Deep Bayesian regression models

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

- 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;
- 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."*

Model specification

The deep Bayesian regression model

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

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

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

- $f(\cdot|\mu,\phi)$ 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 ($\gamma_j = 0$).

Model priors

Latent coefficients:

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

- $a,b \in (0,1)$, $Q \le p$, $R \le r$;
- $|\boldsymbol{\gamma}_{1:K}| = \sum_{j=1}^{K}\gamma_j$ 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;

Model parameters:

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

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

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

Hierarchy of the features

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

$$F_j(\boldsymbol{x}) = v(\boldsymbol{\alpha}^T\mathbf{F}(\boldsymbol{x})). \tag{8}$$

- $v \in \mathcal{G}$ 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.

Posterior evaluation

- **Consider** marginal likelihood $p(\mathbb{D}|\boldsymbol{\gamma}) = \int_\Theta p(\mathbb{D}|,\theta_{\boldsymbol{\gamma}},\boldsymbol{\gamma})p(\theta_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})d\theta_{\boldsymbol{\gamma}}$ are computable [1];
- **Notice that** $p(\boldsymbol{\gamma},\theta_{\boldsymbol{\gamma}}|\mathbb{D}) = p(\theta_{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\mathbb{D})p(\boldsymbol{\gamma}|\mathbb{D})$;
- **Notice that** $p(\boldsymbol{\gamma}|\mathbb{D}) = \frac{p(\mathbb{D}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'\in\Omega_{\boldsymbol{\gamma}}}p(\mathbb{D}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')} \approx \widehat{p}(\boldsymbol{\gamma}|\mathbb{D}) = \frac{p(\mathbb{D}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'\in\mathbb{V}}p(\mathbb{D}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}$;
- $\mathbb{V}$ is the **subspace** of $\Omega_{\boldsymbol{\gamma}}$ to be **efficiently explored**;

In [2] **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 **DBRM**. To solve this problem we present the Genetically Modified MJMCMC (GMJMCMC) [3, 4] algorithm as well as its reversible version - RGMJMCMC[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.

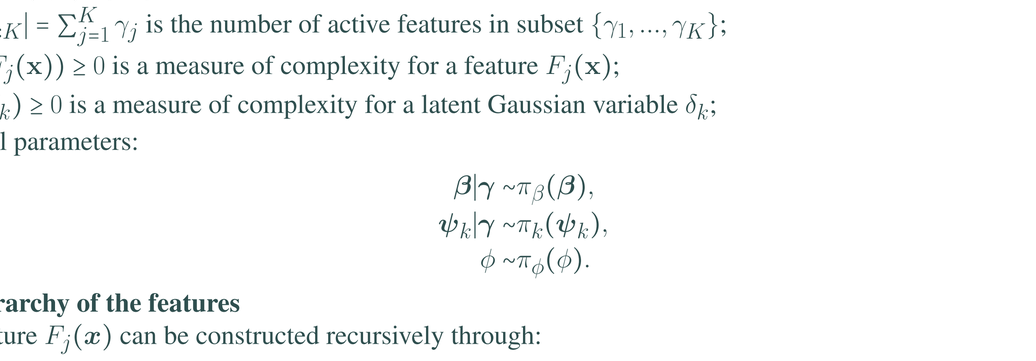

Figure 1: Illustration of locally optimized proposals (left) and MDS (multidimensional scaling) plot of the best 1024 models in terms of PMP in the space of models for some epigenetic data (right).

The GMJMCMC algorithm

Algorithm 1 GMJMCMC

1: Initialize $\mathcal{S}_0$
2: Run the MJMCMC algorithm within search space $\mathcal{S}_0$ for N_{init} iterations and use results to initialize $\mathcal{S}_1$
3: **for** $t = 1,...,T_{max-1}$ **do**
4: Run the MJMCMC algorithm within search space $\mathcal{S}_t$ for N_{expl} iterations.
5: Generate a new population $\mathcal{S}_{t+1}$ using genetic operators
6: **end for**
7: Run the MJMCMC algorithm within search space $\mathcal{S}_{T_{max}}$ for N_{final} iterations.

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

Planetary mass and third Kepler's laws

$$m_p \approx K_1 R_p^3 \times \rho_p \tag{9}$$

$$a \approx K_2\left(P^2 M_h\right)^{\frac{1}{3}}. \tag{10}$$

Here the mass of the hosing star M_h is measured in the unit of Solar mass. We use $Y_i|\mu_i,\sigma^2 \sim N(\mu_i,\sigma^2)$, $h(\mu_i) = \mu_i$, $\pi(\sigma^2) = \sigma^{-2}$, $p(\boldsymbol{\beta}|\boldsymbol{\gamma},\sigma^2) = |J_n^{\boldsymbol{\gamma}}(\boldsymbol{\beta},\sigma^2)|^{\frac{1}{2}}$, $a = e^{-\log n}$, $Q = 15$, $q = 0$, and $\mathcal{G} = \{\mathrm{sigmoid}(x),\sin(x),\tanh(x),\mathrm{atan}(x),|x|^{\frac{1}{3}}\}$.

Table 1: Power, False Positives (FP) and FDR for detecting the mass law based on the decision rule that the posterior probability of a feature is larger than $\eta^* = 0.25$. DBRM is applied using different numbers of parallel threads.

	DBRM_G_PAR			DBRM_R_PAR		
Threads	Power	FP	FDR	Power	FP	FDR
16	1.00	0.00	0.00	0.97	0.06	0.03
4	0.79	0.40	0.21	0.61	0.73	0.39
1	0.42	1.21	0.58	0.33	1.63	0.67

Table 2: Results for detecting Kepler's third law based on the decision rule that the posterior probability of a feature is larger than $\eta^* = 0.25$. The three features $\left(P\times P\times M_h^a\right)^{\frac{1}{3}}$, $(P\times P\times R_h)^{\frac{1}{3}}$ and $(P\times P\times T_h)^{\frac{1}{3}}$ are counted as true positives, all other selected features as false positives. DBRM is applied using different numbers of parallel threads.

	DBRM_G_PAR						DBRM_R_PAR					
Threads	F_1	F_2	F_3	Pow	FP	FDR	F_1	F_2	F_3	Pow	FP	FDR
64	81	71	1	1.00	0.02	0.01	78	75	2	0.99	0.03	0.01
16	34	41	32	0.84	0.46	0.18	31	38	18	0.79	0.68	0.25
1	6	5	3	0.141	0.65	0.86	6	4	2	0.12	1.81	0.88

Breast cancer

Observations from 357 benign and 212 malignant tissues. Covariates include *radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry* and *fractal dimension* of the cell nucleus. For each feature, the mean, standard error, and "worst" value (mean of the three largest values) were computed, resulting in 30 input variables in total. A quarter of the data was used as a training data set, the remaining images as a test set. We use $Y_i|\rho_i \sim \mathrm{Binom}(1,\rho_i)$, $h(\rho_i) = \mathrm{logit}(\rho_i)$, $p(\boldsymbol{\beta}^{\boldsymbol{\gamma}}) = |J_n^{\boldsymbol{\gamma}}(\boldsymbol{\beta}^{\boldsymbol{\gamma}})|^{\frac{1}{2}}$, $a = e^{-1}$, $Q = 20$, $q = 0$, and $\mathcal{G} = \{\mathrm{gauss}(x),\tanh(x),\mathrm{atan}(x),\sin(x)\}$.

Table 3: Comparison of performance (ACC, FPR, FNR) of different algorithms for breast cancer data. 100 runs were performed for each algorithm. For methods with random outcome the median measures (with minimum and maximum in parentheses) are displayed.

Algorithm	ACC	FNR	FPR
DBRM_R_PAR	0.9765 (0.9695,0.9812)	0.0479 (0.0479,0.0479)	0.0074 (0.0000,0.0184)
DBRM_G_PAR	0.9742 (0.9695,0.9812)	0.0479 (0.0479,0.0536)	0.0111 (0.0000,0.0184)
RIDGE	0.9742 (-,-)	0.0592 (-,-)	0.0037 (-,-)
LBRM	0.9718 (0.9648,0.9765)	0.0592 (0.0536,0.0702)	0.0074 (0.0000,0.0148)
DBRM_G	0.9695 (0.9554,0.9789)	0.0536 (0.0479,0.0809)	0.0148 (0.0037,0.0326)
DEEPNETS	0.9695 (0.9225,0.9789)	0.0674 (0.0305,0.1167)	0.0074 (0.0000,0.0949)
DBRM_R	0.9671 (0.9577,0.9812)	0.0536 (0.0479,0.0702)	0.0148 (0.0000,0.0361)
LR	0.9671 (-,-)	0.0479 (-,-)	0.0220 (-,-)
LASSO	0.9577 (-,-)	0.0756 (-,-)	0.0184 (-,-)
LXGBOOST	0.9554 (0.9554,0.9554)	0.0809 (0.0809,0.0809)	0.0184 (0.0184,0.0184)
TXGBOOST	0.9531 (0.9484,0.9601)	0.0647 (0.0536,0.0756)	0.0326 (0.0291,0.0361)
RFOREST	0.9343 (0.9038,0.9624)	0.0914 (0.0422,0.1675)	0.0361 (0.0000,0.1010)
NBAYES	0.9272 (-,-)	0.0305 (-,-)	0.0887 (-,-)

Conclusions

- We introduced the (R)GMJMCMC algorithm for deep Bayesian 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

Forthcoming Research

In future more research on scalability of the approach is needed. In particular developing efficient subsampling techniques are of major interest.

References

[1] A. Hubin and G. Storvik. Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA), 2016. arXiv:1611.01450v1.
[2] A. Hubin and G. Storvik. Mode jumping MCMC for Bayesian variable selection in GLMM, 2018. Accepted for publication in *Computational Statistics and Data Analysis*.
[3] A. Hubin, G. Storvik, and F. Frommlet. A novel algorithmic approach to Bayesian Logic Regression, 2018. Submitted for publication in *Statistics and Computing*.
[4] A. Hubin, G. Storvik, and F. Frommlet. Deep Bayesian regression models, 2018. Submitted for publication in *Journal of American Statistical association*.

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.*