CS5487 Problem Set

Solutions - Tutorials (1-5)

Antoni Chan
Department of Computer Science
City University of Hong Kong

Tutorial Problems (1-5)

Problem 1.6 Multivariate Gaussian

(a) The multivariate Gaussian distribution is

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}. \tag{S.158}$$

Assuming a diagonal covariance matrix, $\Sigma = \begin{bmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_d^2\end{bmatrix}$, and substituting the properties in (1.16),

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}(\prod_{i=1}^d \sigma_i^2)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \begin{bmatrix}1/\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & 1/\sigma_d^2\end{bmatrix}(x-\mu)\right\} \tag{S.159}$$

$$= \frac{1}{(2\pi)^{d/2}\prod_{i=1}^d \sigma_i} \exp\left\{-\frac{1}{2}\sum_{i=1}^d \frac{1}{\sigma_i^2}(x_i-\mu_i)^2\right\} \tag{S.160}$$

$$= \prod_{i=1}^d \left[\frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left\{-\frac{1}{2}\frac{1}{\sigma_i^2}(x_i-\mu_i)^2\right\}\right] \tag{S.161}$$

$$= \prod_{i=1}^d \mathcal{N}(x_i|\mu_i,\sigma_i^2) \tag{S.162}$$

(b) See Figure 7b. The diagonal terms indicate how far the Gaussian stretches in each axis direction.

(c) See Figure 7a.

(d) The eigenvalues and eigenvector pairs (λ_i, v_i) of Σ satisfy

$$\Sigma v_i = \lambda_i v_i, \quad i \in \{1,\cdots,d\}. \tag{S.163}$$

Rewriting using matrix notation,

$$\Sigma\begin{bmatrix}v_1 & \cdots & v_d\end{bmatrix} = \begin{bmatrix}\lambda_1 v_1 & \cdots & \lambda_d v_d\end{bmatrix} \tag{S.164}$$

$$\Sigma\begin{bmatrix}v_1 & \cdots & v_d\end{bmatrix} = \begin{bmatrix}v_1 & \cdots & v_d\end{bmatrix}\cdot\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_d\end{bmatrix} \tag{S.165}$$

$$\Sigma V = V\Lambda, \tag{S.166}$$

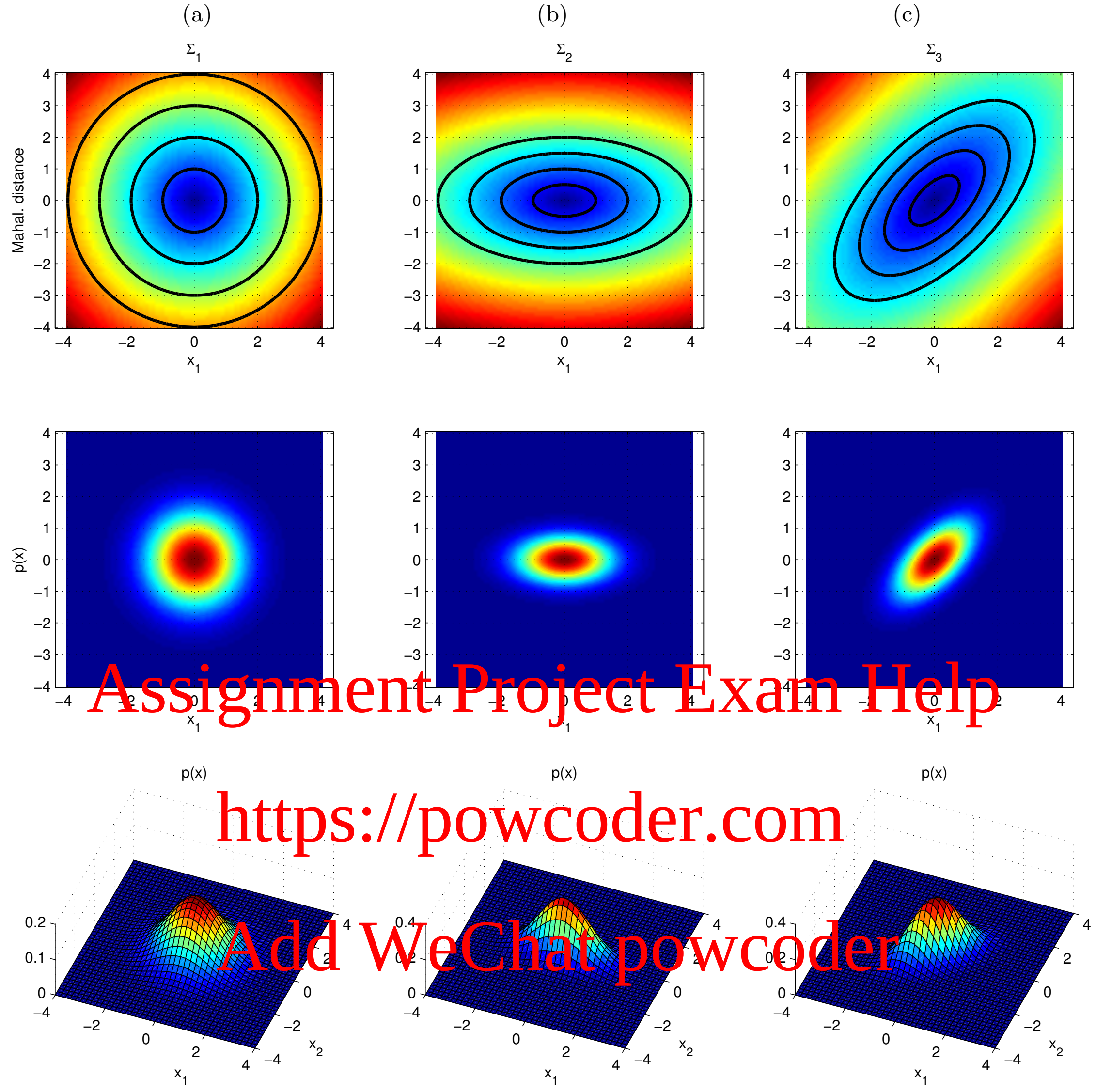

Figure 7: Example of multivariate Gaussians: a) isotropic covariance matrix; b) diagonal covariance matrix; c) full covariance matrix.

where $V = \begin{bmatrix} v_1 & \cdots & v_d \end{bmatrix}$ is a matrix of eigenvectors, and $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$ is a diagonal matrix with the corresponding eigenvalues. Finally, post-multiplying both sides by V^{-1},

$$\Sigma V V^{-1} = V \Lambda V^{-1} \tag{S.167}$$

$$\Sigma = V \Lambda V^T, \tag{S.168}$$

where in the last line we have used the property that the eigenvectors are orthogonal, i.e., $V^T V = I$ and hence $V^{-1} = V^T$.

(e) The inverse of the covariance matrix is

$$\Sigma^{-1} = (V\Lambda V^T)^{-1} = V^{-T}\Lambda^{-1}V^{-1} = V\Lambda^{-1}V^T. \tag{S.169}$$

Hence, the Mahalanobis distance term can be rewritten as

$$\|x-\mu\|_\Sigma^2 = (x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)V\Lambda^{-1}\underbrace{V^T(x-\mu)}_{y} = y\Lambda^{-1}y = \|y\|_\Lambda^2\,, \tag{S.170}$$

where we define $y = V^T(x-\mu)$.

(f) In the transformation $x = Vy + \mu$, first y is rotated according to the eigenvector matrix V, then the result is translated by vector μ.

(g) See Figure 7c. For this Σ, we have $V = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{-\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$ and $\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix}$. The Gaussian is first stretched according to the eigenvalues (e.g., like in Figure 7b). Then it is rotated to match the directions of the eigenvectors.

....

Problem 1.8 Product of Multivariate Gaussian Distributions

The product of two Gaussians is

$$\mathcal{N}(x|a,A)\mathcal{N}(x|b,B) \tag{S.171}$$

$$= \frac{1}{(2\pi)^{d/2}\,|A|^{1/2}}\exp\left\{-\frac{1}{2}(x-a)^TA^{-1}(x-a)\right\}\frac{1}{(2\pi)^{d/2}\,|B|^{1/2}}\exp\left\{-\frac{1}{2}(x-b)^TB^{-1}(x-b)\right\} \tag{S.172}$$

$$= \frac{1}{(2\pi)^{d}\,|A|^{1/2}\,|B|^{1/2}}\exp\left\{-\frac{1}{2}\left[(x-a)^TA^{-1}(x-a)+(x-b)^TB^{-1}(x-b)\right]\right\}. \tag{S.173}$$

Looking at the exponent term, we expand and collect terms,

$$M = (x-a)^TA^{-1}(x-a)+(x-b)^TB^{-1}(x-b) \tag{S.174}$$

$$= x^TA^{-1}x + x^TB^{-1}x - 2a^TA^{-1}x - 2b^TB^{-1}x + a^TA^{-1}a + b^TB^{-1}b \tag{S.175}$$

$$= x^T\underbrace{(A^{-1}+B^{-1})}_{\mathbf{A}}x - 2\underbrace{(a^TA^{-1}+b^TB^{-1})}_{\mathbf{b^T}}x + \underbrace{a^TA^{-1}a+b^TB^{-1}b}_{\mathbf{c}}. \tag{S.176}$$

Using the above definitions of $(\mathbf{A},\mathbf{b},\mathbf{c})$ we complete the square using Problem 1.10:

$$\mathcal{M} = (x-\mathbf{d})^T\mathbf{A}(x-\mathbf{d}) + \mathbf{e}, \tag{S.177}$$

where

$$\mathbf{d} = \mathbf{A}^{-1}\mathbf{b} = (A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b) \tag{S.178}$$

and

$$\begin{aligned}
\mathbf{e} &= \mathbf{c} - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} && \text{(S.179)}\\
&= a^TA^{-1}a + b^TB^{-1}b - (a^TA^{-1} + b^TB^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a + B^{-1}b) && \text{(S.180)}\\
&= a^TA^{-1}a - a^TA^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a \\
&\quad + b^TB^{-1}b - b^TB^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b && \text{(S.181)}\\
&\quad - 2a^TA^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b \\
&= a^T(A+B)^{-1}a + b^T(A+B)^{-1}b - 2a^T(A+B)^{-1}b && \text{(S.182)}\\
&= (a-b)^T(A+B)^{-1}(a-b), && \text{(S.183)}
\end{aligned}$$

where in (S.182) we use the matrix inversion lemma on the first two terms (from Problem 1.15). Finally, defining $C = (A^{-1}+B^{-1})^{-1}$ and $c = C(A^{-1}a + B^{-1}b)$, we obtain for the exponent term

$$\mathcal{M} = (x-c)^TC^{-1}(x-c) + (a-b)^T(A+B)^{-1}(a-b) = \|x-c\|^2_C + \|a-b\|^2_{A+B}\,. \tag{S.184}$$

Next, we look at the determinant term,

$$\begin{aligned}
\mathcal{D} &= \frac{1}{|A|^{1/2}\,|B|^{1/2}} = \frac{1}{|A|^{1/2}\,|B|^{1/2}}\frac{|C|^{1/2}}{|C|^{1/2}} && \text{(S.185)}\\
&= \frac{1}{|A|^{1/2}\,|B|^{1/2}}\frac{\left|(A^{-1}+B^{-1})^{-1}\right|^{1/2}}{|C|^{1/2}} && \text{(S.186)}\\
&= \frac{1}{|A|^{1/2}\,|B|^{1/2}\,|A^{-1}+B^{-1}|^{1/2}}\frac{1}{|C|^{1/2}} && \text{(S.187)}\\
&= \frac{1}{(|A|\,|A^{-1}+B^{-1}|\,|B|)^{1/2}}\frac{1}{|C|^{1/2}} && \text{(S.188)}\\
&= \frac{1}{|A+B|^{1/2}}\frac{1}{|C|^{1/2}} && \text{(S.189)}
\end{aligned}$$

Finally, substituting the derived expressions for $\mathcal{M}$ and $\mathcal{D}$ into (S.173)

$$\begin{aligned}
&\mathcal{N}(x|a,A)\mathcal{N}(x|b,B) && \text{(S.190)}\\
&= \frac{1}{(2\pi)^d\,|A+B|^{1/2}}\frac{1}{|C|^{1/2}}\exp\left\{-\frac{1}{2}\left[\|x-c\|^2_C + \|a-b\|^2_{A+B}\right]\right\} && \text{(S.191)}\\
&= \left[\frac{1}{(2\pi)^{d/2}\,|A+B|^{1/2}}\exp\left\{-\frac{1}{2}\|a-b\|^2_{A+B}\right\}\right]\left[\frac{1}{(2\pi)^{d/2}\,|C|^{1/2}}\exp\left\{-\frac{1}{2}\|x-c\|^2_C\right\}\right] && \text{(S.192)}\\
&= \mathcal{N}(a|b,A+B)\mathcal{N}(x|c,C) && \text{(S.193)}
\end{aligned}$$

....

Problem 1.10 Completing the square

The original form is

$$f(x) = x^TAx - 2x^Tb + c. \tag{S.194}$$

Now expand the desired form,

$$f(x) = (x-d)^T A(x-d) + e = \underbrace{x^T Ax}_{\text{quadratic}} - \underbrace{2x^T Ad}_{\text{linear}} + \underbrace{d^T Ad + e}_{\text{constant}}. \tag{S.195}$$

We need to match the quadratic, linear, and constant terms. The quadratic term already matches. For the linear term, we have by inspection $d = A^{-1}b$ so that $x^T Ad = x^T b$. For the constant term, we set

$$c = d^T Ad + e \quad \Rightarrow \quad e = c - d^T Ad \tag{S.196}$$

$$= c - b^T A^{-1} AA^{-1} b = c - b^T A^{-1} b \tag{S.197}$$

. . . .

Problem 2.6 MLE for a multivariate Gaussian

(a) The log-likelihood of the data is

$$\ell = \log p(X) = \sum_{i=1}^N \log p(x_i) = \sum_{i=1}^N \left\{ -\frac{1}{2} \|x_i - \mu\|_\Sigma^2 - \frac{1}{2} \log|\Sigma| - \frac{d}{2} \log 2\pi \right\} \tag{S.198}$$

$$= -\frac{1}{2} \sum_{i=1}^N \|x_i - \mu\|_\Sigma^2 - \frac{N}{2} \log|\Sigma| \tag{S.199}$$

$$= -\frac{1}{2} \sum_{i=1}^N (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) - \frac{N}{2} \log|\Sigma|, \tag{S.200}$$

where we have dropped constant terms. To find the maximum of the log-likelihood ℓ w.r.t. μ, we take the derivative and set to 0. One way to proceed is to expand the quadratic term and take the derivative. Alternatively, we can use the chain rule. First, removing terms that do not depend on μ,

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{2} \sum_{i=1}^N \frac{\partial}{\partial \mu} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu). \tag{S.201}$$

Letting $z_i = x_i - \mu$ and applying the chain rule $\frac{\partial f}{\partial \mu} = \frac{\partial f}{\partial z} \frac{\partial z}{\partial \mu}$,

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{2} \sum_{i=1}^N \frac{\partial}{\partial \mu} z_i^T \Sigma^{-1} z_i = -\frac{1}{2} \sum_{i=1}^N \left[\frac{\partial}{\partial z_i} z_i^T \Sigma^{-1} z_i \right] \left[\frac{\partial z_i}{\partial \mu} \right] \tag{S.202}$$

$$= -\frac{1}{2} \sum_{i=1}^N \left[2\Sigma^{-1} z_i \right] [-1] = \sum_{i=1}^N \Sigma^{-1} (x_i - \mu). \tag{S.203}$$

Setting the derivative to 0, and solving for μ,

$$\sum_{i=1}^N \Sigma^{-1} (x_i - \mu) = 0. \tag{S.204}$$

Pre-multiplying both sides by Σ,

$$\Sigma \sum_{i=1}^{N} \Sigma^{-1}(x_i - \mu) = \Sigma 0 \quad \Rightarrow \quad \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = 0 \tag{S.205}$$

$$\Rightarrow \quad \sum_{i=1}^{N} x_i - N\mu = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{S.206}$$

(b) To find the maximum of the log-likelihood w.r.t. Σ, we first use the "trace" trick, $x^T A x = \mathrm{tr}[x^T A x] = \mathrm{tr}[A x x^T]$, on the log-likelihood,

$$\ell = -\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) - \frac{N}{2}\log|\Sigma| \tag{S.207}$$

$$= -\frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}[\Sigma^{-1}(x_i - \mu)(x_i - \mu)^T] - \frac{N}{2}\log|\Sigma|, \tag{S.208}$$

$$= -\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T\right] - \frac{N}{2}\log|\Sigma|. \tag{S.209}$$

Taking the derivative w.r.t. Σ (using the helpful derivatives provided in Problem 2.6) and setting to 0,

$$\frac{\partial \ell}{\partial \Sigma} = -\frac{1}{2}\left\{-\Sigma^{-1}\left[\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T\right]\Sigma^{-1}\right\} - \frac{N}{2}\Sigma^{-1} = 0. \tag{S.210}$$

Pre-multiplying and post-multiplying by Σ on both sides gives

$$\frac{1}{2}\left[\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T\right] - \frac{N}{2}\Sigma = 0 \tag{S.211}$$

$$\Rightarrow \quad \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T. \tag{S.212}$$

. . . .

Problem 2.8 Least-squares regression and MLE

(a) We want to find the θ that minimizes the sum-squared error,

$$E = \sum_{i=1}^{N}(y_i - \phi(x_i))^2 = \left\|y - \Phi^T\theta\right\|^2 = (y - \Phi^T\theta)^T(y - \Phi^T\theta) \tag{S.213}$$

$$= y^T y - 2y^T\Phi^T\theta + \theta^T\Phi\Phi^T\theta. \tag{S.214}$$

Next, take the derivative w.r.t. θ (using the derivatives from Problem 2.6), and setting to zero,

$$\frac{\partial E}{\partial \theta} = -2\Phi y + 2\Phi\Phi^T\theta = 0 \tag{S.215}$$

$$\Rightarrow \quad \Phi\Phi^T\theta = \Phi y \quad \Rightarrow \quad \hat{\theta} = (\Phi\Phi^T)^{-1}\Phi y. \tag{S.216}$$

(b) The noise ϵ is zero-mean Gaussian, and hence $y_i = f(x_i) + \epsilon$ is also Gaussian with mean equal to the function value $f(x_i) = \phi(x_i)^T\theta$,

$$p(y_i|x_i, \theta) = \mathcal{N}(y_i|\phi(x_i)^T\theta, \sigma^2). \tag{S.217}$$

Then the log-likelihood of the data is

$$\ell = \sum_{i=1}^{N} \log p(y_i|x_i\theta) = \sum_{i=1}^{N} \left\{ -\frac{1}{2\sigma^2}(y_i - \phi(x_i)^T\theta)^2 - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi \right\}. \tag{S.218}$$

The maximum likelihood solution is then

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \ell = \operatorname*{argmax}_{\theta} -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \phi(x_i)^T\theta)^2 = \operatorname{argmin}\sum_{i=1}^{N}(y_i - \phi(x_i)^T\theta)^2, \tag{S.219}$$

where the last step follows from $\frac{-1}{2\sigma^2}$ being a scalar constant w.r.t. θ. Hence, the maximum likelihood solution is equivalent to the least-squares solution. This is mainly because we assumed Gaussian noise, which introduces the squared-error term.

....

Problem 3.10 Bayesian regression with Gaussian prior

(a) The posterior distribution of the parameters is obtained using Bayes' rule,

$$p(\theta|y, X) = \frac{p(y, |X, \theta)p(\theta)}{\int p(y|X, \theta)p(\theta)d\theta}. \tag{S.220}$$

Here the denominator ensures that the posterior is properly normalized (integrates to 1). Note that the denominator is only a function of the data $\mathcal{D}$ since the parameter θ is integrated out. Hence, it suffices to find the form of θ in the numerator first, and then normalize that equation to obtain the distribution,

$$p(\theta|y, X) \propto p(y|X, \theta)p(\theta) \tag{S.221}$$

or equivalently

$$\log p(\theta|y, X) = \log p(y|X, \theta) + \log p(\theta) + \text{const.} \tag{S.222}$$

Substituting for the data likelihood and prior terms (and ignoring terms not involving θ),

$$\begin{align} \log p(\theta|y, X) &= \log \mathcal{N}(y|\Phi^T\theta, \Sigma) + \log \mathcal{N}(\theta|0, \Gamma) + \text{const.} \tag{S.223} \\ &= -\frac{1}{2}\left\|y - \Phi^T\theta\right\|_{\Sigma}^{2} - \frac{1}{2}\theta^T\Gamma^{-1}\theta + \text{const.} \tag{S.224} \\ &= -\frac{1}{2}\left(-2\theta^T\Phi\Sigma^{-1}y + \theta^T\Phi\Sigma^{-1}\Phi^T\theta\right) - \frac{1}{2}\theta^T\Gamma^{-1}\theta + \text{const.} \tag{S.225} \\ &= -\frac{1}{2}[\theta^T \underbrace{\left(\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1}\right)}_{A}\theta - 2\theta^T \underbrace{\Phi\Sigma^{-1}y}_{b}] + \text{const.} \tag{S.226} \end{align}$$

Next, using the above A and b, we complete the square (see Problem 1.10),

$$\begin{aligned}\log p(\theta|y,X) &= -\frac{1}{2}(\theta - A^{-1}b)^T A(\theta - A^{-1}b) + \text{const.} && \text{(S.227)}\\ &= -\frac{1}{2}\left\|\theta - \hat{\mu}_\theta\right\|^2_{\hat{\Sigma}_\theta} + \text{const.} && \text{(S.228)}\end{aligned}$$

where again constant terms are ignored, and we define

$$\begin{aligned}\hat{\mu}_\theta &= A^{-1}b = (\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})^{-1}\Phi\Sigma^{-1}y, && \text{(S.229)}\\ \hat{\Sigma}_\theta &= A^{-1} = (\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})^{-1}. && \text{(S.230)}\end{aligned}$$

Finally, note that the log-posterior in (S.228) is of the same form of a Gaussian for θ. Hence the posterior is Gaussian,

$$p(\theta|y,X) = \mathcal{N}(\theta|\hat{\mu}_\theta, \hat{\Sigma}_\theta) \qquad \text{(S.231)}$$

(b) The MAP solution is the mean of the Gaussian (i.e., the θ with largest likelihood),

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname*{argmax}_\theta \mathcal{N}(\theta|\hat{\mu}_\theta, \hat{\Sigma}_\theta) && \text{(S.232)}\\ &= \hat{\mu}_\theta = (\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})^{-1}\Phi\Sigma^{-1}y. && \text{(S.233)}\end{aligned}$$

The $\hat{\theta}_{MAP}$ is similar to the weighted least-squares estimate, but has an additional term Γ^{-1}. When $\Gamma^{-1} = 0$, then $\hat{\theta}_{MAP}$ is the same as the weighted least-squares solution. For non-zero values of Γ, then the term serves to regularize the covariance matrix $\Phi\Sigma^{-1}\Phi^T$, which might not be strictly positive definite or nearly singular. E.g., if $\Gamma = I$ is a diagonal matrix, then this would guarantee that the matrix inverse of $(\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})$ can always be performed.

(c) Substituting for $\Gamma = \alpha I$ and $\Sigma = \sigma^2 I$ in $\hat{\theta}_{MAP}$,

$$\hat{\theta} = (\Phi(\tfrac{1}{\sigma^2}I)\Phi^T + \tfrac{1}{\alpha}I)^{-1}\Phi\tfrac{1}{\sigma^2}Iy = (\Phi\Phi^T + \tfrac{\sigma^2}{\alpha}I)^{-1}\Phi y = (\Phi\Phi^T + \lambda I)^{-1}\Phi y, \qquad \text{(S.234)}$$

where $\lambda = \frac{\sigma^2}{\alpha} \geq 0$.

To solve the regularized least-squares problem, first consider the objective function

$$\begin{aligned}R = \left\|y - \Phi^T\theta\right\|^2 + \lambda\left\|\theta\right\|^2 &= (y - \Phi^T\theta)^T(y - \Phi^T\theta) + \lambda\theta^T\theta && \text{(S.235)}\\ &= y^Ty - 2y^T\Phi^T\theta + \theta^T\Phi\Phi^T\theta + \lambda\theta^T\theta && \text{(S.236)}\\ &= y^Ty - 2y^T\Phi^T\theta + \theta^T(\Phi\Phi^T + \lambda I)\theta. && \text{(S.237)}\end{aligned}$$

Taking the derivative and setting to 0,

$$\begin{aligned}&\frac{\partial R}{\partial\theta} = -2\Phi y + 2(\Phi\Phi^T + \lambda I)\theta = 0 \quad \Rightarrow \quad (\Phi\Phi^T + \lambda I)\theta = \Phi y && \text{(S.238)}\\ &\Rightarrow \quad \theta = (\Phi\Phi^T + \lambda I)^{-1}\Phi y && \text{(S.239)}\end{aligned}$$

Hence the regularized least-squares estimate (aka ridge regression) is equivalent to the above MAP estimate for Bayesian regression with Gaussian noise and prior with isotropic covariances.

(d) Substituting for $\Gamma = \alpha I$ and $\Sigma = \sigma^2 I$, the posterior mean and covariance are

$$\hat{\mu}_\theta = (\Phi\Phi^T + \tfrac{\sigma^2}{\alpha} I)^{-1}\Phi y, \qquad \hat{\Sigma}_\theta = (\tfrac{1}{\sigma^2}\Phi\Phi^T + \tfrac{1}{\alpha} I)^{-1}. \tag{S.240}$$

Here are the various cases of interest:

- Setting $\alpha \to 0$ corresponds to setting a very strong prior at $\theta = 0$, since the covariance of the prior will be $\Gamma = 0$. The term $\frac{1}{\alpha} I$ is a diagonal matrix with very large entries, and its inverse is a matrix that is zero. As a result, the posterior of θ is equivalent to the prior, $\hat{\mu}_\theta = 0$ and $\hat{\Sigma}_\theta = 0$.
- Setting $\alpha \to \infty$ yields a very weak prior since the covariance Γ is very large. As a result, the the $\frac{1}{\alpha} I$ term vanishes, leaving an estimate equivalent to ordinary least squares, $\hat{\mu}_\theta = (\Phi\Phi^T)^{-1}\Phi y$ and $\hat{\Sigma}_\theta = (\frac{1}{\sigma^2}\Phi\Phi^T)^{-1}$.
- When $\sigma^2 \to 0$, then this means there is no observation noise. As a result, the prior is ignored and the data term dominates the mean $\hat{\mu}_\theta = (\Phi\Phi^T)^{-1}\Phi y$, resulting in the ordinary least square estimate. The posterior covariance is $\hat{\Sigma}_\theta = 0$, since there is no uncertainty in the observations.
- When $\sigma^2 \to \infty$ this corresponds to observations being very very noisy. As a result, the data term is ignored and the prior dominates, resulting a posterior equivalent to the prior, $\hat{\mu}_\theta = 0$ and $\hat{\Sigma}_\theta = \alpha I$.

(e) Given a novel input x_*, we are interested in the function value f_* conditioned on the data $\{y, X, x_*\}$. Note that the posterior $\theta|y, X$ is a Gaussian random variable. Hence, when conditioning on the data, $f_* = \phi(x_*)^T\theta$ is a linear transformation of a Gaussian random variable, which is also Gaussian. Using the properties in Problem 1.1, which give the mean and variance of the transformed Gaussian, we have

$$p(f_*|y, X, x*) = \mathcal{N}(f_*|\hat{\mu}_*, \hat{\sigma}_*^2), \quad \hat{\mu}_* = \phi(x_*)^T\hat{\mu}_\theta, \quad \hat{\sigma}_*^2 = \phi(x_*)^T\hat{\Sigma}_\theta\phi(x_*). \tag{S.241}$$

Finally, for the predicted y_*, we have obtain the distribution by integrating over f_*,

$$p(y_*|y, X, x_*) = \int p(y_*|f_*)p(f_*|y, X, x_*)df_* = \int \mathcal{N}(y_*|f_*, \sigma^2)\mathcal{N}(f_*|\hat{\mu}_*, \hat{\sigma}_*^2)df_* \tag{S.242}$$

$$= \int \mathcal{N}(f_*|y_*, \sigma^2)\mathcal{N}(f_*|\hat{\mu}_*, \hat{\sigma}_*^2)df_* = \mathcal{N}(y_*|\hat{\mu}_*, \hat{\sigma}_*^2 + \sigma^2). \tag{S.243}$$

where the last line uses Problem 1.9 to calculate the integral (correlation between Gaussian distributions).

. . . .

Problem 3.12 L1-regularized least-squares (LASSO)

(a) The L1-regularized least-squares objective function is

$$E = \frac{1}{2}\left\|y - \Phi^T\theta\right\|^2 + \lambda\sum_{i=1}^{D}|\theta_i|. \tag{S.244}$$

The first term (data term) is the squared-error as in ordinary least squares. The second term (regularization term) is the absolute value of the parameter values. In contrast, regularized least squares (ridge regression) uses the norm of the parameter $\|\theta\|^2$. The objective function E is equivalent to the negative log-likelihood, so the likelihood takes the form of

$$\ell \propto e^{-E} = e^{-\left\|y-\Phi^T\theta\right\|^2 - \lambda\sum_{i=1}^{D}|\theta_i|} = e^{-\sum_{i=1}^{N}(y_i-\phi(x_i)^T\theta)^2 - \lambda\sum_{i=1}^{D}|\theta_i|} \tag{S.245}$$

$$= \left[\prod_{i=1}^{N} e^{-(y_i-\phi(x_i)^T\theta)^2}\right]\left[\prod_{i=1}^{D} e^{-\lambda|\theta_i|}\right]. \tag{S.246}$$

The left term is the data likelihood, which is equivalent to a Gaussian (as in ordinary least squares). The right term is the prior on θ, and takes the form of a Laplacian on each parameter value. Hence, the probabilistic interpretation of L1-regularized least squares is to find the MAP estimate of the parameters θ, using the model

$$y = f(x;\theta) + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2), \quad \theta_i \sim \text{Laplace}(\lambda) = \tfrac{1}{2\lambda}e^{-\frac{|\theta_i|}{\lambda}}. \tag{S.247}$$

(b) In the equivalent optimization problem, we rewrite θ_i as the difference between two positive values, $\theta_i = \theta_i^+ - \theta_i^-$, where $\theta_i^+ \geq 0$ and $\theta_i^- \geq 0$. Note that when $\theta_i^+ = 0$ and $\theta_i^- > 0$, and vice versa ($\theta_i^+ > 0$, $\theta_i^- = 0$), then the absolute value can be rewritten as $|\theta_i^+ - \theta_i^-| = (\theta_i^+ + \theta_i^-)$. Finally, at the optimum indeed one of these two conditions hold, $(\theta_i^+ = 0, \theta_i^- > 0)$ or $(\theta_i^+ > 0, \theta_i^- = 0)$. If it were not the case, i.e., $(\theta_i^+ > 0, \theta_i^- > 0)$, then the term $(\theta_i^+ + \theta_i^-)$ could be further reduced by subtracting $\min(\theta_i^+, \theta_i^-)$ from θ_i^+ and θ_i^-. This would reduce the regularization term $(\theta_i^+ + \theta_i^-)$, but not affect the data term since the data term only depends on the difference $(\theta_i^+ - \theta_i^-)$.

(c) Let $\mathbf{x} = \begin{bmatrix}\theta^+ \\ \theta^-\end{bmatrix}$. The objective function is

$$E = \frac{1}{2}\left\|y - \Phi^T(\theta^+ - \theta^-)\right\|^2 + \lambda\sum_i(\theta_i^+ + \theta_i^-) \tag{S.248}$$

$$= \frac{1}{2}\left\|y - \left[\Phi^T, -\Phi^T\right]\mathbf{x}\right\|^2 + \lambda\mathbf{1}^T\mathbf{x} \tag{S.249}$$

$$= \frac{1}{2}y^Ty - y^T\left[\Phi^T, -\Phi^T\right]\mathbf{x} + \frac{1}{2}\mathbf{x}^T\begin{bmatrix}\Phi \\ -\Phi\end{bmatrix}\left[\Phi^T, -\Phi^T\right]\mathbf{x} + \lambda\mathbf{1}^T\mathbf{x} \tag{S.250}$$

$$\propto \frac{1}{2}\mathbf{x}^T\underbrace{\begin{bmatrix}\Phi\Phi^T & -\Phi\Phi^T \\ -\Phi\Phi^T & \Phi\Phi^T\end{bmatrix}}_{\mathbf{H}}\mathbf{x} + \underbrace{\left(\lambda\mathbf{1} - \begin{bmatrix}\Phi y \\ -\Phi y\end{bmatrix}\right)}_{\mathbf{f}}^T\mathbf{x}, \tag{S.251}$$

where constant terms that do not affect the minimization are dropped.

Figure 8 shows an example of cubic polynomial regression using least-squares, Bayesian regression, and LASSO.

....

Problem 4.12 Lagrange multipliers and equality constraints

(a) The Lagrangian is

$$L(\pi,\lambda) = \sum_{j=1}^{K} z_j \log \pi_j + \lambda\left(\sum_{j=1}^{K}\pi_j - 1\right). \tag{S.252}$$

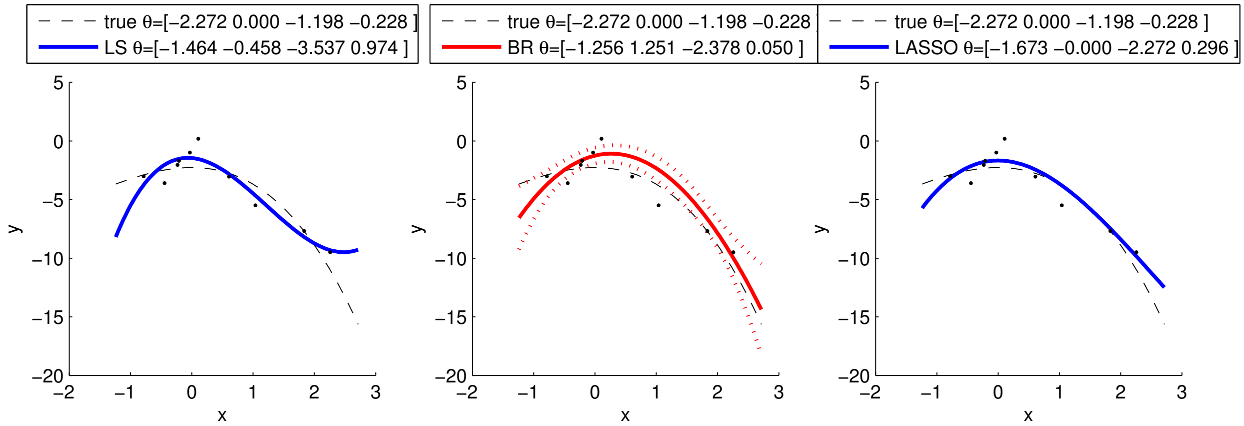

Figure 8: Cubic polynomial regression using (left) least-squares, (middle) Bayesian regression, (right) LASSO. The true function is the dashed line. For Bayesian regression, the dotted-lines show the 2 standard-deviations around the mean. Note that LASSO can find that the linear term θ_1 is 0.

Taking the derivatives and setting to zero gives

$$\frac{\partial L}{\partial \lambda} = \sum_{j=1}^{K} \pi_j - 1 = 0 \quad \Rightarrow \quad \sum_{j=1}^{K} \pi_j = 1, \tag{S.253}$$

$$\frac{\partial L}{\partial \pi_j} = \frac{z_j}{\pi_j} + \lambda = 0 \quad \Rightarrow \quad z_j + \pi_j \lambda = 0. \tag{S.254}$$

Taking (S.254) and summing over j,

$$\sum_{j=1}^{K} (z_j + \pi_j \lambda) = 0 \quad \Rightarrow \quad \lambda \sum_{j=1}^{K} \pi_j = -\sum_{j=1}^{K} z_j \quad \Rightarrow \quad \lambda = -\sum_{j=1}^{K} z_j, \tag{S.255}$$

since $\sum_{k=1}^{K} \pi_k = 1$. Finally substituting λ into (S.254) (and changing the index from j to k to avoid confusion),

$$z_j - \pi_j \sum_{k=1}^{K} z_j = 0 \quad \Rightarrow \quad \pi_j = \frac{z_j}{\sum_{k=1}^{K} z_k}. \tag{S.256}$$

(b) The Lagrangian is

$$L(\pi, \lambda) = \sum_{j=1}^{K} \pi_j (z_j - \log \pi_j) + \lambda \left(\sum_{j=1}^{K} \pi_j - 1 \right). \tag{S.257}$$

Taking the derivatives and setting to zero gives

$$\frac{\partial L}{\partial \lambda} = \sum_{j=1}^{K} \pi_j - 1 = 0 \quad \Rightarrow \quad \sum_{j=1}^{K} \pi_j = 1, \tag{S.258}$$

$$\frac{\partial L}{\partial \pi_j} = z_j - \log \pi_j - \frac{\pi_j}{\pi_j} + \lambda = 0 \quad \Rightarrow \quad \frac{1}{\pi_j} e^{z_j - 1} e^{\lambda} = 0 \quad \Rightarrow \quad \pi_j e^{-\lambda} = e^{z_j - 1} \tag{S.259}$$

Summing over j and noting that $\sum_{j=1}^{K} \pi_j = 1$ gives

$$e^{-\lambda} = \sum_{j=1}^{K} e^{z_j - 1}. \tag{S.260}$$

Finally, substituting (S.260) back into (S.259),

$$\pi_j \sum_{k=1}^{K} e^{z_k - 1} = e^{z_j - 1} \quad \Rightarrow \quad \pi_j = \frac{e^{z_j - 1}}{\sum_{k=1}^{K} e^{z_k - 1}} = \frac{e^{z_j}}{\sum_{j=1}^{K} e^{z_k}} \tag{S.261}$$

....

Problem 4.6 Mixture of exponentials

Define z_i as the hidden assignment variable that assigns sample x_i to mixture component $z_i = j$. The complete data likelihood is

$$p(X, Z) = \prod_{i=1}^{n} p(z_i) p(x_i | z_i) = \prod_{i=1}^{n} \pi_{z_i} p(x_i | z_i). \tag{S.262}$$

Using the indicator variable trick, by defining $z_{ij} = 1$ iff $z_i = j$ and 0 otherwise, we have

$$p(X, Z) = \prod_{i=1}^{n} \prod_{j=1}^{K} \pi_j^{z_{ij}} p(x_i | z_i = j)^{z_{ij}}, \tag{S.263}$$

and taking the log,

$$\log p(X, Z) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \log \pi_j + z_{ij} \log p(x_i | z_i = j). \tag{S.264}$$

For the E-step, we obtain the Q function by taking the expectation of the complete data log-likelihood in (S.264),

$$Q(\theta; \hat{\theta}) = \mathbb{E}_{Z|X,\hat{\theta}} \left[\log p(X, Z | \theta) \right] \tag{S.265}$$

$$= \mathbb{E}_{Z|X,\hat{\theta}} \left[\sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \log \pi_j + z_{ij} \log p(x_i | z_i = j) \right] \tag{S.266}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{K} \hat{z}_{ij} \log \pi_j + \hat{z}_{ij} \log p(x_i | z_i = j), \tag{S.267}$$

where the last line follows because the expectation only applies to variable z_{ij}. The "soft assignment" term $\hat{z}_{ij}$ is calculated using the current parameter estimates $\hat{\theta}$,

$$\hat{z}_{ij} = \mathbb{E}_{Z|X,\hat{\theta}}[z_{ij}] = p(z_i = j | X, \hat{\theta}) \tag{S.268}$$

$$= \frac{p(X | z_i = j) p(z_i = j)}{p(X)} = \frac{p(X_{\neg i}) p(x_i | z_i = j) p(z_i = j)}{p(X_{\neg i}) p(x_i)} \tag{S.269}$$

$$= \frac{\pi_j p(x_i | z_i = j)}{\sum_{k=1}^{K} \pi_k p(x_i | z_i = k)} = p(z_i = j | x_i, \hat{\theta}). \tag{S.270}$$

(S.269) follows from the independence assumption of the samples, and $X_{\neg i}$ is the set of x_k with $k \neq i$. Note that we have not used any properties of the mixture components yet, so this is the general form of the Q function for any mixture model.

For the M-step, we maximize Q with respect to the parameters θ. First, we optimize the component priors,

$$\pi^* = \operatorname*{argmax}_{\pi} Q(\theta; \hat{\theta}). \tag{S.271}$$

We have a constraint that $\sum_j \pi_j = 1$, and $\pi_j \geq 0$. For the equality constraint (the non-negative constraint is naturally satisfied), define the Lagrangian as

$$L(\pi) = \sum_{i=1}^{n} \sum_{j=1}^{K} \hat{z}_{ij} \log \pi_j + \lambda(1 - \sum_{j=1}^{K} \pi_j), \tag{S.272}$$

where λ is the Lagrange multiplier. Taking the derivatives and setting to 0,

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{j=1}^{K} \pi_j = 0 \quad \Rightarrow \quad \sum_{j=1}^{K} \pi_j = 1, \tag{S.273}$$

$$\frac{\partial L}{\partial \pi_j} = \sum_{i=1}^{n} \frac{\hat{z}_{ij}}{\pi_j} + \lambda = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \hat{z}_{ij} + \lambda \pi_j = 0. \tag{S.274}$$

Summing (S.274) over j, we have

$$\sum_{j=1}^{K} \sum_{i=1}^{n} \hat{z}_{ij} + \lambda \sum_{j=1}^{K} \pi_j = 0 \quad \Rightarrow \quad \lambda = -n, \tag{S.275}$$

which follows from $\sum_j \hat{z}_{ij} = 1$ and $\sum_j \pi_j = 1$. Finally, substituting into (S.274), we have

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ij} = \frac{\hat{n}_j}{n}, \tag{S.276}$$

where $\hat{n}_j = \sum_{i=1}^{n} \hat{z}_{ij}$ is the (soft) number of samples assigned to component j. Again this is a standard result for any mixture model. For the M-step, the mixture weights are updated as before,

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ij} = \frac{\hat{n}_j}{n}, \tag{S.277}$$

where $\hat{n}_j = \sum_{i=1}^{n} \hat{z}_{ij}$ is the (soft) number of samples assigned to component j.

Finally, for the exponential parameter λ_j, we optimize:

$$\lambda_j^* = \operatorname*{argmax}_{\lambda_j} Q(\theta; \hat{\theta}). \tag{S.278}$$

We collect the terms of the Q function that depend on λ_j,

$$\ell_j = \sum_{i=1}^{n} \hat{z}_{ij} \log p(x_i | z_i = j) = \sum_{i=1}^{n} \hat{z}_{ij} (\log \lambda_j - \lambda_j x_i). \tag{S.279}$$

Taking the derivative and setting to 0 gives

$$\frac{\partial \ell_j}{\partial \lambda_j} = \sum_{i=1}^n \hat{z}_{ij}(\lambda_j^{-1} - x_i) = 0 \quad \Rightarrow \quad \lambda_j^{-1}\sum_{i=1}^n \hat{z}_{ij} = \sum_{i=1}^n \hat{z}_{ij}x_i \quad \Rightarrow \quad \hat{\lambda}_j^{-1} = \frac{\sum_{i=1}^n \hat{z}_{ij}x_i}{\sum_{i=1}^n \hat{z}_{ij}}. \tag{S.280}$$

In summary, the EM algorithm for mixture of exponentials is

$$\mathrm{E-step:}\ \ \hat{z}_{ij} = p(z_i = j|x_i, \hat{\theta}) = \frac{\hat{\pi}_j p(x_i|z_i = j, \hat{\lambda}_j)}{\sum_{k=1}^K \hat{\pi}_k p(x_i|z_i = k, \hat{\lambda}_j)} \tag{S.281}$$

$$\mathrm{M-step:}\ \ \hat{n}_j = \sum_{i=1}^n \hat{z}_{ij}, \quad \hat{\pi}_j = \frac{\hat{n}_j}{n}, \quad \hat{\lambda}_j^{-1} = \frac{1}{\hat{n}_j}\sum_{i=1}^n \hat{z}_{ij}x_i. \tag{S.282}$$

. . . .

Problem 5.1 Bias and variance of the kernel density estimator

(a) To calculate the bias and variance of the kernel density estimator $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^n \tilde{k}(x - x_i)$, we suppose that the samples $X = \{x_i\}_{i=1}^n$ are distributed according to the true distribution $p(x)$, i.e., $x_i \sim p(x), \forall i$. The mean of the estimator $\hat{p}(x)$ is

$$\mathbb{E}_X[\hat{p}(x)] = \mathbb{E}_X\left[\frac{1}{n}\sum_{i=1}^n \tilde{k}(x - x_i)\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x_i}\left[\tilde{k}(x - x_i)\right] \tag{S.283}$$

$$= \frac{1}{n}\sum_{i=1}^n \int p(x_i)\tilde{k}(x - x_i)dx_i = \int p(z)\tilde{k}(x - z)dz = p(x) * \tilde{k}(x), \tag{S.284}$$

where (S.284) follows from each term in the sum being the same, and $*$ is the convolution operator. The mean of estimator is the true distribution convolved with the kernel. In other words, it is a "blurred" or "smoothed" version of the true distribution.

(b) The variance of the estimator is

$$\mathrm{var}_X(\hat{p}(x)) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^n \tilde{k}(x - x_i)\right) = \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^n \tilde{k}(x - x_i)\right) \tag{S.285}$$

$$= \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}\left(\tilde{k}(x - z)\right) = \frac{1}{n}\mathrm{var}\left(\tilde{k}(x - x_i)\right), \tag{S.286}$$

which follows from $\{x_i\}$ being independent distributions, and hence the variance of the sum is the sum of the variances (see Problem 1.4), and also identical distributions. Noting that $\mathrm{var}(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ and thus $\mathrm{var}(x) \leq \mathbb{E}[x^2]$, then we can place an upper-bound on the variance,

$$\mathrm{var}_X(\hat{p}(x)) \leq \frac{1}{n}\mathbb{E}[\tilde{k}(x - z)^2] = \frac{1}{n}\int \frac{1}{h^d}k\left(\frac{x - z}{h}\right)\tilde{k}(x - z)p(z)dz \tag{S.287}$$

$$\leq \frac{1}{nh^d}\left(\max_x k(x)\right)\int \tilde{k}(x - z)p(z)dz = \frac{1}{nh^d}\left(\max_x k(x)\right)\mathbb{E}[\hat{p}(x)], \tag{S.288}$$

where the last line follows from $k(\frac{x-z}{h}) \leq \max_x k(x)$, i.e., the kernel is upper-bounded by its maximum value.

Figure 9 plots the mean and variance of the KDE for different bandwidths.

. . . .

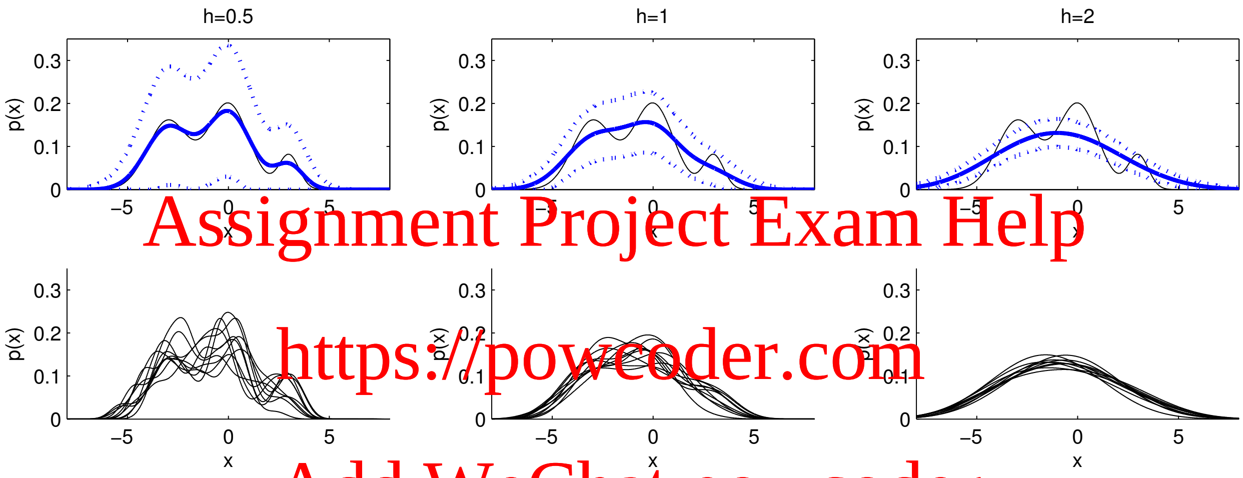

Figure 9: Mean and variance of kernel density estimator. Top row: the true density function (thin line) and the estimator mean (thick line) and 2 standard deviations around the mean (dotted line) for different bandwidths h. Bottom row: Examples of the estimate $\hat{p}(x)$ using 10 different sets of 50 samples drawn from the true density $p(x)$. When the bandwidth is small ($h = 0.5$), the estimates are significantly different for each sample set (i.e., high variance), but the estimator mean is close to the true density. When the bandwidth is large ($h = 2$), the estimates are consistent with each other (i.e., low variance), but the estimator mean is far from the true density.