Home Assignment 5 Solutions

STAT 151A Linear Modelling: Theory and Applications

Ajay Sharma — Spring 2025

Problem 1: PCA (Scaling and Correlation Structure)
Consider a dataset with three variables, X_1, X_2, and X_3. These variables have the following characteristics based on preliminary analysis:

- X_1 and X_2 are strongly positively correlated with each other.
- X_2 has a measurement scale orders of magnitude larger than X_3.
- X_3 is essentially uncorrelated with both X_1 and X_2.

Suppose the sample covariance matrix calculated from the raw, unstandardized data X

$$\Sigma_{\text{raw}} = \begin{pmatrix} 1 & 90 & 0 \\ 90 & 10000 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

Further suppose that after standardizing the data (subtracting the mean and dividing by the standard deviation for each variable) to get Z, the resulting covariance matrix (which is also the correlation matrix of the original data) is

$$\Sigma_{\text{std}} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

(a) *PCA Intuition & Variance*
Stated simply, the objective of the k^{th} principal component is the direction ϕ_k^* that maximizes the variance of the projected data $\langle X, \phi_k \rangle$ subject to orthonormality constraints. In particular, we are interested in the direction where the data varies the most, based on the covariance matrix Σ_{raw}. Notice $\text{Var}(X_1) = 1, \text{Var}(X_2) = 10^4$. Therefore, the first principal component ϕ_1 will point in the direction of X_2 since the variance is much larger.

We should note that PC1 would primarily represent X_2 since the scaling (variance) is simply large, which overplays the role of the correlation between the two variables X_1 and $X_2 : \text{Cov}(X_1, X_2) = 90$ (which is strong). This aspect is misleading, since the scaling of particular variables is emphasized more so than the correlation.

(b) *Computing PCA with Standardized Data*
We would like to perform PCA by computing the eigenvalues ($\lambda_1 \geq \lambda_2 \geq \lambda_3$) and the corresponding normalized eigenvectors ($\phi_1^*, \phi_2^*, \phi_3^*$). To begin with, define the characteristic polynomial $p(\lambda) = \det(\Sigma_{\text{std}} - \lambda I_3)$ and solve $p(\lambda) = 0$.

Then we can write

$$p(\lambda) = (1-\lambda)(1-\lambda^2) - 0.9^2(1-\lambda) = (1-\lambda)[(1-\lambda)^2 - 0.9^2] = 0.$$

This immediately gives $(\lambda_1, \lambda_2, \lambda_3) = (1.9, 1, 0.1)$. The eigenvectors satisfy the property $\phi_k^* = \text{null}(\Sigma_{\text{std}} - \lambda_k I_3), k \in [3]$. Solving for each λ_k gives:

- $\lambda_1 = 1.9 : (\Sigma_{\text{std}} - \lambda_1 I_3)\phi_1 = 0 \Rightarrow \phi_1 = (1,1,0)^T$. Properly normalizing this, we obtain $\phi_1^* = \frac{1}{\sqrt{2}}(1,1,0)^T$.
- $\lambda_2 = 1 : (\Sigma_{\text{std}} - \lambda_2 I_3)\phi_2 = 0 \Rightarrow \phi_2^* = (0,0,1)^T$, which is already normalized.
- $\lambda_3 = 0.1 : (\Sigma_{\text{std}} - \lambda_3 I_3)\phi_3 = 0 \Rightarrow \phi_3 = (1,-1,0)^T$. Properly normalizing this, we obtain $\phi_3^* = \frac{1}{\sqrt{2}}(1,-1,0)^T$.

(c) *Interpretation of PCA*

i. Each eigenvalue λ_k represents the variance of the the k^{th} principal component ϕ_k^* (i.e., the direction along the normalized eigenvector). In particular, we know that the total variance is given by $\sum_{k\in[3]} \text{Var}(X_k) = 3$. So expressing each PC as a percentage of the total variance, we have obtain the quantities $\text{PC1} = 1.9/3 \approx 0.63, \text{PC2} = 1/3 = 0.33, \text{PC3} = 0.1/3 = 0.033$.

ii. Let $x_j = (x_{j1}, x_{j2}, x_{j3})^T \in \mathbb{R}^3$ be the standardized observation $j \in [3]$ and so we can define the PC score $z_{jk} = \langle x_j, \phi_k^* \rangle = (\phi_k^*)^T x_j$. Then using the above formulation, we have $z_{j1} = \langle x_1, \phi_1^* \rangle = \frac{1}{\sqrt{2}}(x_{j1} + x_{j2})$. Thus we can write $\text{PC1} = \frac{1}{\sqrt{2}}(X_1 + X_2)$.

iii. Similarly, we have $z_{j2} = \langle x_j, \phi_2^* \rangle = x_{j3}$. So we can write $\text{PC2} = X_3$. Since the observations $X_{j\in[3]}$ are standardized, we have $\text{Var}(X_3) = 1$, so $\lambda_2 = 1$. In other words, $\text{Cov}(X_3, X_{i\in[2]}) = 0$, so X_3 is an orthogonal component.

iv. Finally, can write $z_{j3} = \langle x_j, \phi_3^* \rangle = \frac{1}{\sqrt{2}}(x_{j1} - x_{j2})$. In other words, we have $\text{PC3} = \frac{1}{\sqrt{2}}(X_1 - X_2)$. Intuitively, we know that X_1 and X_2 are strong, positive correlated. In particular $\text{Cov}(X_1, X_2) = 0.9$. Therefore in the direction of $X_1 - X_2$ we have minimal variance. More formally, we can write $\lambda_3 = \text{Var}(\langle x_j, \phi_3^* \rangle) = \text{Var}(\frac{1}{\sqrt{2}}(X_1 - X_2)) = \frac{1}{2}\text{Var}(X_1 - X_2) = 0.1$.

v. We should choose the two best directions obtained from PCA, we should choose the ones such that that variance along the directions is maximal. In other words, we should choose PC1 and PC2, which capture $(1.9+1)/3 = 0.96$ of the total variance.

(d) *Comparing Outcomes of PCA*
As observed in part (a), without properly scaling the covariance matrix Σ_{raw}, we found that PC1 is largely dominated by X_2 due to scaling (despite them being strongly correlated). When we properly scale, we are able to properly capture the structure between the variables $X_{j\in[3]}$ based on variances and not scaling.

In fact for PC1, we showed that X_1 and X_2 have a joint structure (up to scaling by $1/\sqrt{2}$). Also, as we may expect PC2 is captured by the variation in X_3 and PC3 is orthogonal (in direction) to PC1.

Problem 2: Logistic Regression

Consider the usual regression data with binary (0 or 1) response values $y_1, \dots, y_n$, and explanatory variable values x_{ij}, for $i \in [n]$ and $j \in [p]$. We wish to fit the logistic regression model to the data:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \langle x_i, \beta \rangle \quad \text{for } i \in [n],$$

where $y_1, \dots, y_n$ are independent random variables having the Bernoulli distribution with means $p_1, \dots, p_n$. For notational purpose, let $x_i = (1, x_{i1}, \dots, x_{ip})^\top \in \mathbb{R}^{p+1}$, and thus $X \in \mathbb{R}^{n \times (p+1)}$. Also $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$.

(a) *Expression for Log-likelihood Function $\ell(\beta)$*

We have that each $y_i \overset{\text{iid}}{\sim} \text{Ber}(p_i)$. Therefore we can write

$$\mathcal{L}(\beta) = f(y_1, ..., y_n|\beta) = \prod_{i=1}^{n} f(y_i|\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}.$$

Taking the log, gives the log-likelihood function

$$\ell(\beta) = \log(\mathcal{L}(\beta)) = \sum_{i=1}^{n}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] = \langle y, \log(p)\rangle + \langle 1-y, \log(1-p)\rangle,$$

where $y, p \in \mathbb{R}^n$.

(b) *Fitted Values $\hat{\beta}_{MLE}$*

We already know that for logistic regression

$$\log\left(\frac{p_i}{1-p_i}\right) = \langle x_i, \beta \rangle.$$

Then to find the fitted values $\hat{p}_{i\in[n]}$, we can substitute the estimator $\hat{\beta} := \hat{\beta}_{\text{MLE}}$ i.e.,

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \langle x_i, \hat{\beta} \rangle \Rightarrow \frac{\hat{p}_i}{1-\hat{p}_i} = \exp(\langle x_i, \hat{\beta}\rangle) \Rightarrow \hat{p}_i = \frac{\exp(\langle x_i, \hat{\beta}\rangle)}{1+\exp(\langle x_i, \hat{\beta}\rangle)}.$$

After rearranging, we obtain

$$\hat{p}_i = \frac{1}{1+\exp(-\langle x_i, \hat{\beta}\rangle)} = \sigma(\langle x_i, \hat{\beta}\rangle).$$

(c) *Computing the MLE*

We wish to find the MLE based on the expression derived in part (a). We'll focus on a particular observation and compute the derivative with respect to β_j. Define

- $y_i = \{0, 1\}$
- $z_i = \langle x_i, \beta \rangle = x_i^T \beta$
- $\hat{p}_i = \sigma(z_i)$

Then the log-likelihood for a single observation becomes

$$\ell_i(\beta) = y_i \log(p_i) + (1-y_i)\log(1-p_i).$$

We can write $\frac{\partial \ell_i(\beta)}{\partial \beta_j} = \frac{\partial \ell_i(\beta)}{\partial p_i} \cdot \frac{\partial p_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial \beta_j}$ and obtain the gradient by summing $i \in [n]$.

From the chain rule above, make the following observations:

- Using the expression from in part (a), we have $\frac{\partial \ell_i}{p_i} = \frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}$.
- Since $p_i = \sigma(z_i)$ and $\sigma'(z) = \sigma(z)(1-\sigma(z)) = p_i(1-p_i)$, we have $\frac{\partial p_i}{\partial z_i} = p_i(1-p_i)$.
- Note that $z_i = x_i^T \beta = \sum_{k \in [p]} x_{ik}\beta_k \Rightarrow \frac{\partial z_i}{\partial \beta_j} = x_{ij}$.

Putting this together, we have

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right) \cdot p_i(1-p_i) \cdot x_{ij} = [y_i(1-p_i) - (1-y_i)p_i] \cdot x_{ij} = (y_i - p_i) \cdot x_{ij}.$$

Summing over all $i \in [n]$ gives

$$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} (y_i - p_i) x_{ij} = X^T(y - p).$$

Therefore the MLE is given by solving $\nabla_\beta \ell(\beta) = 0 \Rightarrow X^T y = X^T \hat{p}$. In other words, we have shown that $\hat{\beta}_{\text{MLE}}$ depends only on y through $X^T y$.

(d) *An Additional Calculation*
Using our result from part (c), we know that at $\hat{\beta}_{\text{MLE}}$ the gradient is 0. In particular,

$$\sum_{i=1}^{n} (y_i - \hat{p}_i) x_{ij} = 0 \in \mathbb{R}^{p+1}.$$

Furthermore, we know that the first component of the design matrix column we have each $x_{i0} = 1$. Therefore, our equation above simplifies to

$$\sum_{i=1}^{n} (y_i - \hat{p}_i) \cdot 1 = 0 \Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{p}_i,$$

as claimed.

Problem 3: Additional Property of Logistic Regression
In the logistic regression model, let $\hat{p}$ denote the vector of fitted probabilities. We want to show that $Y - \hat{p}$ is orthogonal to any column of the matrix X.

Note that $X = (x^{(1)}, ..., x^{(p+1)}) \in \mathbb{R}^{n \times (p+1)}$, where each $x^{(j)} \in \mathbb{R}^n$. Additionally, from part (c) we established that at the MLE, the gradient vanishes. In other words

$$X^T(Y - \hat{p}) = 0 \Leftrightarrow \langle X, Y - \hat{p} \rangle = 0.$$

This means each $\langle x^{(j)}, Y - \hat{p} \rangle = 0, j \in [p+1]$. Therefore the residual $Y - \hat{p}$ is orthogonal to the columns of X. In other words, $Y - \hat{p} \perp \text{Col}(X_j), j \in [p+1]$ as a result of the conditions established in parts (c) and (d) in exercise 2, as claimed.