A novel algorithmic approach to Bayesian Logic Regression

Hubin A.A.,Storvik G.O., Frommlet F.

Department of Mathematics, University of Oslo

aliaksah@math.uio.no, *geirs@math.uio.no*
and
Department of Medical Statistics (CEMSIIS)

florian.frommlet@meduniwien.ac.at

UiO : Universitetet i Oslo

ML@UiO,
Oslo 2017

01.06.2017

Introduction. Issues

- Logic regression was developed as a tool to construct predictors from Boolean combinations of binary covariates previously used for
  - Inference
  - Not 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

Introduction. Issues

- 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 (leaves)
  - The search space has numerous sparsely located local extrema
  - Time and computing resources are limited

Bayesian Logic Regression (GLM context)

$$Y_i|\mu_i \sim \mathrm{f}(y|\mu_i), i \in \{1, ..., n\} \quad (1)$$

$$\mu_i = g^{-1}(\eta_i) \quad (2)$$

$$\eta_i = \gamma_0\beta_0 + \sum_{j=1}^{k} \gamma_j\beta_j L_{ij} \quad (3)$$

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

Model Priors

$$p(\gamma) \propto \mathbb{I}\left\{\sum_{j=1}^{k} \gamma_j \leq Q\right\} \prod_{j=1}^{k} v^{\gamma_j c(L_j)}, \tag{4}$$

$c(L_j)$ is a measure for the complexity of term L_j, Q is the maximal allowed number of trees per model, and $0 < v < 1$.

Model Priors

$$p(\gamma) \propto \mathbb{I}\left\{\sum_{j=1}^{k} \gamma_j \leq Q\right\} \prod_{j=1}^{k} v^{\gamma_j c(L_j)}, \tag{4}$$

$c(L_j)$ is a measure for the complexity of term L_j, Q is the maximal allowed number of trees per model, and $0 < v < 1$.

Inference driven choice capable of controlling FDR is:

$$p(\gamma) \propto \mathbb{I}\left\{\sum_{j=1}^{k} \gamma_j \leq Q\right\} \prod_{j=1}^{k} \frac{s_j!}{p^{s_j} 2^{2s_j-2}} \mathbb{I}\left\{s_j \leq C\right\}, \tag{5}$$

s_j is a number of leaves in tree L_j, p is the number of input leaves, and C is the maximal allowed number of leaves per tree.

Model Priors

$$p(\gamma) \propto \mathbb{I}\left\{\sum_{j=1}^{k} \gamma_j \leq Q\right\} \prod_{j=1}^{k} v^{\gamma_j c(L_j)}, \tag{4}$$

$c(L_j)$ is a measure for the complexity of term L_j, Q is the maximal allowed number of trees per model, and $0 < v < 1$.

Inference driven choice capable of controlling FDR is:

$$p(\gamma) \propto \mathbb{I}\left\{\sum_{j=1}^{k} \gamma_j \leq Q\right\} \prod_{j=1}^{k} \frac{s_j!}{p^{s_j} 2^{2s_j-2}} \mathbb{I}\left\{s_j \leq C\right\}, \tag{5}$$

s_j is a number of leaves in tree L_j, p is the number of input leaves, and C is the maximal allowed number of leaves per tree.

Prediction or inference driven prior is:

$$p(\gamma) \propto \mathbb{I}\left\{r \leq Q\right\} \frac{(1-1/r)(1/r)^a}{(1-1/s)(1/s)^b} \prod_{j=1}^{k} \mathbb{I}\left\{s_j \leq C\right\}, a, b > 0, \tag{6}$$

$r = \sum_{j=1}^{k} \gamma_j$ is the model size and s in r plus the total number of logical operators present in the model.

Marginal likelihood approximation

Assume the following prior on regression coefficients:

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \tag{7}$$

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

$$p(\mathbb{D}|\boldsymbol{\gamma}) \approx e^{\log p(\mathbb{D}|\boldsymbol{\gamma}, \hat{\theta}_{\boldsymbol{\gamma}}) - 0.5 \times |\theta_{\boldsymbol{\gamma}}| \log n}, \tag{8}$$

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

Inference on the model

Let:

- $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_k\}$ define a model itself, i.e. which covariates are addressed;
- $\theta_{\boldsymbol{\gamma}}$ define parameters of the model.

Goals:

- $p(\boldsymbol{\gamma}, \theta_{\boldsymbol{\gamma}}|\mathbb{D})$ posterior distribution of parameters and models;
- $p(\boldsymbol{\gamma}|\mathbb{D})$ marginal posterior probabilities of the models;
- $p(\Delta|\mathbb{D})$ marginal posterior probabilities of the quantiles of interest Δ.

But:

- $\exists 2^k$ different models in Ω_γ;
- k is huge;
- Both k and Ω_γ are extremely difficult to specify.

Possible pipeline

- **Notice that** $p(\gamma, \theta_\gamma|\mathbb{D}) = p(\theta_\gamma|\gamma, \mathbb{D})p(\gamma|\mathbb{D})$;
- Here $p(\mathbb{D}|\gamma)$ can be obtained by LA or similarly;
- **Notice that** $p(\gamma|\mathbb{D}) = \frac{p(\mathbb{D}|\gamma)p(\gamma)}{\sum_{\gamma' \in \Omega_\gamma} p(\mathbb{D}|\gamma')p(\gamma')}$;
- **Approximate with**

$$\widehat{p}(\gamma|\mathbb{D}) = \frac{p(\mathbb{D}|\gamma)p(\gamma)}{\sum_{\gamma' \in \mathbb{V}} p(\mathbb{D}|\gamma')p(\gamma')} \tag{9}$$

- $\mathbb{V}$ is the **subspace** of Ω_γ to be **efficiently explored**;
- **Near modal values in terms of "MLIK** $\times$ **prior" are particularly important** for construction of reasonable $\mathbb{V} \subset \Omega_\gamma$, **missing** them **can dramatically influence** posterior in the original space Ω_γ.

MJMCMC is efficient, but...

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**.

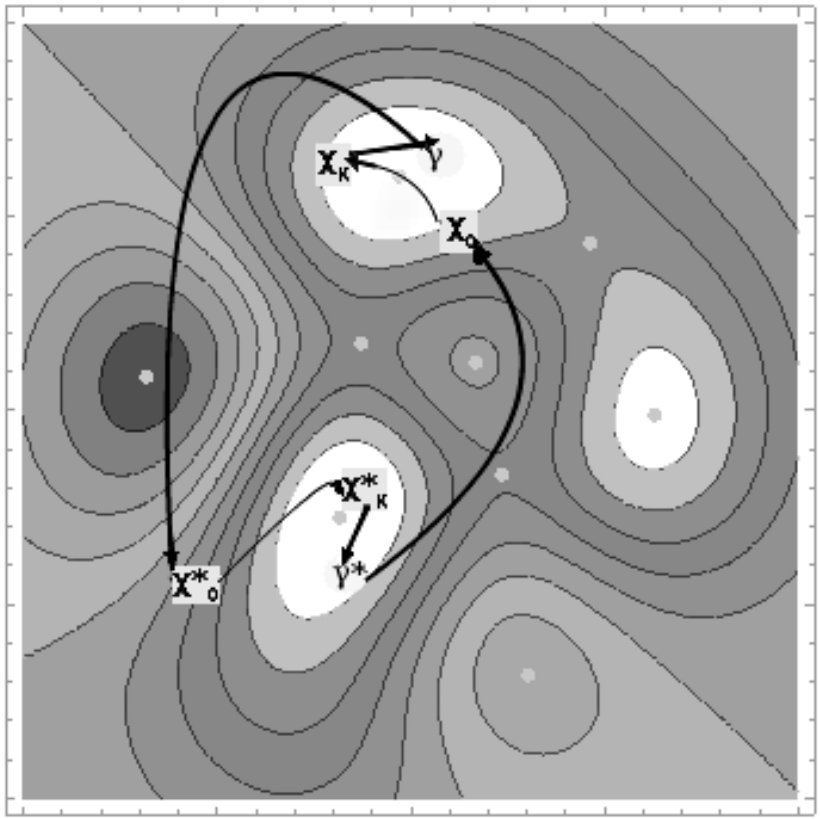

Figure: Locally optimized with randomization proposals

Genetically modified MJMCMC. Idea

- 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 induces a separate **search space** for MJMCMC or in the language of genetic algorithms $\mathcal{S}$ is the **population**;
- $\mathcal{S}$ dynamically evolves as a Markov chain of populations through $\{\mathcal{S}_0, ..., \mathcal{S}_{t_{max}}\}$ to allow MJMCMC explore different reasonable parts of the in-feasibly large total search space;
- Each $\mathcal{S}_t, t \in \{1, ..., t_{max}\}$ is selected from the neighborhood $\mathcal{N}_{t-1}$ of $\mathcal{S}_{t-1}$. $\mathcal{N}_{t-1}$ includes all populations feasible by performing **mutation**, **crossover**, **reduction** and **filtration** operations to the current $\mathcal{S}_{t-1}$;
- Utilization of the approximation (9) allows us to compute marginal inclusion probabilities

$$\widehat{p}(L|\mathbb{D}) = \sum_{\gamma \in \mathbb{V}: L \in T(\gamma)} \widehat{p}(\gamma|\mathbb{D}) \tag{10}$$

;

Genetically modified MJMCMC. Pipeline

- $\mathcal{S}_0$ 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 $d_t \leq d$ 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 $\mathcal{F}$;

Filtration operation. $\mathcal{S}_0$ case

$$\widehat{p}(F_1|\mathbb{D}) \leq \ldots \leq \widehat{p}(F_{p-d_1}|\mathbb{D}) \leq \widehat{p}(L_1^1|\mathbb{D}) \leq \widehat{p}(L_2^1|\mathbb{D}) \leq \ldots \leq \widehat{p}(L_{d_1}^1|\mathbb{D}) \quad (11)$$

$$\widehat{p}(L_1^1|\mathbb{D}) \geq p_s^o \quad (12)$$

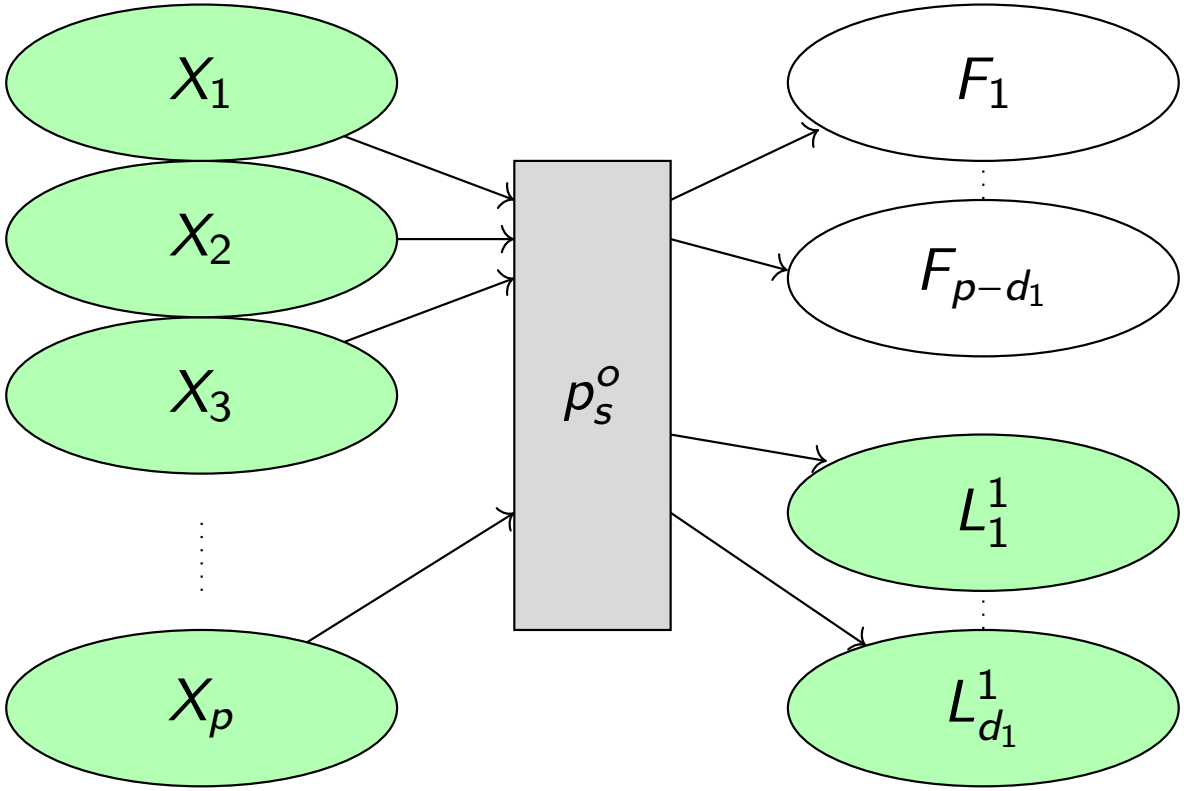

Figure: Feature filtering.

Filtration operation. $\mathcal{S}_t, t \in \{1, ..., t_{max}\}$ case

$$\widehat{p}(D_1^{t+1}|\mathbb{D}) \leq ... \leq \widehat{p}(D_{p-d_{t+1}}^{t+1}|\mathbb{D}) \leq \widehat{p}(L_{d_1+1}^{t+1}|\mathbb{D}) \leq ... \leq \widehat{p}(L_{d_{t+1}}^{t+1}|\mathbb{D}) \quad (13)$$

$$\widehat{p}(L_{d_1+1}^{t+1}|\mathbb{D}) \geq p_s^t \quad (14)$$

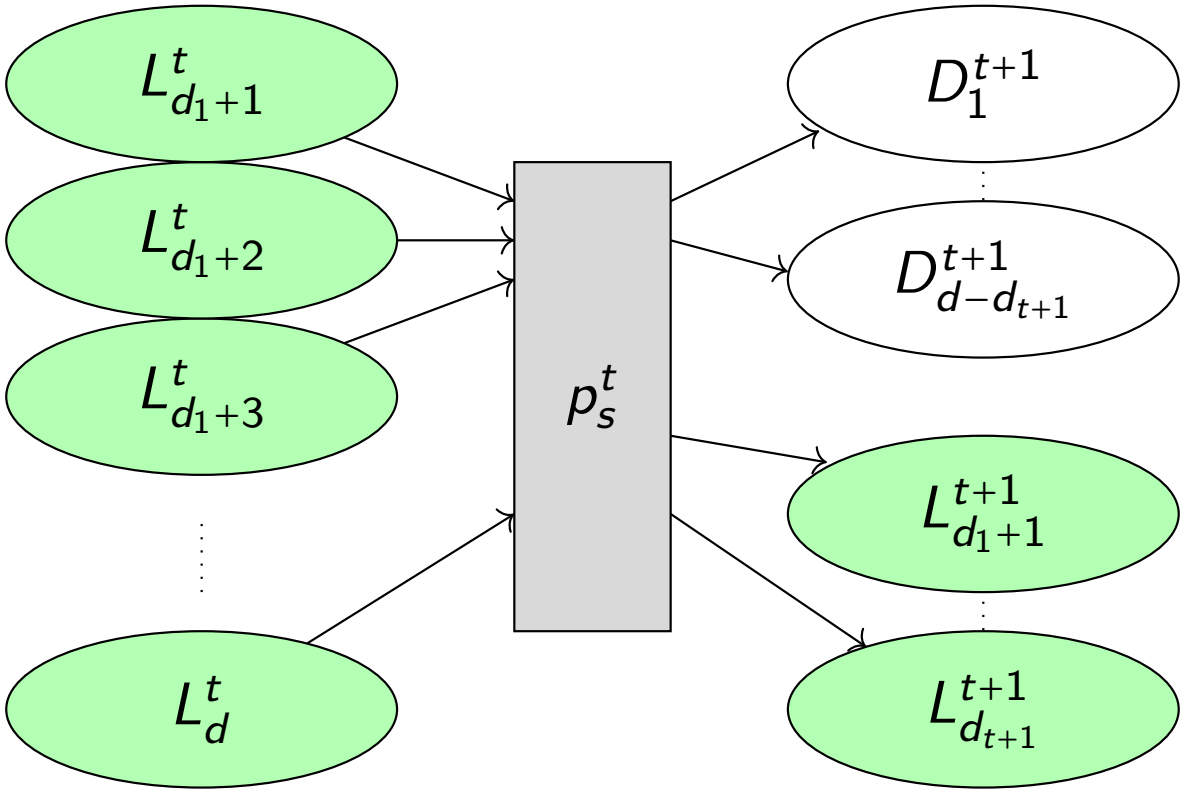

Figure: Feature filtering.

Crossover and mutation operations. Parents selection

Other $d - d_{t+1}$ members are filled with either **crossovers** with probability p_c or **mutations** with probability $1 - p_c$. **Crossovers** inbreed parents from population $\mathcal{S}_t$ only, **mutations** allow parents from $\mathcal{F}$.

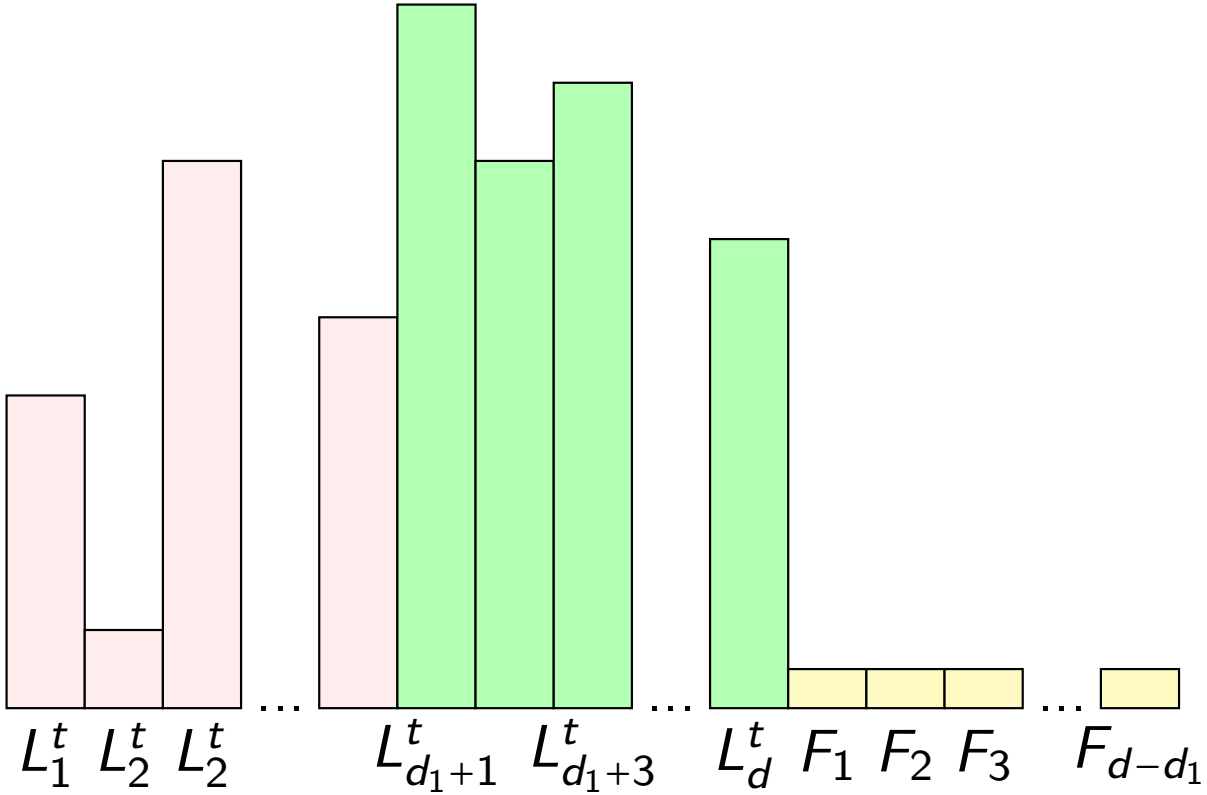

Figure: Class probabilities for selection of parents proportional to current marginal inclusion probabilities

Crossover and mutation operations. Inbreeding of parents

Within each **mutation** or **crossover** $\wedge$ is used for inbreeding with probability p_{and}, and $\vee$ - otherwise. The not ^{c} operator is applied to the parents with probability p_{not}.

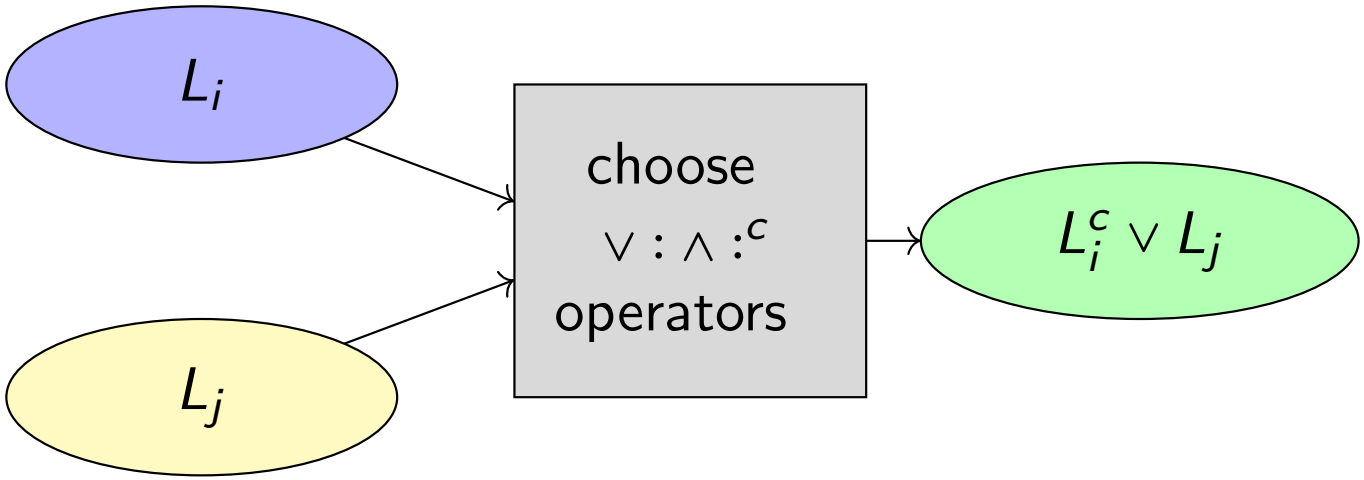

Figure: Tree engineering step illustration.

Reduction operator

- Reductions are applied for the trees greater than C;
- Each leave is independently deleted with Bernoulli probability p_d;
- The survived leaves are sticked together with $\wedge$ with probability p_{and} and $\vee$ - otherwise.

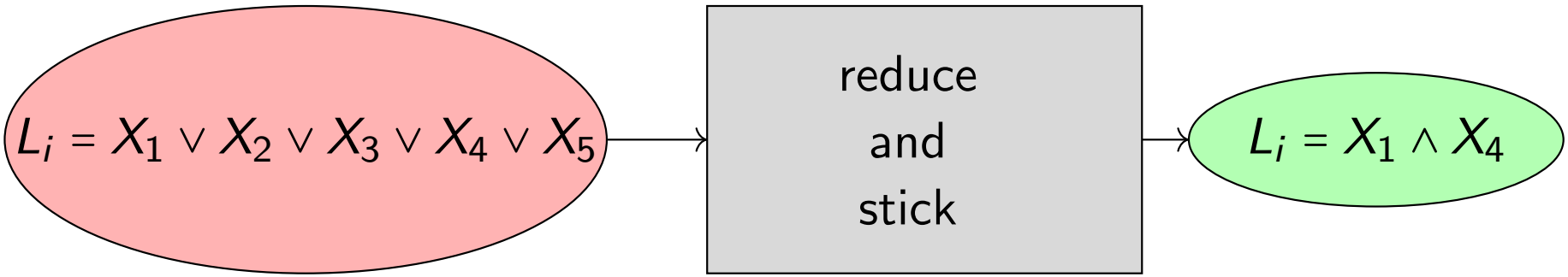

Figure: Tree pruning step illustration.

Genetically modified MJMCMC. Embarrassingly parallelized

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 model posteriors as (9)
4. Compute marginal inclusion probabilities as (10)
5. Compute posteriors of other parameters of interest as

$$\widehat{p}(\Delta|\mathbb{D}) = \sum_{\gamma \in \mathbb{V}} p(\Delta|\gamma, \mathbb{D}) \widehat{p}(\gamma|\mathbb{D}) \tag{15}$$

Simulation scenarios. Binary responses

For this scenario we generated $N = 100$ datasets with $n = 1000$ observations and $p = 50$ binary covariates. The covariates were assumed to be independent and were simulated as $X_j \sim \text{Bernoulli}(0.3)$ for $j \in \{1, \ldots, 50\}$.

Binary responses

Bernoulli observations with individual success probability π:

Scenario 1: $\text{logit}(\pi) = -0.7 + X_1^c \wedge X_4 + X_8 \wedge X_{11} + X_5 \wedge X_9$

Scenario 2: $\text{logit}(\pi) = -0.45 + 0.6 \cdot X_1^c \wedge X_4 + 0.6 \cdot X_8 \wedge X_{11} + 0.6 \cdot X_5 \wedge X_9$

Scenario 3: $\text{logit}(\pi) = 0.4 - 5 \cdot X_2 \wedge X_9 + 9 \cdot X_7 \wedge X_{12} \wedge X_{20} - 9 \cdot X_4 \wedge X_{10} \wedge X_{17} \wedge X_{30}$

Binary responses. Results over 100 datasets

	FBLR	MCLR	GMJMCMC
Scenario 1			
$X_1^c \wedge X_4$	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	0.00	0.00	0.00
Scenario 2			
$X_1^c \wedge X_4$	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	0.01	0.01	0.00
Scenario 3			
$X_2 \wedge X_9$	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	0.90	0.72	0.01

Simulation scenarios. Continuous responses

For this scenario we generated $N = 100$ datasets with $n = 1000$ observations and $p = 50$ binary covariates. The covariates were assumed to be independent and were simulated as $X_j \sim \text{Bernoulli}(0.5)$ for $j \in \{1, \ldots, 50\}$.

Continuous responses

Gaussian observations with error variance $\sigma^2 = 1$ and individual expectations specified as follows for the different scenarios:

Scenario 4: $E(Y) = 1 + 1.43\ X_5 \wedge X_9 + 0.89\ X_8 \wedge X_{11} + 0.7\ X_1 \wedge X_4$

Scenario 5: $E(Y) = 1 + 1.5\ X_{37} + 3.5\ X_2 \wedge X_9 + 9\ X_7 \wedge X_{12} \wedge X_{20} + 7 \times X_4 \wedge X_{10} \wedge X_{17} \wedge X_{30}$

Scenario 6: $E(Y) = 1 + 1.5\ X_7 + 1.5\ X_8 + 6.6\ X_{18} \wedge X_{21} + 3.5\ X_2 \wedge X_9 + 9\ X_{12} \wedge X_{20} \wedge X_{37} + 7\ X_1 \wedge X_3 \wedge X_{27} + 7\ X_4 \wedge X_{10} \wedge X_{17} \wedge X_{30} + 7\ X_{11} \wedge X_{13} \vee X_{19} \wedge X_{50}$

Continuous responses. Results over 100 datasets

Scenario 4	
$X_5 \wedge X_9$	1.00
$X_8 \wedge X_{11}$	0.99
$X_1 \wedge X_4$	0.97
Overall Power	0.99
FP	0.01
FDR	0.005
WL	0.00

Scenario 5	
X_{37}	1.00
$X_2 \wedge X_9$	0.99
$X_4 \wedge X_{10} \wedge X_{17} \wedge X_{30}$	0.89
$X_7 \wedge X_{12} \wedge X_{20}$	0.96
Overall Power	0.96
FP	0.37
FDR	0.06
WL	0.00

Scenario 6	
X_7	0.95
X_8	0.98
$X_2 \wedge X_9$	0.99
$X_{18} \wedge X_{21}$	0.96
$X_1 \wedge X_3 \wedge X_{27}$	1.00
$X_{12} \wedge X_{20} \wedge X_{37}$	0.95
$X_4 \wedge X_{10} \wedge X_{17} \wedge X_{30}$	0.32
$X_{11} \wedge X_{13} \vee X_{19} \wedge X_{50}$	0.21 (0.93)
Overall Power	0.79 (0.88)
FP	4.28 (2.05)
FDR	0.38 (0.19)
WL	0.03

Real data analysis. Arabadopsis thailana QTL study

Phenotype: hypocytl length under different light conditions

Population	Phenotype	Chr	Marker expression	$\widehat{p}(L\|\mathbb{D}) > 0.05$
EstC	Blue Light	4	X44606688	0.7669669331
EstC	Blue Light	5	X44607250	0.3348765393
EstC	Blue Light	2	X21607656	0.3090553596
EstC	Blue Light	4∧2	X44606688∧X44606810	0.2028843229
EstC	Red Light	2	MSAT2.36	0.4410634126
EstC	Red Light	2	PHYB	0.3526598888
EstC	Red Light	2∧1	(1-PHYB)∧X44606541	0.1120190937
EstC	Red Light	2	X21607013	0.0920895303
EstC	Far Red Light	4	MSAT4.37	0.3022248127
EstC	Far Red Light	4	NGA1107	0.3022203524
EstC	White Light	5	X44606159	0.6316546291
EstC	White Light	1	X21607165	0.4271134131

Real data analysis. Drosophila Simulans study

Phenotype: pc1 of the size and shape of the posterior lobe of the male genital arch

marker	chromosome	marker name	posterior	mBIC
m2	X	w	0.9999999347	x
m4	X	v	0.9999928537	x
m7	2	gl	0.9604459511	x
m9	2	cg	0.9999374951	
m10	2	gpdh		x
m14	2	mhc	1	x
m18	2	sli	0.41435673	x
m22	2	zip	0.8376115966	x
m23	2	lsp	0.9975705529	x
m26	3	dbi	1	x
m29	3	fz	1	x
m32	3	rdg		x
m33	3	ht	0.9999616425	
m35	3	ninaE		x
m37	3	mst	0.999989013	x
m40	3	hb	0.9422920378	
m41	3	rox		x
m44	3	jan	1	x
m12, m34	2, 3	glt∗ant		x
m11, m35	2, 3	ninaE $\wedge$ ninaC	0.998469988	

Real data analysis. Drosophila Mauritana study

Phenotype: pc1 of the size and shape of the posterior lobe of the male genital arch

marker	chromosome	marker name	posterior	mBIC
m1	X	ewg		x
m4	X	v	0.9936767813	x
m9	2	cg	0.9999999728	x
m11	2	ninaC	0.3821414014	x
m15	2	ddc	0.9999999565	x
m18	2	sli	0.5232755379	x
m22	2	zip	0.9997288015	x
m24	3	ve	0.9660071153	
m25	3	acr		x
m26	3	dbi	0.9946590631	
m28	3	cyc	0.3975980086	x
m29	3	fz	0.8338732206	
m34	3	ant	1	x
m37	3	mst		x
m39	3	tub	0.9992952335	
m40	3	hb		x
m41	3	rox	0.4200867879	
m44	3	jan	0.9999998898	x
m1, m2	X, X	w∨ewg	0.8552678026	
m2, m36	X, 3	w.fas		x
m29, m40	3, 3	fz.hb		x

Real data analysis. Simulans VS Mauritana prediction

- Short runs of GMJMCMC with 1 thread are addressed;
- Less conservative prior is used, namely $p(\gamma) \propto \frac{(1-1/r)(1/r)^{\log(500)}}{(1-1/s)(1/s)^2}$;
- Prediction is based on marginalized over all models' probabilities, namely $\hat{Y} = \mathbb{I}\{\widehat{p}(Y|\mathbb{D}) \geq 0.5\}, \widehat{p}(Y|\mathbb{D}) = \sum_{\gamma \in \mathbb{V}} \widehat{p}(Y|\gamma, \mathbb{D})\widehat{p}(\gamma|\mathbb{D})$;

Algorithm	min.p	mean.p	max.p	min.fn	mean.fn	max.fn	min.fp	mean.fp	max.fp
RIDGE	0.6341	0.6607	0.6861	0.1973	0.2510	0.2756	0.2284	0.2541	0.2891
GMJMCMC	0.6362	0.6580	0.6965	0.1781	0.2500	0.2817	0.2203	0.2571	0.3006
NAIVEBAYESS	0.6195	0.6572	0.6923	0.1084	0.1522	0.1864	0.2875	0.3380	0.3698
IXGBOOST	0.6195	0.6561	0.6798	0.2079	0.2554	0.2857	0.2310	0.2549	0.2870
MJMCMC	0.6299	0.6549	0.6861	0.1892	0.2535	0.2897	0.2257	0.2573	0.2965
LASSO	0.6279	0.6526	0.6757	0.2079	0.2561	0.2837	0.2363	0.2581	0.2891
LR	0.6029	0.6266	0.6570	0.0476	0.1266	0.1973	0.3153	0.3775	0.4267
DEEPNETS	0.5447	0.6185	0.6590	0.1336	0.1930	0.2857	0.2568	0.3418	0.4303
tXGBOOST	0.5800	0.6152	0.6486	0.2208	0.2724	0.3108	0.2539	0.2812	0.3105
RFOREST	0.5759	0.6131	0.6549	0.1902	0.2484	0.3033	0.2710	0.3062	0.3554
KMEANS	0.3638	0.5279	0.6466	0.2405	0.3095	0.4014	0.2806	0.3225	0.3773

Table: Comparison of performance (Precision, FDR, FNR) of different algorithms for Drosophila classification

Real data analysis. Simulans VS Mauritana study

But no nonlinearities were found either in the inferential study

Table: Drosophila data: whether simulans or mauritiana.

Population	Phenotype	Chr	Marker expression	$\widehat{p}(L\|\mathbb{D}) > 0.5$
Both	S or M	X	run	0.9999456265
Both	S or M	3	ninaE	0.9993791047
Both	S or M	2	ddc	0.9104732061

Real data analysis. Simulated response prediction

Simulate responses as:

$$Y = \mathbb{I}\left\{\text{logit}^{-1}(0.4 - 9v * eve * eip * gpdh + 9gl * egfr * glt - 5cg * w) \geq 0.5\right\}$$

Now perform predictions:

Algorithm	min.p	mean.p	max.p	min.fn	mean.fn	max.fn	min.fp	mean.fp	max.fp
tXGBOOST	0.9792	0.9902	1.0000	0.0000	0.0056	0.0150	0.0000	0.0259	0.0625
GMJMCMC	0.9439	0.9609	0.9813	0.0103	0.0239	0.0415	0.0412	0.0889	0.1826
MJMCMC	0.9335	0.9574	0.9792	0.0026	0.0253	0.0484	0.0538	0.0995	0.1565
IXGBOOST	0.9356	0.9565	0.9688	0.0076	0.0158	0.0391	0.1020	0.1366	0.1826
DEEPNETS	0.9189	0.9514	0.9667	0.0077	0.0354	0.0796	0.0538	0.0883	0.1226
RFOREST	0.9314	0.9484	0.9688	0.0078	0.0250	0.0598	0.0222	0.1337	0.2143
LASSO	0.9314	0.9437	0.9626	0.0224	0.0339	0.0576	0.0928	0.1263	0.1754
RIDGE	0.9064	0.9366	0.9584	0.0102	0.0232	0.0665	0.1287	0.1862	0.2231
NAIVEBAYESS	0.8919	0.9343	0.9626	0.0000	0.0312	0.0903	0.1287	0.1679	0.2540
LR	0.8524	0.8767	0.8960	0.0331	0.0721	0.1129	0.0435	0.2195	0.3662
KMEANS	0.4782	0.4956	0.5343	0.3052	0.3342	0.3548	0.2683	0.3346	0.3952

Table: Comparison of performance (Precision, FDR, FNR) of different algorithms for Drosophila classification

Further (partly current) research

Generalization. Allow general feature engineering instead of only logical trees:

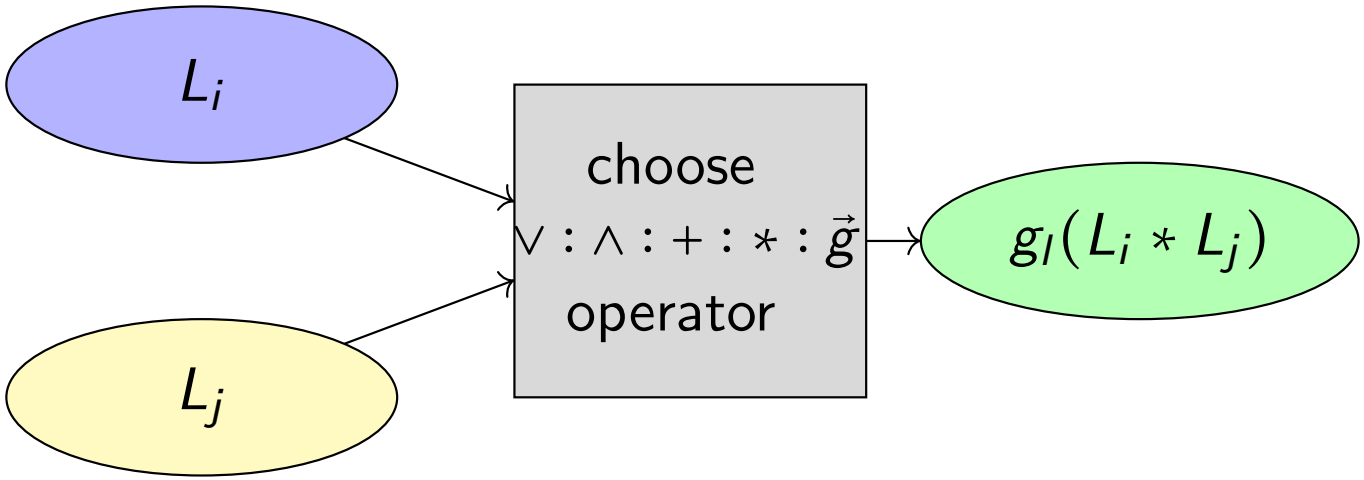

Figure: General feature engineering step illustration.

Further (partly current) research

Generalization will allow (allows) inference like the following:

Table: TP, FP, power and FDR under Kepler's 3rd law model ($\widehat{p}(L|\mathbb{D}) \geq 0.5$).

	Detections			
Models/Last mutation	32 threads 5000/5000	16 threads 5000/5000	4 threads 7500/7500	1 thread 10000/10000
$(HostStarMassSlrMass \times PeriodDays^2)^{\frac{1}{3}}$	70	59	23	7
$(HostStarRadiusSlrRad \times PeriodDays^2)^{\frac{1}{3}}$	29	20	14	7
$(HostStarTempK \times PeriodDays^2)^{\frac{1}{3}}$	1	10	18	15
Other	0	14	85	163
Totally:				
	FDR	FP	Power	Power*
32 threads	0.0000	0	1.0000	0.7000
16 threads	0.1359	14	0.8900	0.5900
4 threads	0.6071	85	0.5500	0.2300
1 thread	0.8624	163	0.2900	0.0700

where the observations *a* are semi major axises of the ellipses of orbits. Explanatory variables include TypeFlag, RadiusJpt, PeriodDays, HostStarMassSlrMass, Eccentricity, PlanetaryMassJpt, HostStarRadiusSlrRad, HostStarMetallicity, and PlanetaryDensJpt.

Concluding remarks

- We introduced the GMJMCMC algorithm for Bayesian logic regression models capable of
  - estimating posterior model probabilities
  - Bayesian model averaging and selection
- *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 hadling 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

References

A. Hubin, G.O. Storvik, F. Frommlet

A novel algorithmic approach to Bayesian Logic Regression.

arXiv:1705.07616v1, 2017.

A. Hubin and G.O. Storvik

Efficient mode jumping MCMC for Bayesian variable selection in GLMM.

arXiv:1604.06398v3, 2016.

C. Kooperberg, and I. Ruczinski.

Identifying Interacting SNPs Using Monte Carlo Logic Regression.

Genetic Epidemiology, 28:157–170, 2005.

A. Fritsch.

A Full Bayesian Version of Logic regression for SNP Data.

PhD thesis, 2006.

The End.

Thank you.