Problem 2.1

(a) $D = \{k_1, \dots, k_n\}$

$P(D) = \prod_{i=1}^{n} P(x = k_i \mid \lambda)$

$\log P(D) = \sum_{i=1}^{n} P(k_i)$

$= -\lambda n + \sum_{i=1}^{n} \log \frac{1}{k_i!} + \sum_{i=1}^{n} k_i \cdot \log \lambda$

$\frac{\partial}{\partial \lambda} \log P(D) = -n + \frac{\sum_{i=1}^{n} k_i}{\lambda}$

let $\frac{\partial}{\partial \lambda} \log P(D) = 0$

$\Rightarrow \quad \hat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{n}$

check second order derivative

$\frac{\partial^2}{\partial \lambda^2} \log P(D) = -\frac{\sum_{i=1}^{n} k_i}{\lambda^2} < 0$

(b) $E(\hat{\lambda} - \lambda)$

$= E\left(\frac{1}{n} \sum_{i=1}^{n} k_i\right) - \lambda$

$= \frac{1}{n} \cdot \sum_{i=1}^{n} E(k_i) - \lambda$

$= \frac{1}{n} \cdot n\lambda - \lambda = 0$

$Var(\hat{\lambda}) = E[(\hat{\lambda} - E(\hat{\lambda}))^2]$

$= E[(\hat{\lambda} - \lambda)^2]$

$= E[(\frac{1}{n} \sum_i (k_i - \lambda))^2]$

$= \frac{1}{n^2} E[(\sum_i (k_i - \lambda))^2]$

$= \frac{1}{n^2} E[\sum_i (k_i - \lambda) \cdot \sum_j (k_j - \lambda)]$

$$\left(\begin{array}{ll} i = j & E[(k_i - \lambda)(k_j - \lambda)] \\ & = E[(k_i - \lambda)^2] = \lambda \\ i \neq j & E[(k_i - \lambda)(k_j - \lambda)] = 0 \end{array} \right)$$

$= \frac{1}{n^2} \cdot n \cdot \lambda = \frac{\lambda}{n}$

(c) $\hat{\lambda} = \frac{1}{n} \cdot \sum_{i=1}^{n} k_i$

$= \frac{1}{576} \cdot (211 + 93\times2 + 35\times3 + 7\times4 + 5)$

≈ 0.9288

(d) $x_1 = P(x=0|\hat{\lambda})\times n = e^{-0.9288} \times 0.9288^0 \times 576 \approx 228$

$x_2 = P(x=1|\hat{\lambda})\times n = e^{-0.9288} \times 0.9288^1 \times 576 \approx 211$

$x_3 = P(x=2|\hat{\lambda})\times n = \frac{1}{2} \times e^{-0.9288} \times 0.9288^2 \times 576 \approx 98$

$x_4 = P(x=3|\hat{\lambda})\times n = \frac{1}{6} \times e^{-0.9288} \times 0.9288^3 \times 576 \approx 30$

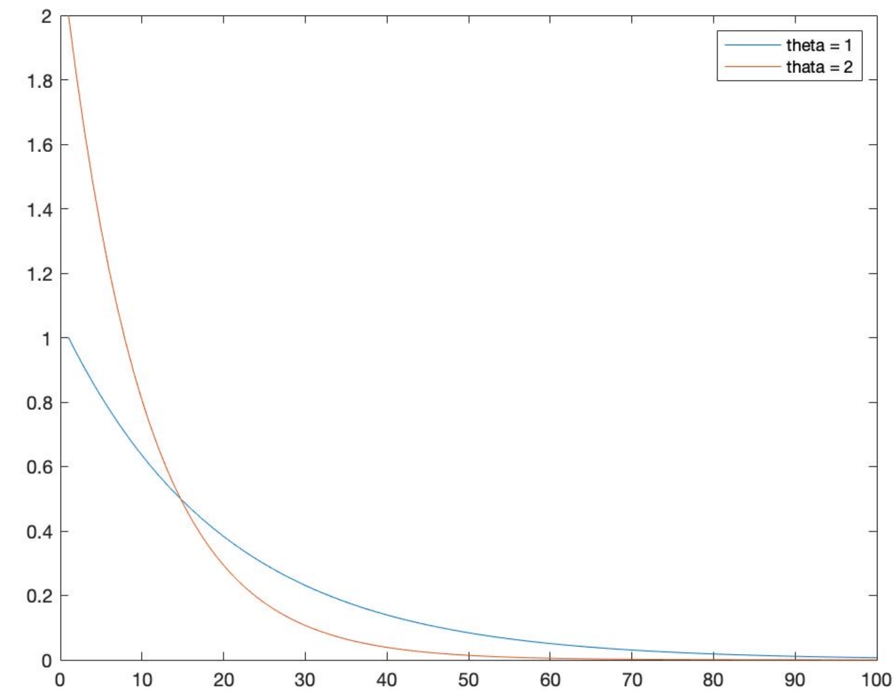
$x_5 = P(x=4|\hat{\lambda})\times n = \frac{1}{24} \times e^{-0.9288} \times 0.9288^4 \times 576 \approx 7$

$x_6 = 576 - \sum_i x_i = 2$

The possion distribution using estimate $\hat{\lambda}$ from MLE fits the actual data very well. Hence, the Germans very likely targeted areas randomly. (a possion distribution)

Promblem 2.2

(a)



(b) $\log P(D) = N \log \theta - \theta \sum_i x_i$

$$\frac{\partial}{\partial \theta} \log P(D) = \frac{N}{\theta} - \sum x_i = 0$$

$$\Rightarrow \quad \hat{\theta} = \frac{N}{\sum_i x_i} = \frac{1}{\frac{1}{N} \cdot \sum_i x_i}$$

$$\frac{\partial^2}{\partial \theta^2} \log P(D) = -\frac{N}{\theta^2} < 0$$

(c) $\log P(D) = -N \log \lambda - \frac{1}{\lambda} \cdot \sum_i x_i$

$$\frac{\partial}{\partial \lambda} \log P(D) = -\frac{N}{\lambda} + \frac{1}{\lambda^2} \sum_i x_i = 0$$

$$\Rightarrow \quad \hat{\lambda} = \frac{\sum_i x_i}{N}$$

$$\hat{\lambda} = \frac{1}{\hat{\theta}}$$

PS 2.3

Consider $g(\theta): \Omega \to \Gamma$

Define $\Omega_\gamma = \{\theta \mid g(\theta) = \gamma\}$ ($\Omega = \bigcup_{\gamma \in \Gamma} \Omega_\gamma$)

($g(\theta)$ is not necessarily one-to-one)

Let $M_x(\gamma) = \max_{\theta \in \Omega_\gamma} L_x(\theta)$

We are to find $\hat{\gamma}$ such that $M_x(\gamma)$ is maximized.

Now $M_x(\hat{\gamma}_0) = \max_{\theta \in \Omega_{\hat{\gamma}_0}} L_x(\theta) \geq L_x(\hat{\theta})$ $\theta \in \Omega_{\hat{\gamma}_0}$

where $\Omega_{\hat{\gamma}_0} = \{\theta : g(\theta) = \hat{\gamma}_0\}$ As $g(\hat{\theta}) = \hat{\gamma}_0$ so $\hat{\theta} \in \Omega_{\hat{\gamma}_0}$

Again $M_x(\hat{\gamma}_0) \leq \max_{\hat{\gamma} \in \Gamma} M_x(\hat{\gamma})$

$$= \max_{\hat{\gamma} \in \Gamma} \max_{\theta \in \Omega_{\hat{\gamma}}} L_x(\theta)$$

$$= \max_{\theta \in \Omega} L_x(\theta) = L_x(\hat{\theta})$$

Therefore $M_x(\hat{\gamma}_0) = L_x(\hat{\theta}) = \max M_x(\gamma)$

PS 2.4

(a) $\log p(D) = -n\log 2\lambda - \frac{1}{\lambda}\cdot\sum_i |x_i-u|$

$= -n\log 2\lambda - \frac{1}{\lambda}\sum_{i=1}^{k} u - x_i$

$-\frac{1}{\lambda}\sum_{i=k+1}^{n} x_i - u \quad (k\in\{0,\dots,n\})$

$\frac{\partial}{\partial\lambda}\log p(D) = -\frac{2n}{2\lambda} + \frac{1}{\lambda^2}\sum_i|x_i-u| = 0$

$\Rightarrow \quad \hat\lambda = \frac{\sum_i |x_i-u|}{n}$

$\log p(D|\hat\lambda) = -n\log 2\lambda - n$

By observation $\hat{u} = med(x_i)$

PS 2.5

(a) $p(D) = \prod_i \frac{1}{\sqrt{2\pi}\alpha}\cdot e^{-\frac{1}{2}\frac{|x_i-u|^2}{\alpha^2}}$

$\log p(D) = -\frac{1}{2}\sum_i\left(\frac{x_i-u}{\alpha}\right)^2 - \log\sqrt{2\pi}n - \log\alpha n$

$\frac{\partial}{\partial u}\log p(D) = -\sum_i\left(\frac{x_i-u}{\alpha}\right)\cdot(-1)$

$= \frac{1}{\alpha}\sum_i(x_i-u) = 0$

$\Rightarrow u = \frac{\sum x_i}{n}$

(b) $$\log P(D) = -\frac{1}{2}\cdot\frac{1}{\alpha^2}\cdot\sum_i (x_i-\mu)^2 - \log \alpha n$$

$$\frac{\partial}{\partial\alpha}\log P(D) = \frac{1}{\alpha^3}\sum_i (x_i-\mu)^2 - \frac{1}{\alpha}n = 0$$

$$\Rightarrow \hat{\alpha}^2 = \frac{\sum_i (x_i-\mu)^2}{n}$$

PS2-6

(a) $$P(D) = \frac{1}{|2\pi|^{\frac{d}{2}}\cdot|\Sigma|^{\frac{1}{2}}}\cdot e^{-\frac{1}{2}\|x_i-\mu\|_\Sigma^2}$$

$$\log P(D) = -\frac{1}{2}\cdot\sum_i \|x_i-\mu\|_\Sigma^2 - n\log|2\pi|^{\frac{d}{2}} - \frac{n}{2}\cdot\log|\Sigma|$$

$$\frac{\partial}{\partial\mu}\log P(D) = \frac{\partial}{\partial\mu} -\frac{1}{2}(x_i-\mu)^T\cdot\Sigma^{-1}\cdot(x_i-\mu)$$

$$= \sum_i -\Sigma^{-1}\cdot(x_i-\mu) = 0$$

$$\Rightarrow \mu = \frac{\sum_i x_i}{n}$$

$$\frac{\partial}{\partial\Sigma}\log P(D) = \frac{\partial}{\partial\Sigma}\sum_i tr\left[-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right] - \frac{n}{2}\cdot\Sigma^{-T}$$

$$= \frac{\partial}{\partial\Sigma}\sum_i tr\left(-\frac{1}{2}(x_i-\mu)(x_i-\mu)^T\cdot\Sigma^{-1}\right) - \frac{n}{2}\cdot\Sigma^{-T}$$

$$= \sum_i \frac{1}{2}\left(\Sigma^{-T}(x_i-\mu)(x_i-\mu)^T\Sigma^{-T}\right) - \frac{n}{2}\cdot\Sigma^{-T} = 0$$

$$\frac{1}{2}\Sigma^{-1}\cdot\left(\sum_i (x_i-\mu)(x_i-\mu)'\cdot\Sigma^{-1}-n\right)=0$$

$$\hat{\Sigma}=\frac{\sum_i (x_i-\mu)(x_i-\mu)^T}{n}$$

PS1

PS 2.7

(a) $\log p(c) = \log n! - \sum_i \log c_i! + \sum_i c_i \log \pi_i$

Constraint $\sum \pi_i = 1$

$$\nabla \log p(c) = \lambda \nabla \sum \pi_i$$

$$\Rightarrow \quad \frac{c_i}{\pi_i} = \lambda$$

$$\pi_i = \frac{c_i}{\lambda}$$

$$\sum_i \pi_i = 1 \ \& \ \frac{\sum c_i}{\lambda} = 1 \Rightarrow \lambda = n$$

$$\hat{\pi}_i = \frac{c_i}{n}$$

(b) $E(\hat{\pi}_i - \pi_i)$

$= E\left(\frac{c_i}{n}\right) - \pi_i$

$= 0$

$E([\hat{\pi}_i - E(\hat{\pi}_i)]^2)$

$= E\left[\left(\frac{c_i}{n} - \pi_i\right)^2\right]$

$= \frac{1}{n^2} E[(c_i - n\pi_i)^2]$

$= \frac{1}{n^2} \cdot n \pi_i (1-\pi_i)$

$= \frac{\pi_i(1-\pi_i)}{n}$

PS 2.8

(a) $f(\theta) = (y - \Phi^T\theta)^T (y - \Phi^T\theta)$

$= y^T y - y^T \Phi^T \theta - \theta^T \Phi \cdot y + \theta^T \Phi \Phi^T \theta$

$\frac{\partial}{\partial\theta} f(\theta) = -\Phi y - \Phi y + 2\Phi\Phi^T\theta = 0$

$\Phi\Phi^T\theta = \Phi y$

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

(inverse exists if columns of Φ are independent)

proof: $\Phi(\Phi^T x) = 0$

① $\boxed{\Phi \text{ row independent}}$

$\Rightarrow$ $\Phi^T x$ in its null space > orthogonal

$\Phi^T x$ in row space.

$\Phi^T x = 0 \Rightarrow x = 0$ (row, independent)

$\Rightarrow$ $\Phi\Phi^T$ invertible.

Why reduce data dimension?

$\Rightarrow$ make it invertible ☆

(b) $y = \Phi^T \theta + \varepsilon$

$$y \sim \mathcal{N}(f(x_i), \alpha^2)$$

$$\log p(D) = -\frac{1}{2} \cdot \Sigma \left(\frac{y - f(x_i)}{\alpha}\right)^2 + g(\alpha)$$

$$\frac{\partial}{\partial \theta} = 0 \Rightarrow \sum_i (y - f(x_i))^2 = 0$$

$$\Downarrow$$

least square

2.9 $\quad y_i \sim \mathcal{N}(f(x_i), \alpha_i^2)$

(a) $\log p(D) = -\frac{1}{2} \sum_i \left(\frac{y_i - f(x_i, \theta)}{\alpha_i}\right)^2$

$$- \log \sqrt{2\pi} \cdot n - \sum_i \log \alpha_i$$

$$\min_\theta \sum_i \frac{1}{\alpha_i^2} (y_i - f(x_i, \theta))^2$$

(b) $f(\theta) = (y - \Phi^T \theta)^T \cdot W (y - \Phi^T \theta)$

$$= y^T W y - y^T W \Phi^T \theta - \theta^T \Phi W y + \theta^T \Phi W \Phi^T \theta$$

$$\frac{\partial}{\partial \theta} f(\theta) = -\Phi W y - \Phi W y + 2 \Phi W \Phi^T \theta = 0$$

$$\theta = (\Phi W \Phi^T)^{-1} \Phi W y$$

$$\left(w_i = \frac{1}{\alpha_i^2} \right)$$

PS 2.10

(a) $\varepsilon = \text{Laplacian}(0, \lambda)$

$$y_i = \text{laplacian}(f(x_i,\theta), \lambda)$$

$$\log P(D) = \log \prod_i \frac{1}{2\lambda} e^{-\left|\frac{y_i - f(x_i,\theta)}{\lambda}\right|}$$

$$= -n\log 2\lambda - \sum_i \left|\frac{y_i - f(x_i,\theta)}{\lambda}\right|$$

$$\Rightarrow \min_\theta \sum_i |y_i - f(x_i,\theta)|$$

square of exponential

A normal distribution has very thin tails, i.e. probability density drops very rapidly as you move further from the middle, like exp(-x^2). The Laplace distribution has moderate tails [1], going to zero like exp(-|x|).

(ii) It assigns more weight to the tail, hence MLE moves shorter towards the outlier.

(i) L_1 norm assigns roughly $=$ weight to each sample where L_2 norm squares the error and hence, the outliers' weights become larger.

(b)

PS 2.11

(a) $E(\hat{u}) = E(x_1) = u$

(b) $E([\hat{u} - E(\hat{u})]^2)$

$= E[(X_1-\mu)^2]$

$= \alpha^2$

(c) never converge

PS 2-12

(a) $E(\hat{\alpha}^2) = E[\frac{1}{N}\sum(x_i-\hat{\mu})^2]$

$= \frac{1}{N}\sum_i E[(x_i-\hat{\mu})^2]$

$= \frac{1}{N}\sum_i E\left[\left(\frac{(n-1)\cdot x_i - \sum_{j\neq i} x_j}{n}\right)^2\right]$

$= \frac{1}{N^3}\cdot\sum_i E\left([(n-1)x_i]^2 + (\sum_{j\neq i} x_j)^2 - 2(n-1)\cdot x_i\cdot\sum_{j\neq i} x_j\right)$

$= \frac{1}{N^3}\Big(N\cdot(n-1)^2\cdot(\alpha^2+\mu^2) + N\cdot(n-1)\cdot(\alpha^2+\mu^2) + N\cdot(n-1)(n-2)\cdot\mu^2 - 2N(n-1)^2\cdot\mu^2\Big)$

$= \frac{1}{N^3}\left(N^2(N-1)\cdot\alpha^2\right)$

$= \frac{N-1}{N}\alpha^2$

(b) $\hat{\alpha}' = \frac{1}{N-1}\cdot\sum_{i=1}^{N}(x_i-\hat{\mu})^2$

PS 2-13

(a) $E(\tilde{\mu}) = \frac{\alpha}{N} E(\sum_i x_i) = \frac{\alpha}{N} \cdot N \cdot \mu = \alpha\mu$

$$
\begin{aligned}
Var(\tilde{\mu}) &= E[(\tilde{\mu} - E(\tilde{\mu}))^2] \\
&= E\left[\left(\frac{\alpha}{N}\sum x_i - \alpha\mu\right)^2\right] \\
&= \frac{\alpha^2}{N^2} \cdot E\left(\left(\sum (x_i - \mu)\right)^2\right) \\
&= \frac{\alpha^2}{N^2} \cdot N \sigma^2 \\
&= \frac{\alpha^2}{N} \cdot \sigma^2
\end{aligned}
$$

PS 2.14

(a) $E[M_n] = E\left(\frac{\sum x_i}{n}\right) = \frac{1}{n} \cdot n \cdot \mu_x = \mu_x$

$$
\begin{aligned}
Var(M_n) &= E[(M_n - E(M_n))^2] \\
&= E\left[\left(\frac{\sum x_i}{n} - \mu_x\right)^2\right] \\
&= \frac{1}{n^2} \cdot E\left[\left(\sum_i (x_i - \mu_x)\right)^2\right] \\
&= \frac{1}{n^2} \cdot n \cdot \sigma_x^2 = \frac{\sigma_x^2}{n}
\end{aligned}
$$

(b) $P(|M_n - \mu_x| \geq \varepsilon) \leq \frac{\sigma_x^2}{\varepsilon^2 n}$

$\longrightarrow 0$

$$\left(+ \begin{array}{l} \text{Chebyshev inequality} \\ \dfrac{1}{\Theta(n)} \cdot \text{variance} \end{array} \right)$$ ☆

$\Rightarrow$ indicate that coverge to mean