Mixed-effects high-dimensional multivariate regression via group-lasso regularization

Regressione multivariata con effetti misti per dati ad alta dimensionalità: un approccio con regolarizzazione di tipo group-lasso

Francesca Ieva, Andrea Cappozzo, and Giovanni Fiorito

Abstract Linear mixed modeling is a well-established technique widely employed when observations possess a grouping structure. Nonetheless, this standard methodology is no longer applicable when the learning framework encompasses a multivariate response and high-dimensional predictors. To overcome these issues, in the present paper a penalized estimation procedure for multivariate linear mixed-effects models (MLMM) is introduced. In details, we propose to regularize the likelihood via a group-lasso penalty, forcing only a subset of the estimated parameters to be preserved across all components of the multivariate response. The methodology is employed to develop novel surrogate biomarkers for cardiovascular risk factors, such as lipids and blood pressure, from whole-genome DNA methylation data in a multi-center study. The described methodology performs better than current state-of-art alternatives in predicting a multivariate continuous outcome.

Abstract *I modelli ad effetti misti sono ampiamente utilizzati nell'analisi di dati che possiedono una struttura a gruppi. Tuttavia, tale metodologia non è applicabile in contesti dove la variabile risposta è multidimensionale ed il numero di regressori elevato. Nel proporre una soluzione ai sopracitati problemi, nel presente lavoro viene introdotta una procedura di stima penalizzata per modelli ad effetti misti con risposta multivariata. In dettaglio, si propone di regolarizzare la verosimiglianza tramite una penalità di tipo group-lasso, forzando solo un sottoinsieme dei parametri stimati ad essere diverso da 0 per ogni componente della variabile risposta. La metodologia proposta viene poi utilizzata per creare nuovi surrogati per fattori di rischio cardiovascolare, come lipidi e pressione sanguigna, dai dati di metilazione del DNA dell'intero genoma in uno studio multicentrico. L'analisi così condotta dimostra risultati migliori rispetto alle attuali alternative nella previsione di un outcome continuo multivariato.*

Francesca Ieva, Andrea Cappozzo
MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano, e-mail: francesca.ieva@polimi.it, andrea.cappozzo@polimi.it

Giovanni Fiorito
Department of Biomedical Sciences, Università di Sassari e-mail: gfiorito@uniss.it

Key words: Mixed-effects models, Multivariate regression, group-lasso penalty, penalized estimation

1 Introduction and motivation

Multivariate regression performs joint learning of a multidimensional response on a common set of predictors. When samples possess a hierarchical/temporal structure, data independence cannot be assumed a-priori and thus a Multivariate Mixed-Effects Model (MLMM) must be adopted [5]. An MLMM framework thus allows for the inclusion of a grouping structure within the model specification, a situation that often arises in multi-centric and/or longitudinal studies. With the advent of modern technologies, it is more and more common nowadays that in such studies a huge number of features is recorded, often greatly exceeding the available sample size. To this extent, regularization methods based on penalized estimation have been fruitfully adopted to overcome the resulting over-parameterization issue [7]. In particular, for univariate mixed-effects models, ℓ_1- penalization schemes have been devised to perform selection of fixed effects when dealing with high-dimensional data [4, 3]. By suitably leveraging the methodology proposed in [3], we extend it to the multivariate response framework including a group-lasso penalty in the model specification.

The remainder of the paper proceeds as follows: in Section 2 we introduce our new proposal and we discuss its main methodological aspects. Section 3 presents an application of our model in creating surrogate scores based on blood DNA methylation. Section 4 summarizes the novel contributions and highlights future research directions.

2 Group-lasso regularized mixed-effects multivariate regression

In an MLMM framework, the data-generating process for the n_j units in group j, with $\sum_{j=1}^{J} n_j = N$ and J the total number of groups, is assumed to be as follows:

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

where $\boldsymbol{Y}_j$, $\boldsymbol{X}_j$, $\boldsymbol{Z}_j$ respectively define the response, fixed and random effects design matrices. Further, $\boldsymbol{B}$ denotes the matrix of fixed coefficients, $\boldsymbol{\Lambda}_j$ the matrix of random effects in group j and $\boldsymbol{E}_j$ the group specific error term. The following distributions are assumed for the random quantities in (1):

$$\operatorname{vec}(\boldsymbol{\Lambda}_j) \sim \mathscr{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \quad \operatorname{vec}(\boldsymbol{E}_j) \sim \mathscr{N}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}), \quad j = 1, \ldots, J$$

with $\operatorname{vec}(\cdot)$ denoting the vec operator, $\boldsymbol{\Psi}$ is a positive semidefinite matrix incorporating variations and covariations between the responses and the random effects and

$\boldsymbol{\Sigma}$ is a covariance matrix capturing column-wise dependence in the multivariate error term $\boldsymbol{E}_j$. Thereupon, the distribution of the vectorized response can be written as follows:

$$\operatorname{vec}(\boldsymbol{Y}_j) \sim N\left((\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}), (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right), \quad j = 1, \ldots, J.$$

When dealing with high-dimensional data, the number of regressors (i.e., the rows of matrix $\boldsymbol{B}$) is generally much larger than the sample size N. Therefore, in order to still be able to make sensible inference on the parameters $\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\}$, we propose to maximize the following penalized log-likelihood:

$$\begin{aligned}\ell_{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \log \phi\left(\operatorname{vec}(\boldsymbol{Y}_j), (\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}), (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right) + \\ - \lambda\left[(1-\alpha)\sum_{c=1}^{r}\sum_{l=2}^{p} b_{lc}^2 + \alpha \sum_{l=2}^{p} ||\boldsymbol{b}_{l\cdot}||_2\right],\end{aligned} \tag{2}$$

where b_{lc} and $\boldsymbol{b}_{l\cdot}$ denote the element in position (l,c) and the l-th row of matrix $\boldsymbol{B}$, respectively. The penalty in (2) behaves like the lasso but on a whole group of coefficients. In details, for each covariate, the estimated parameters are either all zero or none are zero, and this behavior is preserved across all components of the response variable. This characteristic is particularly desirable when it comes to variable selection in multivariate regression, since features that are jointly related to the multidimensional response are automatically identified. The amount of shrinkage is determined by the penalty factor λ, whilst the mixing parameter α controls the weight associated to ridge and group-lasso regularizers. Maximization of (2) is performed via a tailored EM-type algorithm [1], in which standard fixed-effects routines are conveniently exploited within the M-step.

The devised framework is employed to build a multidimensional predictor of systolic and diastolic blood pressure, LDL and HDL cholesterol based on blood DNA methylation (DNAm): results are reported in the next section.

3 Application to DNAm biomarkers creation

DNAm biomarkers are obtained by regressing blood measured quantities (response variables) on methylation levels within CpG sites in the DNA sequence (dependent variables) [6]. The aim of this section is to build a multivariate DNAm biomarker for cardiovascular risk factors and comorbidities, considering Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), High Density Lipoprotein (HDL) and Low Density Lipoprotein (LDL) as responses, regressing them onto 13449 CpG sites (top 1% p-value based ranking) adjusting for sex and age. The employed dataset comes from the Italian component of the European Prospective Investigation into

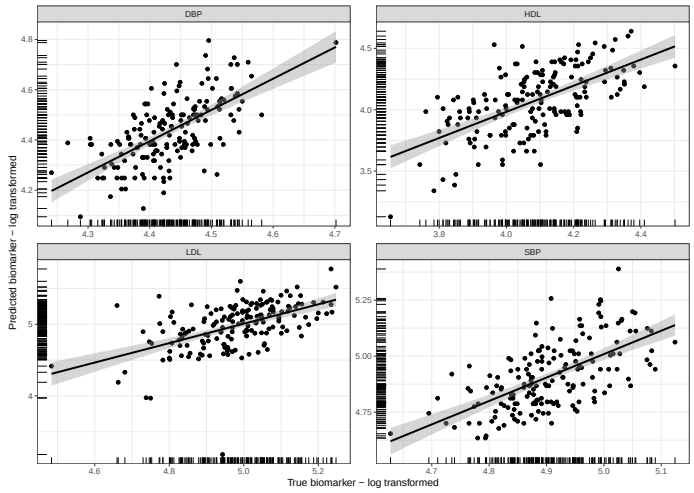

Fig. 1 Observed vs fitted scatterplots for the estimated biomarkers, namely log-transformed Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL) and Systolic Blood Pressure (SBP), EPIC Italy test set. Linearly smoothed conditional means and associated standard deviations are superimposed in each facet.

Cancer and Nutrition (EPIC) study [2], comprised of $J = 4$ geographical sub-cohorts identified by the centre of recruitment. We employ $N_{tr} = 401$ training samples to fit the model in (2) including a random intercept component, validating its performance on $N_{te} = 173$ test units. The root mean squared error (RMSE), computed on the test set for the four-dimensional response, is reported in Table 1. Together with our proposal (denoted as MLMM Group-lasso in the table), results for two competing methods are reported, namely fixed-effect group lasso and univariate elastic-net [8]. For each method, the penalty factor λ was tuned via 10-fold CV on the training set, while the the mixing parameter α was kept fixed and equal to 0.5.

As it clearly stands out from Table 1, our proposal achieves better predictive performances for all components in the response variable with respect to the competing models. The reason behind this result is two-fold. On one hand MLMM Group-lasso performs better than its fixed-effects counterpart as the heterogeneity induced by the centre of recruitment is properly taken into account by means of a random intercept. On the other hand, solving the four regression problems jointly and imposing a group structure on the coefficients leads to better prediction performance than fitting four univariate models separately as done for the elastic-net procedure. The good predictive performance of the proposed model is highlighted in Figure 1, where we report for each biomarker the observed vs fitted scatterplots. All components of the response exhibits positive linear correlation between measured and predicted values in the test set, with Pearson's correlation coefficients always higher than 0.5.

The employment of the MLMM Group-lasso not only produces moderate improvements in terms of prediction accuracy, but it is also supported by biological

Table 1 Root Mean Squared Error (RMSE) for different penalized regression models, EPIC Italy test set. Bold numbers indicate lowest RMSE for each component of the four dimensional response.

Model	DBP	HDL	LDL	SBP
MLMM Group-lasso	**0.102**	**0.2139**	**0.278**	**0.1172**
Group-lasso	0.112	0.2238	0.286	0.1229
Univariate elastic-net	0.1064	0.2292	0.2884	0.1271

reasons. In fact, the pleiotropic effect suggests that multiple correlated phenotypes will likely affect the same set of CpG sites, motivating the adoption of a group-lasso penalty. Furthermore, DNAm biomarkers creation stands on the rationale that the resulting surrogate should be study-invariant: by incorporating a random intercept in the model specification the center effect can still be captured, while maintaining generalizability of the method to external cohorts.

4 Conclusion

The present work has introduced a novel penalized mixed-effects multivariate regression framework, able to model a multidimensional response with high-dimensional covariates and grouped data structure. By means of a group-lasso regularizer, we have achieved excellent predictive accuracy when creating a DNAm surrogate of cardiovascular risk factors, outperforming state-of-the-art alternatives. Such surrogates possess some advantages over their blood-measured counterparts, as they can directly take into account genetic susceptibility and subject specific response to risk factors.

In the devised framework we have implicitly assumed low-dimensionality in the response variable. A direction for future research may concern the inclusion of custom penalties to cope with situations in which both the response and the design matrix are high-dimensional. Feasible solutions are currently being investigated and they will be the object of future work.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the EM Algorithm. J. R. Stat. Soc. Ser. B **39**(1), 1–22 (1977) doi:10.1111/j.2517-6161.1977.tb01600.x
2. Riboli, E., Hunt, K., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., Vineis, P., Palli, D., Bueno-de Mesquita, H., Peeters, P., Lund, E., Engeset, D., González, C., Barricarte, A., Berglund, G., Hallmans, G., Day, N., Key, T., Kaaks, R., Saracci, R.: European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. **5**(6b), 1113–1124 (2002) doi:10.1079/PHN2002394

3. Rohart, F., San Cristobal, M., Laurent, B.: Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. Comput. Stat. Data Anal. **80**, 209–222 (2014) doi:10.1016/j.csda.2014.06.022
4. Schelldorfer, J., Bühlmann, P., De Geer, S.V.: Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. Scand. J. Stat. **38**(2), 197–214 (2011) doi:10.1111/j.1467-9469.2011.00740.x
5. Shah, A., Laird, N., Schoenfeld, D.: A Random-Effects Model for Multiple Characteristics with Possibly Missing Data. J. Am. Stat. Assoc. **92**(438), 775–779 (1997) doi:10.1080/01621459.1997.10474030
6. Singal, R., Ginder, G.D.: DNA Methylation. Blood **93**(12), 4059–4070 (1999) doi:10.1182/blood.V93.12.4059
7. Vinga, S.: Structured sparsity regularization for analyzing high-dimensional omics data. Brief. Bioinform. **22**(1), 77–87 (2021) doi:10.1093/bib/bbaa122
8. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodol. **67**(5), 768–768 (2005) doi:10.1111/j.1467-9868.2005.00527.x