

---

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

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1       The abstract paragraph should be indented ½ inch (3 picas) on both the left- and  
2       right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points.  
3       The word **Abstract** must be centered, bold, and in point size 12. Two line spaces  
4       precede the abstract. The abstract must be limited to one paragraph.

## 5   1   Introduction

### 6   1.1   DNA Methylation

7   Cancers develop via the acquisition of genomic changes. These changes in turn alter the cells  
8   harbouring them, leading to changes in cell state (phenotype). One common effect of these mutations  
9   is to induce a series of modifications to the genome, which do not alter the encoded DNA but rather  
10   the ability of the DNA to be read and processed into protein. These heritable non-genetic changes  
11   are broadly referred to as epigenetic changes. According to Baylin and Jones (2011), "Epigenetic  
12   alterations are leading candidates for the development of specific markers for cancer detection,  
13   diagnosis and prognosis". One such epigenetic change is DNA methylation, which is unambiguously  
14   linked with transcriptional repression. When present in promoter regions, DNA methylation correlates  
15   negatively with gene expression; furthermore, characteristic changes in DNA methylation have  
16   been reported for cancer. Research indicates that gene promoter CpG islands acquire abnormal  
17   hypermethylation resulting in transcriptional silencing in cancer (Bock, 2012). It is important to  
18   note that 70-80% of CpGs in the human genome are affected by DNA methylation. Therefore,  
19   understanding the effects of DNA methylation in cancer can provide valuable information for finding  
20   an effective remedy for cancer in humans. The potential relationship between DNA methylation  
21   and cancer opens up a new avenue of exploration. Advances in next-generation sequencing and  
22   microarray technology allow for analysis on DNA methylation in large samples and genome-wide;  
23   currently, DNA methylation is the only epigenetic mark that can be measured reliably. As a result,  
24   DNA methylation can be profiled accurately using high throughput sequencing and is a potential  
25   feature for categorizing cancer patients into functionally similar groups.

### 26   1.2   Variational Inference

27   Variational inference is an alternative strategy to MCMC sampling which tends to be faster and  
28   easier to scale to larger datasets. The key characteristic of variational inference is that it casts  
29   Bayesian inference as an optimization problem (Salimans et al., 2015). Variational inference attempts  
30   to approximate the posterior with a distribution  $q_{\theta}(z|x)$  by choosing its parameters  $\theta$  to  
31   optimize the evidence lower bound (ELBO) on the marginal likelihood.

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

Coordinate Ascent Mean-Field VI (CAVI) is one of the most common algorithms for solving this optimization problem, and is very similar to the EM algorithm and Gibbs sampling. In short, the main difference between MCMC and VI is that MCMC is a sampling algorithm, while VI is an optimization algorithm; MCMC constructs an ergodic markov chain, while VI approximates the posterior with another distribution.

## 2 Methods

### 2.1 Model

We propose a Variational Inference Dirichlet Process Gaussian Mixture Model based on the Chinese Restaurant Process, to stratify cancer patients based on DNA methylation values. This paper will build upon Blei and Jordan’s work on Variational Inference for Dirichlet Process Mixtures. Our proposed method differs on the basis that we base our model on the Chinese Restaurant Process whereas Blei and Jordan base theirs on the Stick-breaking process. See Figure 1 for the graphical model.

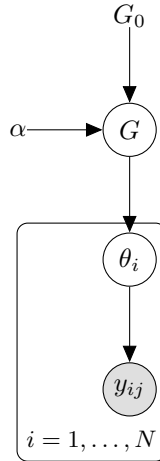


Figure 1: Graphical model of a Dirichlet Process based on the Chinese Restaurant Process

### 2.2 Plots

See Figure 2 for clustermaps that we intend to obtain in this study.

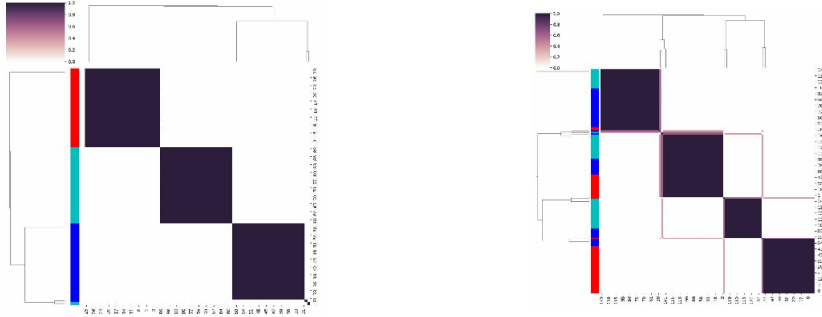


Figure 2: Example clustermap. Color labels indicate cancer types: Type I, Type II, Type III

### 3 Headings: first level

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

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

#### 3.1 Headings: second level

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

##### 3.1.1 Headings: third level

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

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

## 4 Citations, figures, tables, references

These instructions apply to everyone.

### 4.1 Citations within the text

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

The documentation for `natbib` may be found at

<http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf>

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

```
\citet{hasselmo} investigated\dots
```

produces

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

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

```
\PassOptionsToPackage{options}{natbib}
```



Figure 3: Sample figure caption.

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

74 `\usepackage[nonatbib]{neurips_2021}`

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

## 79 4.2 Footnotes

80 Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number<sup>1</sup>  
81 in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote  
82 with a horizontal rule of 2 inches (12 picas).

83 Note that footnotes are properly typeset *after* punctuation marks.<sup>2</sup>

## 84 4.3 Figures

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

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

## 91 4.4 Tables

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

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

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

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

100 This package was used to typeset Table 1.

---

<sup>1</sup>Sample of the first footnote.

<sup>2</sup>As in this example.

Table 1: Sample table title

Part		
Name	Description	Size ( $\mu\text{m}$ )
Dendrite	Input terminal	$\sim 100$
Axon	Output terminal	$\sim 10$
Soma	Cell body	up to $10^6$

## 5 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.

## 6 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}{I\!\!R} %real numbers
\newcommand{\Nat}{I\!\!N} %natural numbers
\newcommand{\CC}{I\!\!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.

### 6.1 Margins in L<sup>A</sup>T<sub>E</sub>X

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}
```

133 See Section 4.4 in the graphics bundle documentation ([http://mirrors.ctan.org/macros/](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf)  
134 [latex/required/graphics/grfguide.pdf](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf))

135 A number of width problems arise when L<sup>A</sup>T<sub>E</sub>X cannot properly hyphenate a line. Please give LaTeX  
136 hyphenation hints using the \- command when necessary.

## 137 References

138 References follow the acknowledgments. Use unnumbered first-level heading for the references. Any  
139 choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font  
140 size to `small` (9 point) when listing the references. Note that the Reference section does not count  
141 towards the page limit.

142 [1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In  
143 G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp.  
144 609–616. Cambridge, MA: MIT Press.

145 [2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the*  
146 *GENeral NEural Simulation System*. New York: TELOS/Springer-Verlag.

147 [3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent  
148 synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

## 149 Checklist

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

- 154 • Did you include the license to the code and datasets? **[Yes]** See Section 2.
- 155 • Did you include the license to the code and datasets? **[No]** The code and the data are  
156 proprietary.
- 157 • Did you include the license to the code and datasets? **[N/A]**

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

### 161 1. For all authors...

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

### 168 2. If you are including theoretical results...

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

### 171 3. If you ran experiments...

- 172 (a) Did you include the code, data, and instructions needed to reproduce the main experi-  
173 mental results (either in the supplemental material or as a URL)? **[TODO]**
- 174 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they  
175 were chosen)? **[TODO]**

- 176 (c) Did you report error bars (e.g., with respect to the random seed after running experi-  
177 ments multiple times)? **[TODO]**
- 178 (d) Did you include the total amount of compute and the type of resources used (e.g., type  
179 of GPUs, internal cluster, or cloud provider)? **[TODO]**
- 180 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 181 (a) If your work uses existing assets, did you cite the creators? **[TODO]**
- 182 (b) Did you mention the license of the assets? **[TODO]**
- 183 (c) Did you include any new assets either in the supplemental material or as a URL?  
184 **[TODO]**
- 185 (d) Did you discuss whether and how consent was obtained from people whose data you're  
186 using/curating? **[TODO]**
- 187 (e) Did you discuss whether the data you are using/curating contains personally identifiable  
188 information or offensive content? **[TODO]**
- 189 5. If you used crowdsourcing or conducted research with human subjects...
- 190 (a) Did you include the full text of instructions given to participants and screenshots, if  
191 applicable? **[TODO]**
- 192 (b) Did you describe any potential participant risks, with links to Institutional Review  
193 Board (IRB) approvals, if applicable? **[TODO]**
- 194 (c) Did you include the estimated hourly wage paid to participants and the total amount  
195 spent on participant compensation? **[TODO]**

## 196 **A Appendix**

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