Example Sheet 2

5 Weighted mean with Gaussian errors

Let measured data values $y_1, y_2, \ldots, y_N$ be given with Gaussian measurement errors $\sigma_1, \sigma_2, \ldots, \sigma_N$. Our goal is to estimate the mean μ. The (i.i.d.) log-posterior reads:

$$\log P(\mu|D) = -\frac{1}{2}\sum_{n=1}^{N}\left(\frac{y_n-\mu}{\sigma_n}\right)^2 - \frac{1}{2}\left(\frac{\mu_0-\mu}{\sigma_0}\right)^2 + \text{const}$$

(a) Interpret the conjugate prior, what does it mean?

(b) Write down explicitly what the constant terms are in $\log P(\mu|D)$. (Contributions from likelihood and prior.)

(c) Differentiate the log-posterior with respect to μ, equate with zero and solve analytically for the maximum a-posteriori estimate $\hat{\mu}_{\mathrm{MPE}}$. (Do *not* use matrix notation.)

(d) $\hat{\mu}_{\mathrm{MPE}}$ has the form of a weighted mean value. What are the weights? Discuss how individual measurement values of low or high measurement errors influence the mean estimate.

(e) How does the prior enter into $\hat{\mu}_{\mathrm{MPE}}$? What happens to $\hat{\mu}_{\mathrm{MPE}}$, if we have no data at all ($N = 0$)?

(f) What happens if we discard the prior and assume that all measurement values have identical errors $\sigma_1 = \sigma_2 = \ldots = \sigma_N$?

6 Fisher analysis

A posterior $P(\theta|D)$ can be *approximated* by a Gaussian PDF at its maximum. The mean of the Gaussian is found by maximising the posterior. The width of the Gaussian is found by the Fisher analysis. For a single parameter, the variance is $\frac{1}{\sigma^2} = -\frac{\partial^2 \log P(\theta|D)}{\partial\theta^2}$.

(a) For the weighted mean estimate with Gaussian errors from the previous problem, give an error estimate using the Fisher analysis. How does the error σ scale with number of data points N?

(b) For the mean estimate with Poisson errors from the lecture, give an error estimate using the Fisher analysis. How does the error σ scale with number of data points N?

(c) For the fraction estimate with binomial errors from the lecture, give an error estimate using the Fisher analysis. How does the error σ scale with number of data points N?

7 Galaxy luminosity function

The distribution of galaxy luminosities is usually described by a so-called Schechter function $f(L) = c(L/L_*)^\alpha e^{-L/L_*}$, where L_* is an exponential cut-off luminosity, $\alpha < 0$ is the faint-end slope, and c is a normalisation constant such that $\int_0^\infty f(L)dL = 1$.

(a) What is the (mathematical) name of distribution functions of the kind $P(x) \propto x^\alpha e^{-x}$?

(b) Show that the Schechter function with $\alpha = -1.25$ (field galaxies) has finite mean but infinite variance.

(c) Convert the Schechter function $f(L)$ from luminosities to magnitudes $f(M)$ where $M = -2.5 \log_{10} L + M_0$.

(d) Let a set of N galaxy luminosities $L_1, L_2, \ldots, L_N$ be given *with no* errors. Write down the likelihood function.

(e) Show that for given data $L_1, L_2, \ldots, L_N$ and *fixed value* of α, the best-fit value of L_* can be found analytically.

(f) Show that the Schechter function is a generalised linear model and find the corresponding functions $a(\eta)$, $b(y)$, and $T(y)$. What does that imply for the likelihood and parameter estimation?

8 Loss functions

Sometimes we are facing data where we know next to nothing about its origin. In particular, we sometimes do not know the data's measurement error distributions, such that we cannot define any likelihood function. Instead of a likelihood function, we perform regression by minimising the *expected loss*

$$\langle L \rangle = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n, \theta))$$

where $L(y, f(x, \theta))$ is called the *loss function* and $f(x, \theta)$ is some model function at position x with parameters θ.

(a) Consider the quadratic loss function $L(y, f(x, \theta)) = (y - f(x, \theta))^2$ and the absolute-error loss function $L(y, f(x, \theta)) = |y - f(x, \theta)|$. Draw a sketch of both loss functions over the interval $-4 \leq y - f(x, \theta) \leq 4$. Which of the two loss functions is more sensitive to outliers? Which of the two loss functions is not differentiable and thus cannot possibly have any analytic minimisation?

(b) Consider the following loss function:

$$L(y, f(x, \theta)) = \begin{cases} (y - f(x, \theta))^2 & \Leftrightarrow \quad |y - f(x, \theta)| \leq 1 \\ 1 + 2 \log |y - f(x, \theta)| & \text{otherwise} \end{cases}$$

Draw a sketch of this loss function over the interval $-4 \leq y - f(x, \theta) \leq 4$. Compare the robustness of this loss function against outliers to the quadratic and the absolute-error loss functions. Is this loss function differentiable? Is an analytic minimisation possible?

(c) In classification settings, one often uses the zero-one loss function

$$L(y, f(x, \theta)) = \begin{cases} 0 & \Leftrightarrow \quad y = f(x, \theta) \\ 1 & \text{otherwise} \end{cases}$$

where y and $f(x, \theta)$ are integers (i.e. class labels). Give an interpretation of what the expected loss $\langle L \rangle$ means in simple words.