Elements of Statistical Learning
Solutions to the Exercises

Yu Zhang, sjtuzy@gmail.com

November 25, 2009

Exercise 2.6 Consider a regression problem with inputs x_i and outputs y_i, and a parameterized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with tied or identical values of x, then the fit can be obtained from a reduced weighted least squares problem.

Proof For known heteroskedasticity (e.g., grouped data with known group sizes), use weighted least squares (WLS) to obtain efficient unbiased estimates. Fig.1 explains "Observations with tied or identical values of x". In

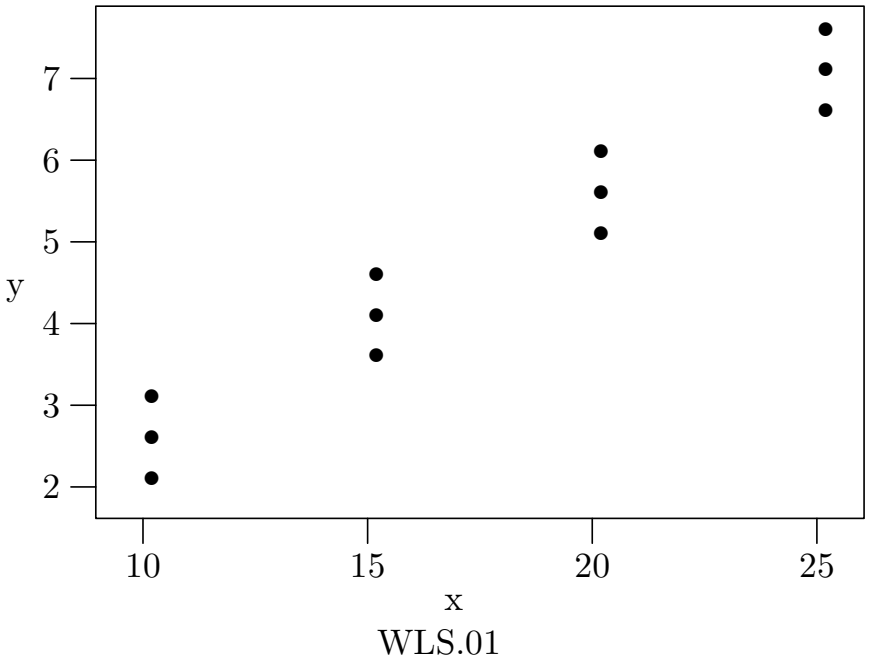

Figure 1: Observations with tied

the textbook, section 2.7.1 also explain this problem.

If there are multiple observation pairs $x_i, y_{il}, l = 1, 2 \ldots, N_i$ at each value of x_i, the risk is limited as follows:

$$\begin{aligned}
\underset{\theta}{\operatorname{argmin}} \sum_i \sum_{l=1}^{N_i} (f_\theta(x_i) - y_{il})^2 &= \underset{\theta}{\operatorname{argmin}} \sum_i (N_i f_\theta(x_i)^2 - 2 \sum_l^{N_i} f_\theta(x_i) y_{il} \\
&+ \sum_l^{N_i} y_{il}^2) \\
&= \underset{\theta}{\operatorname{argmin}} \sum_i N_i \{(f_\theta(x_i) - \overline{y_i})^2 + \text{Constant}\} \\
&= \underset{\theta}{\operatorname{argmin}} \sum_i N_i \{(f_\theta(x_i) - \overline{y_i})^2\} \qquad (1)
\end{aligned}$$

which is a weighted least squares problem.

Exercise 3.19 Show that $||\hat{\beta}^{ridge}||$ increases as its tuning parameter $\lambda \to 0$. Does the same property hold for the lasso and partial least squares estimates? For the latter, consider the "tuning parameter" to be the successive steps in the algorithm.

Proof Let $\lambda_2 < \lambda_1$ and β_1, β_2 be the optimal solution. We denote the loss function as follows

$$f_\lambda(\beta) = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|| + \lambda ||\boldsymbol{\beta}||_2^2$$

Then we have

$$\begin{aligned}
f_{\lambda_1}(\beta_2) + f_{\lambda_2}(\beta_1) &\geq f_{\lambda_2}(\beta_2) + f_{\lambda_1}(\beta_1) \\
\lambda_1 ||\beta_2||_2^2 + \lambda_2 ||\beta_1||_2^2 &\geq \lambda_2 ||\beta_2||_2^2 + \lambda_1 ||\beta_1||^2 \\
(\lambda_1 - \lambda_2) ||\beta_2||_2^2 &\geq (\lambda_1 - \lambda_2) ||\beta_1||_2^2 \\
||\beta_2||_2^2 &\geq ||\beta_1||_2^2
\end{aligned}$$

So $||\hat{\beta}^{\text{ridge}}||$ increases as its tuning parameter $\lambda \to 0$.

Similarly, in lasso case, $||\hat{\beta}||_1$ increase as $\lambda \to 0$. But this can't guarantee the l_2-norm increase. Fig.2 is a direct view of this property.

In partial least square case, it can be shown that the PLS algorithm is equivalent to the conjugate gradient method. This is a procedure that iteratively computes approximate solutions of $|\boldsymbol{\beta} A = b|$ by minimizing the quadratic function

$$\frac{1}{2} \boldsymbol{\beta}^\top A \beta - \boldsymbol{b}^\top \beta$$

along directions that are $|\boldsymbol{A}|$-orthogonal (Ex.3.18). The approximate solution obtained after m steps is equal to the PLS estimator obtained after p iterations. The canonical algorithm can be written as follows:

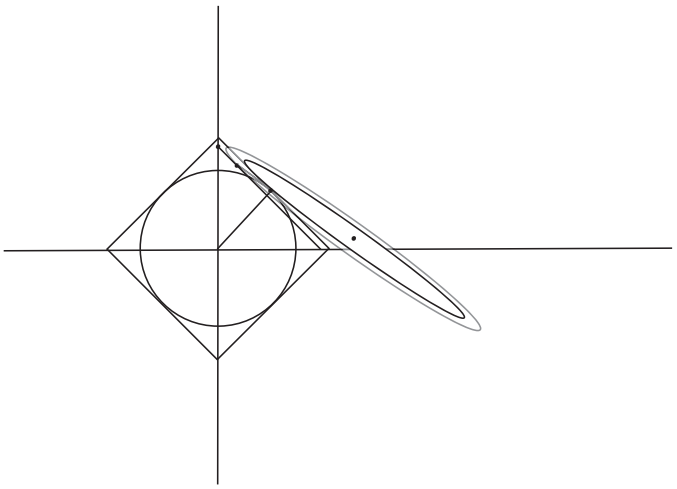

Figure 2: Lasso

1. Initialization $\boldsymbol{\beta}_0 = 0, \boldsymbol{d}_0 = \boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{\beta}_0 = \boldsymbol{b}$
2. $a_i = \frac{\boldsymbol{d}_i^H \boldsymbol{r}_i}{\boldsymbol{d}_i^H \boldsymbol{A}\boldsymbol{d}_i}$
3. $\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + a_i \boldsymbol{d}_i$
4. $\boldsymbol{r}_{i+1} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{\beta}_{i+1} (= \boldsymbol{r}_i - a_i \boldsymbol{A}\boldsymbol{d}_i)$
5. $b_i = -\frac{\boldsymbol{r}_{i+1}^H \boldsymbol{A}\boldsymbol{d}_i}{\boldsymbol{d}_i^H \boldsymbol{A}\boldsymbol{d}_i}$
6. $\boldsymbol{d}_{i+1} = \boldsymbol{r}_{i+1} + b_i \boldsymbol{d}_i$

The squared norm of $\boldsymbol{\beta}_{j+1}$ can be written as

$$||\boldsymbol{\beta}_{j+1}||^2 = ||x_j||^2 + 2a_j^2 ||\boldsymbol{d}_j||^2 + a_j \boldsymbol{d}_j^H \boldsymbol{\beta}_j \tag{2}$$

We need only shown that $a_j \boldsymbol{d}_j^H \boldsymbol{\beta}_j > 0$.

$$\because \quad \boldsymbol{r}_{j+1}^H \boldsymbol{d}_{j+1} = \boldsymbol{r}_{j+1}^H \boldsymbol{r}_{j+1} + b_j \boldsymbol{r}_{j+1}^H \boldsymbol{d}_j \text{ and } \langle \boldsymbol{r}_{j+1}, \boldsymbol{d}_j \rangle = 0 \tag{3}$$

$$\therefore \quad a_j = \frac{\boldsymbol{d}_j^H \boldsymbol{r}_j}{\boldsymbol{d}_j^H \boldsymbol{A}\boldsymbol{d}_j} = \frac{||\boldsymbol{r}_j||^2}{||\boldsymbol{d}_j||_{\boldsymbol{A}}^2} > 0 (\boldsymbol{A} \text{ is positive definite.}) \tag{4}$$

$$\because \quad \boldsymbol{\beta}_j = \sum_{i=0}^{j} a_i \boldsymbol{d}_i \tag{5}$$

$$\therefore \quad \boldsymbol{d}_j^H \boldsymbol{\beta}_j = \sum_{i=0}^{j} a_i \boldsymbol{d}_j^H \boldsymbol{d}_i \tag{6}$$

Now we need to show that $\boldsymbol{d}_j^H \boldsymbol{d}_i > 0, i \neq j$. By Step 6, we have $\boldsymbol{d}_j = \boldsymbol{r}_j + \sum_{i=0}^{j-1}(\prod_{k=i}^{j-1} b_k)\boldsymbol{r}_i$.

$$\begin{aligned}
&\because \quad \langle \boldsymbol{r}_i, \boldsymbol{r}_j \rangle = 0, i \neq j && (7)\\
&\therefore \quad b_k > 0 \rightarrow \boldsymbol{d}_j^H \boldsymbol{d}_i > 0 && (8)\\
&\because \quad \boldsymbol{A}\boldsymbol{d}_i = a_i^{-1}(\boldsymbol{r}_i - \boldsymbol{r}_{i+1}) && (9)\\
&\therefore \quad b_j = -\frac{\boldsymbol{r}_{j+1}^H(\boldsymbol{r}_j - \boldsymbol{r}_{j+1})}{a_j||\boldsymbol{d}_j||_{\boldsymbol{A}}^2} = \frac{||\boldsymbol{r}_{j+1}||^2}{a_j||\boldsymbol{d}_j||_{\boldsymbol{A}}^2} > 0 && (10)
\end{aligned}$$

So in PLS case, the $||\beta||_2^2$ increase with m.

Exercise 4.1 Show how to solve the generalized eigenvalue problem $\max a^\top \boldsymbol{B} a$ subject to $a^\top \boldsymbol{W} a = 1$ by transforming to a standard eigenvalue problem.

Proof Hence $\boldsymbol{W}$ is the the common covariance matrix, we have $\boldsymbol{W}$ is semi-positive definite. WLOG $\boldsymbol{W}$ is positive definite, we have

$$\boldsymbol{W} = \boldsymbol{P}^2 \tag{11}$$

Let $b = \boldsymbol{P}a$, then the problem is

$$a^\top \boldsymbol{B} a = b^\top \boldsymbol{P}^{-1}\boldsymbol{B}\boldsymbol{P}^{-1} b = b^\top \boldsymbol{B}^* b \tag{12}$$

subject to $a^\top \boldsymbol{W} a = 1 = b^\top \boldsymbol{B}^* b = 1$. Now the problem transform to a standard eigenvalue problem.

Exercise 4.2 Suppose we have features $x \in \mathbb{R}^p$, a two-class response, with class sizes N_1, N_2, and the target coded as $-N/N_1, N/N_2$.

(d) Show that this result holds for any (distinct) coding of the two classes.

W.L.O.G any distinct coding y_1', y_2'. Let $\hat{\beta} = (\beta, \beta_0)^\top$. Compute the partial deviation of the RSS$(\hat{\beta})$ we have

$$\begin{aligned}
\frac{\partial \mathrm{RSS}(\hat{\beta})}{\partial \beta} &= -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta^\top x_i)x_i^\top = 0 && (13)\\
\frac{\partial \mathrm{RSS}(\hat{\beta})}{\partial \beta_0} &= -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta^\top x_i) = 0 && (14)
\end{aligned}$$

It yields that

$$\begin{aligned}
\beta^\top \sum_i (x_i - \bar{x_i})x_i^\top &= \sum_i y_i x_i && (15)\\
\beta_0 &= \frac{1}{N}\sum_i (y_i - \beta^\top x_i) && (16)
\end{aligned}$$

Than we get

$$\begin{aligned}
\beta^\top \sum_i (x_i - \bar{x_i}) x_i^\top &= \sum_i (y_i - \frac{1}{N}\sum_i y_i) x_i^\top && (17)\\
&= N_1 y_1' \hat{\mu}_1 + N_2 y_2' \hat{\mu}_2 - \frac{N_1 y_1' + N_2 y_2'}{N_1 + N_2}(N_1\hat{\mu}_1 + N_2\hat{\mu}_2) && (18)\\
&= \frac{1}{N}(N_1 N_2 y_2'(\hat{\mu}_2 - \hat{\mu}_1) - N_1 N_2 y_1'(\hat{\mu}_2 - \hat{\mu}_1)) && (19)\\
&= \frac{N_1 N_2}{N}(y_2' - y_1')(\hat{\mu}_2 - \hat{\mu}_1) && (20)
\end{aligned}$$

Hence the proof in (c) still holds.

Exercise 4.3 Suppose we transform the original predictors X to $\hat{Y}$ via linear regression. In detail, let $\hat{Y} = X(X^\top X)^{-1}X^\top Y = X\hat{B}$, where Y is the indicator response matrix. Similarly for any input $x \in \mathbb{R}^p$, we get a transformed vector $\hat{y} = \hat{B}^\top x \in \mathbb{R}^K$. Show that LDA using $\hat{Y}$ is identical to LDA in the original space.

Proof Prof.Zhang had given the solution about this problem, but only tow student notice this problem is tricky. Here I give the main part of the solution.

- First $y_1, \ldots, y_K$ must independent, then we have $r(\Sigma_y) = K$.
- Assume that $\hat{B}_{p*K}, r(\hat{B}) = K < p, r(\Sigma_x) = p$. Note that

  $$r(\hat{B}(\hat{B}^\top \Sigma_x \hat{B})^{-1}\hat{B}^\top) < r(\hat{B}) < r(\Sigma_x)$$

  . Since $\hat{B}(\hat{B}^\top \Sigma_x \hat{B})^{-1}\hat{B}^\top \neq \Sigma_x^{-1}$ when $dim(Y) < dim(X)$.
- This problem are equal to prove that $\hat{B}(\hat{B}^\top \Sigma_x \hat{B})^{-1}\hat{B}^\top(\mu_k^x - \mu_l^x) = \Sigma_x^{-1}(\mu_k^x - \mu_l^x)$. Let $P = \Sigma_x \hat{B}(\hat{B}^\top \Sigma_x \hat{B})^{-1}\hat{B}^\top$, we have

  $$P^2 = P \tag{21}$$

  Hence P is projection matrix. Note that $Y = \{y_1, y_2 \ldots, y_K\}$ is indi-

cator response matrix, we have

$$\begin{aligned}
\mu_k^x &= \frac{1}{N_k} X^\top y_k \\
\Sigma_x &= \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \mu_i^k)(x_i - \mu_i^k)^\top / (N-K) \\
&= \frac{1}{N-K} (\sum_{k=1}^{K} \sum_{g_i=k} x_i x_i^\top - \sum_{k=1}^{K} N_k \mu_k^x (\mu_k^x)^\top) \\
&= \frac{1}{N-K} (X^\top X - \sum_{k=1}^{K} X^\top y_k y_k^\top) \\
&= \frac{1}{N-K} (X^\top X - X^\top Y Y^\top X)
\end{aligned}$$

Note that

$$P(N_1\mu_1^x, N_2\mu_2^x \ldots, N_K\mu_K^x) = PX^\top Y \tag{22}$$

So we need only to shown that $PX^\top Y = X^\top Y$.

$$PX^\top Y = \Sigma_x B (B^\top \Sigma_x B) B^\top X^\top Y$$

By the definition of B, we have

$$\begin{aligned}
B^\top X^\top Y &= Y^\top X (X^\top X)^{-1} X^\top Y \\
\Sigma_x B &= \frac{1}{N-K} ((X^\top X)(X^\top X)^{-1} X^\top Y - X^\top Y Y^\top X (X^\top X)^{-1} X^\top Y) \\
&= \frac{1}{N-K} X^\top Y (I - Y^\top X (X^\top X)^{-1} X^\top Y) \\
(B^\top \Sigma_x B)^{-1} &= (N-K)(Y^\top X (X^\top X)^{-1} (X^\top X - X^\top Y Y^\top X)(X^\top X)^{-1} X^\top Y)^{-1} \\
&= (N-K)(Y^\top X (X^\top X)^{-1} X^\top Y [I - Y^\top X (X^\top X)^{-1} X^\top Y])^{-1}
\end{aligned}$$

Let $Q = Y^\top X (X^\top X)^{-1} X^\top Y$ we have

$$B^\top X^\top Y = Q \tag{23}$$

$$\Sigma_x B = \frac{1}{N-K} X^\top Y (I - Q) \tag{24}$$

$$(B^\top \Sigma_x B)^{-1} = (N-K)(Q(I-Q))^{-1} \tag{25}$$

Hence $Q(I-Q)$ is invertible matrix, we have

$$K = r(Q(Q-I)) \leq r(Q) \Rightarrow r(Q) = K \tag{26}$$

So Q is invertible matrix, then it yields

$$(B^\top \Sigma_x B)^{-1} = (N-K)(I-Q)^{-1} Q^{-1} \tag{27}$$

Combine with (16),(17),(20), we have

$$\begin{aligned} PX^\top Y &= \frac{1}{N-K} X^\top Y (I-Q)(N-K)(I-Q)^{-1}Q^{-1}Q \quad (28) \\ &= X^\top Y \quad (29) \end{aligned}$$

Since we have $\hat{B}(\hat{B}^\top \Sigma_x \hat{B})^{-1}\hat{B}^\top(\mu_k^x - \mu_l^x) = \Sigma_x^{-1}(\mu_k^x - \mu_l^x)$.