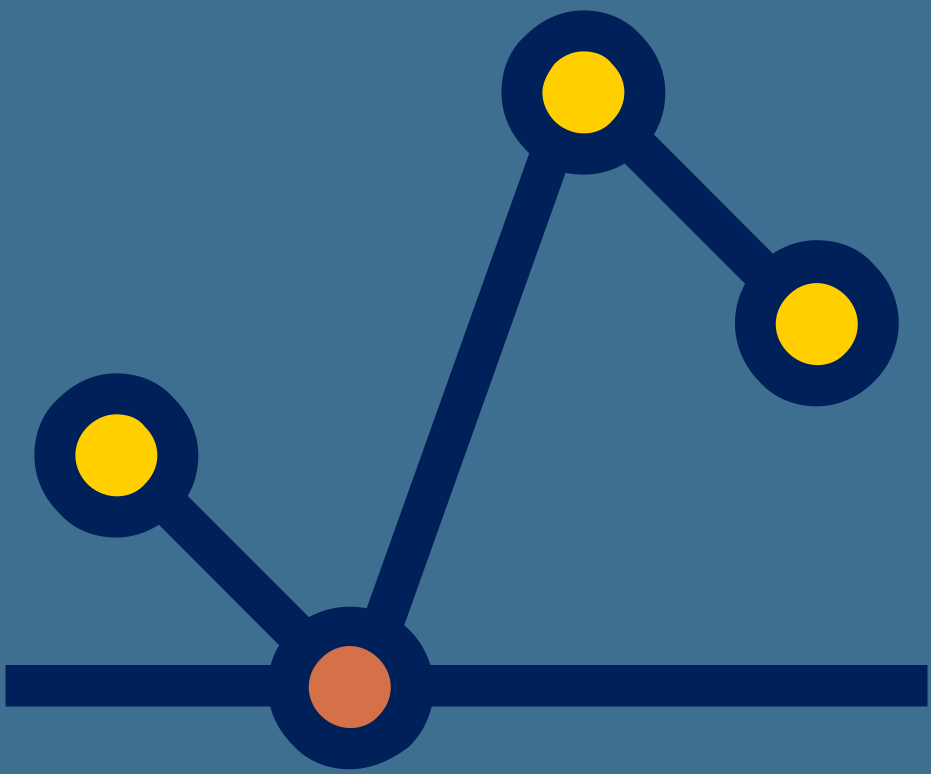
Edited by
Paola Cerchiello · Arianna Agosto
Silvia Osmetti · Alessandro Spelta

Proceedings of the Statistics and Data Science Conference



PaviaUniversityPress

Copertina: Cristina Bernasconi, Milano

Via Salasco, 5 - 20136 Milano
Tel. 02/5836.5751 - Fax 02/5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

Pavia University Press
info@paviauniversitypress.it – www.paviauniversitypress.it

Prima edizione: maggio 2023
ISBN volume 978-88-6952-170-6

Preface

The development of large-scale data analysis and statistical learning methods for data science is gaining more and more interest, not only among statisticians, but also among computer scientists, mathematicians, computational physicists, economists, and, in general, all experts in different fields of knowledge who are interested in extracting insight from data.
Cross-fertilization between the different scientific communities is becoming crucial for progressing and developing new methods and tools in data science.
In this respect, the Statistics & Data Science group of the Italian Statistical Society has organized an international conference held in Pavia on the 27 and 28 of April 2023, attended by over 70 researchers from different scientific fields.
A collection of the presented papers is available in the present Proceedings showing a huge variety of approaches, methods, and data-driven problems, always tackled according to a rigorous and robust scientific paradigm.

The Statistics & Data Science group

Contents

Streamlined Variational Inference for Modeling Italian Educational Data

Gioia Di Credico, Claudia Di Caterina, Francesco Santelli

Abstract The streamlined version of the mean field variational Bayes (MFVB) algorithm for linear mixed models with crossed random effects allows simplifying calculations but may require one group's dimension to be moderate.Data collecting high school students' first term evaluations and INVALSI scores for Italian and Maths subjects perfectly comply with this setting: students are a vast random sample of those who enrolled at the university in 2019/20, while the number of tests is limited to 6. Three different MFVB product restrictions with incremental complexity are evaluated. All of them are convenient with respect to classic MCMC solutions from both a computational and a memory storage viewpoint. Results and interpretation of model coefficients are in line with the literature on educational data.

Key words: Crossed random effects, INVALSI, Mean field variational Bayes.

1 Introduction

Linear mixed-effects models are commonly used to analyze data with a continuous Gaussian outcome and multiple levels of variability arising from a grouped data structure. In order to account for the variability introduced by the nested or crossed structure of the observations, it may be convenient to include random effects treated as random variables in the model.

In the following, we focus on a crossed-data application where two levels of variability exist. Crossed designs imply that each combination of the group levels is represented in the data [1]. Our data refer to a random sample of students who enrolled in a Bachelor program at an Italian university in the academic year 2019/2020. Out-

Gioia Di Credico
Università degli studi di Trieste, Trieste, e-mail: gioia.dicredico@deams.units.it

Claudia Di Caterina
Università degli studi di Verona, Verona, e-mail: claudia.dicaterina@univr.it

Francesco Santelli
Università degli studi di Trieste, Trieste, e-mail: fsantelli@units.it

comes consist of students' evaluations during their 10th and 13th high school grades. Here, students and the type of tests define our two-group crossed-data design. As the groups size increases, model estimation gets slower and may even become unfeasible. Streamlined variational inference has recently been applied to overcome these estimation difficulties in random-effects models, e.g. by [3] for the nested group structure and by [4] for the crossed one. The key idea relies on the sparseness of the matrix to be inverted, which enables a quicker computation and less storage capacity. Nested data imply a so-called arrowhead block structure for this matrix [5], and non-zero sub-blocks can be easily identified to simplify the calculation of its inverse. In the crossed-data case studied here the matrix is less sparse, however, considering the most accurate restriction, the streamlined approach offers advantages when one group is moderate in size. Our motivating data exhibit such a feature: the number of students involved is very large (around 7000), while the tests whose mark is recorded on each student are limited to 6.

2 Methods

For each ith student, we assume the scores $\boldsymbol{y}_{ii'}$ on test i' follows a linear mixed model with two crossed random effects:

$$\begin{aligned} \boldsymbol{y}_{ii'}|\boldsymbol{\beta},\boldsymbol{u}_i,\boldsymbol{u}'_{i'},\sigma^2 &\overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{ii'}\boldsymbol{\beta}+\boldsymbol{Z}_{ii'}\boldsymbol{u}_i+\boldsymbol{Z}'_{ii'}\boldsymbol{u}'_{i'},\sigma^2\boldsymbol{I}), \quad i=1,\dots,m, \\ \boldsymbol{u}_i|\boldsymbol{\Sigma} \overset{\text{ind.}}{\sim} N(0,\boldsymbol{\Sigma}), &\qquad \boldsymbol{u}'_{i'}|\boldsymbol{\Sigma}' \overset{\text{ind.}}{\sim} N(0,\boldsymbol{\Sigma}'), \quad i'=1,\dots,m', \end{aligned} \tag{1}$$

where $\boldsymbol{X}_{ii'}$ is the $n_{ii'}\times p$ design matrix, $\boldsymbol{Z}_{ii'}$ and $\boldsymbol{Z}'_{ii'}$, respectively of dimension $n_{ii'}\times q$ and $n_{ii}\times q'$, are the random effects matrices, $\boldsymbol{\beta}$ is the p-vector of fixed-effect coefficients, $\boldsymbol{u}_i$ and $\boldsymbol{u}'_{i'}$, respectively $q\times 1$ and $q'\times 1$, are the vectors of random effects, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are their $q\times q$ and $q'\times q'$ respective covariance matrices and σ^2 is the error variance.

The joint *a priori* density of the p fixed effects is $\boldsymbol{\beta}\sim N_p(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$. For the error variance σ^2 and the random effects covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$, we consider the following family of marginally non-informative prior distributions [2]:

$$\begin{aligned} \sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(\nu_{\sigma^2},1/a_{\sigma^2}), &\quad a_{\sigma^2}\sim \text{Inverse-}\chi^2(1,1/(\nu_{\sigma^2}s^2_{\sigma^2})), \\ \boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}} &\sim \text{Inverse-G-Wishart}(G_{\text{full}},\nu_{\boldsymbol{\Sigma}}+2q-2,\boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}}), \\ \boldsymbol{\Sigma}'|\boldsymbol{A}_{\boldsymbol{\Sigma}'} &\sim \text{Inverse-G-Wishart}(G_{\text{full}},\nu_{\boldsymbol{\Sigma}'}+2q'-2,\boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}'}), \\ \boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}},1,\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}), &\quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}=\{\nu_{\boldsymbol{\Sigma}}(s^2_{\boldsymbol{\Sigma},1},s^2_{\boldsymbol{\Sigma},2})\}^{-1}, \\ \boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}(G_{\text{diag}},1,\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}'}}), &\quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}'}}=\left\{\nu_{\boldsymbol{\Sigma}'}(s^2_{\boldsymbol{\Sigma}',1},s^2_{\boldsymbol{\Sigma}',2})\right\}^{-1}. \end{aligned} \tag{2}$$

In our application, the first group of random effects $\boldsymbol{u}_i$ ($i=1,\dots,m=7005$) corresponds to students enrolled at an Italian university in 2019/2020, and the second group $\boldsymbol{u}'_{i'}$ ($i'=1,\dots,m'=6$) corresponds to scores from assessments of Italian and

Math skills. Specifically, for each student, a written and oral score was recorded at the end of the first term and one written standardized INVALSI score (see Section 3) was recorded at the end of the final term. Each combination student/test, corresponding to the pair (i, i'), is observed $n_{ii'} = n = 2$ times, namely in the 10th and 13th grades of high school. The design matrix $\boldsymbol{X}_{ii'}$ has $p = 32$ columns, including the intercept. Moreover, $q = q' = 2$ because we consider both random intercepts and random slopes for the two groups, meaning

$$\boldsymbol{Z}_{ii'} = \boldsymbol{Z}'_{ii'} = \left[1 \; x_{1,ii'j}\right]_{j=1,2},$$

where $x_{1,ii'j} = 1,2$ is the year indicator encoding the two high school grades. According to (1) and this set-up, the two scores of the ith student on the i'th test are modeled to be

$$y_{ii'j}|\boldsymbol{\beta}, u_{0i}, u_{1i}, u'_{0i'}, u'_{1i'}, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + u_{0i} + u_{0i'} + (\beta_1 + u_{1i} + u_{1i'})x_{1,ii'j} + \sum_{k=1}^{31} \beta_k x_{k,ii'j}, \sigma^2)$$

for $j = 1,2$. The formula above shows that this modelling strategy allows for a different intercept and slope for every student/test combination. Heterogeneities among intercepts and slopes are defined by appropriate entries of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$.

We consider three product restrictions on the mean field approximation of the joint conditional density function of all parameters in (1) with covariance priors (2) ([4], Sect. 3):

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}') = \begin{cases} q(\boldsymbol{\beta})q(\boldsymbol{u})q(\boldsymbol{u}')\,q(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{restriction I,} \\ q(\boldsymbol{\beta}, \boldsymbol{u})q(\boldsymbol{u}')\,q(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{restriction II,} \\ q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')\,q(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{restriction III.} \end{cases} \quad (3)$$

Product restriction I has the simplest streamlined implementation and scales well to very large problems, but may produce small posterior variances as it sets all posterior correlations between $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{u}'$ to zero. Conversely, product restriction III allows for a full joint posterior covariance matrix of $(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')$, leading to higher inferential accuracy but challenging computing that can be streamlined for limited m'. A compromise is given by product restriction II, which includes posterior correlations between $\boldsymbol{\beta}$ and $\boldsymbol{u}$, for $\boldsymbol{u}$ larger than $\boldsymbol{u}$. For all the product restrictions, the prior distributions specification 2 leads to a fully factorization of the q-densities related to the covariance matrix and auxiliary variables [4].

The q-density parameters can be obtained using a coordinate ascent iterative algorithm. However, if applied naïvely, the potentially prohibitively large matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u})}$ requires storage and inversion. Product restrictions I, II and III lead to streamlined mean field variational Bayes (MFVB) algorithms with varying degrees of storage and computational overhead (see [4], Sect. 4).

3 Italian students' proficiency data

Data drawn from the Italian 'Anagrafe Nazionale della Formazione Superiore' has been processed according to the research project 'From high school to the job market: analysis of the university careers and the university North-South mobility' carried out by the University of Palermo (head of the research program), the Italian 'Ministero Università e Ricerca', and INVALSI. The dataset is known as MOBYSU. In the Italian School System, the scholastic assessment is in charge of the "Italian national institute for the evaluation of the school system" (INVALSI), which uses a set of standardized tests to evaluate the proficiency of students attending different schools at different years. Several domains are tested, and the main domains are Mathematical skills, English language, Italian language, and Science.

Our data regard the cohort of pupils that finished high school, achieved the Diploma in 2018/19, and then enrolled at university in 2019/20. Such students are more than 240000. To be included, students must have never failed a scholastic year, and must attend high school for the first time in the Italian school system. The response variable refers to students marks: four recorded at the end of the first term (Italian and Math, written and oral) and two throughout the INVALSI test (Italian and Math, written), during their 10th and 13th high school grades. Predictors involved in the analysis are listed in Tab.1. They include information on the socio-economic background, parental occupation and demographics.

Table 1 Model predictors, their description and reference categories.

Variable	Description	Reference
Gender	Male, Female	Female
Age	Reception (one year ahead), Regular	Regular
Nation	Foreigner, Italian	Italian
Student escs (EscsStud)	student socio-economic level	
School escs (EscsSch)	school socio-economic level	
School type (SchTy)	13 categories	Classical Lyceum
Work Mother (Work.M)	5 categories	Unemployed
Work Father (Work.F)	5 categories	Unemployed
Year	School year	
NUTS2 classification (NUTS2)	5 areas	Center
School	Private, Public	Public

4 Analysis and results

First term and INVALSI scores were centered to the national means and INVALSI were also scaled to standardize them to a common range and adapt to the prior distributions setting. Furthermore, we excluded students with missing information so that valid and complete data refer to 21228 students. The final model is fitted on a random sample of 33% of the units, corresponding to 7005 total pupils.

As hyperparameters, we set $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^{10}\boldsymbol{I}$, $\nu_{\sigma^2} = 1$, $\nu_{\boldsymbol{\Sigma}} = \nu_{\boldsymbol{\Sigma}'} = 2$, $s_{\sigma^2} = s^2_{\boldsymbol{\Sigma},1} = s^2_{\boldsymbol{\Sigma},2} = s^2_{\boldsymbol{\Sigma}',1} = s^2_{\boldsymbol{\Sigma}',2} = 10^5$. For each product restriction, we run the MFVB algorithm for 100 iterations. Computational times were about 8 minutes for the MFVB with product restriction I, 11 and 21 times longer for the MFVB with product restriction II and III. The model 1 with prior distributions as in 2 was also simulated through MCMC in Stan. In particular, 4 chains of 2000 iterations each (1000 warm-up; 1000 sampling) were simulated. In the following, MCMC inference is based on the sampling step draws. The running time for the MCMC setup was of 19 hours.

As expected, variational inference on the random effects and error variance components are relatively affected by the MFVB product restriction used, giving very similar results (Tab. 2). Differences between MFVB and MCMC on the tests random effects variability are likely due to a slow convergence of the MCMC chains, advised by a low effective sample size on the $\boldsymbol{\Sigma}'$ parameters. The product restriction impact is evident on the estimated variability of the fixed effects. While approximate posterior means of the fixed effects are the same across product restrictions, the least accurate (I) strongly underestimates the variability, while restrictions II and III lead to very similar results on all the coefficients, except for the year variable (see Fig. 1). On average, fixed marginal effects suggest that male students and those from islands and Southern Italy regions perform worse, while the Northeast is the area with best proficiency. The socio-economic status dimension has a significant positive effect, as expected, both at the individual and school levels. Lyceums record the best scores on average, with Scientific lyceum overperforming all the others. When parents are less involved in demanding jobs, students usually perform better. No clear effects are found for Italian nationality and students one year ahead. In both random intercept and slope, the variability carried by the student group is slightly larger than the item group one.

Table 2 Random effects standard deviation estimates (approximate posterior mean). Square root of diagonal entries of $\hat{\boldsymbol{\Sigma}}$ ($\hat{\boldsymbol{\Sigma}}'$) are denoted by $\hat{\sigma}_1$ ($\hat{\sigma}'_1$) and $\hat{\sigma}_2$ ($\hat{\sigma}'_2$). Correlation between random intercept and slope is denoted by $\hat{\rho}$ ($\hat{\rho}'$).

	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\rho}$	$\hat{\sigma}'_1$	$\hat{\sigma}'_2$	$\hat{\rho}'$	$\hat{\sigma}$
MFVB I	0.677	0.114	0.252	0.540	0.087	-0.578	0.931
MFVB II	0.678	0.114	0.251	0.540	0.087	-0.578	0.931
MFVB III	0.678	0.114	0.251	0.591	0.095	-0.523	0.931
MCMC	0.679	0.114	0.251	0.883	0.189	-0.201	0.931

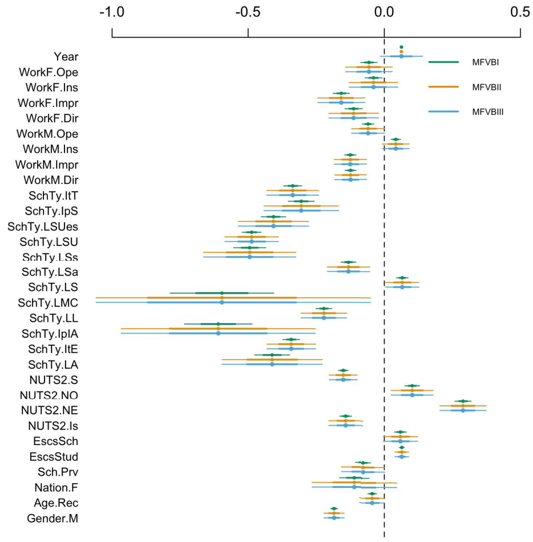


Fig. 1 Fixed effects: approximate posterior means (dots) and 95% credible intervals for the MFVB with the three product restrictions (I in green, II in orange and III in blue).

5 Conclusions

The work analysed Italian students' proficiency data using the streamlined MFVB algorithm based on three product restrictions. The code is not optimized and computational times are reported for comparative purposes only. Even so, the MFVB algorithms are much faster than standard MCMC solutions. Comparing the three MFVB product retrictions, the second appears to be an excellent compromise balancing estimation speed and results accuracy.

References

1. Baayen, R. H., Davidson, D. J., Bates, D. M.: Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. (2008) doi: 10.1016/j.jml.2007.12.005
2. Huang, A., Wand. M. P.: Simple marginally noninformative prior distributions for covariance matrices.Bayesian Anal. (2013) doi: 10.1214/13-BA815
3. Lee, C. Y. Y., Wand, M. P.: Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. Biometrical J. (2016) doi: 10.1002/bimj.201500007
4. Menictas, M., Di Credico, G., Wand, M. P.: Streamlined variational inference for linear mixed models with crossed random effects. J. Comput. Graph. Stat. (2022) doi: 10.1080/10618600.2022.2096622
5. Nolan, T. H., Wand, M. P.: Streamlined solutions to multilevel sparse matrix problems. ANZIAM J. (2020) doi: 10.1017/S1446181120000061