UC Santa Barbara, Department of Statistics & Applied Probability Spring, 2022

PSTAT 220C: Advanced Statistical Methods

Instructor:	Dr. Mengyang (Michael) Gu
Office:	South Hall 5511
Email:	*mengyang@pstat.ucsb.edu*
Lecture:	Tuesday and Thursday 5:00pm - 6:15pm
Format:	The lecture will be held in Phelps 2536. The instructor is attending a conference between Apr 11-15 and recorded videos will be uploaded. Students do not need to come to class for the week of Apr 11-15.
Office Hours:	Tuesday 3:30pm-4:30pm
OH Format:	Hybrid. Students can use the following Zoom link or come to my office.
Zoom Link:	https://ucsb.zoom.us/j/81707945375?pwd=ODlEZGQxY0xwanlOcFFlWlAxS01CZz09 password: pstat_220c
TA:	Xubo Liu, xubo@umail.ucsb.edu
Section:	In person in Phelps 2524, Thursday 8:00am - 8:50am
TA Office Hours:	Monday 12:30pm - 1:30pm
OH Format:	Remote. Students can use the following Zoom link.
Zoom Link:	
Materials:	
Textbook	*Applied Multivariate Statistical Analysis*, 6th ed., by R. A. Johnson and D. W. Wichern.
Reference	1.*Matrix Cookbook*, by Kaare Brandt Petersen, an online version is available at https://www.math.uwaterloo.ca/ hwolkowi/matrixcookbook.pdf 2. *Dynamic linear models with R*, by Giovanni Petris, Sonia Petrone and Patrizia Campagnoli. Springer-Verlag New York. An online version is available at: http://people.bordeaux.inria.fr/pierre.delmoral/dynamics-linear-models.petris_et_al.pdf 3. *Gaussian process for machine learning*, by Carl Edward Rasmussen and Christopher K. I. Williams. MIT press. An online version is available at: http://www.gaussianprocess.org/gpml/chapters/RW.pdf
Software	This course uses the statistical software `R` to perform data analysis. It can be downloaded from http://www.r-project.org/. After installing `R`, you may download and install its integrated development environment `RStudio` from https://www.rstudio.com/products/RStudio/. You may also use Rstudio cloud https://rstudio.cloud.
Course Website:	GauchoSpace Lecture notes, demo code, optional reading materials and homework will be posted on GauchoSpace.

Grading:

Homework:	30%
Lecture and section participation:	10%
Data academy assignment:	10%
Final project or literature review article:	50%

There is no exam or quiz.

Homework:

(1) 4 homework will be posted on GauchoSpace and please submit your homework through GauchoSpace. These 4 homework assignments are due on Wednesday before class.

You can type down the answer of your homework or write it by hand. Upload the homework in Gauchospace on time.

Late homework policy: total grade based on 80% if late within 24 hours; 60% if late within 2 days; 40% if late within 3 days; 20% if late within 4 days.

Participation:

To obtain full scores, students should participate in discussion for not less than 75% of the lecture and section meetings.

Final Literature review:

You will review 2 papers, which can be chosen from [11, 15, 2, 1, 7, 18, 17, 9, 8, 12, 6, 5, 3, 14, 4, 13, 10, 16]. If you need to review other papers, please come to discuss the papers with me during my office hours. Please limit your review to 12 pages in length. For a scientific paper, the supplementary materials are required to review. Font size should be not smaller than 11-point and single space between lines is required. **If you are not sure, please come to my office hours in the first weeks to choose a suitable paper to review.**

Final research article:

Your project report should have the similar format of a research article, which typically contains title, abstract, introduction, a section of research goals, statistical model and data set, numerical analysis, conclusion and references. Please type your project report using latex or word. Computer code is NOT needed, but can be submitted as a separate file

Clearly describe your research question(s), summarize related studies using your own words. Introduce the data set you use. Analyze the data using statistical approaches/models. Explain and interpret the result. You may compare it with the existing methods. Discuss the findings and have appropriate references. The report should be not more than 12 pages in length, including references, tables and figures. Font size should be not smaller than 11-point and single space between lines is required.

Plagiarism is strictly prohibited.

Course Goal: This course has two goals. The first goal is to develop mathematical reasoning and computational skills for multivariate statistics. The second goal is to review research articles, analyze real data set with R software and other software. Computation, algorithms, data visualization and analysis are the focus of this course. This course is to prepare you for a data relevant research project and to write research articles.

Course Content: Multivariate statistical analysis. Topics selected from matrix theory, sampling distribution, multivariate normal distribution, Gaussian process, multivariate linear regression, time series, Kalman filter, principal component analysis and factor analysis. Emphasis on application rather than theory.

Academic Honesty: The assignments are meant to be challenging, and you are encouraged to discuss them with your classmates. However, when you set your work down on paper, I expect it to be your own thoughts. We expect all students at UCSB to maintain the integrity of the academic community. While collaboration is an important part of the learning process, any written work submitted for homework should be your own thoughts and not copied from any other source. **We will take steps to detect cheating, and any students breaking the rules will fail the class and be reported to the appropriate university authorities**.

IP: All course materials (class lectures and discussions, handouts, examinations, web materials) and the intellectual content of the course itself are protected by United States Federal Copyright Law, the California Civil Code. The UC Policy 102.23 expressly prohibits students (and all other persons) from recording lectures or discussions and from distributing or selling lectures notes and all other course materials without the prior written permission of the instructor (See http://policy.ucop.edu/doc/2710530/PACAOS-100).

PSTAT 220C Course Schedule and Related Book Chapters (Tentative)

Spring 2022

	Materials	Assignment Due
Week 1, Mar 28–Apr 3	Syllabus, matrix, decomposition, tensor	/
Week 2, Apr 4–Apr 10	Sampling distribution/models	/
Week 3, Apr 11–Apr 17	Multivariate normal distribution	HW 1, Apr 14 before class
Week 4, Apr 18–Apr 24	Inference of the mean vector	/
Week 5, Apr 25–May 1	Multivariate linear regression	HW 2, Apr 28 before class
Week 6, May 2–May 8	Dynamic linear models (DLM), Kalman filter	
Week 7, May 9–May 15	DLM packages	HW 3, May 12 before class
Week 8, May 16–May 22	Correlated/functional data, statistical emulator	/
Week 9, May 23–May 29	Unsupervised learning, (probabilistic) PCA	HW 4 May 26, before class
Week 10, May 30–June 5	Tensor decomposition, classification or clustering	
Week 11, June 6–June 12	Final report	June 10, 11:59pm PT

Week 1: Textbook Chapter 2.1-2.5, 2.7

Week 2: Textbook Chapter 2.6, 3

Week 3: Textbook Chapter 4

Week 4: Textbook Chapter 5

Week 5: Textbook Chapter 7

Week 6: DLM with R Chapter 2

Week 7: DLM with R Chapter 2

Week 8: GP for ML Chapter 2/4

Week 9: Textbook Chapter 8/9

Week 10: Textbook Chapter 11/12

Please review lecture notes and demo code before working on homework assignments.

References

[1] Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. The Annals of Statistics, 40(1):436–465, 2012.

[2] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. Econometrica, 70(1):191–221, 2002.

[3] Volker L Deringer, Albert P Bartók, Noam Bernstein, David M Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. Chemical Reviews, 121(16):10073–10141, 2021.

[4] Peter I Frazier. A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811, 2018.

[5] Mengyang Gu and James O Berger. Parallel partial Gaussian process emulation for computer models with massive output. Annals of Applied Statistics, 10(3):1317–1347, 2016.

[6] Mengyang Gu, Xubo Liu, Xinyi Fang, and Sui Tang. Scalable marginalization of latent variables for correlated data. arXiv preprint arXiv:2203.08389, 2022.

[7] Mengyang Gu and Weining Shen. Generalized probabilistic principal component analysis of correlated data. Journal of Machine Learning Research, 21(13), 2020.

[8] Jouni Hartikainen and Simo Sarkka. Kalman filtering and smoothing solutions to temporal gaussian process regression models. In Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on, pages 379–384. IEEE, 2010.

[9] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. Journal of basic Engineering, 82(1):35–45, 1960.

[10] Mohammad Emtiyaz Khan and Håvard Rue. The bayesian learning rule. arXiv preprint arXiv:2107.04562, 2021.

[11] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. SIAM review, 51(3):455–500, 2009.

[12] Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. Proc. Natl. Acad. Sci. U.S.A., 116(29):14424–14433, 2019.

[13] Benjamin J Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I Martinez Alvarado, Jacob M Janey, Ryan P Adams, and Abigail G Doyle. Bayesian reaction optimization as a tool for chemical synthesis. Nature, 590(7844):89–96, 2021.

[14] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 25, 2012.

[15] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):611–622, 1999.

[16] William J Wilkinson, Simo Särkkä, and Arno Solin. Bayes-newton methods for approximate bayesian inference with psd guarantees. arXiv preprint arXiv:2111.01721, 2021.

[17] Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. IEEE Transactions on Information Theory, 64(11):7311–7338, 2018.

[18] Anru R Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic pca: Algorithm, optimality, and applications. arXiv preprint arXiv:1810.08316, 2018.