3F3 Example Paper4 Part Solution

Hanchen Wang

Ph.D Candidate in
Engineering, University of Cambridge

January 24, 2020

- **Q1**: check the 'The Matrix Cookbook'[1]
- **Q2**:

  $a)\ \mathbf{x}^T\mathbf{B}\mathbf{x} = \mathbf{G}\mathbf{x}^T\mathbf{G}\mathbf{x}$

  $b)$ this is equally to prove that: $\quad \mathbf{G}\mathbf{x} = 0 \text{ iff } \mathbf{x} = 0 \Rightarrow \mathbf{G}$ is full rank

  $$c)\ \frac{\partial \mathbf{J}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{B}\mathbf{x} - 2\mathbf{b}^T\mathbf{x}) = \left(\mathbf{B} + \mathbf{B}^T\right)\mathbf{x} - 2\mathbf{b} = 0$$

  $$d)(\mathbf{x} - \mathbf{m})^T\mathbf{M}(\mathbf{x} - \mathbf{m}) + C = \mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{m}^T\mathbf{M}\mathbf{m} - \mathbf{m}^T\mathbf{M}\mathbf{x} - \mathbf{x}^T\mathbf{M}\mathbf{m} + C$$
  $$\sim \mathbf{x}^T\mathbf{B}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} \Rightarrow (\mathbf{x} - \mathbf{B}^{-1}\mathbf{b})^T\mathbf{B}(\mathbf{x} - \mathbf{B}^{-1}\mathbf{b}) - \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}$$

  $$e)\ p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \mathcal{N}\left(0, \sigma_e^2\right)\mathcal{N}\left(\mathbf{m}_{\boldsymbol{\theta}}, \mathbf{C}_{\boldsymbol{\theta}}\right)$$

  Proof Omitted(it has been practiced for so many times), it's in lecture note

  - it is worth mentioning that Linear Gaussian model is neither Gaussian Mixture Model(GMM) or Gaussian Process(GP)

[1] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

- **Q3**:

$$\mathcal{L}(x|\theta) = \prod_{i=0}^{N-1} \mathcal{N}(x_i|\mu, \sigma_n^2) = \prod_{i=0}^{N-1} \mathcal{N}(\frac{x-\mu}{\sigma_n}|0,1) \Rightarrow \mu^{(ML)} = \arg\max \log(\mathcal{L}(x|\mu))$$

$$\Rightarrow \mu^{(ML)} = \sum_{i=0}^{N-1} x_i w_i \quad \text{where } w_i = \frac{1}{\sigma_i^2} / \sum_k \frac{1}{\sigma_k^2} \propto \frac{1}{\sigma_i} \Rightarrow \text{larger variance, smaller weight}$$

$$\mu^{(OLS)} = \arg\min \sum_{i=0}^{N-1} (x_i - \mu)^2 \Rightarrow \mu^{(OLS)} = \frac{1}{N} \sum_{i=0}^{N-1} x_i \Rightarrow \text{same weight}$$

  - bias and variance of $\mu^{(OLS)}$:

$$\text{bias}(\mu^{(OLS)}) = \mathbb{E}[\mu^{(OLS)}] - \mu = \frac{1}{N} \sum_i \mathbb{E}[x_i] - \mu = \frac{1}{N} \sum_i (\mu + \mathbb{E}[\sigma_i]) - \mu = 0$$

$$\text{var}(\mu^{(OLS)}) = \mathbb{E}[(\mu^{(OLS)})^2] - \mathbb{E}[\mu^{(OLS)}]^2 = \frac{1}{N^2} \mathbb{E}[(\sum_{i=0}^{N-1} x_i)^2] - \mu^2$$

$$= \frac{1}{N^2} \mathbb{E}[(\sum_{i=0}^{N-1} \sum_{j \neq i} x_i x_j) + \sum_{i=0}^{N-1} x_i^2] - \mu^2$$

$$= \frac{1}{N^2} (N(N-1)\mu^2 + N\mu^2 + \sum_{i=1}^{N-1} \sigma_i^2) - \mu^2 = \frac{1}{N^2} \sum_{i=1}^{N-1} \sigma_i^2$$

- **Q3**:
  - bias and variance of $\mu^{(ML)}$:

$$\text{bias}(\mu^{(ML)}) = \mathbb{E}[\sum_i x_i w_i] - \mu = (\sum_i \frac{1}{\sigma_i^2})/(\sum_k \frac{1}{\sigma_k^2})\mu - \mu = 0$$

$$\begin{aligned}
\text{var}(\mu^{(ML)}) &= \mathbb{E}[(\sum_i x_i w_i)^2] - \mu^2 = \mathbb{E}[(\sum_{i=0}^{N-1}\sum_{j\neq i} w_i w_j x_i x_j) + \sum_{i=0}^{N-1} w_i^2 x_i^2] - \mu^2 \\
&= \sum_{i=0}^{N-1}\sum_{j\neq i} w_i w_j \mathbb{E}[x_i]\mathbb{E}[x_j] + \sum_{i=0}^{N-1} w_i^2 \mathbb{E}[x_i^2] - \mu^2 \\
&= \mu^2 \sum_{i=0}^{N-1}\sum_{j\neq i} w_i w_j + \sum_{i=0}^{N-1} w_i^2(\mu^2 + \sigma_i^2) - \mu^2 \\
&= \mu^2 \sum_i \sum_j w_i w_j + \sum_{i=0}^{N-1} w_i^2 \sigma_i^2 - \mu^2 = \sum_{i=0}^{N-1} w_i^2 \sigma_i^2 = 1/\left(\sum_{i=0}^{N-1} 1/\sigma_i^2\right)
\end{aligned}$$

- both are unbiased estimator, the variance of the ML estimator depends on the smallest variance, $\min(\sigma_i^2)$, while that of the OLS estimator depends on all the variance, $\sum(\sigma_i^2)$.
- OLS estimator is sensitive to outliers in terms of the larger variance of inferred variable when encountering outliers, while ML estimator is more robust with the same criteria[2]

[2] actually ML method is robust to outliers in the response variable(here is Y), but turned out not to be resistant to outliers in the explanatory variables(here is X)(e.g., leverage points)

- **Q4**:

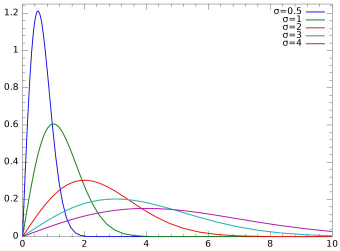
$$a\sin(n\Omega) + b\cos(n\Omega) = \sqrt{a^2+b^2}\sin(n\Omega+\phi) \qquad \text{where } \phi = \arctan(b/a)$$

$\Rightarrow$ to derive $P_{R,\theta}(r,\theta)$, we first derive the Jacobian, $\mathcal{J}$

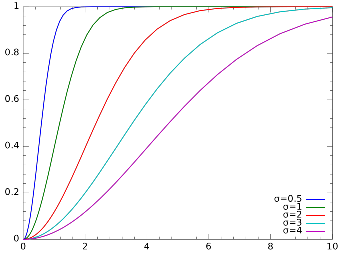
$$\Rightarrow \mathcal{J} = \begin{vmatrix} \partial a/\partial r & \partial a/\partial\theta \\ \partial b/\partial r & \partial b/\partial\theta \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r = \sqrt{a^2+b^2}$$

$$\Rightarrow f_{R,\theta}(r,\theta) = f_{A,B}(r\cos\theta, r\sin\theta) \times r = \frac{r}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}r^2\right)$$

$$\Rightarrow f_R(r) = \int_0^{2\pi} f_{R,\theta}(r,\theta)\mathrm{d}\theta = \frac{r}{\sigma^2}\exp\left(-\frac{1}{2\sigma^2}r^2\right) \qquad \text{Rayleigh Distribution}$$



(a) Rayleigh PDF



(b) Rayleigh CDF

Figure: PDF and CDF of Rayleigh, source: wikipedia

- **Q5**
  - it is straight to calculate the expectation, θ, and variance, θ^2, thus omit
  - just to clarify, t_i is the time interval between i-th and i+1-th trade

$$\mathcal{L} = p(\mathbf{t}|\theta) = \prod_i p(t_i|\theta) = \theta^{-N} \exp\left(-\sum_{i=1}^{N} t_i/\theta\right)$$

$$\theta^{(ML)} = \arg\max_{\theta} \mathcal{L} = \sum_{i=1}^{N} t_i/N$$

$$\mathrm{bias}(\theta^{(ML)}) = \mathbb{E}[\theta^{(ML)}] - \theta = \sum_{i=1}^{N} \theta/N - \theta = 0$$

$$\mathrm{var}(\theta^{(ML)}) = \mathbb{E}[(\theta^{(ML)})^2] - \theta^2 = \theta^2/N$$

  - for the posterior:

$$p(\theta|\mathbf{t}) \propto p(\mathbf{t}|\theta)p(\theta) = \theta^{-(N+2)} \exp\left(-\frac{\sum_{i=1}^{N} t_i + 1}{\theta}\right)$$

$$\Leftrightarrow \text{two additional observations, } t' \text{ and } t'', \text{ where } t' + t'' = 1$$

$$\Rightarrow \theta^{(MAP)} = \arg\max_{\theta} p(\theta|\mathbf{t}) = \frac{1 + \sum_{i=1}^{N} t_i}{N+2}$$

  - when N is very large, likelihood term dominates the posterior, thus $\theta^{(ML)} \approx \theta^{(MAP)}$

- **Q6**

$$
\begin{aligned}
x_n &= a x_{n-1} + e_n \\
\mathcal{L} &= P_{\mathbf{x}}(x_1, x_2, ..., x_{N-1}|a) = P_{\mathbf{x}}(x_{N-1}|x_{N-2}, a) ... P_{\mathbf{x}}(x_1|x_0, a) \\
&= \prod_{i=1}^{N-1} \mathcal{N}(x_i|a x_{i-1}, 1) = \prod_{i=1}^{N-1} \mathcal{N}(x_i - a x_{i-1}|0, 1)
\end{aligned}
$$

$$
\begin{aligned}
a^{(ML)} &= \arg\max_a \log(\mathcal{L}) = \arg\max_a \sum_{i=1}^{N-1} (x_i - a x_{i-1})^2 \\
&= \arg\max_a \sum_{i=1}^{N-1} (x_i^2 + a^2 x_{i-1}^2 - 2 a x_i x_{i-1}) \\
&\Rightarrow a^{(ML)} = \frac{\sum_{n=1}^{N-1} x_{n-1} x_n}{\sum_{n=0}^{N-2} x_n^2}
\end{aligned}
$$

thus we can derive the posterior:

$$
\begin{aligned}
P(a|\mathbf{x}) \propto P(a)P(\mathbf{x}|a) &\propto \exp\left\{-\frac{1}{2}\left[\sum_{n=1}^{N-1}(a x_{n-1} - x_n)^2 + \frac{(a-\mu_a)^2}{\sigma_a^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[a^2\left(\sum_{n=1}^{N-1} x_{n-1}^2 + \frac{1}{\sigma_a^2}\right) - 2a\left(\sum_{n=1}^{N-1} x_n x_{n-1} + \frac{\mu_a}{\sigma_a^2}\right) + f(\mathbf{x})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[a^2\left(\mathcal{S}_2 + \frac{1}{\sigma_a^2}\right) - 2a\left(\mathcal{S}_1 + \frac{\mu_a}{\sigma_a^2}\right) + f(\mathbf{x})\right]\right\} \sim \mathcal{N}\left(a|\hat{\mu}_a, \hat{\sigma}_a^2\right)
\end{aligned}
$$

- **Q6**
  by matching the moments

$$\hat{\sigma}_a^2 = \frac{1}{\mathcal{S}_2 + 1/\sigma_a^2} = 0.005 \qquad \hat{\mu}_a^2 = \frac{\mathcal{S}_1 + \mu_a/\sigma_a^2}{\mathcal{S}_2 + 1/\sigma_a^2} = 0.94$$

probability of unstable filter:

$$P(|a| > 1) = 1 - \Phi\left(\frac{1 - 0.94}{\sqrt{0.005}}\right) + \Phi\left(\frac{-1 - 0.94}{\sqrt{0.005}}\right) = 0.2$$

to sketch the prior, likelihood and posterior:

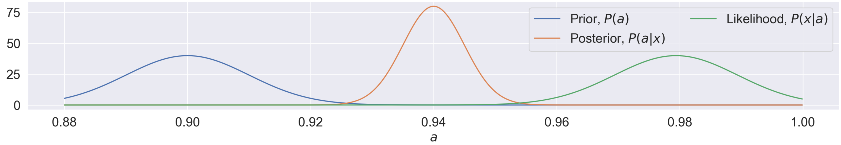


Figure: PDF of Prior, Likelihood and Posterior

note that all of the three are Gaussians, likelihood share the same variance with prior

- **Q6**
  the new posterior is a truncated Gaussian:

$$P(a|\mathbf{x}) = \begin{cases} \frac{1}{1-0.2}\mathcal{N}(a|0.94, 0.005), & \text{if } -1 < a < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$a^{(MAP)} = 0.94$$

For MMSE, require:

$$\begin{aligned}
a^{MMSE} &= \mathbb{E}[a|\mathbf{x}] = \int_{-1}^{+1} aP(a|\mathbf{x})\mathrm{d}a = \frac{1}{0.8}\int_{-1}^{+1} a\mathcal{N}\left(a|\mu,\sigma^2\right)\mathrm{d}a \\
&= \frac{1}{0.8\sqrt{2\pi\sigma^2}}\int_{-1}^{+1} a\exp\left[-\frac{(a-\mu)^2}{2\sigma^2}\right]\mathrm{d}a \qquad \text{set } u = a-\mu \\
&= \frac{1}{0.8\sqrt{2\pi\sigma^2}}\int_{-1-\mu}^{+1-\mu} (u+\mu)\exp\left[-\frac{u^2}{2\sigma^2}\right]\mathrm{d}u \\
&= \frac{1}{0.8\sqrt{2\pi\sigma^2}}\left\{\int_{-1-\mu}^{+1-\mu} u\exp\left[-\frac{u^2}{2\sigma^2}\right]\mathrm{d}u + \int_{-1-\mu}^{+1-\mu} \mu\exp\left[-\frac{u^2}{2\sigma^2}\right]\mathrm{d}u\right\} \\
&= \frac{1}{0.8\sqrt{2\pi\sigma^2}}\left\{\left[-\sigma^2\exp\left(-\frac{u^2}{2\sigma^2}\right)\right]_{-1-\mu}^{1-\mu} + \mu\sqrt{2\pi\sigma^2}\int_{-1-\mu}^{1-\mu}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{u^2}{2\sigma^2}\right)\mathrm{d}u\right\} \\
&= \frac{1.25}{\sqrt{2\pi\sigma^2}}\left\{\sigma^2\exp\left(-\frac{u^2}{2\sigma^2}\right) - \exp\left(-\frac{(1-u)^2}{2\sigma^2}\right)^2\right]\mu\sqrt{2\pi\sigma^2}\left[\Phi\left(\frac{1-\mu}{2\sigma^2}\right) - \Phi\left(\frac{-1-\mu}{2\sigma^2}\right)\right]\right\} \\
&= 0.9154 \qquad \text{fairly small impact on the estimate for } a
\end{aligned}$$