Compressing Parameters In Bayesian High-order Models

Longhai Li* and Radford M. Neal†

*Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Canada. Email: longhai@math.usask.ca
†Department of Statistics and Department of Computer Science, University of Toronto, Toronto, Canada. Email: radford@utstat.toronto.edu

Abstract

Bayesian classification and regression with high-order interactions is largely infeasible because Markov chain Monte Carlo (MCMC) would need to be applied with a great many parameters, whose number increases rapidly with the order considered. We show how to make it feasible by effectively reducing the number of parameters, exploiting the fact that many interactions have the same values for all training cases. Our method uses a single "compressed" parameter to represent the sum of all parameters associated with a set of patterns that have the same value for all training cases. We apply this method to logistic sequence models and demonstrated it with an English text data set.

Predictor variables derived from interaction patterns

Below is a toy example with only 3 training cases and 2 binary (1/2) features, illustrating the indicators whether interaction patterns occur.

Discrete Measurements

i	x_1	x_2
1	1	2
2	2	1
3	1	1

$\Longrightarrow$

Indicators on Interaction Patterns (0=any)

i	$I_{[00]}$	$I_{[10]}$	$I_{[20]}$	$I_{[01]}$	$I_{[02]}$	$I_{[11]}$	$I_{[21]}$	$I_{[12]}$	$I_{[22]}$
1	1	1	0	0	1	0	0	1	0
2	1	0	1	1	0	0	1	0	0
3	1	1	0	1	0	1	0	0	0

Facts:

- The number of predictor variables increases exponentially with the order considered. This brings many difficulties:
  - intensive computation, especially for Bayesian methods implemented by Markov chain Monte Carlo (MCMC)
  - overfitting the data with maximum likelihood method
- Many predictor variables have the same value for all training cases, for example, the predictor variables enclosed by boxes of the same colour in the above example. Namely, these interactions are *expressed* by the same training cases. Particularly, when an interaction pattern of certain order is expressed by only 1 case, all interactions of higher order expressed by this case have the same value. We exploit this fact to reduce the number of 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:

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

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

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

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

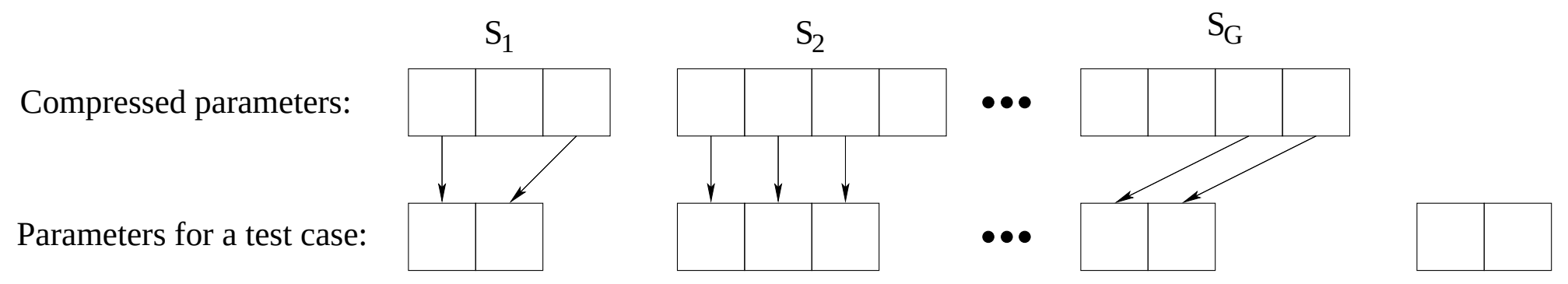
After obtaining the samples of s_g's using MCMC, we recover the original parameters, using the splitting distribution:

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

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.

To save space, we can split s_g temporarily for each test case.



Need only to split s_g into two parts for a particular test case:

$$P(s_g^t \mid s_g) = P_g^{(1)}(s_g^t)\ P_g^{(2)}(s_g - s_g^t)\ /\ P_g^s(s_g)$$

Splitting $\boldsymbol{s}_g$ into two parts

- Split a sum of Gaussian variables:

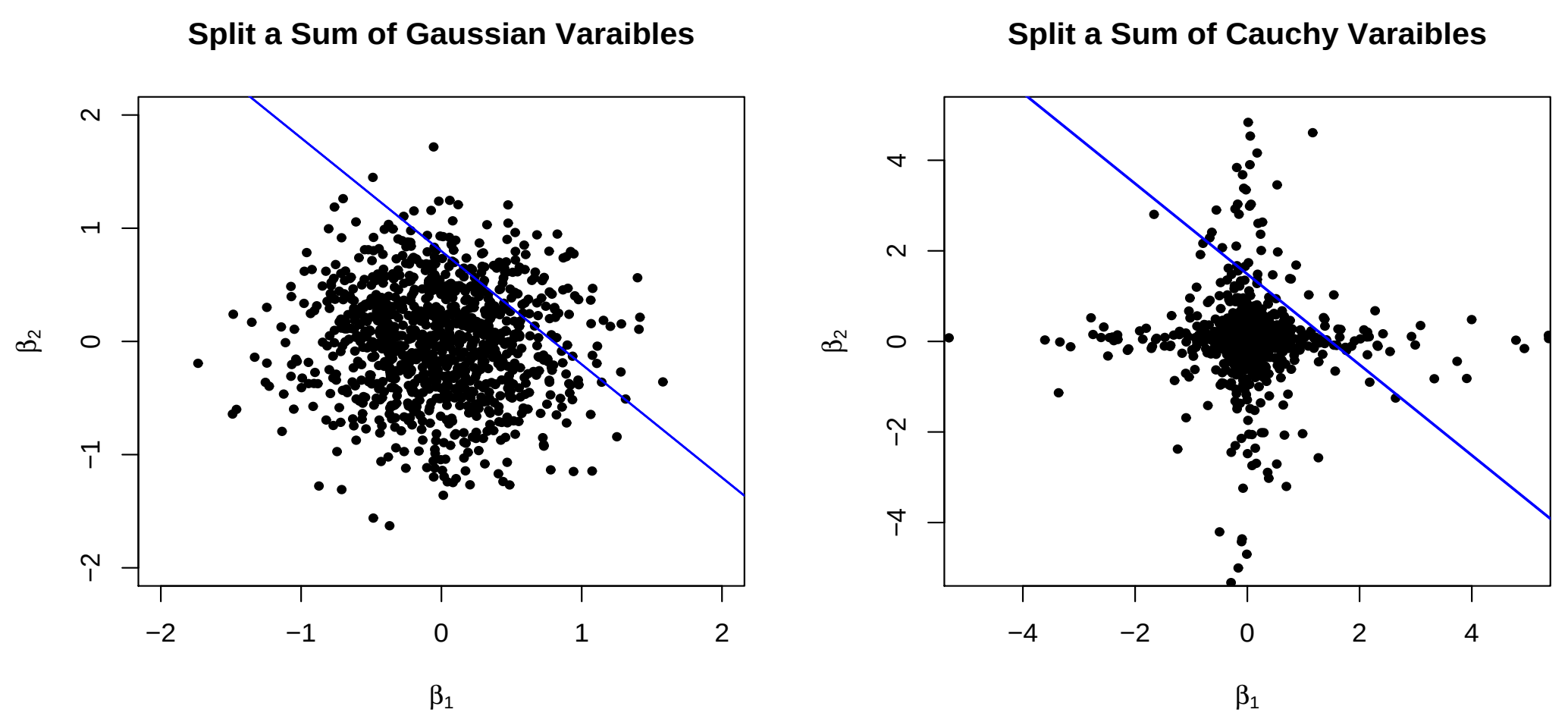
$$s_g^t \mid s_g \ \sim\ N\left(s_g\,\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\,,\ \sigma_1^2\left(1\ -\ \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)\right)$$

- Split a sum of Cauchy variables:

$$\begin{aligned} F(s_g^t\,;\,s_g,\sigma_1,\sigma_2) = \frac{1}{C}\Bigg[& r\log\left(\frac{(s_g^t)^2+\sigma_1^2}{(s_g^t-s_g)^2+\sigma_2^2}\right) + \\ & p_0\left(\arctan\left(\frac{s_g^t}{\sigma_1}\right)+\frac{\pi}{2}\right) + \\ & p_s\left(\arctan\left(\frac{s_g^t-s_g}{\sigma_2}\right)+\frac{\pi}{2}\right)\Bigg] \end{aligned}$$

Being able to compute the CDF, we can use inversion method to sample from the above distribution, with the inverse CDF found numerically.

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



Logistic sequence models

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

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

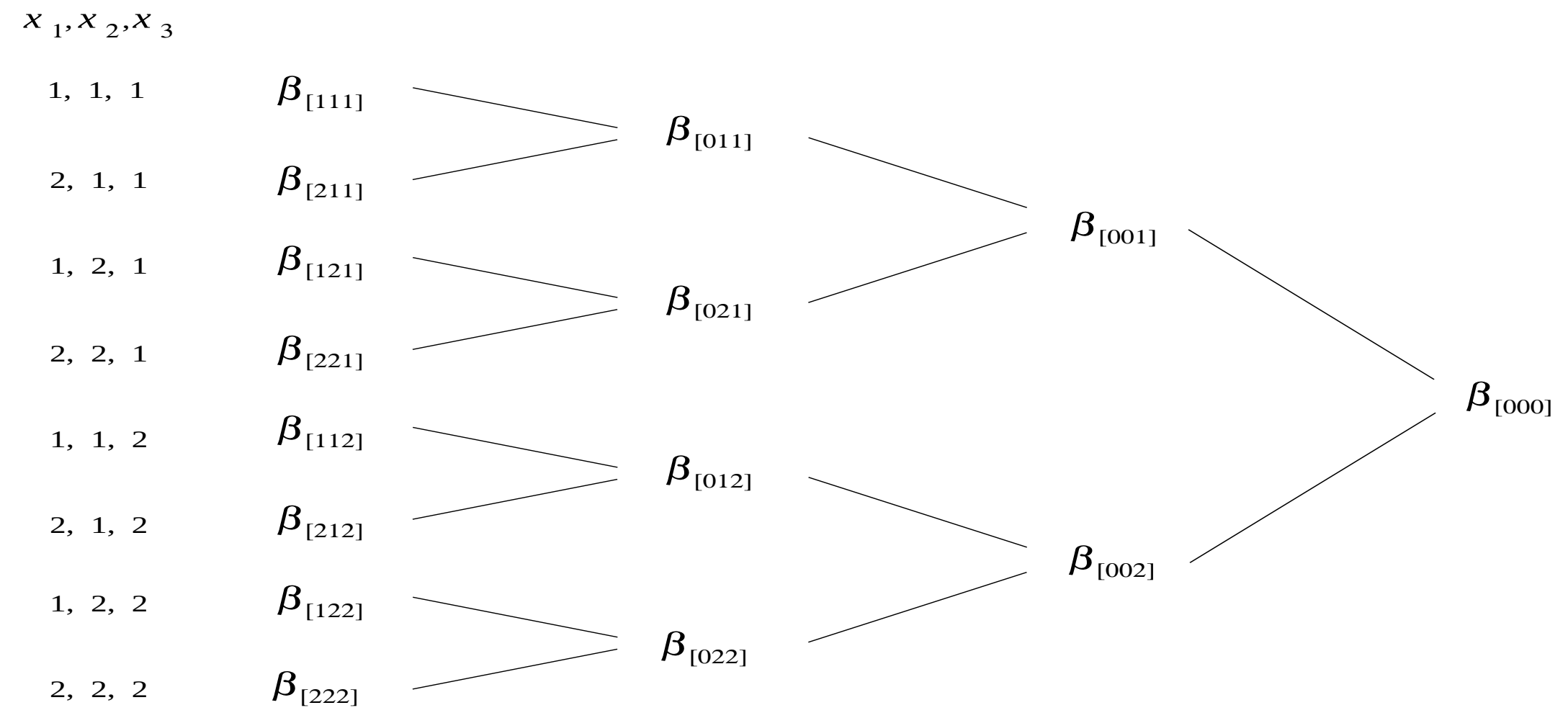
where $\boldsymbol{\mathcal{S}}$ is the set of all patterns of O or fewer of the preceding O symbols.

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

$$\begin{aligned} \sigma_t &\sim \text{Inverse-Gamma}(\alpha_t\,,(\alpha_t+1)\,w_t),\ \text{for } t=0,\ldots,O \\ \beta_{\mathcal{P}}^{(k)} \mid \sigma_{o(\mathcal{P})} &\sim N(0,\sigma_{o(\mathcal{P})}^2) \text{ or } \text{Cauchy}(0,\sigma_{o(\mathcal{P})}),\ \text{for } \mathcal{P}\in\boldsymbol{\mathcal{S}} \end{aligned}$$

Picture of a logistic sequence model

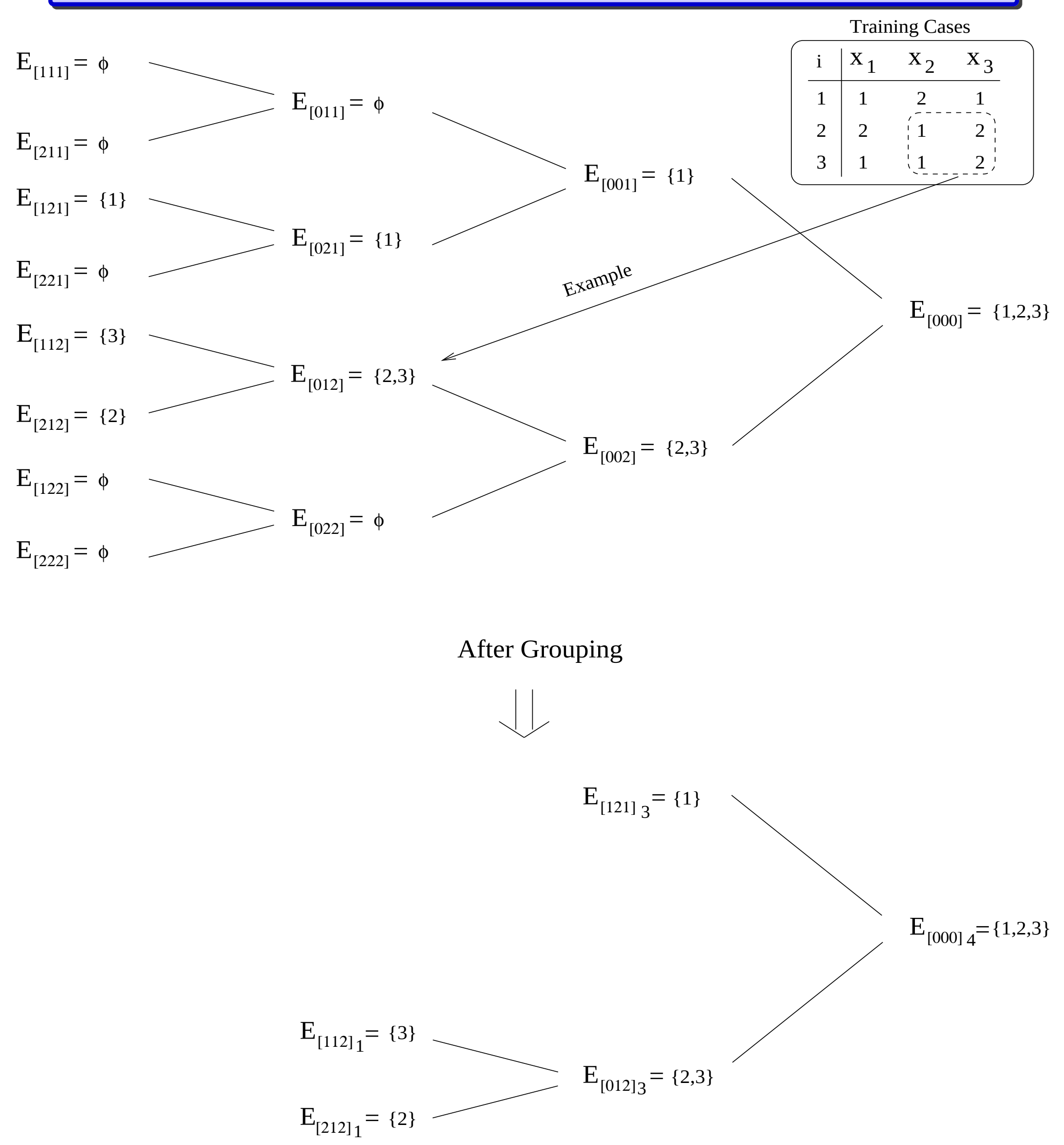
The picture below displays the regression coefficients of a binary sequence model with 3 preceding states. The linear function $l((x_1,x_2,x_3),\boldsymbol{\beta})$ is equal to the sum of β linked by straight lines.



Remarks on logistic sequence models:

- By expressing $l((x_1,x_2,x_3),\boldsymbol{\beta})$ as the sum of parameters for interactions from low order to high order, the predictive distributions given similar preceding sequences are similar. This is a natural prior belief, and helps avoid overfitting.
- 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.

Grouping parameters in logistic sequence models



Training Cases

i	x1	x2	x3
1	1	2	1
2	2	1	2
3	1	1	2

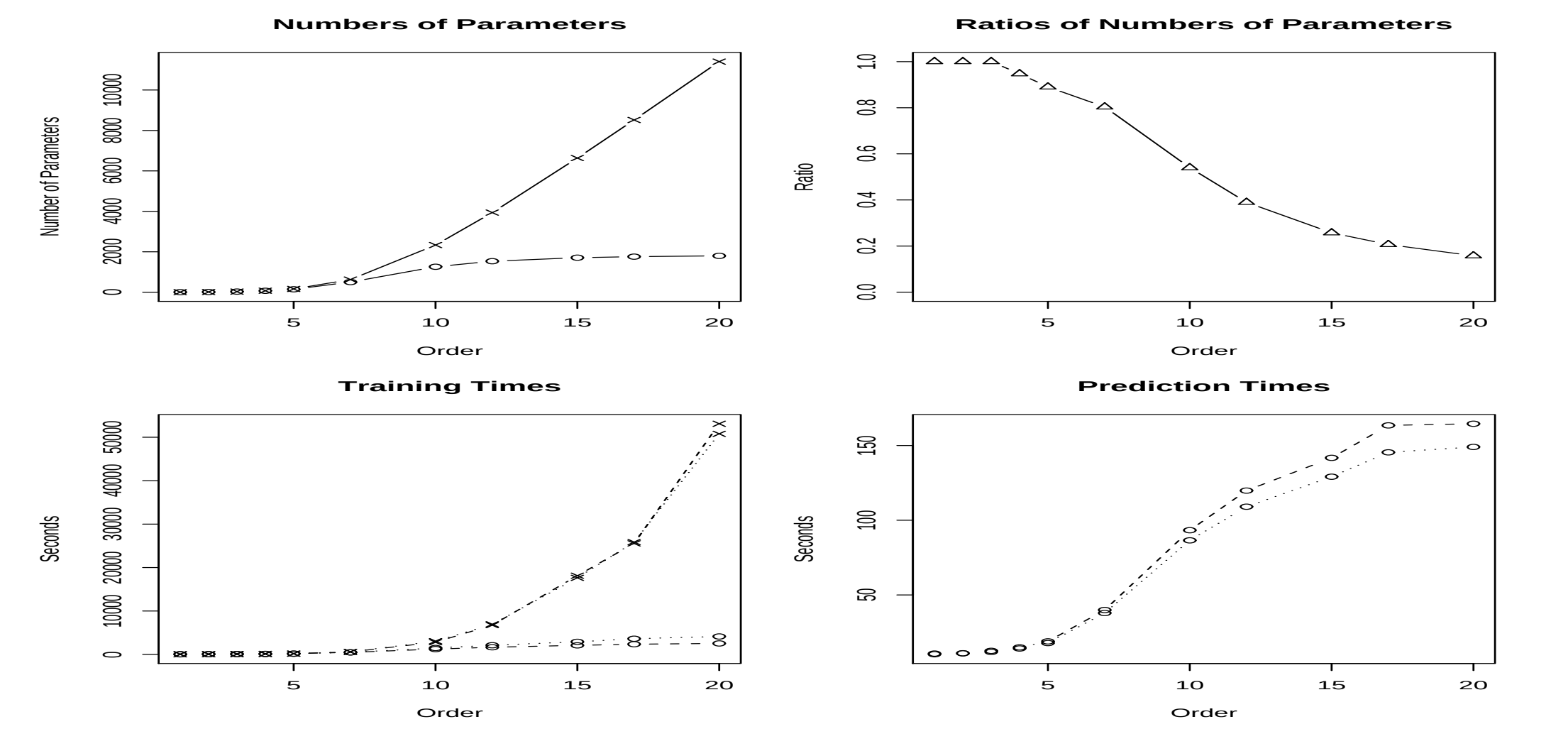
Experiments on English text

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

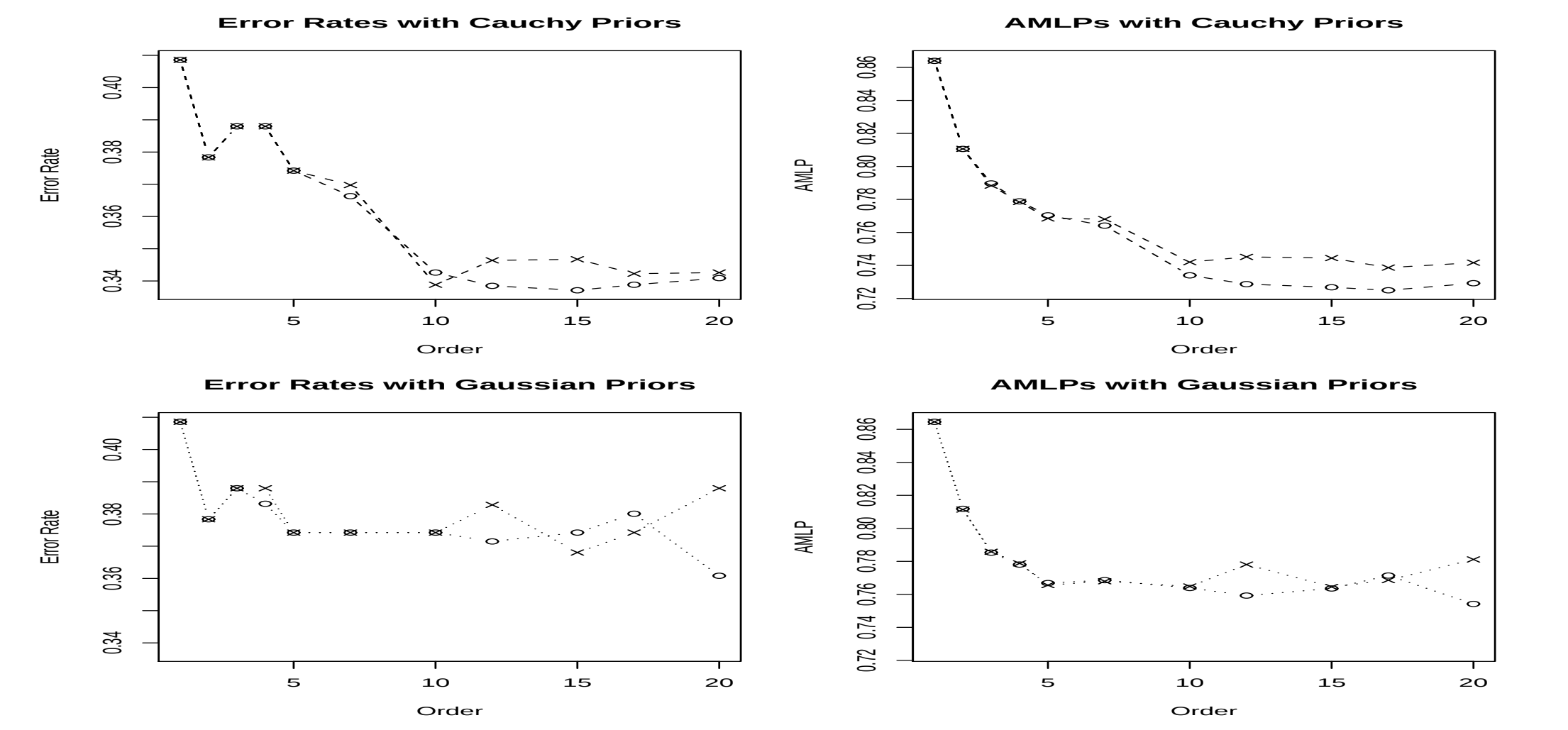
There are a total of 3930 characters, giving 3910 overlapped sequences of length 21. We tested our method by predicting the 21st character based on varying numbers of preceding characters.The first 1000 sequences were used as training cases. The remaining 2910 were used as test cases.

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

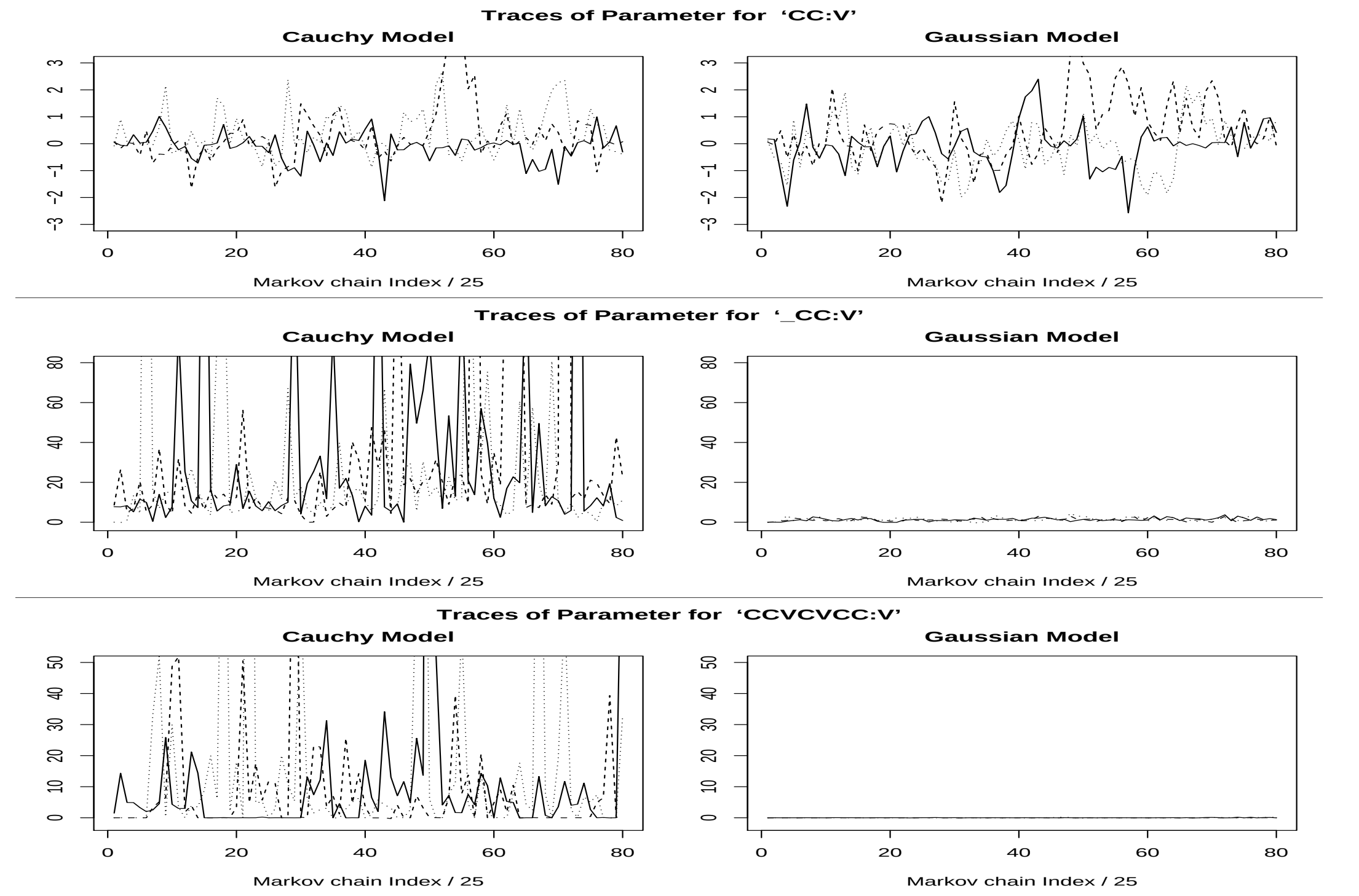


○ = Parameters Compressed, × = Original Parameters
– – – = Using Cauchy Priors, ⋯ = Using Gaussian Priors

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



We show the Markov chain traces (3 independent runs) for some particular β, for example, '_CC:V', the parameter for predicting that a vowel follows "others","consonant","consonant". The posterior of supposedly small β (e.g. 'CC:V') concentrates more around 0 in the Cauchy model than in the Gaussian model, but the posterior of supposedly large β (e.g. '_CC:V') favors much larger value in the Cauchy model than in the Gaussian model.



Conclusions and discussions

- We propose a method to reduce the number of parameters in Bayesian high-order models, with application to logistic sequence models. This method can also be applied to general logistic classification models (see Li, 2007).
- We demonstrate empirically that Cauchy distributions could be better than Gaussian distributions as the priors for the regression coefficients of high-order models for some problems.
- This method could be applied to many diverse problems, such as data compression, speech recognition and bioinformatics.

Reference: Li, L. (2007), *Bayesian Classification and Regression with High Dimensional Features*, Ph.D. thesis, University of Toronto