

A Partial Solution Manual for: *The Elements of  
Statistical Learning* by Jerome Friedman, Trevor  
Hastie, and Robert Tibshirani

Wenhao Wu

wnhwu@ucdavis.edu

Dept. ECE, UC Davis

June 16, 2016

# Contents

Preface	2
Acknowledgment	3
2 Overview of Supervised Learning	4
3 Linear Methods for Regression	5
4 Linear Methods for Classification	6
5 Basis Expansions and Regularization	7
6 Kernel Smoothing Methods	8
7 Model Assessment and Selection	9
8 Model Inference and Averaging	13
9 Additive Models, Trees, and Related Methods	14
10 Boosting and Additive Trees	15
11 Neural Networks	16
12 Support Vector Machines and Flexible Discriminants	18
13 Prototype Methods and Nearest-Neighbors	26
14 Unsupervised Learning	30
15 Random Forests	40
16 Ensemble Learning	44
17 Undirected Graphical Models	46
18 High-Dimensional Problems	51
References	60

# Preface

This work is expected to be used as a supplementary material for Weatherwax and Epstein's solution manual [1], which I found to be very helpful when self-studying this popular textbook. The numbering of chapters and problems are based on the 2nd edition (10th printing with corrections, Jan 2013) available online [2].

The author was not able to solve all the exercises. Even for the solutions included we expect many mistakes and shortcomings. It would be of great help if people could suggest possible solutions or help us find and correct the errors so this solution manual can be continuously improved to benefit more interested readers. We are also open to all comments and criticisms. Our contact information can be found at the website holding this draft [3].

# Acknowledgment

## Chapter 2

### Overview of Supervised Learning

## Chapter 3

### Linear Methods for Regression

## Chapter 4

### Linear Methods for Classification

## Chapter 5

### Basis Expansions and Regularization



## Chapter 6

### Kernel Smoothing Methods

## Chapter 7

### Model Assessment and Selection

**Ex. 7.1**

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}} [\text{Err}_{\text{in}}] &= \frac{1}{N} \mathbb{E}_{\mathbf{y}, \mathbf{y}^0} [\|\mathbf{y}^0 - \hat{f}(\mathbf{x})\|^2 | \mathcal{T}] \\
&= \frac{1}{N} \mathbb{E}_{\mathbf{y}, \mathbf{y}^0} [\|\mathbf{y}^0 - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\|^2 | \mathcal{T}] \\
&= \frac{1}{N} \mathbb{E}_{\mathbf{y}, \mathbf{y}^0} [\|(\mathbf{y}^0 - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}^0) + \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y}^0 - \mathbf{y})\|^2 | \mathcal{T}] \\
&= \frac{1}{N} \mathbb{E}_{\mathbf{y}^0} [\|(\mathbf{y}^0 - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}^0)\|^2] + \frac{1}{N} \mathbb{E}_{\epsilon, \epsilon^0} [\|\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\epsilon^0 - \epsilon)\|^2] \\
&= \mathbb{E}_{\mathbf{y}} (\overline{\text{err}}) + \frac{2\sigma_{\epsilon}^2}{N} \text{tr} (\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \\
&= \mathbb{E}_{\mathbf{y}} (\overline{\text{err}}) + \frac{2d}{N} \sigma_{\epsilon}^2
\end{aligned} \tag{7.1}$$

**Ex. 7.2**

$$\begin{aligned}
\text{Err}(x_0) &= \Pr(Y \neq \hat{G}(x_0) | X = x_0) \\
&= \Pr(Y \neq G(x_0) | X = x_0) \left[ \Pr(G(x_0) = \hat{G}(x_0) | X = x_0) + \Pr(G(x_0) \neq \hat{G}(x_0) | X = x_0) \right] \\
&\quad + [\Pr(Y = G(x_0) | X = x_0) - \Pr(Y \neq G(x_0) | X = x_0)] \Pr(G(x_0) \neq \hat{G}(x_0) | X = x_0) \\
&= \text{Err}_{\text{B}}(x_0) + [\Pr(Y = G(x_0) | X = x_0) - \Pr(Y \neq G(x_0) | X = x_0)] \Pr(G(x_0) \neq \hat{G}(x_0) | X = x_0) \\
&= \text{Err}_{\text{B}}(x_0) + |1 - 2f(x_0)| \Pr(G(x_0) \neq \hat{G}(x_0) | X = x_0)
\end{aligned} \tag{7.2}$$

- If  $f(x_0) \geq 1/2$  therefore  $G(x_0) = 1$ , we have

$$\Pr(G(x_0) \neq \hat{G}(x_0) | X = x_0) = \Pr(\hat{f}(x_0) < 1/2) \approx \Phi \left( \frac{-\mathbb{E} [\hat{f}(x_0)] + 1/2}{\sqrt{\text{var}(\hat{f}(x_0))}} \right) \tag{7.3}$$

- Otherwise  $f(x_0) < 1/2$  thus  $G(x_0) = 0$ , we have

$$\Pr(G(x_0) \neq \hat{G}(x_0)|X = x_0) = \Pr(\hat{f}(x_0) \geq 1/2) \approx \Phi\left(\frac{\mathbb{E}[\hat{f}(x_0)] - 1/2}{\sqrt{\text{var}(\hat{f}(x_0))}}\right) \quad (7.4)$$

Consequently, Eq. (7.63) is proved.

### Ex. 7.3

(a) Denote  $\mathbf{x}_i$  as the column vector representing the  $i$ -th record, i.e. the transpose of the  $i$ -th row of  $\mathbf{X}$ .

$$y_i - \hat{f}^{-i}(\mathbf{x}_i) = y_i - \mathbf{x}_i^T (\mathbf{X}_{-i}^T \mathbf{X}_{-i})^{-1} \mathbf{X}_{-i}^T \mathbf{y}_{-i} \quad (7.5)$$

where

$$\begin{aligned} (\mathbf{X}_{-i}^T \mathbf{X}_{-i})^{-1} &= (\mathbf{X}^T \mathbf{X} - \mathbf{x}_i \mathbf{x}_i^T)^{-1} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i (-1 + \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i)^{-1} \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} + \frac{1}{1 - S_{ii}} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \end{aligned} \quad (7.6a)$$

$$\mathbf{X}_{-i}^T \mathbf{y}_{-i} = \mathbf{X}^T \mathbf{y} - \mathbf{x}_i y_i \quad (7.6b)$$

therefore

$$\begin{aligned} y_i - \hat{f}^{-i}(\mathbf{x}_i) &= y_i - \left[ \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} + \frac{1}{1 - S_{ii}} \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \right. \\ &\quad \left. - \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i y_i - \frac{1}{1 - S_{ii}} \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i y_i \right] \\ &= y_i - \left[ \hat{f}(\mathbf{x}_i) + \frac{S_{ii}}{1 - S_{ii}} \hat{f}(\mathbf{x}_i) - S_{ii} y_i - \frac{S_{ii}^2}{1 - S_{ii}} y_i \right] \\ &= \frac{1}{1 - S_{ii}} (y_i - \hat{f}(\mathbf{x}_i)) \end{aligned} \quad (7.7)$$

(b) Since  $\mathbf{S} = \mathbf{U}\mathbf{U}^T$ , where  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$  is the SVD of  $\mathbf{X}$ ,  $0 \leq S_{ii} \leq 1$ , thus

$$|y_i - \hat{f}^{-i}(\mathbf{x}_i)| \geq |y_i - \hat{f}(\mathbf{x}_i)| \quad (7.8)$$

(c) ???

### Ex. 7.4

Denote  $\hat{f}(\mathbf{x}_i) = \hat{y}_i$  (which of course also depends on  $\mathbf{X}, \mathbf{y}$ ), then

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\text{Err}_{\text{in}}] - \mathbb{E}_{\mathbf{y}}[\overline{\text{err}}] &= \frac{1}{N} \mathbb{E}_{\mathbf{y}, \mathbf{y}^0} \left[ \sum_{i=1}^N \|y_i^0 - \hat{y}_i\|^2 \right] - \frac{1}{N} \mathbb{E}_{\mathbf{y}, \mathbf{y}^0} \left[ \sum_{i=1}^N \|y_i - \hat{y}_i\|^2 \right] \\
&= \frac{2}{N} \sum_{i=1}^N \mathbb{E}_{\mathbf{y}} [y_i \hat{y}_i] - \mathbb{E}_{\mathbf{y}, \mathbf{y}^0} [y_i^0 \hat{y}_i] \\
&= \frac{2}{N} \sum_{i=1}^N \mathbb{E}_{\mathbf{y}} [y_i \hat{y}_i] - \mathbb{E}_{\mathbf{y}} [y_i] \mathbb{E}_{\mathbf{y}} [\hat{y}_i] \\
&= \frac{2}{N} \sum_{i=1}^N \text{cov}(y_i, \hat{y}_i)
\end{aligned} \tag{7.9}$$

**Ex. 7.5**

$$\begin{aligned}
\sum_{i=1}^N \text{cov}(y_i, \hat{y}_i) &= \text{tr} \left( \mathbb{E} [(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{S}(\mathbf{y} - \bar{\mathbf{y}}))^T] \right) \\
&= \text{tr} \left( \mathbb{E} [(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})^T] \mathbf{S} \right) \\
&= \text{tr}(\mathbf{S}) \sigma_{\epsilon}^2
\end{aligned} \tag{7.10}$$

**Ex. 7.6**

For  $k$ -nearest-neighbor regression,  $\mathbf{S}$  equals to  $1/k$  times a  $N$ -by- $N$  binary matrix which

- Each row has exactly  $k$  1s and  $N - k$  0s.
- The diagonal entries are all 1s.

Therefore  $\text{tr}(\mathbf{S}) = N/k$ .

**Ex. 7.7**

$$\begin{aligned}
\text{GCV} &= \frac{1}{N(1 - d/N)^2} \sum_{i=1}^N \left( y_i - \hat{f}(\mathbf{x}_i) \right)^2 \\
&\approx \frac{1}{N} \left( 1 + \frac{2d}{N} \right) \sum_{i=1}^N \left( y_i - \hat{f}(\mathbf{x}_i) \right)^2 \\
&= \overline{\text{err}} + \frac{2d}{N} \hat{\sigma}_{\epsilon}^2 \\
&= C_p
\end{aligned} \tag{7.11}$$

**Ex. 7.8**

$\alpha$  can be constructed as

$$\alpha = \frac{\pi}{2} \left( 10^{l+1} + \sum_{i=1}^l d_i 10^i \right) \quad (7.12)$$

where  $d_i = 1$  if the  $i$ -th point is assigned label 0 and  $d_i = 3$  otherwise. Apparently,  $\sin(\alpha z^i) < 0$  if  $d_i = 1$  and  $\sin(\alpha z^i) > 0$  otherwise, thus  $\sin(\alpha x)$  shatters  $z^1, \dots, z^l$ .

**Ex. 7.9 (Program)****Ex. 7.10**

No, this is not the right way to do CV. One should select a different predictor individually for each validation set and then carries out the CV.

## Chapter 8

### Model Inference and Averaging

## Chapter 9

### Additive Models, Trees, and Related Methods

## Chapter 10

### Boosting and Additive Trees



# Chapter 11

## Neural Networks

### Ex. 11.1

In (11.5), set  $K = 1$ ,  $g_1(T) = T$ , we have

$$f_1(X) = \beta_{01} + \beta_1^T Z = \beta_{01} + \sum_{m=1}^M \beta_{m1} \sigma(\alpha_{0m} + \alpha_m^T X) \quad (11.1)$$

The correspondence between (11.1) and (11.5) becomes clearer, as enumerated in Table 11.1

Table 11.1: Correspondence between the project pursuit regression and the neural network

(11.1)	(11.5)
$\omega_m$	$\alpha_m$
$g_m(\cdot)$	$\beta_{01}, \beta_{m1} \sigma(\alpha_{0m} + \alpha_m^T X)$

### Ex. 11.2

$$\frac{\partial f}{\partial X} = \sum_{m=1}^M \beta_m [\sigma(\cdot)(\sigma(\cdot) - 1)] \alpha_m \quad (11.2)$$

$$\frac{\partial^2 f}{\partial X \partial X^T} = \sum_{m=1}^M \beta_m [(2\sigma(\cdot) - 1)(\sigma(\cdot) - 1)\sigma(\cdot)] \alpha_m \alpha_m^T \quad (11.3)$$

Since  $\sigma(\alpha_{0m} + \alpha_m^T X) \approx 1/2$  when  $\alpha_{0m} \approx 0$  and  $\alpha_m \approx 0$ , therefore  $\frac{\partial^2 f}{\partial X \partial X^T} \approx 0$ , i.e. the resulting model is nearly linear.

### Ex. 11.3

$$R(\theta) = - \sum_{i=1}^N R_i(\theta) = - \sum_{i=1}^N \sum_{j=1}^K y_{ij} \log g_j(T) \quad (11.4)$$

Note that different from regression, each softmax function  $g_j(T)$ ,  $j = 1, \dots, K$  is a function

of all  $T_1, \dots, T_K$ .

$$\frac{\partial R_i}{\partial \beta_{km}} = - \sum_{j=1}^K \frac{y_{ij}}{g_j} \frac{\partial g_j}{\partial T_k} z_{mi} = \delta_{ki} z_{mi} \quad (11.5a)$$

$$\begin{aligned} \frac{\partial R_i}{\partial \alpha_{ml}} &= - \sum_{j=1}^K \frac{y_{ij}}{g_j} \sum_{k=1}^K \frac{\partial g_j}{\partial T_k} \beta_{km} \sigma'(\alpha_m^T x_i) x_{il} \\ &= \left[ \sigma'(\alpha_m^T x_i) \sum_{k=1}^K \beta_{km} \delta_{ki} \right] x_{il} = s_{mi} x_{il} \end{aligned} \quad (11.5b)$$

It is noted that

$$\frac{1}{g_j} \frac{\partial g_j}{\partial T_k} = \begin{cases} 1 - g_j & j = k \\ -g_k / \exp(T_j) & j \neq k \end{cases} \quad (11.6)$$

As a result, although  $g_j(T)$  depends on all  $T_1, \dots, T_K$ ,  $(\partial g_j / \partial T_k) / g_j$  can still be locally evaluated and propagated downward over the link  $(T_k, g_j)$ . Consequently, the forward and backward propagation equations are pretty much the same as those for the square error loss function. In the forward pass for record  $x_i$ ,  $i = 1, \dots, N$ , the weights  $\beta_{km}$  and  $\alpha_{ml}$  are fixed and the predicted  $\hat{g}_j(T_i)$  are evaluated. In the backward pass,  $(y_{ij}/g_j)(\partial g_j / \partial T_k)$  are evaluated and propagated to  $T_k$ , where  $\delta_{ki}$  is computed, and then back-propagated to give  $s_{mi}$  at  $Z_m$ . Then the gradients are evaluated as in Eq. (11.5). The gradient descent update is exactly the same as (11.13).

#### Ex. 11.4

If the network has no hidden layer, we have

$$g_j(x) = \frac{\exp(T_j)}{\sum_{k=1}^K \exp(T_k)} = \frac{\exp(\beta_j^T x)}{\sum_{k=1}^K \exp(\beta_k^T x)}, \quad (11.7)$$

exactly the same as the multinomial logistic model.

#### Ex. 11.5 (Program)

#### Ex. 11.6 (Program)

#### Ex. 11.7 (Program)

## Chapter 12

### Support Vector Machines and Flexible Discriminants

#### Ex. 12.1

Firstly, we prove that for (12.8), the optimal solution must satisfy  $\hat{\xi}_i = [1 - y_i(x_i^T \hat{\beta} + \hat{\beta}_0)]_+$ . To see this, from the constraints in (12.8), we have  $\hat{\xi}_i \geq [1 - y_i(x_i^T \hat{\beta} + \hat{\beta}_0)]_+$ . Assume for contradiction that  $\exists i$  such that  $\hat{\xi}_i > [1 - y_i(x_i^T \hat{\beta} + \hat{\beta}_0)]_+$ , then setting  $\hat{\xi}_i \leftarrow [1 - y_i(x_i^T \hat{\beta} + \hat{\beta}_0)]_+$  results in smaller objective in (12.8), which is in contradiction to the fact that  $\hat{\xi}_i$  is from an optimal solution.

On the other hand,  $\xi_i = [1 - y_i(x_i^T \beta + \beta_0)]_+ \Rightarrow \xi_i \geq 0, y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i$ . Therefore, the solution to (12.8) is the same as

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \xi_i \quad (12.1)$$

$$\text{s.t. } \xi_i = [1 - y_i(x_i^T \beta + \beta_0)]_+, \forall i \quad (12.2)$$

which is exactly the same as (12.25).

#### Ex. 12.2

Define kernel  $K(a, b) = \sum_{j=1}^p a_j b_j$ , i.e.  $\psi_j(x) = x_j, \gamma_j = 1$  for  $j = 1, \dots, p$ . Consequently,  $g(x) = \sum_{j=1}^p \beta_j x_j \Leftrightarrow g(x) \in \mathcal{H}_K$ . Consequently,

$$(12.25) \Leftrightarrow \min_{g, \beta_0} \sum_{i=1}^N [1 - y_i(g(x_i) + \beta_0)]_+ + \frac{\lambda}{2} \|g\|_{\mathcal{H}_K}^2 \quad (12.3)$$

Denote  $L(y_i, g(x_i); \beta_0) = [1 - y_i(g(x_i) + \beta_0)]_+ = L_i(\beta_0)$ , then

$$(12.25) \Leftrightarrow \min_{\beta_0} \left\{ \min_g \sum_{i=1}^N L_i(\beta_0) + \frac{\lambda}{2} \|g\|_{\mathcal{H}_K}^2 \right\}. \quad (12.4)$$

where the inner min must have a solution in the form of  $g(x) = \sum_{i=1}^N \alpha_i K(x, x_i)$  as per

(5.50)(5.51), and we have  $\|g\|_{\mathcal{H}_K}^2 = \alpha^T K \alpha$ . Therefore

$$(12.25) \Leftrightarrow \min_{\beta_0} \left\{ \min_{\alpha} \sum_{i=1}^N [1 - y_i (\sum_{j=1}^N \alpha_j K(x_j, x_i) + \beta_0)]_+ + \frac{\lambda}{2} \alpha^T K \alpha \right\} \quad (12.5)$$

$$\Leftrightarrow \min_{\beta_0, \alpha} \sum_{i=1}^N [1 - y_i (\sum_{j=1}^N \alpha_j K(x_j, x_i) + \beta_0)]_+ + \frac{\lambda}{2} \alpha^T K \alpha \quad (12.6)$$

### Ex. 12.3

Similar to Ex. (12.2). Denote  $g(x) = \sum_{m=1}^M \beta_m h_m(x)$ . Without penalting the constant term, we have

$$H(\beta, \beta_0) = \sum_{i=1}^N V(y_i - \beta_0 - g(x_i)) + \frac{\lambda}{2} \sum_{m=1}^M \beta_m \quad (12.7)$$

Again we break the minimization problem into 2 steps:

$$\min_{\beta_0, \beta} H(\beta, \beta_0) = \min_{\beta_0} \left\{ \min_{\beta | \beta_0} H(\beta, \beta_0) \right\} \quad (12.8)$$

Consider saquare error loss  $V(r) = r^2$ , the inner min problem is in the form of

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - H\beta)^2 + \frac{\lambda}{2} \beta^T \beta \quad (12.9)$$

$$\Leftrightarrow \min_{\alpha} \|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{K}\alpha\|_F^2 + \frac{\lambda}{2} \alpha^T \mathbf{K}\alpha \quad (12.10)$$

whose solution is  $\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I}/2)^{-1} \mathbf{y}_{\beta_0}$ ,  $\mathbf{y}_{\beta_0} = \mathbf{y} - \beta_0 \mathbf{1}$ . Consequently, the outer min problem w.r.t  $\beta_0$  is in the form of

$$\min_{\beta_0} \mathbf{y}_{\beta_0}^T [\mathbf{I} - (\mathbf{K} + \lambda \mathbf{I}/2)^{-1} \mathbf{K}] \mathbf{y}_{\beta_0} \quad (12.11)$$

which is a quadratic problem.

### Ex. 12.4

(a)

$$\text{Left} = (x - \bar{x}_k)^T U U^T (x - \bar{x}_k) - (x - \bar{x}_{k'})^T U U^T (x - \bar{x}_{k'}) \quad (12.12)$$

where  $U = W^{-1/2} V^*$ , the  $L$  columns of  $V^*$  are the eigen vectors of  $B^* = (W^{-1/2})^T B W^{-1/2}$ ,

where  $B$  is the between-class covariance.

$$\text{Right} = (x - \bar{x}_k)^T W^{-1} (x - \bar{x}_k) - (x - \bar{x}_{k'})^T W^{-1} (x - \bar{x}_{k'}) \quad (12.13)$$

Consequently,

$$\begin{aligned} & \text{Left} - \text{Right} \\ &= 2(\bar{x}_k - \bar{x}_{k'})^T (W^{-1} - UU^T)x + (\bar{x}_k - \bar{x}_{k'})^T [W^{-1} - UU^T](\bar{x}_k + \bar{x}_{k'}) \end{aligned} \quad (12.14)$$

$$\begin{aligned} &= 2(\bar{x}_k - \bar{x}_{k'})^T W^{-1/2} (I - V^*(V^*)^T) (W^{-1/2})^T x \\ &+ (\bar{x}_k - \bar{x}_{k'})^T W^{-1/2} (I - V^*(V^*)^T) (W^{-1/2})^T (\bar{x}_k + \bar{x}_{k'}) \end{aligned} \quad (12.15)$$

Since  $(\bar{x}_k - \bar{x}_{k'})^T \in R(M)$  (row space),  $(\bar{x}_k - \bar{x}_{k'})^T W^{-1/2} \in R(M^*)$ , therefore  $(W^{-1/2})^T (\bar{x}_k - \bar{x}_{k'}) \in C(V^*)$  (column space). Therefore

$$(\bar{x}_k - \bar{x}_{k'})^T W^{-1/2} (I - V^*(V^*)^T) = 0 \quad (12.16)$$

thus Left = Right.

(b) ???

### Ex. 12.5 (Program)

### Ex. 12.6

(a) The  $i$ -th row of  $\mathbf{Y}\theta$  is

$$(\mathbf{Y}\theta)_i = \sum_{j=1}^K 1(Y_{ij} = 1)\theta_j = \theta(g_i) \quad (12.17)$$

(since there are exactly one  $j$  where  $Y_{ij} = 1$  for each  $i$ ). The  $i$ -th row of  $\mathbf{H}\beta$  is

$$(\mathbf{H}\beta)_i = \sum_{j=1}^K \beta_j h_j(x_i) \quad (12.18)$$

therefore

$$\sum_{i=1}^N (\theta(g_i) - \beta^T h(x_i))^2 = \|\mathbf{Y}\theta - \mathbf{H}\beta\|^2 \quad (12.19)$$

(b) According to the definition,  $(\mathbf{D}_\pi)_{kk}$  is the empirical frequency of class  $k$ , and  $\theta_k$  is the score for class  $k$ .  $\theta^T \mathbf{D}_\pi \mathbf{1} = 0$  implies that the average score over the  $N$  records is 0;  $\theta^T \mathbf{D}_\pi \theta = 1$  means the variance of the over the  $N$  records is 1.

(c) Fixing  $\theta$  the optimal  $\beta$  is

$$\hat{\beta} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{Y} \theta \quad (12.20)$$

therefore (12.65) can be rewritten as

$$\min_{\theta} \|(\mathbf{I} - \mathbf{S})\mathbf{Y}\theta\|^2 \Leftrightarrow \min_{\theta} \theta^T \mathbf{Y}^T \mathbf{Y} \theta - \theta^T \mathbf{Y}^T \mathbf{S} \mathbf{Y} \theta \quad (12.21)$$

where  $\mathbf{S} = \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$ . Since  $\theta^T \mathbf{Y}^T \mathbf{Y} \theta = N$ , this minimization is equivalent to

$$\max_{\theta} \theta^T \mathbf{Y}^T \mathbf{S} \mathbf{Y} \theta \quad (12.22)$$

(d) Suppose that the SVD of  $\mathbf{H} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$ , then  $\mathbf{S} = \mathbf{U} \mathbf{U}^T$  where  $\mathbf{U}$  is a  $N$ -by- $l$  orthonormal matrix. Therefore  $\mathbf{S}$  has  $L$  eigenvalues of 1 and  $N - L$  eigenvalues of 0. Since constant function is included in  $h_j$ ,  $\mathbf{H} \neq 0$ , therefore  $L > 0$ , so the largest eigenvalue is 1.

(e) (12.53) can be rewritten as

$$ASR = \frac{1}{N} \|\mathbf{Y}\Theta - \mathbf{H}\mathbf{B}\|_F^2 \quad (12.23)$$

Similar to (c) the solution is the same as

$$\max_{\Theta} \text{tr}\{\Theta^T \mathbf{Y}^T \mathbf{S} \mathbf{Y} \Theta\} \quad (12.24)$$

$$\text{s.t. } \Theta^T \mathbf{Y}^T \mathbf{Y} \Theta = \mathbf{I} \quad (12.25)$$

Therefore  $\mathbf{Y}\Theta$  are the  $K$  largest eigenvectors of  $\mathbf{S}$ .

### Ex. 12.7

The penalized optimal scoring problem is in the form of

$$\min_{\Theta, \mathbf{B}} \|\mathbf{Y}\Theta - \mathbf{H}\mathbf{B}\|_F^2 + \lambda \text{tr}(\mathbf{B}^T \mathbf{\Omega} \mathbf{B}) \quad (12.26)$$

Given  $\Theta$ , the optimal  $\mathbf{B}$  is

$$\hat{\mathbf{B}} = (\mathbf{H}^T \mathbf{H} + \lambda \mathbf{\Omega})^{-1} \mathbf{H}^T \mathbf{Y} \Theta \quad (12.27)$$

Substitute into Eq. (12.26), we have

$$\min_{\Theta} \text{tr}(\Theta^T \mathbf{Y}^T \mathbf{Y} \Theta - \Theta^T \mathbf{Y}^T \mathbf{S} \mathbf{Y} \Theta) \quad (12.28)$$

$$\text{s.t. } \Theta^T \mathbf{D}_{\pi} \Theta = \mathbf{I} \quad (12.29)$$

where  $\mathbf{S} = \mathbf{H}(\mathbf{H}^T \mathbf{H} + \lambda \mathbf{\Omega})^{-1} \mathbf{H}^T$ . Therefore  $\mathbf{Y}\mathbf{\Theta}$  are still the eigenvectors of  $\mathbf{S}$ .

### Ex. 12.8

I found the proof to this problem on [4]. I am trying to follow it the best I can and here is my interpretation. Assuming that  $\bar{x} = 0$ . We first perform the generalized SVD:

$$(\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{U} \mathbf{D} \mathbf{V}^T \quad (12.30)$$

$$\text{s.t. } \mathbf{U}^T \mathbf{Y}^T \mathbf{Y} \mathbf{U} = \mathbf{I}, \mathbf{V}^T \mathbf{X}^T \mathbf{X} \mathbf{V} = \mathbf{I} \quad (12.31)$$

Later we will show that both  $\beta_l$  and  $v_l$  are proportional to the columns of  $\mathbf{V}$ . From the GSVD,  $\mathbf{U}$  and  $\mathbf{V}$  satisfy the following 2 equations:

$$\mathbf{U}^T \mathbf{Y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \mathbf{U} = \mathbf{D}^2 \quad (12.32a)$$

$$\mathbf{V}^T \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X} \mathbf{V} = \mathbf{D}^2 \quad (12.32b)$$

of which the proof is trivial. First we show that the LDA's discriminant directions  $v_l$  are parallel to the columns of  $\mathbf{V}$ :

**Proposition 12.1.** *For the LDA problem (Fisher)*

$$\max_{\mathbf{A}} \text{tr}(\mathbf{A}^T \mathbf{B} \mathbf{A}), \text{ s.t. } \mathbf{A}^T \mathbf{W} \mathbf{A} = \mathbf{I} \quad (12.33)$$

where

$$\mathbf{B} = \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X} \quad (12.34a)$$

$$\mathbf{W} = \mathbf{T} - \mathbf{B} \quad (12.34b)$$

$$\mathbf{T} = \mathbf{X}^T \mathbf{X} \quad (12.34c)$$

are the between-class, within-class and total variance (up to normalization), the solution is

$$\hat{\mathbf{A}} = \mathbf{V}(\mathbf{I} - \mathbf{D}^2)^{-1/2} \quad (12.35)$$

*Proof.* From Eq. (12.32b) and the second constraint of the GSVD, it is easy to see  $\hat{\mathbf{A}}^T \mathbf{W} \hat{\mathbf{A}} = \mathbf{I}$ . On the other hand,  $\hat{\mathbf{A}}$  diagonalizes  $\mathbf{B}$  by  $\hat{\mathbf{A}}^T \mathbf{B} \hat{\mathbf{A}} = (\mathbf{I} - \mathbf{D}^2)^{-1} \mathbf{D}^2$ .  $\square$

Next we show that the  $\beta_l$  from optimal scoring are also parallel to the columns of  $\mathbf{V}$

**Proposition 12.2.** *The optimal scoring problem as in Eq. (12.24) has solution  $\hat{\mathbf{\Theta}} = \mathbf{U}$ .*

*Proof.* From the first constraint of GSVD, obviously  $\mathbf{U}^T \mathbf{Y}^T \mathbf{Y} \mathbf{U} = \mathbf{I}$ . from Eq. (12.32a),  $\mathbf{U}$  diagonalizes  $\mathbf{Y}^T \mathbf{S} \mathbf{Y}$  by  $\mathbf{U}^T \mathbf{Y}^T \mathbf{S} \mathbf{Y} \mathbf{U} = \mathbf{D}^2$ .  $\square$

Consequently, we have  $\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \mathbf{U} = \mathbf{V} \mathbf{D}$ . We can see that  $v_l$  (columns of  $\hat{\mathbf{A}}$ ) and  $\beta_l$  (columns of  $\hat{\mathbf{B}}$ ) differ by only a diagonal matrix  $(\mathbf{I} - \mathbf{D}^2)^{-1} \mathbf{D}$ .

**Ex. 12.9**

The reduced features are simply

$$\mathbf{X}^* = \mathbf{X} \hat{\mathbf{B}} = \mathbf{S} \mathbf{Y} \quad (12.36)$$

therefore the optimal scoring can be computed by

$$\max_{\boldsymbol{\Theta}} \boldsymbol{\Theta}^T \mathbf{Y}^T [\mathbf{S} \mathbf{Y} (\mathbf{Y}^T \mathbf{S} \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{S}^T] \mathbf{Y} \boldsymbol{\Theta} \quad (12.37)$$

$$\text{s.t. } \boldsymbol{\Theta}^T \mathbf{Y}^T \mathbf{Y} \boldsymbol{\Theta} = N \mathbf{I} \quad (12.38)$$

with trivial manipulations one can see that the objective function is exactly the same as optimal scoring on original features.

**Ex. 12.10**

The derivation for the general  $K$ -class GDA can be found in [5]. The kernel LDA (Fisher) is in the form of

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}} \quad (12.39)$$

where

$$\mathbf{B} = (\bar{\mathbf{h}}_1 - \bar{\mathbf{h}}_2)(\bar{\mathbf{h}}_1 - \bar{\mathbf{h}}_2)^T \quad (12.40)$$

$$\bar{\mathbf{h}}_1 = \frac{1}{N_1} \sum_{i \in \mathcal{C}_1} \mathbf{h}(x_i) \quad (12.41)$$

$$\bar{\mathbf{h}}_2 = \frac{1}{N_2} \sum_{i \in \mathcal{C}_2} \mathbf{h}(x_i) \quad (12.42)$$

$$(12.43)$$

(up to constant) and  $N_1 = |\mathcal{C}_1|$ ,  $N_2 = |\mathcal{C}_2|$  are the numbers of data points in class 1, 2, respectively. The discriminant vector  $\mathbf{a}$  is a linear combination

$$\mathbf{a} = \sum_{i=1}^N \alpha_i \mathbf{h}(x_i) \quad (12.44)$$

therefore

$$\mathbf{a}^T \mathbf{B} \mathbf{a} = \boldsymbol{\alpha}^T (\mathbf{k}_1 - \mathbf{k}_2)(\mathbf{k}_1 - \mathbf{k}_2)^T \boldsymbol{\alpha} \quad (12.45)$$



where

$$\{\mathbf{k}_1\}_i = \frac{1}{N_1} \sum_{j \in \mathcal{C}_1} K_{ij}, \quad \{\mathbf{k}_2\}_i = \frac{1}{N_2} \sum_{j \in \mathcal{C}_2} K_{ij}, \quad i = 1, \dots, N \quad (12.46)$$

On the other hand,  $\mathbf{W} = \mathbf{W}_h + \gamma \mathbf{I}$ , where

$$\mathbf{W}_h = \sum_{i \in \mathcal{C}_1} (\mathbf{h}(x_i) \mathbf{h}(x_i)^T - \bar{\mathbf{h}}_1 \bar{\mathbf{h}}_1^T) + \sum_{i \in \mathcal{C}_2} (\mathbf{h}(x_i) \mathbf{h}(x_i)^T - \bar{\mathbf{h}}_2 \bar{\mathbf{h}}_2^T) \quad (12.47)$$

(up to constant) therefore

$$\mathbf{a}^T \mathbf{W}_h \mathbf{a} = \boldsymbol{\alpha}^T \mathbf{K}^2 \boldsymbol{\alpha} - N_1 \boldsymbol{\alpha}^T \mathbf{k}_1 \mathbf{k}_1^T \boldsymbol{\alpha} - N_2 \boldsymbol{\alpha}^T \mathbf{k}_2 \mathbf{k}_2^T \boldsymbol{\alpha} \quad (12.48)$$

and  $\mathbf{a}^T \mathbf{a} = \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}$ . Consequently, the model depend on  $h(\cdot)$  only via the  $N$ -by- $N$  matrix  $\mathbf{K}$ .

### Ex. 12.11

(a)

$$\begin{aligned} P(X = x | G = k) &= \frac{P(X = x, G = k)}{\int_x P(X = x, G = k) dx} \\ &= \frac{\sum_{r=1}^R \pi_r P_r(G = k) \phi(x; \mu_r, \Sigma)}{\sum_{r=1}^R \pi_r P_r(G = k) \int_x \phi(x; \mu_r, \Sigma) dx} \\ &= \frac{\sum_{r=1}^R \pi_r P_r(G = k) \phi(x; \mu_r, \Sigma)}{\sum_{r=1}^R \pi_r P_r(G = k)} \end{aligned} \quad (12.49)$$

Compare with (12.59), we can see that, by setting

$$\pi_{kr} = \frac{\sum_{r=1}^R \pi_r P_r(G = k)}{\sum_{r=1}^R \pi_r P_r(G = k)}, \quad R_k = R, \quad \mu_{kr} = \mu_r \quad (12.50)$$

MDA2 is a generalization of MDA.

(b) E-step: compute the responsibility of subclass  $c_{kr}$  within class  $k$  for each class- $k$  observation ( $g_i = k$ ):

$$W(c_{kr} | x_i, g_i) = \frac{\pi_{kr} \phi(x_i; \mu_r, \Sigma)}{\sum_{r=1}^R \pi_{kr} \phi(x_i; \mu_r, \Sigma)} \quad (12.51)$$

M-step: MLE on  $\mu_r$  and  $\Sigma$

$$\mu_r = \frac{\sum_{k=1}^K \sum_{i:g_i=k} W(c_{kr}|x_i, g_i) x_i}{\sum_{k=1}^K \sum_{i:g_i=k} W(c_{kr}|x_i, g_i)} \quad (12.52)$$

$$\Sigma = \frac{\sum_{k=1}^K \sum_{i:g_i=k} \sum_{r=1}^R W(c_{kr}|x_i, g_i) (x_i - \mu_r)(x_i - \mu_r)^T}{\sum_{k=1}^K \sum_{i:g_i=k} \sum_{r=1}^R W(c_{kr}|x_i, g_i)} \quad (12.53)$$

(c) ???

## Chapter 13

### Prototype Methods and Nearest-Neighbors

#### Ex. 13.1

Again  $k = 1, \dots, K$  are the indices of clusters/classes,  $r = 1, \dots, R$  are the indices of cluster centers/Gaussian components.

For each predictor labeled to class  $k$ , namely  $\{x_i | g_i = k\}$ , in terms of the E-step,

- For EM algorithm, its responsibility to component  $r$  is evaluated as

$$\gamma_{kr}(x_i) = \frac{\pi_{kr} \phi(x_i; \mu_{kr}, \Sigma)}{\sum_{s=1}^R \pi_{kr} \phi(x_i; \mu_{ks}, \Sigma)} \quad (13.1)$$

- For k-means algorithm, each predictor belongs to exactly 1 cluster center, therefore its counterpart of responsibility is binary-valued:

$$\gamma_{kr}(x_i) = \begin{cases} 1 & \text{if } \|x_i - \mu_{kr}\| \leq \|x_i - \mu_{ks}\|, s = 1, \dots, R \\ 0 & \text{otherwise} \end{cases} \quad (13.2)$$

In terms of the M-step,

- For EM algorithm, the component means are updated as a weighted average

$$\mu_{kr} = \frac{\sum_{i:g_i=k} \gamma_{kr}(x_i) x_i}{\sum_{i:g_i=k} \gamma_{kr}} \quad (13.3)$$

and the mixing probability is updated as

$$\pi_{kr} = \sum_{i:g_i=k} \frac{\gamma_{kr}(x_i)}{N_k} \quad (13.4)$$

where  $N_k$  is the number of records labeled to class  $k$ .

- For k-means, the cluster center is taken as an unweighted average over all  $x_i$  closest to it.

Using the binary-valued responsibility definition,  $\mu_{kr}$  is exactly the same as Eq. (13.3).

To draw a connection between EM and k-means, as  $\sigma \rightarrow 0$ ,  $\phi(x_i; \mu_{kr}, \Sigma) \gg \phi(x_i; \mu_{ks}, \Sigma)$  if  $\|x_i - \mu_{kr}\| < \|x_i - \mu_{ks}\|$ , therefore the responsibility for EM approaches that of k-means. And  $\pi_{kr}$  becomes the proportion of points in  $\{x_i | g_i = k\}$  that are closer to  $r$  than any other components centers.

**Ex. 13.2**

This problem is similar as Ex 2.3. Denote  $P_0(r, N)$  as the probability of the following event: “Among the  $N$  i.i.d uniformly distributed points, there is none within the ball centered at 0 with radius of  $r$ .” thus  $P_0(r, N) = P_0(r)^N$ , where  $P_0(r) = P_0(r, 1)$ . Since the points are uniformly distributed within the  $p$ -dim cube of edge length 1,  $P_0(r, 1)$  simply equals to the ratio between the volume of spaces out of the  $r$ -ball but within the cube, and the volume of the cube. Consequently,

$$P_0(r, N) = (1 - v_p r^p)^N \quad (13.5)$$

The median  $R_{med}$  satisfy  $P_0(R_{med}, N) = 1/2$ , therefore

$$R_{med} = v_p^{-1/p} (1 - 2^{-1/N})^{1/p} \quad (13.6)$$

**Ex. 13.3**

Since  $\sum_{k \neq k^*} p_k(x) = 1 - p_{k^*}(x)$

$$\begin{aligned} \sum_{k=1}^K p_k(x)(1 - p_k(x)) &= p_{k^*}(x)(1 - p_{k^*}(x)) + \sum_{k \neq k^*} p_k(x)(1 - p_k(x)) \\ &= p_{k^*}(x)(1 - p_{k^*}(x)) + (1 - p_{k^*}(x)) - \sum_{k \neq k^*} p_k(x)^2 \\ &\leq p_{k^*}(x)(1 - p_{k^*}(x)) + (1 - p_{k^*}(x)) - \frac{(\sum_{k \neq k^*} p_k(x))^2}{K-1} \\ &= (1 - p_{k^*}(x)) - \frac{K}{K-1} (1 - p_{k^*}(x))^2 \end{aligned} \quad (13.7)$$

where we made use of Cauchy-Schwarz inequality.

**Ex. 13.4**

(a)

$$\begin{aligned} \mathbf{A} &= \mathbf{Q}\mathbf{R} \\ &= \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix} \end{aligned} \quad (13.8)$$

where  $\theta$  represents the angle of rotation,  $a$ ,  $b$  represent the scale in  $x$  and  $y$  direction and  $\lambda$  represent the shear.

(b) Denote  $\mathbf{J}$  as the Jacobian (1-by-2) of  $F$  at  $c + x_0 + \mathbf{A}(\mathbf{x} - \mathbf{x}_0)$ , we have

$$\frac{\partial F}{\partial \theta} = \mathbf{J} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \mathbf{A}(\mathbf{x} - \mathbf{x}_0) \quad (13.9a)$$

$$\frac{\partial F}{\partial a} = \mathbf{J} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \begin{bmatrix} 1 & \lambda \end{bmatrix} (\mathbf{x} - \mathbf{x}_0) \quad (13.9b)$$

$$\frac{\partial F}{\partial b} = \mathbf{J} \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} (\mathbf{x} - \mathbf{x}_0) \quad (13.9c)$$

$$\frac{\partial F}{\partial \lambda} = \mathbf{J} \begin{bmatrix} a \cos \theta \\ a \sin \theta \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} (\mathbf{x} - \mathbf{x}_0) \quad (13.9d)$$

(c) It seems that the key to this procedure is to evaluate  $\mathbf{J}$  given a coordinate  $\mathbf{x}$ . Denote the 2-D kernel smoother as  $K(\mathbf{u}, \mathbf{v})$ , then we solve the locally weighted regression at  $\mathbf{x}$ :

$$\min_{\alpha(\mathbf{x}), \beta(\mathbf{x})} \sum_{i=1}^{256} K(\mathbf{x}, \mathbf{x}_i) [F(\mathbf{x}_i) - \alpha(\mathbf{x}) - \beta(\mathbf{x})^T \mathbf{x}_i] \quad (13.10)$$

Then we can use  $\beta(\mathbf{x})^T$  as  $\mathbf{J}(\mathbf{x})$  to compute the tangent space.

### Ex. 13.5

Since

$$N \operatorname{tr}(\bar{\mathbf{B}}\bar{\mathbf{B}}^T) = \sum_{i=1}^N \operatorname{tr}(\mathbf{B}_i\bar{\mathbf{B}}^T) = \sum_{i=1}^N \operatorname{tr}(\bar{\mathbf{B}}\mathbf{B}_i^T) \quad (13.11a)$$

$$N \operatorname{tr}(\bar{\mathbf{B}}\mathbf{M}^T) = N \operatorname{tr}(\mathbf{M}\bar{\mathbf{B}}^T) = \sum_{i=1}^N \operatorname{tr}(\mathbf{B}_i\mathbf{M}^T) = \sum_{i=1}^N \operatorname{tr}(\mathbf{M}\mathbf{B}_i^T), \quad (13.11b)$$

it is easy to show that

$$\sum_{i=1}^N \operatorname{tr}[(\mathbf{B}_i - \mathbf{M})^2] = \sum_{i=1}^N \operatorname{tr}[(\mathbf{B}_i - \bar{\mathbf{B}})^2] + N \operatorname{tr}[(\mathbf{M} - \bar{\mathbf{B}})^2] \quad (13.12)$$

Therefore the rank- $L$  approximation of  $\mathbf{B}_i$  is equivalent to the rank- $L$  approximation of  $\bar{\mathbf{B}}$ , namely  $\bar{\mathbf{B}}_{[L]}$ .

### Ex. 13.6

( $\mathbf{L}_j, j = 1, \dots, M$  are the coordinates of the “black” parts of the cursive letter.) As the

optimal  $\mathbf{A}_j = \mathbf{V}^T \mathbf{L}_j$ , we have

$$\begin{aligned}
 \sum_{j=1}^M \min_{\mathbf{A}_j} \|\mathbf{L}_j - \mathbf{V} \mathbf{A}_j\|^2 &= \sum_{j=1}^M \|(\mathbf{I} - \mathbf{V} \mathbf{V}^T) \mathbf{L}_j\|^2 \\
 &= \sum_{j=1}^M \text{tr} [(\mathbf{I} - \mathbf{V} \mathbf{V}^T) \mathbf{L}_j \mathbf{L}_j^T (\mathbf{I} - \mathbf{V} \mathbf{V}^T)] \\
 &= \sum_{j=1}^M \text{tr} [(\mathbf{I} - \mathbf{V} \mathbf{V}^T) \mathbf{L}_j \mathbf{L}_j^T] \\
 &= \sum_{j=1}^M \text{tr} [\mathbf{U}_j \mathbf{\Sigma}_j^2 \mathbf{U}_j^T] - \text{tr} \left[ \mathbf{V}^T \left( \sum_{j=1}^M \mathbf{U}_j \mathbf{\Sigma}_j^2 \mathbf{U}_j^T \right) \mathbf{V} \right] \quad (13.13)
 \end{aligned}$$

where  $\mathbf{L}_j = \mathbf{U}_j \mathbf{\Sigma}_j \mathbf{V}_j^T$  is the SVD of  $\mathbf{L}_j$ . Therefore  $\mathbf{V}$  corresponds to the 2 largest eigenvectors of  $\sum_{j=1}^M \mathbf{U}_j \mathbf{\Sigma}_j^2 \mathbf{U}_j^T$ .

For the alternative approach,

$$\begin{aligned}
 \sum_{j=1}^M \min_{\mathbf{A}_j} \|\mathbf{L}_j \mathbf{A}_j^T - \mathbf{V}\|^2 &= \sum_{j=1}^M \|(\mathbf{I} - \mathbf{S}_j) \mathbf{V}\|^2 \\
 &= \sum_{j=1}^M \text{tr} [\mathbf{V}^T (\mathbf{I} - \mathbf{S}_j) (\mathbf{I} - \mathbf{S}_j)^T \mathbf{V}] \\
 &= \sum_{j=1}^M \text{tr} [\mathbf{V}^T (\mathbf{I} - \mathbf{S}_j) \mathbf{V}] \\
 &= 2M - \text{tr} \left[ \mathbf{V}^T \left( \sum_{j=1}^M \mathbf{S}_j \right) \mathbf{V} \right] \quad (13.14)
 \end{aligned}$$

where  $\mathbf{S}_j = \mathbf{L}_j (\mathbf{L}_j^T \mathbf{L}_j)^{-1} \mathbf{L}_j^T = \mathbf{U}_j \mathbf{U}_j^T$ . Therefore  $\mathbf{V}$  corresponds to the 2 largest eigenvectors of  $\sum_{j=1}^M \mathbf{U}_j \mathbf{U}_j^T$ .

**Ex. 13.7 (Program)**

**Ex. 13.8 (Program)**

# Chapter 14

## Unsupervised Learning

**Ex. 14.1**

$$\begin{aligned}
 d_e(z_i, z_{i'}) &= \sum_{l=1}^p (z_{il} - z_{i'l})^2 \\
 &= \sum_{l=1}^p (z_{il} - z_{i'l})^2 \frac{w_l}{\sum_{j=1}^p w_j} (x_{il} - x_{i'l})^2 \\
 &= \frac{\sum_{l=1}^p w_l (x_{il} - x_{i'l})^2}{\sum_{j=1}^p w_j} \\
 &= d_e^{(w)}(x_i, x_{i'})
 \end{aligned} \tag{14.1}$$

**Ex. 14.2**

(a) The log-likelihood of a given record  $\mathbf{x}_i$  is

$$l(\theta; \mathbf{x}_i) = -\frac{1}{2} \log |\mathbf{L}| - \frac{p}{2} \log 2\pi - \frac{p}{2} \log \sigma^2 + \log \left[ \sum_{k=1}^K \pi_k \exp(\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\sigma^2 \mathbf{L})^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right], \tag{14.2}$$

and the log-likelihood over the entire data set is simply  $l(\theta; \mathbf{X}) = \sum_{i=1}^N l(\theta; \mathbf{x}_i)$

(b) Suppose we enlarge the dataset with latent variable  $\boldsymbol{\Delta}$  ( $N$ -by- $K$ ) such that  $\Delta_{ik} = 1$  if  $\mathbf{x}_i$  is associated with the  $k$ -th component and 0 otherwise. Each  $\mathbf{x}_i$  is associated with exactly one  $k$ . The the loglikelihood on  $\mathbf{x}_i, \boldsymbol{\Delta}_i$  becomes

$$\begin{aligned}
 l(\theta; \mathbf{x}_i, \boldsymbol{\Delta}_i) &= -\frac{1}{2} \log |\mathbf{L}| - \frac{p}{2} \log 2\pi - \frac{p}{2} \log \sigma^2 \\
 &\quad + \sum_{k=1}^K \Delta_{ik} \left[ \log \pi_k - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\sigma^2 \mathbf{L})^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right]
 \end{aligned} \tag{14.3}$$

therefore

$$l(\theta; \mathbf{X}, \boldsymbol{\Delta}) = C - \frac{Np}{2} \log \sigma^2 + \sum_{i=1}^N \sum_{k=1}^K \Delta_{ik} \left[ \log \pi_k - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\sigma^2 \mathbf{L})^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right] \tag{14.4}$$

Now we can formulate the EM-algorithm. We replace  $\Delta_{ik}$  with responsibility  $\gamma_{ik}$ . For the maximization step, we evaluate the MLE of  $\sigma^2$  and  $\mu_k$ . Since

$$\frac{\partial l(\theta; \mathbf{X}, \Delta)}{\partial \sigma^2} = -\frac{Np}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} (\mathbf{x}_i - \mu_k)^T \mathbf{L}^{-1} (\mathbf{x}_i - \mu_k) \quad (14.5)$$

$$\frac{\partial l(\theta; \mathbf{X}, \Delta)}{\partial \mu_k} = \sum_{i=1}^N \gamma_{ik} (\mu_k - \mathbf{x}_i)^T (\sigma^2 \mathbf{L})^{-1} \quad (14.6)$$

By setting the partial derivative to 0, the MLE are

$$\hat{\mu}_k = \frac{\sum_{i=1}^N \gamma_{ik} \mathbf{x}_i}{\sum_{i=1}^N \gamma_{ik}} \quad (14.7)$$

$$\hat{\sigma}^2 = \frac{1}{Np} \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} (\mathbf{x}_i - \mu_k)^T \mathbf{L}^{-1} (\mathbf{x}_i - \mu_k) \quad (14.8)$$

and the MLE for  $\pi_k$  are solved by

$$\max_{\pi_k, l=1, \dots, K} \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} \log \pi_k \quad (14.9)$$

$$\text{s.t. } \sum_{k=1}^K \pi_k = 1 \quad (14.10)$$

therefore  $\hat{\pi}_k = \sum_{i=1}^N \gamma_{ik} / N$ .

For the expectation step, the responsibilities are updated as

$$\hat{\gamma}_{ik} = \frac{\pi_k \phi_k(\mathbf{x}_i)}{\sum_{l=1}^K \pi_l \phi_l(\mathbf{x}_i)} \quad (14.11)$$

where  $\phi_k(\cdot)$  is the PDF of  $\mathcal{N}(\mu_k, \sigma^2 \mathbf{L})$ .

(c) Pretty much the same as Ex 13.1. Now we are not dealing with classification so we don't need to treat  $\mathbf{x}_i$  with different labels separately.

**Ex. 14.3**

???

**Ex. 14.4 (Program)**

**Ex. 14.5 (Program)**



**Ex. 14.6 (Program)****Ex. 14.7**

$$\sum_{i=1}^N \|\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{V}_q \boldsymbol{\lambda}_i\| = \|\mathbf{X} - \mathbf{1}_N \boldsymbol{\mu}^T - \boldsymbol{\Lambda} \mathbf{V}_q^T\|_F^2 \quad (14.12)$$

which is minimized when  $\hat{\boldsymbol{\Lambda}} = (\mathbf{X} - \mathbf{1}_N \boldsymbol{\mu}^T) \mathbf{V}_q$  given  $\mathbf{V}_q$  and  $\boldsymbol{\mu}$ . Denote the null space of  $\mathbf{V}_q$  is represented by  $\tilde{\mathbf{V}}_q$  where  $\tilde{\mathbf{V}}_q^T \tilde{\mathbf{V}}_q = \mathbf{I}_{p-q}$ . Now (14.50) becomes

$$\min_{\boldsymbol{\mu}, \mathbf{V}_q} \|(\mathbf{X} - \mathbf{1}_N \boldsymbol{\mu}^T) \tilde{\mathbf{V}}_q\|_F^2 \quad (14.13)$$

Taking partial derivative w.r.t  $\boldsymbol{\mu}$ , we can see that given  $\mathbf{V}_q$ , the optimal  $\boldsymbol{\mu}$  satisfy

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} + \mathbf{V}_q \mathbf{b} \quad (14.14)$$

therefore  $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$  is an optimal solution for arbitrary  $\mathbf{V}_q$ .

**Ex. 14.8**

Since

$$\frac{\partial \|\mathbf{X}_2 - (\mathbf{X}_1 \mathbf{R}) + \mathbf{1} \boldsymbol{\mu}^T\|_F^2}{\partial \boldsymbol{\mu}} = 2N \boldsymbol{\mu}^T - 2\mathbf{1}^T \mathbf{X}_2 2\mathbf{1}^T \mathbf{X}_1 \mathbf{R} \quad (14.15)$$

by setting the partial derivative to 0, we have  $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}_2 - \mathbf{R} \bar{\mathbf{x}}_1$ . Substitute this result into (14.56) we get

$$\min_{\mathbf{R}} \|\tilde{\mathbf{X}}_2 - \tilde{\mathbf{X}}_1 \mathbf{R}\|_F^2 \quad (14.16)$$

s.t.  $\mathbf{R}$  is orthogonal.

which is a orthogonal procustes problem (wiki). Since

$$\min_{\mathbf{R}} \|\tilde{\mathbf{X}}_2 - \tilde{\mathbf{X}}_1 \mathbf{R}\|_F^2 \Leftrightarrow \max_{\mathbf{R}} \text{tr}(\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_2 \mathbf{R}^T) \Leftrightarrow \max_{\mathbf{R}} \text{tr}(\mathbf{D} \mathbf{V}^T \mathbf{R}^T \mathbf{U}) \quad (14.17)$$

which is maximized when  $\mathbf{R} = \mathbf{U} \mathbf{V}^T$

**Ex. 14.9**

(14.115) should be Procrustes average with scaling

$$\min_{\{\beta_l, \mathbf{R}_l\}_1^L, \mathbf{M}} \sum_{l=1}^L \|\beta_l \mathbf{X}_l \mathbf{R}_l - \mathbf{M}\|_F^2 \quad (14.18)$$

This problem can be solved (sub-optimally) with alternating optimization

(1) Given  $\mathbf{M}$ , the  $L$  pairs of  $(\beta_l, \mathbf{R}_l)$  can be solved independently

$$\min_{\beta_l, \mathbf{R}_l} \|\beta_l \mathbf{X}_l \mathbf{R}_l - \mathbf{M}\|_F^2 \quad (14.19)$$

(a) Given  $\mathbf{R}_l$ ,  $\beta_l$  is optimized as

$$\hat{\beta}_l = \frac{\text{tr}(\mathbf{X}_l \mathbf{R}_l \mathbf{M}^T)}{\text{tr}(\mathbf{X}_l \mathbf{X}_l^T)} \quad (14.20)$$

(b) Given  $\beta_l$ ,  $\mathbf{R}_l$  is optimized (orthogonal procrustes problem) as

$$\hat{\mathbf{R}}_l = \mathbf{U}_l \mathbf{V}_l^T \quad (14.21)$$

where  $\beta_l \mathbf{X}_l^T \mathbf{M} = \mathbf{U}_l \mathbf{D}_l \mathbf{V}_l^T$  is SVD. However, we note that  $\mathbf{U}_l, \mathbf{V}_l$  does not actually depend on  $\beta_l$ . Therefore, we can evaluate  $\hat{\mathbf{R}}_l$  with the SVD of  $\mathbf{X}_l^T \mathbf{M}$  and then evaluate  $\hat{\beta}_l$ .

(2) Given  $\beta_l, \mathbf{R}_l, \mathbf{M}$  is simply optimized as the average

$$\hat{\mathbf{M}} = \frac{1}{L} \sum_{l=1}^L \beta_l \mathbf{X}_l \mathbf{R}_l \quad (14.22)$$

The above 2 steps are taken alternately until convergence.

#### Ex. 14.10

Given  $\mathbf{M}$ ,  $\mathbf{A}_l$  are optimized as  $\hat{\mathbf{A}}_l = (\mathbf{X}_l^T \mathbf{X}_l)^{-1} \mathbf{X}_l^T \mathbf{M}$ . Consequently, (14.60) is equivalent to

$$\min_{\mathbf{M}} \sum_{l=1}^L \|(\mathbf{I} - \mathbf{H}_l) \mathbf{M}\|_F^2 \quad (14.23)$$

s.t.  $\mathbf{M} \mathbf{M}^T = \mathbf{I}$ , where  $\mathbf{H}_l = \mathbf{X}_l (\mathbf{X}_l^T \mathbf{X}_l)^{-1} \mathbf{X}_l^T$ . This in turn is equivalent to

$$\max_{\mathbf{M}} \sum_{l=1}^L \text{tr}(\mathbf{M}^T \mathbf{H}_l \mathbf{M}) \quad (14.24)$$

s.t.  $\mathbf{M} \mathbf{M}^T = \mathbf{I}$ . As a result,  $\hat{\mathbf{M}}$  is the  $p$  largest eigen vectors of  $\sum_{l=1}^L \mathbf{H}_l$ .

**Ex. 14.11**

???

**Ex. 14.12**

(a)

$$\begin{aligned}
& \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{\Theta} \mathbf{V}^T \mathbf{x}_i\|^2 \\
&= \|\mathbf{X}(\mathbf{I} - \mathbf{V} \mathbf{\Theta}^T)\|_F^2 \\
&= \text{tr}[\mathbf{X}(\mathbf{\Theta} - \mathbf{V})(\mathbf{\Theta} - \mathbf{V})^T \mathbf{X}^T] + \text{tr}[\mathbf{X} \mathbf{X}^T - \mathbf{X} \mathbf{\Theta} \mathbf{\Theta}^T \mathbf{X}] \quad (14.25)
\end{aligned}$$

where the second term is not dependent on  $\mathbf{V}$  and the first term equals to  $\sum_{i=1}^N \|\mathbf{\Theta}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{x}_i\|^2$ . Consequently, the minimization of (14.71) w.r.t  $\mathbf{V}$  becomes

$$\min_{\{\mathbf{v}_k\}_{k=1}^K} \sum_{k=1}^K \left[ \sum_{i=1}^N \|\boldsymbol{\theta}_k^T \mathbf{x}_i - \mathbf{v}_k^T \mathbf{x}_i\|_2^2 + \lambda \|\mathbf{v}_k\|_2^2 + \lambda_{1k} \|\mathbf{v}_k\|_1 \right] \quad (14.26)$$

which can be solved as  $K$  separate elastic net regression problems.

(b) We rewrite

$$\begin{aligned}
& \|\mathbf{X}(\mathbf{I} - \mathbf{V} \mathbf{\Theta}^T)\|_F^2 \\
&= \text{tr}[\mathbf{X} \mathbf{X}^T - \mathbf{X} \mathbf{V} \mathbf{V}^T \mathbf{X}^T] - 2 \text{tr}[\mathbf{\Theta}^T \mathbf{X}^T \mathbf{X} \mathbf{V}] \quad (14.27)
\end{aligned}$$

Since the rest part of (14.71) is not dependent on  $\mathbf{\Theta}$ , the minimization of (14.71) w.r.t  $\mathbf{\Theta}$  is equivalent to

$$\begin{aligned}
& \max_{\mathbf{\Theta}} \text{tr}(\mathbf{\Theta}^T \mathbf{M}) \\
& \text{s.t. } \mathbf{\Theta}^T \mathbf{\Theta} = \mathbf{I} \quad (14.28)
\end{aligned}$$

where  $\mathbf{M} = \mathbf{X}^T \mathbf{X} \mathbf{V}$ . This has the same form as the Procrustes problem in Ex 14.8, therefore its solution is  $\mathbf{\Theta} = \mathbf{U} \mathbf{Q}^T$ .

**Ex. 14.13 (Program)****Ex. 14.14**

Denote  $\mathbf{D}_A = \text{diag}(\{\mathbf{a}_j^T \mathbf{a}_j\}_{j=1}^p)$ , then

$$\mathbf{P} = \mathbf{D}_A^{-1/2} \boldsymbol{\Sigma} \mathbf{D}_A^{-1/2} = \mathbf{P}_A + \mathbf{D}_A^{-1} \mathbf{D}_\epsilon \quad (14.29)$$

where  $\mathbf{P}_A$  is the correlation matrix

$$\{\mathbf{P}_A\}_{ij} = \frac{\mathbf{a}_i^T \mathbf{a}_j}{\|\mathbf{a}_i\| \|\mathbf{a}_j\|} \quad (14.30)$$

**Ex. 14.15 (Program)**

**Ex. 14.16**

Since  $\mathbf{Z} = \tilde{\mathbf{K}}\mathbf{U}\mathbf{D}^{-1}$ , therefore

$$z_{im} = \sum_{j=1}^N \tilde{K}(x_i, x_j) u_{jm} d_m^{-1} \quad (14.31)$$

where  $\tilde{K}(x_i, x_j)$  differs from  $K(x_i, x_j)$  only in centering. For a new observation  $x_0$ , its mapping to the  $m$ -th component is

$$\langle \tilde{\phi}(x_0), \sum_{j=1}^N \alpha_{jm} \tilde{\phi}(x_j) \rangle = \sum_{j=1}^N \alpha_{jm} \langle \tilde{\phi}(x_0), \tilde{\phi}(x_j) \rangle = \sum_{j=1}^N \alpha_{jm} \tilde{K}(x_0, x_j) \quad (14.32)$$

which differs from  $\sum_{j=1}^N \alpha_{jm} K(x_0, x_j)$  only in centering. For more details see Ex 18.15.

**Ex. 14.17**

Denote  $\mathbf{c} = [c_1, \dots, c_N]^T$ . First we note

$$\|g_1(x)\|_{\mathcal{H}_K} = \sum_{i=1}^N \sum_{j=1}^N c_i c_j K(x_i, x_j) = \mathbf{c}^T \mathbf{K} \mathbf{c} \quad (14.33)$$

Secondly, we have

$$\begin{aligned} \text{Var}_{\mathcal{T}} g_1(X) &= \frac{1}{N} \sum_{k=1}^N g_1(x_k)^2 \\ &= \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^N \sum_{j=1}^N c_i c_j K(x_k, x_i) K(x_k, x_j) \\ &= \frac{1}{N} \mathbf{c}^T \mathbf{K} \mathbf{K} \mathbf{c} \end{aligned} \quad (14.34)$$

Since  $\mathbf{K} = \mathbf{U}\mathbf{D}^2\mathbf{U}^T$ , (14.66) can be rewritten as

$$\begin{aligned} &\max \mathbf{c}^T \mathbf{U} \mathbf{D}^4 \mathbf{U}^T \mathbf{c} \\ &\text{s.t. } \mathbf{c}^T \mathbf{U} \mathbf{D}^2 \mathbf{U}^T \mathbf{c} = 1 \end{aligned} \quad (14.35)$$

Denote  $\mathbf{a} = \mathbf{D}\mathbf{U}^T\mathbf{c}$ , then the optimal solution must satisfy  $\hat{\mathbf{a}} = [1, 0, \dots, 0]^T$ , therefore  $\hat{c} = \mathbf{u}_1/d_1$ .  $g_2, \dots, g_M$  can be derived in a similar manner.

**Ex. 14.18**

Consider the stationary condition on  $\theta_0$ , we have

$$\frac{\partial l}{\partial \theta_0} = 1 - \int \phi(t) \exp(\theta_0 + \theta_1 t + \theta_2 t^2) dt = 0 \quad (14.36)$$

also since  $\phi(t) \exp(\theta_0 + \theta_1 t + \theta_2 t^2) > 0$ , it is a probability density function.

Consider the stationary condition on  $\theta_1$ , we have

$$\frac{\partial l}{\partial \theta_1} = \frac{1}{N} \sum_{i=1}^N s_i - \int t \phi(t) \exp(\theta_0 + \theta_1 t + \theta_2 t^2) dt = 0 \quad (14.37)$$

since  $\sum_{i=1}^N s_i = 0$ , this condition suggest that  $\phi(t) \exp(\theta_0 + \theta_1 t + \theta_2 t^2)$  has zero mean.

Consider the stationary condition on  $\theta_2$ , we have

$$\frac{\partial l}{\partial \theta_2} = \frac{1}{N} \sum_{i=1}^N s_i^2 - \int t^2 \phi(t) \exp(\theta_0 + \theta_1 t + \theta_2 t^2) dt = 0 \quad (14.38)$$

since  $\sum_{i=1}^N s_i^2/N = 1$ , this condition suggest that  $\phi(t) \exp(\theta_0 + \theta_1 t + \theta_2 t^2)$  has unit variance.  
(???)

**Ex. 14.19**

$$\sum_{j=1}^p \sum_{i=1}^N \log \phi(\mathbf{a}_j^T \mathbf{x}_i) = -\frac{pN}{2} \log 2\pi - \frac{1}{2} \|\mathbf{A}\mathbf{X}\|_F^2. \quad (14.39)$$

Since  $\|\mathbf{A}\mathbf{X}\|_F^2 = \|\mathbf{X}\|_F^2$  for any orthogonal  $\mathbf{A}$ , this term does not depend on  $\mathbf{A}$ .

**Ex. 14.20**

Since

$$\frac{\partial g}{\partial a} = \mathbb{E}[X g'(a^T X)] \quad (14.40)$$

$$\frac{\partial^2 g}{\partial a \partial a^T} = \mathbb{E}[X X^T g''(a^T X)] \approx \mathbb{E}[g''(a^T X)] I \quad (14.41)$$

the Newton update is

$$a \leftarrow a - (\mathbb{E}[g''(a^T X)])^{-1} \mathbb{E}[X g'(a^T X)] \quad (14.42)$$

Since  $a$  needs to be normalized to ensure  $\|a\| = 1$  anyway, the right hand side of the above equation can be multiplied with a positive constant  $-\mathbb{E}[g''(a^T X)]$  (followed by a normalization), resulting in

$$a \leftarrow \mathbb{E}[Xg'(a^T X)] - \mathbb{E}[g''(a^T X)]a \quad (14.43)$$

**Ex. 14.21**

Since there are  $m$  connected components in the graph,  $\mathbf{L}$  can be transformed into a block-diagonal matrix

$$\mathbf{L} = \text{diag}(\mathbf{L}_1, \dots, \mathbf{L}_m) \quad (14.44)$$

where  $\mathbf{L}_j = \mathbf{G}_j - \mathbf{W}_j$ . Since  $\mathbf{L}_m \mathbf{1} = \mathbf{0}$ ,  $\mathbf{L}$  has  $m$  eigenvectors corresponding to eigenvalue of 0, which are the same as the permuted indicator vectors  $I_{A_1}, \dots, I_{A_m}$ .

**Ex. 14.22**

(a)

$$\begin{aligned} \mathbf{1}^T \mathbf{p} &= (1-d)\mathbf{1}^T \mathbf{e} + d\mathbf{1}^T \mathbf{L} \mathbf{D}_c^{-1} \mathbf{p} \\ &= (1-d)N + d\mathbf{c}^T \mathbf{D}_c^{-1} \mathbf{p} \\ &= (1-d)N + d\mathbf{1}^T \mathbf{p} \end{aligned} \quad (14.45)$$

therefore  $\mathbf{1}^T \mathbf{p} = N$ .

(b) (Program)

**Ex. 14.23**

(a) Since  $\log(\cdot)$  is concave, according to Jensen's inequality

$$\begin{aligned} \sum_{k=1}^r c_k \log(y_k/c_k) &= \frac{1}{\sum_{k=1}^r c_k} \sum_{k=1}^r c_k \log(y_k/c_k) \\ &\leq \log\left(\frac{\sum_{k=1}^r y_k}{\sum_{k=1}^r c_k}\right) \\ &= \log\left(\sum_{k=1}^r y_k\right) \end{aligned} \quad (14.46)$$

where equality holds iff  $c_k = 1/r$ .

(b)

$$g(\mathbf{W}, \mathbf{H} | \mathbf{W}^s, \mathbf{H}^s) = \sum_{i=1}^N \sum_{j=1}^p \sum_{k=1}^r x_{ij} \frac{a_{ikj}^s}{b_{ij}^s} \log \left( \frac{b_{ij}^s}{a_{ikj}^s} w_{ik} h_{kj} \right) - \sum_{i=1}^N \sum_{j=1}^p \sum_{k=1}^r w_{ik} h_{kj} \quad (14.47)$$

minorizes  $L(\mathbf{W}, \mathbf{H})$

(c) The stationary conditions are

$$\frac{\partial g}{\partial w_{ik}} = \sum_{j=1}^p x_{ij} \frac{a_{ikj}^s}{b_{ij}^s} - \sum_{j=1}^p h_{kj} = 0 \quad (14.48)$$

$$\frac{\partial g}{\partial h_{kj}} = \sum_{i=1}^N x_{ij} \frac{a_{ikj}^s}{b_{ij}^s} - \sum_{i=1}^N w_{ik} = 0 \quad (14.49)$$

which are equivalent to

$$w_{ik} = \frac{\sum_{j=1}^p x_{ij} a_{ikj}^s / b_{ij}^s}{\sum_{j=1}^p h_{kj}} \quad (14.50)$$

$$h_{kj} = \frac{\sum_{i=1}^N x_{ij} a_{ikj}^s / b_{ij}^s}{\sum_{i=1}^N w_{ik}} \quad (14.51)$$

which are exactly the updating steps (14.74).

**Ex. 14.24**(a) When  $r = 1$ , we have  $a_{ikj}^s / b_{ij}^s = 1$ , therefore the updating steps are simplified as

$$w_i \leftarrow \frac{\sum_{j=1}^p x_{ij}}{\sum_{j=1}^p h_j}, \quad h_j \leftarrow \frac{\sum_{i=1}^N x_{ij}}{\sum_{i=1}^N w_i} \quad (14.52)$$

(b) From Eq. (14.52), for every two steps, the updating becomes

$$h_k \leftarrow \frac{\sum_{i=1}^N x_{ik}}{\sum_{i=1}^N w_i} \leftarrow \frac{\sum_{i=1}^N x_{ik}}{\sum_{i=1}^N \sum_{j=1}^p x_{ij}} \sum_{j=1}^p h_j \quad (14.53)$$

$$w_l \leftarrow \frac{\sum_{j=1}^p x_{lj}}{\sum_{j=1}^p h_j} \leftarrow \frac{\sum_{j=1}^p x_{lj}}{\sum_{i=1}^N \sum_{j=1}^p x_{ij}} \sum_{i=1}^N w_i \quad (14.54)$$

It is easy to see that throughout the updating,  $\sum_{j=1}^p h_j$  and  $\sum_{i=1}^N w_i$  remains constant, thus  $h_k$  and  $w_l$  remain constant. Consequently, the iteration is completely stationary. By enforcing  $\sum_{j=1}^p h_j \sum_{i=1}^N w_i = 1$ , the iteration has the explicit form as (14.122) for any  $c$ .

**Ex. 14.25 (Program)**





# Chapter 15

## Random Forests

### Ex. 15.1

Assuming  $X_b$ ,  $b = 1, \dots, B$  are i.i.d with mean  $\bar{x}$  and variance  $\sigma^2$ . An average of these  $B$  variables are

$$X_B = \frac{1}{B} \sum_{b=1}^B X_b \quad (15.1)$$

therefore

$$\mathbb{E}[X_B] = \bar{x} \quad (15.2)$$

$$\begin{aligned} \mathbb{E}[X_B^2] &= \frac{1}{B^2} \mathbb{E} \left[ \sum_{b=1}^B X_b^2 + \sum_{b=1}^B \sum_{c \neq b}^B X_b X_c \right] \\ &= \frac{1}{B} [\sigma^2 + \bar{x}^2] + \frac{B-1}{B} [\rho \sigma^2 + \bar{x}^2] \end{aligned} \quad (15.3)$$

Therefore

$$\begin{aligned} \text{var}(X_B) &= \mathbb{E}[X_B^2] - \mathbb{E}[X_B]^2 \\ &= \rho \sigma^2 + \frac{1-\rho}{B} \sigma^2 \end{aligned} \quad (15.4)$$

### Ex. 15.2

The  $N$ -fold CV error estimate

$$\frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-i}(x_i)) \quad (15.5)$$

where  $\hat{f}^{-i}()$  denotes the model trained without using  $(y_i, x_i)$ . On the other hand, the OOB error estimate is

$$\frac{1}{N} \sum_{i=1}^N L(y_i, \frac{1}{\tilde{B}_i} \sum_{b \in \tilde{B}_i} \tilde{f}_b(x_i)) \quad (15.6)$$

where  $\tilde{B}_i = |\tilde{\mathcal{B}}_i|$  and  $\tilde{\mathcal{B}}_i$  represent the bootstrap sample sets that does not include  $(y_i, x_i)$ ,

and  $\tilde{f}_b$  denotes the model trained using the  $b$ -th bootstrap sample set.

When  $B \rightarrow \infty$ , each  $\tilde{B}_i \rightarrow \infty$ , the OOB prediction  $1/\tilde{B}_i \sum_{b \in \tilde{B}_i} \tilde{f}_b(x_i)$  is just the non-parametric bootstrap version of  $\hat{f}^{-i}(x_i)$  that is consistent (under some conditions). Therefore the OOB error estimate is asymptotically the same as the  $N$ -fold CV error estimate.

### Ex. 15.3

(Note:  $\sum_{j=1}^J X_j$  follows Irwin-Hall distribution, a spline of degree of  $J - 1$  over knots  $0, 1, \dots, J$ ).

The probability is a function defined over a  $J$  dimensional unit-cube in the  $J$  dimensional space  $\{x_1, x_2, \dots, x_J\}$  separated by the plane  $\sum_{j=1}^J x_j = J/2$ . On one side of the plane the probability  $Pr(Y = 1|X = x) = q$  and on the other side  $Pr(Y = 1|X = x) = 1 - q$ .

Consequently, the Bayesian error rate is

$$\begin{aligned} P_E &= P\left(\sum_{j=1}^J X_j > J/2\right) P(Y = 0|X) + P\left(\sum_{j=1}^J X_j < J/2\right) P(Y = 1|X) \\ &= \frac{1}{2}q + \frac{1}{2}q = q \end{aligned} \quad (15.7)$$

### Ex. 15.4

$$\bar{x}_1^* = \frac{1}{N} \sum_{i=1}^N x_{s_i}, \quad \bar{x}_2^* = \frac{1}{N} \sum_{i=1}^N x_{r_i} \quad (15.8)$$

where  $s_i$  and  $r_i$  are i.i.d uniformly distributed over  $\{1, \dots, N\}$ . Therefore we have

$$\mathbb{E}[\bar{x}_1^*] = \mu \quad (15.9)$$

$$\begin{aligned} \text{var}(\bar{x}_1^*) &= \mathbb{E}[(\bar{x}_1^*)^2] - \mathbb{E}[\bar{x}_1^*]^2 \\ &= \frac{1}{N^2} \sum_{i=1}^N \mathbb{E}[x_{s_i}^2] + \frac{1}{N^2} \sum_{i=1}^N \sum_{j \neq i}^N \mathbb{E}[x_{s_i} x_{s_j}] - \mu^2 \end{aligned} \quad (15.10)$$

Since

$$\mathbb{E}[x_{s_i} x_{s_j}] = P(s_i = s_j) \mathbb{E}[x_{s_i} x_{s_j}] + P(s_i \neq s_j) \mathbb{E}[x_{s_i} x_{s_j}] = \frac{1}{N}(\mu^2 + \sigma^2) + \frac{N-1}{N} \mu^2 \quad (15.11)$$

we have

$$\text{var}(\bar{x}_1^*) = \text{var}(\bar{x}_2^*) = \frac{2N-1}{N^2} \sigma^2. \quad (15.12)$$

On the other hand, we have

$$\begin{aligned}
\text{cov}(\bar{x}_1^*, \bar{x}_2^*) &= \mathbb{E}[\bar{x}_1^* \bar{x}_2^*] - \mathbb{E}[\bar{x}_1^*] \mathbb{E}[\bar{x}_2^*] \\
&= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[x_{s_i} x_{r_j}] - \mu^2 \\
&= \frac{1}{N} (\mu^2 + \sigma^2) + \frac{N-1}{N} \mu^2 - \mu^2 \\
&= \frac{\sigma^2}{N}
\end{aligned} \tag{15.13}$$

Consequently,

$$\text{corr}(\bar{x}_1^*, \bar{x}_2^*) = \frac{\text{cov}(\bar{x}_1^*, \bar{x}_2^*)}{\sqrt{\text{var}(\bar{x}_1^*) \text{var}(\bar{x}_2^*)}} = \frac{N}{2N-1} \tag{15.14}$$

**Ex. 15.5**

$$\begin{aligned}
&\text{var}_{\Theta, \mathbf{Z}}(T(X; \Theta(\mathbf{Z}))) \\
&= \mathbb{E}_{\mathbf{Z}} [\mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))^2]] - \mathbb{E}_{\mathbf{Z}} [\mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))]]^2 \\
&= \mathbb{E}_{\mathbf{Z}} [\text{var}_{\Theta|\mathbf{Z}} (T(X; \Theta(\mathbf{Z}))) + \mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))^2] - \mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))]]^2 \\
&= \mathbb{E}_{\mathbf{Z}} [\text{var}_{\Theta|\mathbf{Z}} (T(X; \Theta(\mathbf{Z}))) + \mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))^2] - \mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))]]^2 \\
&= \mathbb{E}_{\mathbf{Z}} [\text{var}_{\Theta|\mathbf{Z}} (T(X; \Theta(\mathbf{Z}))) + \text{var}_{\mathbf{Z}} (\mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))])]
\end{aligned} \tag{15.15}$$

On the other hand,

$$\begin{aligned}
&\text{cov}_{\Theta, \mathbf{Z}}(T_1(X; \Theta(\mathbf{Z})), T_2(X; \Theta(\mathbf{Z}))) \\
&= \mathbb{E}_{\Theta, \mathbf{Z}} [T_1(X; \Theta(\mathbf{Z})) T_2(X; \Theta(\mathbf{Z}))] - \mathbb{E}_{\Theta, \mathbf{Z}} [T_1(X; \Theta(\mathbf{Z}))] \mathbb{E}_{\Theta, \mathbf{Z}} [T_2(X; \Theta(\mathbf{Z}))] \\
&= \mathbb{E}_{\mathbf{Z}} [\mathbb{E}_{\Theta|\mathbf{Z}} [T_1(X; \Theta(\mathbf{Z})) T_2(X; \Theta(\mathbf{Z}))]] - \mathbb{E}_{\mathbf{Z}} [\mathbb{E}_{\Theta|\mathbf{Z}} [T_1(X; \Theta(\mathbf{Z}))]]^2
\end{aligned} \tag{15.16}$$

Since  $T_1$  and  $T_2$  conditioned on  $\mathbf{Z}$  are independent w.r.t  $\Theta$ , we have

$$\begin{aligned}
&\text{cov}_{\Theta, \mathbf{Z}}(T_1(X; \Theta(\mathbf{Z})), T_2(X; \Theta(\mathbf{Z}))) \\
&= \mathbb{E}_{\mathbf{Z}} [\mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))^2]] - \mathbb{E}_{\mathbf{Z}} [\mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))]]^2 \\
&= \text{var}_{\mathbf{Z}} (\mathbb{E}_{\Theta|\mathbf{Z}} [T(X; \Theta(\mathbf{Z}))])
\end{aligned} \tag{15.17}$$

Consequently (15.12) is proved.

**Ex. 15.6 (Program)**

**Ex. 15.7**

$$RSS = \frac{1}{N} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \quad (15.18)$$

On the other hand,

$$RSS_j^* = \frac{1}{N} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\beta}_j(\mathbf{x}_j - \mathbf{x}_j^*)\|^2 \quad (15.19)$$

where  $\mathbf{x}_j$  represents the  $j$ -th column of  $\mathbf{X}$  and  $\mathbf{x}_j^*$  represents its permuted version. Consequently,

$$\begin{aligned} \mathbb{E}[RSS_j^*] &= RSS + \frac{1}{N} \mathbb{E}[\|\hat{\beta}_j(\mathbf{x}_j - \mathbf{x}_j^*)\|^2] \\ &= RSS + \frac{2}{N} \mathbb{E}[\|\hat{\beta}_j \mathbf{x}_j\|^2] \\ &= RSS + 2\|\hat{\beta}_j\|^2 \end{aligned} \quad (15.20)$$

assuming that the  $j$ -th column of  $\mathbf{X}$  has been standardized so that  $\mathbb{E}\|\mathbf{x}_j\|/N = 1$ .

# Chapter 16

## Ensemble Learning

### Ex. 16.1

For each block of 20, generate independent samples  $v_{0,i}, v_{1,i}, \dots, v_{20,i} \sim \mathcal{N}(0, 1)$ . Then generate the  $i$ -th sample of the 20 variables as  $x_{1,i} = \sqrt{0.95}v_{0,i} + \sqrt{0.05}v_{1,i}, \dots, x_{20,i} = \sqrt{0.95}v_{0,i} + \sqrt{0.05}v_{20,i}$ .

### Ex. 16.2

$$\begin{aligned}\Lambda(t) &= \int_0^t |\dot{\alpha}(t)|_1 dt \\ &\geq \left| \int_0^t \dot{\alpha}(t) dt \right|_1 \\ &= |\alpha(t)|_1\end{aligned}\tag{16.1}$$

Equality holds iff  $\dot{\alpha}(t) \geq 0, \forall t$ , or  $\dot{\alpha}(t) \leq 0, \forall t$ , i.e.  $\alpha(t)$  is monotonic.

### Ex. 16.3

The regressio problem is

$$\min \sum_{i=1}^N [y_i - \beta_1 I_1(x_i) - \beta_4 I_4(x_i) - \beta_6 I_6(x_i) - \beta_7 I_7(x_i)]^2 \tag{16.2}$$

Since  $R_1, R_4, R_6, R_7$  is a partital of the sample space  $\mathcal{X}$ , the above problem can be rewritten as

$$\begin{aligned}&\min \sum_{x_i \in R_1} [y_i - \beta_1]^2 + \sum_{x_i \in R_4} [y_i - \beta_4]^2 + \sum_{x_i \in R_6} [y_i - \beta_6]^2 + \sum_{x_i \in R_7} [y_i - \beta_7]^2 \\ &\leftrightarrow \min \sum_{x_i \in R_1} [y_i - \beta_1]^2 + \min \sum_{x_i \in R_4} [y_i - \beta_4]^2 + \min \sum_{x_i \in R_6} [y_i - \beta_6]^2 + \min \sum_{x_i \in R_7} [y_i - \beta_7]^2\end{aligned}\tag{16.3}$$

i.e. can be decomposed into 4 independent regression problems, of which the solutions are  $\hat{\beta}_1 = \text{mean } y_i | x_i \in R_1$ ,  $\hat{\beta}_4 = \text{mean } y_i | x_i \in R_4$ ,  $\hat{\beta}_6 = \text{mean } y_i | x_i \in R_6$  and  $\hat{\beta}_7 = \text{mean } y_i | x_i \in R_7$ , which are exactly the same as a regression tree.

The 2-class logistic regression problem can be formulated as follows, by encoding  $y_i \in$

$\{0, 1\}$ ,

$$\max \sum_{i=1}^N y_i \boldsymbol{\beta}^T \mathbf{I}(x_i) - \log(1 + \exp(\boldsymbol{\beta}^T \mathbf{I}(x_i))) \quad (16.4)$$

where  $\boldsymbol{\beta} = [\beta_1, \beta_4, \beta_6, \beta_7]^T$ ,  $\mathbf{I}(x_i) = [I_1(x_i), I_4(x_i), I_6(x_i), I_7(x_i)]^T$ . Again this problem can be decomposed into

$$\begin{aligned} & \max \sum_{x_i \in R_1} y_i \beta_1 x_i - \log(1 + \exp(\beta_1 x_i)) + \max \sum_{x_i \in R_4} y_i \beta_4 x_i - \log(1 + \exp(\beta_4 x_i)) \\ & + \max \sum_{x_i \in R_6} y_i \beta_6 x_i - \log(1 + \exp(\beta_6 x_i)) + \max \sum_{x_i \in R_7} y_i \beta_7 x_i - \log(1 + \exp(\beta_7 x_i)) \end{aligned} \quad (16.5)$$

of which the solution is

$$\frac{\exp(\beta_j)}{1 + \exp(\beta_j)} = \frac{\sum_{x_i \in R_j} y_i}{\sum_{x_i \in R_j} 1} \quad (16.6)$$

where  $j = 1, 4, 6, 7$ . This result is equivalent to a classification tree.

# Chapter 17

## Undirected Graphical Models

### Ex. 17.1

A not complete list of conditional independence relations are as follows

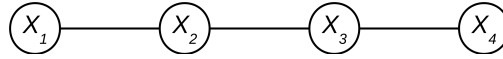
$$\begin{aligned} X_1 \perp X_3 | X_4, \quad X_1 \perp X_5 | X_6, \quad X_2 \perp X_3 | X_4, \quad X_2 \perp X_4 | X_1, \\ X_2 \perp X_5 | X_6, \quad X_3 \perp X_5 | X_6, \quad X_4 \perp X_5 | X_1 \end{aligned} \quad (17.1)$$

The maximal cliques are

$$\{X_1, X_4\}, \{X_3, X_4\}, \{X_5, X_6\}, \{X_1, X_2, X_6\} \quad (17.2)$$

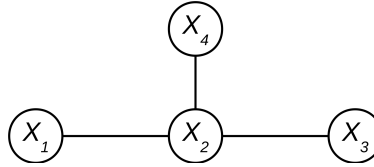
### Ex. 17.2

(a)



(b) Same as (a).

(c)



### Ex. 17.3

(a) Since

$$\begin{bmatrix} \Theta_{aa} & \Theta_{ab} \\ \Theta_{ba} & \Theta_{bb} \end{bmatrix} \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} = \mathbf{I} \quad (17.3)$$

therefore we have

$$\Theta_{aa} \Sigma_{aa} + \Theta_{ab} \Sigma_{ba} = \mathbf{I} \quad (17.4a)$$

$$\Theta_{aa} \Sigma_{ab} + \Theta_{ab} \Sigma_{bb} = \mathbf{0} \quad (17.4b)$$

(17.4a) - (17.4b)  $\Sigma_{bb}^{-1} \Sigma_{ba}$ , we have

$$\Theta_{aa} (\Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}) = \mathbf{I} \quad (17.5)$$

consequently  $\Theta_{aa}^{-1} = \Sigma_{a,b}$ .

(b) Assume that  $\Sigma_{12} = 0$ , i.e.  $\Sigma_{aa}$  is diagonal. From (a)  $\Sigma_{a,b}$  is also diagonal, which suggests that  $\text{cov}(X_1, X_2 | \text{rest}) = 0$ .

(c) Since  $r_{jk} = \Theta_{jk} / \sqrt{\Theta_{jj} \Theta_{kk}}$ , for  $j = k$ , we have  $r_{jk} = 1$ . For  $j \neq k$ , denote  $X_a = (X_k, X_k)$ , then

$$\begin{bmatrix} \Theta_{jj} & \Theta_{jk} \\ \Theta_{kj} & \Theta_{kk} \end{bmatrix} = \Theta_{aa} = \begin{bmatrix} \Sigma_{jj|\text{rest}} & \Sigma_{jk|\text{rest}} \\ \Sigma_{kj|\text{rest}} & \Sigma_{kk|\text{rest}} \end{bmatrix}^{-1} \quad (17.6)$$

therefore

$$\Theta_{jj} \Sigma_{jj|\text{rest}} + \Theta_{jk} \Sigma_{kj|\text{rest}} = 1 \quad (17.7a)$$

$$\Theta_{jj} \Sigma_{jk|\text{rest}} + \Theta_{jk} \Sigma_{kk|\text{rest}} = 0 \quad (17.7b)$$

$$\Theta_{kj} \Sigma_{jk|\text{rest}} + \Theta_{kk} \Sigma_{kk|\text{rest}} = 1 \quad (17.7c)$$

consequently, we have

$$\Theta_{jj} \Sigma_{jj|\text{rest}} = \Theta_{kk} \Sigma_{kk|\text{rest}} \quad (17.8a)$$

$$\Theta_{jk} \Sigma_{kk|\text{rest}} = \Theta_{jj} \Sigma_{jk|\text{rest}} \quad (17.8b)$$

As a result,  $r_{jk} = -\Sigma_{jk|\text{rest}} / \sqrt{\Sigma_{jj|\text{rest}} \Sigma_{kk|\text{rest}}} = -\rho_{jk|\text{rest}}$ .

#### Ex. 17.4

Since

$$f(X_1 | X_2, \text{rest}) = \frac{f(X_1, X_2 | \text{rest})}{f(X_2 | \text{rest})} = f(X_1 | \text{rest}) \quad (17.9)$$

we have

$$f(X_1, X_2 | \text{rest}) = f(X_1 | \text{rest}) f(X_2 | \text{rest}) \quad (17.10)$$

i.e.  $X_1 \perp X_2 | \text{rest}$ .

#### Ex. 17.5

Since there is no missing edges

$$l_C(\Theta) = l(\Theta) = \log |\Theta| - \text{tr}(\mathbf{S}\Theta) \quad (17.11)$$



The gradient equation for maximizing  $l_C(\Theta)$  becomes  $\Theta^{-1} - \mathbf{S} = 0$ , which suggests

$$\mathbf{S}\Theta = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{bmatrix} \begin{bmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \quad (17.12)$$

therefore

$$\mathbf{S}_{11}\theta_{12} + \theta_{22}\mathbf{s}_{12} = \mathbf{0} \quad (17.13)$$

Since  $\beta = -\theta_{12}/\theta_{22}$  as in (17.9), we have  $\mathbf{S}_{11}\beta - \mathbf{s}_{12} = \mathbf{0}$ .

**Ex. 17.6**

Since

$$\begin{bmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & w_{22} \end{bmatrix} \begin{bmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \quad (17.14)$$

we have

$$\mathbf{W}_{11}\theta_{12} + \theta_{22}\mathbf{w}_{12} = \mathbf{0} \quad (17.15a)$$

$$\mathbf{w}_{12}^T\theta_{12} + \theta_{22}w_{22} = 1 \quad (17.15b)$$

therefore

$$\theta_{12} = -\mathbf{W}_{11}^{-1}\mathbf{w}_{12}\theta_{22} = -\hat{\beta}\theta_{22} \quad (17.16a)$$

$$\theta_{22} = \frac{1 - \mathbf{w}_{12}^T\theta_{12}}{w_{22}} \quad (17.16b)$$

Combining these 2 equations, we have

$$\theta_{22} = \frac{1}{w_{22} - \mathbf{w}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{w}_{12}} \quad (17.17)$$

**Ex. 17.7 (Program)**

**Ex. 17.8 (Program)**

**Ex. 17.9**

(a) *E-step*: The missing data (latent variables), given the current estimation  $\hat{\mu}, \hat{\Sigma}$  and the

observed data, follow Gaussian distribution as

$$\mathbf{x}_{i,m_i} \sim \mathcal{N} \left( \hat{\boldsymbol{\mu}}_{m_i} + \hat{\boldsymbol{\Sigma}}_{m_i,o_i} \hat{\boldsymbol{\Sigma}}_{o_i,o_i}^{-1} (\mathbf{x}_{i,o_i} - \hat{\boldsymbol{\mu}}_{o_i}), \hat{\boldsymbol{\Sigma}}_{m_i,m_i} - \hat{\boldsymbol{\Sigma}}_{m_i,o_i} \hat{\boldsymbol{\Sigma}}_{o_i,o_i}^{-1} \hat{\boldsymbol{\Sigma}}_{o_i,m_i} \right) \quad (17.18)$$

(here  $\mathbf{x}_{i,m_i}$ ,  $\mathbf{x}_{i,o_i}$  and  $\hat{\boldsymbol{\mu}}_{m_i}$ ,  $\hat{\boldsymbol{\mu}}_{o_i}$ ) are written as column vectors.) Therefore, the expectation of the log-likelihood of the data over the above conditional distribution of  $\mathbf{x}_{i,m_i}$  is

$$\begin{aligned} \mathbb{E} [l(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}_o, \mathbf{X}_m)] &= C + N \log |\boldsymbol{\Sigma}^{-1}| - \text{tr} \left( \mathbb{E} [(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T) \boldsymbol{\Theta} (\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T)^T] \right) \\ &= C + N \log |\boldsymbol{\Sigma}^{-1}| - \sum_{i=1}^N \sum_{j=1}^p \sum_{k=1}^p \mathbb{E} [(x_{ij} - \mu_j) \Theta_{jk} (x_{ik} - \mu_k)] \end{aligned} \quad (17.19)$$

where  $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$

*M-step:* To maximize the log likelihood, we have

$$\frac{\partial \mathbb{E} [l(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}_o, \mathbf{X}_m)]}{\partial \boldsymbol{\mu}} = \mathbf{1}^T \mathbb{E} [\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T] \boldsymbol{\Theta} = \mathbf{0} \quad (17.20)$$

thus  $\hat{\boldsymbol{\mu}} = \mathbb{E} [\mathbf{X}^T \mathbf{1}] / N = \hat{\mathbf{X}}^T \mathbf{1} / N$ , where  $\hat{\mathbf{X}}$  represents the  $N$ -by- $p$  predictor matrix with the missing entries replaced by the imputed ones, namely the mean of  $\mathbf{x}_{i,m_i}$ . Also we have

$$\frac{\partial \mathbb{E} [l(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}_o, \mathbf{X}_m)]}{\partial \boldsymbol{\Theta}} = N \boldsymbol{\Sigma} - \mathbb{E} [(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T)^T (\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T)] = \mathbf{0} \quad (17.21)$$

therefore, the ML estimation of  $\boldsymbol{\Sigma}$  is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \mathbb{E} [(\mathbf{X} - \mathbf{1}\hat{\boldsymbol{\mu}}^T)^T (\mathbf{X} - \mathbf{1}\hat{\boldsymbol{\mu}}^T)] \quad (17.22)$$

Denote  $E_{ijk} = \mathbb{E} [(x_{ij} - \hat{\mu}_j)(x_{ik} - \hat{\mu}_k)]$ , since

$$E_{ijk} = (\mathbb{E}[x_{ij}] - \hat{\mu}_j)(\mathbb{E}[x_{ik}] - \hat{\mu}_k) + \text{cov}(x_{ij}x_{ik}) \quad (17.23)$$

in which

$$\mathbb{E}[x_{ij}] = \hat{x}_{ij}, \quad \mathbb{E}[x_{ik}] = \hat{x}_{ik} \quad (17.24)$$

whether  $j, k \in m_i$  or not, and

$$\text{cov}(x_{ij}x_{ik}) = \begin{cases} 0 & \text{if } j \in o_i \text{ or } k \in o_i \\ \hat{\Sigma}_{jk} & \text{otherwise} \end{cases} \quad (17.25)$$

therefore (17.44) is proved, in which the correction term  $c_{i,jj'}$  corresponds to the non-zero covariance  $\text{cov}(x_{ij}x_{ij'})$  when both  $j$  and  $j'$  are imputed for  $x_i$ .

(b) (Program)

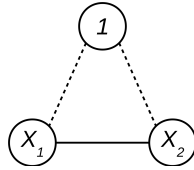
(c) (Program)

**Ex. 17.10**

An absence of the constant node  $X_0 \equiv 1$  ill lead to the following ambiguity

$$p(X_1 = 0, X_2 = 0) = p(X_1 = 1, X_2 = 0) = p(X_1 = 0, X_2 = 1) \quad (17.26)$$

only by including  $X_0 \equiv 1$  the 4 possible values can be uniquely defined.



**Ex. 17.11**

$$\begin{aligned} p(X_j = 1 | X_{\text{rest}} = x_{\text{rest}}, \Theta) &= \frac{p(X_j = 1, X_{\text{rest}} = x_{\text{rest}} | \Theta)}{p(X_{\text{rest}} = x_{\text{rest}} | \Theta)} \\ &= \frac{p(X_j = 1, X_{\text{rest}} = x_{\text{rest}} | \Theta)}{p(X_j = 1, X_{\text{rest}} = x_{\text{rest}} | \Theta) + p(X_j = 0, X_{\text{rest}} = x_{\text{rest}} | \Theta)} \\ &= \frac{C \exp \left( \sum_{k: (j,k) \in E} \theta_{jk} x_k \right)}{C \exp \left( \sum_{k: (j,k) \in E} \theta_{jk} x_k \right) + C} \\ &= \frac{1}{1 + \exp \left( - \sum_{k: (j,k) \in E} \theta_{jk} x_k \right)} \end{aligned} \quad (17.27)$$

where  $C$  is a constant given the value of the rest nodes in the graph. Now this probability has the logistic form as (17.30) (considering a constant node  $X_0 = 1$ ).

**Ex. 17.12**

???

# Chapter 18

## High-Dimensional Problems

**Ex. 18.1**

???

**Ex. 18.2**

To minimize the lasso-style objective function (denoted as  $L$ ), we have

$$\frac{\partial L}{\partial \mu_j} = \sum_{k=1}^K \sum_{i \in C_k} -\frac{x_{ij} - \mu_j - \mu_{jk}}{s_j^2} = 0 \quad (18.1a)$$

$$\frac{\partial L}{\partial \mu_{jk}} = \sum_{i \in C_k} -\frac{x_{ij} - \mu_j - \mu_{jk}}{s_j^2} + \lambda \sqrt{N_k} \frac{\text{sign}(\mu_{jk})}{s_j} = 0 \quad (18.1b)$$

therefore

$$\mu_{jk} = \bar{x}_{jk} - \mu_j - \frac{\lambda s_j \text{sign}(\mu_{jk})}{\sqrt{N_k}} \quad (18.2a)$$

$$\mu_j = \bar{x}_j - \frac{1}{N} \sum_{k=1}^K N_k \mu_{jk} \quad (18.2b)$$

where  $\bar{x}_{jk} = \sum_{i \in C_k} x_{ij}/N_k$  and  $\bar{x}_j = \sum_{i=1}^N x_{ij}/N$ . (Here we note that the condition  $\sum_{k=1}^K \mu_{jk}$  should have been  $\sum_{k=1}^K N_k \mu_{jk} = 0$ .) Consequently

$$\mu_{jk} = \bar{x}_{jk} - \bar{x}_j - \frac{\lambda s_j \text{sign}(\mu_{jk})}{\sqrt{N_k}}, \quad (18.3)$$

therefore

$$\begin{aligned} d'_{kj} &= \frac{\sqrt{N_k} \mu_{jk}}{s_j} \\ &= \frac{\bar{x}_{jk} - \bar{x}_j}{m_k(s_j + s_0)} - \lambda \text{sign}(d'_{kj}) \\ &= d_{kj} - \lambda \text{sign}(d'_{kj}) \end{aligned} \quad (18.4)$$

where  $s_0 = 0$  and  $m_k = 1/\sqrt{N_k}$ . As a result, we have  $d'_{kj} = \text{sign}(d_{kj})(|d_{kj} - \Delta|)_+$ , where  $\Delta = \lambda$ .

**Ex. 18.3**

The penalized log-likelihood objective function in (18.11) is explicitly written as

$$l_P(\beta_0, \mathbf{B}; \mathbf{X}, \mathbf{g}) = \sum_{i=1}^N \left[ \beta_{k_i 0} + \mathbf{x}_i^T \beta_{k_i} - \log \sum_{l=1}^K \exp(\beta_{l0} + \mathbf{x}_i^T \beta_l) \right] - \frac{\lambda}{2} \sum_{k=1}^K \|\beta_k\|^2 \quad (18.5)$$

The necessary condition to maximize the penalized log-likelihood is

$$\frac{\partial l_P(\beta_0, \mathbf{B}; \mathbf{X}, \mathbf{g})}{\partial \beta_k} = \sum_{i: g_i=k} \mathbf{x}_i^T - \sum_{i=1}^N \Pr(k|\mathbf{x}_i) \mathbf{x}_i^T - \lambda \beta_k^T = \mathbf{0} \quad (18.6)$$

for  $k = 1, \dots, K$ . Consequently,

$$\begin{aligned} \sum_{k=1}^K \frac{\partial l_P(\beta_0, \mathbf{B}; \mathbf{X}, \mathbf{g})}{\partial \beta_k} &= \mathbf{0} \\ &= \sum_{i=1}^N \mathbf{x}_i^T - \sum_{i=1}^N \left( \sum_{k=1}^K \Pr(k|\mathbf{x}_i) \right) \mathbf{x}_i^T - \lambda \sum_{k=1}^K \beta_k^T \\ &= -\lambda \sum_{k=1}^K \beta_k^T \end{aligned} \quad (18.7)$$

therefore  $\sum_{k=1}^K \beta_{kj} = 0$ ,  $j = 1, \dots, p$ .  $\beta_{k0}$  should all be set to 0.

**Ex. 18.4**

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{V} \mathbf{R}^T \mathbf{R} \mathbf{V}^T + \lambda \mathbf{I})^{-1} \mathbf{V} \mathbf{R}^T \mathbf{y} \\ &= \left[ \lambda^{-1} \mathbf{I} - \lambda^{-1} \mathbf{V} ((\mathbf{R}^T \mathbf{R})^{-1} + \lambda^{-1} \mathbf{I})^{-1} \mathbf{V}^T \lambda^{-1} \right] \mathbf{V} \mathbf{R}^T \mathbf{y} \\ &= \lambda^{-1} \mathbf{V} \left[ \mathbf{I} - \lambda^{-1} ((\mathbf{R}^T \mathbf{R})^{-1} + \lambda^{-1} \mathbf{I})^{-1} \right] \mathbf{R}^T \mathbf{y} \\ &= \lambda^{-1} \mathbf{V} \left[ \mathbf{I} - \lambda^{-1} (\lambda \mathbf{I} - \lambda^2 (\mathbf{R}^T \mathbf{R} + \lambda \mathbf{I})^{-1}) \right] \mathbf{R}^T \mathbf{y} \\ &= \mathbf{V} (\mathbf{R}^T \mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R}^T \mathbf{y} \end{aligned} \quad (18.8)$$

Through this proof we have made use of the Woodbury matrix identity twice. (What has it to do with the hint?)

**Ex. 18.5**

$\forall \beta, \beta_0$ , let  $\theta_0 = \beta_0$  and decompose  $\beta$  as  $\beta = \mathbf{V} \theta + \mathbf{V}_\perp \theta_\perp$ , where  $\mathbf{V}_\perp$  is a set of orthonormal

vectors representing the complementary space of  $\mathbf{V}$ . Consequently

$$\mathbf{X}\boldsymbol{\beta} + \beta_0\mathbf{1} = \mathbf{R}\boldsymbol{\theta} + \theta_0\mathbf{1} \quad (18.9a)$$

$$\boldsymbol{\beta}^T \boldsymbol{\beta} = \boldsymbol{\theta}^T \boldsymbol{\theta} + \boldsymbol{\theta}_\perp^T \boldsymbol{\theta}_\perp \geq \boldsymbol{\theta}^T \boldsymbol{\theta} \quad (18.9b)$$

This suggests that a solution to (18.16) must have  $\hat{\boldsymbol{\theta}}_\perp = \mathbf{0}$ . Consequently, the solution to (18.16) can be constructed from the solution to (18.17) by  $\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\theta}}$ ,  $\hat{\beta}_0 = \hat{\theta}_0$ .

**Ex. 18.6**

(Not Section 4.14 but equation (4.14).) Write the regularized discriminant analysis (RDA) into the ASR form in (12.57):

$$ASR = \frac{1}{N} \sum_{l=1}^L \left[ \sum_{i=1}^N (\theta_l(g_i) - \beta_{l0} - \mathbf{x}_i^T \boldsymbol{\beta}_l)^2 + \frac{1-\gamma}{\gamma} \hat{\sigma}^2 \boldsymbol{\beta}_l^T \boldsymbol{\beta}_l \right] \quad (18.10)$$

which is now in the form as in Sec.18.3.5. By defining  $\beta_{l0} = u_{l0}$ ,  $\boldsymbol{\beta}_l = \mathbf{V}\mathbf{u}_l$ , we can solve a smaller problem for  $\hat{u}_{l0}$ ,  $\hat{\mathbf{u}}_l$ ,  $l = 1, \dots, L$  then map them back to get  $\hat{\beta}_{l0}$ ,  $\hat{\boldsymbol{\beta}}_l$ .

**Ex. 18.7**

(a) As  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ , we have  $\mathbf{R}^{-1}\mathbf{y} = \mathbf{V}^T \boldsymbol{\beta}$ . Since  $\mathbf{V}^T$  has rank  $N$ , this equation must have at least 1 solution denoted as  $\boldsymbol{\beta}_0$ . Consequently,

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \mathbf{V}_\perp \boldsymbol{\beta}_\perp \quad (18.11)$$

where  $\mathbf{V}_\perp$  (rank  $N - p$ ) represent the complementary space of  $\mathbf{V}$ , is a solution for arbitrary  $\boldsymbol{\beta}_\perp$ .

(b) Same as Ex. (18.4).

(c)

$$\mathbf{X}\hat{\boldsymbol{\beta}}_0 = \mathbf{U}\mathbf{D}\mathbf{V}^T \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T \mathbf{y} = \mathbf{y} \quad (18.12)$$

thus there is zero residual. Denote a solution as  $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\theta} + \mathbf{V}_\perp \boldsymbol{\theta}_\perp$ . Since  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} = \mathbf{R}\boldsymbol{\theta}$ , we have  $\boldsymbol{\theta} = \mathbf{R}^{-1}\mathbf{y}$ , therefore  $\boldsymbol{\beta}$  can be rewritten as

$$\boldsymbol{\beta} = \mathbf{V}\mathbf{R}^{-1}\mathbf{y} + \mathbf{V}_\perp \boldsymbol{\theta}_\perp \quad (18.13)$$

Since  $\|\boldsymbol{\beta}\|^2 = \|\mathbf{R}^{-1}\mathbf{y}\|^2 + \|\boldsymbol{\theta}_\perp\|^2 \leq \|\mathbf{R}^{-1}\mathbf{y}\|^2$ , in which equality holds iff  $\boldsymbol{\beta} = \mathbf{V}\mathbf{R}^{-1}\mathbf{y}$ . As a result, it is unique with the smallest Euclidean norm.

**Ex. 18.8**

(a) Decompose  $\mathbf{X} = \mathbf{R}\mathbf{V}^T$ . Then  $\mathbf{X}\boldsymbol{\beta}$  projects to  $\pm(1 - \alpha)$  for  $\boldsymbol{\beta} = \mathbf{V}\mathbf{R}^{-1}\mathbf{y} + \mathbf{V}_\perp \boldsymbol{\beta}_\perp$

(b) Since

$$\frac{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} = \frac{\pm(1 - \alpha)}{\hat{\beta}} \quad (18.14)$$

therefore the distance is  $2/\hat{\beta}$ .

(c) Same as Ex 18.7. Largest distance achieved by  $\hat{\boldsymbol{\beta}}_0$  which has the smallest Euclidean norm.

### Ex. 18.9

Apparently optimal separating hyperplane makes the widest margin by its definition. Specifically, (4.48) when  $p \gg N$  must have a solution  $\tilde{\boldsymbol{\beta}}$ , which means  $\forall i, j$  such that  $y_i = 1$ ,  $y_j = -1$ , we have

$$\mathbf{x}_i^T \tilde{\boldsymbol{\beta}} + \tilde{\beta}_0 \geq 1, \mathbf{x}_j^T \tilde{\boldsymbol{\beta}} + \tilde{\beta}_0 \leq 1 \quad (18.15)$$

Consequently the margin is at least  $2/\|\tilde{\boldsymbol{\beta}}\|$ . Also we must have  $\|\tilde{\boldsymbol{\beta}}\| \leq \|\hat{\boldsymbol{\beta}}_0\|$ , since the later is also valid for the constraints in (4.48). As a result, optimal separating hyperplane separates data by a wider margin then does the data piling direction.

### Ex. 18.10

Decompose  $\mathbf{X}$  into  $\mathbf{X} = \mathbf{R}\mathbf{V}^T$ . Then we can project  $\bar{\mathbf{x}}_1$  and  $\bar{\mathbf{x}}_{-1}$  onto  $\mathbf{V}$  as

$$\bar{\mathbf{r}}_1 = \mathbf{V}^T \bar{\mathbf{x}}_1, \bar{\mathbf{r}}_{-1} = \mathbf{V}^T \bar{\mathbf{x}}_{-1} \quad (18.16)$$

$$\bar{\mathbf{x}}_1 = \mathbf{V} \bar{\mathbf{r}}_1, \bar{\mathbf{x}}_{-1} = \mathbf{V} \bar{\mathbf{r}}_{-1} \quad (18.17)$$

and the within-class variance matrix for  $\mathbf{R}$  satisfies  $\mathbf{W} = \mathbf{V}\mathbf{W}_R\mathbf{V}^T$ . For any  $\mathbf{x}$ , we decompose it into  $\mathbf{x} = \mathbf{V}\mathbf{r} + \mathbf{V}_\perp \mathbf{r}_\perp$ . Then the discriminant function becomes

$$\begin{aligned} & \mathbf{x}^T (\mathbf{W} + \lambda \mathbf{I})^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_{-1}) \\ &= (\mathbf{r}^T \mathbf{V}^T + \mathbf{r}_\perp^T \mathbf{V}_\perp^T) (\mathbf{V}\mathbf{W}_R\mathbf{V}^T + \lambda \mathbf{I})^{-1} \mathbf{V} (\bar{\mathbf{r}}_1 - \bar{\mathbf{r}}_{-1}) \\ &= (\mathbf{r}^T \mathbf{V}^T + \mathbf{r}_\perp^T \mathbf{V}_\perp^T) \mathbf{V} (\mathbf{W}_R + \lambda \mathbf{I})^{-1} (\bar{\mathbf{r}}_1 - \bar{\mathbf{r}}_{-1}) \\ &= \mathbf{r}^T (\mathbf{W}_R + \lambda \mathbf{I})^{-1} (\bar{\mathbf{r}}_1 - \bar{\mathbf{r}}_{-1}) \end{aligned} \quad (18.18)$$

where the proof is similar to Ex. 18.4. Consequently, the discriminant function can be redefined on  $\mathbf{r}$ :

$$\delta_0(\mathbf{r}) = \lim_{\lambda \rightarrow 0} \delta(\mathbf{r}) = \mathbf{r}^T \mathbf{W}_R^{-1} (\bar{\mathbf{r}}_1 - \bar{\mathbf{r}}_{-1}) \quad (18.19)$$

On the other hand, for the solution  $\hat{\boldsymbol{\beta}}$  to the linear response regression to binary response

$\pm 1$ , we have  $\mathbf{x}^T \hat{\boldsymbol{\beta}} = \mathbf{r}^T \mathbf{R}^{-1} \mathbf{y}$ . According to Ex. 4.2 we have  $\mathbf{R}^{-1} \mathbf{y} \propto \mathbf{W}_R^{-1}(\bar{\mathbf{r}}_1 - \bar{\mathbf{r}}_{-1})$  (Assuming  $N_1 = N_2$ ). Consequently  $\delta_0(\mathbf{r})$  is equivalent to the projection onto the maximal data piling direction up to scaling.

**Ex. 18.11**

The optimal solution is characterized by (4.21)

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \mathbf{x}_i^T \left( y_i - \frac{\exp(\alpha + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{x}_i^T \boldsymbol{\beta})} \right) = \mathbf{0} \quad (18.20)$$

If  $\boldsymbol{\beta}_0$  is a solution, then  $\boldsymbol{\beta}_0 + \Delta \boldsymbol{\beta}$  where  $\mathbf{X} \Delta \boldsymbol{\beta} = \mathbf{0}$  is also a solution. Since  $p \gg N$ , there are infinitely many  $\Delta \boldsymbol{\beta}$ . Consequently,  $\boldsymbol{\beta}$  is undefined.

**Ex. 18.12**

$\mathbf{X} = \mathbf{R} \mathbf{V}^T$  implies that  $\mathbf{X}_B = \mathbf{R}_B \mathbf{V}^T$ , where  $\mathbf{R}_B$  corresponds to the same rows in  $\mathbf{R}$  as does the CV samples  $\mathbf{X}_B$  to  $\mathbf{X}$ . Consequently, we need to reduce  $\mathbf{X}$  to  $\mathbf{R}$  only once, and CV fitting can be done on subsets of rows of  $\mathbf{R}$ .

**Ex. 18.13**

Denote the logit function as  $\text{logit}(\mathbf{x}) = a_0 + \mathbf{x}^T \mathbf{a}$ , then the ridged logistic regression is in the form of

$$\min_{a_0, \mathbf{a}} \sum_{i=1}^N y_i \log \frac{\exp(a_0 + \mathbf{x}_i^T \mathbf{a})}{1 + \exp(a_0 + \mathbf{x}_i^T \mathbf{a})} + (1 - y_i) \log \frac{1}{1 + \exp(a_0 + \mathbf{x}_i^T \mathbf{a})} + \lambda \|\mathbf{a}\|^2 \quad (18.21)$$

Similar to Ex 18.7, denote  $\beta_0 = a_0$ ,  $\mathbf{a} = \mathbf{V} \boldsymbol{\beta}$ , this problem is equivalent to the ridged logistic regression in  $\mathbf{R}$  instead of  $\mathbf{X}$  where  $\mathbf{X} = \mathbf{R} \mathbf{V}^T$ :

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^N y_i \log \frac{\exp(\beta_0 + \mathbf{r}_i^T \boldsymbol{\beta})}{1 + \exp(\beta_0 + \mathbf{r}_i^T \boldsymbol{\beta})} + (1 - y_i) \log \frac{1}{1 + \exp(\beta_0 + \mathbf{r}_i^T \boldsymbol{\beta})} + \lambda \|\boldsymbol{\beta}\|^2 \quad (18.22)$$

Then the predictions are given by

$$\begin{aligned} \hat{f}_0 &= \hat{a}_0 + \mathbf{x}_0^T \hat{\mathbf{a}} \\ &= \hat{\beta}_0 + \mathbf{x}_0^T \mathbf{V} \hat{\boldsymbol{\beta}} \\ &= \hat{\beta}_0 + \mathbf{x}_0^T \mathbf{V} \mathbf{D} \mathbf{U}^T \mathbf{U} \mathbf{D}^{-1} \hat{\boldsymbol{\beta}} \\ &= \hat{\beta}_0 + \mathbf{k}_0^T \mathbf{U} \mathbf{D}^{-1} \hat{\boldsymbol{\beta}} \end{aligned} \quad (18.23)$$

therefore  $\hat{\mathbf{a}} = \mathbf{U} \mathbf{D}^{-1} \hat{\boldsymbol{\beta}}$ .

With the logit function in kernel space  $\text{logit}(\mathbf{x}) = h(\mathbf{x})$ ,  $h \in \mathcal{H}_K$ , the kernel ridged logistic



regression problem is

$$\min_{h \in \mathcal{H}_K} = \sum_{i=1}^N y_i \log \frac{\exp(h(\mathbf{x}_i))}{1 + \exp(h(\mathbf{x}_i))} + (1 - y_i) \log \frac{1}{1 + \exp(h(\mathbf{x}_i))} + \lambda \|h\|_{\mathcal{H}_K}^2 \quad (18.24)$$

According to (5.48), the solution must be in the form of

$$h(\mathbf{x}) = \sum_{i=1}^N \beta_i K(\mathbf{x}, \mathbf{x}_i) \quad (18.25)$$

therefore the ridged regression can be rewritten as

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^N y_i \log \frac{\exp(\mathbf{k}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{k}_i^T \boldsymbol{\beta})} + (1 - y_i) \log \frac{1}{1 + \exp(\mathbf{k}_i^T \boldsymbol{\beta})} + \lambda \boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta} \quad (18.26)$$

where  $\mathbf{k}_i$  is the  $i$ -th column of  $\mathbf{K}$ . Denote  $\mathbf{b} = \mathbf{D}\mathbf{U}^T \boldsymbol{\beta}$ , then the ridged regression is equivalent to

$$\min_{\mathbf{b}} \sum_{i=1}^N y_i \log \frac{\exp(\mathbf{r}_i^T \mathbf{b})}{1 + \exp(\mathbf{r}_i^T \mathbf{b})} + (1 - y_i) \log \frac{1}{1 + \exp(\mathbf{r}_i^T \mathbf{b})} + \lambda \|\mathbf{b}\|^2 \quad (18.27)$$

Assuming the optimal solution  $\mathbf{b}$ , then the prediction is

$$h(\mathbf{x}_0) = \mathbf{k}_0^T \hat{\boldsymbol{\beta}} = \mathbf{k}_0^T \mathbf{U} \mathbf{D}^{-1} \mathbf{b} \quad (18.28)$$

where  $\mathbf{k}_0 = [K(\mathbf{x}_0, \mathbf{x}_1), \dots, K(\mathbf{x}_0, \mathbf{x}_N)]^T$ .

### Ex. 18.14

(a)

$$\begin{aligned} f_+(x_0) &\approx \frac{\text{Total number in class +1 in this region}}{(\text{Total number in class +1})(\text{Volumn of this region})} \\ &= \frac{1}{N_+ d_+(x_0)^p} \end{aligned} \quad (18.29)$$

Therefore the discriminant function

$$\delta(x_0) = \log \frac{p_+(x_0)}{p_-(x_0)} = \log \frac{\pi_+ f_+(x_0)}{\pi_- f_-(x_0)} \quad (18.30)$$

If we estimate the prior distribution as  $\pi_+ = N_+/N$ ,  $\pi_- = N_-/N$ , then

$$\delta(x_0) = p \log \frac{d_-(x_0)}{d_+(x_0)} \quad (18.31)$$

(b) If  $\pi_+$ ,  $\pi_-$  is given, then

$$\delta(x_0) = \log \frac{\pi_+}{\pi_-} + \log \frac{N_-}{N_+} + p \log \frac{d_-(x_0)}{d_+(x_0)} \quad (18.32)$$

(c) Simply redefine  $d_+(x_0)$  as the smallest distance within which there are  $k$  samples in class +1, and  $d_-(x_0)$  as the smallest distance within which there are  $k$  samples in class -1. The results are the same as (a), (b).

**Ex. 18.15**

First we show that the  $m$ -th component  $\mathbf{z}_m$  can be written as  $z_{im} = \sum_{j=1}^N \alpha_{jm} K(\mathbf{x}_i, \mathbf{x}_j)$  up to centering, where  $\alpha_{jm} = u_{jm}/d_m$ . Since  $z_{im}$  is the entry in the  $i$ -th row,  $m$ -th column of  $\mathbf{Z}^T$ , where

$$\mathbf{Z}^T = \mathbf{D}\mathbf{U}^T = \mathbf{D}^{-1}\mathbf{U}^T(\mathbf{I} - \mathbf{M})\mathbf{K}(\mathbf{I} - \mathbf{M}) \quad (18.33)$$

Therefore  $z_{im}$  equals to the product of the  $i$ -th row of  $\mathbf{D}^{-1}\mathbf{U}^T$  and the  $m$ -th column of  $(\mathbf{I} - \mathbf{M})\mathbf{K}(\mathbf{I} - \mathbf{M})$ . Note that the  $j$ -th element of the former is  $u_{jm}/d_m$  and the  $j$ -th element of the latter is  $\langle h(\mathbf{x}_i) - \bar{h}, h(\mathbf{x}_j) - \bar{h} \rangle$  where  $\bar{h} = \sum_{j=1}^N h(\mathbf{x}_j)/N$ .

Denote the centered projection of  $\mathbf{x}_0$  onto the principle component direction as  $\mathbf{z}_0$ , then its  $m$ -th element is

$$\begin{aligned} z_{0m} &= \left\langle h(\mathbf{x}_0) - \bar{h}, \sum_{j=1}^N \frac{u_{jm}}{d_m} (h(\mathbf{x}_j) - \bar{h}) \right\rangle \\ &= \sum_{j=1}^N \frac{u_{jm}}{d_m} [\langle h(\mathbf{x}_0), h(\mathbf{x}_j) \rangle - \langle h(\mathbf{x}_0), \bar{h} \rangle - \langle \bar{h}, h(\mathbf{x}_j) \rangle + \langle \bar{h}, \bar{h} \rangle] \\ &= \sum_{j=1}^N \frac{u_{jm}}{d_m} [k_{0j} - (\mathbf{M}\mathbf{k}_0)_j - (\mathbf{K}\mathbf{1})_j/N + \mathbf{1}^T\mathbf{K}\mathbf{1}] \end{aligned} \quad (18.34)$$

Again  $u_{jm}/d_m$  is the  $m$ -th row,  $j$ -th column of  $\mathbf{D}^{-1}\mathbf{U}^T$ , and the term in the squared bracket equals to the  $j$ -th element of  $(\mathbf{I} - \mathbf{M})[\mathbf{k}_0 - \mathbf{K}\mathbf{1}/N]$ . Consequently,

$$\mathbf{z}_0 = \mathbf{D}^{-1}\mathbf{U}^T(\mathbf{I} - \mathbf{M})[\mathbf{k}_0 - \mathbf{K}\mathbf{1}/N] \quad (18.35)$$

**Ex. 18.16**

(a)

$$\Pr(A) = \Pr(\cup_{j=1}^M A_j) \leq \sum_{j=1}^M \Pr(A_j) = \alpha \quad (18.36)$$

(b) When  $\alpha/M$  is small, the first-order approximation

$$1 - (1 - \alpha/M)^M \approx \alpha \quad (18.37)$$

**Ex. 18.17**

(a) Since  $p_{(1)} \leq \dots \leq p_{(M)}$ ,  $|t_1| \geq \dots \geq |t_M|$ , we have  $|T|_{(L)} = |t_L|$ . By definition in (18.41),

$$p_0 = p_{(L)} = \frac{1}{MK} \sum_{j=1}^M \sum_{k=1}^K I(|t_j^k| > |t_L|) \quad (18.38)$$

thus there are at most  $p_0$  of  $|t_j^k| > |t_L| = |T|_{(L)}$ . For the plug-in estimation, we have

$$R_{\text{obs}} = \sum_{j=1}^M I(|t_j| > |t_L|) = L \quad (18.39)$$

$$\widehat{E(V)} = M \cdot \frac{1}{MK} \sum_{j=1}^M \sum_{k=1}^K I(|t_j^k| > |t_L|) \leq p_0 M \quad (18.40)$$

therefore  $\widehat{\text{FDR}} \leq p_0 M / L = \alpha$ .

(b) According to (18.44)

$$\begin{aligned} p_{(L+1)} &> \alpha \frac{L+1}{M} \\ &= \frac{1}{MK} \sum_{j=1}^M \sum_{k=1}^K I(|t_j^k| > |t_{L+1}|) \\ &= \frac{\widehat{E(V)}}{M} \end{aligned} \quad (18.41)$$

Also we have

$$R_{\text{obs}} = \sum_{j=1}^M I(|t_j| > |t_{L+1}|) = L+1 \quad (18.42)$$

therefore  $\widehat{\text{FDR}} = \widehat{E(V)} / R_{\text{obs}} > \alpha$ .

**Ex. 18.18**

$$\begin{aligned}
\text{pFDR}(\Gamma) &= \frac{\Pr(j\text{-th null hypothesis is true and the null hypothesis is rejected})}{\Pr(j\text{-th null hypothesis is rejected})} \\
&= \frac{\Pr(Z_j = 0, t_j \in \Gamma)}{\Pr(t_j \in \Gamma)} \\
&= \frac{\Pr(Z_j = 0)\Pr(t_j \in \Gamma|Z_j = 0)}{\Pr(Z_j = 0)\Pr(t_j \in \Gamma|Z_j = 0) + \Pr(Z_j = 1)\Pr(t_j \in \Gamma|Z_j = 1)} \\
&= \frac{\pi_0\{\text{Type I error of } \Gamma\}}{\pi_0\{\text{Type I error of } \Gamma\} + \pi_1\{\text{Power of } \Gamma\}} \tag{18.43}
\end{aligned}$$

**Ex. 18.19 (Program)**

**Ex. 18.20**

$$\begin{aligned}
\text{pFDR} &= \mathbb{E} \left[ \frac{V}{R} | R > 0 \right] \\
&= \sum_{k=1}^M \mathbb{E} \left[ \frac{V}{R} | R = k \right] \Pr(R = k | k > 0) \tag{18.44}
\end{aligned}$$

Since  $V$  is binomial distributed from 0 to  $k$  given  $k$ , the expectation of  $V$  given  $k$  is

$$\mathbb{E}_k[V] = k\Pr(H = 0|T \in \Gamma) \tag{18.45}$$

therefore

$$\begin{aligned}
\text{pFDR} &= \sum_{k=1}^M \Pr(H = 0|T \in \Gamma)\Pr(R = k | k > 0) \\
&= \Pr(H = 0|T \in \Gamma) \tag{18.46}
\end{aligned}$$

## References

- [1] J. L. Weatherwax and D. Epstein, “A solution manual and notes for: The elements of statistical learning by jerome friedman, trevor hastie, and robert tibshirani,” June 2013. [Online]. Available: [http://waxworksmath.com/Authors/G\\_M/Hastie/hastie.html](http://waxworksmath.com/Authors/G_M/Hastie/hastie.html)
- [2] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer series in statistics Springer, Berlin, 2009.
- [3] W. Wu, “A partial solution manual for: The elements of statistical learning by jerome friedman, trevor hastie, and robert tibshirani,” 2016. [Online]. Available: <https://github.com/huragok/IDA>
- [4] T. Hastie, R. Tibshirani, and A. Buja, “Flexible discriminant analysis by optimal scoring,” *Journal of the American statistical association*, vol. 89, no. 428, pp. 1255–1270, 1994.
- [5] G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach,” *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.