,

BAYESIAN BINOMIAL REGRESSION MODEL WITH A LATENT GAUSSIAN FIELD FOR ANALYSIS OF EPIGENETIC DATA

A. Hubin[1] , G. Storvik[2], P. Grini[3], M. Butenko[4]
[1,2,3,4] *University of Oslo,* [1,2] *Norwegian Computing Center*
Oslo, NORWAY
e-mail: [1] aliaksandr.hubin@nr.no, [2] geirs@math.uio.no,
[3] paul.grini@ibv.uio.no, [4] m.a.butenko@ibv.uio.no

Abstract

Epigenetic observations are represented by the total amount of reads from a particular cell and the amount of methylated reads, making it reasonable to model this data by a binomial distribution. There are numerous factors that can influence probability of success from a particular region. We might also expect spatial dependence of these probabilities. We incorporate dependence on the covariates and spatial dependence of methylation probability for observation from a particular cell by means of a binomial regression model with a latent Gaussian field. We run Mode Jumping Markov Chain Monte Carlo algorithm (MJMCMC) across different choices of covariates in order to obtain the joint posterior distribution of parameters and models. This also allows to find the best set of covariates to model methylation probability within the genomic region of interest.

Keywords: Binomial regression, Gaussian field, epigenetic data, data science

1 Introduction

Natural epigenetic variation provides a source for the generation of phenotypic diversity, but to understand its contributions to such diversity and its interaction with genetic variation requires further investigation [4]. Epigenetic changes are crucial for the development and differentiation of various cell types in an organism, as well as for normal cellular processes. High-throughput epigenetics experiments have enabled researchers to measure genome-wide epigenetic profiles. Epigenome-wide association studies (EWAS) hold promise for the detection of new regulatory mechanisms that may be susceptible to modification by environmental and lifestyle factors [3]. At the same time, epigenetic data are often spatially correlated with high noise levels, which requires careful spatial-temporal statistical modeling.

A major task today is the development of models and statistical methods for linking epigenetic patterns to genetic and/or environmental variables and interpreting them. Due to the availability of data, our focus will be on the plant *Arabidopsis.* [1] previously analysed Arabidopsis data consisting of epigenetic observations on a set of 10 lines, which were separately propagated in a common environment for 30 generations. These were compared with two independent lines propagated for only three generations (because of missing ancestor). Their analysis aimed at global summaries of structures

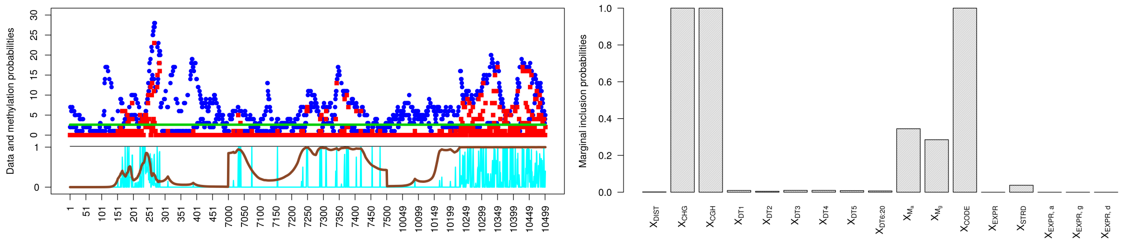

Figure 1: **Left** graph depicts epigenetic observations, where blue dots are total number of reads, red dots - number of methylated reads, green line corresponds to 2 total reads distinguishing the inference and the identification data, light blue line gives naïve probabilities as rates, brown line - probabilities as the posterior mean of the probability of success parameter from the posterior mode model. **Right** graph depicts barplots of RM estimates [2] of marginal inclusion probabilities of the covariates.

but was based on individual and (site-wise) hypothesis testing methods combined with FDR control methodology.

In this paper we limit ourselves to finding a pattern of signals appearing along genome that significantly influences methylation probability. We additionally take into account spatial dependence between the observations as well as the unexplained by the exogenous variables variability of the epigenetic observations. This is done by means of applying the MJMCMC algorithm developed by [2] to the Bayesian binomial regression with a random walk of order one, denoted as $RW(1)$, and independent Gaussian, denoted as IG, latent processes.

2 Mathematical model

We model the number of methylated reads $Y_t \in \{1, ..., n_t\}$ per position in the genome (nucleobase) to be binomially distributed with the number of trials equal to the number of reads for this position $n_t \in \mathbb{N}$ and probability of success $p_t \in \mathbb{R}_{[0,1]}$ modeled via logit link to the covariates $X_t = \{X_{t1}, ..., X_{tM}\}, t \in \{t_1, ..., t_T\}$, where T is the total number of genomic positions in the addressed genomic region. These covariates might be a position within a gene, indicator of the underlying genetic structure, and others (our choice of the covariates is given in Section 3). A latent Gaussian $RW(1)$ process $\delta_t \in \mathbb{R}$ is included into the model in order to take into account spatial dependence of methylation probabilities along the genome, whilst a latent independent Gaussian process (IG) ζ_t is used to model the variance of the observations, which is not explained by the covariates. This gives the following model formulation:

$$\Pr(Y_t = y|n_t, p_t) = \binom{n_t}{y} p_t^y (1-p_t)^{n_t - y}, \tag{1}$$

$$p_t = \frac{e^{\beta_0 + \sum_{i=1}^{M} \gamma_i \beta_i X_{ti} + \delta_t + \zeta_t}}{1 + e^{\beta_0 + \sum_{i=1}^{M} \gamma_i \beta_i X_{ti} + \delta_t + \zeta_t}}, \tag{2}$$

$$\delta_t = \delta_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2), \tag{3}$$

$$\zeta_t \sim N(0, \sigma_\zeta^2), \tag{4}$$

where $\beta_i \in \mathbb{R}, i \in \{0, ..., M\}$ are regression coefficients of the covariates of the model showing whether and in which way the corresponding covariate influences the probability of methylation on average, $\gamma_i \in \{0, 1\}, i \in \{1, ..., M\}$ are latent indicators, defining if covariate i is included into the model($\gamma_i = 1$) or not ($\gamma_i = 0$), ϵ_t are the error terms of $RW(1)$ process δ_t, which are normally distributed with zero mean and variance ${\sigma_\epsilon}^2$. Finally, σ_ζ^2 is the variance term of the IG process ζ_t. We then put the following priors for the parameters of the model:

$$\gamma_i \sim \text{Bernoulli}(q), \beta_i|\gamma_i \sim \mathbb{1}(\gamma_i = 1)N(\mu_\beta, \sigma_\beta^2), \psi_j \sim \log\Gamma(1, 5 \cdot 10^{-5}), j \in \{1, 2\}, \tag{5}$$

where the log Gamma distributed $\psi_1 = \log \frac{1}{\sigma_{\epsilon,t}^2}$ and $\psi_2 = \log \frac{1}{\sigma_{\zeta,t}^2}$ are the scaled hyperparameters of the latent models, $q = 0.5$ is the prior Bernoulli probability of including a covariate into the model. We perform analysis for the model defined by Equations (1)-(5) by means of the MJMCMC algorithm [2]. The algorithm is capable of efficiently moving in the defined model space by means of both accurately exploring the modes of the probability mass and switching between these modes using large jumps combined with local optimization and randomization [2].

3 Data description

The addressed data set consists of 1502 observations from the first chromosome of Arabadopsis plant belonging to five predefined groups of genes. This data set was divided into 950 observations (with more than 2 reads, see Figure 1) for inference and 552 observations (with less than 3 reads) for model based identification of methylation probabilities for the positions with the lack of data.

Apart from the observations represented by the methylated versus total amount of reads we have data on various exogenous variables (covariates). Among these covariates we address the factor with 3 levels corresponding to whether the location belongs to CGH, CHH or CHG genetic region, where H is either A, C or T and thus generating two covariates X_{CGH} and X_{CHH}. The second group of factors indicates whether the distance to the previous cytosine nucleobase (C) in DNA is 1, 2, 3, 4, 5, from 6 to 20 or greater than 20 inducing six binary covariates $X_{DT1}, X_{DT2}, X_{DT3}, X_{DT4}, X_{DT5}$, and $X_{DT6:20}$. We also include such 1D distance as a continuous covariate X_{DIST}. The third addressed group of factors corresponds to whether the location belongs to a gene from a particular group of genes of biological interest. These groups are indicated as

M_a, M_g and M_d, yielding two additional covariates X_{M_a}, X_{M_g}. Additionally we have a covariate X_{CODE} indicating if the corresponding nucleobase is in the coding region of a gene and a covariate X_{STRD} indicating if the nucleobase is on a "+" or a "-" strand. Finally, we have a continuous covariate $X_{EXPR} \in \mathbb{R}^+$ representing expression level for the corresponding gene and interactions between expression levels and gene groups $X_{EXPR,a}, X_{EXPR,g}, X_{EXPR,d} \in \mathbb{R}^+$. Thus multiple predictors with respect to a strict choice of the reference model in our example induced $M = 17$ potentially important covariates.

4 Results and discussion

MJMCMC algorithm was run until around 10000 unique models (7.6% of the model space) were explored. We parallelized the search on 10 CPUs. Default frequencies of large jumps and corresponding local optimizers from [2] were used. Also the default radiuses of proposals of global moves and local moves were addressed.

According to the marginal inclusion probabilities reported in the right graph of Figure 1, only factors X_{CHG}, X_{CGH} and X_{CODE} are clearly significant for inference on the methylation patterns for the addressed epigenetic region, factors X_{M_a} and X_{M_g} also have some significance. In Table 1 one can find marginal posterior model probability and posterior means of the parameters for the best model in the explored subset of models from the model space. Based on the best model we carried out computations of methylation probabilities of the locations in both the inference set and the identification set. Furthermore, we compared the results with the naïve approach based on computing the proportion of methylated reads, which is currently addressed in the biological literature as a standard way to evaluate methylation probability of a given nucleobase. These results are summarized in the left graph of Figure 1. The results show that the naïve approach should not be trusted in the presence of spatially correlated data and the corresponding to it probabilities are strongly biased.

Table 1: Posterior means for the best model in terms of marginal posterior probability (PMP)

PMP	β_0	β_{CHG}	β_{CGH}	β_{CODE}	σ^2_ϵ	σ^2_ζ
0.4276	-8.8255	2.4717	5.2122	6.4240	0.1332	0.8258

In future it would be of interest to obtain additional covariates such as whether the corresponding nucleobase belongs to a particular part of the non-coding gene region like promoter, intron or tranposone, and whether the nucleobase is within a CpG island.

References

[1] Becker C. [et al.] (2011). Spontaneous epigenetic variation in the arabidopsis thaliana methylome. *Nature*. Vol. 480, No. 7376, pp. 245–249.

[2] Hubin A., Storvik G. (2018). Mode jumping MCMC for Bayesian variable selection in GLMM. *Computational Statistics and Data Analysis*. Vol. 127, pp. 281–297.

[3] Michels K. B. [et al.] (2013). Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*. Vol. 10, No. 10, pp. 949–955.

[4] Schmitz R. J. [et al.] (2013). Patterns of population epigenomic diversity. *Nature*. Vol. 495, No. 7440, pp. 193–198.