Deep Bayesian regression models

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

Department of Mathematics, University of Oslo

aliaksah@math.uio.no

UiO : Universitetet i Oslo

NORDSTAT 2018,
Tartu

28.06.2018

Introduction. Issues

- Regression models are addressed for inference and prediction in a wide range of applications;
- More and more sources of data are becoming available introducing a variety of hypothetical explanatory variables for these models to be considered;
- Model selection and averaging of different combinations of covariates in this context becomes extremely important for both good inference and prediction;

Introduction. Issues

- It is often the case that linear relations between the explanatory variables and the response are not sufficient;
- One has to avoid unreasonably deep non-linearities to avoid overfitting;
- Ideally models should remain as transparent and dense as possible, or quoting Einstein's famous *"It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience."*

In this work we introduce a class of deep Bayesian regression models and suggest algorithmic approaches for fitting them.

Deep Bayesian regression models

$$Y_i|\mu_i, \phi \sim \mathfrak{f}(y|\mu_i; \phi), \quad i \in \{1, ..., n\} \tag{1}$$

$$\mu_i = h^{-1}\left(\beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j F_j(\boldsymbol{x}) + \sum_{k=1}^{r} \gamma_{k+p} \delta_{ik}\right), \tag{2}$$

$$\boldsymbol{\delta_k} = (\delta_{1k}, ..., \delta_{nk}) \sim N_n\left(\mathbf{0}, \boldsymbol{\Sigma}_k\right). \tag{3}$$

- $f(\cdot|\mu, \phi)$ is a density/distribution with expectation μ and dispersion parameter ϕ;
- $F_j(\boldsymbol{x})$ are all features based on the input explanatory variables ordered w.r.t. complexity, p is the finite number of allowed features;
- $\beta_j \in \mathbb{R}, j \in \{1, ..., p\}$ are regression coefficients of the features;
- $h(\cdot)$ is a proper link function;
- $\gamma_j \in \{0, 1\}, j \in \{1, ..., q = r + p\}$ are latent indicators defining if a feature is included into the model $(\gamma_j = 1)$ or not $(\gamma_j = 0)$.

Hierarchy of the features

A feature $F_j(\boldsymbol{x})$ can be constructed recursively through:

$$F_j(\boldsymbol{x}) = \begin{cases} v(F_k(\boldsymbol{x})), & \text{for a modification;} \\ F_k(\boldsymbol{x}) * F_l(\boldsymbol{x}), & \text{for a crossover;} \\ v(\alpha^T \mathbf{F}(\boldsymbol{x})), & \text{for a projection;} \end{cases}$$

- $F_k(\boldsymbol{x})$ and $F_l(\boldsymbol{x})$ are previously defined features $(k, l < j)$;
- $v \in \mathcal{G}$ is one of the allowed basic function from set $\mathcal{G}$;
- $\mathbf{F}(\boldsymbol{x})$ is a sub-vector of all possible features with indexes lower than j;
- A constraint on the complexity of feature $F_j(\boldsymbol{x})$ is defined by a finite number p of all possible features;
- Projections include modifications and crossovers as particular cases.

Types and meaning of functions in $\mathcal{G}$

- ANN: $\text{logit}(x)$, $\tanh(x)$, $\text{erf}(x)$, $\text{ReLU}(x)$;
- Polynomials: $F_k(\boldsymbol{x}) * F_l(\boldsymbol{x}) = \exp\left(\log(F_k(\boldsymbol{x})) + \log(F_l(\boldsymbol{x}))\right)$;
- Logical *AND* and *OR*: $L_k \wedge L_l = L_k * L_l$ and $L_k \vee L_l = L_k + L_l - L_k * L_l$;
- CART: $\mathrm{I}(x \geq 1)$;
- Fourier series: $\sin(Ax)$ and $\cos(Bx)$;
- Fractional polynomials: $x^{\frac{1}{a}} = \exp\left(b\log(x)\right), b = \frac{1}{a}$;
- RNN and Lagged features: $\text{lag}^k(x)$.

The universal approximation theorem by *Hornik (1991)* is applicable if at least one of v functions is strictly monotonous and bounded.

Model priors

$$p(\gamma) \propto \mathrm{I}\left(|\gamma_{1:p}| \le Q\right)\mathrm{I}\left(|\gamma_{p+1:q}| \le R\right)\prod_{j=1}^{p} a^{\gamma_j w_j c(F_j(\mathbf{x}))} \prod_{k=p+1}^{q} b^{\gamma_k \omega_k c(\delta_k)}. \quad (4)$$

- $a, b \in (0,1)$, $Q \le p$, $R \le r$;
- $|\gamma_{1:K}| = \sum_{j=1}^{K} \gamma_j$ is the number of active features in subset $\{\gamma_1, ..., \gamma_K\}$;
- $c(F_j(\mathbf{x})) \ge 0$ is a measure of complexity for a feature $F_j(\mathbf{x})$;
- $c(\delta_k) \ge 0$ is a measure of complexity for a latent Gaussian variable δ_k;
- w_j are weights for complexities of the corresponding features;
- ω_k are weights for the complexities of the corresponding latent Gaussian variables.

Parameter priors

Particular choices are given in the application to follow:

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \pi_{\beta}(\boldsymbol{\beta}), \tag{5}$$

$$\boldsymbol{\psi}_k|\boldsymbol{\gamma} \sim \pi_k(\boldsymbol{\psi}_k), \tag{6}$$

$$\phi \sim \pi_{\phi}(\phi). \tag{7}$$

Prior distributions on $\boldsymbol{\beta}_j|\gamma$, ϕ (if present) and $\boldsymbol{\psi}_k|\gamma$ (if latent Gaussian variables are present) are usually selected in a way to efficiently compute marginal likelihoods of the models (by for example specifying conjugate priors) and should be carefully specified for the applications of interest.

Inference on the model

Let:

- $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_q\}$ define a model itself, i.e. which components(features or latent variables) are addressed;
- $\theta_{\boldsymbol{\gamma}} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \phi)$ define parameters of the model $\boldsymbol{\gamma}$.

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:

- $\exists 2^q$ different models in Ω_γ;
- q is huge;
- Ω_γ is extremely difficult to specify.

Marginal likelihood approximation.

Marginal likelihood:

$$p(\mathbb{D}|\gamma) = \int_{\Theta} p(\mathbb{D}|, \theta_\gamma, \gamma) p(\theta_\gamma|\gamma) d\theta_\gamma \tag{8}$$

can be obtained analytically or approximated by various methods:

- Laplace approximations;
- Integrated nested Laplace approximations;
- Chib's method (via Gibbs sampling);
- Chib and Jeliazkov's method (via MH algorithm);
- Variational Inference.

Computation of the marginal likelihood is the major computational bottleneck currently. No efficient methods with subsampling are currently available.

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 × 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 must be clearly specified for MJMCMC. The later is **not feasible** in **Deep Regression Models**.

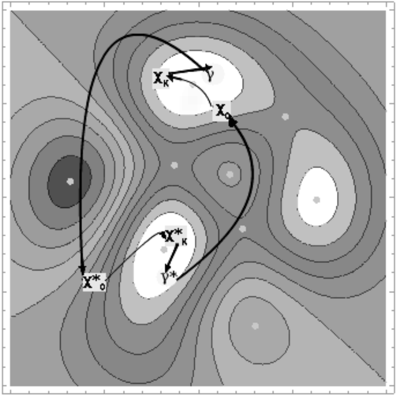


Figure: Locally optimized with randomization proposals

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 features (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**, **modification**, **projection** and **filtration** operations to the current $\mathcal{S}_{t-1}$ and a special set $\mathcal{F}$;
- Utilization of the approximation (9) allows us to compute marginal inclusion probabilities

$$\widehat{p}(\gamma_j = 1|\mathbb{D}) = \sum_{\boldsymbol{\gamma}\in\mathbb{V}} \mathbf{I}(\gamma_j = 1)\widehat{p}(\boldsymbol{\gamma}|\mathbb{D}). \quad (10)$$

Reversible version of GMJMCMC - RGMJMCMC

- For every search space we suggest running MJMCMC for a sufficient amount of time;
- Otherwise a transition $\gamma \to \mathcal{S}' \to \gamma^1 \to \dots \to \gamma^k \to \gamma'$ is considered with a given probability kernel;
  - $q(\mathcal{S}'|\gamma)$ is the proposal for the new search space;
  - Transitions $\gamma^1 \to \dots \to \gamma^k$ are generated by local MJMCMC $\mathcal{S}'$;
  - Transition $\gamma^k \to \gamma'$ is some randomization at the end of the procedure.
- Acceptance probability for such a procedure is $r_m = \min\{1, \alpha_m\}$

$$\alpha_m = \frac{\pi(\gamma')q(\gamma|\gamma'_k)}{\pi(\gamma)q(\gamma'|\gamma_k)}. \tag{11}$$

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

$$\widehat{p}(\Delta|\mathbb{D}) = \sum_{\boldsymbol{\gamma}\in\mathbb{V}} p(\Delta|\boldsymbol{\gamma},\mathbb{D})\widehat{p}(\boldsymbol{\gamma}|\mathbb{D}) \ . \tag{12}$$

The model

For this example as well as 2 following examples we have addressed the following DBRM with conditionally independent Gaussian observations:

$$Y_i|\mu_i,\sigma^2 \sim N(\mu_i,\sigma^2), i \in \{1,...,n\}, \tag{13}$$

$$\mu_i = \gamma_0\beta_0 + \textstyle\sum_{j=1}^{p} \gamma_j\beta_j F_j(\boldsymbol{x}_i). \tag{14}$$

$$p(\gamma) \propto \mathsf{I}\left(|\gamma_{1:p}| \leq 15\right) \textstyle\prod_{j=1}^{p} \exp(-2\log n\gamma_j c(F_j(\mathbf{x}))) \tag{15}$$

$$p(\boldsymbol{\beta}|\gamma) = |J_n^{\gamma}(\boldsymbol{\beta})|^{\frac{1}{2}} \ , \tag{16}$$

$$\pi(\sigma^2) = \sigma^{-2}, \tag{17}$$

where $c(F_j(\mathbf{x}))$ is the sum of the number additive, multiplicative and non-linear complications in feature j and $|J_n^{\gamma}(\boldsymbol{\beta})|$ is the determinant of the corresponding Fisher information matrix. The non-linear modifications in this example include cos(x), sigmoid(x), tanh(x), atan(x), sin(x), and $|x|^{\frac{1}{3}}$.

Simulation scenario. Results over 100 datasets

Table: Results for the three simulation scenario. Power for individual expression, overall power, expected number of false positives (FP), and FDP.

	GMJ	RGMJ	GMJ(logic)
X_7	1.0000	1.0000	0.9900
X_8	1.0000	1.0000	1.0000
$X_2 * X_9$	1.0000	0.9600	1.0000
$X_{18} * X_{21}$	1.0000	1.0000	0.9600
$X_1 * X_3 * X_{27}$	1.0000	1.0000	1.0000
$X_{12} * X_{20} * X_{37}$	1.0000	1.0000	0.9900
$X_4 * X_{10} * X_{17} * X_{30}$	0.9900	0.9200	0.9100
$X_{11} * X_{13} * X_{19} * X_{50}$	0.9800	0.8900	0.3800
Overall Power	0.9963	0.9712	0.9038
FP	0.5100	1.1400	1.0900
FDP	0.0601	0.1279	0.1310

Ground physical laws inference

From the **ground physical** laws:

$$m_p \propto R_p^3 \times \rho_p,$$

m_p is *PlanetaryMassJpt*, R_p^3 is *RadiusJpt*^{3}, and ρ_p is *PlanetaryDensJpt*. And:

$$a = \left(\frac{GP^2}{4\pi^2}(M_h + m_p)\right)^{\frac{1}{3}} \approx \left(\frac{GP^2}{4\pi^2}(M_s M_h^a)\right)^{\frac{1}{3}},$$

a is semi major axises of the ellipses of the orbits, M_h^a is *HostStarMassSlrMass*, a is *SemiMajorAxisAU*, and P is *PeriodDays*.

Generally the data has the **following other variables**: *TypeFlag*, *RadiusJpt*, *PeriodDays*, *PlanetaryMassJpt*, *Eccentricity*, *HostStarMassSlrMass*, *HostStarRadiusSlrRad*, *HostStarMetallicity*, *HostStarTempK*
PlanetaryDensJpt denoted as x_1-x_{10}

Planet mass inference. Results over 100 simulations

Table: Power, False Positives (FP) and FDP based on the decision rule that the posterior probability of a feature is larger than 0.25. The feature *PlanetaryRadiusJpt*^{3} *PlanetaryDensJpt* is counted as true positive, all other selected features as false positive.

	GMJMCMC			RGMJMCMC		
Threads	Power	FP	FDP	Power	FP	FDP
16	1.00	0.00	0.00	0.97	0.06	0.058
4	0.79	0.40	0.34	0.61	0.73	0.54
1	0.43	1.21	0.74	0.32	1.67	0.84

3rd Kepler's law inference. Results over 100 simulations

Table: Comparison of Results on Example 3 for GMJMCMC and RGMJMCMC using different number of threads. F_1, F_2 and F_3 refer to the number of times the specific features $(\mathit{HostStarMassSlrMass} \times \mathit{PeriodDays}^2)^{\frac{1}{3}}$, $(\mathit{HostStarRadiusSlrRad} \times \mathit{PeriodDays}^2)^{\frac{1}{3}}$ and $(\mathit{HostStarTempK} \times \mathit{PeriodDays}^2)^{\frac{1}{3}}$ had a posterior probability larger than 0.25. Power gives the percentage of runs where at least one of these three features was detected. FP counts the number of other features and FDP is the corresponding false discovery proportion.

	GMJMCMC						RGMJMCMC					
Th	F_1	F_2	F_3	Power	FP	FDP	F_1	F_2	F_3	Power	FP	FDP
64	81	71	1	1.00	0.02	0.013	78	75	2	0.99	0.03	0.019
32	63	58	11	0.99	0.14	0.11	55	57	9	0.95	0.12	0.09
16	34	41	32	0.84	0.46	0.30	31	38	18	0.79	0.68	0.44
4	15	10	16	0.38	1.05	0.62	8	14	8	0.29	1.47	0.83
1	6	5	3	0.13	1.46	0.82	6	4	2	0.12	1.81	0.94

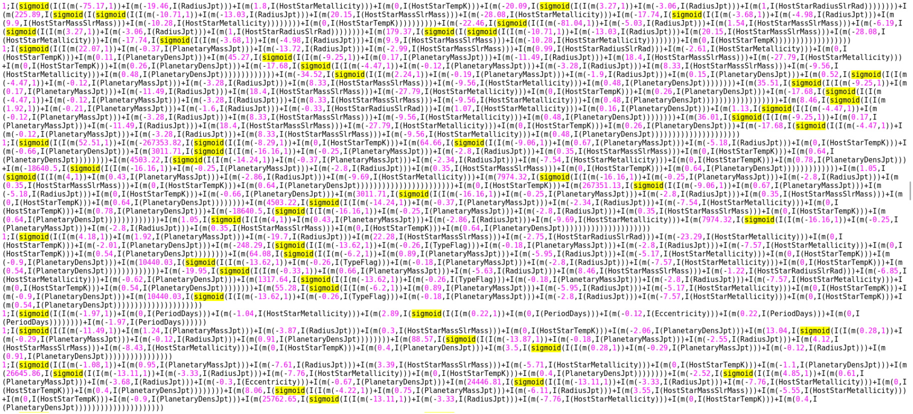
What would have happened with simpler ANN

1. When $\mathcal{G} = \{\text{sigmoid}(x)\}$;
2. When $\mathcal{G} = \{\text{sigmoid}(x)\}$, $D_{max} = 300$, and $P_c = 0$;
3. When $\mathcal{G} = \{\text{sigmoid}(x)\}$, $D_{max} = 300$, and $P_c = 0$ and $p(\gamma_j) \propto 1$.

Table: 10 most frequent features detected under scenarios 1, 2 and 3

Fq	Feature	Fq	Feature	Fq	Feature
99	x_3	100	x_3	100	x_3
98	x_3*x_3	72	g_σ(-10.33+0.24x_4-8.83x_8)	54	x_2
93	x_3*x_{10}	64	x_{10}	21	g_σ(-16.91-4.94x_2)
4	x_3*x_3*x_{10}	62	x_2	19	x_9
1	x_9*x_3	16	g_σ(0.21+0.01x_3+0.20x_7)	16	x_5
1	x_9*x_3*x_3	9	x_4	14	x_{10}
1	x_{10}*x_{10}*x_3	7	g_σ(-13.11-7.76x_8-3.33x_2+0.40x_{10})	10	g_σ(6.88×10^9-3.92x_2+ 3.44×$10^9$$g_\sigma$(-13.57-0.17$x_4-2.84x_2-7.66x_8$+0.54$x_{10}$) -13.76×$10^9$$g_\sigma$($g_\sigma$(-13.57-0.17$x_4-2.84x_2-7.66x_8$+ 0.54$x_{10}$)))
1	x_7*x_3*x_3	5	g_σ(-3.36+2.83x_3+0.21x_3-3.36x_9)	9	x_4
1	x_6*x_3*x_3	3	g_σ(g_σ(-10.33+0.24x_4)-8.83x_8)	8	g_σ(-13.57-0.17x_4-2.84x_2-7.66x_8+0.54x_{10})
1	x_3*x_3*x_3	3	g_σ(0.15+0.05x_4-0.01x_3+0.15x_7)	7	g_σ(0.21+0.21x_3)
0	Others	4	Others	> 300	Others

Trash features under the non regularized case



```
1;I(sigmoid(I(I(m(-75.17,1))+I(m(-19.46,I(RadiusJpt)))+I(m(1.8,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(-20.09,I(sigmoid(I(I(m(3.27,1))+I(m(-3.06,I(RadiusJpt)))+I(m(1,I(HostStarRadiusSlrRad)))))))+I
(m(225.89,I(sigmoid(I(sigmoid(I(I(m(-10.71,1))+I(m(-13.03,I(RadiusJpt)))+I(m(20.15,I(HostStarMassSlrMass)))+I(m(-28.08,I(HostStarMetallicity)))+I(m(-17.74,I(sigmoid(I(I(m(-3.68,1))+I(m(-4.98,I(RadiusJpt)))+I(m
(9.9,I(HostStarMassSlrMass)))+I(m(-10.28,I(HostStarMetallicity))))))))+I(m(0,I(HostStarTempK)))))))))+I(m(-22.46,I(sigmoid(I(I(m(-81.04,1))+I(m(-5.03,I(RadiusJpt)))+I(m(1.54,I(HostStarMassSlrMass)))+I(m(-6.19,I
(sigmoid(I(I(m(3.27,1))+I(m(-3.06,I(RadiusJpt)))+I(m(1,I(HostStarRadiusSlrRad))))))))+I(m(179.37,I(sigmoid(I(sigmoid(I(I(m(-10.71,1))+I(m(-13.03,I(RadiusJpt)))+I(m(20.15,I(HostStarMassSlrMass)))+I(m(-28.08,I
(HostStarMetallicity)))+I(m(-17.74,I(sigmoid(I(I(m(-3.68,1))+I(m(-4.98,I(RadiusJpt)))+I(m(9.9,I(HostStarMassSlrMass)))+I(m(-10.28,I(HostStarMetallicity))))))))+I(m(0,I(HostStarTempK))))))))))))))))
1;I(sigmoid(I(I(m(22.07,1))+I(m(-0.37,I(PlanetaryMassJpt)))+I(m(-13.72,I(RadiusJpt)))+I(m(-2.99,I(HostStarMassSlrMass)))+I(m(0.99,I(HostStarRadiusSlrRad)))+I(m(-2.61,I(HostStarMetallicity)))+I(m(0,I
(HostStarTempK)))+I(m(0.11,I(PlanetaryDensJpt)))+I(m(45.27,I(sigmoid(I(I(m(-9.25,1))+I(m(0.17,I(PlanetaryMassJpt)))+I(m(-11.49,I(RadiusJpt)))+I(m(18.4,I(HostStarMassSlrMass)))+I(m(-27.79,I(HostStarMetallicity)))
+I(m(0,I(HostStarTempK)))+I(m(0.26,I(PlanetaryDensJpt)))+I(m(-17.68,I(sigmoid(I(I(m(-4.47,1))+I(m(-0.12,I(PlanetaryMassJpt)))+I(m(-3.28,I(RadiusJpt)))+I(m(8.33,I(HostStarMassSlrMass)))+I(m(-9.56,I
(HostStarMetallicity)))+I(m(0.48,I(PlanetaryDensJpt))))))))))))+I(m(-34.52,I(sigmoid(I(I(m(2.24,1))+I(m(-0.19,I(PlanetaryMassJpt)))+I(m(-1.9,I(RadiusJpt)))+I(m(0.15,I(PlanetaryDensJpt)))+I(m(0.52,I(sigmoid(I(I(m
(-4.47,1))+I(m(-0.12,I(PlanetaryMassJpt)))+I(m(-3.28,I(RadiusJpt)))+I(m(8.33,I(HostStarMassSlrMass)))+I(m(-9.56,I(HostStarMetallicity)))+I(m(0.48,I(PlanetaryDensJpt))))))))+I(m(35.51,I(sigmoid(I(I(m(-9.25,1))+I(m
(0.17,I(PlanetaryMassJpt)))+I(m(-11.49,I(RadiusJpt)))+I(m(18.4,I(HostStarMassSlrMass)))+I(m(-27.79,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(0.26,I(PlanetaryDensJpt)))+I(m(-17.68,I(sigmoid(I(I(m
(-4.47,1))+I(m(-0.12,I(PlanetaryMassJpt)))+I(m(-3.28,I(RadiusJpt)))+I(m(8.33,I(HostStarMassSlrMass)))+I(m(-9.56,I(HostStarMetallicity)))+I(m(0.48,I(PlanetaryDensJpt))))))))))))))))+I(m(8.46,I(sigmoid(I(I(m
(1.92,1))+I(m(-0.21,I(PlanetaryMassJpt)))+I(m(-1.6,I(RadiusJpt)))+I(m(-0.33,I(HostStarRadiusSlrRad)))+I(m(1.07,I(HostStarMetallicity)))+I(m(0.16,I(PlanetaryDensJpt)))+I(m(1.13,I(sigmoid(I(I(m(-4.47,1))+I(m
(-0.12,I(PlanetaryMassJpt)))+I(m(-3.28,I(RadiusJpt)))+I(m(8.33,I(HostStarMassSlrMass)))+I(m(-9.56,I(HostStarMetallicity)))+I(m(0.48,I(PlanetaryDensJpt))))))))+I(m(36.01,I(sigmoid(I(I(m(-9.25,1))+I(m(0.17,I
(PlanetaryMassJpt)))+I(m(-11.49,I(RadiusJpt)))+I(m(18.4,I(HostStarMassSlrMass)))+I(m(-27.79,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(0.26,I(PlanetaryDensJpt)))+I(m(-17.68,I(sigmoid(I(I(m(-4.47,1))+I
(m(-0.12,I(PlanetaryMassJpt)))+I(m(-3.28,I(RadiusJpt)))+I(m(8.33,I(HostStarMassSlrMass)))+I(m(-9.56,I(HostStarMetallicity)))+I(m(0.48,I(PlanetaryDensJpt)))))))))))))))))))
1;I(sigmoid(I(I(m(52.51,1))+I(m(-267353.82,I(sigmoid(I(I(m(-8.29,1))+I(m(0,I(HostStarTempK)))+I(m(64.66,I(sigmoid(I(I(m(-9.06,1))+I(m(0.67,I(PlanetaryMassJpt)))+I(m(-5.18,I(RadiusJpt)))+I(m(0,I(HostStarTempK)))+I
(m(-0.66,I(PlanetaryDensJpt)))+I(m(3011.71,I(sigmoid(I(I(m(-16.16,1))+I(m(-0.25,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(0.35,I(HostStarMassSlrMass)))+I(m(0,I(HostStarTempK)))+I(m(0.64,I
(PlanetaryDensJpt))))))))+I(m(4503.22,I(sigmoid(I(I(m(-14.24,1))+I(m(-0.37,I(PlanetaryMassJpt)))+I(m(-2.34,I(RadiusJpt)))+I(m(-7.54,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(0.78,I(PlanetaryDensJpt)))
+I(m(-18640.5,I(sigmoid(I(I(m(-16.16,1))+I(m(-0.25,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(0.35,I(HostStarMassSlrMass)))+I(m(0,I(HostStarTempK)))+I(m(0.64,I(PlanetaryDensJpt))))))))))))+I(m(1.05,I
(sigmoid(I(I(m(4,1))+I(m(0.43,I(PlanetaryMassJpt)))+I(m(-2.86,I(RadiusJpt)))+I(m(-9.69,I(HostStarMetallicity)))+I(m(7974.32,I(sigmoid(I(I(m(-16.16,1))+I(m(-0.25,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m
(0.35,I(HostStarMassSlrMass)))+I(m(0,I(HostStarTempK)))+I(m(0.64,I(PlanetaryDensJpt))))))))))))))))))))))+I(m(0,I(HostStarTempK)))+I(m(267351.13,I(sigmoid(I(I(m(-9.06,1))+I(m(0.67,I(PlanetaryMassJpt)))+I(m
(-5.18,I(RadiusJpt)))+I(m(0,I(HostStarTempK)))+I(m(-0.66,I(PlanetaryDensJpt)))+I(m(3011.71,I(sigmoid(I(I(m(-16.16,1))+I(m(-0.25,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(0.35,I(HostStarMassSlrMass)))+I(m
(0,I(HostStarTempK)))+I(m(0.64,I(PlanetaryDensJpt))))))))+I(m(4503.22,I(sigmoid(I(I(m(-14.24,1))+I(m(-0.37,I(PlanetaryMassJpt)))+I(m(-2.34,I(RadiusJpt)))+I(m(-7.54,I(HostStarMetallicity)))+I(m(0,I
(HostStarTempK)))+I(m(0.78,I(PlanetaryDensJpt)))+I(m(-18640.5,I(sigmoid(I(I(m(-16.16,1))+I(m(-0.25,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(0.35,I(HostStarMassSlrMass)))+I(m(0,I(HostStarTempK)))+I(m
(0.64,I(PlanetaryDensJpt))))))))))))+I(m(1.05,I(sigmoid(I(I(m(4,1))+I(m(0.43,I(PlanetaryMassJpt)))+I(m(-2.86,I(RadiusJpt)))+I(m(-9.69,I(HostStarMetallicity)))+I(m(7974.32,I(sigmoid(I(I(m(-16.16,1))+I(m(-0.25,I
(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(0.35,I(HostStarMassSlrMass)))+I(m(0,I(HostStarTempK)))+I(m(0.64,I(PlanetaryDensJpt))))))))))))))))))))
1;I(sigmoid(I(I(m(4.18,1))+I(m(1.92,I(PlanetaryMassJpt)))+I(m(-19.7,I(RadiusJpt)))+I(m(22.28,I(HostStarMassSlrMass)))+I(m(-2.75,I(HostStarRadiusSlrRad)))+I(m(-23.29,I(HostStarMetallicity)))+I(m(0,I
(HostStarTempK)))+I(m(-2.01,I(PlanetaryDensJpt)))+I(m(-248.29,I(sigmoid(I(I(m(-13.62,1))+I(m(-0.26,I(TypeFlag)))+I(m(-0.18,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(-7.57,I(HostStarMetallicity)))+I(m(0,I
(HostStarTempK)))+I(m(0.54,I(PlanetaryDensJpt))))))))+I(m(64.08,I(sigmoid(I(I(m(-6.2,1))+I(m(0.89,I(PlanetaryMassJpt)))+I(m(-5.95,I(RadiusJpt)))+I(m(-5.17,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m
(-0.9,I(PlanetaryDensJpt)))+I(m(10440.03,I(sigmoid(I(I(m(-13.62,1))+I(m(-0.26,I(TypeFlag)))+I(m(-0.18,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(-7.57,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m
(0.54,I(PlanetaryDensJpt))))))))))))+I(m(-19.95,I(sigmoid(I(I(m(-0.33,1))+I(m(0.66,I(PlanetaryMassJpt)))+I(m(-5.63,I(RadiusJpt)))+I(m(8.46,I(HostStarMassSlrMass)))+I(m(-1.22,I(HostStarRadiusSlrRad)))+I(m(-6.85,I
(HostStarMetallicity)))+I(m(-0.62,I(PlanetaryDensJpt)))+I(m(1317.64,I(sigmoid(I(I(m(-13.62,1))+I(m(-0.26,I(TypeFlag)))+I(m(-0.18,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(-7.57,I(HostStarMetallicity)))+I
(m(0,I(HostStarTempK)))+I(m(0.54,I(PlanetaryDensJpt))))))))+I(m(55.28,I(sigmoid(I(I(m(-6.2,1))+I(m(0.89,I(PlanetaryMassJpt)))+I(m(-5.95,I(RadiusJpt)))+I(m(-5.17,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I
(m(-0.9,I(PlanetaryDensJpt)))+I(m(10440.03,I(sigmoid(I(I(m(-13.62,1))+I(m(-0.26,I(TypeFlag)))+I(m(-0.18,I(PlanetaryMassJpt)))+I(m(-2.8,I(RadiusJpt)))+I(m(-7.57,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I
(m(0.54,I(PlanetaryDensJpt))))))))))))))))))))
1;I(sigmoid(I(I(m(-1.97,1))+I(m(0,I(PeriodDays)))+I(m(-1.04,I(HostStarMetallicity)))+I(m(2.89,I(sigmoid(I(I(m(0.22,1))+I(m(0,I(PeriodDays)))+I(m(-0.12,I(Eccentricity)))+I(m(0.22,I(PeriodDays)))+I(m(0,I
(PeriodDays))))))))+I(m(-1.97,I(PeriodDays))))))
1;I(sigmoid(I(I(m(-11.49,1))+I(m(1.24,I(PlanetaryMassJpt)))+I(m(-3.87,I(RadiusJpt)))+I(m(0.3,I(HostStarMassSlrMass)))+I(m(0,I(HostStarTempK)))+I(m(-2.06,I(PlanetaryDensJpt)))+I(m(13.04,I(sigmoid(I(I(m(0.28,1))+I
(m(-0.29,I(PlanetaryMassJpt)))+I(m(-0.12,I(RadiusJpt)))+I(m(0.91,I(PlanetaryDensJpt))))))))+I(m(88.57,I(sigmoid(I(I(m(-13.87,1))+I(m(-0.18,I(PlanetaryMassJpt)))+I(m(-2.55,I(RadiusJpt)))+I(m(4.12,I
(HostStarMassSlrMass)))+I(m(-8.43,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(0.4,I(PlanetaryDensJpt)))+I(m(3.5,I(sigmoid(I(I(m(0.28,1))+I(m(-0.29,I(PlanetaryMassJpt)))+I(m(-0.12,I(RadiusJpt)))+I(m
(0.91,I(PlanetaryDensJpt))))))))))))))))
1;I(sigmoid(I(I(m(-1.08,1))+I(m(0.95,I(PlanetaryMassJpt)))+I(m(-7.61,I(RadiusJpt)))+I(m(3.39,I(HostStarMassSlrMass)))+I(m(-5.71,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(-1.1,I(PlanetaryDensJpt)))+I(m
(26645.86,I(sigmoid(I(I(m(-13.11,1))+I(m(-3.33,I(RadiusJpt)))+I(m(-7.76,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(0.4,I(PlanetaryDensJpt))))))))+I(m(-2.52,I(sigmoid(I(I(m(4.85,1))+I(m(0.61,I
(PlanetaryMassJpt)))+I(m(-3.68,I(RadiusJpt)))+I(m(-0.3,I(Eccentricity)))+I(m(-0.67,I(PlanetaryDensJpt)))+I(m(24446.81,I(sigmoid(I(I(m(-13.11,1))+I(m(-3.33,I(RadiusJpt)))+I(m(-7.76,I(HostStarMetallicity)))+I(m(0,I
(HostStarTempK)))+I(m(0.4,I(PlanetaryDensJpt))))))))+I(m(8.06,I(sigmoid(I(I(m(-4.22,1))+I(m(0.75,I(PlanetaryMassJpt)))+I(m(-6.11,I(RadiusJpt)))+I(m(3.55,I(HostStarMassSlrMass)))+I(m(-5.55,I(HostStarMetallicity)))
+I(m(0,I(HostStarTempK)))+I(m(-0.9,I(PlanetaryDensJpt)))+I(m(25762.65,I(sigmoid(I(I(m(-13.11,1))+I(m(-3.33,I(RadiusJpt)))+I(m(-7.76,I(HostStarMetallicity)))+I(m(0,I(HostStarTempK)))+I(m(0.4,I
(PlanetaryDensJpt))))))))))))))))))))
```

Figure: A snapshot of features detected under Scenario 3

Application. Breast cancer classification. Problem

- **Observations**: 357 benign, 212 malignant tissues
- **Covariates**: 10 different covariates describing objects
- **Logistic deep Bayesian regression** addressed

$$y_i = y|\rho_i \sim \text{Binom}(1, \rho_i) \tag{18}$$

$$\rho_i = \frac{e^{\gamma_0\beta_0+\sum_{j=1}^{p}\gamma_j\beta_jF_j(\mathbf{x}_i)}}{1+e^{\gamma_0\beta_0+\sum_{j=1}^{q}\gamma_j\beta_jF_j(\mathbf{x}_i)}} \tag{19}$$

$$p(\boldsymbol{\gamma}) \propto \mathrm{I}\left(|\gamma_{1:p}| \leq 20\right)\prod_{j=1}^{p}\exp(-\gamma_j 2c(F_j(\mathbf{x}))) \tag{20}$$

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \left|J_n^{\gamma}(\boldsymbol{\beta})\right|^{\frac{1}{2}}, \tag{21}$$

Input variables include such parameters of the tissues as:
Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension.
The **allowed modifications** are:sigmoid(x), $\mathrm{I}(x > 1)$, ReLu(x), $x^{\frac{1}{3}}$, $x^{\frac{1}{5}}$.

Real data analysis. Breast cancer classification

- Runs of GMJMCMC with 1 thread and 32 threads are addressed;
- $||$**training set**$||$ = 142, $||$**test set**$||$ = 427;
- Prediction is based on marginalized over all models' probabilities, namely $\hat{Y} = \mathbf{I}\left\{\widehat{p}(Y|\mathbb{D}) \geq 0.5\right\}, \widehat{p}(Y|\mathbb{D}) = \sum_{\gamma\in\mathbb{V}} \widehat{p}(Y|\gamma,\mathbb{D})\widehat{p}(\gamma|\mathbb{D})$;

Algorithm	min.p	med.p	max.p	min.fn	med.fn	max.fn	min.fp	med.fp	max.fp
PARRGMJ	**0.9695**	**0.9765**	**0.9812**	0.0479	0.0479	0.0536	**0.0000**	0.0074	0.0184
PARGMJ	0.9671	0.9742	0.9812	0.0479	0.0479	0.0536	0.0000	0.0111	0.0220
RIDGE	0.9742	0.9742	0.9742	0.0592	0.0592	0.0592	0.0037	**0.0037**	**0.0037**
GMJMCMC	0.9437	0.9695	**0.9812**	0.0479	0.0536	0.1067	**0.0000**	0.0148	0.0361
DEEPNETS	0.9225	0.9695	0.9789	**0.0305**	0.0674	0.1167	0.0000	0.0074	0.0949
RGMJMCMC	0.9554	0.9683	0.9789	0.0479	0.0536	0.0809	0.0037	0.0148	0.0361
LR	0.9671	0.9671	0.9671	0.0479	0.0479	0.0479	0.0220	0.0220	0.0220
LASSO	0.9577	0.9577	0.9577	0.0756	0.0756	0.0756	0.0184	0.0184	0.0184
LXGBOOST	0.9554	0.9554	0.9554	0.0809	0.0809	0.0809	0.0184	0.0184	0.0184
TXGBOOST	0.9484	0.9531	0.9601	0.0536	0.0647	0.0756	0.0291	0.0326	0.0361
RFOREST	0.9038	0.9343	0.9624	0.0422	0.0914	0.1675	0.0000	0.0361	0.1010
NAIVEBAYESS	0.9272	0.9272	0.9272	0.0305	**0.0305**	**0.0305**	0.0887	0.0887	0.0887

Application. Epigenetic study

- **Observations**: Epigenetically modified reads
- **Covariates**: 14 different covariates
- **Poisson deep Bayesian mixed regression** addressed

$$y_i|\mu_i \sim \text{Poisson}(\mu_i) \tag{22}$$

$$\mu_i = e^{\gamma_0\beta_0+\sum_{j=1}^{p}\gamma_j\beta_jF_j(\boldsymbol{x}_i)+\sum_{k=1}^{r}\gamma_{k+p}\delta_{ik}} \tag{23}$$

$$p(\boldsymbol{\gamma}) \propto \mathbf{I}\left(|\boldsymbol{\gamma}_{1:p}| \le 30\right)\mathbf{I}\left(|\boldsymbol{\gamma}_{p+1:q}| \le 2\right)\prod_{j=1}^{p}\exp(-\log n\gamma_j c(F_j(\mathbf{x}))) \tag{24}$$

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_{p_{\boldsymbol{\gamma}}}(\boldsymbol{\mu}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}). \tag{25}$$

Latent variables include: ***AR*(1)**, ***RW*(1)**, ***OU***, ***IG*** processes

Input variables include: *3 levels corresponding to whether a location belongs to a CGH, CHH or CHG, whether a distance to the previous (C) in DNA is 1, 2, 3, 4, 5, from 6 to 20 or greater than 20, whether a location belongs to a gene from a particular group of genes of biological interest,whether expression is greater than 3000 or greater than 10000, the offset for the number of total bases per location offset(log(total.bases)).* .

The **allowed modifications** are: cos(x), sigmoid(x), tanh(x), atan(x),

Epigenetic study. Results

No non-linearities found!?

Variable	Posterior
f(model="rw1")	1.00000e+00
offset(log(total.bases))	1.00000e+00
CG	9.990546e-01
CHG	9.515581e-01

Table: Features with posterior probability above 0.25 found by GMJMCMC with 16 threads for the epigenetic data example

Concluding remarks

- We introduced the (R)GMJMCMC algorithm for deep 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 handling is incorporated
- Results showed that (R)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:1705.07616v2, 2018.

A. Hubin and G.O. Storvik
Mode jumping MCMC for Bayesian variable selection in GLMM.
Journal of Computational Statistics and Data Analysis; 2018 November; 127:281-297.

A. Hubin, G.O. Storvik, F. Frommlet
Deep Bayesian regression models.
arXiv:1806.02160v2, 2018.

The End.

Thank you.