mathematical methods - week 14

Math for experimentalists

Georgia Tech PHYS-6124

Homework HW #14 due Tuesday, November 24, 2020

== show all your work for maximum credit,
== put labels, title, legends on any graphs
== acknowledge study group member, if collective effort
== if you are LaTeXing, here is the source code

Exercise 14.1 *A "study" of stress and life satisfaction* a) to d) 10 points

Bonus points
Exercise 14.1 *A "study" of stress and life satisfaction* e) 4 points
Exercise 14.2 *Unbiased sample variance* 5 points
Exercise 14.3 *Standard error of the mean* 5 points
Exercise 14.4 *Bayesian statistics*, by Sara A. Solla 10 points

Total of 10 points = 100 % score. Extra points accumulate, can help you later if you miss a few problems.

edited November 25, 2020

Week 14 syllabus

November 17, 2020

For this week's lectures read about the binomial theorem, Poisson and Gaussian distributions in AWH Chapter 23 *Probability and statistics* (click here). The fastest way to watch any week's lecture videos is by letting YouTube run the course playlist (click here).

- Sara A Solla: *Linear and nonlinear dimensionality reduction: applications to neural data, or - The unreasonable effectiveness of linear algebra. (1:08 hour)*

  - Sara A. Solla's lecture notes: Neural recordings; Principal Components Analysis (PCA); Singular Value Decomposition (SVD); ISOMAP nonlinear dimensionality reduction; Multidimensional scaling.

- *Clip 2 Ignacio Taboada- Probability, Uncertainty, probability density functions, error matrix (20 min)*

  - Ignacio Taboada — lecture notes.

- *Clip 3 Distributions: binomial, normal, uniform, moments, quantiles. Monte Carlo (why you need the uniform distribution) (19 min)*

  - Ignacio Taboada — lecture notes.

- *Clip 4 central limit theorem (why you need normal dist) (2 min)*

- *Clip 5 Multi-dimensional PDFs (13 min)*

- *Clip 6 Error propagation: Covariances add! Covariances add! Covariances add! Three times :) (18 min)*

  - Ignacio Taboada — lecture notes.

Optional reading

- Discussion 1, Sara A Solla: *A physicist turns neuroscientist. You can do anything. Progress in brain science. What is consciousness. What we know. The rest is speculation. Much speculation. (57 min)*

- *Rant - This is a kindergarten course. A professional should teach it, so I can teach you stuff that nobody teaches you here. (4 min)*

- *Sermon - Thanksgiving is upon us, don't be stupid (3 min)*

14.1 Optional reading: Bayesian statistics

Sara A. Solla

Natural sciences aim at abstracting general principles from the observation of natural phenomena. Such observations are always affected by instrumental restrictions and limited measurement time. The available information is thus imperfect and to some extent unreliable; scientists in general and physicists in particular thus have to face the task of extracting valid inferences from noisy and incomplete data. Bayesian probability theory provides a systematic framework for quantitative reasoning in the face of such uncertainty.

In this lecture (not given in the Fall 2020 course) we will focus on the problem of inferring a probabilistic relationship between a dependent and an independent variable. We will review the concepts of joint and conditional probability distributions, and justify the commonly adopted Gaussian assumption on the basis of maximal entropy arguments. We will state Bayes' theorem and discuss its application to the problem of integrating prior knowledge about the variables of interest with the information provided by the data in order to optimally update our knowledge about these variables. We will introduce and discuss Maximum Likelihood (ML) and Maximum A Posteriori (MAP) for optimal inference. These methods provide a solution to the problem of specifying optimal values for the parameters in a model for the relationship between independent and dependent variables. We will discuss the general formulation of this framework, and demonstrate that it validates the method of minimizing the sum-of-squared-errors in the case of Gaussian distributions.

- A quick but superficial read: Matthew R. Francis, *So what's all the fuss about Bayesian statistics?*
- Reading: Lyons [3], *Bayes and Frequentism: a particle physicist's perspective* (click here)

14.2 Statistics for experimentalists: desiderata

I have solicited advice from my experimental colleagues. You tell me how to cover this in less than two semesters :)

2012-09-24 Ignacio Taboada Cover least squares. To me, this is the absolute most basic thing you need to know about data fitting - and usually I use more advanced methods.

For a few things that particle and astroparticle people do often for hypothesis testing, read Li and Ma [2], *Analysis methods for results in gamma-ray astronomy*, and Feldman and Cousins [1] *Unified approach to the classical statistical analysis of small signals*. Both papers are too advanced to cover in this course, but the idea of hypothesis testing can be studied in simpler cases.

2012-09-24 Peter Dimon thoughts on how to teach math methods needed by experimentlists:

1. Probability theory
   (a) Inference
   (b) random walks
   (c) Conditional probability

   (d) Bayes rule (another look at diffusion)
   (e) Machlup has a classic paper on analysing simple on-off random spectrum. Hand out to students. (no Baysians use of information that you do not have) (Peter takes a dim view)

2. Fourier transforms
3. power spectrum - Wiener-Kitchen for correlation function
   (a) works for stationary system
   (b) useless on drifting system (tail can be due to drift only)
   (c) must check whether the data is stationary
4. measure: power spectrum, work in Fourier space
   (a) do this always in the lab
5. power spectra for processes: Brownian motion,
   (a) Langevin $\rightarrow$ get Lorenzian
   (b) connect to diffusion equation
6. they need to know:
   (a) need to know contour integral to get from Langevin power spectrum, to the correlation function
7. why is power spectrum Lorenzian - look at the tail $1/\omega^2$
   (a) because the cusp at small times that gives the tails
   (b) flat spectrum at origin gives long time lack of correlation
8. position is not stationary
   (a) diffusion
9. Green's function
   (a) δ fct $\rightarrow$ Gaussian + additivity
10. Nayquist theorem
    (a) sampling up to a Nayquist theorem (easy to prove)
11. Other processes:
    (a) what signal you expect for a given process
12. Fluctuation-dissipation theorem
    (a) connection to response function (lots of them measure that)
    (b) for Brownian motion power spectrum related to imaginary part of response function
13. Use *Numerical Recipes* (stupid on correlation functions)
    (a) zillion filters (murky subject)
    (b) Kalman (?)
14. (last 3 lecturs)
    (a) how to make a measurement

(b) finite time sampling rates (be intelligent about it)

PS: Did I suggest all that? I thought I mentioned, like, three things.

Did you do the diffusion equation? It's an easy example for PDEs, Green's function, etc. And it has an unphysically infinite speed of information, so you can add a wave term to make it finite. This is called the Telegraph Equation (it was originally used to describe damping in transmission lines).

What about Navier-Stokes? There is a really cool exact solution (stationary) in two-dimensions called Jeffery-Hamel flow that involves elliptic functions and has a symmetry-breaking. (It's outlined in Landau and Lifshitz, *Fluid Dynamics*).

2012-09-24 Mike Schatz .

1. 1D bare minimum:
   (a) temporal signal, time series analysis
   (b) discrete Fourier transform, FFT in 1 and 2D - exercises
   (c) make finite set periodic
2. Image processing:
   (a) Fourier transforms, correlations,
   (b) convolution, particle tracking
3. PDEs in 2D (Matlab): will give it to Predrag (Predrag is still waiting)

References

[1] G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals", Phys. Rev. D **57**, 3873–3889 (1998).

[2] T.-P. Li and Y.-Q. Ma, "Analysis methods for results in gamma-ray astronomy", Astrophys. J. **272**, 317–324 (1983).

[3] L. Lyons, "Bayes and Frequentism: a particle physicist's perspective", Contemporary Physics **54**, 1–16 (2013).

Exercises

14.1. **A "study" of stress and life satisfaction.**
Participants completed a measure on how stressed they were feeling (on a 1 to 30 scale) and a measure of how satisfied they felt with their lives (measured on a 1 to 10 scale). Participants' scores are given in table 14.1.

You can do this homework with pencil and paper, in Excel, Python, whatever:

Participant	Stress score (X)	Life Satisfaction (Y)
1	11	7
2	25	1
3	19	4
4	7	9
5	23	2
6	6	8
7	11	8
8	22	3
9	25	3
10	10	6

Table 14.1: Stress vs. satisfaction for a sample of 10 individuals.

a) Calculate the average stress and satisfaction.
b) Calculate the variance of each.
c) Plot Y vs. X.
d) Calculate the correlation coefficient matrix and indicate the value of the covariance.
e) Bonus: Read the article on "The Economist" (if you can get past the paywall), or, more seriously, D. Kahneman and A. Deaton -the 2002 Nobel Memorial Prize in Economic Sciences- about the correlation between income and happiness. Report on your conclusions.

14.2. **Unbiased sample variance.** Empirical estimates of the mean $\hat{\mu}$ and the variance $\hat{\sigma}^2$ are said to be "unbiased" if their expectations equal the exact values,

$$\mathbb{E}[\hat{\mu}] = \mu\,, \qquad \mathbb{E}[\hat{\sigma}^2] = \sigma^2\,. \tag{14.1}$$

(a) Verify that the empirical mean

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} a_i \tag{14.2}$$

is unbiased.

(b) Show that the naive empirical estimate for the *sample variance*

$$\bar{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - \hat{\mu})^2 = \frac{1}{N}\sum_{i=1}^{N} a_i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} a_i\right)^2$$

is biased. Hint: note that in evaluating $\mathbb{E}[\cdots]$ you have to separate out the diagonal terms in

$$\left(\sum_{i=1}^{N} a_i\right)^2 = \sum_{i=1}^{N} a_i^2 + \sum_{i\neq j}^{N} a_i a_j\,. \tag{14.3}$$

(c) Show that the empirical estimate of form

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \hat{\mu})^2\,, \tag{14.4}$$

is unbiased.

(d) Is this empirical sample variance unbiased for any finite sample size, or is it unbiased only in the $n \to \infty$ limit?

Sara A. Solla

14.3. **Standard error of the mean.**

Now, estimate the empirical mean (14.2) of observable a by $j = 1, 2, \cdots, N$ attempts to estimate the mean $\hat{\mu}_j$, each based on M data samples

$$\hat{\mu}_j = \frac{1}{M} \sum_{i=1}^{M} a_i \, . \tag{14.5}$$

Every attempt yields a different sample mean.

(a) Argue that $\hat{\mu}_j$ itself is an idd random variable, with unbiased expectation $\mathbb{E}[\hat{\mu}] = \mu$.

(b) What is its variance

$$\text{Var}[\hat{\mu}] = \mathbb{E}[(\hat{\mu} - \mu)^2] = \mathbb{E}[\hat{\mu}^2] - \mu^2$$

as a function of variance expectation (14.1) and N, the number of $\hat{\mu}_j$ estimates? Hint; one way to do this is to repeat the calculations of exercise 14.2, this time for $\hat{\mu}_j$ rather than a_i.

(c) The quantity $\sqrt{\text{Var}[\hat{\mu}]} = \sigma/\sqrt{N}$ is called the *standard error of the mean* (SEM); it tells us that the accuracy of the determination of the mean μ. How does SEM decrease as the N, the number of estimate attempts, increases?

Sara A. Solla

14.4. **Bayes.** *Bayesian statistics*.