A general framework for penalized mixed-effects multitask learning with application on DNAm biomarkers creation

Andrea Cappozzo * Francesca Ieva* Giovanni Fiorito †

Abstract

The creation of non invasive biomarkers from blood DNA methylation profiles is a cutting-edge achievement in personalized medicine: DNAm epimutations have been demonstrated to be tightly related to lifestyle and 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 model, with elastic net being the standard way-to-go when accomplishing the task. Nonetheless, more advanced modeling procedures are required when the response is multivariate in nature and the samples showcase a structured dependence pattern. In this work, with the aim of developing a multivariate DNAm biomarker from a multi-centric study, we propose a general framework for high-dimensional, mixed-effects multitask learning. A penalized estimation scheme based on an EM algorithm is devised, in which any penalty criteria for fixed-effects models can be conveniently incorporated in the fitting process. The methodology is then employed to create a novel surrogate of cardiovascular and high blood pressure comorbidities, showcasing better results, both in terms of predictive power and epidemiological interpretation, than state-of-the-art alternatives.

1 Introduction

DNA methylation (DNAm) is an epigenetic process that regulates gene expression, typically occurring in cytosines within CpG sites in the DNA sequence. The development of surrogate scores based on blood DNA methylation has received thriving attention in recent years: impressive epidemiological evidence has been established between DNAm epimutations and long term exposure to lifestyle and environmental risk factors (Zhong et al., 2016; Gagliardi et al., 2020). 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). The reason 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, with elastic-net (Zou and Hastie, 2005) being the routinely employed technique when accomplishing the task. Indeed, the associated learning problem entirely falls within the “p bigger than N” framework: DNA methylation levels are measured at approximately half million CpG sites for each sample, with the dimension of the latter generally not exceeding the order of thousands in most studies. The afore-described procedure is shown to be widely effective in building DNAm biomarkers, with two very recent contributions including a surrogate score for cumulative lead exposure (Colicino et al., 2021) and the identification of CpG sites associated with clinical severity of COVID-19 disease (Castro de Moura et al., 2021). Nonetheless, elastic-net may not allow for the required degree of flexibility when dealing with particularly delicate tasks.

A first layer of complexity is encountered when a multi-dimensional DNAm biomarker needs to be created, to model multiple risk factors jointly and to coherently account for the correlation structure among the response variables. Such a multivariate problem, also known as multi-task regression in the machine learning literature, can be fruitfully untangled only if dedicated care is devoted in choosing the most appropriate penalty required for the analysis. For instance, one may opt for the incorporation of \( \ell_1/\ell_2 \) type of regularizers (Obozinski et al., 2010, 2011), that extend the lasso (Tibshirani, 1996), group-lasso (Yuan and Lin, 2006) and sparse group-lasso (Simon et al., 2013; Laria et al., 2019) to the multiple response framework. Another option could contemplate the inclusion, within the estimation, of prior information related to the association structure among CpG sites: this is effectively achieved by means of graph-based penalties (Li and Li, 2010; Kim et al., 2013; Cheng et al., 2014; Dirmeier et al., 2018). Furthermore, tree-based regularization methods have also been recently introduced in the literature, to account for hierarchical structure over the responses in a single study (Kim and Xing, 2012) as well as when multiple data sources are at our disposal (Zhao and Zucknick, 2020). For a thorough and up-to-date survey on the analysis of high-dimensional omics data via structured regularization we refer the interested reader to Vinga (2021), while the monograph of Hastie et al. (2015) provides a general introduction to statistical learning with sparsity.

A second layer of complexity is introduced when DNA samples and related blood-measured biomarkers are collected in a study comprising multiple cohorts. In such a situation, an unknown degree of heterogeneity may be included in the data, with patients coming from the same cohort sharing some degree of com-
monality. Observations in the dataset are thus no longer independent and the cohort-wise covariance structure needs to be properly estimated. Linear Mixed-Effects Models (LMM) provide a convenient solution to this problem by adding a random component to the model specification (see, e.g., Demidenko, 2013, for an introduction on the topic). Whilst being able to capture unobserved heterogeneity standard mixed models, very much like their fixed counterpart, cannot directly handle situations in which the number of predictors exceeds the sample size. In order to overcome this issue Schelldorfer et al. (2011) introduced a procedure for estimating high-dimensional LMM via an $\ell_1$-penalization. More recently, Rohart et al. (2014) devised a general-purpose ECM algorithm (Meng and Rubin, 1993) for solving the same issue, but achieving greater flexibility as the proposed framework can be combined with any penalized structure previously developed for linear fixed-effects models.

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

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

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

2 EPIC Italy data and study design

The considered dataset belongs to the Italian cohort 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 coun-
tries and followed for almost 15 years (Riboli et al., 2002). For each participant, lifestyle and personal history questionnaires were recorded, together with anthropomorphic measures and blood samples for DNA extraction. The EPIC Italy dataset is comprised of four geographical sub-cohorts: the provinces of Ragusa, Varese and the cities of Turin and Naples. The latter centre became associated with EPIC in later times through the Progetto ATENA study (Panico et al., 1992).

By profiting from the information recorded in the aforementioned sub-cohorts, we aim at creating a multi-dimensional DNAm biomarker for cardiovascular and hypertension comorbidities. To this extent, we consider a multivariate response comprised of \( r = 5 \) measures, namely Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL) and Triglycerides (TG). These characteristics were chosen as they represent the major risk factors for cardiovascular diseases (Wu et al., 2015). In building a DNAm biomarker, the response variables are regressed on DNA methylation values for each CpG site, adjusted for sex and age. A total of \( N = 574 \) patients in the \( J = 4 \) cohorts showcase non-missing values for each and every response variable: they comprise the sample onto which all subsequent analyses will be performed. An epigenome-wide association study (Campagna et al., 2021) was performed as a pre-screening procedure: out of the whole set of CpG sites, 13449 DNA methylation features have been retained for subsequent modeling. Together with sex and age, this amounts to a total of \( p = 13451 \) predictors and a 5-dimensional response. The boxplots in Figure 1 emphasize the effect induced by the Centre grouping on the responses. To capture the centre-wise variability and to maintain generalizability of the devised DNAm biomarker outside the Italy EPIC cohorts, a partial pooling random-intercept model must be adopted. That is, a \( q = 1 \) random effect component is included in the model specification. This challenging modeling task requires an ad-hoc formulation for multivariate mixed effect framework applicable to high-dimensional predictors. A novel approach, based on penalized estimation, is presented in the next Section.

3 Penalized mixed-effects models for multitask learning

3.1 Model definition

The multivariate linear mixed-effects model (Shah et al., 1997) expresses the \( n_i \times r \) response matrix \( Y_i \) for the \( i \)-th group as:

\[
Y_i = X_iB + Z_i\Lambda_i + E_i
\]  

(1)

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

- \( B \) is the \( p \times r \) matrix of fixed effects (including the intercept)
Figure 1: Boxplots of log-transformed Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Systolic Blood Pressure (SBP) and Triglycerides (TG) for different Center, Italy EPIC dataset.

- $\Lambda_i$ is the $q \times r$ matrix of random effects
- $X_i$ is the $n_i \times p$ fixed-effects regressor matrix (the first column is an all-one vector)
- $Z_i$ is the $n_i \times q$ random-effects regressor matrix
- $E_i$ is the $n_i \times r$ within-group error matrix
- $i = 1, \ldots, J$ with $J$ total number of groups

By employing the vec operator, we have that:

$$\text{vec} (\Lambda_i) \sim \mathcal{N} (0, \Psi)$$

where $\Psi$ 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:

$$\text{vec} (E_i) \sim \mathcal{N} (0, \Sigma \otimes I_{n_i}),$$

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

$$\text{vec} (Y_i) \sim \mathcal{N} \left( (I_r \otimes X_i) \text{vec} (B), (I_r \otimes Z_i) \Psi (I_r \otimes Z_i)^\prime + \Sigma \otimes I_{n_i} \right).$$
Given a sample of $N = \sum_{i=1}^{J} n_i$, the log-likelihood of model (1) reads:

$$\ell(\theta) = \sum_{i=1}^{J} -\frac{n_i}{2} \log 2\pi - \frac{1}{2} \log |(I_r \otimes Z_i) \Psi (I_r \otimes Z_i)^t + \Sigma \otimes I_{n_i}| +$$

$$- \frac{1}{2} (\text{vec} (Y_i) - (I_r \otimes X_i) \text{vec} (B))' \left( (I_r \otimes Z_i) \Psi (I_r \otimes Z_i) + \Sigma \otimes I_{n_i} \right)^{-1} (\text{vec} (Y_i) - (I_r \otimes X_i) \text{vec} (B))$$

(3)

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

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

(4)

with $p(B; \lambda)$ being a penalty term employed to regularize the fixed-effects $B$ 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 $p(B; \lambda)$ (see Section 3.3 for further details).

3.2 Model estimation

Direct maximization of (4) is unfeasible, as the quantities $\text{vec} (A_i)$ are unknown. We therefore construct an EM algorithm (Dempster et al., 1977) in which in the E-step the conditional expectations $\hat{\text{vec}} (A_i) = \mathbb{E}(\text{vec} (A_i) | Y_i; \theta)$, $i = 1, \ldots, J$ are computed, while in the M-step a complete penalized log-likelihood is maximized.

3.2.1 E-step

The E-step requires the computation of $\mathbb{E}(\text{vec} (A_i) | Y_i; \theta)$ and $\mathbb{E}(\text{vec} (A_i) \text{vec} (A_i)^t | Y_i; \theta)$. This is achieved by noticing that the conditional density $p(\text{vec} (A_i) | Y_i; \theta)$ is Normal. Updating formulae for the quantities of interest are thus easily derived as follows:

$$\hat{\Gamma}_i = \mathbb{V}(\text{vec} (A_i) | Y_i; \theta) = \left[ (I_r \otimes Z_i) (\Sigma \otimes I_{n_i})^{-1} (I_r \otimes Z_i) + \Psi^{-1} \right]^{-1}$$

(5)

$$\hat{\text{vec}} (A_i) = \mathbb{E}(\text{vec} (A_i) | Y_i; \theta) = \hat{\Gamma}_i \left( I_r \otimes Z_i \right) (\Sigma \otimes I_{n_i})^{-1} (\text{vec} (Y_i) - (I_r \otimes X_i) \text{vec} (B))$$

(6)

And, consequently, the second moment $\hat{R}_i = \mathbb{E}(\text{vec} (A_i) \text{vec} (A_i)^t | Y_i; \theta)$ reads:

$$\hat{R}_i = \hat{\Gamma}_i + \text{vec}(A_i) \text{vec}(A_i)^t$$

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

$$Y_i|\Lambda_i \sim mN(X_iB + Z_iA_i, I_{n_i}, \Sigma),$$

where $X_iB + Z_iA_i$ is the $n_i \times r$ mean matrix, and $I_{n_i}, \Sigma$ respectively identify the row and column covariance matrices (Dawid, 1981). Such a representation will be useful in specifying the update for $B$ in the devised M-step: details are provided in the next section.

### 3.2.2 M-step

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

$$\ell_{C\text{ pen}}(\theta) = \sum_{i=1}^{J} \log(p(\text{vec}(Y_i) | \text{vec}(\Lambda_i); B, \Sigma)) + \log(p(\text{vec}(\Lambda_i); \Psi)) - p(B; \lambda) =$$

$$= \sum_{i=1}^{J} -\frac{n_i}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma \otimes I_{n_i}| - \frac{1}{2} \mathbb{E}(e_i'(\Sigma \otimes I_{n_i})^{-1}e_i|Y_i, \theta) +$$

$$- \frac{n_i}{2} \log(2\pi) - \frac{1}{2} \log|\Psi| - \frac{1}{2} \mathbb{E}(\text{vec}(\Lambda_i)'\Psi^{-1}\text{vec}(\Lambda_i)|Y_i, \theta) - p(B; \lambda),$$

where $e_i = \text{vec}(Y_i) - (I_r \otimes X_i)\text{vec}(B) - (I_r \otimes Z_i)\text{vec}(A_i)$ and the maximization is performed wrt $\theta = \{B, \Sigma, \Psi\}$.

The updating formula for $B$ clearly depends on the considered $p(B; \lambda)$ 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 $B$ reads:

$$Q_B(B) = -\frac{1}{2} \sum_{i=1}^{J} \text{tr} \left( \Sigma^{-1/2} \left( \hat{Y}_i - X_iB \right)' \left( \hat{Y}_i - X_iB \right) \right) - p(B; \lambda),$$

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

$$\minimize_{B \in \mathbb{R}^{p \times r}} \frac{1}{2} \sum_{i=1}^{J} \left\| \Sigma^{-1/2} \left( \hat{Y}_i - X_iB \right)' \right\|_{F}^{2} + p(B; \lambda)$$

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

$$
\hat{B} = \arg \min_B \frac{1}{2} \sum_{i=1}^{J} \left\| \Sigma^{-1/2} \tilde{Y}_i - X_i B \right\|_F^2 + p(B; \lambda). \tag{12}
$$

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

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

where $\hat{B}$ maximizes (10). This procedure stems from the rationale outlined in Rohart et al. (2014), where, contrarily to their original solution, in our context the updating steps are made more complex by the multidimensional nature of $Y$. The devised updating scheme allows to easily incorporate any $p(B; \lambda)$ 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 $\Psi$ and $\Sigma$ agree with those of the unpenalized framework, namely

$$
\hat{\Psi} = \frac{1}{J} \sum_{i=1}^{J} \hat{R}_i, \tag{14}
$$

and

$$
\hat{\Sigma}_{hk} = \frac{1}{N} \sum_{i=1}^{J} \left[ \mathbb{E}(E_{ih}|Y_i)^{\prime} \mathbb{E}(E_{ik}|Y_i) \right] + \text{tr} \left[ \text{cov}(E_{ih}, E_{ik}|Y_i) \right], \quad h, k = 1, \ldots, r. \tag{15}
$$

where $E_{ih}$ denotes the $h$-th column of the matrix $E_i = Y_i - Z_i \hat{\Lambda}_i - X_i B_i$, $h = 1, \ldots, r$.

### 3.3 On the choice of $p(B; \lambda)$

The EM algorithm devised in the previous section defines a general-purpose optimization strategy for penalized mixed-effects multitask learning. Nonetheless, in practice, a functional form for $p(B; \lambda)$ must be chosen when performing the analysis. While any penalty type can in principle be defined, three notable examples, commonly used in this context, are the elastic net penalty, the group-lasso penalty for multivariate regression (Example 4.2 in Hastie et al. (2015)) and the netReg routines for Network-regularized linear models (Dirmeier et al., 2018): each of them is briefly described in the next subsections.

#### 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 $\alpha$, $0 \leq \alpha \leq 1$. In details, the penalty expression
reads:

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

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

### 3.3.2 Group-lasso penalty

This type of penalty imposes a group structure on the coefficients, forcing the same subset of predictors to be preserved across all \( r \) components of the response matrix. This feature is particularly desirable when building multivariate DNAm biomarkers, as reported in Section 4, since it automatically identifies the CpG sites that are \textit{jointly} related to the considered risk factors. Such a penalty is defined as follows:

\[ p(B; \lambda, \alpha) = \lambda \left[ (1 - \alpha) \sum_{c=1}^{r} \sum_{l=2}^{p} b_{lc}^2 + \alpha \sum_{l=2}^{p} \|b_l\|_2 \right], \quad (17) \]

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

### 3.3.3 Network-regularized penalty

The last penalty we consider allows for the inclusion of biological graph-prior knowledge in the estimation by accounting for the contribution of two non-negative adjacency matrices \( G_X \in \mathbb{R}^{(p-1) \times (p-1)} \) and \( G_Y \in \mathbb{R}^{r \times r} \), respectively related to \( X \) and \( Y \). In this case, \( p(B; \lambda) \) assumes the following functional form:
Table 1: Root Mean Squared Error (RMSE) and active number of CpG sites for different penalized regression models, EPIC Italy test set. Bold numbers indicate lowest RMSE for each of the \( r = 5 \) dimension of the response matrix.

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

\[
p(B; \lambda, \lambda_X, \lambda_Y) = \lambda_1|B_0||1 + \lambda_X \text{ tr} \left(B_0'(D_{G_X} - G_X)B_0\right) + \lambda_Y \text{ tr} \left(B_0'(D_{G_Y} - G_Y)B_0\right)
\]

where \( B_0 \) is the \((p - 1) \times r \) matrix of coefficients without the intercepts and \( D_{G_X}, D_{G_Y} \) indicate the degree matrices of \( G_X \) and \( G_Y \), respectively (Chung and Graham, 1997). \( G_X \) and \( G_Y \) encode a biological similarity, forcing rows and columns of \( B_0 \) to be similar. The netReg R package provides a convenient implementation of (18) (Dirmeier et al., 2018).

### 3.4 Further aspects

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

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

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

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

where \( \hat{\theta}^{(t)} = \{B^{(t)}, \Sigma^{(t)}, \Psi^{(t)}\} \) is the set of estimated values at the end of the \( t \)-th iteration. In our analyses, \( \varepsilon \) is set equal to \( 10^{-6} \).

- **Model selection**: a standard 10-fold cross validation (CV) strategy is employed for selecting the tuning factors in the application presented in Section 4. Alternatively, as suggested in Rohart et al. (2014), one could employ a modified version of the Bayesian Information Criterion (BIC,
where \( \ell(\hat{\theta}) \) is the log-likelihood evaluated at \( \hat{\theta} \), obtained maximizing (4), and \( d_0 \) is the number of non-zero parameters resulting from the penalized estimation. Another option would be to rely on an interval search algorithm, like the efficient parameter selection via global optimization (Frohlich and Zell, 2005): an implementation is available in the c060 R package (Sill et al., 2014).

- **Implementation**: routines for fitting the penalized mixed-effects multitask learning method have been implemented in R (R Core Team, 2021), and the source code is freely available at https://github.com/AndreaCappozzo/emlmm in the form of an R package. The three penalties described in Section 3.3 are included in the software, and can be selected via the `penalty_type` argument of the `ecmmlmmpenalized` 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 effect**: model in (1) assumes that each and every response requires a random-effect component. Whilst in principle reasonable, it may happen in specific applications that only a subset of the \( r \) characteristics in \( Y \) enjoys group-dependent heterogeneity. The occurrence of such a scenario can be unveiled by looking at the \( r \) diagonal elements of dimension \( q \) in \( \hat{\Psi} \): an heuristic procedure for detecting group-independent responses is to locate elements in \( \text{diag}(\hat{\Psi}) \) whose magnitude is significantly lower than the remaining ones. Thus, by setting a user-defined threshold, the impact random effects have on the different characteristics is retrieved as a by-product of the modeling procedure.

## 4 DNAm biomarkers creation from EPIC dataset

The methodology described in the previous section is employed to build a 5-dimensional DNAm biomarker of hypertension and hyperlipidemia. As mentioned in the introduction, DNAm surrogates possess extensive advantages over their blood-measured counterparts, since they directly account for subject-specific response to risk factors. Furthermore, once the DNAm biomarkers have been created (i.e., model parameters have been estimated), their values can immediately be predicted for patients not directly involved in the study, even coming from an external cohort: it is only sufficient to acquire a blood sample and information about sex and age. Lastly, through a pathway enrichment analysis (Reimand et al., 2019), deep epidemiological understanding can be uncovered by investigating which regressors are involved in the surrogate construction (i.e., the CpG sites whose associated parameters are not shrunk to 0).
To reconstruct the process of DNAm surrogates creation and validation, the EPIC Italy data is randomly split into two sets: 70% ($N_{tr} = 401$) of it is employed for model fitting, while the remaining 30% ($N_{te} = 173$) acts as test set for assessing predictive performance. Several estimation strategies are contemplated varying penalty and modeling type. For each model, the penalty factor $\lambda$ was tuned via 10-fold CV on the training set, while the mixing parameter $\alpha$ was kept fixed and equal to 0.5. Results are summarized in Table 1, where the Root Mean Squared Error (RMSE) and the number of active CpG sites are reported. The first two rows are related to the novel methodology introduced in the present paper, for which we consider 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. We immediately notice that our proposal outperforms the state-of-the-art approach (univariate elastic-net) for 4 out of 5 dimensions of the response variable. The reason being that our method takes advantage of the borrowing information asset typical of multivariate models (the correlation between SBP and DBP is equal to 0.77 in the training set), whilst allowing for center-wise difference to be captured by the random intercept. Furthermore, thanks to the behavior of the group-lasso penalty, our proposal 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. The network-regularized penalty has not been included in the comparison as the incorporation of prior knowledge through graph-based regularizers does not seem to be suited for this context, with predictive metrics being much worse for both Random-effects and Fixed-effects models. Figure 2 reports the residuals vs fitted plots for the model in the first row of Table 1:
each dimension displays a more than satisfactory diagnostic pattern.

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

5 Discussion and further work

In the present paper we have proposed a novel framework for mixed-effects multitask learning suitable for high-dimensional data. By resorting to penalized likelihood estimation, we have devised a general purpose EM algorithm capable of accommodating any penalty type that has been previously developed 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 effect, and other more general issues concerning initialization, convergence and model selection.

The work has been motivated by the problem of developing a multivariate DNA methylation biomarker of cardiovascular and high blood pressure comorbidities from a multi-centric sample. The EPIC Italy dataset has been analyzed using Diastolic Blood Pressure, Systolic Blood Pressure, High Density Lipoprotein, Low Density Lipoprotein and Triglycerides as response variables, regressing them on 13449 CpG sites and accounting for between-center heterogeneity. Our modeling framework, coupled with a group-lasso penalty, has demonstrated to outperform the state-of-the-art alternative, both in terms of predictive power and epidemiological interpretation. Such a result may thereupon favor the adoption of our methodological approach for building DNA methylation surrogates.

A direction for future research concerns promoting the application of the proposed procedure in creating additional multi-dimensional DNA methylation biomarkers, conveniently embedding random effects and customized penalty types. Further validating the applicability of this methodology will aid the data-driven progress of personalized medicine, ultimately meliorating the life quality of the entire society.

References

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

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-Santamaría V, Rodríguez-Palmero A, Villar-Garcia J, Horcajada JP, Albà S, Casasnovas C, Rull A, Reverte L, Dietl B, Dalmau D, Arranz MJ, Llucià-Carol L,
Planas AM, Pérez-Tur J, Fernandez-Cadenas I, Villares P, Tenorio J, Cob- 
bran R, Martín-Nalda A, Soler-Palacin P, Vidal F, Pujol A, Esteller M (2021) 
Epigenome-wide association study of COVID-19 severity with respiratory fail- 
ure. EBioMedicine 66:103339

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

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

Chung FRK, Graham FC (1997) Spectral graph theory. 92, American Mathe- 
matical Soc.

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

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

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

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

Dempster A, N Laird, Rubin D (1977) Maximum likelihood from incomplete 
data via the EM algorithm. Journal of the Royal Statistical Society 39(1):1– 
38

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

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

Gagliardi A, Dugué PA, Nøst TH, Southey MC, Buchanan DD, Schmidt DF, 
Makalic E, Hodge AM, English DR, Doo NW, Hopper JL, Severi G, Baglietto 
L, Naccarati A, Turallo S, Pace L, Krog N, Pulli D, Panico S, Sacerdote C, 
Tumino R, Lund E, Giles GG, Pardini B, Sandanger TM, Milne RL, Vineis P, 
Polidoro S, Fiorito G (2020) Stochastic Epigenetic Mutations Are Associated 
with Risk of Breast Cancer, Lung Cancer, and Mature B-cell Neoplasms. 
Cancer Epidemiology Biomarkers & Prevention 29(10):2026–2037
Guida F, Sandanger TM, Castagné R, Campanella G, Polidoro S, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, Severi G, Kyrtopoulos SA, Georgiadis P, Vermeulen RC, Lund E, Vincis P, Chadeau-Hyam M (2015) Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Human Molecular Genetics 24(8):2349–2359

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

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

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

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

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

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

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

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

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

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

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

Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, Merico D, Bader GD (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols 14(2):482–517
Reinsel G (1984) Estimation and Prediction in a Multivariate Random Effects Generalized Linear Model. Journal of the American Statistical Association 79(386):406–414

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

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

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

Schelldorfer J, Bühlmann P, De Geer SV (2011) Estimation for High-Dimensional Linear Mixed-Effects Models Using L1-Penalization. Scandinavian Journal of Statistics 38(2):197–214

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

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

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

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

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

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

Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1):267–288
Vinga S (2021) Structured sparsity regularization for analyzing high-dimensional omics data. Briefings in Bioinformatics 22(1):77–87

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

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

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

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

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

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