

MATH5472 Assignment 3

October 8, 2025

Problem 1

Consider the following probabilistic model: $z_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \dots, N$.

(a) Suppose we know a large proportion of μ_i is zero and only a small proportion of μ_i is non-zero. Let γ_i be the indicator, i.e., $\gamma_i = 1$ indicates μ_i is non-zero and $\gamma_i = 0$ indicates μ_i is zero. Then we assume

$$\begin{cases} \mu_i \sim \mathcal{N}(0, \sigma^2), & \text{if } \gamma_i = 1, \\ \mu_i = 0, & \text{if } \gamma_i = 0, \end{cases}$$

where $\pi = p(\gamma_i = 1)$ and $1 - \pi = p(\gamma_i = 0)$.

(a) Derive an algorithm to estimate parameters σ^2 and π . Obtain posterior mean of μ_i , denoted as $\hat{\mu}_i^S$ using the estimated parameters.

(b) Use simulation to compare $\hat{\mu}_i^S$ with James-Stein estimator and Tweedie's formula. In the simulation, you can fix $\sigma^2 = 1$, and vary π_1 to evaluate their performance. Note that sample size N may also affect your comparison results.

(c) Suppose we have observed an additional vector $\mathbf{a} = [a_1, \dots, a_N]^T$ which could be relevant with γ . A probabilistic model to describe their relationship is given as

$$\log \frac{p(\gamma_i = 1|a_i)}{p(\gamma_i = 0|a_i)} = \beta_0 + \beta a_i,$$

where β_0 and β are two scalar. Can you design a method to estimate β_0 and β ? If yes, can you examine whether $\beta = 0$ or not. Further, can you obtain $p(\gamma_i|z_i, a_i, \beta_0, \beta, \pi)$ and $p(\mu_i|z_i, a_i, \beta_0, \beta, \pi)$? Do you have a better result compared with $\hat{\mu}_i^S$ given in (b)? Justify your answer with simulation.

Problem 2

Consider a linear model,

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

where \mathbf{y} is an $n \times 1$ response vector, \mathbf{X} is an $n \times p$ design matrix, \mathbf{b} is a $p \times 1$ coefficient vector, and $\mathbf{e} = [e_1, \dots, e_n]^T$ is an $n \times 1$ vector of errors with $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.

(a) Suppose that we know there is only one non-zero coefficient in \mathbf{b} . Then we can write $\mathbf{b} = b\boldsymbol{\gamma}$, where b is a scalar and $\boldsymbol{\gamma}$ is a $p \times 1$ binary vector. Let us specify the prior as follows

$$b \sim \mathcal{N}(0, \sigma_0^2), \quad p(\gamma_i = 1) = 1/p,$$

for all i . Can you design an algorithm to estimate σ^2 and σ_0^2 ? Can you compute the posterior of $\mathbf{b} = b\boldsymbol{\gamma}$, i.e., $p(\gamma_i | \mathbf{X}, \mathbf{y}, \sigma_0^2, \sigma^2)$ and $p(b | \mathbf{X}, \mathbf{y}, \sigma_0^2, \sigma^2, \gamma_i = 1)$?

Problem 3

In the paper “Evidence Contrary to the Statistical View of Boosting” <https://www.jmlr.org/papers/volume9/mease08a/mease08a.pdf>, the author provides some empirical evidence that raises questions about the statistical perspective of boosting algorithms. Please read the paper and choose one question to provide your own comments. For example, do you agree with the author? Why? You need to present empirical evidence to support your comments.

Problem 4

Besides the gradient boosting, Bagging (bootstrap aggregating of multiple trees) and Random Forest are two alternative approaches for ensembling learning. Let p be the number of variables for classification or regression problems. The only difference between Bagging and Random Forest is the *mtry* parameter, where *mtry* = p in Bagging and *mtry* = \sqrt{p} and $p/3$ for classification and regression in Random Forest, respectively. There is a claim that the *mtry* parameter in Random Forest plays a role of inexplicit regularization. Please provide your own view with some supporting evidence. You may use “Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success” <https://jmlr.org/papers/v21/19-905.html> as a reference. The answer to this question is quite open.