Problems, Astrostatistics.

1. Consider the coin tossing example, discussed in the first lecture. Simulate 1000 tosses of the coins, setting $H = 0.3$. Consider a uniform prior and update the posterior at each toss. Plot the resulting posterior after 1, 50, 100, 300, 700, 1000 tosses. Repeat the simulated experiment by setting a Gaussian prior centered in $H = 0.5$, with standard deviation $\sigma = 0.1$. Do both posteriors converge a similar distribution in the end? What does that mean? Which posterior converges faster and why?

2. Politician A makes a statement about some issue you knew nothing about before. Let's call such proposition S and assume your starting prior on S is uniform with 0.5 probability of S being either true or false. Update your probability of S being true, knowing that you trust Mr. A to tell the truth with probability $prob(A_T) = 4/5$. At this point Mr B - another politician - declares that he agrees with Mr A on S being true. You trust Mr. B much less, and believe that the probability of him to lie is $prob(B_T) = 3/4$. What is your final degree of belief in proposition S?

3. You are tested for a dangerous disease named "Bacillum Bayesianum" (BB). You test positive to BB. You know that the general incidence of BB in the population is 1%. Moreover, you know that your test has a false negative probability of 5% (false negative: you have BB but the test scores negative), and a false positive rate also of 5% (false positive: you do not have BB, but the test scores positive). What is the probability that you have actually contracted BB?

4. We define the characteristic function ϕ, for a given probability distribution, as $\phi \equiv E\left[\exp\left(-i\mathbf{k}\cdot\mathbf{x}\right)\right]$. Therefore, ϕ is the Fourier transform of the distribution.
$$\phi(\mathbf{k}) \equiv \int d^n x \exp\left(-i\mathbf{k}\cdot\mathbf{x}\right) p(\mathbf{x}). \tag{1}$$
Show that the characteristic function of a multivariate Gaussian distribution, $N(\mu, C)$, is equal to:
$$\phi(\mathbf{k}) = \exp\left(-i\mu^T \cdot \mathbf{k} - \frac{1}{2}\mathbf{k}^T C \mathbf{k}\right). \tag{2}$$
Do this in two different ways. First. Start from the original multivariate Gaussian and perform the Fourier transform by completing the square in the integrand. Second: perform a suitable rotation to diagonalize the covariance, and transform the resulting distribution.

5. The characteristic function is useful because it can be used to generate moments of the distribution via differentiation.
$$E\left[\right]x_\alpha^{n_\alpha} \dots x_\beta^{n_\alpha}\Big] = \left[\frac{\partial^{n_\alpha \dots n_\beta}\phi(\mathbf{k})}{\partial(-ik_\alpha)^{n_\alpha} \dots \partial(-ik_\beta)^{n_\beta}}\right]_{\mathbf{k}=0}. \tag{3}$$

Apply this to a Multivariate Gaussian, $N(\mu, \Sigma)$, to find its mean and covariance.

6. Show that the characteristic function of a multivariate Gaussian $N(\mu, \Sigma)$ is another (unnormalized) multivariate Gaussian.

7. Consider a bivariate Gaussian distribution $N(\mu, \Sigma)$, where $\mu = (4, 2)$, $\Sigma_{11} = 1.44$, $\Sigma_{22} = 0.81$, $\Sigma_{12} = \Sigma_{21} = -0.702$. Assume that such distribution describes the posterior of two parameters X_1 and X_2. Plot contours of this distribution and tell what are the boundaries of the 95% credibile interval for each parameter, after marginalizing over the other. Tell how these boundaries change if, instead of marginalizing, we fix either parameter to a known value. Now, re-obtain the same boundaries via Monte Carlo sampling of the posterior, in three ways: a) by directly drawing values of (X_1, X_2), by mean of a a Cholesky decomposition of the covariance (check in the literature how to do this); b) Via Metropolis-Hastings sampling of the posterior, c) Via Gibbs sampling of the posterior. In all cases, do not rely on pre-made code, but write your own (i.e., you can surely use libraries to generate known distributions, do Cholesky decomposition, and so on, but you should not use a pre-made library with a command like "Gibbs-sampling"). In all cases, provide final contour plots, and marginalized posteriors. In the MCMC analysis, produce also trace plots to test for convergence of the chains.

8. We consider a data vector $\vec{d}$ of length N. We want to fit the data using a linear combination of M templates $\vec{t}$ with unknown amplitudes $\vec{A}$, i.e. we assume that the average of the data is given by:

$$\langle \vec{d} \rangle = \sum_{i=1}^{M} A_i \vec{t_i} \equiv \vec{A}^t \cdot T. \tag{4}$$

If we further assume that the data are Gaussian distributed with covariance C, the best-fit amplitude parameter vector is found via maximization of the chi-squared statistics:

$$\chi^2 = \left(\vec{d} - \vec{A}^t T\right)^t C^{-1} \left(\vec{d} - \vec{A}^t T.\right). \tag{5}$$

Perform such maximization to get an estimate of $\vec{A}$.

9. Consider a dataset $\vec{d} = (d_1, \dots, d_N)$, where the measurement d_i is taken at time/position x_i. The average of $\vec{d}$ is assumed to be $\langle \vec{d} \rangle = \omega \vec{x} + b$ (i.e., you are going to fit the data with a straight line). Noise is Gaussian and uncorrelated. Find the MAP estimate of parameters ω, b and the parameter covariance matrix, assuming a uniform prior (note: you can take the result of the previous exercise as a starting point; of course the result yields the usual linear regression formulae).

10. Consider a dataset $(d_1, \dots, d_n)$, where the measurement d_i is taken at time/position $\vec{x}_i$. The average of d_i is assumed to be $\langle d_i \rangle = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$. Noise is Gaussian and uncorrelated, therefore $d_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Assuming uniform priors on the parameters β_i $(i = 0, \dots, p)$, it is well known and easy to see that the MAP estimate of $\vec{\beta} = (\beta_0, ..., \beta_p)$ is found via least square fitting, i.e., by minimizing the function $\mathrm{RSS} \equiv \sum_{i=1}^{n} (d_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$ (RSS: "residual sum of squares"). We want to see how this result is modified by different, specific choices of prior.

   (a) Assume the following prior for $\vec{\beta}$: all β_i are i.i.d., according to a double exponential (Laplace) distribution: $p(\beta_i) = \frac{1}{2b} \exp\left(-\frac{|\beta_i|}{b}\right)$, where b is some fixed scale parameter. Show in this case that MAP estimation leads to the following solution ("LASSO regression"):

$$\hat{\vec{\beta}} = \underset{\vec{\beta}}{\operatorname{argmin}} \left(\mathrm{RSS} + \frac{2\sigma^2}{b} \sum_{j=1}^{p} |\beta_j| \right). \tag{6}$$

   (b) Assume the following prior for $\vec{\beta}$: all β_i are i.i.d., according to a Gaussia distribution with mean 0 and variance c. Show in this case that MAP estimation leads instead to the following solution ("ridge regression"):

$$\hat{\vec{\beta}} = \underset{\vec{\beta}}{\operatorname{argmin}} \left(\mathrm{RSS} + \frac{\sigma^2}{c} \sum_{j=1}^{p} \beta_j^2 \right). \tag{7}$$

11. Consider a likelihood with the average depending on some amplitude A, just rescaling the mean

$$L(\Theta, x) = \frac{1}{(2\pi)^{\frac{N}{2}} |C|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (x - A\bar{x})^T C^{-1} (x - A\bar{x}) \right] \tag{8}$$

    In the above $\bar{x}$ is the theoretical expectation value, depending on the parameter vector Θ, in which we set the amplitude parameter A to $A = 1$. Marginalize over A assuming a uniform (unnormalized) $p(A) = 1$ prior.

12. You have collected a data time series $D = \{d_1 \equiv d(t_1), d_2 \equiv d(t_2), \dots, d_n \equiv d(t_N)\}$. Your data are evenly time spaced ($t_{i+1} - t_i = \Delta$ for any i, $1 \leq i \leq N$). You assume that the time series contains a single stationary sinusoidal signal $f(t)$ plus noise. You also assume that noise is Gaussian with known variance σ^2, and that different measurements are statistically independent. Your data are therefore modeled as

$$d_i = f(t_i) + n_i = B_1 \cos(\omega t_i) + B_2 \sin(\omega t_i) + n_i, \tag{9}$$

where B_1 and B_2 are unknown amplitudes, ω is the unknown frequency of the signal, and n_i are the noise contributions. In the following, always assume to work in a large frequency limit for your signal.

- Show that the likelihood $\mathcal{L}(D|\omega, B_1, B_2, I)$ can be written in the form:

$$\mathcal{L} = (2\pi\sigma)^{-\frac{N}{2}} \exp\left(-\frac{Q}{2\sigma^2}\right), \tag{10}$$

$$Q \equiv N\bar{d^2} - 2[B_1 R(\omega) + B_2 I(\omega)] + B_1^2 c + B_2^2 s, \tag{11}$$

$$R(\omega) = \sum_{i=1}^{N} d_i \cos(\omega t_i) \tag{12}$$

$$I(\omega) = \sum_{i=1}^{N} d_i \sin(\omega t_i) \tag{13}$$

$$s \equiv \sum_{i=1}^{N} \cos^2(\omega t_i), \tag{14}$$

$$c \equiv \sum_{i=1}^{N} \sin^2(\omega t_i), \tag{15}$$

  where $\bar{d^2}$ is the arithmetic average of the square of the data in the series.

- Show that s and c, defined above, are well approximated as ω-independent constants.

- Consider uniform (proper or improper) priors on the three parameters of the problem (ω, B_1, B_2). Treating the amplitudes B_1 and B_2 as nuisance parameters, find the marginalized 1-parameter posterior for ω. Show that the marginalized posterior can be written as:

$$P(\omega|D, I) \propto \frac{1}{\sqrt{cs}} \exp\left[\frac{R(\omega)^2/c + I(\omega)^2/c}{2\sigma^2}\right], \tag{16}$$

  (note that marginalization amounts to solving Gaussian integrals, which can be done analytically via standard techniques).

- Finally show that the MAP estimate of the frequency of your signal is obtained as the value of ω which maximizes the discrete Fourier transform power spectrum of the signal (periodogram), $C(\omega)$:

$$C(\omega) = \frac{2}{N} \left| \sum_{k=1}^{N} d_k \exp(-i\omega t_k) \right|^2 . \tag{17}$$

  Summarize all the assumptions you made along the way, to reach this result.

- Under which condition on ω the MAP estimate above would have been obtained via least square fitting?