Statistical Inference 1 Winter Term 2024/2025

Lecture: Prof. Dr. Thomas Augustin, Prof. Dr. Christian Heumann
Exercises: Sergio Buttazzo, Jan Anders
Tutorial, Corrections: Sophie Hopp, Michael Speckbacher

Suggested Solution

Exercise Sheet 2

Exercise 2.1 - KL-Divergence

(a) Give the definition of the Kullback-Leibler Divergence and show why

$$KL(g(.), f(.)) \neq KL(f(.), g(.))$$

except when $f(.) = g(.)$.

We prove this for the discrete case; the proof for the continuous case is analogous. Analytically, we have that

$$\begin{aligned} KL(g,f) - KL(f,g) &= \sum_{i=1}^{n} \log\left(\frac{g(x_i)}{f(x_i)}\right) g(x_i) - \sum_{i=1}^{n} \log\left(\frac{f(x_i)}{g(x_i)}\right) f(x_i) \\ &= \sum_{i=1}^{n} \log\left(\frac{g(x_i)}{f(x_i)}\right) g(x_i) + \sum_{i=1}^{n} \log\left(\frac{g(x_i)}{f(x_i)}\right) f(x_i) \\ &= \sum_{i=1}^{n} \log\left(\frac{g(x_i)}{f(x_i)}\right) (f(x_i) + g(x_i)) \neq 0 \end{aligned}$$

In the trivial case of $f = g$, we get that $\log\left(\frac{f(x_i)}{g(x_i)}\right) = log(1) = 0 \; \forall i$. The whole difference thus evaluates to 0.

In addition to the algebraic reason, we can motivate this intuitively. One should think of the two arguments of the KL divergence as different kinds of things: The first argument is empirical data, and the second argument is a model we're comparing the data to. In this sense, the KL-Divergence measures the "expected excess surprise" from using f as a model when the actual distribution is g.

(b) Under real world conditions, the true data generating process G(.) is rarely known. Assume we have a sample $x = (2,2,3,2,3,0,2,1,4,3)^T$, but we don't know the underlying data generating process, so we make the assumption $X_i \sim Bin(4, 0.5)$. Calculate the Kullback Leibler Divergence between the empirical distribution of the sample, $g(.)$, and the assumed distribution, $f(.)$.

First, construct the empirical distribution function from the given sample and calculate the corresponding densities for the assumption using `dbinom(c(0,1,2,3,4), 4, 0.5)`:

x	g(x)	f(x)
0	1/10	1/16
1	1/10	1/4
2	4/10	3/8
3	3/10	1/4
4	1/10	1/16

For the discrete case, we have:

$$KL(g, f) = \sum_{i=1}^{n} \log \frac{g(x_i)}{f(x_i)} g(x_i)$$

Plugging into the formula gives:

$$\begin{aligned} KL(g, f) &= \frac{1}{10} \cdot \log \frac{16}{10} + \frac{1}{10} \cdot \log \frac{4}{10} + \frac{4}{10} \cdot \log \frac{4 \cdot 8}{10 \cdot 3} + \frac{3}{10} \cdot \log \frac{3 \cdot 4}{10} + \frac{1}{10} \cdot \log \frac{16}{10} \\ &= \frac{1}{10} \cdot (\log \frac{16}{10} + \log \frac{4}{10} + 4 \cdot \log \frac{32}{30} + 3 \cdot \log \frac{12}{10} + \log \frac{16}{10}) \\ &= \frac{1}{10} \cdot (\log 4 + 3 \cdot \log 12 + 6 \cdot \log 16 - 4 \cdot \log 15 - 6 \cdot \log 10) \approx 0.0829 \end{aligned}$$

Or, in R:

```
g <- c(0.1, 0.1, 0.4, 0.3, 0.1)
f <- dbinom(c(0,1,2,3,4), 4, 0.5)
sum(log(g/f) * g)
[1] 0.08288353
```

(c) Assume that G(.) is actually hypergeometrically distributed with parameters $N = 9$, $K = 5$ and $n = 4$ (i.e. parameters $m = 5$, $k = 4$ and $n = 4$ in R). Calculate the KL-Divergence between the two densities. You may do the entire calculation in R. Was our assumed distribution a good approximation to the data and/or the true distribution?

Reminder: The hypergeometric distribution describes the probability of obtaining k successes in a sample of size n from a finite population of size N containing a total of K successes:

$$p_X(k) = P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Analytical solution is analogue to (b), just with different values for g(.). We first compare the data to the true data generating process in R, to determine which distribution provides a better model:

```
g <- c(0.1, 0.1, 0.4, 0.3, 0.1)
f <- dhyper(c(0,1,2,3,4), 5, 4, 4)
sum(log(g/f) * g)
[1] 0.2128796
```

In this case, the KL value obtained by using the binomial model at point (b) is lower than that given by the hypergeometric model, calculated in point (c). We can thus conclude that the proposed binomial model describes the observed data (slightly) better than the hypergeometric one, even if the latter is the true data generating process. While it may seem a contradiction, it is entirely possible for this to happen in the real world, as we don't observe the complete distribution of the data, but only a probabilistic sample.

The same holds true if we compare the assumed binomial distribution to the actual data generating process, as asked in the exercise:

```
g <- dhyper(c(0,1,2,3,4), 5, 4, 4)
f <- dbinom(c(0,1,2,3,4), 4, 0.5)
> sum(log(g/f) * g)
[1] 0.08308822
```

We observe that here too, the KL divergence between the assumed binomial model and the true data generating process is larger than its "distance" to the actual data, adding evidence that the actual obtained sample is described better by our binomial model.

Exercise 2.2 - Other divergence measures

Let $f(y)$ and $g(y)$ be two normal densities, with mean μ_f, μ_g respectively, and variance σ_f^2, σ_g^2.

(a) Use R to obtain a plot of the two density functions, for different values of the means and variances.

Setting $\mu_f = 0, \mu_g = 0.5$ and variance $\sigma_f^2 = 2, \sigma_g^2 = 2$, we obtain the following plot:

See `Ex_2_2.R`.

We now define different measures of divergence between two densities, by **integrating** the two following quantities:

- **Log-ratio** $= \log\left(\frac{f(y)}{g(y)}\right)$
- **Absolute difference** $= |f(y) - g(y)|$.

(b) Plot the log-ratio and absolute difference functions, for different values of $\mu_f, \mu_g, \sigma_f, \sigma_g$. To make things easier, you could use the `manipulate` package, for example.

See `Ex_2_2.R`.

(c) Define appropriate divergence measures, using the two quantities above. Evaluate them for some values of $\mu_f, \mu_g, \sigma_f, \sigma_g$.

We can define a divergence measure using the log-ratio by integrating over y:

$$LR(f(.), g(.)) = \int_{-\infty}^{\infty} \log\left(\frac{f(y)}{g(y)}\right) dy \, .$$

This results in the integral diverging, as the further we move away from the center of mass of the distributions, the more extreme the ratio and thus the log-ratio of the two distributions gets.

A simple, symmetric distance measure using the absolute differences would be to determine the maximum difference between the two densities:

$$\sup_y |f(y) - g(y)| \, .$$

However, this ignores the differences between the densities over the entire domain and does not involve the overall shape of the distributions. It is sensitive to local anomalies. A better approach is thus to again integrate the absolute differences over the entire domain. We realize that for a difference in the probability mass to occur in one place, the same difference must occur in the other direction in another place, since both densities must each integrate to one. We should thus halve the result so we don't end up counting the differences twice:

$$\delta(f(.), g(.)) = \frac{1}{2} \int_{-\infty}^{\infty} |f(y) - g(y)| \, dy \, .$$

Some evaluations are easily obtained by running the Shiny app in `Ex_2_2.R`.

(d) In your opinion, are the new versions of divergence better or worse than the KL divergence at evaluating a distance between two densities? Explain why, and support your thesis with appropriate values of $\mu_f, \mu_g, \sigma_f, \sigma_g$ in the example above.

As reference we consider the KL-divergence, that is defined as the integral of a weighted log-ratio. The problem with the unweighted log ratio is that the integral between $-\infty$ and ∞ doesn't converge, except when the two distributions are the same. This is shown in the plot from `Ex_2_2.R` for any value of μ or σ.

The absolute difference solves this problem: since the difference between the distributions grows small at exponential speed (at least for normal distributions), the integral will converge. The problem here is that when the difference in mean between the two distributions is very large, this divergence measure stops discriminating between them, it has an upper bound of 2. For example the divergence between two normal densities with variance 1 and $\mu_f = -1, \mu_g = 1$ is almost the same as between $\mu_f = -1000, \mu_g = 1000$.

Again, this can be checked in the plots from `Ex_2_2.R`..

Exercise 2.3 - Likelihood and log-likelihood

The number of people withdrawing money from an ATM machine in 10 minute intervals was counted. Four 10-minute intervals resulted in the following observations: 3, 1, 2, 2. Let us assume that these are realisations of a poisson-distributed variable $X \sim Po(\lambda)$.

(a) Write down the (log) likelihood function, and calculate the log-likelihood for $\widehat{\lambda} = 2$.
Hint: In the lecture, the log-likelihood and it's relation to the likelihood function were discussed (lecture 4, slide 26). It may be worthwhile to start with the likelihood here.

Given: i.i.d. sample $x = (3,1,2,2)^\top$ of a poisson-distributed variable.

$$\begin{aligned} L(\lambda) = \prod_{i=1}^{4} f(x_i;\lambda) &= \frac{\lambda^3}{3!}\exp(-\lambda) \cdot \frac{\lambda^1}{1!}\exp(-\lambda) \cdot \frac{\lambda^2}{2!}\exp(-\lambda) \cdot \frac{\lambda^2}{2!}\exp(-\lambda) \\ &= \frac{\lambda^8}{24}\exp(-4\lambda) \\ \Rightarrow l(\lambda) &= 8\log(\lambda) - \log(24) - 4\lambda \end{aligned}$$

So, for $\widehat{\lambda} = 2$, estimated from $\bar{x}$

$$\Rightarrow l(\widehat{\lambda}) = 8\log(2) - \log(24) - 8 \approx -5.633$$

In `R`:

```
sum(log(dpois(c(3,1,2,2), 2)))
[1] -5.632876
```

(b) How does the log-likelihood function change if the sample takes the values

$$x = (3,1,2,2,3,1,2,2,3,1,2,2,3,1,2,2)^\top$$

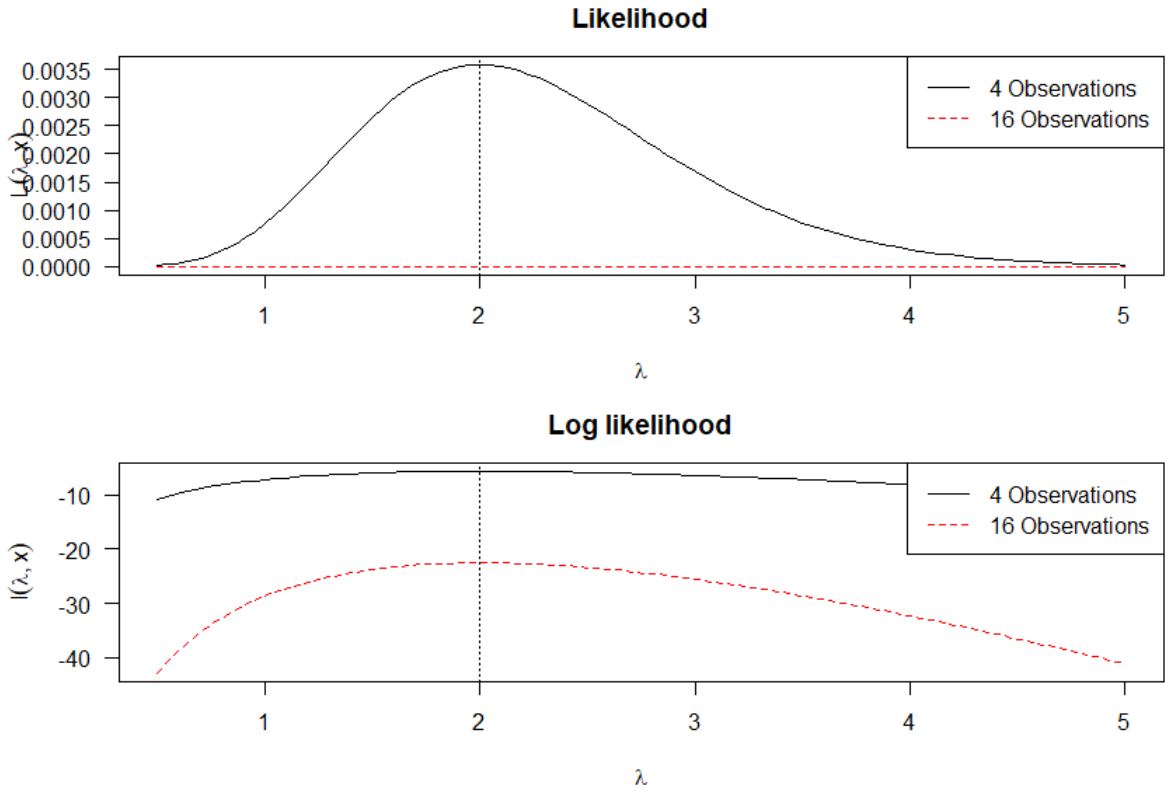
Interpret your results.

The second sample turns out to be a 4 times concatenation of the former sample from (a).

$$\begin{aligned} L_1(\lambda) = [L(\lambda)]^4 &= \frac{\lambda^{4\cdot 8}}{24^4}\exp(-4\cdot 4\lambda) \\ \Rightarrow l_1(\lambda) &= \log(L_1(\lambda)) = \log\left([L(\lambda)]^4\right) = 4\log\left(L(\lambda)\right) = 4l(\lambda) \\ \Rightarrow l_1(\widehat{\lambda}) &= 4l(\widehat{\lambda}) \end{aligned}$$

As we can see, the likelihood is taken to the fourth power which results in the log likelihood being multiplied by four, as there is four times the evidence for the estimate $\hat{\lambda}$.

(c) Let's now visualise how the value of the likelihood and log-likelihood varies with lambda. Write two functions in R that calculate the likelihood and log-likelihood for the sample given in (a) with λ as an input parameter. Use the functions to plot the likelihood function with λ in the range [0.5, 5]. Make use of the relationship discovered in (b) to display the likelihood for the samples given in (a) and (b) in each plot. What can you observe? Was the value $\widehat{\lambda}$ provided in (a) reasonable?



We can see that, while the likelihood function takes on much smaller values as the sample size increases, the log-likelihood is only linearly transformed. The maximum - the optimal parameter $\widehat{\lambda}$ maximising the likelihood - of both functions is the same, thanks to the strict monotonicity of the logarithm operator. We can also observe that the estimate for λ in (a) was optimal, as it maximizes the likelihood given the sample (3,1,2,2) (or multiples of it).

Note: For computers, the likelihood shrinking (or exploding) as a power as the sample gets larger can quickly become problematic, since floating point numbers have limited resolution, which can lead to rounding inaccuracies for very small or very large values. As a result, the log-likelihood is not only advantageous due to its analytical properties on paper, but also in digital format due to its less extreme reaction to growing sample sizes.