A novel algorithmic approach to Bayesian Logic Regression

Aliaksandr Hubin, Geir Storvik & Florian Frommlet
Department of Mathematics, University of Oslo
& Department of Medical Statistics (CEMSIIS), Medical University of Vienna
aliaksah@math.uio.no, geirs@math.uio.no, florian.frommlet@meduniwien.ac.at

Introduction

- Logic regression is tool to construct predictors from Boolean combinations of binary covariates previously used for inference but never for predictions (before);
- Among the main applications were
  - Modeling epistatic effects in genetic association studies;
  - Regulatory motif finding;
  - Identifying target populations for screening or not screening;
- Has not become widely known because of
  - Combinatorial complexity;
  - Fit algorithms were not performing sufficiently well;
  - Few applications were addressed;
- Efficient fit algorithms for model probabilities in space of logic regression are required, since
  - The number of models to select from is doubly exponential in the number of input boolean variables;
  - The search space has numerous sparsely located local extrema;
- We introduce the GMJMCMC algorithm
  - It is able identify three-way and even four-way interactions with large power and low FDR;
  - We apply GMJMCMC to Recombinant Inbred Lines in *Arabidopsis thaliana* and in *Drosophila*.

Model specification

The logic regression model

$$Y_i|\mu_i \sim \mathrm{f}(y|\mu_i), i \in \{1,...,n\} \quad (1)$$
$$\mu_i = g^{-1}(\eta_i) \quad (2)$$
$$\eta_i = \gamma_0\beta_0 + \sum_{j=1}^{k}\gamma_j\beta_j L_{ij} \quad (3)$$

- $L_{ij} \in \{0,1\}, j \in \{1,...,k\}$ are all feasible logical expressions (trees). E.g. $L_{i1} = (X_{i1} \wedge X_{i2}) \vee X_{i3}^c$, where
  1. $\wedge$ is logical *and*;
  2. $\vee$ is logical *or*;
  3. ^{c} is logical *not*;
- k is the total number of all possible trees of size up to K based on p input leaves;
- Q is the maximal allowed number of trees per model;
- $\beta_j \in \mathbb{R}, j \in \{0,...,k\}$ are regression coefficients of these trees;
- $g(\cdot)$ is a proper link function;
- $\gamma_j \in \{0,1\}, j \in \{0,...,k\}$ are indicators defining if a tree L_{ij} is included into the model;

Model priors

Latent coefficients:

$$p(\boldsymbol{\gamma}) \propto \mathbb{I}\left\{\sum_{j=1}^{k}\gamma_j \le Q\right\}\prod_{j=1}^{k}\frac{s_j!}{p^{s_j}2^{2s_j-2}}\mathbb{I}\left\{s_j \le K\right\}, \quad (4)$$

Slope coefficients:

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \quad (5)$$

Other parameters (if present)

$$\boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}) \quad (6)$$

Posterior evaluation

Then Laplace approximations of the marginal likelihood can be obtained in the GLM context:

$$p(\mathbb{D}|\boldsymbol{\gamma}) \approx e^{\log p(\mathbb{D}|\boldsymbol{\gamma},\hat{\theta}_\gamma) - 0.5\times|\theta_\gamma|\log n}, \quad (7)$$

where $p(\mathbb{D}|\boldsymbol{\gamma},\hat{\theta}_\gamma)$ is the likelihood evaluated at the maximum likelihood estimate $\hat{\theta}_\gamma$ of the parameters for model $\boldsymbol{\gamma}$ (the corresponding regression coefficients and possibly a variance parameter) while n is the number of observations.

- **Notice that** $p(\boldsymbol{\gamma},\theta_\gamma|\mathbb{D}) = p(\theta_\gamma|\boldsymbol{\gamma},\mathbb{D})p(\boldsymbol{\gamma}|\mathbb{D})$;
- **Notice that** $p(\boldsymbol{\gamma}|\mathbb{D}) = \frac{p(\mathbb{D}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'\in\Omega_\gamma}p(\mathbb{D}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')} \approx \widehat{p}(\boldsymbol{\gamma}|\mathbb{D}) = \frac{p(\mathbb{D}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'\in\mathbb{V}}p(\mathbb{D}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}$;
- $\mathbb{V}$ is the **subspace** of Ω_γ to be **efficiently explored**;

In **Hubin and Storvik** [6] **we suggested efficient mode jumping proposals** in the **discrete parameter** spaces. But Ω_γ and k must be clearly specified for MJMCMC. The later is **not feasible** in **logic regression**. To solve this problem we present the Genetically Modified MJMCMC (GMJMCMC) algorithm [4], where MJMCMC is embedded in the iterative setting of a genetic algorithm. In each iteration only a given set $\mathcal{S}$ of trees (of fixed size d) is considered. Each $\mathcal{S}$ then becomes a separate *search space* for MJMCMC.

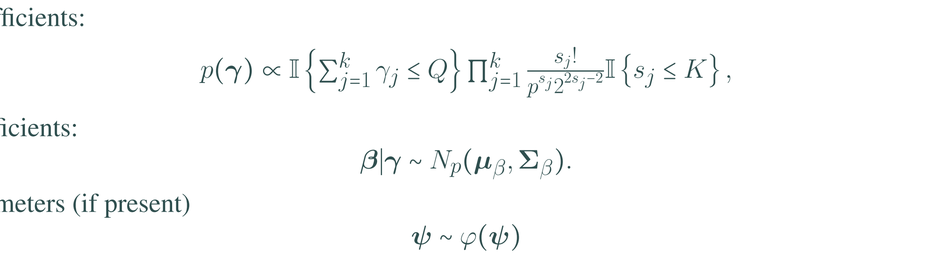

Figure 1: Illustration of locally optimized proposals (left) and MDS (multidimensional scaling) plot of the best 1024 models in terms of PMP in the space of models for some epigenetic data (right).

The GMJMCMC algorithm

- $\mathcal{S}_0$ is the set of p input binary leaves;
- $\mathcal{S}_1$ is constructed by:
  1. Running MJMCMC for a given number of iterations N_{init} on $\mathcal{S}_0$;
  2. The first $d_1 < d$ members of population $\mathcal{S}_1$ are then defined by **filtration** operation, whilst $p - d_1$ filtered leaves from $\mathcal{S}_0$ are kept in $\mathcal{F}$;
  3. The remaining $d - d_1$ members of $\mathcal{S}_1$ are obtained by means of the **crossover** operation applied to $\mathcal{S}_0$;
- All other $\mathcal{S}_t, t \in \{2,...,t_{max}\}$ are constructed by:
  1. Running MJMCMC for a given number of iterations N_{expl} on $\mathcal{S}_{t-1}$;
  2. The first $d_t \le d$ members of population $\mathcal{S}_t$ are then defined by **filtration** operation;
  3. The remaining $d - d_t$ members of $\mathcal{S}_t$ are obtained by means of the **crossover**, **mutation** and **reduction** operations applied to $\mathcal{S}_{t-1}$ and $\mathcal{F}$;

Parallelization

1. Run B GMJMCMC chains in parallel with different seeds on separate CPUs or clusters;
2. Combine all unique models visited by all B chains into $\mathbb{V}$;
3. Compute posteriors of other parameters of interest as $\widehat{p}(\Delta|\mathbb{D}) = \sum_{\boldsymbol{\gamma}\in\mathbb{V}}p(\Delta|\boldsymbol{\gamma},\mathbb{D})\widehat{p}(\boldsymbol{\gamma}|\mathbb{D})$.

Results

Simulation study. Binary responses

Scenario 1: $\mathrm{logit}(\pi) = -0.7 + X_1^c \wedge X_4 + X_8 \wedge X_{11} + X_5 \wedge X_9$
Scenario 2: $\mathrm{logit}(\pi) = -0.45 + 0.6\cdot X_1^c \wedge X_4 + 0.6\cdot X_8 \wedge X_{11} + 0.6\cdot X_5 \wedge X_9$
Scenario 3: $\mathrm{logit}(\pi) = 0.4 - 5\cdot X_2 \wedge X_9 + 9\cdot X_7 \wedge X_{12} \wedge X_{20} - 9\cdot X_4 \wedge X_{10} \wedge X_{17} \wedge X_{30}$

Table 1: *Results for the three simulation scenarios for binary responses. Power for individual trees, overall power, expected number of false positives (FP) and FDR are compared between Monte Carlo logic regression (MCLR) [5] and the full Bayesian version of logic regression (FBLR) by [1] and parallel GMJMCMC [4].*

S. 1	FBLR	MCLR	GMJMCMC
$X_1^c \wedge X_4$	0.30	≤ 0.67	0.97
$X_5 \wedge X_9$	0.42	≤ 0.61	1.00
$X_{11} \wedge X_8$	0.33	≤ 0.59	0.91
Overall Power	0.35	≤ 0.62	0.96
FP	3.88	≥ 2.70	0.25
FDR	0.77	≥ 0.06	0.06
WL	0	0	0

S. 2	FBLR	MCLR	GMJMCMC
$X_1^c \wedge X_4$	0.32	≤ 0.66	0.97
$X_5 \wedge X_9$	0.40	≤ 0.67	0.99
$X_{11} \wedge X_8$	0.37	≤ 0.60	0.86
Overall Power	0.36	≤ 0.64	0.94
FP	3.83	≥ 2.58	0.38
FDR	0.75	≥ 0.06	0.09
WL	1	1	0

S. 3	FBLR	MCLR	GMJMCMC
$X_2 \wedge X_9$	0.93	≤ 0.93	1.00
$X_7 \wedge X_{12} \wedge X_{20}$	0.04	≤ 0.67	0.91
$X_4 \wedge X_{10} \wedge X_{17} \wedge X_{30}$	0.00	≤ 0.19	1.00
Overall Power	0.32	≤ 0.60	0.97
FP	6.40	≥ 2.98	0.15
FDR	0.54	≥ 0.06	0.04
WL	90	72	1

Real data analysis. Drosophila Simulans

Genotype data from 45 markers is available for 471 samples from Drosophila Simulans. Six markers are located on chromosome X, 16 markers on chromosome 2 and 23 markers on chromosome 3. Imputation of the few missing genotypes was performed by a simple maximum likelihood approach based on flanking markers.

Table 2: *Posterior probabilities for additive and epistatic effects detected with GMJMCMC (column posterior) for Drosophila Simulans are presented for the trait pc1 (first principal component of the size and shape of male genital arc).*

marker	chromosome	marker name	posterior
m2	X	w	1.000
m4	X	v	1.000
m7	2	gl	0.960
m9	2	cg	1.000
m14	2	mhc	1.000
m18	2	sli	0.414
m22	2	zip	0.838
m23	2	lsp	0.999
m26	3	dbi	1.000
m29	3	fz	1.000
m33	3	ht	1.000
m37	3	mst	1.000
m40	3	hb	0.942
m44	3	jan	1.000
m11, m35	2, 3	ninaE ∧ ninaC	0.998

Conclusions

- We introduced the GMJMCMC algorithm for Bayesian logic regression models capable of
  - estimating posterior model probabilities;
  - Bayesian model averaging and selection;
- The *EMJMCMC* R-package is available
  - http://aliaksah.github.io/EMJMCMC2016/;
  - flexibility in the choice of methods
    - marginal likelihoods;
    - model selection criteria;
  - extensive parallel computing is available;
  - vectorized predictions with NA handling is incorporated;
- Results showed that GMJMCMC
  - performs well in terms of the search speed and quality;
  - addresses a more general class of models than competitors;
  - provides nice predictive and inferential performance in the applications.

Forthcoming Research

In future we are going to extend GMJMCMC to the settings of a more general than Logic regression non-linear regression settings to be able to address the cases with continuous input covariates. Another direction of the further research is the notion of true and false positives in the context of Logic regression.

References

[1] A. Fritsch. *A Full Bayesian Version of Logic regression for SNP Data*. PhD thesis, Diploma Thesis, 2006.
[2] A. Fritsch and K. Ickstadt. Comparing Logic Regression Based Methods for Identifying SNP Interactions. *Springer Berlin / Heidelberg, Lecture Notes in Computer Science*, 4414:90–103, 2007.
[3] A. Hubin and G. Storvik. Efficient mode jumping MCMC for Bayesian variable selection in GLMM, 2016. arXiv:1604.06398v3.
[4] A. Hubin, G. Storvik, and F. Frommlet. A novel algorithmic approach to Bayesian Logic Regression, 2017. arXiv:1705.07616v1.
[5] C. Kooperberg and I. Ruczinski. Identifying Interacting SNPs Using Monte Carlo Logic Regression. *Genetic Epidemiology*, 28:157–170, 2005.
[6] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *J. Comput Graphical Statist.*, 12(3):474–511, 2003.

Acknowledgments

*The authors gratefully acknowledge the **CELS** project at the University of Oslo, http://www.mn.uio.no/math/english/research/groups/cels/index.html, for giving us the opportunity, inspiration and motivation to perform our research.*