Statistical Learning
Assignment 4

Exercises

Problem 1: Knockoffs

1. What are knockoffs?
2. The vector of W statistics for the knockoffs procedure is equal to:

$$W = (8, -4, -2, 2, -1.2, -0.6, 10, 12, 1, 5, 6, 7).$$

Which variables would be considered important if we use knockoffs at the false discovery rate (FDR) level $q = 0.4$?

Problem 2: PCA – Eigenvalue Decomposition and Projection

You are given a centered data matrix $X \in R^{4\times 2}$:

$$X = \begin{bmatrix} 2 & 0 \\ 0 & 2 \\ -2 & 0 \\ 0 & -2 \end{bmatrix}$$

(a) Compute the sample covariance matrix of X.

(b) Find the eigenvalues and eigenvectors of the covariance matrix.

(c) Project the data onto the first principal component.

(d) What is the variance explained by the first component?

(e) Reconstruct the original data (approximate reconstruction) using only the first principal component.

(c) Compute the reconstruction error (sum of squared Euclidean distances between original and reconstructed data points).

Problem 3: PPCA – Log-Likelihood and Parameter Estimation

Suppose a single data point $x \in R^2$ is generated by the PPCA model:

$$x = Wz + \mu + \epsilon$$

with:

$$W = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \sigma^2 = 1$$

Latent variable $z \sim \mathcal{N}(0, 1)$, and noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

(a) What is the marginal distribution of x? (mean and covariance matrix)

(b) Compute the log-likelihood of observing $x = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.

(c) Briefly explain how maximum likelihood estimation would be used to estimate W and σ^2 given a dataset.

Computer projects

Project 1: Knockoffs

Generate the design matrix $X_{500\times 450}$ such that its elements are independent and identically distributed (iid) random variables from $\mathcal{N}(0, \sigma = \sqrt{\frac{1}{n}})$. Then generate the vector of the response variable according to the model:

$$Y = X\beta + \epsilon,$$

where $\epsilon \sim 2\mathcal{N}(0, I)$, $\beta_i = 10$ for $i \in \{1, \ldots, k\}$, $\beta_i = 0$ for $i \in \{k+1, \ldots, 450\}$, and $k \in \{5, 20, 50\}$.

For 100 replications of the above experiments, estimate the regression coefficients and/or identify important variables using:

i) Least squares.

ii) Ridge regression and LASSO with the tuning parameters selected by cross-validation.

iii) Knockoffs with ridge and LASSO at the nominal false discovery rate (FDR) equal to 0.2.

Perform the following analyses:

a) Estimate the false discovery rate (FDR) and the power of the cross-validated LASSO and the knockoffs with ridge and LASSO.

b) For all three methods in i) and ii), estimate the mean square errors of the estimators of β and $\mu = X\beta$.

Project 2: Exploratory Data Analysis with PCA

Objective: Use PCA to analyze the structure and reduce dimensionality of a real-world dataset.

Suggested datasets:

- Wine quality dataset (see Kaggle)
- Iris dataset (built into most statistical libraries).
- MNIST handwritten digits (see Kaggle)

Tasks:

1. Standardize features and perform PCA.
2. Plot the explained variance ratio and cumulative variance.
3. Visualize data in 2D PCA space, color-coded by class labels (if applicable).
4. Interpret principal components by examining the top loading vectors.

Project 3: Probabilistic PCA vs Classical PCA – A Simulation Study

Objective: Compare PCA and PPCA in terms of reconstruction accuracy and robustness to noise.

Tasks: For $n = 200$, $p = 20$ and $k = 3$ generate n rows of synthetic data from the PPCA model:

$x = Wz + \mu + \epsilon, \quad z \sim \mathcal{N}(0, I_k), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_p), W$ is some matrix $p \times k$ and $\mu \in R^p$.

1. Fit both PCA and PPCA models.
2. Compare:
   (a) Reconstruction error
   (b) Estimated latent variables z
   (c) Estimated covariance matrices
3. Estimate the number of Principal Components using Minka's BIC.
4. Explore performance across varying noise levels σ^2.
5. Apply to a real dataset selected in the previous project and compare results qualitatively.