Infinite Latent Feature Models Indexed by

Pairwise Distance Information

Arthur Lui

A selected Project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

David B. Dahl, Chair
Committee 1
Committee 2

Department of Statistics

Brigham Young University

April 2014

ABSTRACT

Infinite Latent Feature Models Indexed by
Pairwise Distance Information

Arthur Lui
Department of Statistics, BYU
Master of Science

Abstract goes here

Keywords: Bayesian nonparametrics, Indian buffet process, latent feature models

ACKNOWLEDGMENTS

Acknowledgement goes here

CHAPTER 1

INTRODUCTION

CHAPTER 2

LITERATURE REVIEW

We will first review the Indian buffet process (IBP). We will then review the distance dependent Indian buffet process, formulated by Gershman. Finally, we will finally review existing McMC methods for sampling from the IBP.

The Indian buffet process

The Indian buffet process is a distribution on infinite sparse (left-ordered) binary matrices. (i.e. matrices with finite number of rows and infinite number of columns, ordered by the magnitude of their columns when expressed as a binary number.) The IBP is a multivariate extension of the Chinese restaurant process. For an N x K matrix, Z, that follows an IBP(α) distribution, element z_{ik} is 1 if observation (customer) i possess feature (dish) k, and 0 otherwise. And α is a parameter that determines the sparsity of the matrix. The larger α is, the more likely Z will be sparse. The process can be generated as follows:

N customers enter a buffet one after another. The buffet line contains an infinite number of dishes. The first customer takes the first Poisson(α) number of dishes. The i^{th} customer then takes previously sampled dishes with probability proportional to their popularity, serving himself with probability $\frac{m_k}{i}$, where m_k is the number of people that had previously taken dish k. After customer i has sampled all the previously sampled dishes, he samples Poisson($\frac{\alpha}{i}$) new dishes. The probability of any matrix being generated by this process is

$$P(\boldsymbol{Z}) = \frac{\alpha^{K_+}}{\prod_{i=1}^{N} K_1^{(i)}!} exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}, \tag{2.1}$$

where H_N is the harmonic number, $\sum_{i=1}^{N} \frac{1}{i}$, K_+ is the number of non-zero columns in $\boldsymbol{Z}$, and $K_1^{(i)}$ is the number of new dishes sampled by customer i.

Gibbs sampler

CHAPTER 3

METHODS