G., PEDRON, S., OCHOA-ROSALES, C., MCCRORY, C., POLIDORO, S., ZHANG, Y., DUGUÉ, P.-A., RATLIFF, S., ZHAO, W. N., MCKAY, G. J., COSTA, G., SOLINAS, M. G., HARRIS, K. M., TUMINO, R., GRIONI, S., RICCERI, F., PANICO, S., BRENNER, H., SCHWETTMANN, L., WALDENBERGER, M., MATIAS-GARCIA, P. R., PETERS, A., HODGE, A., GILES, G. G., SCHMITZ, L. L., LEVINE, M., SMITH, J. A., LIU, Y., KEE, F., YOUNG, I. S., MCGUINNESS, B., MCKNIGHT, A. J., VAN MEURS, J., VOORTMAN, T., KENNY, R. A., VINEIS, P. and CARMELI, C. (2022). The Role of Epigenetic Clocks in Explaining Educational Inequalities in Mortality: A Multicohort Study and Meta-analysis. *The Journals of Gerontology*: *Series A* **77** 1750–1759.

FIORITO, G., VLAANDEREN, J., POLIDORO, S., GULLIVER, J., GALASSI, C., RANZI, A., KROGH, V., GRIONI, S., AGNOLI, C., SACERDOTE, C., PANICO, S., TSAI, M.-Y., PROBST-HENSCH, N., HOEK, G., HERCEG, Z., VERMEULEN, R., GHANTOUS, A., VINEIS, P. and 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** 234–246.

FROHLICH, H. and 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. **3** 1431–1436. IEEE.

GAŁECKI, A. and BURZYKOWSKI, T. (2013). *Linear Mixed-Effects Models Using R*. *Springer Texts in Statistics*. Springer New York, New York, NY.

GUIDA, F., SANDANGER, T. M., CASTAGNÉ, R., CAMPANELLA, G., POLIDORO, S., PALLI, D., KROGH, V., TUMINO, R., SACERDOTE, C., PANICO, S., SEVERI, G., KYRTOPOULOS, S. A., GEORGIADIS, P., VERMEULEN, R. C. H., LUND, E., VINEIS, P. and CHADEAU-HYAM, M. (2015). Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human Molecular Genetics* **24** 2349–2359.

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*. Chapman and Hall/CRC.

HIDALGO, B. A., MINNIEFIELD, B., PATKI, A., TANNER, R., BAGHERI, M., TIWARI, H. K., ARNETT, D. K. and IRVIN, M. R. (2021). A 6-CpG validated methylation risk score model for metabolic syndrome: The HyperGEN and GOLDN studies. *PLOS ONE* **16** e0259836.

HILLARY, R. F. and MARIONI, R. E. (2021). MethylDetectR: a software for methylation-based health profiling. *Wellcome Open Research* **5** 283.

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

JORDAN, M. I. (2013). On statistics, computation and scalability. *Bernoulli* **19** 1378–1390.

KIM, S., PAN, W. and SHEN, X. (2013). Network-Based Penalized Regression With Application to Genomic Data. *Biometrics* **69** 582–593.

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *The Annals of Applied Statistics* **6** 1095–1117.

LANGFELDER, P. and HORVATH, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9** 559.

LARIA, J. C., CARMEN AGUILERA-MORILLO, M. and LILLO, R. E. (2019). An Iterative Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* **28** 722–731.

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

LI, Y., NAN, B. and ZHU, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363.

LOZA, M. J., MCCALL, C. E., LI, L., ISAACS, W. B., XU, J. and CHANG, B.-L. (2007). Assembly of Inflammation-Related Genes for Pathway-Focused Genetic Analysis. *PLoS ONE* **2** e1035.

LU, A. T., QUACH, A., WILSON, J. G., REINER, A. P., AVIV, A., RAJ, K., HOU, L., BACCARELLI, A. A., LI, Y., STEWART, J. D., WHITSEL, E. A., ASSIMES, T. L., FERRUCCI, L. and HORVATH, S. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11** 303–327.

MARABITA, F., ALMGREN, M., LINDHOLM, M. E., RUHRMANN, S., FAGERSTRÖM-BILLAI, F., JAGODIC, M., SUNDBERG, C. J., EKSTRÖM, T. J., TESCHENDORFF, A. E., TEGNÉR, J. and GOMEZ-CABRERO, D. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics* **8** 333–346.

MCCULLOCH, C. E. and NEUHAUS, J. M. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science* **26** 388–402.

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2*E. Wiley Series in Probability and Statistics* **54**. John Wiley & Sons, Inc., Hoboken, NJ, USA.

MENG, X.-L. and RUBIN, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* **80** 267.

NGUYEN, T. M., LE, H. L., HWANG, K.-B., HONG, Y.-C. and KIM, J. H. (2022). Predicting High Blood Pressure Using DNA Methylome-Based Machine Learning Models. *Biomedicines* **10** 1406.

OBOZINSKI, G., TASKAR, B. and JORDAN, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing* **20** 231–252.

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2009). High-dimensional support union recovery in multivariate regression. *Advances in Neural Information Processing Systems* 21—*Proceedings of the* 2008 *Conference* 1217–1224.

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics* **39** 1–47.

ODINTSOVA, V. V., REBATTU, V., HAGENBEEK, F. A., POOL, R., BECK, J. J., EHLI, E. A., VAN BEIJSTERVELDT, C. E. M., LIGTHART, L., WILLEMSEN, G., DE GEUS, E. J. C., HOTTENGA, J.-J., BOOMSMA, D. I. and VAN DONGEN, J. (2021). Predicting Complex Traits and Exposures From Polygenic Scores and Blood and Buccal DNA Methylation Profiles. *Frontiers in Psychiatry* **12** 1–17.

PANICO, S., DELLO IACOVO, R., CELENTANO, E., GALASSO, R., MUTI, P., SALVATORE, M. and 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** 601–608.

PHIPSON, B., MAKSIMOVIC, J. and OSHLACK, A. (2016). missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32** 286–288.

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

RAULUSEVICIUTE, I., DRABLØS, F. and RYE, M. B. (2020). DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Medical Genomics* **13** 6.

REIMAND, J., ISSERLIN, R., VOISIN, V., KUCERA, M., TANNUS-LOPES, C., ROSTAMIANFAR, A., WADI, L., MEYER, M., WONG, J., XU, C., MERICO, D. and BADER, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols* **14** 482–517.

REINSEL, G. (1984). Estimation and Prediction in a Multivariate Random Effects Generalized Linear Model. *Journal of the American Statistical Association* **79** 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. and SARACCI, R. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutrition* **5** 1113–1124.

RICHARD, M. A., HUAN, T., LIGTHART, S., GONDALIA, R., JHUN, M. A., BRODY, J. A., IRVIN, M. R., MARIONI, R., SHEN, J., TSAI, P.-C., MONTASSER, M. E., JIA, Y., SYME, C., SALFATI, E. L., BOERWINKLE, E., GUAN, W., MOSLEY, T. H., BRESSLER, J., MORRISON, A. C., LIU, C., MENDELSON, M. M., UITTERLINDEN, A. G., VAN MEURS, J. B., FRANCO, O. H., ZHANG, G., LI, Y., STEWART, J. D., BIS, J. C., PSATY, B. M., CHEN, Y.-D. I., KARDIA, S. L. R., ZHAO, W., TURNER, S. T., ABSHER, D., ASLIBEKYAN, S., STARR, J. M., MCRAE, A. F., HOU, L., JUST, A. C., SCHWARTZ, J. D., VOKONAS, P. S., MENNI, C., SPECTOR, T. D., SHULDINER, A., DAMCOTT, C. M., ROTTER, J. I., PALMAS, W., LIU, Y., PAUS, T., HORVATH, S., O'CONNELL, J. R., GUO, X., PAUSOVA, Z., ASSIMES, T. L., SOTOODEHNIA, N., SMITH, J. A., ARNETT, D. K., DEARY, I. J., BACCARELLI, A. A., BELL, J. T., WHITSEL, E., DEHGHAN, A., LEVY, D., FORNAGE, M., HEIJMANS, B. T., 'T HOEN, P. A. C., VAN MEURS, J., ISAACS, A., JANSEN, R., FRANKE, L., BOOMSMA, D. I., POOL, R., VAN DONGEN, J., HOTTENGA, J. J., VAN GREEVENBROEK, M. M. J., STEHOUWER, C. D. A., VAN DER KALLEN, C. J. H., SCHALKWIJK, C. G., WIJMENGA, C., ZHERNAKOVA, A., TIGCHELAAR, E. F., SLAGBOOM, P. E., BEEKMAN, M., DEELEN, J., VAN HEEMST, D., VELDINK, J. H., VAN DEN BERG, L. H., VAN DUIJN, C. M., HOFMAN, A., UITTERLINDEN, A. G., JHAMAI, P. M., VERBIEST, M., SUCHIMAN, H. E. D., 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, D. V., LUIJK, R., BONDER, M. J., VAN DIJK, F., ARINDRARTO, W., KIELBASA, S. M., SWERTZ, M. A. and VAN ZWET, E. W. (2017). DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. *The American Journal of Human Genetics* **101** 888–902.

RODOSTHENOUS, T., SHAHREZAEI, V. and EVANGELOU, M. (2020). Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. *Bioinformatics* **36** 4616–4625.

ROHART, F., SAN CRISTOBAL, M. and 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, J. L. and YUCEL, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *Journal of Computational and Graphical Statistics* **11** 437–457.

SCHELLDORFER, J., BÜHLMANN, P. and DE GEER, S. V. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using $\ell1$-Penalization. *Scandinavian Journal of Statistics* **38** 197–214.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.

SENONER, T. and DICHTL, W. (2019). Oxidative stress in cardiovascular diseases: Still a therapeutic target? *Nutrients* **11**.

SHAH, A., LAIRD, N. and SCHOENFELD, D. (1997). A Random-Effects Model for Multiple Characteristics with Possibly Missing Data. *Journal of the American Statistical Association* **92** 775–779.

SIGRIST, F. (2022). Latent Gaussian Model Boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1.

SILL, M., HIELSCHER, T., BECKER, N. and ZUCKNICK, M. (2014). c060: Extended Inference with Lasso and Elastic-Net Regularized Cox and Generalized Linear Models. *Journal of Statistical Software* **62**.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* **22** 231–245.

SINGAL, R. and GINDER, G. D. (1999). DNA Methylation. *Blood* **93** 4059–4070.

STEVENSON, A. J., MCCARTNEY, D. L., HILLARY, R. F., CAMPBELL, A., MORRIS, S. W., BERMINGHAM, M. L., WALKER, R. M., EVANS, K. L., BOUTIN, T. S., HAYWARD, C., MCRAE, A. F., MCCOLL, B. W., SPIRES-JONES, T. L., MCINTOSH, A. M., DEARY, I. J. and MARIONI, R. E. (2020). Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clinical Epigenetics* **12** 113.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. and MESIROV, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102** 15545–15550.

TAY, J. K., NARASIMHAN, B. and HASTIE, T. (2021). Elastic Net Regularization Paths for All Generalized Linear Models.

R CORE TEAM (2022). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

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

TYLER, A. L., CRAWFORD, D. C. and PENDERGRASS, S. A. (2013). Detecting and characterizing pleiotropy: new methods for uncovering the connection between the complexity of genomic architecture and multiple phenotypes. In *Biocomputing* 2014 183–187. WORLD SCIENTIFIC.

VAN EIJK, K. R., DE JONG, S., BOKS, M. P. M., LANGEVELD, T., COLAS, F., VELDINK, J. H., DE KOVEL, C. G. F., JANSON, E., STRENGMAN, E., LANGFELDER, P., KAHN, R. S., VAN DEN BERG, L. H., HORVATH, S. and OPHOFF, R. A. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13** 636.

VINGA, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. *Briefings in Bioinformatics* **22** 77–87.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

WITTEN, D. M. and TIBSHIRANI, R. J. (2009). Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology* **8** 1–27.

WU, C.-Y., HU, H.-Y., CHOU, Y.-J., HUANG, N., CHOU, Y.-C. and LI, C.-P. (2015). High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults. *Medicine* **94** e2160.

YI, Y., FANG, Y., WU, K., LIU, Y. and ZHANG, W. (2020). Comprehensive gene and pathway analysis of cervical cancer progression. *Oncology Letters* **19** 3316–3332.

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

YUAN, T., EDELMANN, D., FAN, Z., ALWERS, E., KATHER, J. N., BRENNER, H. and HOFFMEISTER, M. (2022). Machine learning in the identification of prognostic DNA methylation biomarkers among patients with cancer: a systematic review of epigenome-wide studies. *medRxiv*.

ZHANG, Y., ELGIZOULI, M., SCHÖTTKER, B., HOLLECZEK, B., NIETERS, A. and BRENNER, H. (2016). Smoking-associated DNA methylation markers predict lung cancer incidence. *Clinical Epigenetics* **8** 1–12.

ZHAO, Z., BANTERLE, M., BOTTOLO, L., RICHARDSON, S., LEWIN, A. and ZUCKNICK, M. (2021a). BayesSUR: An R Package for High-Dimensional Multivariate Bayesian Variable and Covariance Selection in Linear Regression . *Journal of Statistical Software* **100**.

ZHAO, Z., BANTERLE, M., LEWIN, A. and ZUCKNICK, M. (2021b). Structured Bayesian variable selection for multiple related response variables and high-dimensional predictors. *arXiv preprint arXiv*:2101.05899 1–33.

ZHAO, Z., WANG, S., ZUCKNICK, M. and AITTOKALLIO, T. (2022). Tissue-specific identification of multi-omics features for pan-cancer drug response prediction. *iScience* **25** 104767.

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

ZHONG, J., AGHA, G. and BACCARELLI, A. A. (2016). The Role of DNA Methylation in Cardiovascular Risk and Disease. *Circulation Research* **118** 119–131.

ZHONG, W., WANG, J. and CHEN, X. (2021). Censored mean variance sure independence screening for ultrahigh dimensional survival data. *Computational Statistics & Data Analysis* **159** 107206.

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

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation
Author : Andrea Cappozzoimspdfauthor 1 contains more than one orcidTry typing <return> to proceed.If that doesn't work, type X <return> to quit., Francesca Ievaimspdfauthor 2 contains more than one orcidTry typing <return> to proceed.If that doesn't work, type X <return> to quit., Giovanni Fioritoimspdfauthor 3 contains more than one orcidTry typing <return> to proceed.If that doesn't work, type X <return> to quit.
Subject : The Annals of Applied Statistics, 0, Vol. 0, No. 00, 1-26
Keywords: Mixed-effects models, multitask learning, EM algorithm, penalized estimation, multivariate regression, personalized medicine

THE LIST OF URI ADDRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page: http://www.e-publications.org/ims/support/urihelp.html.

```
200 https://imstat.org/journals-and-publications/annals-of-applied-statistics/ [2:pp.1,1] OK
301 http://www.imstat.org [2:pp.1,1] Moved Permanently // http://www.imstat.org/
200 https://orcid.org/0000-0001-9348-710X [2:pp.1,1] OK
200 https://orcid.org/0000-0003-0165-1983 [2:pp.1,1] OK
200 https://orcid.org/0000-0002-7651-5452 [2:pp.1,1] OK
--- mailto:andrea.cappozzo@polimi.it [2:pp.1,1] Check skip
--- mailto:francesca.ieva@polimi.it [2:pp.1,1] Check skip
--- mailto:giovannifiorito@gaslini.org [2:pp.1,1] Check skip
200 https://github.com/AndreaCappozzo/emlmm [4:pp.3,3,11,11] OK
404 https://doi.org/10.1214/23-AOAS1760SUPPA [4:pp.22,22,22,22] Not Found
404 https://doi.org/10.1214/23-AOAS1760SUPPB [4:pp.22,22,22,22] Not Found
301 http://arxiv.org/abs/arXiv:2101.05899 [2:pp.26,26] Moved Permanently
```