Statistical and Discrete Methods for Scientific Computing

The 2014 Concepts Study Page lists all the questions that may be asked in the oral exam.

You can generate your own practice exam at the 2014 Practice Exam Machine.

CSE383M (65280) and CS395T (53715), Spring 2014

Welcome to the course! The instructor is Professor William Press (Bill), and the TA is Jeff Hussmann (Jeff). We meet Mondays and Wednesdays, 1:30 - 3:00 p.m. in CBA 4.344 with Bill, and Fridays, 1:30 - 3:00 p.m. in CBA 4.348 with Jeff. The course is aimed at first or second year graduate students, especially in the CSEM, CS, and ECE programs, but others are welcome. You'll need math at the level of *at least* 2nd year calculus, plus linear algebra, plus either more continuous math (e.g., CSEM students) or more discrete math (e.g., CS and ECE students). You'll also need to be able to program in some known computer language.

Mechanics of the Course

The last two years, we have tried the experiment of a "flipped" course. This has worked so well that we are doing this again this year. "Flipped" means that the lectures are all on the web as recorded webcasts. You **must** watch the assigned webcasts **before** the class for which they are scheduled; maybe watch them more than once if there are parts that you don't easily understand. Then, you will be ready for the active learning that we do in class. The class activities will **not** "cover the material". Rather, class is supposed to be for "aha moments" and for "fixing" the material in your learning memory. We'll thus do various kinds of "active learning" activities that will test and improve your understanding of the material in the lecture. Such in-class activities, often done in *randomized* groups of two or three, may include

- group computer programming exercises
- group working of problems
- group writing assignments
- discussing concepts (and communicating ideas back to the whole class)
- "quiz show" style activities
- short surprise quizzes (generally at the beginning of class -- no makeups allowed)
- whatever else we all think of

Problems for Each Segment

Every lecture segment home page has one or two relatively easy "skill" problems. You should work these after watching the segment, before class. (You might be asked to discuss your solution with your small group in class.) Also on the segment's page are one or two concept thought problems. One or another of these will sometimes be the basis of in-class activities, so you might want to think about them before class.

Student Wiki Pages

Every student will have a wiki page (and as many linked pages as you want). You can post your solutions to as many problems as you wish to your wiki page. You can do this either before the relevant class or afterwards. You can also make up, and solve, additional problems. Problems won't be individually graded. However, at the end of the course, the completeness and quality of you wiki page(s) will be a part of your course grade. Your wiki page can include discussion of the thought problems, as well as the skill problems.

You can also post signed comments on any other student's wiki pages. To the extent that these are generally helpful, they will add credit to your reputation and for your grade.

Student Pages

Laptops or Tablets

You **must** bring your laptop computer or full-sized tablet to every class, so that you can (i) look things up during group discussions or problem sessions and (ii) do in-class programming exercises. You can program in any language you want. For Python, which we recommend as the best choice for this course, you can either install it on your machine, or else use the IPython notebook server described in class. The course will include several lectures of Python workshop by Jeff.

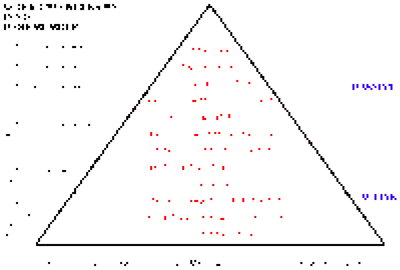
If you instead want to use MATLAB or Mathematica, that is fine, but please be sure that it is installed on your computer before the first class. (The MATLAB Student Edition is a real bargain.) For C, C++, Java, etc., please be sure that you have a fully working environment for compiling and running small pieces of code.

Course Requirements and Grading

Grades will be based on these factors

- in-class attendance and participation
- an in-class midterm exam
- completeness and quality of your individual wiki page(s)
- relevance and usefulness of your comments on other people's wiki pages (or on the main wiki)
- an individual 30-minute final oral exam

In previous years there was a term project, but not this year. Your working the problems and posting solutions on your wiki page is this year's substitute.



Click image to see a legible version.

What is Active Learning?

Much research shows that lecture courses, where students listen passively as the instructor talks, are inefficient ways to learn. What works is so-called active learning, a broad term that, for us, basically means that class time is too valuable to waste on lectures. (See image at right.)

The lectures are all recorded as webcasts, but webcasts are not active learning. However, they are a starting point as a "linear" introduction to the material.

Feedback

What has worked well in class so far? What hasn't worked? How could things be improved? Please leave feedback.

Resources and Links

There is no textbook for the course. A list of recommended supplementary books is here.

Some resources for learning Python can be found here.

Some MATLAB resources can be found here.

Webcast Lecture Segments *(Opinionated Lessons in Statistics)*

All of the lectures are in the form of webcasts, divided into segments of about 15-30 minutes each (occasionally a bit longer). Each segment, has a wiki page, page links below. You can view the lecture on its wiki page, which also has additional stuff about the segment (including the **skill and thought problems**, or by clicking directly to YouTube, where they are all on Bill's "Opinionated Lessons" channel.

Watch segments BEFORE class on the indicated dates:

Mon Jan 13	**First Day of Class** (no segment due)
Wed Jan 15	Segment 1. Let's Talk about Probability (or YouTube)
Fri Jan 17	Python Set-up Tutorial and Workshop (no segment due)

Mon Jan 20	**Martin Luther King Day HOLIDAY** (no segment due)
Wed Jan 22	Segment 2. Bayes (or YouTube)
Fri Jan 24	Segment 3. Monty Hall (or YouTube)
Mon Jan 27	Segment 4. The Jailer's Tip (or YouTube)
Wed Jan 29	Segment 5. Bernoulli Trials (or YouTube)
Fri Jan 31	Segment 6. The Towne Family Tree (or YouTube)
Mon Feb 3	Segment 7. Central Tendency and Moments (or YouTube)
Wed Feb 5	Segment 8. Some Standard Distributions (or YouTube)
Fri Feb 7	Segment 9. Characteristic Functions (or YouTube)
Mon Feb 10	Segment 10. The Central Limit Theorem (or YouTube)
Wed Feb 12	Segment 11. Random Deviates (or YouTube)
Fri Feb 14	Segment 12. P-Value Tests (or YouTube)
Mon Feb 17	Segment 13. The Yeast Genome (or YouTube)
Wed Feb 19	Segment 14. Bayesian Criticism of P-Values (or YouTube)
Fri Feb 21	Segment 16. Multiple Hypotheses (or YouTube) [note order!]
Mon Feb 24	Segment 15. The Towne Family - Again (or YouTube) [note order!]
Wed Feb 26	Segment 17. The Multivariate Normal Distribution (or YouTube)
Fri Feb 28	Review Session for Mid-Term Exam (no new segment due)

Monday, March 3. MIDTERM EXAM

(Exam)   (Bill's solutions)   (Histogram of grades)

Wed Mar 5	Segment 18. The Correlation Matrix (or YouTube)
Fri Mar 7	Segment 19. The Chi Square Statistic (or YouTube)

Monday, March 10 through Friday, March 14: SPRING BREAK

Watch segments BEFORE class on the indicated dates:

Mon Mar 17	Segment 20. Nonlinear Least Squares Fitting (or YouTube)
Wed Mar 19	Segment 21. Marginalize or Condition Uninteresting Fitted Parameters (or YouTube)
Fri Mar 21	Segment 22. Uncertainty of Derived Parameters (or YouTube)
Mon Mar 24	Segment 23. Bootstrap Estimation of Uncertainty (or YouTube)
Wed Mar 26	Segment 24. Goodness of Fit (or YouTube)
Fri Mar 28	Segment 27. Mixture Models (or YouTube)
Mon Mar 31	Segment 28. Gaussian Mixture Models in 1-D (or YouTube)
Wed Apr 2	Segment 29. GMMs in N-Dimensions (or YouTube)
Fri Apr 4	Segment 30. Expectation Maximization (EM) Methods (or YouTube)

Mon Apr 7	Segment 31. A Tale of Model Selection (or YouTube)
Wed Apr 9	Segment 32. Contingency Tables: A First Look (or YouTube)
Fri Apr 11	Segment 33. Contingency Table Protocols and Exact Fisher Test (or You Tube)
Mon Apr 14	Segment 34. Permutation Tests (or YouTube)
Wed Apr 16	Segment 37. A Few Bits of Information Theory (or YouTube)
Fri Apr 18	Segment 38. Mutual Information (or YouTube)
Mon Apr 21	Segment 39. MCMC and Gibbs Sampling (or YouTube)
Wed Apr 23	Segment 40. Markov Chain Monte Carlo, Example 1 (or YouTube)
Fri Apr 25	Segment 41. Markov Chain Monte Carlo, Example 2 (or YouTube)
Mon Apr 28	Segment 47. Low-Rank Approximation of Data (or YouTube)
Wed Apr 30	Segment 48. Principal Component Analysis (PCA) (or YouTube)
Fri May 2	**Review Session for Oral Exams**

Monday, May 5 and Tuesday, May 6: ORAL FINAL EXAMS

Extra Credit Segments (segment number indicates intended sequence)

Segment 25. Fitting Models to Counts (or YouTube)
Segment 26. The Poisson Count Pitfall (or YouTube)
Segment 35. Ordinal vs. Nominal Contingency Tables (or YouTube)
Segment 36. Contingency Tables Have Nuisance Parameters (or YouTube)
Segment 49. Eigenthingies and Main Effects (or YouTube)

Segments with Slides But Not Yet Recorded

(links are to PDF files)

Segment 15.5. Poisson Processes and Order Statistics
Segment 42. Wiener Filtering
Segment 43. The IRE Lady
Segment 44. Wavelets
Segment 45. Laplace Interpolation
Segment 46. Interpolation On Scattered Data
Segment 50. Binary Classifiers
Segment 51. Hierarchical Classification
Segment 52. Dynamic Programming

Segment 1. Let's Talk about Probability

Watch this segment

The direct YouTube link is http://youtu.be/H5WjVgL6Nh4

Links to the slides: PDF file or PowerPoint file

Bill's comments on this segment

Well, I do sound nervous! This was one of my first webcasts. The production values get a little better with later segments. However, the material here is important, so be sure you understand it before going on.

Here is a link to the paper by R.T. Cox, discussed on slide 2. It's surprisingly readable for something so fundamental.

Problems

To Calculate

1. Prove that $P(ABC)=P(B)P(C|B)P(A|BC)$.

2. What is the probability that the sum of two dice is odd with neither being a 4?

To Think About

1. First-order logic is a type of propositional calculus with propositions a,b,c and quantifier symbols $\forall$ and $\exists$. This allows statements like "Socrates is a philosopher", "Socrates is a man", "There exists a philosopher who is not a man", etc. Can you use first-order logic as a calculus of inference? Is it the same as using the probability axioms? If not, then which of Cox's suppositions is violated?

2. You are an oracle that, when asked, says "yes" with probability P and "no" with probability $1-P$. How do you do this using only a fair, two-sided coin? As we did in class. Represent P as a binary number. Whenever

3. For the trout/minnow problem, what if you want to know the probability that the Nth fish caught is a trout, for N=1,2,3,... What is an efficient way to set up this calculation? (Hint: If you ever learned the word "Markov", this might be a good time to remember it!)

Class Activity

http://wpressutexas.net/coursewiki/images/9/92/ActivityWedJan15.pdf

Python Set-up Tutorial and Workshop

Motivation

Relevant XKCD

Why Python?

> **It's EASY.**
>
> Python is a well-designed high level language with a large standard library and an extensive ecosystem of 3rd party libraries.
>
> **It's READABLE.**
>
> Readability of code is important. Python is designed to force you to write readable code.
>
> It can be fun to write obscure, clever code. It feels like solving a puzzle. You should almost always fight this impulse.
>
> Collaborators (especially your most frequent collaborator - your future self) will thank you.
>
> **It's FREE.**
>
> You can use Python on any machine, anywhere, without any licensing constraints.

Why IPython Notebook server?

- Good for learning

  The user doesn't have to install anything or set anything up.
  It provides convenient access to documentation.

- Great for sharing

  It provides a centralized place to share code with each other.
  Notebooks with integrated plots and text/LaTeX annotation are a great way to tell a story.

- Data and package distribution

We can make large datasets available without you needing to download anything. We can manage the installation of any needed packages.

Activity

For most of class today, we will be using the class IPython server through a browser, but towards the end of class, it will be useful to have Python installed on your local machine.

If you are running Linux or OS/X, you almost certainly already do. Type 'python --version' at a command line to confirm.

If you are running Windows, you can download and install Python from http://www.python.org/ftp/python/2.7.6/python-2.7.6.msi

Go to http://rosalind.info and create an account. (A convenient way to do this is to click 'Log in' and use an OpenID like a Google account.)

If you are have never used Python before or just want some review

Start with Rosalind's Python tutorial excercises:

1. Installing Python - We need to do this problem because the website won't give us access to the rest until we do. Ignore the contents of the 'click to expand' box for this one.
2. Variables and Some Arithmetic
3. Strings and Lists
4. Conditions and Loops
5. Dictionaries - Notice that we skipped 'Working with Files' for now.

If you are already comfortable in Python

Jump straight into some basic bioinformatics:

1. Counting DNA Nucleotides
2. Transcribing DNA into RNA
3. Complementing a Strand of DNA
4. Counting Point Mutations - Use the built-in function zip() in your code.
5. Finding a Motif in DNA - For extra challenge: use the re module. How do you deal with overlapping motifs?
6. Translating RNA into Protein

   Hints for 6
   Use the Biopython module, which is already installed on the class server.

For guidance, execute "from Bio.Seq import Seq", then "Seq?" and "Seq.translate?".

The remaining problems are best done using Python installed on your local machine instead of through the browser.

7. Enumerating Gene Orders - Use a function from the itertools module in your code.
8. Computing GC Content
9. Rosalind: Finding a Protein Motif

   Hints for 8
   In your 'Computing GC Content' code, separate the FASTA parsing functionality into a function. Import this file into your current code as a module and use it to do your fasta parsing. Some guidance: modules and __name__ == 'main' idiom.

   Use the urllib module to read data from the web.

   Use the re module to do the motif finding. (Note: some approaches to this will have to take extra care to make sure overlapping motifs can be found.)

Jeff's Solutions

Note: these are a non-interactive snapshot of notebooks that you can also find on the class server.

- The Python tutorial exercises
- The basic bioinformatics problems

Segment 2. Bayes

Watch this segment

The direct YouTube link is http://youtu.be/FROAk4AFKHk

Links to the slides: PDF file or PowerPoint file

Bill's Comments

Here is a link to the Efron paper mentioned.

Problems

To Calculate

1. If the knight had captured a Gnome instead of a Troll, what would his chances be of crossing safely?

2. Suppose that we have two identical boxes, A and B. A contains 5 red balls and 3 blue balls. B contains 2 red balls and 4 blue balls. A box is selected at random and exactly one ball is drawn from the box. What is the probability that it is blue? If it *is* blue, what is the probability that it came from box B?

To Think About

1. Do you think that the human brain's intuitive "inference engine" obeys the commutativity and associativity of evidence? For example, are we more likely to be swayed by recent, rather than older, evidence? How can evolution get this wrong if the mathematical formulation is correct?

2. How would you simulate the Knight/Troll/Gnome problem on a computer, so that you could run it 100,000 times and see if the Knights probability of crossing safely converges to 1/3?

3. Since different observers have different background information, isn't Bayesian inference useless for making social decisions (like what to do about climate change, for example)? How can there ever be any consensus on probabilities that are fundamentally subjective?

Class Activity

Media:ActivityWedJan22.pdf

Jeff's gnomes and trolls simulation

Segment 3. Monty Hall

Watch this segment

The direct YouTube link is http://youtu.be/Rxb8JG8nUFA

Links to the slides: PDF file or PowerPoint file

Bill's Comments

You might enjoy reading some of the correspondence that Marilyn vos Savant received, on her web site.

Problems

To Calculate

1. The slides used a symmetry argument ("relabeling") to simplify the calculation. Redo the calculation without any such relabeling. Assume that the doors have big numbers "1", "2", and "3" nailed onto them, and consider all possibilities. Do you still have to make an assumption about Monty's preferences (where the slide assumed 1/2)?

To Think About

1. Lawyers are supposed to be able to argue either side of a case. What is the best argument that you can make that switching doors can't possibly make any difference? In other words, how cleverly can you hide some wrong assumption?

2. We stated the problem as *requiring* the host to offer the contestant a chance to switch. But what if the host can offer that chance, or not, as he sees fit? Then, when offered the chance, should you still switch? (Spoiler alert: see this New York Times interview with Monte Hall.)

3. Mr. and Mrs. Smith tell you that they have two children, one of whom is a girl.
(a) What is the probability that the other child is a girl?
Mr. Smith then shows you a photo of his children on his iPhone. One is clearly a girl, but the other one's face is hidden behind the family dog, and you can't tell their gender.
(b) What is the probability that the hidden child is a girl?
(c) If your answers to (a) and (b) are different, explain why there is a difference.

Class Activity

Class got cancelled due to a snow day (very unusual in Austin!). We would have done:

Generalized Monty Hall

Segment 4. The Jailer's Tip

Watch this segment

The direct YouTube link is http://youtu.be/425D0CjLLLs

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Evaluate $\int 10 \delta(3x-2)dx$

2. Prove that $\delta(ax)=1a\delta(x)$.

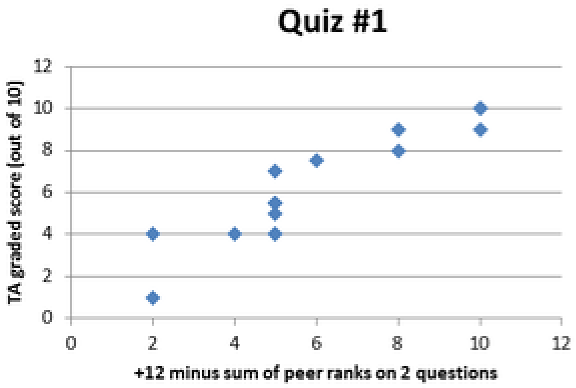
3. What is the numerical value of $P(A|S_B I)$ if the prior for $p(x)$ is a massed prior with half the mass at $x=1/3$ and half the mass at $x=2/3$?

To Think About

1. With respect to problem 3, above, since x is a probability, how can choosing x=1/3 half the time, and x=2/3 the other half of the time be different from choosing x=1/2 all the time?

2. Suppose A is some event that we view as stochastic with P(A), such as "will it rain today?". But the laws of physics (or meteorology) say that A actually depends on other weather variables X, Y, Z, etc., with conditional probabilities P(A|XYZ...). If we repeatedly sample just A, to naively measure P(A), are we correctly marginalizing over the other variables?

Class Activities



Comparison of peer scores to TA grades on quiz

Surprise Quiz (with Bill's solutions here) (Notice in the figure the almost perfect correlation between the peer ranks that the teams assigned and the TA's separate grading.)

We also discussed Mr. and Mrs. Smith and their daughter(s) -- see Think About Question 3 in Segment 3.

We also did some variants of Expected values and continuous distributions

Segment 5. Bernoulli Trials

Watch this segment

The direct YouTube link is http://youtu.be/2T3KP2LleFg

Links to the slides: PDF file or PowerPoint file

Problems

To Compute

1. You throw a pair of fair dice 10 times and, each time, you record the total number of spots. When you are done, what is the probability that exactly 5 of the 10 recorded totals are prime?

2. If you flip a fair coin one billion times, what is the probability that the number of heads is between 500010000 and 500020000, inclusive? (Give answer to 4 significant figures.)

To Think About

1. Suppose that the assumption of independence (the first "i" in "i.i.d.") were violated. Specifically suppose that, after the first Bernoulli trial, every trial has a probability Q of simply reproducing the immediately previous outcome, and a probability (1-Q) of being

an independent trial. How would you compute the probability of getting n events in N trials if the probability of each event (when it is independent) is p?

2. Try the Mathematica calculation on slide 5 without the magical "GenerateConditions -> False". Why is the output different?

Class Activity

http://projecteuler.net/problem=267

Part 2: The problem as stated multiplies your wager by 2 on a win. What is the smallest this factor can be while still leaving the probability of ending up above one billion greater than 0.5, assuming that you play with an optimal f given the value of the factor?

Jeff's solution - updated in response to group 2's excellent work

Segment 6. The Towne Family Tree

Watch this segment

The direct YouTube link is http://youtu.be/y_L2THpv5Jg

Links to the slides: PDF file or PowerPoint file

Problems

To Compute

1. Write down an explicit expression for what the slides denote as bin(n,N,r).

2. There is a small error on slide 7 that carries through to the first equation on slide 8 and the graph on slide 9. Find the error, fix it, and redo the graph of slide 9. Does it make a big difference? Why or why not?

To Think About

1. Suppose you knew the value of r (say, r = 0.0038). How would you simulate many instances of the Towne family data (e.g., the tables on slides 4 and 5?

2. How would you use your simulation to decide if the assumption of ignoring backmutations (the red note on slide 7) is justified?

3. How would you use your simulation to decide if our decision to trim T2, T11, and T13 from the estimation of r was justified? (This question anticipates several later discussions in the course, but thinking about it now will be a good start.)

Class Activity

Multinomial parameter estimation

Jeff's solution

Segment 7. Central Tendency and Moments

Watch this segment

The direct YouTube link is http://youtu.be/ZWOmsKWQ7Fw

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Prove the result of slide 3 the "mechanical way" by setting the derivative of something equal to zero, and solving.

2. Give an example of a function $p(x)$, with a maximum at $x=0$, whose third moment M_3 exists, but whose fourth moment M_4 doesn't exist.

3. List some good and bad things about using the median instead of the mean for summarizing a distribution's central value.

To Think About

1. This segment assumed that $p(x)$ is a known probability distribution. But what if you know $p(x)$ only experimentally. That is, you can draw random values of x from the distribution. How would you estimate its moments?

2. High moments (e.g., 4 or higher) are algebraically pretty, but they are rarely useful because they are very hard to measure accurately in experimental data. Why is this true?

3. Even knowing that it is useless, how would you find the formula for I_8 , the eighth semi-invariant?

Class Activity

Activity20140203

Segment 8. Some Standard Distributions

Watch this segment

The direct YouTube link is http://youtu.be/EDYDC7iNGTg Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. In Segment 6 (slide 8) we used the improper prior $1/r$. Show that this is just a limiting case of a (completely proper) Lognormal prior.

2. Prove that $\text{Gamma}(\alpha,\beta)$ has a single mode at $(\alpha-1)/\beta$ when $\alpha\geq1$.

3. Show that the limiting case of the Student distribution as $\nu\to\infty$ is the Normal distribution.

To Think About

1. Suppose you have an algorithm that can compute a CDF, $P(x)$. How would you design an algorithm to compute its inverse (see slide 9) $x(P)$?

2. The lifetime t of a radioactive nucleus (say Uranium 238) is distributed as the Exponential distribution. Do you know why? (Hint: What is the distribution of an Exponential(β) random variable *conditioned on* its being greater than some given value?)

Class Activity

problem statement: ClassActivity20130204.pdf

data file for class activity: events20130204.txt

Jeff's solution (Python)

Bill's solution (MATLAB)

Segment 9. Characteristic Functions

Watch this segment

The direct YouTube link is http://youtu.be/NJL-BX6HuxY

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Use characteristic functions to show that the sum of two independent Gaussian random variables is itself a Gaussian random variable. What is its mean and variance?

2. Calculate (don't just look up) the characteristic function of the Exponential distribution.

To Think About

1. Learn enough about contour integration to be able to make sense of Saul's explanation at the bottom of slide 7. Then draw a picture of the contours, label the pole(s), and show how you calculate their residues.

2. Do you think that characteristic functions are ever useful computationally (that is, not just analytically to prove theorems)?

Class Activity

Probability blitz

Alternate to problem 6(a): Show that a process with constant-rate exponentially distributed waiting times between events is a Poisson process (that has a Poisson distribution for the number of events in any fixed time interval).

Segment 10. The Central Limit Theorem

Watch this segment

The direct YouTube link is [http://youtu.be/IpuYGsKplSw http://youtu.be/IpuYGsKplSw

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Take 12 random values, each uniform between 0 and 1. Add them up and subtract 6. Prove that the result is close to a random value drawn from the Normal distribution with mean zero and standard deviation 1.

2. Invent a family of functions, each different, that look like those in Slide 3: they all have value 1 at x = 0; they all have zero derivative at x = 0; and they generally (not necessarily monotonically) decrease to zero at large x. Now multiply 10 of them together and graph the result near the origin (i.e., reproduce what Slide 3 was sketching).

3. For what value(s) of ν does the Student distribution (Segment 8, Slide 4) have a convergent 1st and 2nd moment, but divergent 3rd and higher moments?

To Think About

1. A distribution with moments as in problem 3 above has a well-defined mean and variance. Does the CLT hold for the sum of RVs from such a distribution? If not, what goes wrong in the proof? Is the mean of the sum equal to the sum of the individual means? What about the variance of the sum? What, qualitatively, does the distribution of the sum of a bunch of them look like?

2. Give an explanation of Bessel's correction in the last expression on slide 5. If, as we see, the MAP calculation gives the factor 1/N, why would one ever want to use 1/(N-1) instead? (There are various wiki and stackoverflow pages on this. See if they make sense to you!)

Just for fun

A fun problem that ties in to 'To Calculate' 1 above and problem 6 from the Probability Blitz:

1. What is the expected number of Uniform[0,1] draws you need to add up before the sum exceeds 1? Prove your answer analytically and confirm it by simulation.

Jeff's solution

Segment 11. Random Deviates

Watch this segment

The direct YouTube link is http://youtu.be/4r1GlyisB8E

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. For the Cauchy distribution (Segment 8, Slide 3), find the inverse function of the CDF.

2. In your favorite programming language, write a function that returns independent Cauchy deviates.

To Think About

1. Suppose you want a function that returns deviates for Student(ν) . Could you use the Cauchy pdf (or some scaling of it) as a bounding function in a rejection method? How efficient is this (i.e., what fraction of the time does it reject)?

2. Explain the three inequality tests in the "while" statement in Leva's algorithm (slide 7) and why they are hooked together with logical operators in the way shown.

Class Activity

Build your own random number generator

Segment 12. P-Value Tests

Watch this segment

The direct YouTube link is http://youtu.be/2Ul7TI0B5ek

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. What is the critical region for a 5% two-sided test if, under the null hypothesis, the test statistic is distributed as Student$(0,\sigma,4)$? That is, what values of the test statistic disprove the null hypothesis with p < 0.05? (OK to use Python, MATLAB, or Mathematica.)

2. For an exponentially distributed test statistic with mean μ (under the null hypothesis), when is the the null hypothesis disproved with p < 0.01 for a one-sided test? for a two-sided test?

To Think About

1. P-value tests require an initial choice of a test statistic. What goes wrong if you choose a poor test statistic? What would make it poor?

2. If the null hypothesis is that a coin is fair, and you record the results of N flips, what is a good test statistic? Are there any other possible test statistics?

3. Why is it so hard for a Bayesian to do something as simple as, given some data, disproving a null hypothesis? Can't she just compute a Bayes odds ratio, P(null hypothesis is true)/P(null hypothesis is false) and derive a probability that the null hypothesis is true?

Segment 13. The Yeast Genome

Watch this segment

The direct YouTube link is http://youtu.be/QSgUX-Do8Tc

Links to the slides: PDF file or PowerPoint file

Link to the file mentioned in the segment: SacSerChr4.txt

Link to all yeast chromosomes: UCSC

Problems

To Calculate

1. With p=0.3, and various values of n, how big is the largest discrepancy between the Binomial probability pdf and the approximating Normal pdf? At what value of n does this value become smaller than 10^{-15} ?

2. Show that if four random variables are (together) multinomially distributed, each separately is binomially distributed.

To Think About

1. The segment suggests that $A \neq T$ and $C \neq G$ comes about because genes are randomly distributed on one strand or the other. Could you use the observed discrepancies to estimate, even roughly, the number of genes in the yeast genome? If so, how? If not, why not?

2. Suppose that a Bayesian thinks that the prior probability of the hypothesis that "$P_A = P_T$ " is 0.9, and that the set of all hypotheses that "$P_A \neq P_T$ " have a total prior of 0.1. How

might he calculate the odds ratio $\text{Prob}(P_A = P_T)/\text{Prob}(P_A \neq P_T)$? Hint: Are there nuisance variables to be marginalized over?

Class Activity

Yeast chromosome 4

Activity slides

Segment 14. Bayesian Criticism of P-Values

Watch this segment

The direct YouTube link is http://youtu.be/IKV6Pn18C7o

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Suppose the stopping rule is "flip exactly 10 times" and the data is that 8 out of 10 flips are heads. With what p-value can you rule out the hypothesis that the coin is fair? Is this statistically significant?

2. Suppose that, as a Bayesian, you see 10 flips of which 8 are heads. Also suppose that your prior for the coin being fair is 0.75. What is the posterior probability that the coin is fair? (Make any other reasonable assumptions about your prior as necessary.)

3. For the experiment in the segment, what if the stopping rule was (perversely) "flip until I see five consecutive heads followed immediately by a tail, then count the total number of heads"? What would be the p-value?

To Think About

1. If biology journals require $p<0.05$ for results to be published, does this mean that one in twenty biology results are wrong (in the sense that the uninteresting null hypothesis is actually true rather than disproved)? Why might it be worse, or better, than this? (See also the provocative paper by Ioannidis, and this blog in Technology Review (whose main source is this article). Also this news story about ESP research. You can Google for other interesting references.)

Segment 15. The Towne Family - Again

Watch this segment

The direct YouTube link is http://youtu.be/Y-i0CN15X-M

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. In slide 4, we used "posterior predictive p-value" to get the respective p-values 1.0e-13, .01, .12, and .0013. What if we had mistakenly just used the maximum likelihood estimate r=0.003, instead of integrating over r? What p-values would we have obtained?

To Think About

1. Can you think of a unified way to handle the Towne family problem (estimating r and deciding which family members are likely "non-paternal") without trimming the data? We'll show one such method in a later segment, but there is likely more than one possible good answer.

Class Activity

We divided into three teams. Each team prepared a single solution set for last year's surprise quiz of around this date.

Here are the three solutions:

Team 1

Group Two: The Towne Family - Again, Class Activity

Team 3 scanned PDF

Every class member gets to vote for TWO of these for which is best, your own team and one other. You must vote for two, not just 1. Please edit this page to add your (screen) name to two of the following lists:

Team 1 votes: Vsub, Jonathan, Daniel, Aaron, Deepesh, Sanmit, Todd, Nick, Eleisha, Andrea, Rene

Team 2 votes: Todd, Eleisha, Elad

Team 3 votes: Vsub, Jonathan, Daniel, Aaron, Deepesh, Sanmit, Nick, Andrea, Elad, Rene

Here is Bill's solution set from last year. (I wasn't trying to be as complete or neat as I expect this year's teams to be.) Solutions

Voting Comments

Vsub: Team-1: 8/9 (lacking in explanations). Team-2: 6.5/9 (lacks explanation, nice plots, too concise on 6,7) Team-3: 7/9 (incomplete ans3; incorrect answers 6,7; good detailed derivations). Vote preferenceorder: Team1 > Team3 > Team2

Segment 16. Multiple Hypotheses

Watch this segment

The direct YouTube link is http://youtu.be/w6AjduOEN2k

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Simulate the following: You have M=50 p-values, none actually causal, so that they are drawn from a uniform distribution. Not knowing this sad fact, you apply the Benjamini-Hochberg prescription with $\alpha=0.05$ and possibly call some discoveries as true. By repeated simulation, estimate the probability of thus getting N wrongly-called discoveries, for N=0, 1, 2, and 3.

2. Does the distribution that you found in problem 1 depend on M? On α ? Derive its form analytically for the usual case of $\alpha \ll 1$?

To Think About

1. Suppose you have M independent trials of an experiment, each of which yields an independent p-value. Fisher proposed combining them by forming the statistic

$$S=-2\sum_{i=0}^{i=M}\log(p_i)$$

Show that, under the null hypothesis, S is distributed as Chisquare$(2M)$ and describe how you would obtain a combined p-value for this statistic.

2. Fisher is sometimes credited, on the basis of problem 1, with having invented "meta-analysis", whereby results from multiple investigations can be combined to get an overall more significant result. Can you see any pitfalls in this?

Class Activity

P-value follow-ups

- Team 1 - Feb 21 Activity
- [Team Girls + Sanmit - Feb 21 Activity]
- Team3-021714part2
- Feb20-Team4-P-value follow up

Here is John's written up solution: Pvalue Examples Solutions.

Segment 17. The Multivariate Normal Distribution

Watch this segment

The direct YouTube link is http://youtu.be/t7Z1a_BOkN4

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Calculate the Jacobian determinant of the transformation of variables defined by

$$y_1 = x_1/x_2, y_2 = x_2^2$$

2. Consider the 3-dimensional multivariate normal over (x_1, x_2, x_3) with $\mu = (-1, -1, -1)$ and

$\Sigma^{-1} = ((\ 5\ \ -1\ \ 2\ \ //\ \ -1\ \ \ 8\ \ \ 1 //\ \ \ \ 2\ \ 1\ \ 4\ ))$. (Note the matrix inverse notation.)

What are 2-dimensional μ and Σ^{-1} for

(a) the distribution on the slice $x_3 = 0$?

(b) the marginalization over x_3 ?

Hint: The answers are all simple rationals, but I had to use Mathematica to work them out.

To Think About

1. Prove the assertions in slide 5. That is, implement the ideas in the blue text.

2. How would you plot an error ellipsoid in 3 dimensions? That is, what would be the 3-dimensional version of the code in slide 8? (You can assume the plotting capabilities of your favorite programming language.)

Class Activity

Some 3x3 Matrices

MVN Exercise

Bill's Mathematica notebook for problem 2 (above). (Download file, rename as MultivarGaussExample.nb, then open in Mathematica.)

Review Session for Mid-Term Exam

Generalized Monty Hall

Probability blitz

MVN Exercise

Solutions to generalized Monty Hall and blitz

Solutions to MVN Exercise

Monday, March 3. MIDTERM EXAM

(Exam) (Bill's solutions) (Histogram of grades)

Segment 18. The Correlation Matrix

Watch this segment

The direct YouTube link is http://youtu.be/aW5q_P0it9E

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Random points i are chosen uniformly on a circle of radius 1, and their (x_i, y_i) coordinates in the plane are recorded. What is the 2x2 covariance matrix of the random variables X and Y ? (Hint: Transform probabilities from θ to x . Second hint: Is there a symmetry argument that some components must be zero, or must be equal?)

2. Points are generated in 3 dimensions by this prescription: Choose λ uniformly random in $(0,1)$. Then a point's (x,y,z) coordinates are $(\alpha\lambda,\beta\lambda,\gamma\lambda)$. What is the covariance matrix of the random variables (X,Y,Z) in terms of $\alpha,\beta,$ and γ ? What is the linear correlation matrix of the same random variables?

To Think About

1. Suppose you want to get a feel for what a linear correlation $r=0.3$ (say) looks like. How would you generate a bunch of points in the plane with this value of r ? Try it. Then try for different values of r . As r increases from zero, what is the smallest value where you would subjectively say "if I know one of the variables, I pretty much know the value of the other"?

2. Suppose that points in the (x,y) plane fall roughly on a 45-degree line between the points (0,0) and (10,10), but in a band of about width w (in these same units). What, roughly, is the linear correlation coefficient r ?

Class Activity

Class Activity 3/5/14

Segment 19. The Chi Square Statistic

Watch this segment

The direct YouTube link is http://youtu.be/87EMhmPkOhk

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Prove the assertion on lecture slide 5, namely that, for a multivariate normal distribution, the quantity $(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)$, where $\mathbf{x}$ is a random draw from the multivariate normal, is χ^2 distributed.

To Think About

1. Why are we so interested in t-values? Why do we square them?

2. Suppose you measure a bunch of quantities x_i , each of which is measured with a measurement accuracy σ_i and has a theoretically expected value μ_i . Describe in detail how you might use a chi-square test statistic as a p-value test to see if your theory is viable? Should your test be 1 or 2 tailed?

Class Exercise

Class Exercise

Data file: Media:mv_chi.txt

Segment 20. Nonlinear Least Squares Fitting

Watch this segment

The direct YouTube link is http://youtu.be/xtBCGPHRcb0

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. (See lecture slide 3.) For one-dimensional x , the model $y(x|\mathbf{b})$ is called "linear" if $y(x|\mathbf{b})=\sum_k b_k X_k(x)$, where $X_k(x)$ are arbitrary known functions of x . Show that minimizing χ^2 produces a set of linear equations (called the "normal equations") for the parameters b_k .

2. A simple example of a linear model is $y(x|\mathbf{b})=b_0+b_1 x$, which corresponds to fitting a straight line to data. What are the MLE estimates of b_0 and b_1 in terms of the data: x_i 's, y_i 's, and σ_i 's?

To Think About

1. We often rather casually assume a uniform prior $P(\mathbf{b})$=constant on the parameters $\mathbf{b}$. If the prior is not uniform, then is minimizing $\chi 2$ the right thing to do? If not, then what should you do instead? Can you think of a situation where the difference would be important?

2. What if, in lecture slide 2, the measurement errors were $e_i \sim \text{Cauchy}(0,\sigma_i)$ instead of $e_i \sim N(0,\sigma_i)$? How would you find MLE estimates for the parameters $\mathbf{b}$?

Class Activity

Here is some data: Media:Chisqfitdata.txt

In class we will work on fitting this to some models as explained here.

Here are Bill's numerical answers, so that you can see whether you are on the right track (or whether Bill got it wrong!): Media:Chisqfitanswers.txt

Segment 21. Marginalize or Condition Uninteresting Fitted Parameters

Watch this segment

The direct YouTube link is http://youtu.be/yxZUS_BpEZk

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Consider a 2-dimensional multivariate normal distribution of the random variable (b_1,b_2) with 2-vector mean (μ_1,μ_2) and 2x2 matrix covariance Σ . What is the distribution of b_1 given that b_2 has the particular value b_c ? In particular, what is the mean and standard deviation of the conditional distribution of b_1 ? (Hint, either see Wikipedia "Multivariate normal distribution" for the general case, or else just work out this special case.)

2. Same, but marginalize over b_2 instead of conditioning on it.

To Think About

1. Why should it be called the Fisher *Information* Matrix? What does it have to do with "information"?

2. Go read (e.g., in Wikipedia or elsewhere) about the "Cramer-Rao bound" and be prepared to explain what it is, and what it has to do with the Fisher Information Matrix.

Class Activity

Today we'll do Find the Volcano.

Segment 22. Uncertainty of Derived Parameters

Watch this segment

The direct YouTube link is http://youtu.be/ZoD3_rov--w

Links to the slides: PDF file or PowerPoint file

Problems

To Compute

1. In lecture slide 3, suppose (for some perverse reason) we were interested in a quantity $f = b_3/b_5$ instead of $f = b_3 b_5$. Calculate a numerical estimate of this new f and its standard error.

2. Same set up, but plot a histogram of the distribution of f by sampling from its posterior distribution (using Python, MATLAB, or any other platform).

To Think About

1. Lecture slide 2 asserts that a function of normally distributed RVs is not, in general, normal. Consider the product of two independent normals. Is it normal? No! But isn't the product of two normal distribution functions (Gaussians) itself Gaussian? So, what is going on?

2. Can you invent a function of a single normal N(0,1) random variable whose distribution has two separate peaks (maxima)? How about three? How about ten?

Segment 23. Bootstrap Estimation of Uncertainty

Watch this segment

The direct YouTube link is http://youtu.be/1OC9ul-1PVg

Links to the slides: PDF file or PowerPoint file

Problems

To Compute

1. Generate 100 i.i.d. random draws from the beta distribution Beta$(2.5,5.)$, for example using MATLAB's betarnd or Python's random.betavariate. Use these to estimate this statistic of the underlying distribution: "value of the 75% percentile point minus value of the 25th percentile point". Now use statistical bootstrap to estimate the distribution of uncertainty of your estimate, for example as a histogram.

2. Suppose instead that you can draw any number of desired samples (each 100 draws) from the distribution. How does the histogram of the desired statistic from these samples compare with the bootstrap histogram from problem 1?

3. What is the actual value of the desired statistic for this beta distribution, computed numerically (that is, not by random sampling)? (Hint: I did this in Mathematica in three lines.)

To Think About

1. Suppose your desired statistic (for a sample of N i.i.d. data values) was "minimum of the N values". What would the bootstrap estimate of the uncertainty look like in this case? Does this violate the bootstrap theorem? Why or why not?

2. If you knew the distribution, how would you compute the actual distribution for the statistic "minimum of N sampled values", not using random sampling in your computation?

3. For N data points, can you design a statistic so perverse (and different from one suggested in the segment) that the statistical bootstrap fails, even asymptotically as N becomes large?

Class Activity

Download the data set given below. It contains 100 draws from a 4 dimensional distribution i.e. each draw returns a 4 dimensional vector, $[x_1,x_2,x_3,x_4]$. The statistic which we are interested in is,

$$t=[\langle x_1 x_2\rangle,\langle x_3 x_4\rangle].$$

Carry out the following tasks:

- Give a point estimate of the statistic.
- Carry out bootstrapping and visualize the uncertainty in the statistic using a scatter plot.

Data Set

Dataset_txtfile

Segment 24. Goodness of Fit

Watch this segment

The direct YouTube link is http://youtu.be/EJleSVf0Z-U

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Let X be an R.V. that is a linear combination (with known, fixed coefficients α_k) of twenty $N(0,1)$ deviates. That is, $X=\sum_{k=1}^{20}\alpha_k T_k$ where $T_k \sim N(0,1)$. How can you most simply form a t-value-squared (that is, something distributed as Chisquare(1) from X ? For some particular choice of α_k 's (random is ok), generate a sample of x 's, plot their histogram, and show that it agrees with Chisquare(1) .

2. From some matrix of known coefficients α_{ik} with $k=1,\ldots,20$ and $i=1,\ldots,100$, generate 100 R.V.s $X_i=\sum_{k=1}^{20}\alpha_{ik}T_k$ where $T_k \sim N(0,1)$. In other words, you are expanding 20 i.i.d. T_k 's into 100 R.V.'s. Form a sum of 100 t-values-squareds obtained from these variables and demonstrate numerically by repeated sampling that it is distributed as Chisquare(ν) ? What is the value of ν ? Use enough samples so that you could distinguish between ν and $\nu-1$.

3. Reproduce the table of critical $\Delta\chi^2$ values shown in slide 7. Hint: Go back to segment 21 and listen to the exposition of slide 7. (My solution is 3 lines in Mathematica.)

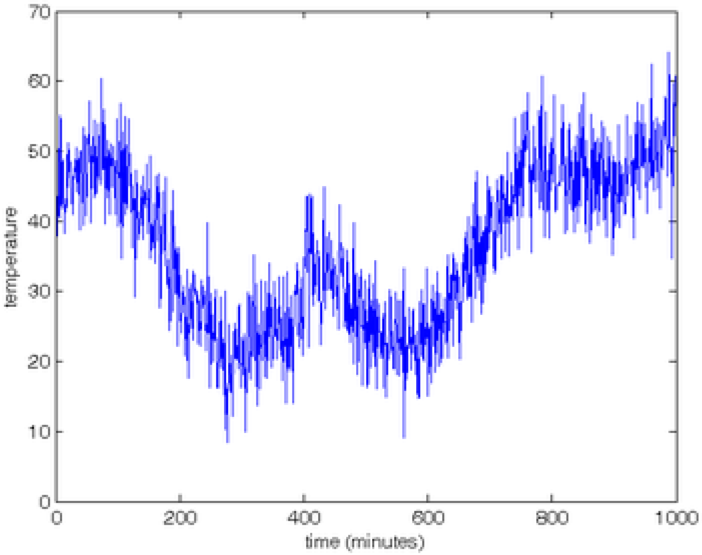
To Think About

1. Design a numerical experiment to exemplify the assertions on slide 8, namely that $\chi 2min$ varies by $\pm 2\nu--\surd$ from data set to data set, but varies only by $\pm O(1)$ as the fitted parameters $\mathbf{b}$ vary within their statistical uncertainty?

2. Suppose you want to estimate the central value μ of a sample of N values drawn from Cauchy(μ,σ) . If your estimate is the mean of your sample, does the "universal rule of thumb" (slide 2) hold? That is, does the accuracy get better as $N-1/2$? Why or why not? What if you use the median of your sample as the estimate? Verify your answers by numerical experiments.

Class Activity

I measured the temperature of my framitron manifold every minute for 1000 minutes, with the same accuracy, $\sigma=5$, for each measurement. The data is plotted below (with data points connected by straight lines), and is in the file Modelselection1.txt.



It's a contest! Which group can write down a model $T(t|\mathbf{b})$, where $\mathbf{b}$ is a vector of parameters, that gives the best fit to the data in a least squares sense.

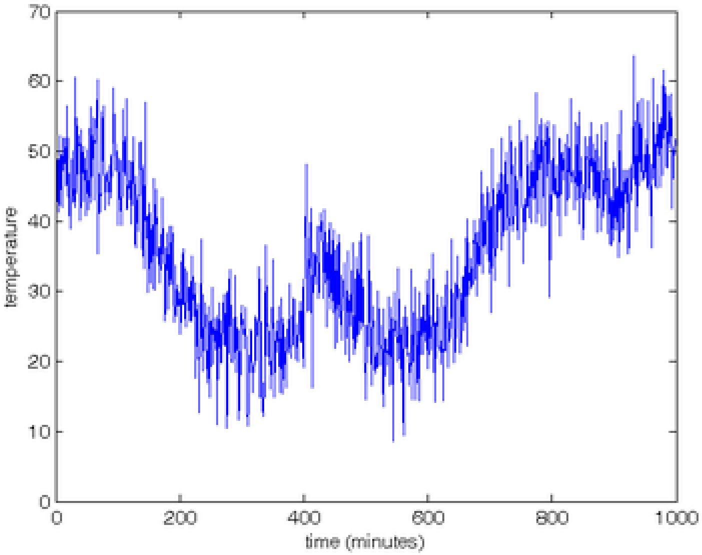
Part 1. Any number of parameters in $\mathbf{b}$ are allowed.

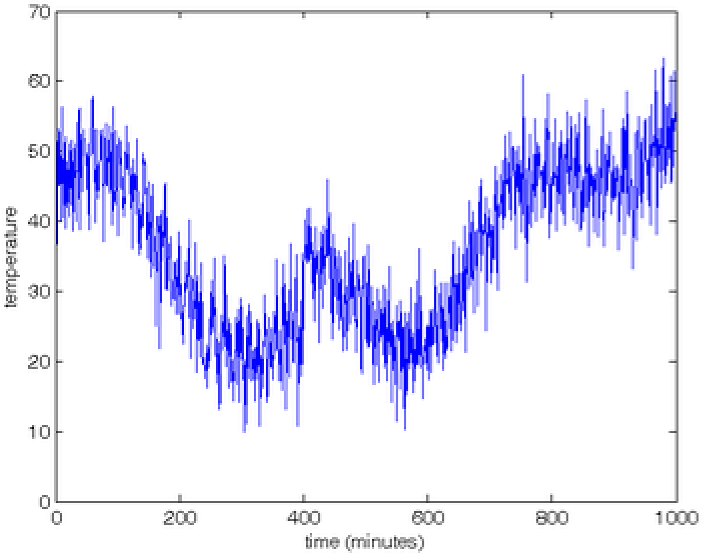
Part 2. At most 20 parameters are allowed.

Part 3. At most 10 parameters are allowed.

Part 4. At most 4 parameters are allowed.

And, oh by the way, we'll actually test your model on a different realization of the same process, possibly one of the ones shown below.





Modelselection2.txt

Modelselection3.txt

Modelselection4.txt

Modelselection5.txt

Modelselection6.txt

Segment 27. Mixture Models

Watch this segment

The direct YouTube link is http://youtu.be/9pWnZcpYh44

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

The file Media:Mixturevals.txt contains 1000 values, each drawn either with probability c from the distribution Exponential(β) (for some constant β), or otherwise (with probability $1-c$) from the distribution $p(x)=(2/\pi)/(1+x^2), x>0$.

1. Write down an expression for the probability of the file's data given some values for the parameters β and c .

2. Calculate numerically the maximum likelihood values of β and c .

3. Estimate numerically the Bayes posterior distribution of β , marginalizing over c as a nuisance parameter. (You'll of course have to make some assumption about priors.)

To Think About

1. In problem 3, above, you assumed some definite prior for c . What if c is itself drawn (just once for the whole data set) from a distribution Beta(μ,ν) , with unknown hyperparameters μ,ν . How would you now estimate the Bayes posterior distribution of β , marginalizing over everything else?

Segment 28. Gaussian Mixture Models in 1-D

Watch this segment

The direct YouTube link is http://youtu.be/n7u_tq0I6jM

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Draw a sample of 100 points from the uniform distribution $U(0,1)$. This is your data set. Fit GMM models to your sample (now considered as being on the interval $-\infty<x<\infty$) with increasing numbers of components K , at least $K=1,\ldots,5$. Plot your models. Do they get better as K increases? Did you try multiple starting values to find the best (hopefully globally best) solutions for each K ?

2. Multiplying a lot of individual likelihoods will often underflow. (a) On average, how many values drawn from $U(0,1)$ can you multiply before the product underflows to zero? (b) What, analytically, is the distribution of the sum of N independent values $\log(U)$, where $U\sim U(0,1)$? (c) Is your answer to (a) consistent with your answer to (b)?

To Think About

1. Suppose you want to approximate some analytically known function $f(x)$ (whose integral is finite), as a sum of K Gaussians with different centers and widths. You could pretend that $f(x)$ (or some scaling of it) was a probability distribution, draw N points from it and do the GMM thing to find the approximating Gaussians. Now take the limit $N\rightarrow\infty$, figure out how sums become integrals, and write down an iterative method for fitting Gaussians to a given $f(x)$. Does it work? (You can assume that well-defined definite integrals can be done numerically.)

Class Activity

Let's explore a data set and try to make sensible statements about it.

netflixishdata.txt

Rows are 200 movie watchers, columns are 100 movies, entries are their ratings on a scale of 1 (I hated it!) to 5 (I loved it!). This is not real data, of course, so it is only Netflixish, not Netflix.

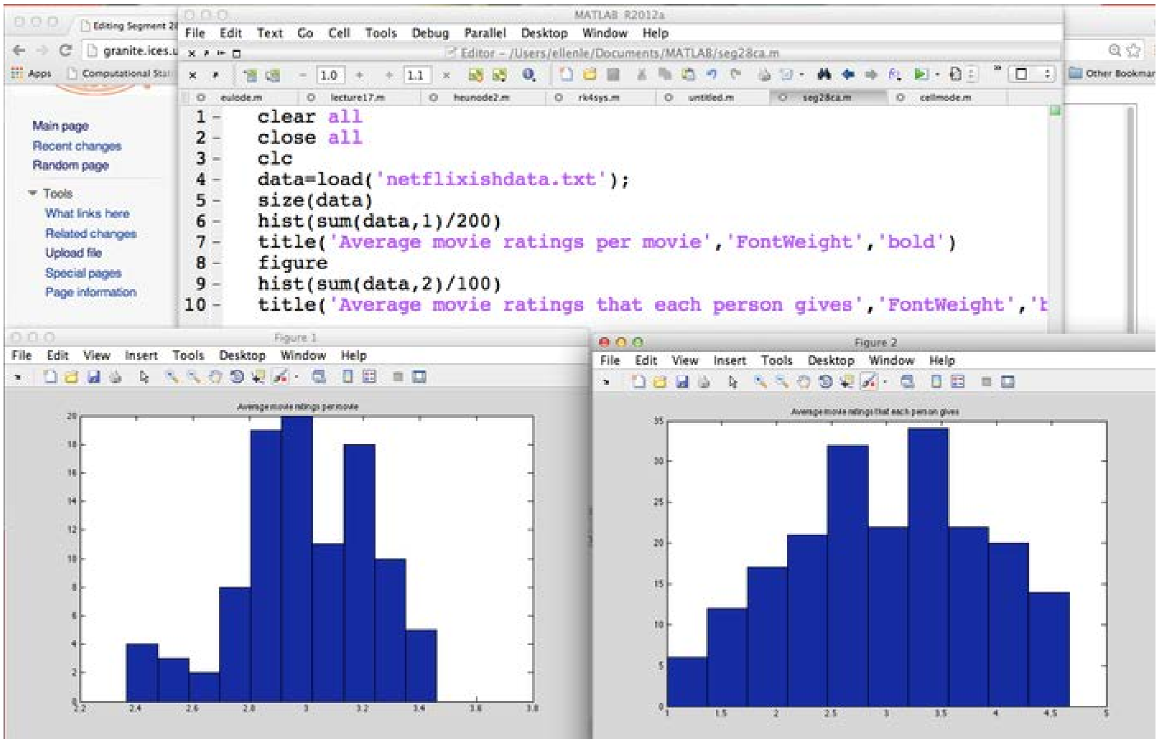
Questions to explore

How much are people alike?
How much are movies alike?
Distribution of the data in various ways?

By summing over all the columns and dividing by number of entires, we got the average rating for each movie. Something surprising was that the max of all the mean ratings was

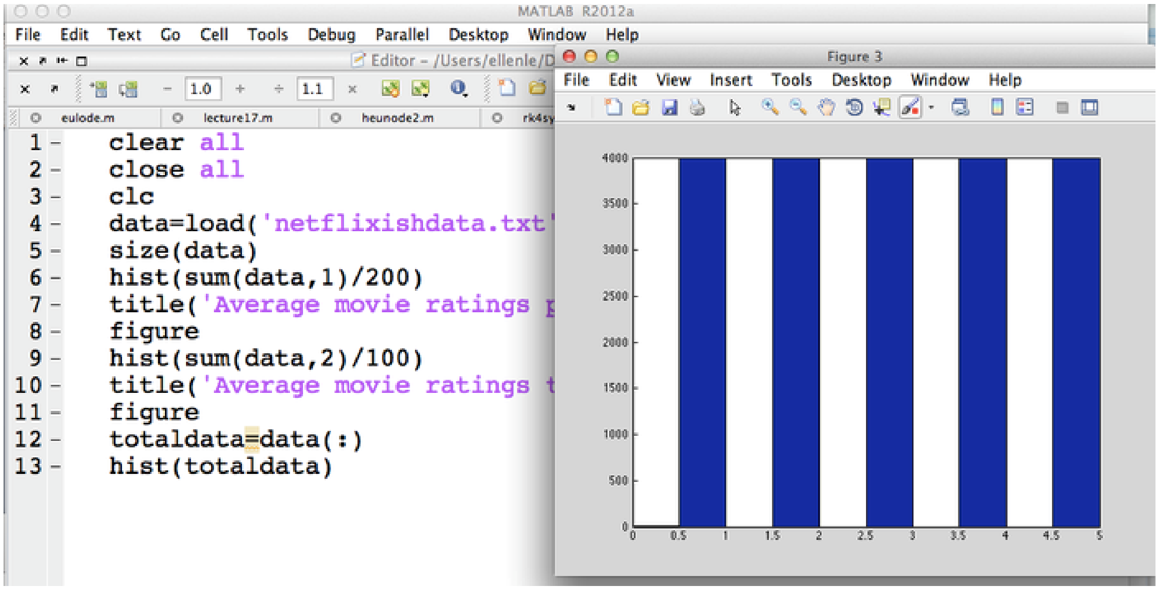
3.46 (so there was no mean rating greater than 3.46 stars!), the min was 2.3650, the mean of all the mean movie ratings is 2.9998, the median of these mean ratings was 2.9925.

Insight #1:
Looking at the actual data set, we see that there are a lot of "haters" i.e. there are a lot of people who gave a lot of 1 ratings.



```
clear all
close all
clc
data=load('netflixishdata.txt');
size(data)
hist(sum(data,1)/200)
title('Average movie ratings per movie','FontWeight','bold')
figure
hist(sum(data,2)/100)
title('Average movie ratings that each person gives','FontWeight','b
```

Insight #2:
Overall there are exactly 4000 (+ or - 1) of each rating.

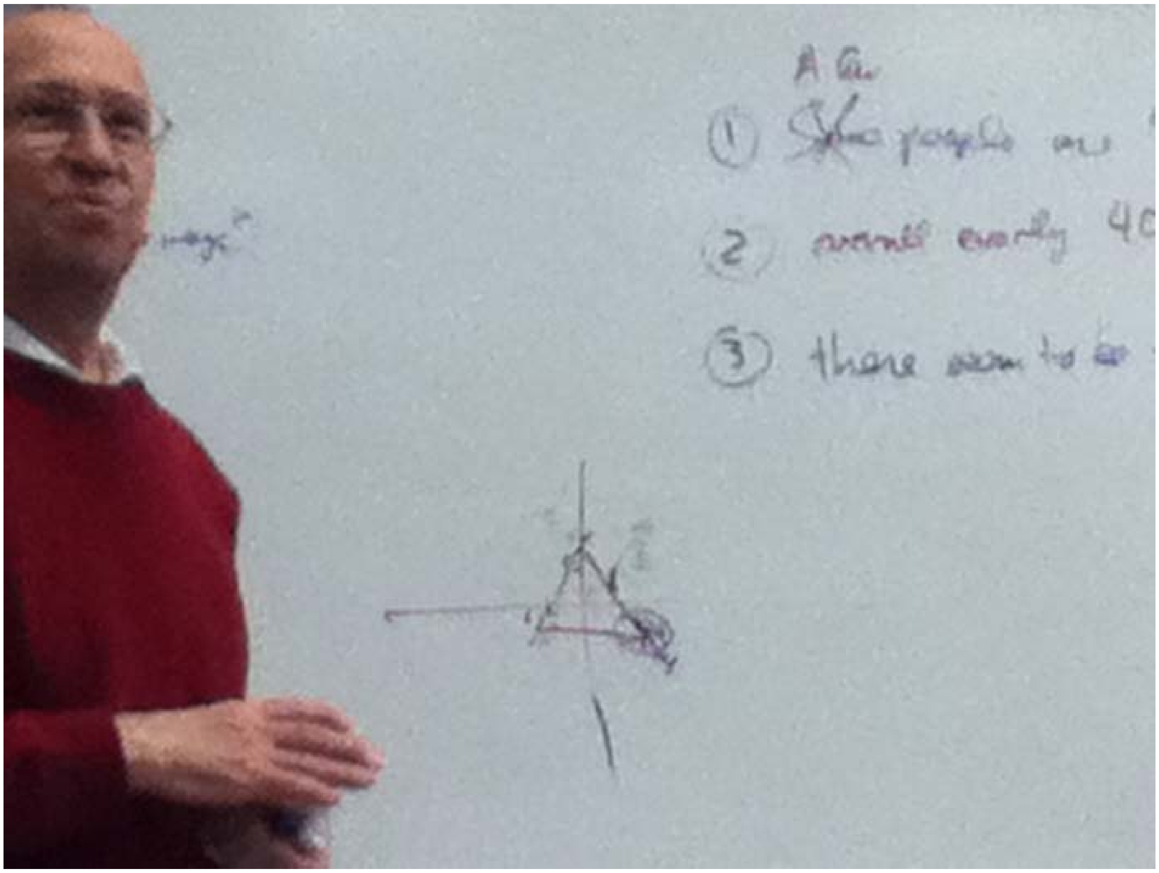


```
clear all
close all
clc
data=load('netflixishdata.txt'
size(data)
hist(sum(data,1)/200)
title('Average movie ratings p
figure
hist(sum(data,2)/100)
title('Average movie ratings t
figure
totaldata=data(:)
hist(totaldata)
```

MATLAB R2012a
File Edit Text Go Cell Tools Debug Parallel Desktop Window Help
Editor - /Users/ellenle/D
eulode.m
lecture17.m
heunode2.m
Figure 3
File Edit View Insert Tools Desktop Window Help
0
500
1000
1500
2000
2500
3000
3500
4000
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Insight #3 There seem to be exactly 4 kinds of movies:

Why? Movie ratings were generated from the sides of a regular tetrahedron



Segment 29. GMMs in N-Dimensions

Watch this segment

The direct YouTube link is http://youtu.be/PH8_qqDTCYY

Links to the slides: PDF file or PowerPoint file

Problems

To Compute

The file twoexondata.txt contains data similar to that shown in slide 6, as 3000 (x,y) pairs.

1. In your favorite computer language, write a code for K-means clustering, and cluster the given data using (a) 3 components and (b) 8 components. Don't use anybody's K-means clustering package for this part: Code it yourself. Hint: Don't try to do it as limiting case of GMMs, just code it from the definition of K-means clustering, using an E-M iteration. Plot your results by coloring the data points according to which cluster they are in. How sensitive is your answer to the starting guesses?

2. In your favorite computer language, and either writing your own GMM program or using any code you can find elsewhere (e.g., Numerical Recipes for C++, or scikit-learn, which is installed on the class server, for Python), construct mixture models like those shown in slide 8 (for 3 components) and slide 9 (for 8 components). You should plot 2-sigma error ellipses for the individual components, as shown in those slides.

To Think About

1. The segment (or the previous one) mentioned that the log-likelihood can sometimes get stuck on plateaus, barely increasing, for long periods of time, and then can suddenly increase by a lot. What do you think is happening from iteration to iteration during these times on a plateau?

Class Activity

GMM activity

Segment 30. Expectation Maximization (EM) Methods

Watch this segment

The direct YouTube link is http://youtu.be/StQOzRqTNsw

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. For a set of positive values $\{x_i\}$, use Jensen's inequality to show (a) the mean of their square is never less than the square of their mean, and (b) their (arithmetic) mean is never less than their harmonic mean.

2. Sharpen the argument about termination of E-M methods that was given in slide 4, as follows: Suppose that $g(x) \geq f(x)$ for all x , for some two functions f and g . Prove that, at any local maximum x_m of f , one of these two conditions must hold: (1) $g(x_m) > f(x_m)$ [in which case the E-M algorithm has not yet terminated], or (2) $g(x_m)$ is a local maximum of g [in which case the E-M algorithm terminates at a maximum of g , as advertised]. You can make any reasonable assumption about continuity of the functions.

To Think About

1. Jensen's inequality says something like "any concave function of a mixture of things is greater than the same mixture of the individual concave functions". What "mixture of things" is this idea being applied to in the proof of the E-M theorem (slide 4)?

2. So slide 4 proves that some function is less than the actual function of interest, namely $L(\theta)$. What makes this such a powerful idea?

Activity

The class activity for Friday can be found at EM activity.

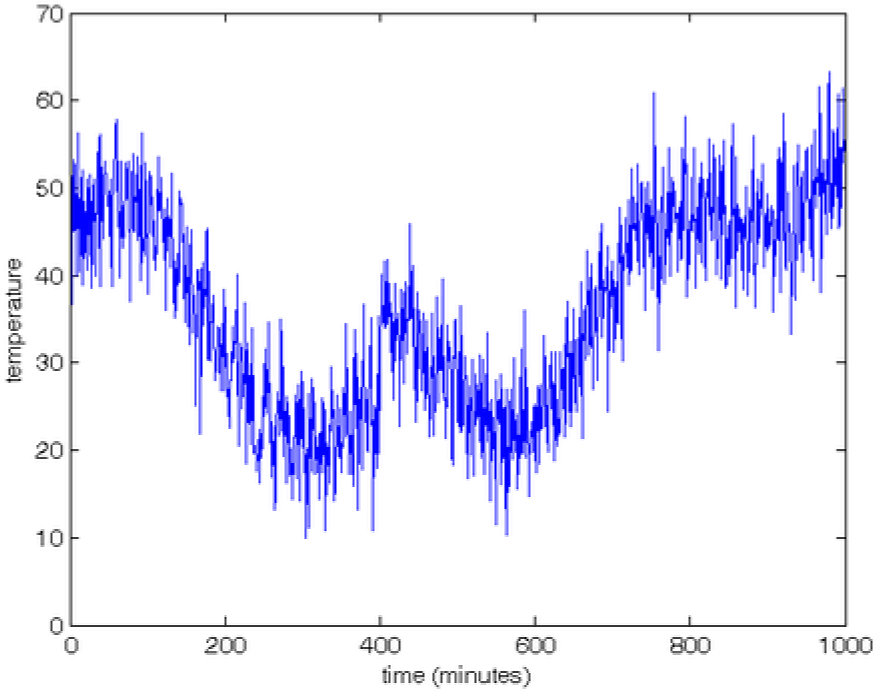
Segment 31. A Tale of Model Selection

Watch this segment

The direct YouTube link is http://youtu.be/_G1gzqQzbuM

Links to the slides: PDF file or PowerPoint file

Problems



To Calculate

(These problems will be the class activity on Monday, but you can get a head start on them if you want.)

I measured the temperature of my framitron manifold every minute for 1000 minutes, with the same accuracy for each measurement. The data is plotted on the right (with data points connected by straight lines), and is in the file Modelselection.txt.

1. From the data, estimate the measurement error σ . (You can make any reasonable assumptions that follow from looking at the data.)

2. Write down a few guesses for functional forms, with different (or adjustable) numbers of parameters that might be good models for the data. Order these by their model complexity (number of parameters) from least to most.

3. Fit each of your models to the data, obtaining the parameters and $\chi 2 min$ for each. (Hint: write your code generally enough that you can change from model to model by changing only one or two lines.)

4. Which of your models "wins" the model selection contest if you use AIC? Which for BIC?

To Think About

1. Both AIC and BIC decide whether to allow a new parameter based on a $\Delta\chi 2$. So it is possible to think about each as a p-value test for whether a null hypothesis ("no new parameter") is ruled out at some significance level. Viewed in this way, what are the critical p-values being used by each test?

2. Can you give a reasonable rationale, that might be used by a proponent of BIC, for why its $\Delta\chi 2$ should be larger in magnitude as N (the number of data points) increases?

Segment 32. Contingency Tables: A First Look

Watch this segment

The direct YouTube link is http://youtu.be/NvCdN2RFufY

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. 20 our of 100 U.S. Senators are women, yet when the Senate formed an intramural baseball team of 9 people only 1 woman was chosen for the team. What is the probability of this occurring by chance? What is the p-value with which the null hypothesis "there is no discrimination against women Senators" can be rejected?

2. A large jelly bean jar has 20% red jelly beans, 30% blue, and 50% yellow. If 6 jelly beans are chosen at random, what is the chance of getting exactly 2 of each color? What is the name of this distribution?

3. A small jelly bean jar has 2 red jelly beans, 3 blue, and 5 yellow. If 6 jelly beans are chosen at random, what is the chance of getting exactly 2 of each color? What is the name of this distribution?

To Think About

1. Suppose that, in the population, 82% of people are right-handed, 18% left handed; 49% are male, 51% female; and that handedness and sex are independent. Repeatedly draw samples of N=15 individuals, form the contingency table, and apply the chi-square test for significance to get a p-value, exactly as described in the lecture segment. How often is your p-value less than 0.05? If you get an answer that is different from 0.05, why? Try larger values of N until the answer converges to 0.05. (How are you handling zero draws when they occur?)

Class Activity

There was a surprise quiz. Bill's solutions are here.

We will analyze these contingency tables, asking (i) What is $\chi 2$? (ii) What is the p-value? (iii) Is there a significant association? (iv) If so, can you describe the main effect(s) seen?

	Vanilla	**Strawberry**	**Chocolate**
Texas Tech	1	1	13
A&M	16	4	15
UT	45	32	80

Grades	**A**	**B**	**C**	**D**
A&M	[illegible]	24	32	1
UT	[illegible]	80	50	18

Segment 33. Contingency Table Protocols and Exact Fisher Test

Watch this segment

The direct YouTube link is http://youtu.be/9Qrkw5UfAmQ

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. How many distinct m by n contingency tables are there that have exactly N total events?

2. For every distinct 2 by 2 contingency table containing exactly 14 elements, compute its chi-square statistic, and also its Wald statistic. Display your results as a scatter plot of one statistic versus the other.

To Think About

1. Suppose you want to find out of living under power lines causes cancer. Describe in detail how you would do this (1) as a case/control study, (2) as a longitudinal study, (3) as a snapshot study. Can you think of a way to do it as a study with all the marginals fixed (protocol 4)?

2. For an m by n contingency table, can you think of a systematic way to code "the loop over all possible contingency tables with the same marginals" in slide 8?

Activity

Chess contingency tables

Segment 34. Permutation Tests

Watch this segment

The direct YouTube link is http://youtu.be/_4BUS1NGNHA

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Use the permutation test to decide whether the contingency table

5	3	2
2	3	6
0	2	3

shows a significant association. What is the p-value?

2. Repeat the calculation using the Pearson chi-square statistic instead of the Wald statistic, or vice versa.

To Think About

1. Is slide's 7 suggestion, that you figure out how to implement the permutation test without "expanding all the data", actually possible? If so, what is your method?

Class Activity

Media:somecontingencies3.txt

Segment 37. A Few Bits of Information Theory

Watch this segment

The direct YouTube link is http://youtu.be/ktzYOLDN3u4

Links to the slides: PDF file or PowerPoint file

Class Activity

There is no general way to estimate the entropy of a (non i.i.d.) process from the data it generates, because you may or may not be able to recognize its entropy-lowering internal structure. So, in general, even an accurate "estimate" is only an upper bound on the entropy.

Let's see how well we can do at estimating the true entropy of five different strings in the alphabet A, C, G, T. (Bill knows the answer, because he knows how they were generated. But he's not telling!)

The more you study the data, the better you'll do! (If you know how to use Hidden Markov Models, which we didn't have room for in this course, you might do even better.)

Media:entropystring1.txt

Media:entropystring2.txt

Media:entropystring3.txt

Media:entropystring4.txt

Media:entropystring5.txt

Segment 38. Mutual Information

Watch this segment

The direct YouTube link is http://youtu.be/huNPh1mkJHM

Links to the slides: PDF file or PowerPoint file

Class activity

Information theory blitz
http://wpressutexas.net/coursewiki/images/f/fa/Information_theory_blitz.pdf

Segment 39. MCMC and Gibbs Sampling

Watch this segment

The direct YouTube link is http://youtu.be/4gNpgSPal_8

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Suppose the domain of a model are the five integers $x=\{1,2,3,4,5\}$, and that your proposal distribution is: "When $x_1=2,3,4$, choose with equal probability $x_2=x_1\pm1$. For $x_1=1$ always choose $x_2=2$. For $x_1=5$ always choose $x_2=4$. What is the ratio of q 's that goes into the acceptance probability $\alpha(x_1,x_2)$ for all the possible values of x_1 and x_2 ?

2. Suppose the domain of a model is $-\infty<x<\infty$ and your proposal distribution is (perversely),

$$q(x_2|x_1)=\{\tfrac{7}{2}\exp[-7(x_2-x_1)],\tfrac{5}{2}\exp[-5(x_1-x_2)],x_2\geq x_1 x_2<x_1$$

Sketch this distribution as a function of x_2-x_1 . Then, write down an expression for the ratio of q 's that goes into the acceptance probability $\alpha(x_1,x_2)$.

To Think About

1. Suppose an urn contains 7 large orange balls, 3 medium purple balls, and 5 small green balls. When balls are drawn randomly, the larger ones are more likely to be drawn, in the proportions large:medium:small = 6:4:3. You want to draw exactly 6 balls, one at a time without replacement. How would you use Gibbs sampling to learn: (a) How often do you get 4 orange plus 2 of the same (non-orange) color? (b) What is the expectation (mean) of the product of the number of purple and number of green balls drawn?

2. How would you do the same problem computationally but without Gibbs sampling?

3. How would you do the same problem non-stochastically (e.g., obtain answers to 12 significant figures)? (Hint: This is known as the Wallenius non-central hypergeometric distribution.)

[Answers: 0.155342 and 1.34699]

Class Activity

There's a story here, about diagnosing rats by which branches they pick in a maze. Bill will explain in class. Unless he thinks up a better story.

Mathematically, it's another one of these amazing Gibbs sampling examples. Suppose 2 unknown distributions over the digits 0..9, that is $p_0,p_1,\ldots,p_9$ and $q_0,q_1,\ldots,q_9$, of course with $\sum_i p_i=1$ and $\sum_i q_i=1$. This data file has 1000 lines, each with 10 i.i.d. draws of digits, either from the p 's or the q 's -- but, for each line, you don't know which.

1. Estimate $p_0,p_1,\ldots,p_9$ and $q_0,q_1,\ldots,q_9$ from the data. If you are ambitious, do this by two different methods: First, by Gibbs sampling. Second, by an E-M method. (Although these are conceptually different, my code for them differs by only a few lines.)

2. Estimate a probability for each line in the data file as to whether it is drawn from the p_i 's (as opposed to the q_i 's.

3. Plot histograms that show the uncertainties of your Gibbs estimate for the p_i 's. Do your E-M estimates appear to be at the modes of your Gibbs histograms? Should they be?

Media:gibbs_data.txt

Jeff's solution

Segment 40. Markov Chain Monte Carlo, Example 1

Watch this segment

The direct YouTube link is http://youtu.be/nSKZ02ZWzsY

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. The file Twoexondata.txt has 3000 pairs of (first, second) exon lengths. Choose 600 of the first exon lengths at random. Then, in your favorite programming language, repeat the calculation shown in the segment to model the chosen first exon lengths as a mixture of two Student distributions. That is (see slide 2): "6 parameters: two centers, two widths, ratio of peak heights, and Student t index." After running your Markov chain, plot the posterior distribution of the ratio of areas of the two Student components, as in slide 6.

2. Make a histogram of the 2nd exon lengths. Do they seem to require two separate components? If so, repeat the calculations of problem 1. If not, use MCMC to explore the posterior of a model with a single Student component. Plot the posterior distribution of the Student parameter ν .

To Think About

1. As a Bayesian, how would you decide whether, in problem 2 above, you need one vs. two components? What about 7 components? What about 200? Can you think of a way to enforce model simplicity?

2. After you have given a good "textbook" answer to the preceding problem, think harder about whether this can really work for large data sets. The problem is that even tiny differences in log-likelihood *per data point* become huge log-odds differences when the number of data points is large. So, given the opportunity, models are almost always driven to high complexity. What do you think that practical Bayesians actually do about this?

Activity

Urns with MCMC

Segment 41. Markov Chain Monte Carlo, Example 2

Watch this segment

The direct YouTube link is http://youtu.be/FnNckBLWJ24

Links to the slides: PDF file or PowerPoint file

Problems

To Calculate

1. Show that the waiting times (times between events) in a Poisson process are Exponentially distributed. (I think we've done this before.)

2. Plot the pdf's of the waiting times between (a) every other Poisson event, and (b) every Poisson event at half the rate.

3. Show, using characteristic functions, that the waiting times between every Nth event in a Poisson process is Gamma distributed. (I think we've also done one before, but it is newly relevant in this segment.)

To Think About

1. In slide 5, showing the results of the MCMC, how can we be sure (or, how can we gather quantitative evidence) that there won't be another discrete change in k_1 or k_2 if we keep running the model longer. That is, how can we measure convergence of the model?

2. Suppose you have two hypotheses: H1 is that a set of times t_i are being generated as every 26th event from a Poisson process with rate 26. H2 is that they are every 27th event from a Poisson process with rate 27. (The mean rate is thus the same in both cases.) How would you estimate the number N of data points t_i that you need to clearly distinguish between these hypotheses?

Class Activity

Urns with MCMC

Segment 47. Low-Rank Approximation of Data

Watch this segment

The direct YouTube link is http://youtu.be/M0gsHNS_5FE

Links to the slides: PDF file or PowerPoint file

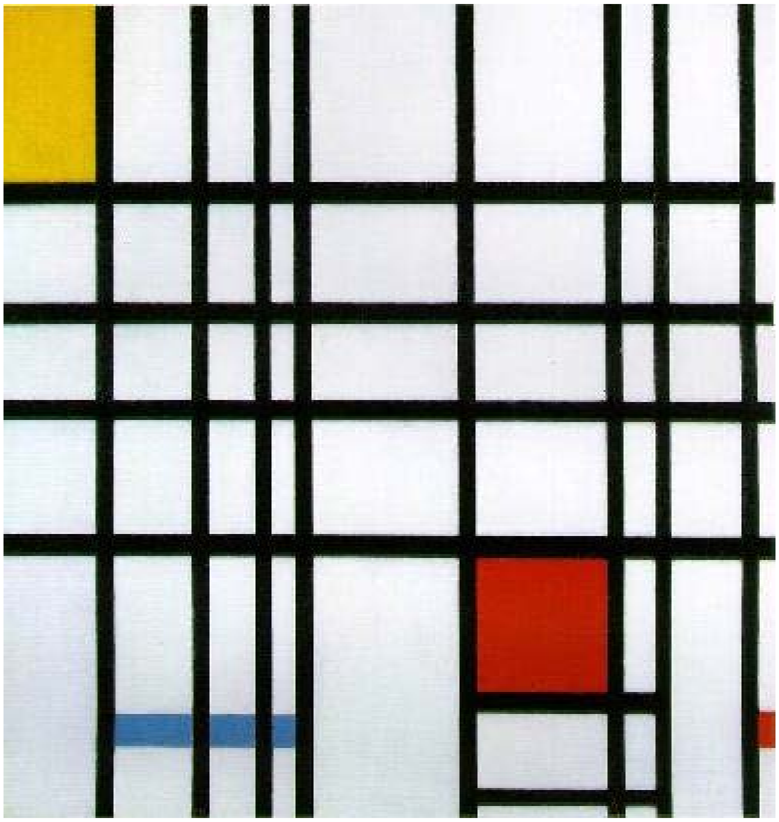
Class Activity

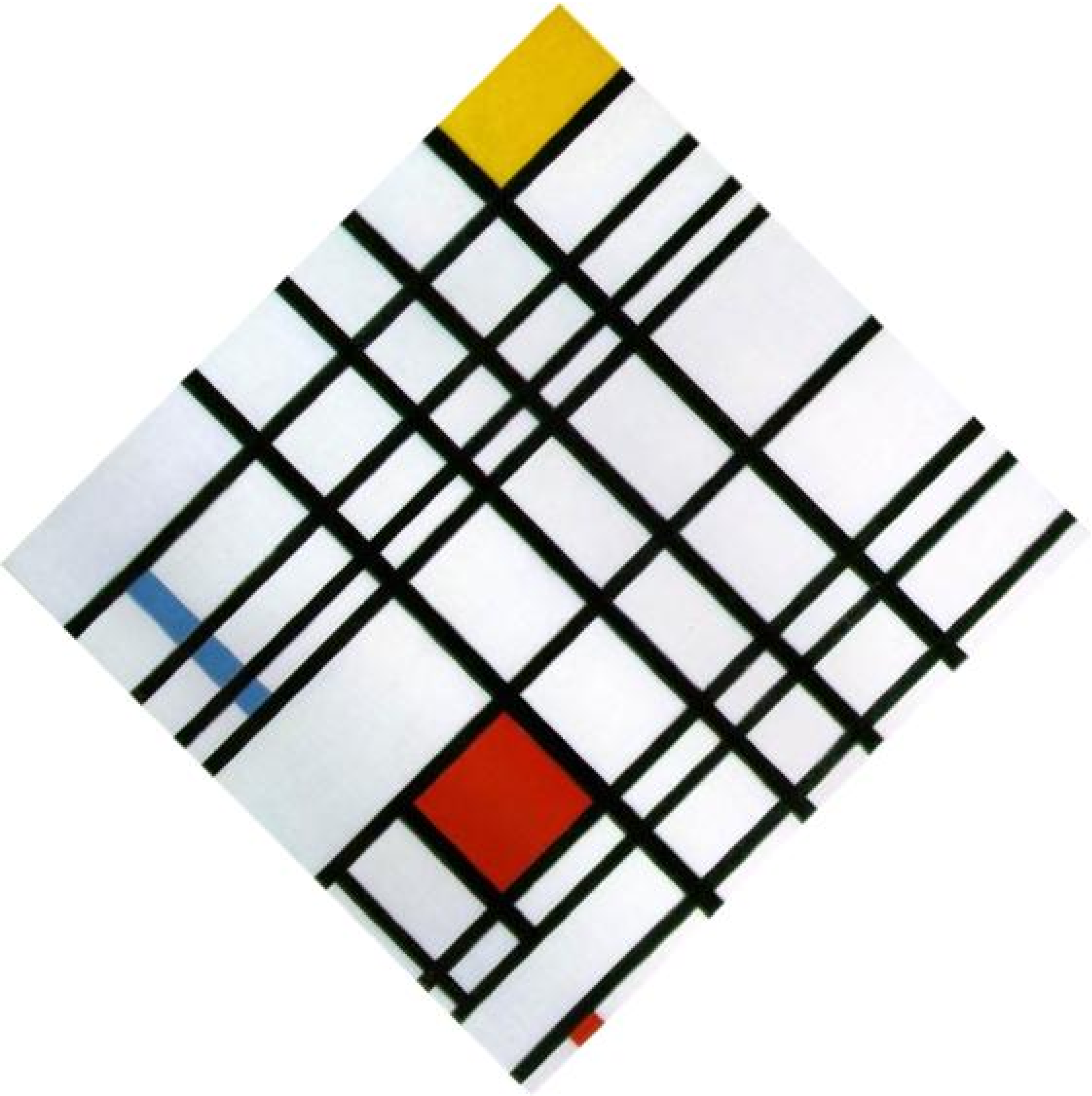
Class activity description: File:Svd exercise.pdf

Required files: File:Svd exercise.m.txt, File:GetCongress.txt

Last resort file: File:CongressSVD.m.txt

Pictures:





Segment 48. Principal Component Analysis (PCA)

Watch this segment

The direct YouTube link is http://youtu.be/frWqIUpIxLg

Links to the slides: [PDF file](#) or [PowerPoint file](#)

Problems

To Compute

1. Suppose that only one principal component is large (that is, there is a single dominant value s_i). In terms of the matrix $\mathbf{V}$ (and anything else relevant), what are the constants a_j and b_j that make a one-dimensional model of the data? This would be a model where $x_{ij} \approx a_j \lambda_i + b_j$ with each of the data points (rows) having its own value of an independent variable λ_i and each of the responses (columns) having it's own constants a_j, b_j .

2. The file dataforpca.txt has 1000 data points (rows) each with 3 responses (columns). Make three scatter plots, each showing a pair of responses (in all 3 possible ways). Do the responses seem to be correlated?

3. Find the principal components of the data and make three new scatter plots, each showing a pair of principal coordinates of the data. What is the distribution (histogram) of the data along the largest principal component? What is a one-dimensional model of the data (as in problem 1 above)?

To Think About

1. Although PCA doesn't require that the data be multivariate normal, it is most meaningful in that case, because the data is then completely defined by its principal components (i.e., covariance matrix) and means. Can you design a test statistic that measures "quality of approximation of a data set by a multivariate normal" in some quantitative way? Try to make your statistic approximately independent of N , the number of data points.

Extra Credit Segments Follow This

Segment 25. Fitting Models to Counts

Watch this segment

The direct YouTube link is http://youtu.be/YXaq2PVCGZQ

Links to the slides: PDF file or PowerPoint file

Segment 26. The Poisson Count Pitfall

Watch this segment

The direct YouTube link is http://youtu.be/rPO3N5GI-3I

Links to the slides: PDF file or PowerPoint file

Segment 35. Ordinal vs. Nominal Contingency Tables

Watch this segment

The direct YouTube link is http://youtu.be/fYUbj78aguk

Links to the slides: PDF file or PowerPoint file

Segment 36. Contingency Tables Have Nuisance Parameters

Watch this segment

The direct YouTube link is http://youtu.be/bHK79WKOX-Y

Links to the slides: PDF file or PowerPoint file

Segment 49. Eigenthingies and Main Effects

Watch this segment

The direct YouTube link is http://youtu.be/LpGQnvvGLMQ

Links to the slides: PDF file or PowerPoint file

Segments with Slides But Not Yet Recorded

(links are to PDF files)

Segment 15.5. Poisson Processes and Order Statistics
Segment 42. Wiener Filtering
Segment 43. The IRE Lady
Segment 44. Wavelets
Segment 45. Laplace Interpolation
Segment 46. Interpolation On Scattered Data
Segment 50. Binary Classifiers
Segment 51. Hierarchical Classification
Segment 52. Dynamic Programming

Team Randomizer

Link to the team randomizer