Bayesian Analysis (0000) **00**, Number 0, pp. 1

Discussion of "A Novel Algorithmic Approach to Bayesian Logic Regression" by A. Hubin, G. Storvik and F. Frommlet

Malgorzata Bogdan[*,‡], Blazej Miasojedow[†] and Jonas Wallin[‡]

First of all we would like to congratulate the authors for a very interesting and important article. Logic regression model introduced in Ruczinski (2000); Ruczinski et al. (2003, 2004) is a Generalized Linear Model where individual predictors take form of logic expressions dependent on binary explanatory variables. This model arises naturally in the context of identifying epistatic effects in genetic studies. Following Bateson and Mendel (1909), biological epistasis is usually understood as a phenomenon in which "a variant or allele at one locus [...] prevents the variant or allele at another locus from manifesting its effect" (see Cordell (2002)), or more generally as a situation when the effect of one allele can only be observed when a second allele is also present. Such epistatic effects can be naturally expressed using logic expressions of the binary variables dependent on the genotypes of genetic markers. While each logic expression can be also represented in the form of the regular linear model, this usually requires many main effects and lower interaction terms. For example, a single "logic interaction" involving four variables $(x_1 \vee x_2) \wedge (x_3 \vee x_4)$ in classical representation takes the form

$$x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 - x_1x_3x_4 - x_2x_3x_4 - x_1x_2x_3 - x_1x_2x_4 + x_1x_2x_3x_4 \ , \quad (0.1)$$

and its natural interpretation is lost in the large number of classical interaction terms. Moreover, the possible causal influence of this "logic interaction" is practically impossible to recover by the regular linear model, where the regression coefficients by each component of (0.1) are estimated separately.

Logic regression seems to be particularly useful for the analysis of outbred populations (like humans), where the number of genetic variants is often much larger than in controlled populations (like e.g. domesticated animals or experimental crosses). Also, it can be applied in a much wider context, like e.g. for the natural representation of the joint influence of general qualitative variables or for the model selection for discrete multicolored graphical models, like the Potts model, in the spirit of (Miasojedow and Rejchel, 2018; Banerjee et al., 2008; Höfling and Tibshirani, 2009; Ravikumar et al., 2010). In case of multicolored graphical models logic expressions can naturally describe dependence between nodes of the graph.

Application of logic regression in real life problems requires solving complex computational and statistical issues, resulting from the large number of possible logic expression models and the possibility of writing a single logic expression in many equivalent

[*]Department of Mathematics, University of Wroclaw, Plac Grunwaldzki 2/4, 50-384 Wroclaw, Poland malgorzata.bogdan@uwr.edu.pl

[†]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland B.Miasojedow@mimuw.edu.pl

[‡]Department of Statistics, Lund University, Box 743, 220 07 Lund, Sweden jonas.wallin@stat.lu.se

 DOI: 0000

tautological forms. For example, the logicFS program of Schwender and Ickstadt (2008) uses simulated annealing (Kirkpatrick et al. (1983)) to maximize the likelihood function over all logic regression models with a given number of leaves. After selecting the "best" model, each logic expression is transformed into a disjunctive normal form (DNF) i.e., OR combination of AND combinations (i.e. prime implicants or logic interactions). The importance of individual interactions is estimated by repeating the whole procedure using many bootstrap samples from the original data and taking into account both the frequency with which a given interaction appears in bootstrap replications as well as its contribution to a total model likelihood.

The disadvantage of the importance measures proposed in Schwender and Ickstadt (2008) is that their values depend on the size of the data set and there exist no natural thresholds which would allow to separate important interactions from false predictors. However, these importance measures can be used for ranking the potential interactions. Concerning model selection strategies, Malina et al. (2014) use logicFS importance measures to build a GLM model including a moderate number of most important interactions. Then the "statistically significant" interactions are selected using the backward elimination procedure based on the multiplicity adjusted p-values. The multiplicity adjustment takes into account that the number of interactions in the space searched by logicFS increases with the interaction complexity.

In Hubin et al. (2020) the issue of identifying important logic interactions is addressed within a Bayesian framework, where the importance of a given logic expression is measured by the sum of posterior probabilites of GLM models which contain this expression as one of predictors. The algorithm in Hubin et al. (2020) calculates the posterior probability for a GLM model M by an approximation to the Bayes rule. The marginal likelihood of the data given M is calculated using the analytical formulas or Laplace approximations for Jeffrey's or robust g-priors. This allows to avoid computational burden of MCMC over the space of model parameters. Similarly as in Bogdan et al. (2004); Baierl et al. (2006), the prior for each model M depends on its complexity and is selected in such a way that the prior expected numbers of logic expressions of different lenghts are approximately the same and do not depend on the number of predictors m. Since the number of complex interactions increases with m at a higher rate than the number of simple interactions, this effectively introduces the additional penalty on the model complexity, which depends on m. The arguments presented in Bogdan et al. (2008b,a) illustrate that this penalty is related to the Bonferroni-type correction for multiplicity, similar to the multiple testing correction used in Malina et al. (2014).

To calculate the posterior probability of a model M the authors use the Bayes rule

$$P(M|Y) = \frac{P(Y|M)P(M)}{\sum_{\Omega} P(Y|M)P(M)} \ , \tag{0.2}$$

where Ω contains all possible logic regression models. Since it is not possible to visit all these models, the main computational challenge relies on designing a search algorithm which can visit most of the likely models, thus well approximating the denominator of (0.2). Similar problem appears also when fitting regular regression models and in

Frommlet et al. (2012a) it was approached by the application of the genetic algorithm supplied with the "local" research in the neighborhood of promising models. In Hubin et al. (2020) the authors propose an iterative algorithm, where in each iteration some new predictors are formed using the specifically designed crossover, mutation and reduction operators on the selected set of logic expressions and then apply the Mode Jumping MCMC of Hubin and Storvik (2018) to search the space of GLM models based on these predictors.

While we believe that the article of Hubin et al. (2020) is an interesting and important contribution to the research on the logic regression, we are rather reserved with respect to the proposed algorithm.

In Section 2.3 of Hubin et al. (2020) it is mentioned that a proper MCMC algorithm is not needed if the main purpose is to visit many highly probable models. We agree with the authors and believe that the reversibility of MJMCMC is actually not desired, since it creates unnecessary loops and increases the time of visiting many distinct models. In our opinion a better performance could be obtained by contructing an irreducible and well mixing algorithm of walking over the space of GLM models. In the recent years non-reversible MCMC algorithms received large attention (see e.g. Bouchard-Côté et al. (2018); Bierkens et al. (2019)) due to the fact that non-reversible chains are able to explore the state space much faster than the reversible algorithms. For example, let us consider a uniform distribution on the set $0, 1, \ldots, N$. In this case the standard reversible MCMC algorithm reduces to a random walk. Hence, after n steps the expected number of explored states is proportional to $\sqrt{n}$ and the number of moves to explore the whole space is proportional to N^2. Instead, we could construct a simple, non-reversible algorithm; i.e. we remember the direction of the previous move and go in the same direction until we hit 0 or N, where the direction is reversed. Then we can explore the whole space in at most $2N$ steps. In case of the problem discussed in Hubin et al. (2020), the construction of the non-reversible MCMC algorithm would be rather simple, since the convergence to the stationary measure is not needed. The only requirement is that the algorithm is irreducible and aperiodic. One solution here would be to define the global and local moves and accept the new state with probability $(\pi(y)/\pi(x))^\alpha$ with some $\alpha > 0$. The parameter α would control the permissible deviations of the posterior with respect to its maximum. Another solution could rely on storing the visited states in a priority queue, with priority proportional to the posterior probability. Then the elements from the queue could be modified by some kernel and placed back to the queue. Such an approach would allow us to explore the space starting from the more promising candidates. Also, this method could be easily parallelized without the need of post processing.

Further, we are concerned with the lack of treatment for tautologies. It seems to us that this might lead to the dilution of the posterior probability among many tautological representations of a given interaction and the loss of power of identification of this interaction. While at the final stage of the algorithm this problem can be solved by post-processing of the output, it is not clear what is the impact of this dilution on elimination of interesting interactions at the earlier stages of the algorithm. It is also important to observe that the number of tautological representations increases with the

interaction complexity. Thus, if one merges all tautologies to a single logical expression in a post-processing step, the total prior probability assigned to this unique expression effectively increases with its length and counterbalances the effect of the multiplicity correcting priors suggested by the authors.

The authors estimate the posterior probabilities of different models using (0.2). It is not clear to us why the sum in the denominator of (0.2) contains only $M_{fin} = 10000$ models based on d trees from the final stage of the algorithm. Why not use the information from the earlier stages? Further, in some of the reported simulation examples the authors use $d = 15$. Thus the final search is performed only over $d = 2^{15} = 32768$ models, which could be easily looked at without application of the MCMC algorithm.

Also, a huge random reduction of the final model space leads to substantially different results for different parallel runs of the algorithm. Therefore the authors aggregate results from different runs using a weighting scheme specified in equation (15) of their paper. In our opinion it seems more reasonable to estimate the posterior probabilities of different models simply by including all models visited in different runs in denominator of (0.2). Also, as we mentioned above, it seems to us that the priority queues would allow for some synchronization between different runs and more efficient search through the model space.

Concerning implementation issues - we observed that the denominator of (0.2) calculated by the currently implemented algorithm includes only the models accepted by MJMCMC. Taking into account that the acceptance rate is usually below 0.1, storing all the models proposed rather than only accepted would give a better estimate of the denominator of (0.2). Further, it seems to us that in the current implementation the denominator of (0.2) increases every time the model is accepted by MJMCMC, without checking if this model already appeared in the sum. However, the detailed analysis of the hidden duplication problems would require a more careful analysis of the code, which is rather difficult due to its structure.

The authors conclude that there is almost no difference between the results when the Jeffrey's or the robust g-prior is used when calculating the model marginal likelihood. However, it seems important to note that the simulations justifying this claim were performed using rather simple GLM models with independent predictors. Actually, it seems that many of the solutions proposed by the authors are specifically designed for this case. For example, consider the case when a given predictor is strongly correlated with other explanatory variables. Then the posterior probability of a "true" model including this predictor will be diluted between "neighboring" models and this predictor might easily miss the threshold for inclusion in the subsequent populations. As noted by the authors, the dilution of posterior probabilities actually occurred in the real data from the experimental recombinant inbred line, where the neighboring markers are rather strongly correlated. We simulated similar spatially correlated data and had a substantial difficulty with identifying a simple two-way logic interaction. Actually, the dilution issue seems to be even more problematic for interactions than for the main effects since the number of correlated interaction terms is substantially higher than the number of respective correlated markers.

Also, one of important features of the algorithm is the initial selection of d_1 important binary variables, which stay as the single trees in the spaces S_i in all iterations of the algorithm. The initial space S_1 is formed by including logic expressions dependent only on these predictors. Other variables can enter the search space only during the mutation, which occurs with a relatively low probability. Thus, the selection of these initial predictors effectively reduces the search space. This approach again seems to be very well suited for the situation when explanatory variables are independent but might lead to missing important predictors otherwise.

Another interesting property, worth studying, is the scaling of the algorithm with respect to the number of explanatory variables m. This number seems to hinder the speed of MJMCMC only at the first step, where d_1 important main effects are selected. However, the magnitude of m probably strongly influences the power of identifying logic interactions. Since the number of possible logic interactions increases rapidly with m, the prior probability for each of them quickly diminishes, which results in decrease of posterior probabilities.

To summarize: it appears to us that the usefulness of the proposed algorithm and the GLM logic regression model is rather restricted to the case when predictors are roughly independent and $n >> m$. This is however still of a great value in genetic studies, where the raw data are often pre-processed and only relatively few candidate genetic markers are used for building more sophisticated predictive models. If such markers are sufficiently distant, they are almost not correlated. The candidate markers are usually selected using the prior biological knowledge. Since logic interactions usually have strong main effects components, the candidate markers could also be selected using the classical Genome Wide Association Studies (see e.g. Frommlet et al. (2012b) or Brzyski et al. (2017)).

References

Baierl, A., Bogdan, M., Frommlet, F., and Futschik, A. (2006). "On locating multiple interacting quantitative trait loci in intercross designs." *Genetics*, 173: 1693–1703. 2

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data." *Journal of Machine Learning Research*, 9: 485–516. 1

Bateson, W. and Mendel, G. (1909). *Mendel's principles of heredity*. Cambridge University Press: New York, G.P. Putnam's Sons. 1

Bierkens, J., Fearnhead, P., and Roberts, G. (2019). "The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data." *Annals of Statistics*, 47: 1288–1320. 3

Bogdan, M., Frommlet, F., Biecek, P., Cheng, R., Ghosh, J., and R.W., D. (2008a). "Extending the Modified Bayesian Information Criterion (mBIC) to dense markers and multiple interval mapping." *Biometrics*, 64: 1162–1169. 2

Bogdan, M., Ghosh, J., and R.W., D. (2004). "Modifying the Schwarz Bayesian In-

formation Criterion to locate multiple interacting quantitative trait loci." *Genetics*, 167: 989–999. 2

Bogdan, M., Ghosh, J., and Żak-Szatkowska, M. (2008b). "Selecting explanatory variables with the modified version of Bayesian Information Criterion." *Quality and Reliability Engineering International*, 24: 627–641. 2

Bouchard-Côté, A., Vollmer, S. J., and Doucet, A. (2018). "The Bouncy Particle Sampler: A Nonreversible Rejection-Free Markov Chain Monte Carlo Method." *Journal of the American Statistical Association*, 113(522): 855–867.
URL https://doi.org/10.1080/01621459.2017.1294075 3

Brzyski, D., Peterson, C., Sobczyk, P., Candès, E., Bogdan, M., and Sabatti, C. (2017). "Controlling the rate of GWAS false discoveries." *Genetics*, 205: 61–75. 5

Cordell, H. (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans." *Hum. Mole. Genet.*, 11: 2463–2468. 1

Frommlet, F., Ljubic, I., Arnardottir, H., and Bogdan, M. (2012a). "QTL Mapping Using a Memetic Algorithm with Modifications of BIC as Fitness Function." *Statistical Applications in Genetics and Molecular Biology*, 11: Art.2. 3

Frommlet, F., Ruhaltinger, F., Twarog, P., and Bogdan, M. (2012b). "Modified versions of Bayesian Information Criterion for genome-wide association studies." *Computational Statistics and Data Analysis*, 56(5): 1038–1051. 5

Höfling, H. and Tibshirani, R. (2009). "Estimation of sparse binary pairwise Markov networks using pseudolikelihoods." *Journal of Machine Learning Research*, 10: 883–906. 1

Hubin, A. and Storvik, G. (2018). "Mode jumping MCMC for Bayesian variable selection in GLMM." *Computational Statistics and Data Analysis*, 127: 281–297. 3

Hubin, A., Storvik, G., and Frommlet, F. (2020). "A Novel Algorithmic Approach to Bayesian Logic Regression." *Bayesian Analysis*. 2, 3

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). "Optimization by simulated annealing." *Science*, 220: 671–680. 2

Malina, M., Ickstadt, K., Schwender, H., Posch, M., and Bogdan, M. (2014). "Detection of epistatic effects with logic regression and a classical linear regression model." *Statistical Applications in Genetics and Molecular Biology*, 13: 83–104. 2

Miasojedow, B. and Rejchel, W. (2018). "Sparse Estimation in Ising Model via Penalized Monte Carlo Methods." *Journal of Machine Learning Research*, 19(75): 1–26.
URL http://jmlr.org/papers/v19/16-554.html 1

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010). "High-dimensional Ising model selection using l_1-regularized logistic regression." *The Annals of Statistics*, 38: 1287–1319. 1

Ruczinski, I. (2000). "Logic Regression and Statistical Issues Related to the Protein

Folding Problem." Dissertation, Department of Statistics, University of Washington, Seattle, WA. 1

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003). "Logic Regression." *Journal of Computational and Graphical Statistics*, 12: 475–511. 1

— (2004). "Exploring Interactions in High-Dimensional Genomic Data: An Overview of Logic Regression, with Applications." *Journal of Multivariate Analysis*, 90: 178–195. 1

Schwender, H. and Ickstadt, K. (2008). "Identification of snp interactions using logic regression." *Biostatistics*, 9: 187–198. 2