40.017 Probability and Statistics

4D Project

S.U.T.D.

2025*

Motivation

This project integrates key concepts previously covered in Bayesian modeling, bivariate normal distributions, prediction, practical regression (as introduced in MU last year), and the marginalization of probability density functions. Together, these tools form a foundational part of modern statistical analysis and are applied in a coherent framework to deepen your understanding of probabilistic reasoning and drawing inference under uncertainty.

Applications. These statistical methods are invaluable for decision-making in real-life scenarios involving uncertainty. Bayesian modeling and regression help predict outcomes in healthcare, finance, and environmental science, while bivariate normal distributions analyze correlated phenomena such as health factors or consumer behaviors. Recently, these foundational methods have also fueled the rapid development of state-of-the-art generative models, particularly *diffusion models*, whose name reflects the mathematical connection to diffusion processes and the normal distributions underlying their probabilistic formulation. Diffusion models have greatly advanced image, audio, and text generation, demonstrating superior performance in data synthesis and restoration tasks.

Instructions

- Please work in your DBA project groups. Each group *must* work independently, without collaborating with any other groups. Each group *must* write their own solutions; do not share your solutions with another group.
- As early as you can, please think about how to divide the various tasks among your group members (e.g. assign everyone clearly defined, non-overlapping tasks based on individual strengths). Your group should (physically) meet at least twice (once at the start to divide the tasks, and once at the end to finalize the report/check answers).
- You are encouraged to use *Python* for all the computations.
- You are encouraged to use Overleaf to prepare your report in LaTeX.
- If you use methods, ideas, or external resources beyond this course (including software or personal assistance), clearly describe them, justify their use, and *cite* all references.
- Any further instructions and a submission link will be uploaded on eDimension. If you have any questions, please reach out to the TAs.

*Issue: YL; 23 March, YL-YC-BD; 25 March 2025

Submission

- Each group should submit a single zip file, named 'Group$x$' (with x replaced by your group number). The zip file should contain a typed report in PDF format, limited to five A4 sized pages, as well as any relevant *Python* computations, suitably annotated. (Any cover page or reference section does not count towards the page limit.)
- You should also submit **a hard copy of your project report** to me, or place it in my pigeonhole (details to follow).
- The report should include a brief and descriptive summary of how each group member contributed to the project. Please do not just say 'each member contributed equally'. If any members did not contribute, please clearly indicate this in the report.
- The report should look more professional than a typical homework submission. Please explain your logic (in words, not just with maths) for each solution, and *show all relevant steps*, so that a reader can easily follow and reconstruct your reasoning.
- The project will be graded on:
  - Correctness, clarity, and quality of your solutions,
  - Appropriate level of working and explanation shown,
  - Evidence of mastery of the course material, even creativity,
  - Presentation, readability, and appropriate citations.
- **Deadline:** ~~20 April 2025, 11.59 PM~~, **21 April 2025, 11.59 PM**

Question 1 (10 marks)

Suppose we have a multiple linear regression model with K input features:

$$y_n = w_0 + w_1 x_{1n} + w_2 x_{2n} + \cdots + w_K x_{Kn} + \epsilon_n, \tag{1}$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$. In vector form, this corresponds to:

$$y_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n$$

where $\mathbf{w} = [w_0, w_1, \ldots, w_K]^\top$ and $\mathbf{x}_n = [1, x_{1n}, x_{2n}, \ldots, x_{Kn}]^\top$. Stack all responses into one vector $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ and all input vectors into a data matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix},$$

we obtain the full model for the dataset:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$$

with $\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^\top$ and $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. Assume we know the value of σ^2. We fix a Gaussian prior for $\mathbf{w} := [w_0, w_1, \ldots, w_K]$. We place a Gaussian prior for $\mathbf{w} := [w_0, \ldots, w_K]$. In particular we have

$$p(\mathbf{w} \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{2}$$

> Recall that the multivariate normal distribution is parameterized as $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean vector and covariance matrix are given by:
>
> $$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}$$
>
> where $\boldsymbol{\mu}$ is the d-dimensional mean vector, and $\boldsymbol{\Sigma}$ is the $d \times d$ covariance matrix.

(a) (**5 marks**) Show that the posterior distribution is given by

$$p(\mathbf{w} \,|\, \mathbf{y}, \mathbf{X}, \sigma^2) = N(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w}) \tag{3}$$

with $\boldsymbol{\Sigma}_\mathbf{w}$ and $\boldsymbol{\mu}_\mathbf{w}$ take the form

$$\boldsymbol{\Sigma}_\mathbf{w} := \left(\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1}, \tag{4}$$

$$\boldsymbol{\mu}_\mathbf{w} := \boldsymbol{\Sigma}_\mathbf{w} \left(\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{y} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right). \tag{5}$$

(b) (**5 marks**) We now turn our attention to making predictions. Given a new observation $\mathbf{x}_\flat$, we are interested in the *posterior predictive distribution*

$$p(y_\flat \,|\, \mathbf{x}_\flat, \mathbf{X}, \mathbf{y}, \sigma^2) \tag{6}$$

with $y_\flat$ being the unobserved data. Show that this density is a Gaussian distribution with parameters

$$\mu_\flat := \mathbf{x}_\flat^\top \boldsymbol{\mu}_\mathbf{w}, \tag{7}$$
$$\sigma_\flat^2 := \mathbf{x}_\flat^\top \boldsymbol{\Sigma}_\mathbf{w} \mathbf{x}_\flat + \sigma^2. \tag{8}$$

Hint: You may wish to leverage the following standard results for marginal and conditional Gaussians:

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = N(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{9}$$
$$p(\mathbf{y} \,|\, \mathbf{x}) = N(\mathbf{y} \,|\, \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{10}$$

then the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is given by

$$p(\mathbf{y}) = N(\mathbf{y} \,|\, \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top) \tag{11}$$

Question 2 (40 marks)

Suppose we take a Bayesian approach to modeling a simplified linear relationship between input and output data:

$$y_n = w_0 + w_1 x_n + \epsilon_n, \tag{12}$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$. In this model, w_0 is the **intercept** — a constant bias term that shifts the line up or down, and w_1 is the **slope** — it determines how strongly, and in what direction, the output y_n changes with the input x_n. This can be written in compact vector form as:

$$y_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n,$$

where $\mathbf{w} = [w_0, w_1]^\top$ is the weight vector, and $\mathbf{x}_n = [1, x_n]^\top$ is the input vector (including a bias term).

Let the observed input and output data be:

$$x^\top = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]$$
$$y^\top = [11.99, 11.73, 11.49, 11.25, 11.02, 10.77, 10.52, 10.24, 9.95, 9.64,$$
$$9.32, 8.98, 8.65, 8.33, 8.06, 7.71, 7.41, 7.13, 6.90, 6.59,$$
$$6.39, 6.10, 5.87, 5.67, 5.47, 5.29, 5.12, 4.95, 4.79, 4.64]$$

(a) (**5 marks**) This task concerns sampling from the prior distribution — i.e., computing the corresponding model outputs $\hat{y}_n^{(i)}$, and plotting these outputs together with the observed data. Follow these steps:

Define a prior distribution over the weight vector:

$$\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),$$

where

$$\mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma_0 = \begin{bmatrix} 100 & 0 \\ 0 & 5 + \frac{\sin(x)}{x} \end{bmatrix} \tag{13}$$

where x is your DBA group number. The following specific values for the prior (in Python syntax) is as follows:

```
mu0 = np.array([0, 0])
si0 = np.array([[100, 0], [0, 5 + sin(x)/x]])
```

Sample from the prior: Use the following Python function

```
np.random.multivariate_normal(mu0, si0, num_samples)
```

to draw 10 samples of $\mathbf{w}$ from this distribution.

Plot the observed data (x, y) as points. Then, for each sampled weight vector $\mathbf{w}^{(i)}$ (for $i = 1, \ldots, 10$), compute the predicted outputs using the linear model:

$$\hat{y}_n^{(i)} = \mathbf{w}^{(i)\top} \mathbf{x}_n,$$

where each input vector is $\mathbf{x}_n = [1, x_n]^\top$. Overlay all 10 predicted functions $\hat{y}^{(i)}$ on the same plot, together with the real data, to visualize the variability in predictions implied by the prior over weights.

(b) (**15 marks**) Using $\sigma^2 = 8$, compute the posterior distribution after observing one data point, i.e.,

$$\mathbf{x}_1 = [1, 0]^\top, \quad y_1 = 11.99.$$

In particular:

(a) Write down the posterior distribution $p(\mathbf{w} \mid \mathbf{x}_1, y_1)$ explicitly, specifying its mean vector and covariance matrix.

(b) Plot 10 samples from the posterior predictive distribution together with the real data.

(c) Where is w_0 centered, and why? Explain the figure: why do all sampled lines intersect at a common point?

(c) (**10 marks**) Using $\sigma^2 = 8$, compute the posterior distribution after observing all data points. Plot 10 samples from the posterior predictive distribution together with the real data.

(a) Write down the posterior distribution $p(\mathbf{w} \mid \mathbf{X}, \mathbf{y})$ explicitly, specifying its mean vector and covariance matrix. You may use matrix notation for compactness.

(b) Plot 10 samples from the posterior predictive distribution over the input range, and overlay these predictions with the observed data to visualize how the posterior has been updated after observing all data points.

(d) (**10 marks**) Using $\sigma^2 = 8$, and given a new data point of $x_{\text{new}} = 31$, write down the predictive posterior distribution, specifying its mean and covariance. Then, plot the predictive posterior distribution.