Homework 5
Mouselinos Spyridon
February 2020

Exersize 1

(a) Erlang Distribution: $p(x) = \theta^2 x \exp(-\theta x) u(x)$
where $u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$

- Given a set of N, $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ independent, measurements / data points, we can calculate the likelihood as the product of per-point pdf. The joint pdf of $\mathcal{X}$ is:

$$p(\mathcal{X};\theta) = \prod_{i=1}^{N} p(x_i)$$

The log-likelihood is:

$$L(\mathcal{X};\theta) = \ln(p(\mathcal{X};\theta)) = \sum_{i=1}^{N} \ln(p(x_i;\theta))$$

substituting:

$$L(\theta) = \sum_{i=1}^{N} \ln(\theta^2 x_i \exp(-\theta x_i)) \quad \text{for } x_i \geq 0$$

$$L(\theta) = \sum_{i=1}^{N} \left[\ln(\theta^2 x_i) + \ln(\exp(-\theta x_i))\right] = \sum_{i=1}^{N} \left[\ln(\theta^2 x_i) + (-\theta x_i)\right]$$

$$L(\theta) = \sum_{i=1}^{N} \ln(\theta^2) + \ln(x_i) - \theta \sum_{i=1}^{N} x_i = 2N\ln(\theta) + \sum_{i=1}^{N} \ln(x_i) - \theta \sum_{i=1}^{N} x_i$$

in order to find the MLE we need to find θ' such as $\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\theta'} = 0$. So: $\frac{\partial L(\theta)}{\partial \theta} = 0 \Rightarrow$

$$\Rightarrow (2N\ln(\theta))' + 0 + \left(-\theta\sum_{i=1}^{N} x_i\right)' = 0 \Rightarrow \frac{2N}{\theta'} - \sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow \theta_{ML} = \theta' = \frac{2N}{\sum_{i=1}^{N} x_i}$$

(b) Now $N=5$ / $x_1=2$ / $x_2=2.2$ / $x_3=2.7$ / $x_4=2.4$ / $x_5=2.6$

First of all let's plug our values in our equation:

$$\hat{\theta} = \theta_{ML} = \frac{2 \cdot N}{\sum_{i=1}^{N} x_i} = \frac{2 \cdot 5}{2+2{,}2+2{,}7+2{,}4+2{,}6} = \frac{10}{11{,}9} = 0{,}8403$$

The Erlang distribution that best "explains the data" is

$$p(x) = \theta_{ML}^2 x \exp(-\theta_{ML} x) u(x)$$

$$p(x) = (0{,}84)^2 x \exp(-0{,}84 x) u(x)$$

Now in order to calculate the mean of the X RV. we have

$$E[X] = \int_{-\infty}^{+\infty} x\, p(x)\, dx = \int_{0}^{+\infty} x \cdot \theta^2 x \exp(-\theta x)\, dx$$

$$= \theta^2 \int_0^{+\infty} x^2 \exp(-\theta x)\, dx = \theta^2 \frac{2}{\theta^3} = \frac{2}{\theta}$$

If we want to take $\theta = \theta_{ML}$ it would be: $E[X] = \frac{2}{\theta_{ML}} = 2{,}38$

Exersize 1

(c) The pdf value of $x_1' = 2.1$ would be

$$p(x_1') = 0{,}84^2 \, x_1' \exp(-0{,}84 x_1') \cdot 1 = 0{,}253912$$

Same: $x_2' = 2.3$

$$p(x_2') = 0{,}84^2 \, x_2' \exp(-0{,}84 x_2') \cdot 1 = 0{,}235087$$

Same: $x_3' = 2.9$

$$p(x_3') = (0{,}84)^2 \, x_3' \exp(-0{,}84 x_3') \cdot 1 = 0{,}179067$$

The joint probability of $\mathcal{X} = \{x_1', x_2', x_3'\}$ would be:

$$\mathrm{jpdf}(\mathcal{X}) = \mathrm{jpdf}(\mathcal{X}; \theta_{ML}) = \prod_{i=1}^{3} p(x_i'; \theta_{ML}) = p(x_1') p(x_2') p(x_3')$$

$$= 0{,}01068.$$

The log livelihood:

$$L(\theta_{ML}) \text{ on } \mathcal{X} = \ln(0{,}01068) = -4{,}538562.$$

Exersize 2

(a) We now have the same scenario $\overset{\text{Erlang}}{p(x) = \theta^2 x \exp(-\theta x) u(x)}$ but we know a prior-probability for $\theta \sim N_{prior}(\theta_0, \sigma_0^2)$ with known θ_0 / σ_0.

The MAP estimate is simply $\theta_{MAP} = \arg\max \; p(\mathcal{X}|\theta)\, p(\theta)$

Again, assuming independent measurements / data-points (x_i) we have:

- $p(\mathcal{X}|\theta) = \prod_{i=1}^{N} p(x_i;\theta)$ (1)
- $p(\theta) = N(\theta_0, \sigma_0^2)$ (2)

jointly: $\theta_{MAP} = \arg\max \left(\prod_{i=1}^{N} p(x_i;\theta) \; N(\theta_0, \sigma_0^2) \right)$, however

we can monotonically restate the problem as its log equivalent.

$$\theta_{MAP} = \arg\max_{\theta} \sum_{i=1}^{N} \ln(p(x_i|\theta)) + \ln(p(\theta)) \quad (3)$$

From (3):

Let $f(\theta)$ be $f(\theta) = \sum_{i=1}^{N} \ln(p(x_i|\theta)) + \ln(p(\theta))$, then

$$f(\theta) = \ln\left[\frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right)\right] + \sum \ln p(x_i|\theta) = \ln\left[\frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right)\right]$$

$$+ \sum_{i=1}^{N} \ln(x_i) - \theta \sum_{i=1}^{N} x_i + 2N\ln(\theta)$$

Exerzize 2

(a) Taking the derivative wrt θ, we get:

$$\frac{\partial f(\theta)}{\partial \theta} = -\frac{(\theta-\theta_0)}{\sigma_0^2} + \frac{2N}{\theta} - \sum_{x=1}^{N} x_i$$

We want $\theta_{MAP} = \arg\max f(\theta) \Rightarrow \theta'$ such as $\left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\theta'} = 0$

$$\frac{\partial f(\theta)}{\partial \theta} = 0 \Rightarrow \theta^2 + \left(\sigma_0^2 \sum_{i=1}^{N} x_i - \theta_0\right)\theta - 2N\sigma_0^2 = 0$$

That means that:

$$\theta_{MAP} = \theta' = \frac{\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i \pm \sqrt{\left(\sigma_0^2 \sum_{i=1}^{N} x_i - \theta_0\right)^2 + 8N\sigma_0^2}}{2}$$

However not both of theese solutions are correct. to limit down to 1 solution we notice that the "$-$" solution:

$$\theta_{MAP}^{\text{"}-\text{"}} = \frac{1}{2}\left[\theta_0 - \sigma_0^2 \sum x_i - \sqrt{(\sigma_0^2 \sum x_i - \theta_0)^2 + 8N\sigma_0^2}\right]$$

Let $z = \theta_0 - \sigma_0^2 \sum x_i$ then

$$\theta_{MAP}^{\text{"}-\text{"}} = \frac{1}{2}\left[z - \sqrt{z^2 + 8N\sigma_0^2}\right] \leq \frac{1}{2}\left[z - \sqrt{z^2} - \sqrt{8N\sigma_0^2}\right]$$

$$\theta_{MAP}^{\text{"}-\text{"}} \leq \frac{1}{2}\left[z - z - \sqrt{8N\sigma_0^2}\right] \leq \frac{1}{2}\left[-\sqrt{8N\sigma_0^2}\right] \leq 0.$$

Meaning the second solutions leads to a negative value thus we dismiss it and keep the "+" one.

(b) We have that $\theta_{MAP} = \frac{1}{2}\left(\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right) + \sqrt{\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right)^2 + 8N\sigma_0^2}\right)$

(i) In case where $N \to \infty$:

Let's recall the equation that lead us to θ_{MAP}:

$$\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\theta_{MAP}} = 0 \iff -\frac{(\theta_{MAP} - \theta_0)}{\sigma_0^2} + \frac{2N}{\theta_{MAP}} - \sum_{i=1}^{N} x_i = 0 \quad (1)$$

if $N \to \infty$ then by dividing with N:

$$(1) \Rightarrow -\frac{(\theta_{MAP} - \theta_0)}{N\sigma_0^2} + \frac{2}{\theta_{MAP}} - \frac{1}{N}\sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow (N \to +\infty) \quad 0 + \frac{2}{\theta_{MAP}} - \frac{1}{N}\sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow \theta_{MAP} = \frac{2N}{\sum x_i} = \theta_{ML}$$

This derives from the fact that MAP is nothing more than a "tradeoff" between the current data distribution and the prior we are presented with. For largre values of N the "data" distribution has greater effect and leads θ_{MAP} to converge to θ_{ML}.

(ii) In case $\sigma_0^2 \gg$, meaning the confidence of our prior knowledge on the distribution of the data is low we have:

$$\frac{2N}{\theta_{MAP}} - \sum_{i=1}^{N} x_i - \frac{(\theta_{MAP} - \theta_0)}{\sigma_0^2} = 0$$

$\sigma_0^2 \gg$ gives us:

$$\frac{2N}{\theta_{MAP}} = \sum_{i=1}^{N} x_i \Rightarrow \theta_{MAP} = \frac{2N}{\sum_{i=1}^{N} x_i} = \theta_{ML}$$

That means that both in $N \to \infty$ and $\sigma_0^2 \gg$, our MAP estimation will converge to the ML estimation.

(iii) In case $\sigma_0^2 \ll$, meaning that we are very confident about the prior distribution of our data we have:

$$\frac{2N}{\theta_{MAP}} - \sum_{i=1}^{N} x_i - \frac{(\theta_{MAP} - \theta_0)}{\sigma_0^2} = 0$$

$$\frac{2N}{\theta_{MAP}} \sigma_0^2 - \sigma_0^2 \sum_{i=1}^{N} x_i - (\theta_{MAP} - \theta_0) = 0$$

$\sigma_0^2 \ll 0$ gives us:

$$0 - 0 - (\theta_{MAP} - \theta_0) = 0$$

$$\Rightarrow \theta_{MAP} = \theta_0$$

that means our MAP estimation is strongly confident to our prior knowledge and will be the same as the known θ_0, the prior expected value.

(c) Let $\theta_0 = 1.1$ / $\sigma_0^2 = 1$

$N = 5$ with $x_1 = 2$ / $x_2 = 2.2$ / $x_3 = 2.7$ / $x_4 = 2.4$ / $x_5 = 2.6$

As solved in question (a).

$$\theta_{MAP} = \hat{\theta} = \frac{1}{2}\left[z + \sqrt{z^2 + 8N\sigma_0^2}\right]$$

where $z = \theta_0 - \sigma_0^2 \sum x_i = 1{,}1 - 1\sum x_i = 1{,}1 - 11{,}9 = -10{,}8$

$$\theta_{MAP} = \hat{\theta} = \frac{1}{2}\left[-10{,}8 + \sqrt{116{,}64 + 8 \cdot 5}\right] = 0{,}8577$$

The erlang formula that best explains the data is

$$p(x) = \theta_{MAP}^2\, x \exp(-\theta_{MAP} x)\, u(x)$$
$$= (0{,}8577)^2 x \exp(-0{,}8577 x) \quad \forall x \geq 0$$

The mean of RV X is:

$$E[X] = \int_{-\infty}^{+\infty} x\, p(x)\, dx = \int_0^{+\infty} x\, \theta_{MAP}^2\, x \exp(-\theta_{MAP} x)\, dx$$
$$= \theta_{MAP}^2 \frac{2}{\theta_{MAP}^3} = \frac{2}{\theta_{MAP}} = 2{,}33$$

Exersize 3

We have the model $x = \mu + n$ where $(x, \mu, n) \in \mathbb{R}$ and a set of measurements $Y = \{x_1, x_2, x_3, \dots x_N\}$
Assuming prior knowledge on μ, we claim that is close to μ_0. As a ridge regression problem we know that:

$$\min J(\mu) = \sum_{n=1}^{N} (x_n - \mu)^2 \quad \text{subject to } (\mu - \mu_0^2) \leq \rho$$

Assuming Gaussian Prior as well as Gaussian Pdf we can reformulate the Ridge Regression Problem to a MAP/Bayesian Inference one, where the prior acts as a regularizer. Keeping that in mind, we have:
$Y = \{x_1, x_2, \dots x_N\}$ a set of independent observations following an Gaussian pdf, with known mean μ.
The prior knowledge is also of a Gaussian prior with mean μ_0 that lies close to μ. If we didn't have any prior knowledge the equivallent of this problem would be the unconstraint LS where $J(\mu) = \sum_{1}^{N} (x_n - \mu)^2$. However now we know that $(\mu - \mu_0)^2 \leq \rho$, where $\rho <<$. That gives us the constraint LS => Ridge Regression problem:

$$\min L(\mu) = J(\mu) + \lambda\left((\mu - \mu_0)^2 - \rho\right)$$

Let λ be the Lagrange multiplier corresponding to the constraint. Then we look for $\hat{\mu}_{RR}$ such as

$$\operatorname{argmin}_{\hat{\mu}_{RR}} L(\mu)$$

$$\frac{\partial L(\mu)}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{n=1}^{N} (x_n - \mu)^2 + \frac{\partial}{\partial \mu} \lambda \left((\mu - \mu_0)^2 - \rho \right)$$

$$\frac{\partial L(\mu)}{\partial \mu} = 2N\mu - 2\sum_{n=1}^{N} x_n + 2\lambda(\mu - \mu_0)$$

we want

$$\left.\frac{\partial L(\mu)}{\partial \mu}\right|_{\mu = \hat{\mu}_{RR}} = 0 \Rightarrow 2N\hat{\mu}_{RR} - 2\sum x + 2\lambda \mu_{RR} - 2\lambda\mu_0 = 0$$

$$\Rightarrow \hat{\mu}_{RR} = \frac{\sum x_i + \lambda\mu_0}{N + \lambda}$$

Now if we choose $\lambda = \sigma_0^2 / \sigma^2$, where σ_0^2 the variance of the prior and σ^2 the variance of the assumed pdf of the data, we would get

$$\hat{\mu}_{RR} = \hat{\mu}_{MAP} = \hat{\mu}_{BI}$$

(d) The pdf value for $x_1' = 2.1$ is

$$p(x_1') = \theta_{MAP}^2 x_1 \exp(-\theta_{MAP} x_1) = 0{,}255$$

// $x_2' = 2.3$:

$$p(x_2') = 0{,}235$$

// $x_3' = 2.9$:

$$p(x_3') = 0{,}177$$

The joint probability of $X = \{x_1', x_2', x_3'\}$ is

$$jpdf(X) = \prod_{i=1}^{3} p(x_i'; \theta_{MAP}) = 0{,}01060$$

The log likelihood:

$$L(\theta_{MAP}) = -4{,}546267$$

(e) By comparing theese results with those of ex1 we can observe a slight per point difference in the probabilities but not a significant chage. That means the θ_{MAP} is close to θ_{ML} thus the "prior" knowledge is not confident enough to shift θ_{MAP} close to $\theta_0 = 1.1$