Compressing Parameters in Bayesian High-order Models

Longhai Li

Joint Work with Radford Neal

longhai@math.usask.ca
http://math.usask.ca/~longhai

Department of Mathematics and Statistics

University of Saskatchewan

Saskatoon, Saskatchewan, Canada

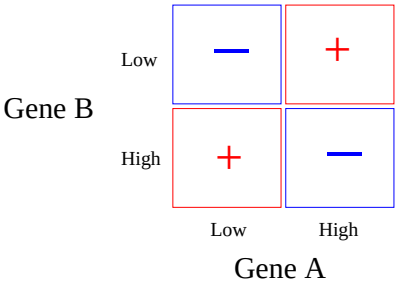
JOINT MEETING OF STATISTICAL SOCIETY of CANADA AND THE SOCIÉTÉ FRANCAISE DE STATISTIQUE

Ottawa, Canada, 29 May 2008

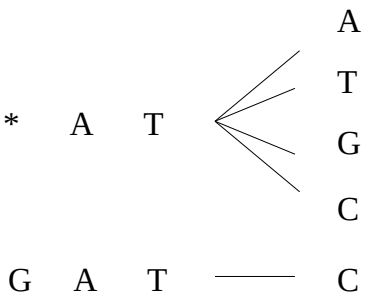
Problem description

Motivating Examples:

1) Many human complex traits may be related to interactions of multilocus genes and environmental exposures. As reported in the literature, the examples of such traits include *breast cancer, post-PTCA stenosis, essential hypertension, atrial fibrillation and type 2 diabetes*.



2) It is believed that there exists long-range dependency among nucleotides in "non-coding" region of human genome. Considering this dependency in modeling nucleotide sequences will improve many statistical applications in genome, such as haplotype inference and discovery of transcription-factor binding sites.



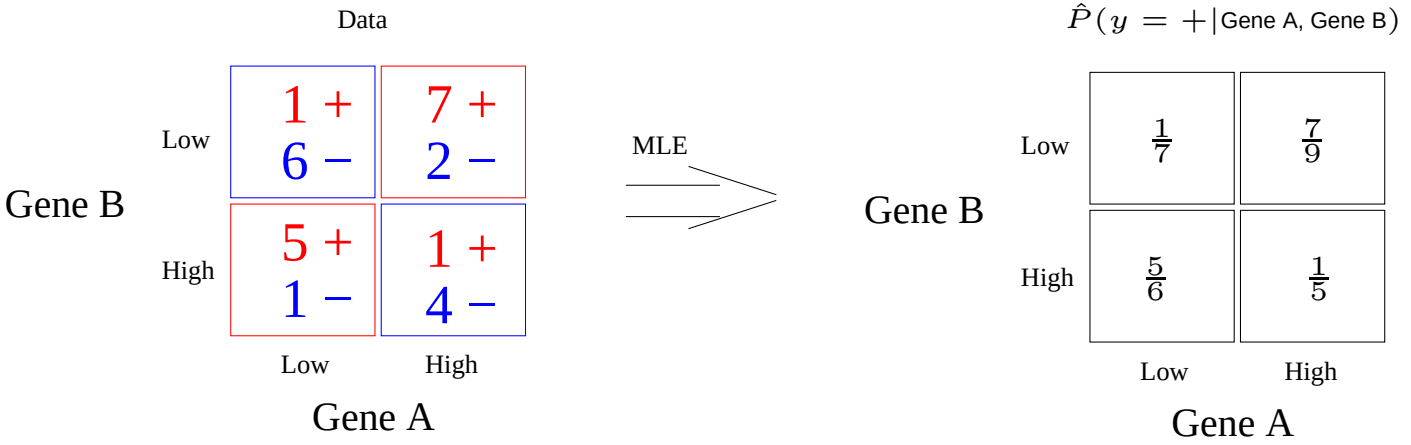
Statistical problem:

For discrete variables $y, x_1, \ldots, x_p$, we want to model the predictive probability:

$$P(y|x_1, \ldots, x_p)$$

Difficulties with a naive method

A naive method for considering interactions: estimate the probability of y for each combination of $x_1, \ldots, x_p$:



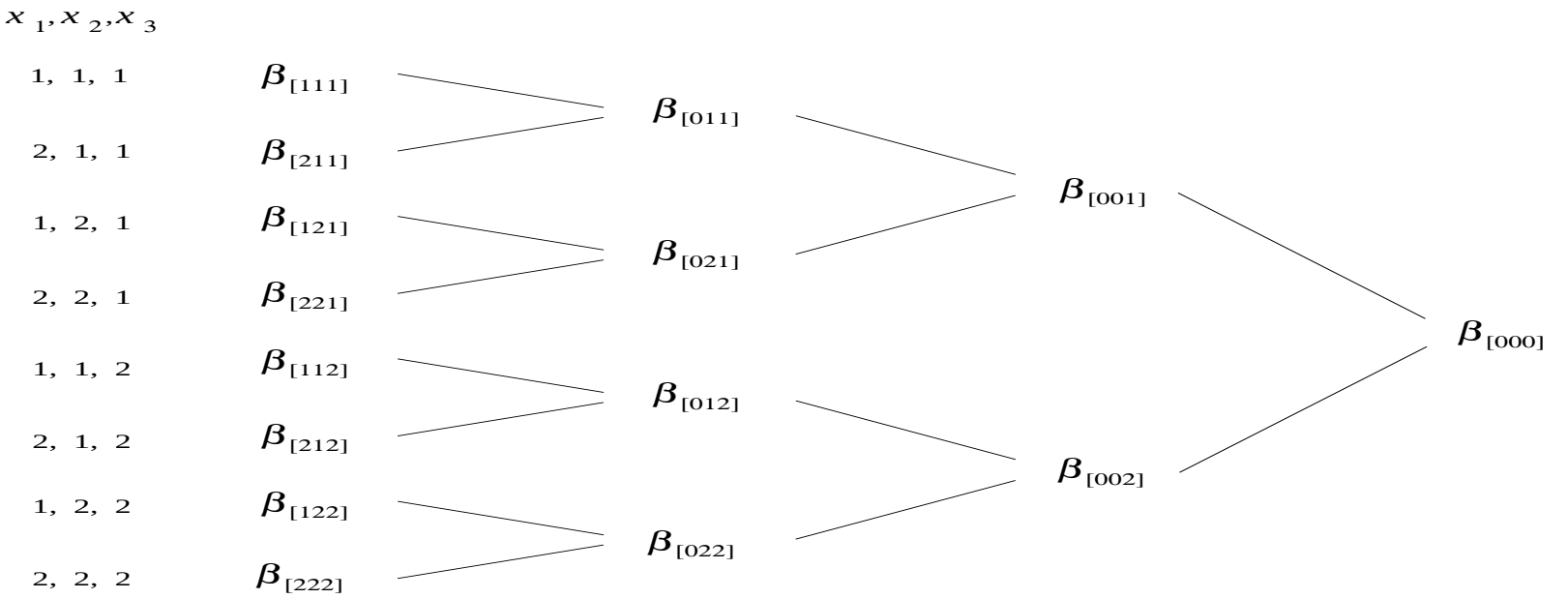
Difficulties: The number of combinations increases exponentially as p. When p is large, there are very few observations in each cell. The estimates of the probabilities are therefore very inaccurate; Considering models of lower order makes the estimates more accurate, but at the risk of omitting useful high-order interactions; Useful interaction patterns may have different orders.

Regression models using interaction patterns

A toy example with only 3 cases and 3 binary (1/2) features

				Indicators on Interaction Patterns (0=any)									
x_1	x_2	x_3	x_4	$I_{[000]}$	$I_{[001]}$	$I_{[002]}$	$I_{[011]}$	$I_{[012]}$	$I_{[021]}$	$I_{[211]}$	$I_{[212]}$	$I_{[121]}$	...
2	1	1	?	1	1	0	1	0	0	1	0	0	0
2	1	2	?	1	0	1	0	1	0	0	1	0	0
1	2	1	?	1	1	0	0	0	1	0	0	1	0

Graphical representation of parameters:



Bayesian logistic sequence model

We want to model $P(x_{O+1} \mid x_1, \ldots, x_O)$, where $x_1, \ldots, x_O, x_{O+1}$ is a discrete sequence. We use a linear logistic model:

$$P(x_{O+1} = k \mid \boldsymbol{x}_{1:O}, \boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(K)}) = \frac{\exp(l\,(\boldsymbol{x}_{1:O}, \boldsymbol{\beta}^{(k)}))}{\sum_{j=1}^{K} \exp(l\,(\boldsymbol{x}_{1:O}, \boldsymbol{\beta}^{(j)}))}$$

where

$$l\,(\boldsymbol{x}_{1:O}, \boldsymbol{\beta}^{(k)}) = \sum_{\mathcal{P} \in \boldsymbol{\mathcal{S}}} \beta_{\mathcal{P}}^{(k)}\; I(\boldsymbol{x}_{1:O} \in \mathcal{P}) = \beta_{[0\cdots0]}^{(k)} + \sum_{t=1}^{O} \beta_{[0\cdots x_t \cdots x_O]}^{(k)}$$

where $\boldsymbol{\mathcal{S}}$ is the set of all patterns of O or fewer of the preceding O symbols.

We use the following priors, where $o(\mathcal{P})$ is the order of pattern $\mathcal{P}$:

$$\begin{aligned} \log(\sigma_t) &\sim \text{Normal}(\mu_t, w_t), \text{ for } t = 0, \ldots, O \\ \beta_{\mathcal{P}}^{(k)} \mid \sigma_{o(\mathcal{P})} &\sim N(0, \sigma_{o(\mathcal{P})}^2) \text{ or Cauchy}(0, \sigma_{o(\mathcal{P})}), \text{ for } \mathcal{P} \in \boldsymbol{\mathcal{S}} \end{aligned}$$

Remarks on logistic sequence model

- By expressing $l((x_1, x_2, x_3), \boldsymbol{\beta})$ as the sum of parameters for interactions from low order to high order, we actually add a prior information that the predictive probabilities $P(x_4|x_1, x_2, x_3)$ are closer for similar x_1, x_2, x_3.

- We are not forced to use a short sequence for avoiding overfitting. Useful high-order interactions can be discovered if some do exist. The model will automatically adjust the complexity of the relationship.

- When order O is large, the number of parameters is huge in this model. **But, we notice that many predictor variables have the same value for all cases in data. We will use this fact to compress parameters.**

Compressing parameters

When groups of predictor variables have the same value for all training cases, the likelihood function of a linear regression model depends only on the sums over groups:

$$\begin{aligned} L^{\beta}(\beta_{11},\ldots,\beta_{1,n_1},\ \ldots\ ,\beta_{G1},\ldots,\beta_{G,n_G}) &= L\left(\sum_{k=1}^{n_1}\beta_{1k},\ \ldots,\ \sum_{k=1}^{n_G}\beta_{Gk}\right) \\ &= L(s_1,\ \ldots\ ,s_G) \end{aligned}$$

Since we use priors as $\beta_{gk} \sim N(0,\sigma^2_{gk})$ or $\beta_{gk} \sim$ Cauchy$(0,\sigma_{gk})$, the priors of the s_g's can be found easily:

$$s_g \sim N\left(0,\ \sum_{k=1}^{n_g}\sigma^2_{gk}\right) \quad \text{or} \quad s_g \sim \text{Cauchy}\left(0,\ \sum_{k=1}^{n_g}\sigma_{gk}\right)$$

The posterior of the s_g's given the training data $\mathcal{D}$:

$$P(\boldsymbol{s} \mid \mathcal{D}) = \frac{1}{c(\mathcal{D})}\, L(s_1,\ \ldots\ ,s_G)\ P_1^s(s_1)\ \cdots\ P_G^s(s_G)$$

where P_g^s is the prior density function of the compressed parameter s_g.

Splitting compressed parameters

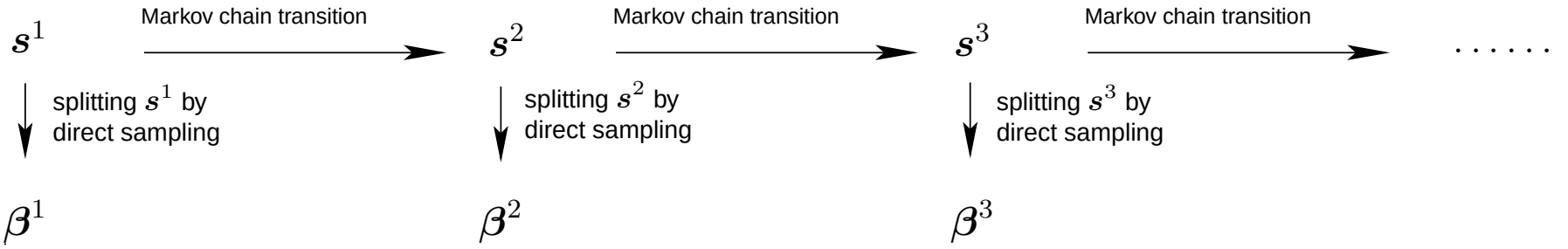
After obtaining the samples of s_g's using MCMC, we recover the original parameters, using the splitting distribution:

$$P(\beta_{g1}, \ldots, \beta_{g,n_g-1} \mid s_g) = \frac{\left(\prod_{k=1}^{n_g-1} P_{gk}(\beta_{gk})\right) P_{g,n_g}\left(s_g - \sum_{k=1}^{n_g-1} \beta_{gk}\right)}{P_g^s(s_g)}$$

where P_{gk} is the prior density function of the original parameter β_{gk}.

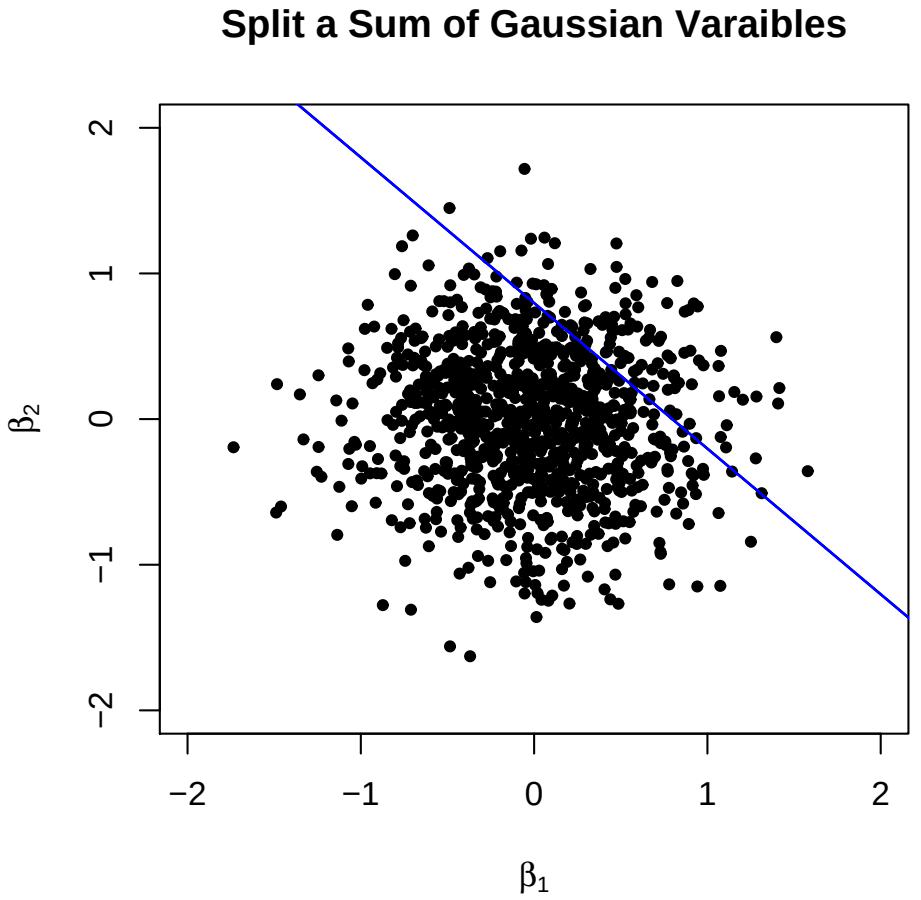
The splitting distribution is unrelated to $\mathcal{D}$. We can directly sample from it.

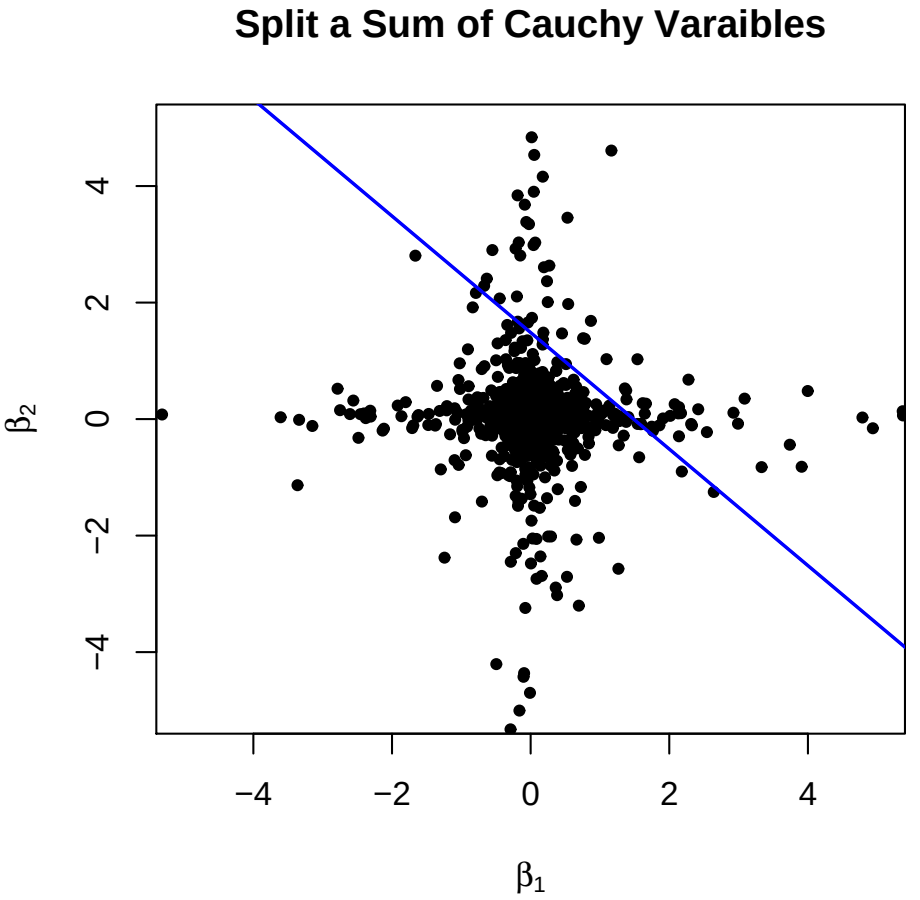
The sampling procedure can be depicted as follows:



Splitting s_g into two parts

The following graph demonstrates the splitting distribution when sampling for independent β_1 and β_2 constrained to lie on the blue line.





An English text data

An online article, which introduces the Department of Statistics at the University of Toronto, is encoded:

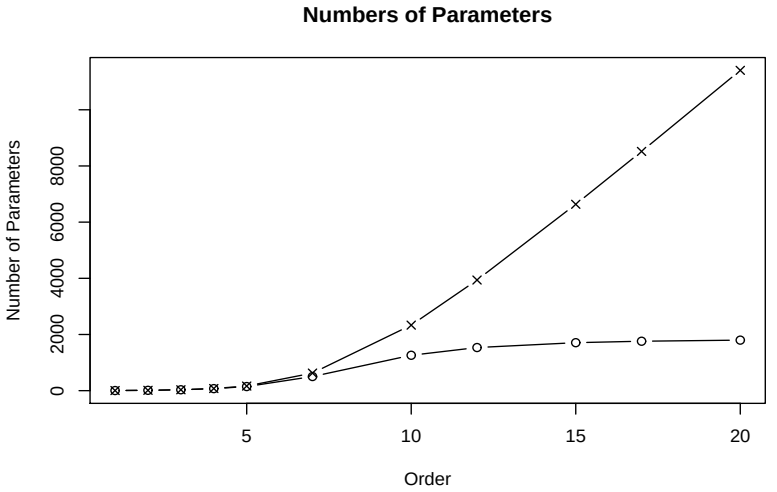
1 = vowel letters, 2 = consonant letters, 3 = all other characters

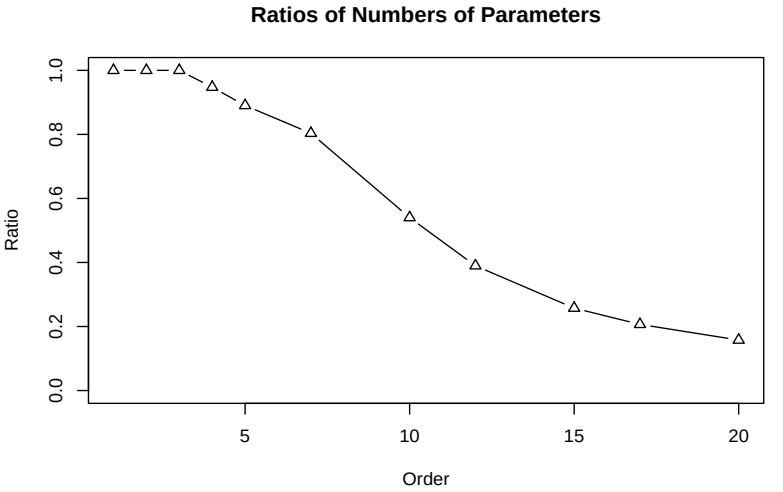
There are a total of 3930 characters, giving 3910 overlapped sequences of length 21.

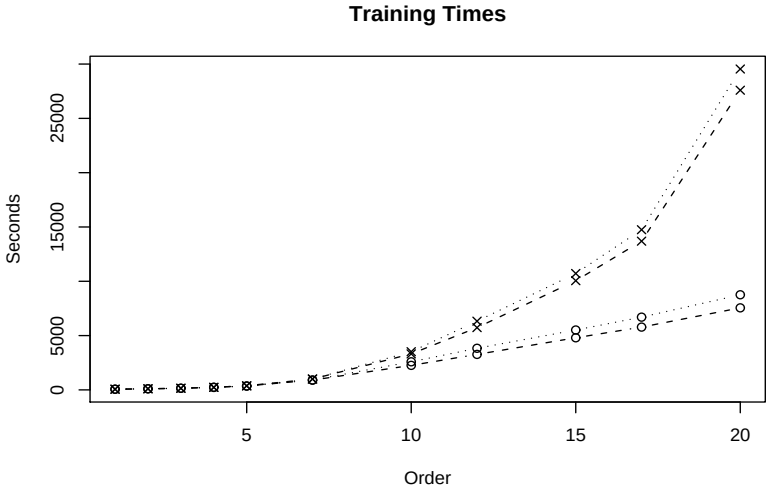
We tested our method by predicting the 21st character based on varying numbers of preceding characters.The first 1000 sequences were used as training cases. The remaining 2910 were used as test cases.

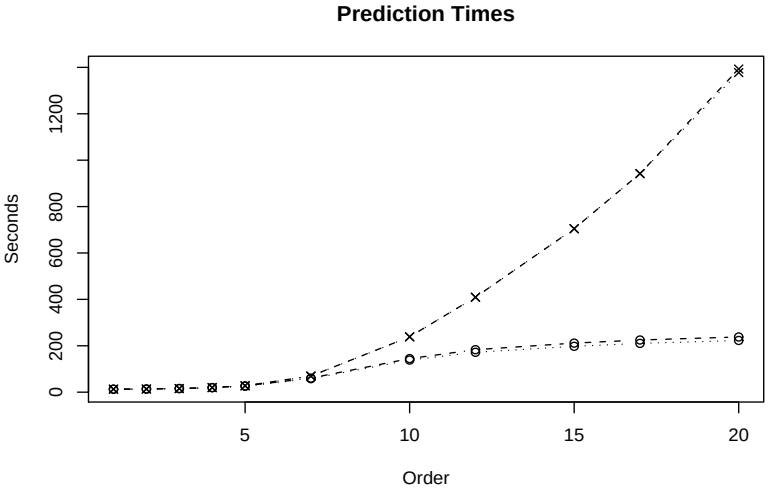
Parameter reduction

The following graph displays the reduction of the number of parameters and training time by MCMC.



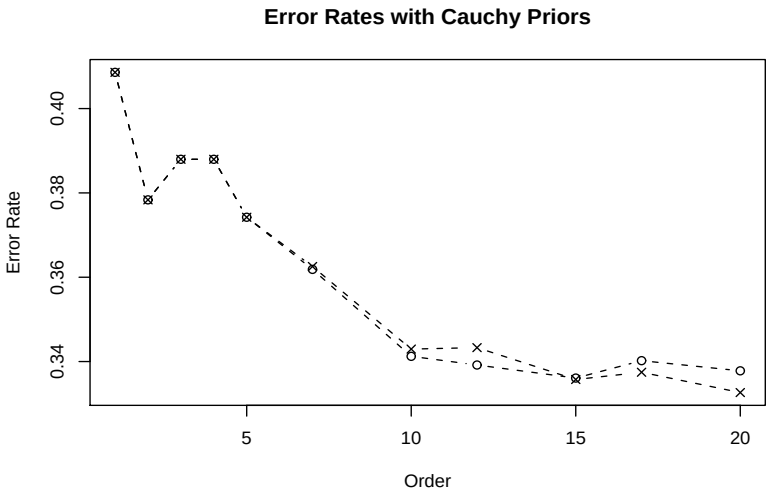


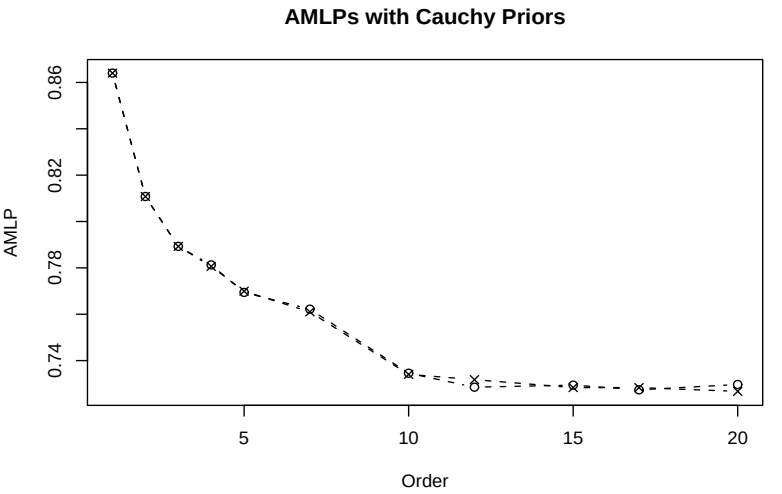


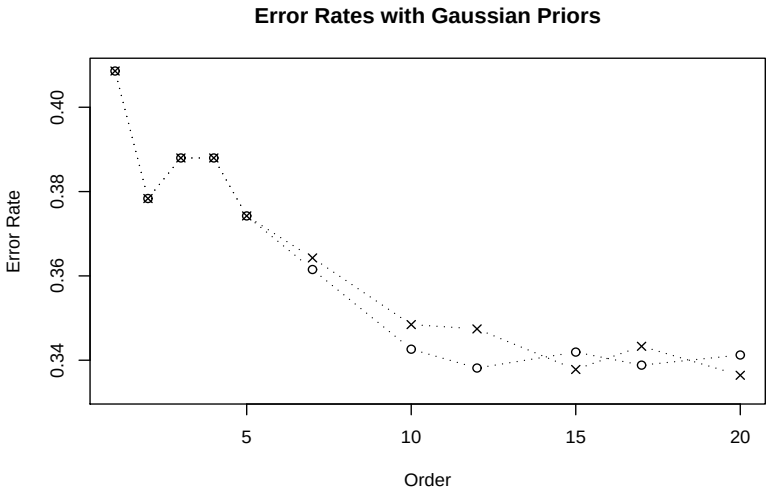


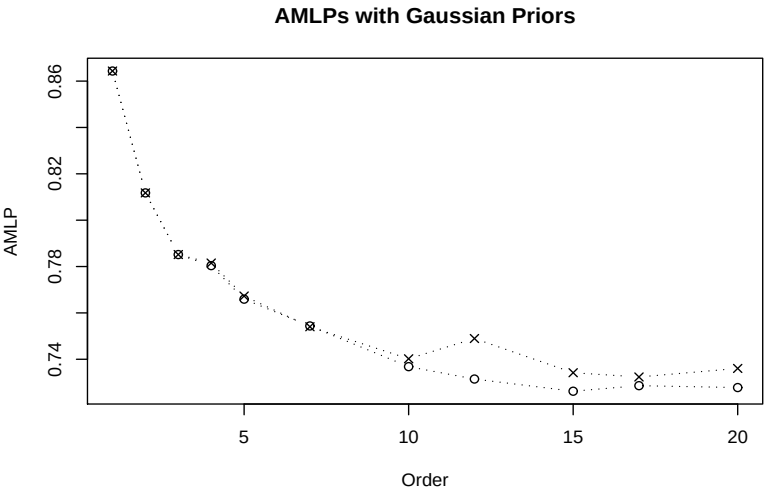
Prediction performance

The following graph displays the prediction performance on test set measured by error rate and average minus log probability.









Concluding remarks

- We propose a method to reduce the number of parameters in Bayesian high-order models, with application to logistic sequence models.

- It is unnecessary to restrict the model complexity in Bayesian high-order models for statistical reason. With our compression method, restricting the model complexity for computational reason is also unnecessary.