Bayesian Statistics III/IV (MATH3361/4071) Michaelmas term 2019

Homework 4: Point Estimation and inference

Lecturer: Georgios Karagiannis georgios.karagiannis@durham.ac.uk

- You are requested to address only the Bayesian part from the following Exercise.
- You are suggested to have a look at both the Frequentist and the Bayesian treatments.

Exercise. ($\star\star$) [1]This exercise is based on a problem that arises in image processing. Look at the first row of Fig 1. If we were to observe the sunflower field from above, the sunflowers would be spread uniformly over it. Viewed from an angle, the sunflowers cluster at the top of the picture due to the effect of perspective. We would like to be able to tell from this clustering at what angle the camera was pointing and its height above the ground. We will not solve this problem here (it is rather difficult in general), but instead look at an idealized and simplified version of it.

Consider the left hand image in the second row of the figure. It shows 200 points sampled at random uniformly from the unit square. On the right, is a transformation of these points similar to that undergone by the sunflower image, except that here only the 'y-coordinate', the vertical position in the image, has been affected.

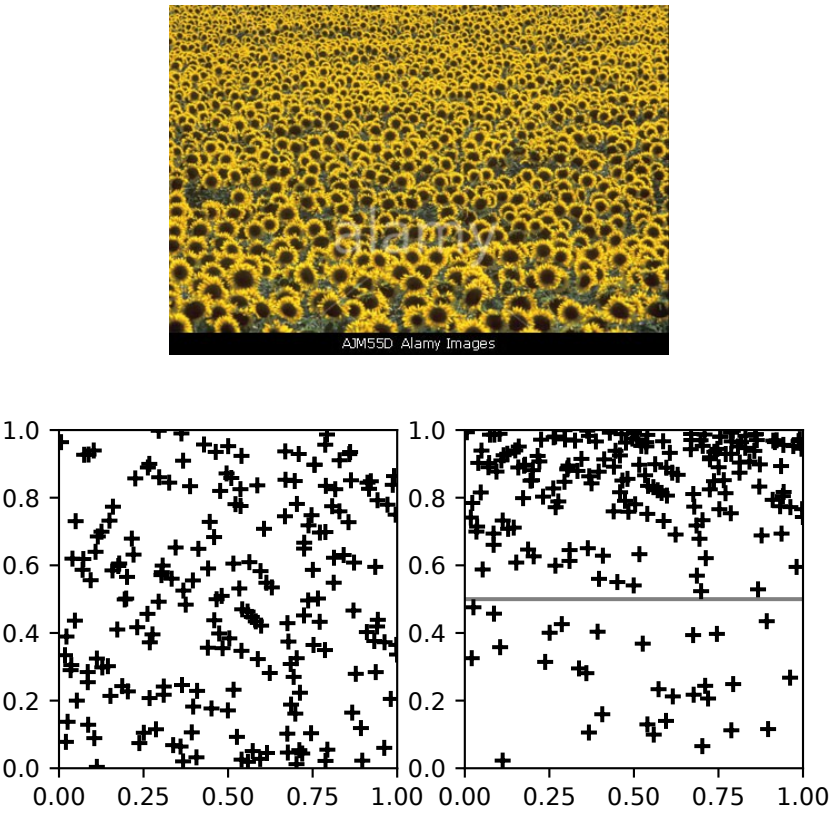

Figure 1: Two sampled point patterns. (Sunflower image © Soren Breiting / Alamy.)

If we were asked whether the right hand image had been sampled from a uniform distribution on the unit square, I am sure we would all say 'no'. The question is how we can justify this response. The first part of the exercise is

[1]This exercise is a modified version of the one from Dr Jermyn's lecture notes in Bayesian statistics 2015-2016

an alternative to the examples given in lectures, showing the need to use Bayesian and not ad hoc methods to obtain sensible answers in many inference problems.

The second part of the exercise is about inferring the camera angle given the data points using Bayesian methods, and tests various technical issues.

1. **Classical treatment.**

   A classical statistical technique to address this problem might go like this. Let's define a 'statistic', a quantity to be calculated from the data, and whose properties we will study to create a test. For example, in a coin tossing experiment, we will get some sequence of heads and tails. A statistic might be the number of heads.

   In this case, one possibility is to divide the unit square into two halves, top and bottom, and to count the number r of points in the bottom half.

   (a) If we assume that both point patterns were sampled from a uniform distribution on the unit square, which of the two sets of points is more probable? Does this help to justify the inference that the right hand image was not sampled from a uniform distribution?

   (b) If the total number of points is n, what is the probability distribution for r (the 'sampling distribution') under the assumption of sampling from the uniform distribution on the unit square?

   (c) What are the mean and variance of this distribution?

   (d) For the right hand image in the bottom row of Fig 1, by visual inspection, extract (approximately) the value of the statistic r.

   (e) Using a normal approximation to the sampling distribution, perform a significance test under the null hypothesis H_0 that the sampling was uniform. What is your conclusion?

   (f) How do we reconcile the answer to Sub-question 1a with the result of the hypothesis test? When we calculate the probability of observing r points in the bottom half of the unit square, what are we actually calculating?

   (g) What would be the result of the hypothesis test if we defined r as the number of points in the left half of the unit square?

   (h) Is this a reasonable thing to do? Why?

   What is being introduced in the last question is a possible alternative hypothesis, specifying the nature of the non-uniformity. The problem is that this alternative hypothesis has no place in the classical statistical testing methodology: all we have is H_0.

   As a result of this deficiency of standard hypothesis testing, other methods have been developed. Likelihood methods in classical statistics take alternatives into account by using two (or more) hypotheses and comparing the probabilities of the data under each of them.

   In our example, we could take a non-uniform model, and then compute the probabilities of the data under both uniform and non-uniform models. This would work in one sense: the probability of the data would be higher under some non-uniform models than under the uniform model. Unfortunately, there are many non-uniform models. Some of them, those with probability densities concentrated around the data points, assign very high probability to the observed data, and yet we do not accept them as valid explanations. This is a usuall phenomenon in Frequentist statistics: there are many hypotheses that predict the observed data with near certainty, and maximum likelihood is powerless to discount them.

2. **Bayesian treatment.**

   To disallow these extreme possibilities (if indeed they are unreasonable), we have to assign probabilities to the possible non-uniformities. One way to do this is via a choice of a parameterized family of non-

uniformities. Any parameterized model implicitly assigns probability zero to any non-family member, and hence is Bayesian by default.

In the case of the image processing problem, we know a lot about the types of distortion that arise (essentially perspective), and we can construct a reasonable family quite easily. Without going into details, and making several approximations, the coordinates $(x, y) \in [0, 1]^2$ of a point in the image distorted by camera viewing angle are related to the coordinates $(u, v) \in [0, 1]^2$ of the same point in the undistorted image that would be taken by a camera looking vertically downwards, by the following equations:

$$x = u \tag{1}$$

$$y = \frac{v(1+t)}{1+tv} \tag{2}$$

where $t = \tan(\alpha)$ is the tangent of the angle α between the camera viewing direction and vertically downwards—in fact, this was the exact transformation used to convert the left hand image in Fig 1 to the right hand image.

(a) Suppose we knew t. Derive the probability density $f(u, v|t)$ for a point (x, y) in the distorted image, given t, and given that the sampling density was uniform on the unit square in the undistorted image, i.e.

$$f(u, v|t) = 1 \tag{3}$$

(b) Write down the corresponding probability density, given t, of a set of points (x_1, y_1), ..., (x_n, y_n) sampled independently from the non-uniform density.

(c) Suppose we have no reason to favour any particular value of $\alpha \in [0, \pi/2]$ before we see the data. Write down the prior probability density for α.

(d) From the answer to Sub-question 2c, derive the prior probability density of t. [Hint: $\frac{\mathrm{d}}{\mathrm{d}t}(\tan^{-1} t) = \frac{1}{1+t^2}$.]

(e) Hence write down the posterior probability density for t up to an overall normalization factor, given the data points (x_1, y_1), ..., (x_n, y_n).

(f) From this result, derive the equation satisfied by the MAP estimate of t. (Taking logarithms makes things easier.)

(g) By expanding the log posterior probability density about 0 to second order in t, find the MAP estimate for t when t is small.

(h) Find the MAP estimate of α when α is small.

Solution.

1. **Classical treatment.**

(a) The probability that a single point falls in the infinitesimal element $\mathrm{d}u\mathrm{d}v$ at point (u, v) is:

$$\mathrm{d}F(u, v) = \mathsf{P}_F(u \in \mathrm{d}u, v \in \mathrm{d}v) = \mathrm{d}u\mathrm{d}v \tag{4}$$

The probability that n points sampled independently fall in the infinitesimal elements $\mathrm{d}u_1\mathrm{d}v_1$, ..., $\mathrm{d}u_n\mathrm{d}v_n$ at points (u_1, v_1), ..., (u_n, v_n) is therefore

$$\mathrm{d}F(u_1, v_1, \ldots, u_n, v_n) = \prod_{i=1}^{n} \mathrm{d}F(u_i, v_i) = \prod_{i=1}^{n} \mathrm{d}u_i\mathrm{d}v_i \tag{5}$$

This does not depend on the data points (u_i, v_i) and so is the same for both patterns.[2]

If we want to justify the idea that the set of points in the right hand image did not come from a uniform distribution, this obviously does not help, since both cases are the same.

(b) The probability of one point landing in the bottom half of the square is $\frac{1}{2}$ for a uniform distribution. The probability of any *given* set of r points lying in the bottom half (and thus $n - r$ lying in the top half) is then $(\frac{1}{2})^r(\frac{1}{2})^{n-r}$. The probability of some set of r points lying in the bottom half is thus given by the binomial distribution:

$$\mathsf{P}(r|n, \text{uniform}) = \binom{n}{r} \frac{1}{2^n} \tag{6}$$

(c) The mean of a binomial distribution is np and the variance is $np(1-p)$, meaning, in this case, that the mean is $\frac{n}{2}$ and the variance is $\frac{n}{4}$. For the given example, these are mean 100 and standard deviation $\sqrt{50} \simeq 7$.

(d) There are about $r = 30$ points in the bottom half of the square. (The exact number is not so relevant here.)

(e) The standardized value of the statistic r is $z = (r - 100)/7$, so for $r = 30$ we get $z = -10$, i.e. 10 standard deviations below the mean. The probability of finding this or a more extreme value in the bottom half of the square is then $2(1 - \Phi(10))$, a number that is vanishingly small. The null hypothesis is thus convincingly rejected.

(f) Superficially there seems to be a contradiction between the fact that the probabilities of the two sets of data are the same, but the hypothesis test so strongly rejects the null hypothesis. In fact, of course, the probability of the data and the probability computed in the hypothesis test are completely different. The former is the probability of a particular set of positions for points. The latter is a different in two ways. Remember that the probability of an individual configuration is constant under the uniform hypothesis. Then first, the probability of a particular value of r is given by the integral of this constant over all possible positions of the points that keeps the same number in the bottom half of the square. Second, the probability in the hypothesis test is then the sum of these probabilities for all values of r 'more extreme', i.e. further from the mean, than the value we observe. For the hypothesis test, then, we calculate the probability of a large, indeed in this case infinite, set of conceivable data sets, none of which we have actually observed.

Naturally, with such a difference in the probabilities, the conclusions to be drawn are different.

(g) The test does not justify rejecting the null hypothesis.

(h) It seems unreasonable because we know or suspect that the non-uniformity is in the vertical direction, and the statistic should be some measure of this non-uniformity. However, this alternative hypothesis has no place in the classical statistical testing methodology: all we have is H_0.

2. **Bayesian treatment.**

(a) Note that

$$u(x, y) = x \tag{7}$$

$$v(x, y) = \frac{y}{1 + t - ty} \tag{8}$$

[2]There is a subtlety here. The above probability is that the points fall in the given elements with the given labelling. Because there are $n!$ ways to label n points, strictly speaking, the probability of a configuration is $n!$ times the above, with the understanding that now the distribution is defined on sets of unlabelled points.

which is an 1-1 transformation. Thus the Jacobian is [3]

$$J = \frac{\mathrm{d}(u,v)}{\mathrm{d}(x,y)} = \det \begin{bmatrix} 1 & 0 \\ 0 & \frac{(1+t-ty)1-y(-t)}{(1+t-ty)^2} \end{bmatrix} = \frac{1+t}{(1+t-ty)^2} \tag{9}$$

Since the pdf on (U,V) is uniform, we find that the pdf is

$$f(x,y|t) = f(u(x,y),v(x,y)|t) \left| \frac{\mathrm{d}(u,v)}{\mathrm{d}(x,y)} \right| = \frac{(1+t)}{(1+t-ty)^2} \tag{10}$$

and hence the distribution is such that

$$\mathrm{d}F(x,y|t) = \underbrace{\frac{1+t}{(1+t-ty)^2}}_{=f(x,y|t)} \mathrm{d}(x,y) \tag{11}$$

(b) Since the points are sampled independently, the resulting pdf is simply the product of the individual pdfs:

$$f(x_1,y_1,\ldots,x_n,y_n|t) = \prod_{i=1}^{n} \frac{(1+t)}{(1+t-ty_i)^2} = \frac{(1+t)^n}{\prod_{i=1}^{n}(1+t-ty_i)^2} \tag{12}$$

(c) If we have no reason to favour any particular value of α, it would make sense to use a uniform distribution over α, i.e. the probability density with respect to α will be constant. Since $\alpha \in [0,\pi/2]$, we have probability distribution density

$$\pi(\alpha) = \frac{2}{\pi} \tag{13}$$

which is normalized; and hence probability such that

$$\mathrm{d}\Pi(\alpha) = \underbrace{\frac{2}{\pi}}_{=\pi(\alpha)} \mathrm{d}\alpha \tag{14}$$

(d) Since $\alpha(t) = \tan^{-1}(t)$, we have[4]

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t}(t) = \frac{1}{1+t^2} \tag{15}$$

so the pdf is

$$\pi(t) = \pi(\alpha(t)) \left| \frac{\mathrm{d}\alpha}{\mathrm{d}t}(t) \right| = \underbrace{\frac{2}{\pi}\frac{1}{1+t^2}}_{=\pi(t)} \tag{16}$$

and the distribution such that

$$\mathrm{d}\Pi(t) = \underbrace{\frac{2}{\pi}\frac{1}{1+t^2}}_{=\pi(t)} \mathrm{d}t \tag{17}$$

[3] Remember how to calculate the PDF or PMF when we perform transformation of random variables (in multivariate case). $\mathrm{d}(u(x,y),v(x,y))$ is an informal way of writing a transformed infinitesimal area, i.e. it denotes the area of the parallelogram covered by $(u(x,y),v(x,y))$ when both x and y are changed infinitesimally over a rectangle at (x,y) with dimensions $(\delta x,\delta y)$:

$$|J| = |\frac{\mathrm{d}(u(x,y),v(x,y))}{\mathrm{d}(x,y)}| = \left| \det \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} \right|$$

[4] Remember how we can calculate the PDF or the PMF when we perform a transformation of random variables

(e) The posterior density of t given the data points is

$$\pi(t|x_1, y_1, \ldots, x_n, y_n) \propto f(x_1, y_1, \ldots, x_n, y_n|t)\pi(t) \tag{18}$$

$$\propto \frac{(1+t)^n}{1+t^2} \prod_{i=1}^{n} \frac{1}{(1+t-ty_i)^2} \tag{19}$$

(f) The log posterior probability density is, up to an additive constant,

$$\log\left(\pi(t|x_1, y_1, \ldots, x_n, y_n)\right) = C + n\ln(1+t) - \ln(1+t^2) - 2\sum_{i=1}^{n} \ln(1+tz_i) \tag{20}$$

with $z_i = 1 - y_i$.
Differentiating with respect to t, and setting the result equal to zero gives the equation satisfied by the MAP estimate:

$$\frac{n}{1+t} - 2\sum_i \frac{z_i}{1+z_i t} = \frac{2t}{1+t^2} \tag{21}$$

(g) Consider that

$$g(t) = C + n\ln(1+t) - \ln(1+t^2) - 2\sum_{i=1}^{n} \ln(1+tz_i)$$

then

$$\frac{\mathrm{d}}{\mathrm{d}t} g(t) = \frac{n}{1+t} - \frac{2t}{1+t^2} - 2\sum_{i=1}^{n} \frac{z_i}{1+tz_i}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} g(t) = \frac{-n}{(1+t)^2} + \frac{2t^2-2}{(1+t^2)^2} + 2\sum_{i=1}^{n} \frac{z_i^2}{(1+tz_i)^2}$$

Then by using Taylor expansion of 2nd order around 0, it is

$$\tilde{g}(t) + g(0) + \frac{1}{1!}\frac{\mathrm{d}}{\mathrm{d}t} g(t)|_{t=0}(t-0) + \frac{1}{2!}\frac{\mathrm{d}^2}{\mathrm{d}t^2} g(t)|_{t=0}(t-0)^2 + \text{tiny stuff...}$$

$$\approx C + (n - 2n\bar{z})t + (-n - 2 + 2n\overline{z^2})t^2 + \text{tiny stuff...}$$

So, I will try to find the MAP estimate $\hat{t}^*$ by using this crude Taylor expansion approximation valid for values of t around zero 0, right... it is

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}(t)|_{t=\hat{t}^*} = 0$$

$$(n - 2n\bar{z}) - (n + 2 - 2n\overline{z^2})2t = 0$$

$$\hat{t}^* = \frac{\bar{z} - \frac{1}{2}}{\overline{z^2} - \frac{1}{2} - \frac{1}{n}}$$

Hence the MAP estimate is for small t approximately

$$\hat{t}^* = \frac{\bar{z} - \frac{1}{2}}{\overline{z^2} - \frac{1}{2} - \frac{1}{n}} \tag{22}$$

Some interpretation story ... Notice that the $\frac{1}{n}$ term comes from the prior pdf for t with respect to dt. The more data we have, the less significant it is. The rest of the formula measures the deviation of the mean point (note that $\bar{z} = 1 - \bar{y}$) from the central $z = \frac{1}{2}$ point, which makes sense. The value of $\hat{t}^*$ will only be small, and hence the approximation will only be consistent, if $\bar{z}$ is closer to $\frac{1}{2}$ than $\overline{z^2}$.

In addition, for those who checked, this is only a maximum (as opposed to a minimum), within the scope of this approximation, when $\overline{z^2} < \frac{1}{2} + \frac{1}{n}$.

(h)

$$\pi(\alpha|x_1, y_1, \ldots, x_n, y_n) \propto f(x_1, y_1, \ldots, x_n, y_n|\alpha)\pi(\alpha) \tag{23}$$

$$\propto (1 + t(\alpha))^n \prod_{i=1}^{n} \frac{1}{(1 + t(\alpha) - t(\alpha)y_i)^2} \tag{24}$$

where $t(\alpha) = \tan(\alpha)$—this is essentially the same as before up to the Jacobian factor $1/(1 + t^2)$. To find the equation satisfied by the MAP estimate we can now take logarithms:

$$\log\left(\pi(\alpha|x_1, y_1, \ldots, x_n, y_n)\right) = C + n\ln(1 + t(\alpha)) - 2\sum_{i=1}^{n} \ln(1 + t(\alpha)z_i) \tag{25}$$

and differentiate with respect to α. By the chain rule, the derivative is zero when

$$\left(\frac{n}{1 + t(\alpha)} - 2\sum_i \frac{z_i}{1 + z_i t(\alpha)}\right) \frac{\mathrm{d}t}{d\alpha}(\alpha) = 0 \tag{26}$$

On first order expansion, now also using first order approximation $t(\alpha) = \alpha$,

$$n(1 - \alpha) - 2\sum_i z_i(1 - z_i\alpha) = 0 \tag{27}$$

and thus the MAP estimate is

$$\hat{\alpha}^* = \frac{\bar{z} - \frac{1}{2}}{\overline{z^2} - \frac{1}{2}} \tag{28}$$

Note that $\hat{t}^* \neq \alpha^*$ even though $t = \alpha$ for small t (or equivalently, for small α). In general, the MAP estimate does not respect transformations.