BIOENG-210: Biological Data Science I: Statistical Learning

Theoretical Exercise Week 9
Prof. Gioele La Manno

April 2024

1 MLE and MAP for Linear Models

In all of the following parts, write your answer as the solution to a norm minimization problem, potentially with a regularization term. **You do not need to solve the optimization problem.** Simplify any sums using matrix notation for full credit.

Hint: Recall that the MAP estimator maximizes $P(\boldsymbol{\theta}|Y)$.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} P(Y|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

The difference between MAP and MLE is the inclusion of a prior distribution on $\boldsymbol{\theta}$ in the objective function.

For the following problems assume you are given $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$ as your data.

(a) Let $y = X\boldsymbol{\theta} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \Sigma)$ for some positive definite, diagonal Σ. Write the MLE estimator of $\boldsymbol{\theta}$ as the solution to a weighted least squares problem, potentially with a regularization term.

Solution:

We are given:

$$y \sim \mathcal{N}(X\boldsymbol{\theta}, \Sigma)$$

The likelihood of the data is:

$$p(y|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y - X\boldsymbol{\theta})^\top \Sigma^{-1}(y - X\boldsymbol{\theta})\right)$$

The log-likelihood is:

$$\log p(y|\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(y - X\boldsymbol{\theta})^\top \Sigma^{-1}(y - X\boldsymbol{\theta})$$

Equivalently, the optimization problem becomes:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} (y - X\boldsymbol{\theta})^\top \Sigma^{-1} (y - X\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\| \Sigma^{-1/2} (y - X\boldsymbol{\theta}) \right\|_2^2$$

This is a **weighted least squares** problem, where the weights are determined by Σ.

Extra: How to derive the solution! To find the MLE, we take the gradient of the log-likelihood with respect to $\boldsymbol{\theta}$ and set it to zero:

$$\nabla_{\boldsymbol{\theta}} \log p(y|\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \left(-\frac{1}{2} (y - X\boldsymbol{\theta})^\top \Sigma^{-1} (y - X\boldsymbol{\theta}) \right)$$

Using matrix calculus:

$$\nabla_{\boldsymbol{\theta}} \left[(y - X\boldsymbol{\theta})^\top \Sigma^{-1} (y - X\boldsymbol{\theta}) \right] = -2X^\top \Sigma^{-1} (y - X\boldsymbol{\theta})$$

Setting the gradient to zero:

$$X^\top \Sigma^{-1} (y - X\boldsymbol{\theta}) = 0 \Rightarrow X^\top \Sigma^{-1} y = X^\top \Sigma^{-1} X\boldsymbol{\theta} \Rightarrow \hat{\boldsymbol{\theta}} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y$$

This is the closed-form solution.

(b) Let $y|\boldsymbol{\theta} \sim \mathcal{N}(X\boldsymbol{\theta}, \Sigma)$ for some positive definite, diagonal Σ. Let $\boldsymbol{\theta} \sim \mathcal{N}(0, \lambda I_d)$ for some $\lambda > 0$ be the prior on $\boldsymbol{\theta}$. Write the MAP estimator of $\boldsymbol{\theta}$ as the solution to a weighted least squares minimization problem, potentially with a regularization term.

Solution:

We wish to find the MAP (maximum a posteriori) estimator:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^d} P(\boldsymbol{\theta} \mid y) = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^d} P(y \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})$$

Since both the likelihood and prior are Gaussian, we can write them explicitly:

$$P(y \mid \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (y - X\boldsymbol{\theta})^\top \Sigma^{-1} (y - X\boldsymbol{\theta}) \right)$$

$$P(\boldsymbol{\theta}) = \frac{1}{(2\pi\lambda)^{d/2}} \exp\left(-\frac{1}{2} \boldsymbol{\theta}^\top (\lambda I)^{-1} \boldsymbol{\theta} \right) = \frac{1}{(2\pi\lambda)^{d/2}} \exp\left(-\frac{1}{2\lambda} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right)$$

Taking the negative log of the posterior (dropping constants that don't depend on $\boldsymbol{\theta}$), the MAP estimator becomes:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left[\frac{1}{2} (y - X\boldsymbol{\theta})^\top \Sigma^{-1} (y - X\boldsymbol{\theta}) + \frac{1}{2\lambda} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right]$$

This is a regularized weighted least squares problem. To make this explicit, observe that:

$$(y - X\boldsymbol{\theta})^\top \Sigma^{-1}(y - X\boldsymbol{\theta}) = \left\|\Sigma^{-1/2}(y - X\boldsymbol{\theta})\right\|_2^2$$

So the optimization problem becomes:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\|\Sigma^{-1/2}(y - X\boldsymbol{\theta})\right\|_2^2 + \frac{1}{\lambda}\|\boldsymbol{\theta}\|_2^2$$

This is **weighted ridge regression**.

(c) Let $y = X\boldsymbol{\theta} + \boldsymbol{\epsilon}$ where $\epsilon_i \overset{i.i.d.}{\sim}$ Laplace$(0, 1)$. Recall that the pdf for Laplace(μ, b) is $p(x) = \frac{1}{2b}\exp\left(-\frac{1}{b}|x - \mu|\right)$. Write down the MLE estimator of $\boldsymbol{\theta}$ as the solution to a norm minimization optimization problem.

Solution:

Since $\epsilon_i \sim$ Laplace$(0, 1)$, each observation y_i is distributed as:

$$y_i \sim \text{Laplace}(x_i^\top\boldsymbol{\theta}, 1) \quad \text{with density} \quad p(y_i \mid \boldsymbol{\theta}) = \frac{1}{2}\exp\left(-|y_i - x_i^\top\boldsymbol{\theta}|\right)$$

Assuming the observations are independent, the likelihood function is:

$$P(y \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{2}\exp\left(-|y_i - x_i^\top\boldsymbol{\theta}|\right) = \left(\frac{1}{2}\right)^n \exp\left(-\sum_{i=1}^{n}|y_i - x_i^\top\boldsymbol{\theta}|\right)$$

To find the maximum likelihood estimate, we maximize the log-likelihood, or equivalently, minimize the negative log-likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log P(y \mid \boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}|y_i - x_i^\top\boldsymbol{\theta}|$$

This can be written more compactly using the ℓ_1 norm:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|y - X\boldsymbol{\theta}\|_1$$

This is a **least absolute deviations (LAD)** estimator.

(d) Let $y|\boldsymbol{\theta} \sim \mathcal{N}(X\boldsymbol{\theta}, \Sigma)$ for some positive definite, diagonal Σ. Let $\theta_i \overset{i.i.d.}{\sim}$ Laplace$(0, \lambda)$ for some positive scalar λ. Write the MAP estimator of $\boldsymbol{\theta}$ as the solution to a weighted least squares minimization problem, potentially with a regularization term.

Solution:

As before, the likelihood is Gaussian:

$$P(y \mid \boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}(y - X\boldsymbol{\theta})^\top \Sigma^{-1}(y - X\boldsymbol{\theta})\right)$$

The prior on each θ_i is independent Laplace:

$$P(\boldsymbol{\theta}) = \prod_{i=1}^{d} \frac{1}{2\lambda} \exp\left(-\frac{1}{\lambda}|\theta_i|\right) = \left(\frac{1}{2\lambda}\right)^d \exp\left(-\frac{1}{\lambda}\|\boldsymbol{\theta}\|_1\right)$$

So, ignoring the constants which are not dependent on $\boldsymbol{\theta}$, the posterior is proportional to:

$$P(y \mid \boldsymbol{\theta})P(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}(y - X\boldsymbol{\theta})^\top \Sigma^{-1}(y - X\boldsymbol{\theta}) - \frac{1}{\lambda}\|\boldsymbol{\theta}\|_1\right)$$

Taking the negative log and dropping constants:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left[\frac{1}{2}(y - X\boldsymbol{\theta})^\top \Sigma^{-1}(y - X\boldsymbol{\theta}) + \frac{1}{\lambda}\|\boldsymbol{\theta}\|_1\right]$$

To write this as a norm minimization problem, observe that:

$$(y - X\boldsymbol{\theta})^\top \Sigma^{-1}(y - X\boldsymbol{\theta}) = \left\|\Sigma^{-1/2}(y - X\boldsymbol{\theta})\right\|_2^2$$

Thus, the MAP estimator becomes:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\|\Sigma^{-1/2}(y - X\boldsymbol{\theta})\right\|_2^2 + \frac{2}{\lambda}\|\boldsymbol{\theta}\|_1$$

This is **weighted LASSO regression**.

2 Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be independent samples from the following distribution:

$$P(x \mid \theta) = \theta x^{-\theta-1} \quad \text{where } \theta > 1,\ x \geq 1$$

Find the maximum likelihood estimator of θ.

Solution:

$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \theta x_i^{-\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{-\theta-1}$$

$$\ln L(x_1, x_2, \ldots, x_n \mid \theta) = n \ln \theta - (\theta + 1) \sum_{i=1}^{n} \ln x_i$$

$$\frac{\delta \ln L}{\delta \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i = 0$$

$$\theta_{\text{mle}} = \frac{n}{\sum_{i=1}^{n} \ln x_i}$$

Since $\theta > 1$, any $\theta_{\text{mle}} \leq 1$ has a zero probability of generating any data, so our best estimate of θ when $\theta_{\text{mle}} \leq 1$ is $\theta_{\text{mle}} = 1$. Therefore, the final answer is $\theta_{\text{mle}} = \max\left(1, \frac{n}{\sum_{i=1}^{n} \ln x_i}\right)$.

However, we will still accept $\theta_{\text{mle}} = \frac{n}{\sum_{i=1}^{n} \ln x_i}$.

3 Linear models and linear transformation

In this exercise we are going to see how the solution to the least squares problem changes when a linear transformation is applied to the input features X. Recall that in linear regression given $X \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ we aim to find the set of coefficients $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ that minimizes:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\hat{\boldsymbol{\beta}}} ||\boldsymbol{y} - X\hat{\boldsymbol{\beta}}||_2^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} X_{ij} \beta_j)^2 \tag{1}$$

For simplicity, we can define $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$. Recall that the solution is given by:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \tag{2}$$

For all the exercises, assume that X is full rank and therefore $X^T X$ is invertible and also $n > d$. We would like to linearly transform our features, so that we obtain a new set of features $X' \in \mathbb{R}^{n \times d'}$. If our the matrix defining our linear transformation is $A \in \mathbb{R}^{d \times d'}$, the transformed features X are simply given by:

$$X' = XA \tag{3}$$

Now we would like to find the set of coefficients $\hat{\boldsymbol{\beta}}'$ that minimize:

$$\hat{\boldsymbol{\beta}}' = \text{argmin}_{\hat{\boldsymbol{\beta}}'} ||\boldsymbol{y} - X'\hat{\boldsymbol{\beta}}'||_2^2 \tag{4}$$

a) Write down the solution to 4, that is, what is the optimal $\hat{\boldsymbol{\beta}}'$ in terms of X' and $\boldsymbol{y}$.

Solution: The problem in 4 is the same as 1. Therefore, the solution is analogous, substituing X by X':

$$\hat{\boldsymbol{\beta}}' = (X'^T X')^{-1} X'^T \boldsymbol{y}$$

We will now try to relate this solution to 2.

b) Substitute $X' = XA$ to the expression found for $\hat{\boldsymbol{\beta}}'$. You will not be able to simply much. Hint: Remember that given two matrices A, B, $(AB)^T = B^T A^T$.

Solution: If we substitute:

$$\hat{\hat{\boldsymbol{\beta}}}' = (A^T X^T X A)^{-1} A^T X^T \boldsymbol{y}$$

From now on, assume that $d = d'$ and that A is full rank (thus invertible). This assumption is equivalent to saying that we transform the data "without loss of information".

c) Show that $\hat{\boldsymbol{\beta}}' = A^{-1}\hat{\boldsymbol{\beta}}$ Hint: The same property as before also holds for the inverse $(AB)^{-1} = B^{-1}A^{-1}$ if A and B are full rank and squared.

Solution: Following the hint $(ABC)^{-1} = C^{-1}(AB)^{-1} = C^{-1}B^{-1}A^{-1}$. We apply this to the previous expression:

$$(A^T X^T X A)^{-1} A^T X^T \boldsymbol{y} = A^{-1}(X^T X)^{-1}(A^T)^{-1} A^T X^T \boldsymbol{y} = A^{-1}(X^T X)^{-1} X^T \boldsymbol{y}$$

Using 2:

$$\hat{\boldsymbol{\beta}}' = A^{-1}\hat{\boldsymbol{\beta}}$$

d) Show that the predictions of the model do not change if we fitted with the transformed data X'.

Solution: The predictions of the model fitted with the original data are $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$. The predictions of the model fitted with X' are $bm\hat{y}' = X'\hat{\boldsymbol{\beta}}'$

We substitute from the previous exercise and $X' = XA$:

$$\hat{\boldsymbol{y}}' = X'\hat{\boldsymbol{\beta}}' = (XA)A^{-1}\hat{\boldsymbol{\beta}} = X\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}$$

e) In part c, we have assumed $d' = d$. What would happen to the solution to the least squares problem in the case $d' > d$? Hint: $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$

Solution: In the case $d' > d$, the matrix X' would be at most of $\text{rank}(d)$ while having d' columns. That is, one of the could be expressed as a linear combination of the rest. This is a case of multicollinearity (see Notes 6) and the least squares problem would not have a unique solution.

f) Finally, what would you intuitively think happens in the case $d' < d$. Try to reason in terms of model performance when comparing the model fitted with X and XA (with $d' < d$).

Solution: Applying a linear transformation resulting in a lower amount of features can be seen as a "loss of information". We would expect the solution to the problem with X' to have a higher mean squared error than the one with X. (In the best case, if we lose information that is not useful to predict $\boldsymbol{y}$ we might get the same MSE than with X).

4 Statistical Properties of the Uniform Distribution

Consider a continuous uniform distribution defined on the interval $[a, b]$ with length $L = b - a$.

1. Derive the probability density function (PDF) of this uniform distribution.

   For a continuous uniform distribution on $[a, b]$, the PDF is given by:

$$f(x) = \begin{cases} \frac{1}{L} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

2. Calculate the expectation value (mean) of this distribution and express it as a function of L and a.

$$\begin{aligned} \mathbb{E}[X] &= \int_a^b x \cdot \frac{1}{L}\, dx = \frac{1}{L} \cdot \left[\frac{x^2}{2}\right]_a^b = \frac{1}{L} \cdot \left(\frac{b^2 - a^2}{2}\right) \\ &= \frac{(b+a)(b-a)}{2L} = \frac{(a+b)}{2} = a + \frac{L}{2} \end{aligned}$$

3. Calculate the variance of this distribution and express it as a function of L only.

$$\mathbb{E}[X^2] = \int_a^b x^2 \cdot \frac{1}{L}\, dx = \frac{1}{L} \cdot \left[\frac{x^3}{3}\right]_a^b = \frac{b^3 - a^3}{3L}$$

$$\begin{aligned} \text{Since } \mathbb{E}[X] = \frac{a+b}{2}, \text{ then } \mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{(b-a)^2}{12} = \frac{L^2}{12} \end{aligned}$$

5 James-Stein estimator

Problem 1: Setup the Multivariate Normal Model

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ where each $X_i \sim N(\theta_i, \sigma^2)$ independently.

a) What is the MLE for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$? (**Hint**: Note that each variable X_i has a different mean.)

b) Show that the risk (mean squared error) of the MLE, $R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2] = p\sigma^2$, where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

Problem 2: Introduce the James-Stein Estimator

Define a James-Stein estimator:

$$\hat{\boldsymbol{\theta}}^{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|\boldsymbol{X}\|^2}\right) \boldsymbol{X},$$

where $\|\boldsymbol{X}\|^2 = \sum_{i=1}^{p} X_i^2$. Compute the condition that p should satisfy so that the shrinkage factor is positive?

Problem 3: Classical vs. Shrinkage Estimators

a) Mention the trade-off we make in terms of bias and variance between the JS estimator and the MLE.

b) Explain the importance of the James-Stein estimator in practical applications. Where might we expect it to outperform traditional methods, and why?

Solution

1a. We have $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$. Therefore, the joint likelihood is,

$$L(\theta) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(X_i - \theta_i)^2}{2\sigma^2})$$

Taking logs,

$$log(L(\theta)) = -\frac{p}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{p} (X_i - \theta_i)^2$$

The log likelihood is maximized when $\sum_{i=1}^{p} (X_i - \theta_i)^2$ is minimized. This happens when $\theta_i = X_i$ for all i. Therefore the MLE is,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{X}$$

1b. Since $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \sigma^2 I)$, the MSE is:

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{\theta}\|^2] = \sum_{i=1}^{p} \mathbb{E}[(X_i - \theta_i)^2] = p\sigma^2.$$

2. The shrinkage factor $1 - \frac{(p-2)\sigma^2}{\|\boldsymbol{X}\|^2}$ is positive when $p > 2$ and $\|\boldsymbol{X}\|^2 > 0$.

3a. The JS estimator introduces some bias, but reduces the overall variance in high-dimensional data.

3b. The JS estimator is useful when estimating many related parameters, such as gene expression levels or image intensities, where pooling or shrinkage helps control overfitting and noise.

6 Pen-and-Paper PCA Exercise

Exercise

Consider the following dataset of four observations in two dimensions:

Obs.	x	y
1	1	2
2	2	1
3	3	4
4	4	3

1. Compute the sample means $\bar{x}$ and $\bar{y}$.
2. Center the data by subtracting $(\bar{x}, \bar{y})$ from each point.
3. Form the sample covariance matrix

$$S = \frac{1}{n-1} \sum_{i=1}^{4} \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix} (x_i - \bar{x},\ y_i - \bar{y}).$$

4. Solve for the eigenvalues λ_1, λ_2 of S.
5. Find corresponding (unit) eigenvectors $v^{(1)}, v^{(2)}$.
6. Compute the proportion of total variance explained by each principal component.
7. Project each centered point onto the first principal component.
8. Sketch the centered data, overlay the PC axes, and draw the ellipse with semi-axes $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$.

Solution

(1) Sample means:

$$\bar{x} = \frac{1+2+3+4}{4} = 2.5, \quad \bar{y} = \frac{2+1+4+3}{4} = 2.5.$$

(2) Centered data:

$$(x_i - \bar{x}, y_i - \bar{y}) \;=\; \{(-1.5,-0.5),\, (-0.5,-1.5),\, (0.5,1.5),\, (1.5,0.5)\}.$$

(3) Covariance matrix:

$$S = \frac{1}{3}\sum_{i=1}^{4} \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix} (x_i - \bar{x},\; y_i - \bar{y}) = \frac{1}{3}\begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix} = \begin{pmatrix} 5/3 & 1 \\ 1 & 5/3 \end{pmatrix}.$$

(4) Eigenvalues of $\begin{pmatrix} a & b \\ b & a \end{pmatrix}$ are $\lambda = a \pm b$. Here $a = 5/3, b = 1$, so

$$\lambda_1 = \frac{5}{3} + 1 = \frac{8}{3}, \quad \lambda_2 = \frac{5}{3} - 1 = \frac{2}{3}.$$

(5) Eigenvectors solve $(S - \lambda I)v = 0$:

$$\lambda_1 = \tfrac{8}{3}: \; (5/3 - \tfrac{8}{3})v_1 + 1 \cdot v_2 = 0 \;\Longrightarrow\; -v_1 + v_2 = 0 \;\Rightarrow\; v^{(1)} \propto (1,1).$$

$$\lambda_2 = \tfrac{2}{3}: \; (5/3 - \tfrac{2}{3})v_1 + 1 \cdot v_2 = 0 \;\Longrightarrow\; v_1 + v_2 = 0 \;\Rightarrow\; v^{(2)} \propto (1,-1).$$

Normalizing gives $v^{(1)} = \frac{1}{\sqrt{2}}(1,1),\; v^{(2)} = \frac{1}{\sqrt{2}}(1,-1)$.

(6) Total variance $= \lambda_1 + \lambda_2 = \frac{8}{3} + \frac{2}{3} = \frac{10}{3}$. Proportions:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{8/3}{10/3} = 0.80, \quad \frac{\lambda_2}{10/3} = 0.20.$$

(7) Projection onto PC1:

$$\text{score}_{i,1} = v^{(1)\,T}(x_i - \bar{x}, y_i - \bar{y}) = \frac{1}{\sqrt{2}}\big[(x_i - \bar{x}) + (y_i - \bar{y})\big].$$

For obs. 1: $\frac{-1.5+(-0.5)}{\sqrt{2}} = -\frac{2}{\sqrt{2}} \approx -1.414$.