Chapter 4

Linear Methods for Classification

Overview

Suppose there are K classes, and the method for classification models the posterior probability $\mathbb{P}(G = k|X = x)$ for any input x and class $k \in \{1, 2, \ldots, K\}$. The classification boundary in the input space is linear as long as some monotone transformation of the modeled posterior distribution $\mathbb{P}(G = k|X = x)$ is linear in x, with the log-odds function being a common transformation:

$$\log \frac{\mathbb{P}(G = k|x)}{\mathbb{P}(G = K|x)} = \beta_k^T x.$$

Two different methods results in linear log-odds: logistic regression and linear discriminant analysis. Logistic regression directly model $\mathbb{P}(G = k|x)$ as

$$\begin{aligned}
\mathbb{P}(G = k|x;\theta) =& \frac{e^{\beta_k^T x}}{1 + \sum_{l=1}^{K-1} e^{\beta_l^T x}} \text{ for } k = 1, 2, \ldots, K-1 \text{ and} \\
\mathbb{P}(G = K|x;\theta) =& \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_l^T x}} \\
\theta =& \{\beta_1, \beta_2, \ldots, \beta_{K-1}\}.
\end{aligned}$$

and fits the posterior probability to the testing data with cross entropy (or equivalently, negative likelihood) as the loss function. There are in total $(p+1) \times (K-1)$ degree-of-freedom in the model. On the other hands, linear discriminant analysis models the distribution of input for any given class as Gaussian:

$$\begin{aligned}
\log \mathbb{P}(X = x|G = k) =& -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma_k) - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k), \\
\log \mathbb{P}(G = k) =& \log \pi_k.
\end{aligned}$$

which leads to linear log-odds for posterior distribution with uniform Σ_k (aka LDA) and quadratic log-odds for non-uniform Σ_k (aka QDA) across different classes. The parameter set $\theta = \{\Sigma_k, \mu_k, \pi_k, k = 1, 2, \ldots, K\}$ are estimated with an aim to maximize the joint likelihood:

$$\sum_{i=1}^{N} \log \mathbb{P}(X = x_i, G = g_i; \theta), \text{ where } g_i \text{ is the class label of input } x_i.$$

Given the additional model assumption, LDA yields more stable (less variance, likely more bias) estimation compared with logistic regression: the additional model assumption can be viewed as regularization. On the other hands, logistic regression relies on fewer assumption about the data and is generally considered as a safer choice.

Linear/Quadratic Discriminant Analysis

The parameter estimation for $\theta = \{\Sigma_k, \mu_k, \pi_k, k = 1, 2, \dots, K\}$ amounts to classic MLE problem for multi-variant normal distribution with the following solution:

$$\begin{aligned}
\hat{\pi}_k &= N_k/N \\
\hat{\mu}_k &= \sum_{i=1}^{N} \mathbf{1}(g_i = k) x_i / N \\
\text{LDA: } \hat{\Sigma} &= \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{1}(g_i = k)(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K) && \text{(trace trick)} \\
\text{QDA: } \hat{\Sigma}_k &= \sum_{i=1}^{N} \mathbf{1}(g_i = k)(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N_k - 1).
\end{aligned}$$

LDA results in the following discriminant function

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

by noting that

$$\begin{aligned}
\log \frac{\mathbb{P}(G = k, X = x)}{\mathbb{P}(G = k, X = l)} &= \log \frac{\pi_k}{\pi_l} + \frac{1}{2}(x - \mu_l)^T \Sigma^{-1} (x - \mu_l) - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \\
&= \log \frac{\pi_k}{\pi_l} - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k + \frac{1}{2} \mu_k \Sigma^{-1} \mu_k + \mu_k^T \Sigma^{-1} x - \mu_l^T \Sigma^{-1} x.
\end{aligned}$$

Similarly, the discriminant function for QDA is

$$\delta_k(x) = -\frac{1}{2} \log \det \Sigma_k - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k.$$

Quadratic classification boundary

To achieve quadratic classification boundary, we can either use LDA with quadratic-extended input space, or use QDA directly. For the LDA plus extension approach, it is easy to see that there are $(K-1) \times (\binom{p+1}{2} + p + 1) = (K-1) \times (p(p+3)/2 + 1)$ parameters. For QDA approach, each mean has p parameters and each covariance matrix has $p(p+1)/2$ parameters, so in total there are $p(p+3)/2$ parameters per class.

Computation of LDA

LDA classification can be implemented by first sphering/whitenning the data according to the uniform covariance matrix, and then search for the class with the closest centroid:

$$\begin{aligned}(x-\mu_k)^T\Sigma^{-1}(x-\mu-k) =&(x-\mu_k)^T\Sigma^{-1/2T}\Sigma^{-1/2}(x-\mu_k)\\ =&(\Sigma^{-1/2}x-\Sigma^{-1/2}\mu_k)^T(\Sigma^{-1/2}x-\Sigma^{-1/2}\mu_k),\end{aligned}$$

where $\Sigma^{-1/2}$ can either be obtained using Cholesky decomposition $\Sigma = LL^T$ with $\Sigma^{-1/2} = L^{-1}$ or using Eigen-component decomposition $\Sigma = U\Lambda U^T$ with $\Sigma^{-1/2} = \Lambda^{-1/2}U^T$.

The above results suggest a natural dimension reduction: after sphering the data, we can focus on the subspace spanned by $\{\Sigma^{-1/2}\mu_k, k = 1, 2, \ldots, K\}$, since only the distance to the centroid matters and we can ignore any directions that are orthogonal to the centroid-spanned subspace. This reduces the dimension from p to $K-1$.

Reduced rank LDA

The dimension of the data can be future compressed by applying the principal component analysis on the sphered centroid. Let us assume the following data model

$$x_p = c_p + n_p \tag{4.1}$$

where c_p represents the class centroid, which can take any one of $\{\mu_k, k = 1, 2, \ldots, K\}$ depending on which class the data belongs, n_p represents the noise on top of the centroid, and x_p represents the observed data. The subscript p here emphasizes that the dimension is p. We denote

$$\mathbb{E}\left[n_p n_p^T\right] = W, \mathbb{E}\left[(c_p - \mathbb{E}\left[c_p\right])(c_p - \mathbb{E}\left[c_p\right])^T\right] = B.$$

The principal component analysis can of viewed as an effort in trying to find a p-dimension vector a_p such that $a_p^T x_p$ has the maximum SNR, by viewing the centroid as the embedded signal in the data and the dispersion from the class centroid as noise.

$$\text{SNR} = \frac{\mathbb{E}\left[a_p^T c_p c_p^T a_p\right]}{\mathbb{E}\left[a_p^T n_p n_p^T a_p\right]} = \frac{a_p^T W a_p}{a_p^T B a_p} = \text{Rayleigh-Quotient}.$$

From the solution to Problem 4.1, we know that the solution to the above problem is the strongest eigen component of $W^{-1}B$. To find the $m-$dimensional principal component, we can simply take the strongest n eigen-vectors of $W^{-1}B$, denoted as $A_{p\times m}$, and then transform the original input data space X to $XA_{p\times m}$. Equivalently, we would sphere the data first by transforming X to $XW^{-1/2}$, and then project $XW^{-1/2}$ onto the m-dimensional principal subspace of the space spanned by the sphered centroid and get $XW^{-1/2}A'_{p\times m}$, where $A'_{p\times n}$ captures the strongest m eigenvectors of $W^{-1/2T}BW^{-1/2}$.

Logistic Regression

The negative loss function for logistic regression is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} \log \mathbb{P}\left(g_i|x_i;\theta\right) && \left(\begin{array}{c}\text{log-likelihood}\\ \text{function}\end{array}\right) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \underbrace{\mathbf{1}\left(g_i = k\right)}_{\text{true distribution}} \log \underbrace{\mathbb{P}\left(k|x_i;\theta\right)}_{\substack{\text{estimated}\\ \text{distribution}}} && \left(\text{negative cross-entropy}\right) \end{aligned}$$

Two-class logistic regression

In the two-class classification case with $y_i = (g_i = 1)$ and $\theta = \{\beta\}$, we have

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} y_i \log \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} + (1-y_i)\log\frac{1}{1+e^{\beta^T x_i}} \\ &= \sum_{i=1}^{N} y_i \log e^{\beta^T x_i} + \log\frac{1}{1+e^{\beta^T x_i}} = \sum_{i=1}^{N} y_i \beta^T x_i - \log\left(1+e^{\beta^T x_i}\right) \\ \frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^{N} y_i x_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} x_i = x_i \left(y_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) = X^T(y-p) \\ \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} &= -\sum_{i=1}^{N} x_i \frac{e^{\beta^T x_i} x_i^T \left(1+e^{\beta^T x_i}\right) - e^{\beta^T x_i} x_i^T e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2} \\ &= -\sum_{i=1}^{N} x_i x_i^T \frac{1}{1+e^{\beta^T x_i}} \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} = -X^T W X, \end{aligned}$$

where $X = [x_1, x_2, \ldots, x_N]^T$, $y = [y_1, y_2, \ldots, y_N]^T$, $p = \left[\mathbb{P}\left(1|x_1;\beta\right), \mathbb{P}\left(1|x_2;\beta\right), \ldots, \mathbb{P}\left(1|x_N;\beta\right)\right]^T$, and W is a $N \times N$ diagonal matrix with $W_{ii} = \mathbb{P}\left(1|x_i;\beta\right)\mathbb{P}\left(2|x_i;\beta\right)$.

$$\begin{aligned} l(\beta) \approx& \beta^{\text{old}} + (\beta-\beta^{\text{old}})^T \frac{\partial l(\beta)}{\partial\beta}|_{\beta=\beta^{\text{old}}} + (\beta-\beta^{\text{old}})^T \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}|_{\beta=\beta^{\text{old}}}(\beta-\beta^{\text{old}}) \\ \Rightarrow \beta^{\text{new}} =& \beta^{\text{old}} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial l(\beta)}{\partial\beta} \\ =& \beta^{\text{old}} + (X^T W X)^{-1} X^T (y-p) \\ =& (X^T W X)^{-1} X^T W X \beta^{\text{old}} + (X^T W X)^{-1} X^T (y-p) \\ =& (X^T W X)^{-1} X^T W \left[X\beta^{\text{old}} + W^{-1} X^T (y-p)\right] \\ =& \arg\min_{\beta} (z - X\beta)^T W (z - X\beta) \text{ with } z = X\beta^{\text{old}} + W^{-1} X^T (y-p). \end{aligned}$$

Multi-class logistic regression

For the general multi-class case, we have

$$\begin{aligned}
\log \mathbb{P}\left(k|x_i;\theta\right) =& \beta_k^T x_i - \log\left(1+\sum_{q=1}^{K-1} e^{\beta_q^T x_i}\right)\\
\text{for } j\neq k,\ \frac{\partial \log \mathbb{P}\left(k|x_i;\theta\right)}{\partial \beta_j} =& -\mathbb{P}\left(j|x_i;\theta\right)x_i\\
\frac{\partial \log \mathbb{P}\left(k|x_i;\theta\right)}{\partial \beta_k} =& \left(1-\mathbb{P}\left(k|x_i;\theta\right)\right)x_i\\
\text{for } j\neq k,\ \frac{\partial \mathbb{P}\left(k|x_i;\theta\right)}{\partial \beta_j^T} =& \mathbb{P}\left(k|x_i;\theta\right)\mathbb{P}\left(j|x_i;\theta\right)x_i^T\\
\frac{\partial \mathbb{P}\left(k|x_i;\theta\right)}{\partial \beta_k^T} =& \left(\mathbb{P}\left(k|x_i;\theta\right)^2-\mathbb{P}\left(k|x_i;\theta\right)\right)x_i^T,
\end{aligned}$$

and thus

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \beta_j} =& \sum_{i=1}^{N}\left[\mathbf{1}(g_i=j)x_i-\mathbb{P}\left(j|x_i;\theta\right)x_i\right]\\
\frac{\partial^2 l(\theta)}{\partial \beta_j \partial \beta_s^T} =& -\sum_{i=1}^{N}\left[\mathbb{P}\left(s|x_i;\theta\right)\mathbb{P}\left(j|x_i;\theta\right)-\mathbf{1}(s\neq j)\mathbb{P}\left(j|x_i;\theta\right)\right]x_i x_i^T = -X^T W^{(s,j)}X,
\end{aligned}$$

where $W^{(s,j)}$ is a $N\times N$ diagonal matrix with $W_{ii}^{(s,j)} = \mathbb{P}\left(s|x_i;\theta\right)\mathbb{P}\left(j|x_i;\theta\right)-\mathbf{1}(s\neq j)\mathbb{P}\left(j|x_i;\theta\right)$.

Similarly as the two-class case, we can apply Newton-Raphson algorithm to reduce the multi-class logistic regression problem to a weighted least square problem with an expanded version of X, as is evident from the following equation:

$$\begin{aligned}
&\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T} = X^{\exp T}W^{\exp}X^{\exp}\\
&=\begin{bmatrix} X^T & 0 & \dots & 0\\ 0 & X^T & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & X^T\end{bmatrix}
\begin{bmatrix} W^{(1,1)} & W^{(1,2)} & \dots & W^{(1,K-1)}\\ W^{(2,1)} & W^{(2,2)} & \dots & W^{(2,K-1)}\\ \vdots & \vdots & \ddots & \vdots\\ W^{(K-1,1)} & W^{(K-1,2)} & \dots & W^{(K-1,K-1)}\end{bmatrix}
\begin{bmatrix} X & 0 & \dots & 0\\ 0 & X & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & X\end{bmatrix}.
\end{aligned}$$

Problem 4.1

$$\begin{aligned}
&\max_{a^TWa=1}\ a^TBa\\
\left(\tilde{a}=W^{1/2}a\right) &= \max_{||\tilde{a}||=1} \tilde{a}^T {W^{-1/2}}^T B W^{-1/2}\tilde{a}\\
&= \max_{||\tilde{a}||=1} \tilde{a}^T B^* \tilde{a}\\
\left(\substack{{W^{-1/2}}^T B W^{-1/2} \text{ is symmetric and thus}\\ \text{a real eigen-decomposition exists}}\right) &= \max_{||\tilde{a}||=1} \tilde{a}^T V^* D_B {V^*}^T\tilde{a}.
\end{aligned}$$

The solution to the above problem is $\tilde{a}=v_1^* \Rightarrow a = W^{-1/2}v_1^*$.

Next, let us show that $W^{-1}B$ and $W^{-1/2T}BW^{-1/2}$ have the same eigenvalue, and that for any eigenvector v^* of $W^{-1/2T}BW^{-1/2}$, $W^{-1/2}v^*$ is the eigenvector of $W^{-1}B$ with the same eigenvalue, which, by combining with the previous equation, finishes the proof:

$$\begin{aligned}
&W^{-1/2T}BW^{-1/2}v^* = \lambda v^* \\
\Rightarrow &W^{-1/2}W^{-1/2T}BW^{-1/2}v^* = \lambda W^{-1/2}v^* \\
\Rightarrow &W^{-1}B\left(W^{-1/2}v^*\right) = \lambda\left(W^{-1/2}v^*\right).
\end{aligned}$$

Problem 4.2

(a) Linear discriminant function

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k).$$

x is classified as class 2 if

$$\begin{aligned}
&\delta_2(x) - \delta_1(x) = x^T\Sigma^{-1}(\mu_2 - \mu_1) - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log(\pi_2/\pi_1) \geq 0 \\
\Rightarrow &x^T\Sigma^{-1}(\mu_2 - \mu_1) \geq \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \log((N_2/N)/(N_1/N)) \\
\Rightarrow &x^T\Sigma^{-1}(\mu_2 - \mu_1) \geq \frac{1}{2}(\mu_2 + \mu_1)^T\Sigma^{-1}(\mu_2 - \mu_1) - \log(N_2/N_1). \qquad (4.2)
\end{aligned}$$

(b) With linear regression, we have

$$(X^TX)\beta = Xy$$

$$\begin{bmatrix} \mathbf{1}^\mathbf{T} & \mathbf{1}^\mathbf{T} \\ X^{(2)T} & X^{(1)T} \end{bmatrix} \begin{bmatrix} \mathbf{1} & X^{(2)} \\ \mathbf{1} & X^{(1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{1}^\mathbf{T} & \mathbf{1}^\mathbf{T} \\ X^{(2)T} & X^{(1)T} \end{bmatrix} \begin{bmatrix} N/N_2\mathbf{1} \\ -N/N_1\mathbf{1} \end{bmatrix}$$

$$\begin{bmatrix} N & N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T \\ N_1\hat{\mu}_1 + N_2\hat{\mu}_2 & X^{(2)T}X^{(2)} + X^{(1)T}X^{(1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ N(\hat{\mu}_2 - \hat{\mu}_1) \end{bmatrix}.$$

$$\begin{aligned}
X^{(2)T}X^{(2)} + X^{(1)T}X^{(1)} =&\bar{X}^{(2)T}\bar{X}^{(2)} + \bar{X}^{(1)T}\bar{X}^{(1)} + N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T \\
=&(N-2)\hat{\Sigma} + N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T.
\end{aligned}$$

Combining the above two equations, we have

$$(N-2)\hat{\Sigma}\beta + \frac{N_1N_2}{N}\left(\hat{\mu}_2 - \hat{\mu}_1\right)\left(\hat{\mu}_2 - \hat{\mu}_1\right)^T\beta = N\left(\hat{\mu}_2 - \hat{\mu}_1\right)$$

(c) Given that $(\hat{\mu}_2 - \hat{\mu}_1)^T\beta$ is a scalar, according to the result in (b), we have $\hat{\Sigma}\beta \propto (\hat{\mu}_2 - \hat{\mu}_1)$, and thus $\hat{\beta} \propto \hat{\Sigma}^{-1}\left(\hat{\mu}_2 - \hat{\mu}_1\right)$.

(d) If the two classes are coded the same, then $\beta_0 = 1$ and $\beta = 0$. Otherwise, $\hat{\beta} \propto \hat{\Sigma}^{-1}\left(\hat{\mu}_2 - \hat{\mu}_1\right)$ still holds.

(c) Based on the previous results, there exists a scalar a with

$$\begin{aligned}
\hat{\beta} =&a\hat{\Sigma}^{-1}\left(\hat{\mu}_2 - \hat{\mu}_1\right) \\
\hat{\beta}_0 =&-\frac{a}{N}\left(N_2\hat{\mu}_2 + N_1\hat{\mu}_1\right)^T\hat{\Sigma}^{-1}\left(\hat{\mu}_2 - \hat{\mu}_1\right).
\end{aligned}$$

It is easy to see that $\hat{\beta}_0 + x^T\hat{\beta}_1 \geq 0$ is different from the LDA classification rule in Equation (4.2) as long as $N_1 \neq N_2$.

Problem 4.4

See Multi-class logistic regression

Problem 4.5

Consider a K-class classification problem with $x \in \mathbb{R}^p$. Suppose the samples x_i for the K classes are linearly separable in the sense that there exist linear discriminant functions $f_k(x) = \beta_k^T x + \beta_{k0}$ such that for any sample x_i, $f_k(x_i) \geq \max\{f_1(x_i), \dots, f_{k-1}(x_i), f_{k+1}(x_i), \dots, f_K(x_i)\}$ if and only if $g_i = k$.

Let us model the posterior probability for a sample x belonging to class k as

$$\mathbb{P}(k|x; a) = \frac{e^{af_k(x)}}{\sum_{l=1}^K e^{af_l(x)}},$$

for which the log-likelihood function becomes

$$l(a) = \sum_{i=1}^N \sum_{k=1}^K \mathbf{1}(g_i = k)\left(af_k(x_i) - \log\sum_{l=1}^K e^{af_l(x_i)}\right).$$

It is easy to see that $l'(a)$ is a non-negative increasing function, which indicates that $l(a) \to \infty$ as $a \to \infty$.

Problem 4.6

(a) It is easy to see that separability implies that $y_i\beta^T x_i \geq \epsilon$ for some ϵ with any i, which implies that $y_i\beta_{\text{sep}}^T z_i \geq 1$ for any i.

(b)

$$\begin{aligned}||\beta_{\text{new}} - \beta_{\text{sep}}||^2 =& ||\beta_{\text{old}} - \beta_{\text{sep}}||^2 + 2(\beta_{\text{old}} - \beta_{\text{sep}})^T y_i z_i + ||y_i z_i||^2 \\ =& ||\beta_{\text{old}} - \beta_{\text{sep}}||^2 + 2\beta_{\text{old}}^T z_i y_i - 2\beta_{\text{sep}}^T y_i z_i + 1 \\ \leq& ||\beta_{\text{old}} - \beta_{\text{sep}}||^2 - 2\beta_{\text{sep}}^T y_i z_i + 1 \\ \leq& ||\beta_{\text{old}} - \beta_{\text{sep}}||^2 - 1.\end{aligned}$$

Problem 4.7

The criterion tries to find a classification hyperplane where the sum distance of all the correctly classified points to the hyperplane minus the sum distance of all the wrongly classified points is maximized. It does not solve the optimal separating hyperplane problem.

Simple example: $\{(x_1 = (0, 10), y_1 = 1), (x_2 = (1, 0), y_2 = 1), (x_3 = (0, -10), y_4 = -1), (x_4 = (-1, 0), y_4 = -11)\}$. Optimal β is $(0, 1)$, which is not a separating hyperplane.

Problem 4.8

$$\mathbb{P}\left(G=k, X=x\right) = \mathbb{P}\left(X=x|G=k\right)\mathbb{P}\left(G=k\right) = \frac{\pi_k}{\sqrt{2\pi \det K}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}.$$

log-likelihood function

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}(g_i=k)\log\mathbb{P}\left(G=k, X=x_i\right)\\
=&\text{constant} + \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}(g_i=k)\left[-\frac{1}{2}\log\det\Sigma - \frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k) + \log(\pi_k)\right].
\end{aligned}$$

The constraint MLE problem can be expressed as

$$\begin{aligned}
\text{max.}\ &\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}(g_i=k)\left[\log\det\Sigma + (x_i-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right]\\
\text{s.t. }&\text{rank}\{\mu_k\}_{k=1}^{K} = L < \max\{K-1, p\}.
\end{aligned}$$

unfinished

Unfinished Problems

Problem 4.3

- Problem 4.8
- Problem 4.9