# A comparison of scalable estimation methods for large-scale logistic regression models with crossed random effects

Ruggero Bellio and Cristiano Varin

**Abstract** Parameter estimation of generalized linear models with crossed random effects for large-scale settings is hampered by challenging numerical hindrances. This contribution focuses on logistic regression with crossed-random intercepts and it investigates the properties of two estimation methods for which a scalable software implementation exists, namely the all-row-column and penalized quasi-likelihood methods. The results of a simulation study for sparse settings inspired by e-commerce data, with sample sizes up to $10^6$, suggest that the all-row-column method is preferable over penalized quasi-likelihood.

**Key words:** all-row-column estimation, e-commerce, penalized quasi-likelihood.

## 1 Introduction

Generalized linear models with crossed random effects arise in a variety of settings such as Agriculture, Biology, Epidemiology, Education, Genetics, and Linguistics. Here we are mostly concerned with the analysis of user-generated content arising in e-commerce or recommender systems, where the typical sample sizes largely exceed what is normally encountered in other applications of crossed random effects.

An illustrative example of the setting of interest is reported in Bellio et al. [1] where it is analysed a sample of data from an online personal styling service. The data include a binary response variable $y_{ij}$ that denotes whether customer $i$ rates item $j$ as top fit or not. In the sample data there are around 5,000,000 user ratings,

Ruggero Bellio
University of Udine (Italy), Department of Economics and Statistics, e-mail: ruggero.bellio@uniud.it

Cristiano Varin
Ca' Foscari University of Venice, Department of Environmental Sciences, Informatics and Statistics. e-mail: cristiano.varin@unive.it

produced by more than 700,000 customers for about more than 3,000 clothes. A reasonable statistical model for this data assumes a generalized linear model for binary responses with linear predictor

$$\eta_{ij} = \mathbf{x}_{ij}^{\top}\beta + a_i + b_j \tag{1}$$

where $i = 1,\ldots,R$ is the index for customers, $j = 1,\ldots,C$ is the index for clothes, $\mathbf{x}$ is a vector of covariates with coefficients $\beta$ and, finally, $a_i$ and $b_j$ are mutually independent normally-distributed random intercepts for customers and clothes with variances $\sigma_A^2$ and $\sigma_B^2$, respectively.

The estimation of a generalized linear model for binary responses with crossed random effects is notoriously challenging. The corresponding likelihood function requires to integrate out the random effects, entailing an integral of dimension $R+C$, which is very hard to approximate in broad generality, and even more so when $R$ or $C$ is very large as in e-commerce applications. In such settings, methods requiring Monte Carlo approximations are readily ruled out, since they usually imply a computational cost exceeding $O(N)$, where $N$ is the sample size. Some scalable alternatives to Monte Carlo methods have been developed recently:

1. the pseudo quasi-likelihood method [2, 3] using the scalable algorithm proposed in Ghosh et al. [4];
2. the all-row-col method proposed by Bellio et al. [1];
3. the combination of Gaussian variational approximation and composite likelihood proposed in Xu et al. [5].

While there is software available for the first two methods, there is not yet any software provided for the variational approximation of Xu et al. Moreover, Xu et al. do not give any example of application of their variational approximation for large-scale problems.

We have not included in the list of methods for inference in crossed-random effects the popular first-order Laplace's approximation implemented in very efficient R packages such as `lme4` [6]. Indeed, the results in [1] show that such method may have a computational cost exceeding $O(N)$ in sparse settings, coupled with some possible (minor) estimation bias arising even for very large sample sizes. The latter is far less critical than the computational issue, which rules out this method for the large datasets met in e-commerce applications. Therefore, in this short paper we compare penalized quasi-likelihood and all-row-call for logistic regression with crossed random effects, since they are both scalable methods with public software available.

## 2 Two scalable estimation methods

Penalized quasi-likelihood is a well-established estimation method, which consists in maximising an objective function corresponding to a simplified version of Laplace's approximation [2]. The method entails the iterated resolution of two kinds

of estimating equations, one which jointly updates fixed effects and random effects, while the other one updates the variance components. The method is consistent only when both the number of random effects and the number of observations per random effect increase, and for binary data substantial finite-sample bias may arise. The method is rather general, and can be applied to virtually any mixed-effect model.

The application of the method to large-dimensional settings illustrated in [4] requires some ingenuity to handle the large number of random effects. Indeed, the authors proposed a modified version of Schall's estimation approach [3], making use of an ad-hoc version of backfitting to keep the computational complexity within the $O(N)$ bound. The code supporting [4] is for logistic regression only.

The all-row-column method was recently proposed by Bellio et al. [1] for probit regression with crossed-random effects. The method exploits the fact that the likelihood for the fixed effects in the marginalized model has a simple form in the case of probit link. The all-row-column method proceeds in three steps:

1. the *all step*. The starting point is the marginal regression model
$$Pr(Y_{ij} = 1) = \Phi(\mathbf{x}_{ij}^\top \gamma), \tag{2}$$
where $\gamma = \beta/(1+\sigma_A^2+\sigma_B^2)^{1/2}$. The $\gamma$ coefficients are estimated by fitting a probit regression model for independent data;
2. the *row step*. A probit regression model with only the $a_i$ (the 'rows') effects is considered
$$Pr(Y_{ij} = 1|a_i) = \Phi(\mathbf{x}_{ij}^\top \gamma_{\mathbf{A}} + u_i), \tag{3}$$
where $u_i = a_i/(1+\sigma_B^2)^{1/2} \sim \mathcal{N}(0,\tau_A^2)$, with $\tau_A^2 = \sigma_A^2/(1+\sigma_B^2)$, $\gamma_{\mathbf{A}} = \gamma(1+\tau_A^2)^{1/2}$. The parameter $\tau_A^2$ is estimated by fitting a random intercept model including only the $u_i$ effects with $\gamma$ held fixed at the value $\widehat{\gamma}$ estimated at the all step;
3. the *col step*. A probit regression model with only the $b_j$ (the 'columns') effects is considered
$$Pr(Y_{ij} = 1|b_j) = \Phi(\mathbf{x}_{ij}^\top \gamma_{\mathbf{B}} + v_j), \tag{4}$$
where $v_j = b_j/(1+\sigma_A^2)^{1/2} \sim \mathcal{N}(0,\tau_B^2)$, with $\tau_B^2 = \sigma_B^2/(1+\sigma_A^2)$, $\gamma_{\mathbf{B}} = \gamma(1+\tau_B^2)^{1/2}$. The parameter $\tau_B^2$ is estimated by fitting a random intercept model including only the $v_j$ effects with again $\gamma = \widehat{\gamma}$. The estimates $\widehat{\sigma}_A^2$ and $\widehat{\sigma}_B^2$ are obtained from those of $\tau_A^2$ and $\tau_B^2$. Finally, the estimate of $\beta$ is computed as $\widehat{\beta} = \widehat{\gamma}(1+\widehat{\sigma}_A^2+\widehat{\sigma}_B^2)^{1/2}$.

Bellio et al. [1] proved that method is consistent and with a computational complexity of order $O(N)$. To proceed with a direct comparison of all-row-column with penalized quasi-likelihood, however, one needs either to implement the probit version of penalized quasi-likelihood, or to adapt the all-row-column method to logistic regression. In this short paper, we consider the latter route because we think the logit link may be of interest to practitioners.

In principle, the extension of all-row-column to logit link may proceed by following the results about marginalized models (see for example [7]), but the resulting

numerical burden might be demanding, since $N$ one-dimensional equations involving a convolution integral would have to be solved numerically for every evaluation of the marginalized likelihoods obtained by omitting one of the two random effects. Even though Bellio et al. [1] conjectures that the resulting computation complexity may be still within the $O(N)$ bound, such a route appears impractical at best, so here we suggest an approximate method.

Like before, let $\gamma$ be the coefficients of the marginalized model, namely the logit counterpart of the probit marginalized model (2). Even though there is no exact explicit relation linking $\gamma$ to the parameters of the linear predictor (1), some approximate relations exist. For example, one may follow Wang and Luis [8], obtaining

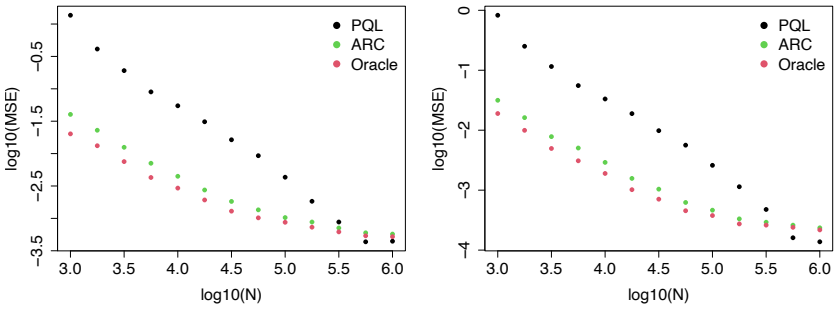
$$\gamma \doteq \frac{\beta}{\sqrt{1+(3/\pi^2)(\sigma_A^2+\sigma_B^2)}}\,. \tag{5}$$

Similar relationships can be defined for the models omitting one of the two random effects that are used in the row and column steps of the all-row-column method, in this way allowing to extend the all-row-column method to logistic regression.

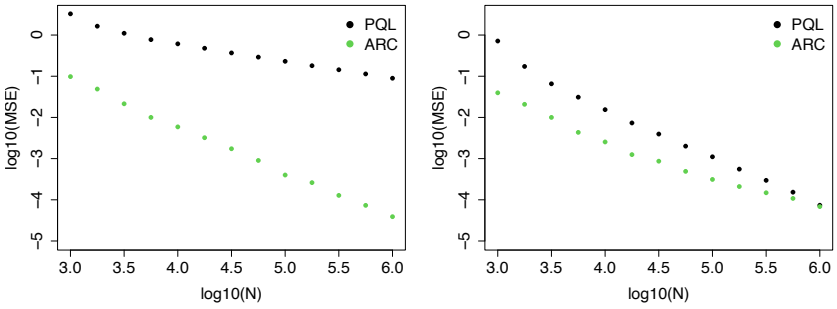
## 3 Some simulation results

Some simulations following closely the studies in Ghosh et al. [4] and Bellio et al. [1] have been carried out. We considered sample sizes $N$ ranging from $10^3$ to $10^6$ and generated 1,000 datasets from model (1) with logit link at each sample size. We simulated seven standard normal correlated predictors and set the intercept to $\beta_0 = -2$ and the other regression coefficients to $\beta_\ell = -2 + 0.5\ell$ for $\ell = 1,\ldots,7$. For the random effects, we considered a sparse design with the number of random effects given by $R = N^{0.88}$ and $C = N^{0.53}$. This is a much more demanding setting than that considered in Ghosh [4] to investigate the properties of their penalized quasi-likelihood estimator. Finally, the random effects standard deviations were set to $\sigma_A = 0.8$ and $\sigma_B = 0.4$, two realistic values for applications in e-commerce. Besides penalized quasi-likelihood and all-row-column, also an *approximate oracle* method was considered, where the $\widehat{\gamma}$ coefficients estimated by a standard logistic regression were adjusted by inverting the relationship (5) using the true values of the variance components.

The analysis of the average estimation times (not reported here) confirmed that the two methods scale to $O(N)$. The statistical performances instead show some marked differences. Figure 1 displays the scatterplots of the $\log_{10}$ of the mean squared errors against $\log_{10} N$ for the three methods under comparison. The scatterplots are reported for the intercept, one of the $\beta$ coefficients and the two standard deviations $\sigma_A$ and $\sigma_B$. Results for the other six $\beta$ coefficients are similar to those of the $\beta$ coefficient show in Figure 1. Below we summarize the interesting findings we obtained from this simulation study:




(a) $\beta_0$ (b) $\beta_7$




(c) $\sigma_A$ (d) $\sigma_B$

Fig. 1: Simulation study. Scatterplots of $\log_{10}$ of the mean squared errors against $\log_{10} N$ for the penalized quasi-likelihood (PQL), all-row-column (ARC) and oracle estimators of four selected parameters.

- the all-row-column and oracle methods provide rather similar results for the $\beta$ coefficients. Indeed, all-row-column estimates well the variance components that are needed to rescale the marginal estimates and thus it can match the oracle results to a quite good extent;
- penalized quasi-likelihood gives estimates of $\beta$ coefficients with much higher mean square error than all-row-column, except for very large sample sizes, where the approximate nature of relation (5) kicks in, resulting in some asymptotic bias for the all-row-column and oracle estimators. From other plots (not shown here), it is also apparent that the all-row-column bias is however always small in size;

- penalized quasi-likelihood exhibits a very large bias and variance in the estimation of $\sigma_A$. This bias follows from the fact that with $R = N^{0.88}$, each $a_i$ is estimated using very few observations; indeed, on the average, $N/R = N^{0.12} = 3.98$ for $N = 10^5$ and $N^{0.12} = 5.25$ for $N = 10^6$. Things are better for estimation of $\sigma_B$, since each $b_j$ is estimated using larger groups.

Some further simulations were carried out for another setting with much larger variances. As expected the results were worse for all the estimators, but with a pattern quite similar to that discussed above. The mean square error of the all-row-column estimator for the $\beta$ coefficients was higher than that of the penalized quasi-likelihood for $N > 10^5$. This is not surprising since (5) is valid under *small-sigma asymptotics*. Yet all-row-column was still preferable to penalized quasi-likelihood for the estimation of variance components.

## 4 Conclusions and ongoing research

The results of this note support the use of the all-row-column estimator also for the logit case. Despite its approximate nature, all-row-column seems preferable to penalized quasi-likelihood for practical use. Ongoing work will focus on the extension of the all-row-column methodology to crossed random effects with random slopes, which may be valuable for large-scale analyses of user-generated content.

**Acknowledgements** Ruggero Bellio was supported by the research project *Latent Variable Models for Complex Data* founded by the European Union - NextGenerationEU (MUR DM funds 737/2021).

## References

1. Bellio, R., Ghosh, S., Owen, A.B., Varin, V.: Scalable estimation of probit models with crossed random effects. arXiv:2308.15681 (2023). https://arxiv.org/abs/2308.15681
2. Breslow, N.E., Clayton, D.G: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88, 9–25 (1993).
3. Schall, R.: Estimation in generalized linear models with random effects. Biometrika 78, 719—727 (1991).
4. Ghosh, S., Hastie, T., Owen, A.B.: Scalable logistic regression with crossed random effects. Electron. J. Stat. 16, 4604–4635 (2022).
5. Xu, L., Reid, N., Kong,D.: Gaussian variational approximation with composite likelihood for crossed random effect models. arXiv:2310.12485 (2023). https://arxiv.org/abs/2310.12485
6. Bates, D., Maechler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using `lme4`. J. Stat. Softw. 67, 1–48 (2015).
7. Heagerty, P.J.: Marginally specified logistic-normal models for longitudinal binary data. Biometrics 55, 688–698 (1999).
8. Wang, Z., Louis, T.A.: Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. Biometrika 90, 765—775 (2003).