# Bayesian Inference I - Exam questions

Roberto Trotta

February 2025

**Instructions**

Each student selects one problem from the list below, using an algorithm of your choice but making sure that each student works on a different problem from everybody else (i.e., there is a bijection between problems and students). The group collectively assigns a weight to each problem: $\star$ for easier problems; $\star\star$ for average difficulty problems (compared to the rest); $\star\star\star$ for harder problems. In the exam, problems with higher difficulty rating will be weighted more.

1. **Bayesian evidence for counting experiment**

   We consider the same 'on/off' problem we investigated in the lectures, but this time from a Bayesian evidence perspective. In order to obtain analytical results, instead of adopting the uniform prior of Eq. (3.40), use instead the conjugate prior for the Poisson likelihood, i.e., the Gamma distribution. Consider the following cases:

   (a) Start from a fixed background rate, $b = 1$, assumed known. For a given signal rate $s$ (you may assume everywhere the integration time to be unity) and a given choice of the parameters of the Gamma distribution, compute the Bayes factor:

   $$B_{01} = \frac{\Pr(n_{\rm on} \mid M_0)}{\Pr(n_{\rm on} \mid M_1)}. \tag{1}$$

   where $M_0$ is a model with $s = 0, b = 1$ and $M_1$ is a model with $b = 1$ and $s$ distributed according to the Gamma prior.

   (b) Plot $B_{01}(n_{\rm on})$ for the following choices of the Gamma distribution shape parameter, $\alpha$, and scale parameter, $\theta$: $\{\alpha = 1, \theta = 2\}$; $\{\alpha = 3, \theta = 3\}$. In each case, compare the posterior probability for $M_1$ (the model with non-zero $s$) with the conclusion that would be drawn on the value of $s$ from its posterior distribution under $M_1$.

   (c) Generalize point (a) above to the case where $b$ has been measured in an 'off' region, and having obtained $n_{\rm off}$ counts. Use the same Gamma prior for $b$ as for $s$. Now carry out the comparison between the Bayes factor result and the marginal posterior for $s$ in the 2D plane spanned by $n_{\rm on}, n_{\rm off}$.

2. **Linear regression in the presence of selection**

We consider the errors-in-variables model seen in the Lab exercises, but now including intrinsic dispersion and with additional selection on the dependent variable, $y_i$. The model has therefore the following conditional structure, with known $\sigma_x, \sigma_y$ for $i = 1, \ldots, N$:

$$y_i^{\text{obs}} \mid \boldsymbol{\theta}, x_i \sim \mathcal{N}\left(\theta_0 + \theta_1 x_i, \sigma_y^2 + \sigma_{\text{int}}^2\right), \tag{2}$$

$$x_i^{\text{obs}} \mid x_i \sim \mathcal{N}\left(0, \sigma_x^2\right), \tag{3}$$

$$x_i \sim \mathcal{N}(x_0, R_x^2), \tag{4}$$

$$x_0 \sim \mathcal{N}(\mu_{x_0}, \sigma_{x_0}^2), \tag{5}$$

$$R_x^2 \sim \text{Inv-Gamma}(\alpha_R, \beta_R), \tag{6}$$

$$\sigma_{\text{int}} \sim \text{Inv-Gamma}(\alpha_\sigma, \beta_\sigma), \tag{7}$$

where we have chosen conjugate priors for the population-level parameters $x_0, R_x^2$ (with fixed hyper-parameters $\mu_{x_0}, \sigma_{x_0}^2, \alpha_R, \beta_R, \alpha_\sigma, \beta_\sigma$) to enable analytical computation of the relevant conditional distributions.

(a) Set $x_0 = 0$ and choose fiducial values of the variables $\boldsymbol{\theta}, R_x, \sigma_{\text{int}}$ (and sensible values of $\sigma_x^2, \sigma_y^2$, ensuring that $\sigma_x \approx R_x$, as well as $\sigma_y \approx \sigma_{\text{int}}$) and generate synthetic data from the model for $N = 100$. Modify the Gibbs sampler seen in the Lab to also include the intrinsic dispersion, and produce posterior distributions for all the parameters in the model.

(b) Now we add an additional complication in the form of a selection function affecting the response variable, $y_i$: we model this effect by assuming a sigmoid as selection probability:

$$\Pr\left(I_i^{\text{obs}} | y_i\right) = \frac{1}{1 + \exp\left(-\gamma y_i\right)}, \tag{8}$$

where in order to obtain a sufficiently strong selection effect, we choose $\gamma = R_x/2$.

To sample from this model, use the complete data likelihood seen in the lecture and add an indicator variable $I_i$ to each latent object, including it in the sampling of the model. As above, produce posterior distributions for all the population variables. Produce a graph showing the simulated data (both observed and unobserved) as well as the reconstructed linear relation (with uncertainty).

3. **Bayes factor vs Likelihood Ratio Test**

We consider the problem of deciding whether to reject a null model $H_0$, with parameter $\mu = 0$, in favour of an alternative model $H_1$ with $\mu \neq 0$. Under each model, the iid random variable $X_i \sim \mathcal{N}\left(\mu, 1\right)$, $i = 1, \ldots, N$ is the observable. We would like to compare the performance of the Bayes factor (considered as a frequentist test statistics) and of the likelihood ratio test (LRT).

(a) Defining $\mathbf{X} = \{X_1, \ldots, X_N\}$, consider the Bayes factor:

$$B = \frac{\Pr(\mathbf{X} \mid H_1)}{\Pr(\mathbf{X} \mid H_0)}$$

and derive analytically the sampling distribution of $\ln B$ under $H_0$, i.e., compute the distribution of $\ln B$ over repeated data realizations when $H_0$ is the true model. Verify your analytical distribution with some numerical experiments. Compute from this the *type I error rate* (or false positive rate), $\alpha$, of a test employing $\ln B$ as test statistics: $\alpha = \Pr(\ln B > \ln B_T | H_0)$, where $\ln B_T$ is the threshold chosen to reject $H_0$.

(b) Consider now the *type II error rate* (false negatives), $\beta$, and the *power* of the test, power $= 1 - \beta = \Pr(\ln B < \ln B_T \mid H_1)$. Compute the power conditional on the choice of $\mu$ (i.e., for a fixed $\mu$) for a few choices of $\mu$ drawn from its Gaussian prior: $\mu \sim \mathcal{N}(\mu_0, \sigma_0)$ (after judicious choice of hyperparameters $\mu_0, \sigma_0$). Plot the ROC curve (i.e., power vs false positives) for those choices of $\mu$, and indicate on the graph the location of the Jeffreys' scale thresholds of evidence. Do this for a few representative choices of $\mu_0, \sigma_0$ and $N$, and discuss the results.

(c) In reality, we don't know what $\mu$ is, so the Bayesian solution is to marginalize over its prior density. Derive the power of the test marginalized over $\Pr(\mu)$ (again, for fixed hyperparameters $\mu_0, \sigma_0$), and produce the same ROC plot as in the previous step (for the same choices of $\mu_0, \sigma_0$ and $N$). Explain why the observed power of the test is small for values of $\mu_0 \approx 0$.

(d) Compare the above results with the classical LRT, testing $H_0$ vs $H_1 : \mu \neq 0$. Based on the comparison of the ROC curves (or any other plot you might think of), suggest a criterion by which to judge which test is 'preferable', and conduct a quantitative comparison between the two on this basis.

4. **SBI vs likelihood-based**

We investigate the use of Simulation-Based Inference (SBI) in a simplified setting. Consider a sinusoidal time series, with observations $y_i$ generated from the model:

$$y_i = A \sin(\omega t_i + \phi) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\boldsymbol{\theta} = \{A, \omega, \phi\}$ are the parameters of interest and $i = 1, \ldots, N$.

(a) After choosing fiducial (i.e., true) values for the parameters, the noise level $\sigma$ and $N$, generate a dataset (for given $t_i$, which can be taken to be equally spaced) and obtain the posterior distribution for $\boldsymbol{\theta}$ using an MCMC method of your choice, adopting uniform priors on the parameters. This will constitute the likelihood-based comparison.

(b) Using a Multi-Layer Perceptron (MLP), create a neural ratio estimation (NRE) network to produce marginal posteriors for each parameter in $\boldsymbol{\theta}$ and each pair of parameters. Train the network using data-parameter

pairs as described in the lecture, and once trained deploy it on the same simulated data as MCMC. In the process, you may have to use truncation to 'zoom in' the parameter ranges if your chosen prior boxes are too large for the network to learn the ratio estimator everywhere in parameter space. Produce a triangle plot figure comparing the posterior distributions obtained by the two methods.

(c) Once the NRE network has been trained to a satisfactory level, exploit its amortized nature to evaluate the coverage of its marginal posteriors. First, produce a Bayesian p-p plot for each of the parameters. Then, focusing on the 2 parameters $A, \omega$, create a map of 2-dimensional coverage, i.e., for every point in a grid in $A, \omega$ space (restricted to the truncated prior box, if necessary), generate 1000 data realizations, each with a value of $\phi$ randomly drawn from its prior; analyze each by querying the trained network and derive its posterior; then plot the observed (empirical) coverage as a function of the nominal coverage (i.e., the credibility level) for each parameter.

5. **Borrowing of strength in Bayesian hierarchical model**

We consider the problem of estimating the average number of patients admitted to ER per day from various hospitals across a city. A Bayesian hierarchical model (BHM) allows us to partially pool information across hospitals, reducing variance in small-sample groups and leveraging strength across hospitals to estimate the population-level parameters.

(a) Generate a dataset where the logarithm of the expected number of admissions per days for hospital $j$, $\ln \lambda_j$, comes from a Normal distribution:

$$\ln \lambda_j \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right), \text{ for } j = 1, \ldots, J.$$

The admission counts data for each hospital for day $i$ follow a Poisson distribution:

$$y_{ij} = \text{Poisson}(\lambda_j), \text{ for } i = 1, \ldots, n_j,$$

where each hospital $j$ has a different reporting frequency, which leads to different $n_j$ for each. To model this, you may draw $n_j$ randomly with equal probability from the set $\{52, 24, 12\}$, representing respectively weekly, bi-monthly and monthly reporting.

The DAG for this model is shown in Fig. 1. Choose appropriate hyperpriors for $\mu_0, \sigma_0$, and use an MCMC algorithm of your choice to sample from the model and produce posterior distribution on all its free parameters (a sensible choice might be $J = 10$ hospitals).

(b) Determine the marginal posterior distribution for the admission rate for each hospital, and compare it with the posterior estimate in a model with no pooling (i.e., where each hospital's rate is inferred exclusively from its observed counts). Verify that the posterior means for hospitals with a smaller number of records (i.e., $n_j = 12$) are shrunk towards the global

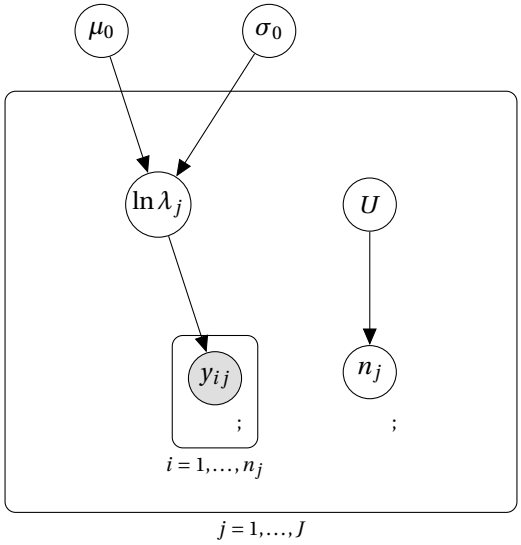


Figure 1: DAG for the Bayesian hierarchical model for hospital admissions.

mean, thus demonstrating borrowing of strength. Verify that the 68% HPD credible intervals from the BHM are shorter than in the no-pooling model, particularly for hospitals with smaller sample size. You may use a suitable violin plot to make this comparison.

(c) Consider the prior predictive distribution for possible data within the BHM:

$$\Pr\left(y_{ij} \mid \text{priors}\right) = \int \Pr\left(y_{ij} \mid \lambda_j\right) \Pr\left(\lambda_j \mid \mu_0, \sigma_0\right) \Pr\left(\mu_0, \sigma_0\right) \, d\lambda_j \, d\mu_0 \, d\sigma_0$$

and simulate from it predictions for the possible counts $y_{ij}$ from the model (before you see any data). By evaluating the ensuing spread of such counts as a function of your prior choices, determine which hyperparameter has the most effect on the degree of shrinkage, and whether such priors are too diffuse (i.e., the predictive spread is unreasonably wide) or too narrow (i.e., the predictive spread is over-constraining).

6. **Bayesian outlier detection**

In several scientific applications, measured data points can be contaminated by 'outliers' – samples that come from another distribution than the one that is assumed for the regression task. This question explores a Bayesian approach to outlier detection in a simple linear regression setting. Each of the $N$ observed data points comes from one of two different populations, called $A$ (the target population) and $B$ (the contaminants), with different (and unknown) intrinsic scatter. However, we don't know which data come from which population.

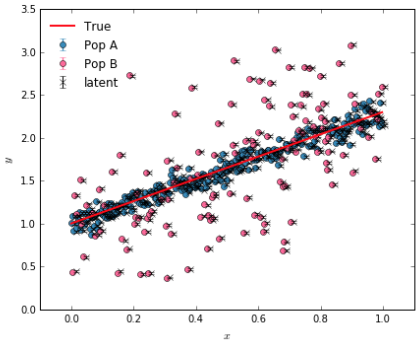


Figure 2: Example data set for the Bayesian outlier detection problem. The observed data (shown as circles) are drawn from two populations with different intrinsic scatter, and the colour indicates the true –but unknown– population membership of each datum. The latent data are shown with crosses.

The model is as follows:

$$y_i \mid \boldsymbol{\theta} = \mathcal{N}\left(\theta_0 + \theta_1 x_i, \sigma_X^2\right) \quad (i = 1, \ldots, N),$$

where $X = A, B$, and we shall assume for illustration purposes that the intrinsic dispersion of the contaminants in population B is $\sigma_B \gg \sigma_A$.

The observed values (at locations $\hat{x}_i$, assumed known), $\hat{y}_i$, have iiid Gaussian random errors of known standard deviation:

$$\hat{y}_i \mid y_i = \mathcal{N}\left(y_i, \varepsilon^2\right) \quad (i = 1, \ldots, N).$$

We have a set of data $d_i \equiv \{\hat{x}_i, \hat{y}_i\}$, $i = 1, \ldots, N$ and we want to infer $\boldsymbol{\theta} = \{\theta_0, \theta_1, \sigma_A, \sigma_B\}$. An example dataset is shown in Figure 2.

Assuming each datum is independent, after marginalising analytically over the latent $y_i$ the likelihood is:

$$\Pr(d_i \mid \boldsymbol{\theta}, X) = \frac{1}{\sqrt{2\pi}(\sigma_X^2 + \varepsilon^2)^{1/2}} \exp\left(-\frac{1}{2}\frac{(\theta_0 + \theta_1 \hat{x}_i - \hat{y}_i)^2}{\sigma_X^2 + \varepsilon^2}\right)$$

for $X = A, B$.

Introduce a latent indicator variable $\mathbf{J}$, with $\dim(\mathbf{J}) = N$ and $J_i = 1$ if observation $i$ has been generated from the distribution $A$ and $J_i = 0$ if it originates from distribution $B$. We are interested in the posterior $\Pr(\boldsymbol{\theta} \mid \boldsymbol{d})$ which can be obtained by marginalising over the latent vector $\mathbf{J}$ as:

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}) = \sum_{\mathbf{J}} \Pr(\boldsymbol{\theta}, \mathbf{J} \mid \boldsymbol{d}).$$

The joint posterior is given by:

$$\Pr(\boldsymbol{\theta}, \mathbf{J} \mid \boldsymbol{d}) = \Pr(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathbf{J}) \Pr(\boldsymbol{\theta}) \Pr(\mathbf{J}) \tag{9}$$

$$= \prod_{i \in \alpha} \Pr(d_i \mid \boldsymbol{\theta}, X = A) \Pr(J_i = 1) \prod_{j \in \overline{\alpha}} \Pr\big(d_j \mid \boldsymbol{\theta}, X = B\big) \Pr\big(J_j = 0\big) \Pr(\boldsymbol{\theta}) \,, \tag{10}$$

where $\alpha = \{1 \leq k \leq N | J_k = 1\}$ and $\overline{\alpha} = \{1 \leq k \leq N | J_k = 0\}$, and $\Pr(\boldsymbol{\theta})$ is the prior on the model's parameters. Therefore the marginal posterior for the parameters can be written:

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}) = \Pr(\boldsymbol{\theta}) \prod_i \Pr(d_i \mid \boldsymbol{\theta}, X = A) \Pr(J_i = 1) + \Pr(d_i \mid \boldsymbol{\theta}, X = B) \left(1 - \Pr(J_i = 1)\right).$$

Consider the following scenarios for the prior $\Pr(\mathbf{J})$.

(a) Assume that nothing is known about the relative probability of an observation belonging to distribution $A$ or $B$, and therefore take $\Pr(J_i = 1) = 1/2$ for all $i$. Find the posterior distribution for $\boldsymbol{\theta}$ (adopting and justifying suitable priors) in this no-pooling scenario. Plot the posterior for the parameters for a few judiciously chosen values of the quantities in the model.

(b) In a more realistic situation, we want to include the information that, on average, a fraction $\nu$ of the data points comes from population $A$, where the value of $\nu$ is also to be inferred. This can be included with the hierarchical prior:

$$\Pr(\mathbf{J} | \nu) = \nu^J (1 - \nu)^{N - J} \,, \tag{11}$$

where $J = \sum_i J_i$. For the hyperparameter $\nu$ we adopt a Beta distribution:

$$\Pr(\nu \mid \Xi) = \frac{\nu^{a-1}(1-\nu)^{b-1}}{B(a,b)} \tag{12}$$

and $\Xi = \{a, b\}$ are the parameters of the Beta distribution ($a, b > 0$) and $B(a,b)$ is the Beta function. Choosing $a = b = 1$ gives a uniform prior between 0 and 1. Discuss the difference between this scenario (with $a = b = 1$) and the one considered above (no-pooling), and compare the posterior distribution for $\boldsymbol{\theta}$ in the two cases, when $\nu = 1/2$ and $\nu = 0.8$ (after choosing sensible values for the fiducial parameters in the model; for example: $\varepsilon = 0.1$, $N = 30$, $\sigma_A = 0.05$, $\sigma_B = 0.5$, $\theta_0 = 1.0$, $\theta_1 = 1.3$, $\nu = 0.8$).

(c) Sample numerically from the joint posterior under the assumption of a hierarchical prior for $\mathbf{J}$, and compare the posterior distribution for the class membership[1] of each observation with the ground truth. Evaluate the influence of the choice of the hyper-parameters $\Xi$ on the result.

[1] To this end, a useful visualization is the posterior mean and standard deviation for $J_i$ as a function of the point's distance from the true regression line.

7. **Bayesian clustering**

We investigate a Bayesian approach to unsupervised clustering: consider a 2-dimensional dataset of samples $\mathbf{x}_i$ ($i = 1, \ldots, N$), generated from a finite Gaussian mixture with $K$ components:

$$\mathbf{x}_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mu_k, \Sigma_k\right)$$

where the mixture weights $\pi_k$ sum to unity and each mixture component has its own mean, $\mu_k$, and covariance matrix $\Sigma_k$. The total number of cluster, $K$, is unknown.

(a) Generate synthetic data for reasonable fiducial choices of parameters, using $K = 3$. For each fixed value of $K = 1, 2, \ldots, 5$, infer with Gibbs sampling (or any other sampler of your choice) the values of $\{\mu_k, \Sigma_k\}_{k=1,\ldots,K}$, using (and justifying) a suitable prior for the parameters in the problem.

(b) Consider the problem of model comparison, and denote by $M_k$ a model with $k$ mixture components. Using nested sampling, the Laplace approximation or any other suitable method, compute (or approximate) the value of the posterior probability for each model (under different separations of the $K = 3$ modes), and check whether you can recover the correct model. Perform a sensitivity analysis on the choice of priors, particularly for $\Sigma_k$.

(c) Consider now a non-parametric Dirichlet process, which allows to consider an infinite mixture model where the number of components $K \to \infty$. In this approach, each data point is given a categorical variable $c_i$, which assigns it to one of the clusters currently active, or to a new cluster, so that the model can be written as:

$$\mathbf{x}_i \mid c_i \sim \mathcal{N}\left(\mu_i, \Sigma_i\right) \tag{13}$$

$$c_i \mid \pi \sim \text{Discrete}(\pi_1, \ldots, \pi_K) \tag{14}$$

$$(\mu_i, \Sigma_i) \sim \text{NIW}(\mu_0, \lambda, \mathbf{S}, \nu) \tag{15}$$

$$\pi \mid \alpha \sim \text{Dir}(\alpha/K, \ldots, \alpha/K), \tag{16}$$

where $\alpha > 0$ is a concentration parameter that controls how often new clusters are spawned; $\text{NIW}(\mu_0, \lambda, \mathbf{S}, \nu)$ is the Normal-Inverse-Wishart distribution (the conjugate distribution to the multivariate normal with unknown mean and covariance matrix) with location parameter $\mu_0$ and inverse scale matrix $\mathbf{S}$; and $\text{Dir}(\alpha/K, \ldots, \alpha/K)$ is a Dirichlet distribution. This model is known as the 'conjugate Dirichlet process Gaussian Mixture Model (DPGMM)'.

By following the detailed steps presented in Li et al [2], build a Gibbs sampler to sample from this model (for the bivariate data generated above)

[2] Li Y, Schofield E, Gönen M., "A tutorial on Dirichlet Process mixture modeling". J Math Psychol. 2019 Aug;91:128-144, available from: `https://pmc.ncbi.nlm.nih.gov/articles/PMC6583910/`.

and compare the resulting clustering with what you obtained from the finite mixture model.

8. **Bayesian experiment optimization**

We investigate the use of a Bayesian approach to optimization of the experimental configuration to maximise a suitably defined utility. We assume that data $\boldsymbol{d}$ have been collected, so that our present knowledge of the system before we build the new experiment is described by the posterior $\Pr(\theta \mid \boldsymbol{d})$. We introduce the utility $U(\theta, \boldsymbol{d}, e)$ as an inverse loss function, which depends on what we have observed so far (the data $\boldsymbol{d}$), on the value of the model's parameters $\theta$, and on the experimental configuration of the future experiment, $e$. From the utility we can build the expected utility, by averaging the utility over the posterior obtained from the past observations, $\Pr(\theta \mid \boldsymbol{d})$.

$$\mathrm{E}[U \mid \boldsymbol{d}, e] = \int \mathrm{d}\theta \, U(\theta, \boldsymbol{d}, e) \Pr(\theta \mid \boldsymbol{d}) \, . \tag{17}$$

Consider the linear model

$$y = \theta x + \varepsilon \, ,$$

where $\varepsilon$ is normally distributed. The past experiment has been performed by measuring the quantity $y$ at two locations, $x_0, x_1$, resulting in data $\boldsymbol{d} = \{y_0, y_1\}$. We assume that the two measurements are uncorrelated, and that the noise from the past experiment is described by a Gaussian with zero mean and variance $\sigma^2$. The standard deviation of the future experiment, $e$, is described by a noise function $\tau(x)$, describing how the future experiment's accuracy varies with distance along the horizontal axis.

(a) Adopt as utility function the inverse of the posterior variance on $\theta$ from both the past experiment and $e$, and assuming constant noise, $\tau(x) = \tau$, determine what is the best choice of a single future measurement, $x_f$, in the range $[0, x_{\max}]$.

(b) Consider now the case where we could build a more accurate experiment, $a$, which could measure $y$ with noise $\tau^\star \ll \tau$, but only if the dependent variable falls around a certain value, $y^\star$. We can model this by writing for the noise of $a$ as:

$$\tau_a^2 = \tau^{\star 2} \exp\left( \frac{(y - y^\star)^2}{2\Delta^2} \right). \tag{18}$$

We consider the parameters $y^\star, \tau^\star, \Delta$ as fixed quantities. Find the optimal location $x_f$ for a new measurement and compare the performance of $a$ with the experiment considered above, $e$. Discuss how and why the result depends on the ratio $\Delta / \Delta y$, where $\Delta y$ is the present-day uncertainty in the value of $y$ at the location $x_f$. Produce numerical results for the following choices: $x_0 = 0.5$, $x_1 = 1.0$, $x_{\max} = 3$, $\sigma = 0.1$, $\theta = 1.0$, $\tau = 0.1$, $\tau_* = \sigma/5$, $y_* = 1.5$, $\Delta = 0.1$. Plot the expected utility as a function of $\Delta$

(all other quantities constant) and determine the cross-over value for $\Delta$ at which the experimental configuration to be preferred changes from $e$ to $a$.

(c) Consider now the task of optimizing the future experiment $a$ to distinguish between two models for the data: Model 1 is the linear model above, Model 2 has an additional quadratic term and is given by:

$$y = \theta x + \psi x^2 + \varepsilon\,,$$

with $\psi$ an additional parameter with a suitably chosen prior. As a utility function, adopt the volume of parameter space where the absolute value of the logarithm of the Bayes factor between the two models exceeds the threshold value of 2.5. This means that in such a region of parameter space the outcome of model comparison (using both past data and the future data point from $a$) would be above the 'moderate evidence' threshold (for either model) under the Jeffreys' scale. Determine the value of $y^\star$ (for fixed $\Delta$) that maximises such a utility function and discuss the result by comparing with maximisation of the previous utility function.