CS 760 HW 2

Name: Kefan Zheng

Student ID: 908 617 5008

Email: kzheng58@wisc.edu

Problem 2.1

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|y - X\theta\|_2^2$$

$$= \underset{\theta}{\operatorname{argmin}} (y - X\theta)^T (y - X\theta)$$

$$= \underset{\theta}{\operatorname{argmin}}\ y^T y - y^T X\theta - \theta^T X^T y + (X\theta)^T X\theta$$

$$= \underset{\theta}{\operatorname{argmin}}\ y^T y - 2\theta^T X^T y + \theta^T X^T X \theta$$

Taking derivative:

$$\frac{d}{d\theta}(y^T y - 2\theta^T X^T y + \theta^T X^T X\theta) = 2X^T X\theta - 2X^T y$$

Setting to zero:

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Problem 2.2

$$P(y, X | \theta, \Sigma^*) = \frac{1}{\sqrt{(2\pi)^D |\Sigma^*|}} e^{-\frac{1}{2}(y - X\theta)^T \Sigma^{*-1} (y - X\theta)}$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ P(y, X | \theta, \Sigma^*)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} -\frac{1}{2}(y - X\theta)^T \Sigma^{*-1} (y - X\theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ (y - X\theta)^T \Sigma^{*-1} (y - X\theta)$$

Let $f(\theta) = (y - X\theta)^T \Sigma^{*-1} (y - X\theta) = (y^T \Sigma^{*-1} - \theta^T X^T \Sigma^{*-1})(y - X\theta)$

$$= y^T \Sigma^{*-1} y - y^T \Sigma^{*-1} X\theta - \theta^T X^T \Sigma^{*-1} y + \theta^T X^T \Sigma^{*-1} X\theta$$

Taking derivative:

$$\frac{df}{d\theta} = -y^T \Sigma^{*-1} X - X^T \Sigma^{*-1} y + 2X^T \Sigma^{*-1} X\theta$$

$$= 2X^T \Sigma^{*-1} X\theta - 2X^T \Sigma^{*-1} y$$

Setting to zero:

$$2X^T \Sigma^{*-1} X\theta = 2X^T \Sigma^{*-1} y$$

So, $\hat{\theta} = (X^T \Sigma^{*-1} X)^{-1} X^T \Sigma^{*-1} y$

Problem 2.3

$\epsilon \sim N(0, \Sigma^*)$, $y = X\theta^* + \epsilon$, so $y \sim N(X\theta^*, \Sigma^*)$

$$E(\hat\theta) = E[(X^T\Sigma^{*-1}X)^{-1}X^T\Sigma^{*-1}y] = (X^T\Sigma^{*-1}X)^{-1}X^T\Sigma^{*-1}E(y)$$
$$= (X^T\Sigma^{*-1}X)^{-1}X^T\Sigma^{*-1}X\theta^* = \theta^*$$

$$Cov(\hat\theta) = Cov((X^T\Sigma^{*-1}X)^{-1}X^T\Sigma^{*-1}y)$$
$$= (X^T\Sigma^{*-1}X)^{-1}X^T\Sigma^{*-1}\,Cov(y)\,[(X^T\Sigma^{*-1}X)^{-1}X^T\Sigma^{*-1}]^T$$
$$= X^{-1}\Sigma^*X^{T^{-1}}X^T\Sigma^{*-1}\Sigma^*(X^{-1}\Sigma^*X^{T^{-1}}X^T\Sigma^{*-1})^T$$
$$= X^{-1}\Sigma^*X^{T^{-1}} = X^{-1}(X^T\Sigma^{*-1})^{-1}$$
$$= (X^T\Sigma^{*-1}X)^{-1}$$

So $\hat\theta \sim N(\theta^*, (X^T\Sigma^{*-1}X)^{-1})$

Problem 2.4

$\hat y = X\hat\theta = x^T\hat\theta$

Problem 2.5

Because $\hat\theta \sim N(\theta^*, (X^T\Sigma^{*-1}X)^{-1})$, so $\hat y \sim N$

$E(\hat y) = x^T E(\hat\theta) = x^T\theta^*$

$Var(\hat y) = Var(x^T\hat\theta) = x^T Cov(\hat\theta)x = x^T(X^T\Sigma^{*-1}X)^{-1}x$

So $\hat y \sim N(x^T\theta^*, x^T(X^T\Sigma^{*-1}X)^{-1}x)$

Problem 2.6

$$P(y, X \mid \theta^*, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(y - X\theta^*)^T \Sigma^{-1} (y - X\theta^*)}$$

$$\underset{\Sigma}{\operatorname{argmax}} P(y, X \mid \theta^*, \Sigma) = \underset{\Sigma}{\operatorname{argmax}} \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(y - X\theta^*)^T \Sigma^{-1} (y - X\theta^*)}$$

$$= \underset{\Sigma}{\operatorname{argmax}} -\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\theta^*)^T \Sigma^{-1} (y - X\theta^*)$$

$$= \underset{\Sigma}{\operatorname{argmax}} -\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \operatorname{trace}\left[(y - X\theta^*)^T \Sigma^{-1} (y - X\theta^*)\right]$$

$$= \underset{\Sigma}{\operatorname{argmax}} -\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\Sigma^{-1}| - \frac{1}{2} \operatorname{trace}\left[\Sigma^{-1} (y - X\theta^*)^T (y - X\theta^*)\right]$$

Taking derivative:

according to properties of trace: $\frac{d}{dX} \operatorname{trace}(XY) = Y^T$

and determinant: $\frac{d}{dX} \log |X| = X^{-1^T}$

$$\frac{d}{d\Sigma^{-1}} \log P(y, X \mid \theta^*, \Sigma) = \frac{1}{2} \Sigma - \frac{1}{2} (y - X\theta^*)(y - X\theta^*)^T$$

Setting to zero:

$$\frac{1}{2} \Sigma = \frac{1}{2} (y - X\theta^*)(y - X\theta^*)^T$$

So $\hat{\Sigma} = (y - X\theta^*)(y - X\theta^*)^T$

Problem 2.7

(a) Add an intercept to the model first

Because we don't know the θ^* and Σ^* here, and their MLE expressions contain each other. So choose to use the iterative EM algorithm to estimate θ and Σ.

Use the $\hat{\theta}$ obtained by MSE $\left(\hat{\theta} = (X^T X)^{-1} X^T y\right)$ and the $\hat{\Sigma}$ calculated from this $\hat{\theta}$ $\left(\hat{\Sigma} = (y - X\theta^*)(y - X\theta^*)^T\right)$ as the iteration initial value. The final iterative convergence result is the same as the $\hat{\theta}$ obtained by MSE.

$\hat{\Sigma}$ is numerically close to non-invertible, use regularization $\Sigma' = \Sigma + \lambda I$, $\lambda = 1e{-6}$ to make it invertible. So, finally

$\hat{\theta} =$

```
theta_star:
[[  0.11595569]
 [  1.06057035]
 [-50.45722366]]
```

, $\hat{\Sigma} =$

```
sigma_star:
[[ 380.34555245  244.45529197 -697.15027709   72.43119763]
 [ 244.45529197  157.23159558 -448.16803534   46.56291277]
 [-697.15027709 -448.16803534 1278.19060653 -132.79052901]
 [  72.43119763   46.56291277 -132.79052901   13.8781836 ]]
```

(b) Because we add an intercept to the Linear Regression model,

so $\hat{y} = x^T \hat{\theta} = \begin{bmatrix} 175 & 170 & 1 \end{bmatrix} \begin{bmatrix} 0.12 \\ 1.06 \\ -50.46 \end{bmatrix} =$

```
y_hat_new:
[[150.13198209]]
```

(c) confidence $= 95\%$, so $\alpha = 0.05$

```
alpha = 0.05
tau = -norm.ppf(alpha/2, scale=np.sqrt(var_y_hat_new))
```

$$\tau = \Phi_N^{-1}\left(\frac{\alpha}{2} \,\middle|\, 0,\ x^T (X^T \Sigma^{*-1} X)^{-1} x\right) = 0.30 =$$

```
tau:
[[0.30285235]]
```

$(\hat{y} - \tau,\ \hat{y} + \tau) =$

```
confidence interval:
[[149.82912973]] [[150.43483444]]
```

(d) assume $\alpha = 0.05$

For height $\hat{\theta}_j^2 = (0.1159)^2$

$\mathrm{cov}(\hat{\theta}) = (X^T \Sigma^{*-1} X)^{-1} =$

```
cov_theta_hat:
[[ 1.20248315e-04 -1.69580956e-05 -1.76248571e-02]
 [-1.69580956e-05  4.43222957e-05 -4.85232789e-03]
 [-1.76248571e-02 -4.85232789e-03  3.88785542e+00]]
```

So $v_j = 1.2e\text{-}04$,

```
alpha = 0.05
value = chi2.ppf(1-alpha, df=1)
```

$\Phi_\chi^{-1}(\alpha) =$

```
chi2_0.05:
3.841458820694124
```

So $\hat{\theta}_j^2 > v_j^2 \Phi_\chi^{-1}(\alpha)$, so height is significant.

(e) assume $\alpha = 0.05$, for weight $\hat{\theta}_j^2 = (1.06)^2$

$v_j = 4.43e\text{-}05$, $\Phi_\chi^{-1}(\alpha) =$

```
chi2_0.05:
3.841458820694124
```

So $\hat{\theta}_j^2 > v_j^2 \Phi_\chi^{-1}(\alpha)$, so weight is significant.