*Mail*

*We recognize and acknowledge that McMaster University meets and learns on the traditional territories of the Mississauga and Haudenosaunee nations, and within the lands protected by the “Dish With One Spoon” wampum, an agreement amongst all allied Nations to peaceably share and care for the resources around the Great Lakes.*

# STATS 4T03 A/B – Senior Research Project

# 2022 -23

**Supervisor:**  Dr. Pratheepa Jeganathan |  **E-mail:** [jeganp1@mcmater.ca](mailto:jeganp1@mcmater.ca) | **Office: HH-212**

**Co-supervisor:**  Jamie McNicol |  **E-mail:** [jmcnic@mcmater.ca](mailto:jmcnic@mcmater.ca) | **Office: YYY**

**Course coordinator:**  Associate Chair (Undergraduate) | **Email:** [**mathstatsugadvisor@mcmaster.ca**](mailto:mathstatsugadvisor@mcmaster.ca)

**Student:** Shiheng Huang

## Location and time:

**Weekly meetings:** Tuesday (in person) 2.30 -3.30 PM.

**Bi-weekly meetings:** Thursday 9.30-10.30 AM (Virtual)

**Course Website**

Github repository (<https://github.com/PratheepaJ/Topic_models_MIBI>) – private repository. Reading lists, workflows, reports, presentations, will be shared through GitHub.

## Course Description

6 unit(s)

## Build upon an existing Image Analysis and Differential Discovery Pipeline by incorporating Topic Modeling to characterize the tissue microenvironment (TME). The TME is recognized as an important determinant of patient outcome. This course will leverage spatial topic modeling to better report the elements within the TME that correlate with patient outcomes.

## A written report and oral presentation will be required. This course includes a scientific communication component.

## Prerequisite(s):

Registration in Level IV of any Honours Mathematics and Statistics program; and a GPA of at least 9.0; and permission of the Chair of the Department

Not open to students with credit or registration in MATH 4P06 A/B S, PHYSICS 4P06 A/B or ISCI 4A12 A/B.

## Course and Learning Objectives

In this course, the student will:

1. Experience doing research in the mathematical sciences. This may involve solving one or more challenging mathematical or statistical problems, an in-depth analysis of a large data set, or reading and understanding a research paper that involves absorbing a significant amount of new mathematics or statistics.
2. Improve their mathematical writing by writing a document explaining their work.
3. Enhance their oral presentation skills by giving a public presentation explaining their work.
4. Develop additional skills deemed relevant by their supervisor, e.g., coding, probabilistic programming, high-performance computing.

### Coursework and Course Delivery

### Weekly meetings with the supervisor and bi-weekly meetings with the co-supervisors; Occasional tutorial during weekly meetings; two terms. The students will attend the supervisor’s weekly group meetings and contribute to the meetings at least two times per term. The student will learn relevant advanced material (Bayesian modeling and computation) and work on problems to understand that material. The weekly meetings may involve the student presenting solutions to problems, explaining the material they have been reading, writing 2-3 pages of report bi-weekly, submitting computational solutions through GitHub, and asking questions. The meetings may be either in-person or virtual. Final presentations are normally held in person.

### Required Materials/ Resources

Reading lists are selected during weekly meetings.

1. BDA3: Bayesian Data Analysis (Gelman et al., 2013)
2. Online course materials – Bayesian Data Analysis course (Aki Vehtari) <https://avehtari.github.io/BDA_course_Aalto/index.html>
3. Online blog on probability theory, modeling and inference (Michael Betancourt) <https://betanalpha.github.io/writing/>
4. Journal articles.
5. The Elements of Statistical Learning [(Hastie et al., 2009)](https://web.stanford.edu/~hastie/Papers/ESLII.pdf)
6. Online course material – Foundation of graphical models (Blei, 2020). <http://www.cs.columbia.edu/~blei/fogm/2020F/>

### Software:

This course uses R, RStudio, Python, and Python Notebook, which are free. We recommend to set-up the computing environment earlier. From the first day of the class, you must ensure that you are up and running with the computing environment. There are three ways to do computing.

1. Use your personal computing device (laptop/desktop)
   1. Install R (required): [https://www.r-project.org/.](https://www.r-project.org/)
   2. Install RStudio: highly recommended for syntax highlighting, package management, document generation, and more: <https://www.rstudio.com/>.

* The newest version of RStudio is highly recommended.
  1. Install Latex, which will enable you to create PDFs directly from the RMarkdown in RStudio.
* Install [TinyTex](https://yihui.name/tinytex/) package: install.packages("tinytex", repos = "https://cloud.r-project.org/").
* After installing TinyTex, close RStudio.
* Reopen RStudio.
  1. Install [Python](https://www.python.org/downloads/) (required).
  2. Install [Anaconda](https://www.anaconda.com/).

1. Use interactive computing notebooks hosted by the Digital Research Alliances of Canada (previously known as Compute Canada).
   1. Visit <https://mcmaster.syzygy.ca/>
   2. Sign in with your McMaster credentials.
   3. Choose R or Python.
   4. You can download the PDF of your workflow.
2. Google Collab and RStudio Cloud.
   1. Visit <https://rstudio.cloud/> and create an account.
   2. Visit <https://colab.research.google.com/> and create an account or use your Google account to sign up.

Other recommended tools

1. Install git <https://git-scm.com/> for version control.
2. Create an account at <https://github.com/>. Use your McMaster email to get an education account at <https://education.github.com/pack/join>.
3. Create an account at <https://www.shinyapps.io/> to deploy your Shiny applications.
4. Stan for probabilistic programming (installing R package -rstan) <https://mc-stan.org/users/interfaces/rstan>.

## Course Overview and Assessment

### Topics:

1. Bayesian modeling
2. Bayesian computation
3. Latent variable modeling
4. Gaussian mixture models
5. Expectation and maximization algorithm
6. Latent variable modeling for microbiome data
7. Latent Dirichlet Allocation (LDA)
8. Variational expectation-maximization algorithm
9. Collapsed Gibbs sampling
10. Stan – probabilistic programming language
11. Spatial LDA
12. Empirical Bayes
13. Spatial feature engineering
14. Goodness-of-fit for latent variable models

### Course Schedule (tentative):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Week** | **Topics** | **Readings** | **Notes** |
| 09/06/2022 | Week 1 | Prepared reading lists  Probability and inference | Bayesian data analysis course – Lecture 1  <https://avehtari.github.io/BDA_course_Aalto/> |  |
| 09/12/2022 | Week 2 | Setting up computing environment,  Single-parameter models | Git, Github, RStudio  Bayesian data analysis – Chapter 2  Bayesian data analysis course R demos – Lecture 2  <https://avehtari.github.io/BDA_course_Aalto/> Michael Betancourt’s blog on probability theory <https://betanalpha.github.io/writing/> | Weekly meeting – frequentist versus Bayesian inference, Bayes’ theorem,  Grid approximation,  acceptance-rejection sampling |
| 09/19/2022 | Week 3 | Multiparameter models | Bayesian data analysis – Chapter 3  Bayesian data analysis course R demos – Lecture 3  <https://avehtari.github.io/BDA_course_Aalto/> Michael Betancourt’s blog on probability theory <https://betanalpha.github.io/writing/> | Report 1 (Week 1 and Week 2) |
| 09/26/2022 | Week 4 | Introduction to Bayesian computation | Bayesian data analysis – Chapter 10  Bayesian data analysis course R demos – Lecture 4  <https://avehtari.github.io/BDA_course_Aalto/> |  |
| 09/30/2022 | No classes |  |  |  |
| 10/03/2022 | Week 5 | Markov chain Monte Carlo | Bayesian data analysis – Chapters 11-12  Bayesian data analysis course R demos – Lecture 5-6  <https://avehtari.github.io/BDA_course_Aalto/> Michael Betancourt’s blog on MCMC <https://betanalpha.github.io/writing/> | Report 2 (Week 3 and Week 4) |
| 10/10/2022 | Week 6 | Midterm recess |  |  |
| 10/17/2022 | Week 7 | Introduction to Stan | Bayesian data analysis course R demos – Lecture 6 Michael Betancourt’s blog on introduction to Stan <https://betanalpha.github.io/writing/> |  |
| 10/24/2022 | Week 8 | Latent variable modeling | Build, compute, critique, repeat: Data analysis with latent variable models [(Blei 2014)](http://www.cs.columbia.edu/~blei/fogm/2020F/readings/Blei2014.pdf) | Report 3 (Week 5 and 7) |
| 10/31/2022 | Week 9 | Gaussian mixture models,  Expectation and maximization algorithm | The Element of Statistical Learning [(Hastie et al., 2009)](https://web.stanford.edu/~hastie/Papers/ESLII.pdf)  Section 6.8 (Gaussian mixture models), Chapter 8 (EM algorithm) |  |
| 11/07/2022 | Week 10 | Latent variable modeling for omics data | Latent variable modeling for the microbiome ([Sankaran et al., 2019](https://academic.oup.com/biostatistics/article/20/4/599/5032578)) | Report 4 (Start to prepare midterm report – draft 1) |
| 11/14/2022 | Week 11 | Latent Dirichlet Allocation (LDA) |  |  |
| 11/21/2022 | Week 12 | Variational expectation-maximization algorithm |  | Report 5 (Start to prepare midterm report – draft 2) |
| 11/28/2022 | Week 13 | Collapsed Gibbs sampling |  | Midterm report due – December 1, 2022 |
| 12/05/2022 | Week 14 | Stan – probabilistic programming language |  |  |
|  |  |  |  |  |
| Winter 2023 |  | Stan – probabilistic programming language  Spatial LDA  Empirical Bayes  Spatial feature engineering  Goodness-of-fit for latent variable models |  |  |

### Assessment

Final marks will be based upon two major components:

* **Thesis (40%)**: The final thesis should be 15-30 pages in length. It should be written in LaTeX, or a professional document preparation system appropriate for the presentation of scientific material. At a minimum, there should be an abstract, an introduction, and a bibliography.
* **Bi-weekly reports (40%):** Student will write 1-2 pages of report bi-weekly and submitting computational solutions through GitHub.
  + Eight - Ten reports (two terms) each weigh 5% - 4%.
* **Presentation (20%)**: There will be a forum during the April 2023 final exam period, within seven days of the end of classes, in which all students writing a senior research project in the department will give a final presentation. Specific details will be determined by that year's examining committee, but presentations of 15 minutes are suggested.

The final marks will be given by an examining committee of at least three faculty members which consists of all of the supervisors of the Math 4P06/Stats 4T06 students in that year and chaired by the associate chair (undergraduate). If this description does not provide enough people, additional members will be co-opted by the associate chair. The presentation will count for 20% of the final mark. The distribution of the remaining 80% will be at the discretion of the supervisor; 40% for coursework and 40% for the thesis is suggested. The student must receive a passing mark for the thesis in order to pass the course overall.

## Timelines

## December 1, 2022 (approximately): Before the end of the December exam period, the student must submit a midterm report to the supervisor and the associate chair. The midterm report should detail the material covered so far and give a proposed outline of the thesis. If the midterm report is not submitted, the student will be disqualified from continuing with the course.

## February 15, 2023 (approximately): The student should submit a preliminary draft of the thesis. The supervisor should read and comment on the draft within two weeks.

## March 5, 2023 (approximately): The thesis should be mostly completed and the student should start work on the presentation. The final version of the thesis should be submitted on a date agreed between student and supervisor, but in time for a final grade to be submitted to the department within two days of the presentations.

## April 15, 2022 (approximately): Students give their final presentations.

## Requests for Relief for Missed Academic Term Work

McMaster Student Absence Form (MSAF): In the event of an absence for medical or other reasons, students should review and follow the Academic Regulation in the Undergraduate Calendar “Requests for Relief for Missed Academic Term Work”.

**MSAF Course Specific Information**

Absences for a longer duration or for other reasons must be reported to your Faculty/Program office, with documentation, and relief from term work may not necessarily be granted. Please note that the MSAF may not be used for any part of the project work.

**The weight (4-5%) for the missed report due to MSAF will be used for the next report to calculate the final grade. We will not accept MSAF for thesis and presentation.**

## Academic Accommodation of Students with Disabilities

Students with disabilities who require academic accommodation must contact [Student Accessibility Services (SAS](https://sas.mcmaster.ca/)) at 905-525-9140 ext. 28652 or [sas@mcmaster.ca](mailto:sas@mcmaster.ca) to make arrangements with a Program Coordinator. For further information, consult McMaster University’s [*Academic Accommodation of Students with Disabilities*](https://secretariat.mcmaster.ca/app/uploads/Academic-Accommodations-Policy.pdf) policy.

## Academic Accommodation for Religious, Indigenous Or Spiritual Observances (Riso)

Students requiring academic accommodation based on religious, indigenous or spiritual observances should follow the procedures set out in the [RISO](https://secretariat.mcmaster.ca/app/uploads/2019/02/Academic-Accommodation-for-Religious-Indigenous-and-Spiritual-Observances-Policy-on.pdf) policy. Students should submit their request to their Faculty Office ***normally within 10 working days*** of the beginning of term in which they anticipate a need for accommodation or to the Registrar's Office prior to their examinations. Students should contact their instructors as soon as possible to make alternative arrangements for classes, assignments, and tests.

## Course with An On-Line Element

Some courses may use on-line elements (e.g. e-mail, Avenue to Learn (A2L), LearnLink, web pages, capa, Moodle, ThinkingCap, etc.). Students should be aware that, when they access the electronic components of a course using these elements, private information such as first and last names, user names for the McMaster e-mail accounts, and program affiliation may become apparent to all other students in the same course. The available information is dependent on the technology used. Continuation in a course that uses on-line elements will be deemed consent to this disclosure. If you have any questions or concerns about such disclosure, please discuss this with the course instructor.

## Academic Integrity

You are expected to exhibit honesty and use ethical behaviour in all aspects of the learning process. Academic credentials you earn are rooted in principles of honesty and academic integrity.

**It is your responsibility to understand what constitutes academic dishonesty.**

Academic dishonesty is to knowingly act or fail to act in a way that results or could result in unearned academic credit or advantage. This behaviour can result in serious consequences, e.g. the grade of zero on an assignment, loss of credit with a notation on the transcript (notation reads: “Grade of F assigned for academic dishonesty”), and/or suspension or expulsion from the university. For information on the various types of academic dishonesty please refer to the [*Academic Integrity Policy.*](https://secretariat.mcmaster.ca/app/uploads/Academic-Integrity-Policy-1-1.pdf)

**The following illustrates only three forms of academic dishonesty:**

* plagiarism, e.g. the submission of work that is not one’s own or for which other credit has been obtained.
* improper collaboration in group work.
* copying or using unauthorized aids in tests and examinations.

## Authenticity / Plagiarism Detection

***Some courses may*** use a web-based service (Turnitin.com) to reveal authenticity and ownership of student submitted work. For courses using such software, students will be expected to submit their work electronically either directly to Turnitin.com or via an online learning platform (e.g. A2L, etc.) using plagiarism detection (a service supported by Turnitin.com) so it can be checked for academic dishonesty.

Students who do not wish their work to be submitted through the plagiarism detection software must inform the Instructor before the assignment is due. No penalty will be assigned to a student who does not submit work to the plagiarism detection software. **All submitted work is subject to normal verification that standards of academic integrity have been upheld** (e.g., on-line search, other software, etc.). For more details about McMaster’s use of Turnitin.com please go to the [McMaster Office of Academic Integrity’s](https://www.mcmaster.ca/academicintegrity/) webpage.

## Conduct Expectations

As a McMaster student, you have the right to experience, and the responsibility to demonstrate, respectful and dignified interactions within all our living, learning and working communities. These expectations are described in the [*Code of Student Rights & Responsibilities* (the “Code”).](https://secretariat.mcmaster.ca/app/uploads/Code-of-Student-Rights-and-Responsibilities.pdf) All students share the responsibility of maintaining a positive environment for the academic and personal growth of all McMaster community members, **whether in person or online**.

It is essential that students be mindful of their interactions online, as the Code remains in effect in virtual learning environments. The Code applies to any interactions that adversely affect, disrupt, or interfere with reasonable participation in University activities. Student disruptions or behaviours that interfere with university functions on online platforms (e.g. use of Avenue 2 Learn, WebEx or Zoom for delivery), will be taken very seriously and will be investigated. Outcomes may include restriction or removal of the involved students’ access to these platforms.

## Copyright and Recording

Students are advised that lectures, demonstrations, performances, and any other course material provided by an instructor include copyright protected works. The Copyright Act and copyright law protect every original literary, dramatic, musical and artistic work, **including lectures** by University instructors.

The recording of lectures, tutorials, or other methods of instruction may occur during a course. Recording may be done by either the instructor for the purpose of authorized distribution, or by a student for the purpose of personal study. Students should be aware that their voice and/or image may be recorded by others during the class. Please speak with the instructor if this is a concern for you.

## Extreme Circumstances

The University reserves the right to change the dates and deadlines for any or all courses in extreme circumstances (e.g., severe weather, labour disruptions, etc.). Changes will be communicated through regular McMaster communication channels, such as McMaster Daily News, A2L and/or McMaster email.

## Important Dates

|  |  |
| --- | --- |
| **Description** | **Dates** |
| Classes Begin | Sep 6, 2022 |
| Last day for enrollment and course changes (drop/add) | Sep 14, 2022 |
| National Day for Truth and reconciliation; No classes | Sep 30, 2022 |
| Mid-term recess | Oct 10-16, 2022 |
| Last day for canceling courses without failure by default | Nov 11, 2022 |
| Test and Examination Restriction | Dec 2-8, 2022 |
| Classes End | Dec 8, 2022 |
| Midterm report due | Dec 1, 2022 |
|  |  |

*Disclaimer: This outline provides a general plan; deviations may be necessary. Any change in the course outline will be posted on the course website, and the details will be announced in class.*