

Homework Solution 3

1.

$$p(x | \eta) = \left(\prod_{n=1}^N h(x_n) \right) \exp \left\{ \eta^T \left(\sum_{n=1}^N T(x_n) \right) - NA(\eta) \right\}$$

$$p(\eta) \propto \exp\{\tau^T \eta - \tau_0 A(\eta)\}$$

$$p(\eta | x) \propto \exp \left\{ \left(\tau + \sum_{n=1}^N T(x_n) \right)^T \eta - (\tau_0 + N)A(\eta) \right\}$$

$$\begin{aligned} 0 &= \nabla \int p(\eta | x) d\eta \\ &= \int \nabla p(\eta | x) d\eta \\ &= \int \left(\tau + \sum_{n=1}^N T(x_n) - (\tau_0 + N)\nabla A(\eta) \right) p(\eta | x) d\eta \\ &= \tau + \sum_{n=1}^N T(x_n) - (\tau_0 + N)E[\nabla A(\eta) | x] \end{aligned}$$

which implies

$$E[\nabla A(\eta) | x] = \frac{\tau + \sum_{n=1}^N T(x_n)}{\tau_0 + N}$$

Similarly, using the prior $p(\eta)$ in the place of the posterior $p(\eta | x)$, which simply involves substituting τ for $\tau + \sum_{n=1}^N T(x_n)$, and τ_0 for $\tau_0 + N$, we obtain:

$$E[\nabla A(\eta)] = \frac{\tau}{\tau_0}$$

Putting these two results together, and writing μ for $\nabla A(\eta)$, we have:

$$\begin{aligned} E[\mu | x] &= \frac{\tau + \sum_{n=1}^N T(x_n)}{\tau_0 + N} \\ &= \left(\frac{\tau_0}{\tau_0 + N} \right) \frac{\tau}{\tau_0} + \left(\frac{N}{\tau_0 + N} \right) \frac{1}{N} \sum_{n=1}^N T(x_n) \\ &= \left(\frac{\tau_0}{\tau_0 + N} \right) E[\mu] + \left(\frac{N}{\tau_0 + N} \right) \hat{\mu}_{ML}, \end{aligned}$$

where $\hat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^N T(x_n)$ is the maximum likelihood estimator of μ .

2. (a)

$$\begin{aligned} p(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \ln \sigma\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \left(\frac{\mu^2}{2\sigma^2} + \ln \sigma\right)\right) \end{aligned}$$

Thus the sufficient statistic is:

$$T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

The relationship between the canonical parameters (η_1, η_2) and the parameter (μ, σ) is:

$$\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2}$$

and

$$\mu = -\frac{\eta_1}{2\eta_2}, \quad \sigma^2 = -\frac{1}{2\eta_2}$$

(b) The cumulant generating function

$$\begin{aligned} A(\eta_1, \eta_2) &= A(\mu, \sigma) \\ &= \frac{\mu^2}{2\sigma^2} + \ln \sigma \\ &= -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-2\eta_2) \end{aligned}$$

(c)

$$\begin{aligned} \sigma^2 &= \text{Var}(X) \\ &= E(X^2) - E(X)^2 \\ &= \frac{\partial A}{\partial \eta_2} - \left(\frac{\partial A}{\partial \eta_1}\right)^2 \end{aligned}$$

(d) Since we know $\nabla^2 A(\eta) = \text{Var}(T(X))$, we have

$$\begin{aligned} \sigma^2 &= \text{Var}(X) \\ &= \frac{\partial^2 A(\eta_1, \eta_2)}{\partial \eta_1^2} \end{aligned}$$

3. (a) The Dirichlet density is:

$$p(\theta | \alpha) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \theta_j^{\alpha_j - 1}$$

So the expectation of θ_i is

$$\begin{aligned} E(\theta_i) &= \int_0^1 \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \theta_j^{(\alpha_j-1)} \theta_i d\theta_i \\ &= \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \int_0^1 \left(\prod_{j=1}^{i-1} \theta_j^{(\alpha_j-1)}\right) \theta_i^{\alpha_i} \left(\prod_{j=i+1}^k \theta_j^{(\alpha_j-1)}\right) d\theta_i \end{aligned}$$

The integrand here is an unnormalized Dirichlet density with parameters $\alpha_1, \dots, \alpha_{i-1}, (\alpha_i + 1), \alpha_{i+1}, \dots, \alpha_k$. So the integral here is the reciprocal of the normalization constant for the density.

$$\begin{aligned} E(\theta_i) &= \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \frac{\left(\prod_{j=1}^{i-1} \Gamma(\alpha_j)\right) \Gamma(\alpha_i + 1) \left(\prod_{j=i+1}^k \Gamma(\alpha_j)\right)}{\Gamma\left(1 + \sum_{j=i+1}^k \alpha_j\right)} \\ &= \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right) \Gamma(\alpha_i + 1)}{\Gamma(\alpha_i) \Gamma\left(1 + \sum_{j=1}^k \alpha_j\right)} \end{aligned}$$

Use the fact that $\Gamma(x+1) = x\Gamma(x)$, we get:

$$E(\theta_i) = \frac{\alpha_i}{\sum_{j=1}^k \alpha_j}$$

(b)

$$\begin{aligned} p(\theta | \alpha) &= \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \theta_j^{(\alpha_j-1)} \\ &= \prod_{i=1}^k \theta_i^{-1} \exp\left\{\sum_{i=1}^k \alpha_i \log \theta_i - \left(\log \Gamma\left(\sum_{i=1}^k \alpha_i\right) - \sum_{i=1}^k \Gamma(\alpha_i)\right)\right\} \end{aligned}$$

So it is an exponential family with

$$\begin{aligned} h(\theta) &= \prod_{i=1}^k \theta_i^{-1} \\ T(\theta) &= (\log \theta_1, \dots, \log \theta_k) \\ A(\eta) &= \sum_{i=1}^k \log \Gamma(\eta_i) - \log \Gamma\left(\sum_{i=1}^k \eta_i\right) \end{aligned}$$

where $\eta_i = \alpha_i$. We know that $ET(\theta) = \nabla_{\eta} A(\eta)$. So

$$\begin{aligned} E(\log \theta_i) &= \frac{\partial A(\eta)}{\partial \eta_i} \\ &= \frac{\partial}{\partial \eta_i} \left(\sum_{j=1}^k \log \Gamma(\eta_j) - \log \Gamma\left(\sum_{j=1}^k \eta_j\right) \right) \\ &= \frac{\Gamma'(\eta_i)}{\Gamma(\eta_i)} - \frac{\Gamma'\left(\sum_{j=1}^k \eta_j\right)}{\Gamma\left(\sum_{j=1}^k \eta_j\right)} \end{aligned}$$

4. For each data point X_n , we introduce an auxiliary variable $Z_n = (Z_n^1, \dots, Z_n^k)$, where Z_n^i is 1 if X_n comes from the i th mixture component. We also assume:

$$\begin{aligned}\sigma_i^2 &\sim \text{InvGamma}(a, b) \\ \mu_i &\sim N(\eta, \tau^2) \\ \pi &\sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)\end{aligned}$$

So the joint density is:

$$\begin{aligned}p(X, Z, \mu, \sigma, \pi) &= p(X | Z, \mu, \sigma) \cdot p(Z | \pi) \cdot p(\pi) \cdot p(\mu) \cdot p(\sigma) \\ &= \left(\prod_{n=1}^N \prod_{i=1}^k N(x_n | \mu_i, \sigma_i^2)^{z_n^i} \right) \left(\prod_{n=1}^N \prod_{i=1}^k (\pi_i)^{z_n^i} \right) \cdot \text{Dirichlet}(\pi | \alpha) \\ &\quad \times \prod_{i=1}^k \left(N(\mu_i | \eta, \tau^2) \cdot \text{InvGamma}(\sigma_i^2 | a, b) \right)\end{aligned}$$

Now let's compute the conditional probability for π by dropping the factors that do not include π , we have

$$\begin{aligned}p(\pi | X, Z, \mu, \sigma, \pi_{-j}) &\propto \prod_{n=1}^N \prod_{i=1}^k (\pi_i)^{z_n^i} \cdot \text{Dirichlet}(\pi | \alpha) \\ &\propto \left(\prod_{i=1}^k (\pi_i)^{\sum_{n=1}^N z_n^i} \right) \cdot \prod_{i=1}^k (\pi_i)^{(\alpha_i - 1)} \\ &= \text{Dirichlet} \left(\left(\alpha_1 + \sum_{n=1}^N z_n^1 \right), \dots, \left(\alpha_k + \sum_{n=1}^N z_n^k \right) \right)\end{aligned}$$

Now for μ_j

$$\begin{aligned}p(\mu_j | X, Z, \mu_{-j}, \sigma, \pi) &\propto \left(\prod_{n=1}^N N(x_n | \mu_j, \sigma_j^2)^{z_n^j} \right) \cdot N(\mu_j | \eta, \tau) \\ &\propto \exp \left(-\frac{\sum_{n=1}^N z_n^j (x_n - \mu_j)^2}{2\sigma_j^2} \right) \cdot \exp \left(-\frac{(\mu_j - \eta)^2}{2\tau^2} \right) \\ &\propto \exp \left\{ -\left(\frac{\sum_{n=1}^N z_n^j}{2\sigma_j^2} + \frac{1}{2\tau^2} \right) \cdot \left(\mu_j - \frac{\sum_{n=1}^N z_n^j x_n \tau^2 + \eta \sigma_j^2}{\sum_{n=1}^N z_n^j \tau^2 + \sigma_j^2} \right)^2 \right\} \\ &= N \left(\frac{\sum_{n=1}^N z_n^j x_n \tau^2 + \eta \sigma_j^2}{\sum_{n=1}^N z_n^j \tau^2 + \sigma_j^2}, \frac{1}{\frac{\sum_{n=1}^N z_n^j}{\sigma_j^2} + \frac{1}{\tau^2}} \right)\end{aligned}$$

For σ_j

$$p(\sigma_j | X, Z, \mu, \sigma_{-j}, \pi) \propto \left(\prod_{n=1}^N N(x_n | \mu_j, \sigma_j^2)^{z_n^j} \right) \cdot \text{InvGamma}(\sigma_j | a, b)$$

$$\begin{aligned} &\propto \sigma_j^{-\sum_{n=1}^N z_n^j} \cdot \exp \left\{ -\frac{\sum_{n=1}^N z_n^j (x_n - \mu_j)^2}{2\sigma_j^2} \right\} \cdot \sigma_j^{-2(a+1)} \exp \left\{ -\frac{1}{b\sigma_j^2} \right\} \\ &= \text{InvGamma} \left(a + \frac{1}{2} \sum_{n=1}^N z_n^j, \frac{2b}{b \sum_{n=1}^N z_n^j (x_n - \mu_j)^2 + 2} \right) \end{aligned}$$

Finally the conditional for Z_j

$$\begin{aligned} p(Z_j | X, Z_{-j}, \mu, \sigma, \pi) &\propto \prod_{i=1}^k (\pi_i N(x_n | \mu_i, \sigma_i^2))^{z_j^i} \\ &= \prod_{i=1}^k \left(\frac{\pi_i \cdot N(x_n | \mu_i, \sigma_i^2)}{\sum_{i=1}^k \pi_i \cdot N(x_n | \mu_i, \sigma_i^2)} \right)^{z_j^i} \end{aligned}$$

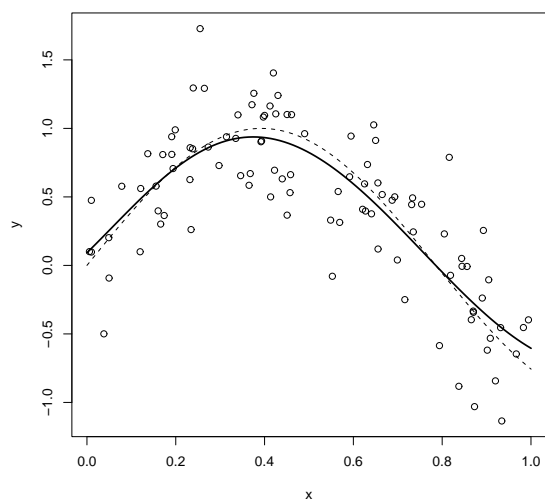


Figure 1: Fitted Curve(solid line), true curve(dash line), with Gaussian kernel as covariance kernel

5. (d) Both Gaussian kernel $k(x, x_i)$ and equivalent kernel $b(x, x_i)$ have large values near x_i . Equivalent kernel $b(x, x_i)$ actually depends on all the data points, whereas Gaussian kernel $k(x, x_i)$ depends only on x_i

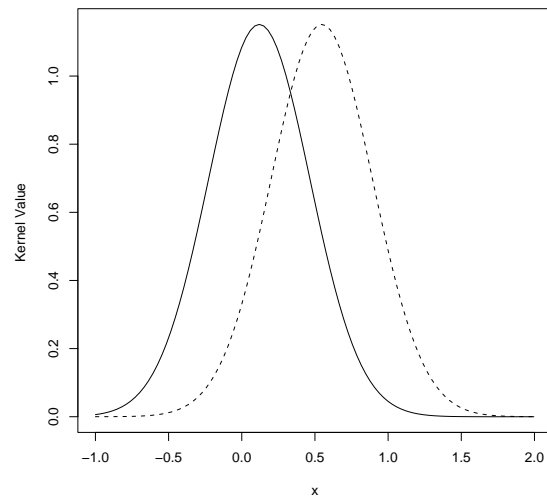


Figure 2: Gaussian kernel function $k(x, x_i)$ evaluated at $x = 0.1210$ and $x = 0.5488$

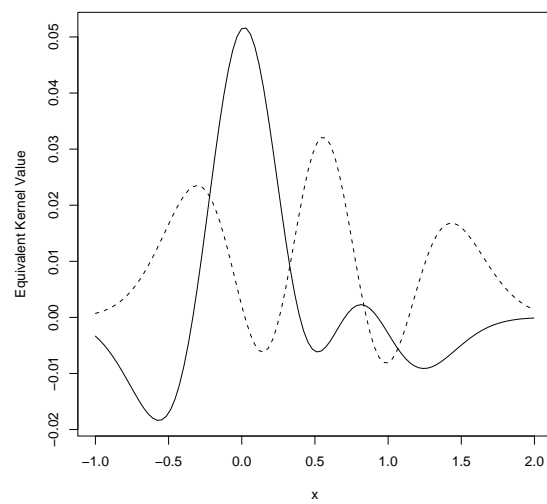


Figure 3: Equivalent kernel function $b(x, x_i)$ evaluated at $x = 0.1210$ and $x = 0.5488$

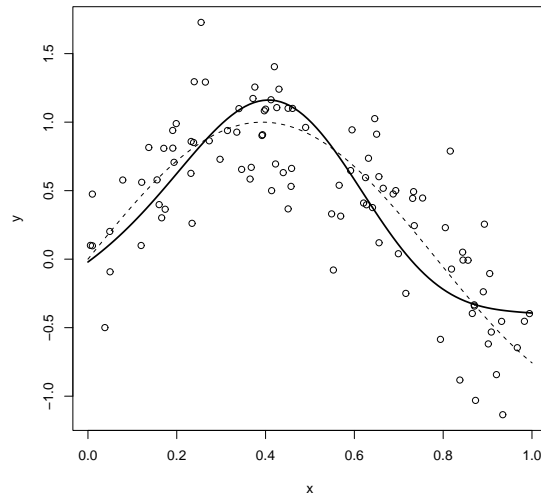


Figure 4: Fitted Curve(solid line), true curve(dash line), with multiquadric kernel as covariance kernel

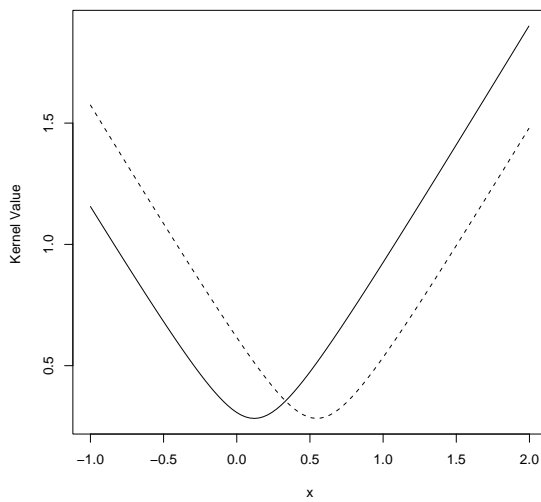


Figure 5: Gaussian kernel function $k(x, x_i)$ evaluated at $x = 0.1210$ and $x = 0.5488$

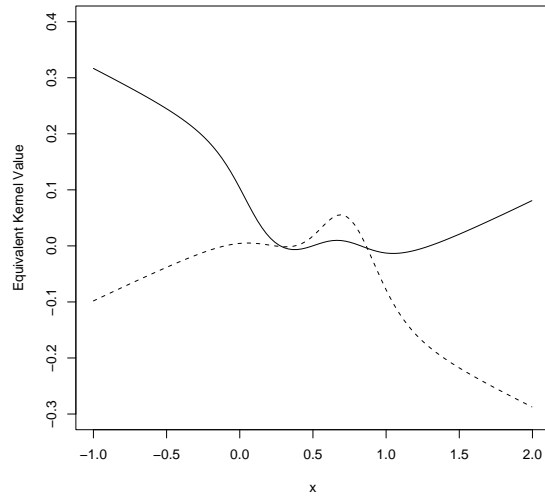


Figure 6: Equivalent kernel function $b(x, x_i)$ evaluated at $x = 0.1210$ and $x = 0.5488$