
Variational Inference for Dirichlet Process to Stratify Cancer Patients Using DNA Methylation

Anonymous Author(s)

Affiliation

Address

email

Abstract

Variational Inference (VI) is an alternative strategy to Markov Chain Monte Carlo (MCMC) which tends to be faster and easier to scale to larger datasets (Blei et al., 2016). This is especially advantageous in applications that interact with high dimensional data. Previous work has been done on a VI method for Dirichlet Process Mixture Models (DPMMs) based on the well-known stick-breaking process (Blei et al., 2016). In contrast, we propose a similar model, but based on the Chinese Restaurant Process (CRP) instead. This model has fewer parameters to estimate; therefore, we hypothesize that our model will perform faster than the model by Blei et al.

DNA methylation is an epigenetic mark that is associated with transcriptional repression and may be closely related to cancer. A common objective is to identify a latent structure shared across cancers from different tissue types reflecting commonly altered gene pathways. These latent structures stratify cancer patients into functionally similar groups and can inform therapy decisions in clinical applications. This study will apply our VI model on DNA methylation data for stratification.

1 Introduction

1.1 DNA Methylation

Cancers develop via the acquisition of genomic changes. These changes in turn alter the cells harbouring them, leading to changes in cell state (phenotype). One common effect of these mutations is to induce a series of modifications to the genome, which do not alter the encoded DNA but rather the ability of the DNA to be read and processed into protein. These heritable non-genetic changes are broadly referred to as epigenetic changes. According to Baylin and Jones (2011), "Epigenetic alterations are leading candidates for the development of specific markers for cancer detection, diagnosis and prognosis". One such epigenetic change is DNA methylation, which is unambiguously linked with transcriptional repression. When present in promoter regions, DNA methylation correlates negatively with gene expression; furthermore, characteristic changes in DNA methylation have been reported for cancer. Research indicates that gene promoter CpG islands acquire abnormal hypermethylation resulting in transcriptional silencing in cancer (Bock, 2012). It is important to note that 70-80% of CpGs in the human genome are affected by DNA methylation. Therefore, understanding the effects of DNA methylation in cancer can provide valuable information for finding an effective remedy for cancer in humans. The potential relationship between DNA methylation

and cancer opens up a new avenue of exploration. Advances in next-generation sequencing and microarray technology allow for analysis on DNA methylation in large samples and genome-wide; currently, DNA methylation is the only epigenetic mark that can be measured reliably. As a result, DNA methylation can be profiled accurately using high throughput sequencing and is a potential feature for categorizing cancer patients into functionally similar groups.

1.2 Variational Inference

Variational inference is an alternative strategy to Markov Chain Monte Carlo (MCMC) sampling which tends to be faster and easier to scale to larger datasets (Blei et al., 2016). The key characteristic of variational inference is that it casts Bayesian inference as an optimization problem (Salimans et al., 2015). Variational inference attempts to approximate the posterior with another distribution $q_\theta(z|x)$ by choosing its parameters θ to optimize the evidence lower bound (ELBO) on the marginal likelihood,

$$\begin{aligned}\log p(x) &\geq \log p(x) - D_{KL}(q_\theta(z|x)||p(z|x)) \\ &= E_{q_\theta(z|x)}(\log p(x, z) - \log q_\theta(z|x))\end{aligned}$$

In recent years, there have been many advances in the field of VI, which are summarized in a review by Zhang et al., (2019).

1.3 Dirichlet Process

The Dirichlet process is a stochastic process used for Bayesian nonparametric regression; in particular, for constructing Dirichlet process mixture models (Neal, 2000). A Dirichlet process G is a distribution of distributions consisting of a base distribution G_0 and a positive real number α , and can be written as,

$$G \sim \text{DP}(G_0, \alpha)$$

where G_0 is a continuous distribution such that the probability of any two samples generated from this distribution being equal is zero, whereas G is a discrete distribution consisting of infinitely many number of point masses, so the probability of two samples colliding is non-zero. For any measurable finite k partitions $\{B_i\}_{i=1}^k$,

$$G(B_1), \dots, G(B_k) \sim \text{Dir}(\alpha G(B_1), \dots, \alpha G(B_k))$$

then

$$\Pr[X_1, \dots, X_k] = \int \Pr[G] \prod_{i=1}^k \Pr[X_i | G] dG$$

which represents the dependencies among $\{X_i\}_{i=1}^k$ by marginalizing out G . The Chinese restaurant process, stick-breaking process and Polya urn scheme are three common perspectives regarding the Dirichlet Process, and only the first method will be focused on in this project. Assuming a restaurant has infinitely many tables, and let $\{X_i\}_{i=1}^k$ be the customers of the restaurant. Then $\{X_i\}_{i=1}^k$ are partitioned determined by the same table the represented customers sitting at. Considering the behavior of X_n given the previous $n-1$ observations:

$$X_n | (X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}) = \begin{cases} x_n & \text{with probability } \frac{|\{j: x_j = x_n\}|}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

where $|\{j: x_j = x_n\}|$ is the number of times the value x_n occurs in $\{x_1, x_2, \dots, x_n\}$.

The fact of the Chinese restaurant process is that people are more likely to sit at the tables where many people are already sitting, however, the customers will sit at a new table with probability

67 proportional to α .

68

69 Dirichlet Process mixtures assumes that the data originally come from a mixture of an infinite number
 70 of distributions. The well-known EM algorithm is used for inference in mixture models by optimizing
 71 likelihood, but MCMC and Variational inference are better options as G is nonparametric.

72 2 Related Work

73 3 Methods

74 3.1 Model

75 We propose a Variational Inference Dirichlet Process Gaussian Mixture Model based on the Chinese
 76 Restaurant Process, to stratify cancer patients based on DNA methylation values. This paper will
 77 build upon Blei and Jordan’s (2006) work on Variational Inference for Dirichlet Process Mixtures.
 78 Our method differs from theirs in that we base our model on a Chinese Restaurant Process as opposed
 79 to a stick-breaking process. This Variational Inference model will be implemented in the probabilistic
 80 programming language Pyro. See Figure 1 for the graphical models for each variant of the model.

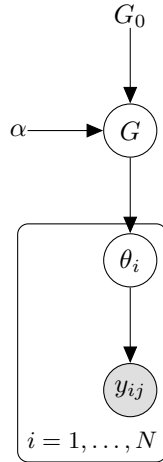


Figure 1: Graphical model for DPMM based on CRP

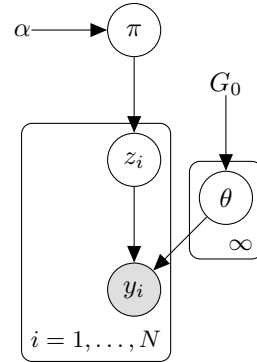


Figure 2: Graphical model for DPMM based on the stick-breaking process

81 3.2 Plots

82 We intend to use clustermaps to visualize the main results, as our study is solving a clustering
 83 problem. Color bars on the left hand side indicates cancer type and purple squares represent the
 84 clusters. The more data points in the cluster, the larger the squares. The intensity of a square describes
 85 the proportion of iterations the data point of interest appeared in that cluster. We expect that the
 86 clustermaps in this study will resemble Figure 3, because data points should cluster by cancer types.

87 4 Conclusion

88 Variational Inference (VI) is an alternative strategy to Markov Chain Monte Carlo (MCMC) which
 89 tends to be faster and easier to scale to larger datasets (Blei et al., 2016). This is especially advan-
 90 tageous in applications that interact with high dimensional data, such as DNA methylation cluster
 91 analysis. We intend to build off the VI for DPMM model based on the stick-breaking process by
 92 Blei et al. and implement the same model, based on the Chinese Restaurant Process (CRP); the
 93 model will be written in the probabilistic programming language Pyro. We hypothesize that our

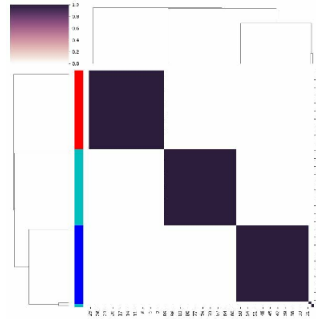


Figure 3: Example of a clustermap. Color labels indicate cancer types: Type I, Type II, Type III

94 model will be faster because it contains fewer parameters to estimate. Our model will be applied
 95 to a bioinformatic problem - stratifying cancer patients into functionally similar groups using DNA
 96 methylation. Consequently, these functionally similar groups can inform therapy decisions in clinical
 97 applications.

98 **5 Headings: first level**

99 All headings should be lower case (except for first word and proper nouns), flush left, and bold.

100 First-level headings should be in 12-point type.

101 **5.1 Headings: second level**

102 Second-level headings should be in 10-point type.

103 **5.1.1 Headings: third level**

104 Third-level headings should be in 10-point type.

105 **Paragraphs** There is also a `\paragraph` command available, which sets the heading in bold, flush
106 left, and inline with the text, with the heading followed by 1 em of space.

107 **6 Citations, figures, tables, references**

108 These instructions apply to everyone.

109 **6.1 Citations within the text**

110 The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as
111 long as you maintain internal consistency. As to the format of the references themselves, any style is
112 acceptable as long as it is used consistently.

113 The documentation for `natbib` may be found at

114 `http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

115 Of note is the command `\citet`, which produces citations appropriate for use in inline text. For
116 example,

117 `\citet{hasselmo}` investigated\dots

118 produces

119 Hasselmo, et al. (1995) investigated...

120 If you wish to load the `natbib` package with options, you may add the following before loading the
121 `neurips_2021` package:

122 `\PassOptionsToPackage{options}{natbib}`

123 If `natbib` clashes with another package you load, you can add the optional argument `nonatbib`
124 when loading the style file:

125 `\usepackage[nonatbib]{neurips_2021}`

126 As submission is double blind, refer to your own published work in the third person. That is, use “In
127 the previous work of Jones et al. [4],” not “In our previous work [4].” If you cite your other papers
128 that are not widely available (e.g., a journal paper under review), use anonymous author names in the
129 citation, e.g., an author of the form “A. Anonymous.”

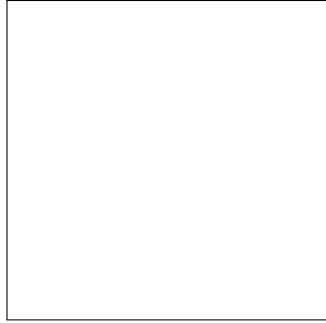


Figure 4: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	~ 100
Axon	Output terminal	~ 10
Soma	Cell body	up to 10^6

130 6.2 Footnotes

131 Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number¹
 132 in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote
 133 with a horizontal rule of 2 inches (12 picas).

134 Note that footnotes are properly typeset *after* punctuation marks.²

135 6.3 Figures

136 All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction.
 137 The figure number and caption always appear after the figure. Place one line space before the figure
 138 caption and one line space after the figure. The figure caption should be lower case (except for first
 139 word and proper nouns); figures are numbered consecutively.

140 You may use color figures. However, it is best for the figure captions and the paper body to be legible
 141 if the paper is printed in either black/white or in color.

142 6.4 Tables

143 All tables must be centered, neat, clean and legible. The table number and title always appear before
 144 the table. See Table 1.

145 Place one line space before the table title, one line space after the table title, and one line space after
 146 the table. The table title must be lower case (except for first word and proper nouns); tables are
 147 numbered consecutively.

148 Note that publication-quality tables *do not contain vertical rules*. We strongly suggest the use of the
 149 booktabs package, which allows for typesetting high-quality, professional tables:

150 <https://www.ctan.org/pkg/booktabs>

151 This package was used to typeset Table 1.

¹Sample of the first footnote.

²As in this example.

7 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

8 Preparing PDF files

Please prepare submission files with paper size “US Letter,” and not, for example, “A4.”

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF file uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdf fonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see <http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf>
- `xfig` “patterned” shapes are implemented with bitmap fonts. Use “solid” shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

```
\usepackage{amsfonts}
```

followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for \mathbb{R} , \mathbb{N} or \mathbb{C} . You can also use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{\mathbb{R}} %real numbers
\newcommand{\Nat}{\mathbb{N}} %natural numbers
\newcommand{\CC}{\mathbb{C}} %complex numbers
```

Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

8.1 Margins in L^AT_EX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the `graphics` bundle documentation (<http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf>)

A number of width problems arise when L^AT_EX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font

size to small (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural Simulation System*. New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to **[Yes]**, **[No]**, or **[N/A]**. You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? **[Yes]** See Section 3.
- Did you include the license to the code and datasets? **[No]** The code and the data are proprietary.
- Did you include the license to the code and datasets? **[N/A]**

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...

- (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
- (b) Did you describe the limitations of your work? **[TODO]**
- (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
- (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**

2. If you are including theoretical results...

- (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
- (b) Did you include complete proofs of all theoretical results? **[TODO]**

3. If you ran experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

- (a) If your work uses existing assets, did you cite the creators? **[TODO]**
- (b) Did you mention the license of the assets? **[TODO]**

- 234 (c) Did you include any new assets either in the supplemental material or as a URL?
235 **[TODO]**
- 236 (d) Did you discuss whether and how consent was obtained from people whose data you're
237 using/curating? **[TODO]**
- 238 (e) Did you discuss whether the data you are using/curating contains personally identifiable
239 information or offensive content? **[TODO]**
- 240 5. If you used crowdsourcing or conducted research with human subjects...
- 241 (a) Did you include the full text of instructions given to participants and screenshots, if
242 applicable? **[TODO]**
- 243 (b) Did you describe any potential participant risks, with links to Institutional Review
244 Board (IRB) approvals, if applicable? **[TODO]**
- 245 (c) Did you include the estimated hourly wage paid to participants and the total amount
246 spent on participant compensation? **[TODO]**

247 **A Appendix**

248 Optionally include extra information (complete proofs, additional experiments and plots) in the
249 appendix. This section will often be part of the supplemental material.