Assignment 2, due September 16, before class

Please follow instructions on the assignments page

About this assignment: This assignment is about stochastic modeling and optimization for maximum likelihood estimation.

1. (*Multiple "best" fits*) Here is an example illustrating that the likelihood surface may have more than one local maximum. Download the datafile

   `Assignment2_Exercise1_dataset1.parquet`

   You can read it into a Python pandas dataframe using

   `Assignment2_Exercise1_read_data.py`

   This exercise is to apply maximum likelihood estimation to this dataset with a hypothesized Student t distribution with PDF given by equation (15) of `ProbabilityModelsForStatistics.pdf`. Make plots of $\ell(\mu)$ (the log likelihood sum) as a function of μ for various values of r (the length scale parameter) and ν (the number of degrees of freedom). When ν is large, there should be a single local max at $\mu = 0$. The shape of the curve will depend on r in some way. Find smaller ν values (always larger than 1) for which there are multiple local maxima. This is an exercise in computational exploration and experimentation, not theory. It illustrates the point that there may be different ways to decide which datapoints are important and which others are outliers. These can lead to different "best fits" of the data. This can't happen, in this particular example, for large ν because large ν is like Gaussian and the Gaussian distribution doesn't recognize outliers.

2. (*Probabilistic modeling*) The probability that a person age t dies in the interval $[t, t + dt]$ is $\lambda(t)dt$. The function $\lambda(t)$ is called the *force of morbidity* (or force of *mortality*). It is an increasing function of t related to life expectancy. For this exercise, use the model $\lambda(t) = a + bt$ with a and b positive numbers.[1] Let T be the time of death with this linear force of morbidity model.

   (a) Find an explicit formula for $p(t)$, the PDF of T, up to an unknown normalization constant. As we did in class, use Bayes' rule and $P'(t) = p(t)$, P being the CDF, together with $P(0) = 0$ for find the functional form of $p(t)$. *Hint.* It might be easier to solve the differential equation, once you have it, using the quantity $S(t) = 1 - P(t) = \Pr(T > t)$. This is the *survival probability* for time t. The boundary condition for S is $S(0) = 1$ (from the problem, $T > 0$ always). The PDF is $p(t) = -S'(t)$.

   (b) Suppose we have N independent measurements T_k. Find a formula for $L(t_1, \cdots, t_N|a, b)$ and $\ell(t_1, \cdots, t_N|a, b)$ in terms of the data.

3. (*Central limit theorem via simulation*) The distribution of a random variable created from a large number of other random variables has a tendency to be approximately Gaussian. Theorems that

[1] Models like this are used by people selling life insurance or annuities (fixed income until death), many of whom are actuaries. They use more complex force of morbidity models, but the mathematical principles are the same.

say this, with various hypotheses, are *central limit theorems*. Many of these involve the sum of i.i.d. random variables. Suppose $X_k \sim g(\cdot)$ are the i.i.d. summands and the sum is

$$Y_n = \sum_{k=1}^{n} X_k \ .$$

The theorem is that the distribution of Y_n is approximately gaussian when n is large.[2] This exercise is a graphical exploration of the basic theorem. *Normalizing* will let you compare distributions visually. We will normalize using the mean and variance of Y_n. The notation is standard but inconsistent. For example, μ_n is the mean of Y_n but μ_X is the mean of X. Also, since X_k all have the same distribution, we write X for a random variable with this distribution.

$$\mu_n = \mathrm{E}[\, Y_n \,] = n\mu_X = n\mathrm{E}[X] \ ,$$
$$\sigma_n^2 = \mathrm{var}(\, Y_n \,) = n\sigma_X^2 = n\,\mathrm{var}(X) \ .$$

We normalize Y_n by subtracting the mean then dividing by the standard deviation:

$$Z_n = \frac{Y_n - \mu_n}{\sigma_n} \ .$$

If Y_n is approximately normal, then Z_n is approximately normal too. The normalization gives Z_n mean zero and variance 1, which makes approximately a *standard normal*.

(a) Take $X_k = U_k^2$, where U_k is uniformly distributed[3] in $[0, 1]$. Calculate μ_X and σ_X. *Hint.* You could do this by first finding the PDF of X, which involves a square root, and then integrating x and x^2 with this PDF. It may be simpler to use $X = U^2$ and then find $\mathrm{E}\big[\, U^2 \,\big]$ by integrating with a constant density over the interval $[0, 1]$.

(b) Make a plot with histograms of the distribution of Z_n for various values of n. The plot should also include the standard normal density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \ .$$

It will take some ingenuity and persistence to make plots that show the accuracy and convergence clearly. Comment on your results. You will be graded on the quality of your work, not just on completing the tasks listed.

- It will take many samples of Z_n to estimate the PDF accurately. The number of samples needed depends on the number of bins. More bins mean smaller bins and smaller bin counts and, relatively speaking, less accurate estimates. Fewer bins gives only a coarse picture of the PDF of Z_n. Ideally, you should use very small bins and very many samples to get a high resolution low noise histogram. These runs can take a long time.
- You need to decide the range of z values for the histogram. If it's too wide, the end estimates are poor. If it's too narrow, you miss the tails, which can be most interesting and where the gaussian approximation can be poor.

[2] Wikipedia describes fancier theorems of this kind. One is the *Lindeberg Feller* theorem, in which the X_k are still independent but may have different distributions. The Lindeberg Feller (two people) condition limits how different the distributions of the X_k can be. Another is the *martingale* central limit theorem, in which the X_k are not independent. Exercise describes a case in which the X_k combine to determine Y_n in a in a way that is not a simple sum. All these theorems have in common some hypothesis that limits the influence any single X_k can have on the outcome, Y_n.

[3] The X distribution is pretty arbitrary, except that using U^2 instead of U removes a symmetry that would have made the central limit theorem approximation look more accurate than it usually is. When we have covered *sampling* methods, we will be able to do this experiment with X distributed differently.

- You might think about plotting the potential as in Assignment 1 in addition to the PDF. The issue with tails may be more difficult but the result more interesting.
- Consider plotting $f_n(z) - f(z)$ rather than just putting f_n and f in the same plot (f_n is the PDF of Z_n, as estimated using the histogram). This might make easier to see the differences when the differences are small. This may be particularly if you want to want to see quantitatively[4] how $f_n(z) - f(z)$ depends on n and z.

4. (*Law of large numbers via simulation*) "The" *law of large numbers*[5] says that the *sample mean* converges to the *population mean* as the sample size grows. In formulas, let X_k be i.i.d. samples of a random variable X with expected value μ_X. This is the *population mean*. The *sample mean* of N samples is

$$\overline{X}_N = \frac{1}{N} \sum_{k=1}^{N} X_k \ .$$

   The theorem is that if N is large then $\overline{X}_N \approx \mu_X$ with high probability. This exercise illustrates the theorem in an example via simulation and plotting. Let $X = U^2$ where U is uniformly distributed in $[0, 1]$ (as in the previous exercise). Choose a few values of N, some not so large and some larger. For each value of N, do n independent computational experiments to generate n independent sample means (total number of samples of X needed $= n \sum N$). Plot the histograms of $\overline{X}_N$ on the same graph. The curves for smaller N should be broader and with lower peaks. The curves for larger N should be narrow with high peaks, all centered around μ_X. Of course, the area under each histogram should be (approximately) equal to one (because they're estimates of probability density functions). *Comment.* This exercise is similar to exercise 3. The difference is the scaling. The plots in exercise 3 are scaled to examine the distribution of the sample mean. The plots here are scaled to emphasize the *concentration phenomenon*, which is that the distribution "concentrates" around a specific value, the population mean in this case. Concentration and distribution both play a role in exercise 5.

5. (*Variance of $\widehat{\theta}_{MLE}$ and Fisher information, one variable version*) Suppose $d = 1$ so that there is just one parameter being estimated. The frequentist analysis of estimation error models maximum likelihood estimate $\widehat{\theta}_{\rm MLE}$ as a random variable, because it depends on the data, which are modeled as random. This exercise outlines an explanation of a central limit theorem for $\widehat{\theta}_{\rm MLE}$ when the observations X_k are i.i.d. and n is large. It is what pure mathematicians call a *heuristic* (or informal) argument, because neither the hypotheses not the theorem are given precisely. The important thing that comes out is the formula for the variance of $\widehat{\theta}_{\rm MLE}$, which depends on the *Fisher information*.

   The calculations have the philosophy that we neglect fluctuations for random variables whose mean is not zero. We assume that the random value *concentrates* about its mean. When the mean is zero, we apply the central limit theorem and model it as gaussian. A longer analysis would justify these optimistic modeling assumptions. If Y_n is a random variable with non-zero mean, we just replace Y_n by its mean in the formulas. If Y_n has mean zero, we calculate its variance, σ_n^2. We then replace Y_n with a Gaussian random variable with mean zero and variance σ_n^2. A convenient way to do this in formulas is to replace Y_n with $\sigma_n Z$, where $Z \sim \mathcal{N}(0, 1)$ (Z is the standard normal).

   In the terminology of the notes, suppose that the probability model is that the PDF of a single

[4] There is a theory of this. The difference between f_n and f is roughly proportional to $\frac{1}{\sqrt{n}}$ multiplied by an *Hermite polynomial*, $H_3(z)$, with a coefficient that depends on the third moment of X. Your AI will know about this.

[5] As with "the" central limit theorem, there are several mathematical theorems that make this claim in various ways with various specific hypotheses.

X observation is $r(x|\theta)$. The log likelihood function is (see notes)

$$\ell(x_1, \cdots, x_n|\theta) = \sum_{k=1}^{n} \log(\, r(X_k|\theta)\,) \ .$$

The maximum likelihood estimator maximizes $\ell(\cdots|\theta)$, so it satisfies

$$\ell'(\cdots|\theta) = 0 \ . \tag{1}$$

Here ℓ' is the derivative with respect to θ. This may be written

$$S(\theta) = \sum_{k=1}^{n} \frac{r'(X_k|\theta)}{r(X_k|\theta)} \quad , \quad S(\widehat{\theta}) = 0 \ . \tag{2}$$

The frequentist model is that there is a true θ value, which we call θ_*, and the observations X_k are i.i.d. samples: $X_k \sim r(\cdot|\theta_*)$.

The estimate is not exact, $\widehat{\theta} \neq \theta_*$, because $S(\theta_*) \neq 0$. The difference between $\widehat{\theta}$ and θ_* depends on how far $S(\theta_*)$ is from zero. In symbols, write $\Delta\theta = \widehat{\theta} - \theta_*$ for the estimation error and $\Delta S = S(\widehat{\theta}) - S(\theta_*) = -S(\theta_*)$. The relation between $\Delta\theta$ and ΔS is given (if ΔS is small enough) by[6]

$$\Delta S \approx S'(\theta_*)\, \Delta\theta \ . \tag{3}$$

It shows that $\Delta\theta \neq 0$ because $\Delta S \neq 0$.

We infer the distribution of $\Delta\theta$ using (3) and the distributions of ΔS and S'.

(a) Show that

$$\mathrm{E}_{\theta_*}\left[\, \frac{r'(X|\theta_*)}{r(X|\theta_*)} \,\right] = 0 \ . \tag{4}$$

You can do this by writing the integral formula for the expected value, which involves the PDF of X. *Comment.* The notation $\mathrm{E}_{\theta_*}[\cdot]$ means to take the expectation when X has the distribution given by θ_*. It also would be possible to consider

$$\mathrm{E}_{\theta_1}\left[\, \frac{r'(X|\theta_*)}{r(X|\theta_*)} \,\right] = 0 \ ,$$

for a different value $\theta_1 \neq \theta_*$. This would mean sample $X \sim r(\cdot|\theta_1)$ and put that X into the fraction in the expectation. The expected value is probably not zero if $\theta_1 \neq \theta_*$.

(b) Show that

$$\mathrm{var}(\, \Delta S\,) = n\, \sigma^2_{\mathrm{MLE}} \quad , \quad \sigma^2_{\mathrm{MLE}} = \mathrm{E}_{\theta_*}\left[\, \left(\frac{r'(X|\theta_*)}{r(X|\theta_*)} \right)^2 \,\right] \ , \tag{5}$$

According to the philosophy above, we should model ΔS as

$$\Delta S \approx \sqrt{n}\, \sigma_{\mathrm{MLE}} Z \ .$$

(c) Show that

$$S' = \sum_{k=1}^{n} (\, V_k + W_k\,) \ ,$$

[6] A mathematically rigorous version of this argument would stop here to assess whether the first derivative approximation is accurate enough, whether $\Delta\theta$ is small enough, etc. These are good questions to ask, but please get through the informal argument first.

where

$$V_k = \frac{r''(X_k|\theta_*)}{r(X_k|\theta_*)} \quad , \quad \mathrm{E}_{\theta_*}[V_k] = 0 \; ,$$

and

$$W_k = -\frac{\left(r'(X_k|\theta_*)\right)^2}{r(X_k|\theta_*)^2} \quad , \quad \mathrm{E}_{\theta_*}[W_k] = -\sigma^2_{\mathrm{MLE}} \; . \tag{6}$$

According to the philosophy, we should model S' as

$$S' \approx -n\,\sigma^2_{\mathrm{MLE}} \; .$$

(d) Combine the results of parts (b) and (c) to show that the estimation error is approximately normal with mean (approximately) zero and variance (approximately)

$$\mathrm{var}(\theta_{\mathrm{MLE}}) \approx \frac{1}{n}\sigma^2_{\mathrm{MLE}} \; . \tag{7}$$

For this, you have to recognize that $-Z$ has the same distribution as Z because the normal distribution is symmetric. The result involves a remarkable cancellation of on factor of σ^2MLE in the ratio. It seems like a coincidence that the variance of S and the expected value of S' both involve σ^2_{MLE}.