Teacher : Gioele La Manno
Mock Exam BIOENG-210: Biological data science
I - MA
22th May 2025
1h30

AC-345

Student 1

SCIPER : **999000** Room : **R-A** Signature :

Do not turn the page before the start of the exam. This document is double-sided, has 10 pages, the last ones possibly blank. Do not unstaple.

- Place your student card on your table.
- **No other paper materials** are allowed to be used during the exam.
- **You may use a calculator.**
- The exam only contains muliple choice questions. Each of them has a unique single correct answer. The point distribution is the following :

  1 points if your answer is correct,
  0 points if you give no answer or more than one,
  0 points if your answer is incorrect.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question is wrong, the teacher may decide to nullify it.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

choisir une réponse | select an answer Antwort auswählen	ne PAS choisir une réponse | NOT select an answer NICHT Antwort auswählen	Corriger une réponse | Correct an answer Antwort korrigieren
☒ ☑ ■	□	⬚

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

▩ ☒ ○ ⊡ ▨ □

Formulas

Probability Mass Function for a random variable X following a Poisson distribution where λ is both mean and variance

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

James-Stein Estimator: if estimating simultaneously $p \geq 3$ means noted Y_i then the following estimation is more accurate than Maximum Likelihood Estimation:

$$\hat{\mu}_i^{JS} = \left(1 - \frac{(p-2)}{\sum_{j=1}^{p} Y_j^2}\right) \cdot Y_i$$

Complete Linkage: distance between farthest points

$$d(C_i, C_j) = \max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})$$

Sample covariance matrix: with $\mathbf{X}_c$ the centered data matrix and n the number of observations

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \mathbf{X}_c^T \mathbf{X}_c$$

Multiple choice questions

For each question, mark the box corresponding to the correct answer.

Question [SCQ-01] In multivariate statistics, covariance matrices play a fundamental role in capturing the relationships between random variables. For 2 variables, the covariance matrix will be a 2×2 matrix. However, not all 2×2 matrices are valid covariance matrices. Which of these 2×2 matrices is NOT a valid covariance matrix?

■ $\begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}$

□ $\begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix}$

□ $\begin{pmatrix} 4 & 3 \\ 3 & 9 \end{pmatrix}$

□ $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$

Question [SCQ-02] Let Σ be a covariance matrix and Σ_{ij} its entry at column i, row j. Given $\Sigma_{12} = 4$, $\Sigma_{11} = 16$ and $\Sigma_{22} = 9$, the correlation ρ_{12} is:

■ $\frac{1}{3}$

□ $\frac{4}{25}$

□ $\frac{1}{9}$

□ $\frac{4}{144}$

Question [SCQ-03] In multivariate statistics, we work with the sample covariance matrix Σ calculated from a dataset with n observations across p variables. What happens to the sample covariance matrix when p (number of variables) exceeds n (number of observations)?

■ It becomes singular (non-invertible) as $\mathrm{rank}(\Sigma) \leq n - 1 < n < p$)

□ It remains full rank but loses positive definiteness (meaning that some eigenvalues become negative)

□ It inverts to give a well-defined precision matrix

□ It factors into a product of lower-dimensional covariances

Question [SCQ-06] $\mathbf{I_p}$ is the identity matrix in p dimensions. If $\boldsymbol{\Sigma} = \mathbf{I}_p$ for a p-dimensional normal distribution, the level sets of the distribution are:

■ A p-dimensional sphere (hypersphere)

□ Arbitrarily oriented ellipsoids

□ A block-diagonal “box”

□ A single point at the mean

Question [SCQ-07] Given the Singular Value Decomposition of a given data matrix $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$, which is true about the right singular vectors $\mathbf{V}$?

■ They are eigenvectors of $\mathbf{X}^T\mathbf{X}$ (the feature covariance)

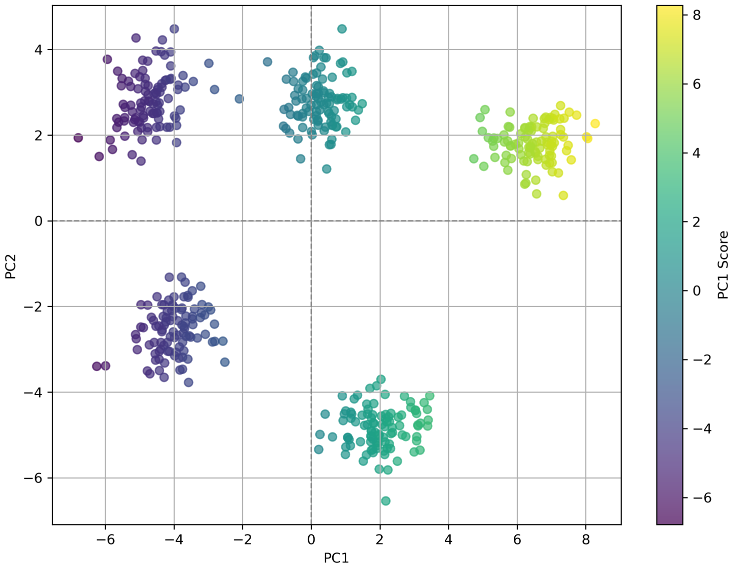
□ They are eigenvectors of $\mathbf{X}\mathbf{X}^T$

□ They are eigenvectors of both $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$

□ They diagonalize $\mathbf{X}\mathbf{X}^T$

□ They equal the left singular vectors $\mathbf{U}$

Question [SCQ-09] A researcher performs Principal Component Analysis on a data set containing the expression of several genes across many cells. The plot of PC scores shows cells as points.



In this plot, the fact that the clusters are tight (cells seem to group close together within each cluster, and the different clusters seem well-separated) indicates:

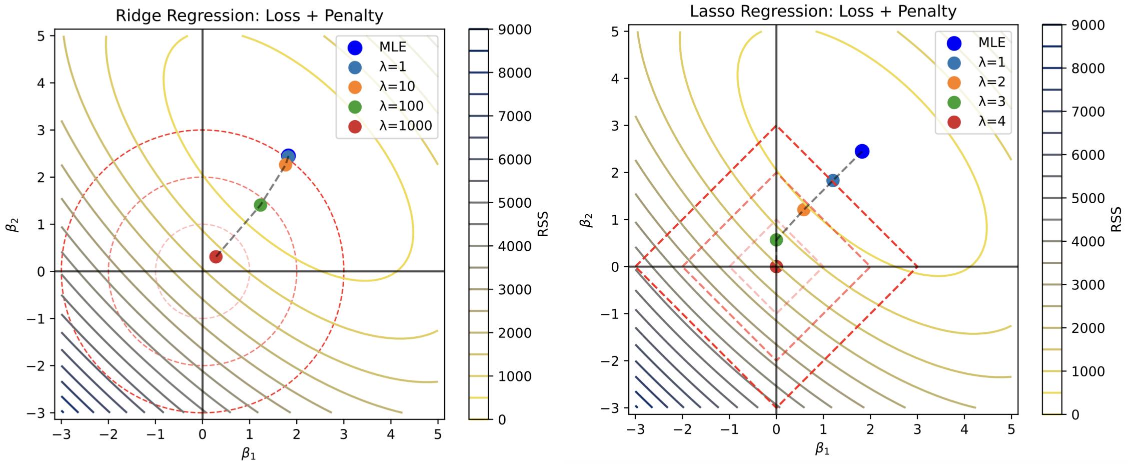
■ That biological groups differ along the principal axes

□ That technical noise dominates the first principal components

□ That Principal Component Analysis failed to reduce dimensionality

□ That genes have identical expression across cells

Question [SCQ-13] The following plots show (in low dimensions) one of properties that characterizes Lasso regularization compared to Ridge regularization.



What is this property?

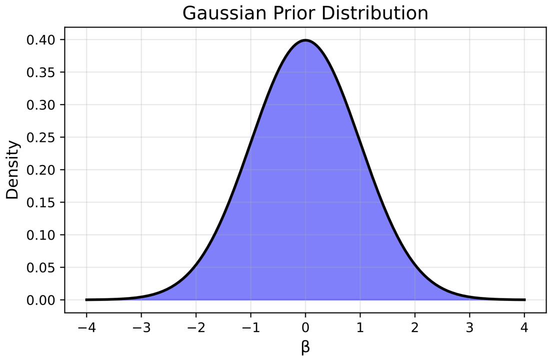
■ Lasso tends to return more coeffients equal to zero.

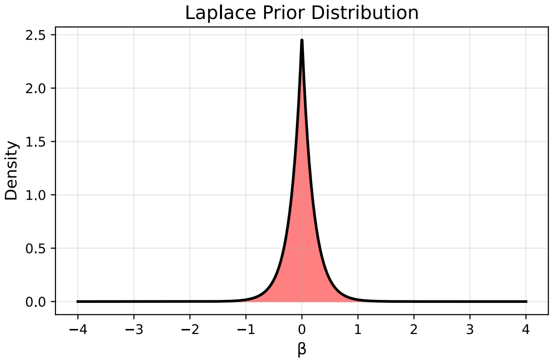
□ Lasso always yields a lower Mean Squared Error than ridge penalty.

□ Lasso does not penalize $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ via absolute value, while ridge does.

□ Lasso uses a gaussian prior, while ridge does not.

Question [SCQ-14] In the context of Maximum A Posteriori Estimation, the Laplace prior density has a sharper peak at zero than the Gaussian prior.





This geometric feature explains why:

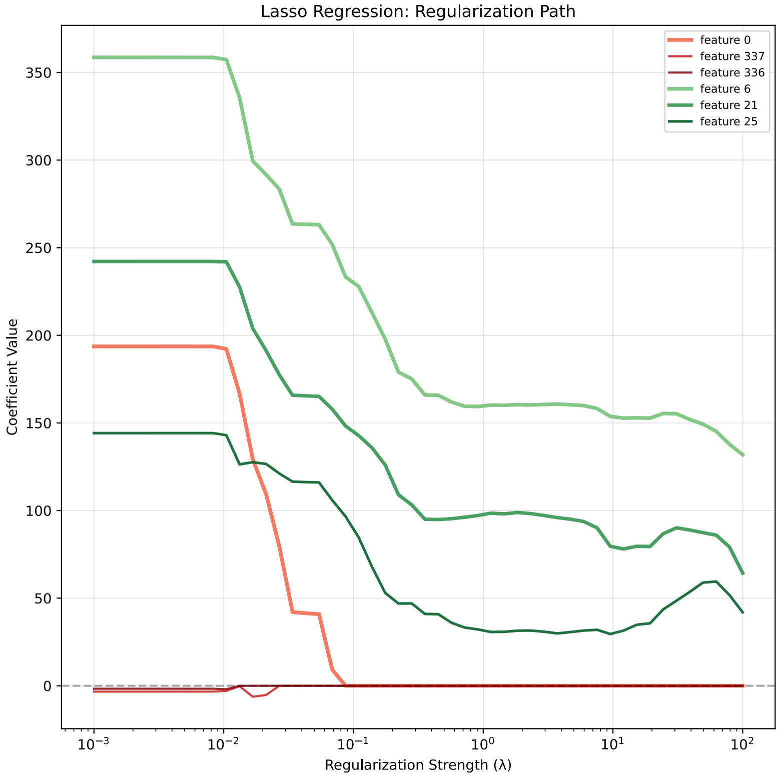
■ A Laplace prior tends to return fewer non-zero coefficients.

□ Ridge yields exact zeros, lasso does not

□ Both yield identical sparsity

□ Lasso always outperforms ridge in prediction

Question [SCQ-18] The lasso regularization path shows many coefficients hitting zero at different λ.



A feature that reaches zero very early (small λ) is likely:

- ■ Less predictive, easily penalized out
- □ Highly predictive, protected by the penalty
- □ Constant across samples
- □ Perfectly correlated with another feature

Question [SCQ-19] Given d-dimensional data $\{\mathbf{x}_i\}_{i=1}^N$, you run principal component analysis and pick P principal components. You want to try to rebuild the original data points using only the information captured by the P principal components. To perform this reconstruction, you express each data point as a linear combination of the P principal components. Can you always reconstruct any data point $\mathbf{x}_i$ for $i \in \{1, \dots, N\}$ from the P principal components with zero reconstruction error?

- □ Yes, if $P < d$
- □ Yes, if $P < N$
- ■ Yes, if $P = d$
- □ No, it's impossible to reconstruct with zero error.

Question [SCQ-21] You have a sample mean $\bar{x}$ from n observations and an unknown population variance. Which test statistic and null distribution should you use to test $H_0 : \mu = \mu_0$?

- □ $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ with $N(0, 1)$
- □ $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with $t(n)$
- ■ $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with $t(n-1)$
- □ $F = \frac{s^2}{\sigma^2}$ with F-distribution

Question [`SCQ-22`] In a linear regression where we aim to predict values of Y based on the predictor X, the width of a confidence interval for the mean response at a specific predictor value $X = x_0$, denoted $\hat{y}_0$, depends on several factors. Which of the following changes would not lead to a narrower 95% confidence interval for the mean response at $X = x_0$?

□ Increasing the sample size n

□ Reducing the estimated error variance $\hat{\sigma}^2$

□ Moving x_0 closer to the sample mean $\bar{x}$

■ Decreasing the variability of the predictor $\sum(x_i - \bar{x})^2$

Question [`SCQ-23`] A biologist collects gene expression data from 50 cells and wants to determine whether the expression of a particular gene follows a Poisson distribution. After fitting a Poisson model using Maximum Likelihood Estimation (MLE), she finds that the MLE for the rate parameter λ is 10.5. Which of the following statements is correct about this estimate?

■ The MLE of $\lambda = 10.5$ represents the value that maximizes the probability of observing the given data under a Poisson model

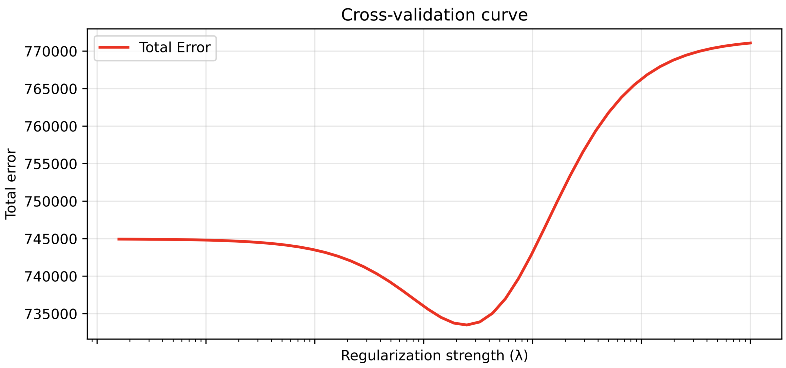
□ The MLE of $\lambda = 10.5$ means that exactly 10.5 molecules are expressed in each cell

□ The MLE of $\lambda = 10.5$ guarantees that the gene follows a Poisson distribution

□ The MLE of $\lambda = 10.5$ means that the log-likelihood function equals 10.5 at its maximum

□ The MLE of $\lambda = 10.5$ indicates that the variance of the expression data must be 10.5

Question [`SCQ-17`] A cross-validation curve shows error as a fonction of regularization strength λ in the validation set, a different set of data points than the ones used to train the model. A typical cross-validation curve is U-shaped:



The left side (small λ) corresponds to:

■ Overfitting—low bias but high variance

□ Underfitting—high bias but low variance

□ The irreducible error plateau

□ The one-standard-error rule region

Catalog

Question [SCQ-24] Consider a bivariate dataset from a biological experiment where measurements of two variables X and Y appear to follow a joint normal distribution. If the mutual information $I(X;Y)$ is computed to be nearly zero, which of the following is the correct interpretation?

■ The joint distribution can be well approximated by the product of the marginals

□ The variance of variable X must be approximately equal to the variance of variable Y

□ The conditional distribution of Y given $X = x$ has exactly the same variance regardless of the value of x

□ The relationship between X and Y must be non-linear and cannot be detected with correlation analysis

Question [SCQ-25] A researcher conducting a genomics study tests 5000 genes for differential expression and uses the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR) at 0.10. After applying this correction, 800 genes are declared significant. Which of the following statements correctly interprets these results?

■ Approximately 80 of the 800 significant findings are expected to be false positives

□ 10% of all 5000 genes tested are expected to be false positives

□ The probability that any individual significant result is a false positive is exactly 0.10

□ If the same experiment were repeated, exactly 800 genes would be found significant again

Question [SCQ-26] A biologist studying the relationship between temperature (X, in °C) and enzyme reaction rate (Y, in μmol/min) obtains the following regression equation: $\hat{Y} = 2.5 + 0.8X$ with $R^2 = 0.65$. The standard error of the slope coefficient is 0.2. Which of the following statements is accurate regarding the interpretation of this model?

■ For each 1°C increase in temperature, we would expect a 0.8 μmol/min increase in the enzyme reaction rate, on average.

□ The model explains 80% of the variability in enzyme reaction rates.

□ If the temperature is 0°C, the enzyme reaction rate will be exactly 2.5 μmol/min.

□ The relationship between temperature and reaction rate is not statistically significant at $\alpha = 0.05$.

□ From the collected data, the sample mean of the enzyme reaction rate is 2.5.

Question [SCQ-28] When comparing two nested regression models using the likelihood ratio test (LRT), what is the correct interpretation if the test yields a p-value of 0.03?

■ There is sufficient evidence to reject the null hypothesis that the additional predictors do not improve model fit

□ The simpler model fits the data better than the more complex model

□ There is a 3% chance that the simpler model is correct

□ The more complex model explains 3% more variance than the simpler model

□ The simpler model is preferred because it has fewer parameters and is more parsimonious

□ There is enough evidence to reject that the most complex model performs better than the simpler one.

Question [SCQ-16] In the context of ridge regression, we note the regularization strength λ. If $\lambda \to \infty$, the fitted coefficients $\hat{\beta_{\text{ridge}}}$ approach:

■ Zero for all predictors

□ The OLS solution

□ The eigenvector associated with the largest eigenvalue

□ Infinity

Question [SCQ-29] A researcher applies logistic regression and Poisson regression to analyze different biological datasets. Which of the following statements correctly describes these models within the GLM framework?

■ Both models transform from bounded response space to unbounded linear space, but logistic regression handles binary outcomes bounded between 0 and 1, while Poisson regression handles counts bounded at 0.

□ Both logistic and Poisson regression use the log link function, but differ in their random component distributions (Bernoulli vs. Poisson).

□ The logistic regression model uses the logit link function and assumes variance equals the mean, while Poisson regression uses the log link and assumes constant variance.

□ In logistic regression, a coefficient $\beta_1 = 0.5$ means the probability of the outcome increases by 0.5 for each unit increase in X_1, while in Poisson regression, the same coefficient means the count increases by 0.5.

Question [SCQ-30] Three researchers are modeling the number of mutations per cell as a Poisson process with rate parameter λ. From their pilot data, each proposes a different fixed rate:

(a) Dr. Adams: $\lambda = 2$

(b) Dr. Baxter: $\lambda = 4$

(c) Dr. Collins: $\lambda = 6$

They then observe a new cell and count 3 mutations. Based on this new data point, whose model fits the data better?

□ Dr. Adams

■ Dr. Baxter

□ Dr. Collins

Question [SCQ-31] A bioinformatician is simultaneously estimating the ($\log_2$) mean expression levels of 5 genes. Each measurement has known variance $\sigma^2 = 1.0$. The observed $\log_2$-fold-changes (relative to control) are:

(a) Gene A: 3.10

(b) Gene B: -1.75

(c) Gene C: 2.42

(d) Gene D: 0.87

(e) Gene E: -0.39

These estimates were obtained using the standard approach of computing the sample mean. However, the researcher recalls a statistical paper from the 1960s showing that this estimation method is inadmissible when simultaneously estimating three or more means. The paper described an alternative approach guaranteed to have lower overall mean squared error (MSE), but the researcher cannot recall the exact formula. Which of the following sets of estimates would improve the MSE of the researcher's estimates?

□ $[2.72, -1.53, 2.12, 0.76, -0.34]$

□ $[2.30, -1.30, 1.80, 0.65, -0.29]$

■ $[2.62, -1.48, 2.05, 0.74, -0.33]$

□ $[2.75, -1.35, 2.18, 0.87, -0.20]$

Question [SCQ-32] Consider 4 points in a high dimensional space with pairwise distances. The distance matrix between these points is:

	a	b	c	d
a	0.0	1.0	4.0	6.1
b	1.0	0.0	3.0	5.1
c	4.0	3.0	0.0	2.2
d	6.1	5.1	2.0	0.0

Using hierarchical agglomerative clustering with complete linkage, what is the correct sequence of merges?

□ First merge (a,b), then ((a,b),c), then (((a,b),c),d)

■ First merge (a,b), then merge (c,d), then merge ((a,b),(c,d))

□ First merge (c,d), then ((c,d),b), then (((c,d),b),a)

□ First merge (c,d), then ((c,d),a), then (((c,d),a),b)