Nicholas Trieu
MATH 189r
Homework 1
September 12, 2016

3 (Murphy 2.11 and 2.16)

(a) Derive the normalization constant (Z) for a one dimensional zero-mean Gaussian

$$\mathbb{P}(x;\sigma^2) = \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

such that $\mathbb{P}(x;\sigma^2)$ becomes a valid density.

(b) Suppose $\theta \sim \text{Beta}(a,b)$ such that

$$\mathbb{P}(\theta;a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function and $\Gamma(x)$ is the Gamma function. Derive the mean, mode, and variance of θ.

(a) For $\mathbb{P}(x;\sigma^2)$ to be a valid density, it must be that

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} \mathbb{P}(x;\sigma^2)dx \\
1 &= \int_{-\infty}^{\infty} \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)dx \\
1 &= \frac{1}{Z}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx \\
Z &= \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx \\
Z^2 &= \left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx\right)^2
\end{aligned}$$

We now find $\left(\int_{-\infty}^{\infty}\exp\left(-\frac{x^2}{2\sigma^2}\right)dx\right)^2$.

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx\right)^2 &= \left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx\right)\left(\int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\sigma^2}\right)dy\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{(x^2+y^2)}{2\sigma^2}\right)dxdy \\
&= \int_{0}^{2\pi}\int_{0}^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right)\cdot rdrd\theta \\
&= 2\pi\sigma^2
\end{aligned}$$

So $Z = \sigma\sqrt{2\pi}$.

(b) We first note that:

$$\begin{aligned}
\mathbb{E}[\theta^k] &= \int_0^1 \theta^k \cdot \mathbb{P}(\theta)\, d\theta \\
&= \int_0^1 \theta^k \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a+k-1}(1-\theta)^{b-1}\, d\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot B(a+k, b) \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+k)\Gamma(b)}{\Gamma(a+b+k)}
\end{aligned}$$

It follows that the mean of θ is

$$\begin{aligned}
\mathbb{E}[\theta] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{a}{a+b}
\end{aligned}$$

and the variance is

$$\begin{aligned}
\text{var}[\theta] &= \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2 \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} - \left(\frac{a}{a+b}\right)^2 \\
&= \left(\frac{a(a+1)}{(a+b)(a+b+1)}\right) - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

The mode of θ is when $\nabla_\theta \mathbb{P}(\theta; a, b) = 0$:

$$\begin{aligned}
0 &= \nabla_\theta \mathbb{P}(\theta; a, b) \\
&= \frac{1}{B(a,b)} \theta^{a-2}(1-\theta)^{b-2}\left((a-1) - (a+b-2)\theta\right).
\end{aligned}$$

Since $\mathbb{P}(\theta) = 0$ when $\theta = 0$ or $\theta = 1$, it follows that the mode of θ is

$$\frac{a-1}{a+b-2}.$$

■

5 (Linear Transformation) Let $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ be a random vector. show that expectation is linear:

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[A\mathbf{x} + \mathbf{b}] = A\mathbb{E}[\mathbf{x}] + \mathbf{b}.$$

Also show that

$$\operatorname{cov}[\mathbf{y}] = \operatorname{cov}[A\mathbf{x} + \mathbf{b}] = A\operatorname{cov}[\mathbf{x}]A^\top = A\mathbf{\Sigma}A^\top.$$

We first show that expectation is linear. Let M_i denote the ith row of any matrix M.

$$\mathbb{E}[\mathbf{y}] = (\mathbb{E}[y_1], \mathbb{E}[y_2], \dots)$$

since $\mathbf{y}$ is a vector. We now solve for $\mathbb{E}[y_i]$.

$$\begin{aligned}
\mathbb{E}[y_i] &= \mathbb{E}[A_i\mathbf{x} + b_i] \quad \text{by definition of } \mathbf{y} \\
&= \mathbb{E}[A_i\mathbf{x}] + b_i \quad \text{by linearity of scalars} \\
&= \mathbb{E}[\textstyle\sum A_{ij}x_j] + b_i \quad \text{where } A_{ij} \text{ is the element in the } i\text{th row and } j\text{th column} \\
&= \textstyle\sum \left(A_{ij}\mathbb{E}[x_j]\right) + b_i \\
&= A_i\mathbb{E}[\mathbf{x}] + b_i.
\end{aligned}$$

Plugging back into $\mathbb{E}[\mathbf{y}]$, we get:

$$\begin{aligned}
\mathbb{E}[\mathbf{y}] &= (\mathbb{E}[y_1], \mathbb{E}[y_2], \dots) \\
&= (A_1\mathbb{E}[\mathbf{x}] + b_1, A_2\mathbb{E}[\mathbf{x}] + b_2, \dots) \\
&= (A_1\mathbb{E}[\mathbf{x}], A_2\mathbb{E}[\mathbf{x}], \dots) + \mathbf{b} \\
&= A\mathbb{E}[\mathbf{x}] + \mathbf{b}
\end{aligned}$$

as desired.

We now show that

$$\operatorname{cov}[\mathbf{y}] = \operatorname{cov}[A\mathbf{x} + \mathbf{b}] = A\operatorname{cov}[\mathbf{x}]A^\top = A\mathbf{\Sigma}A^\top.$$

Let $\operatorname{cov}[\mathbf{y}]_{ij}$ be the ijth element of the covariance matrix of $\mathbf{y}$. By definition of the covariance matrix, we have

$$\operatorname{cov}[\mathbf{y}]_{ij} = \operatorname{cov}[\operatorname{cov}[y_i, y_j]].$$

We now solve for $\operatorname{cov}[y_i, y_j]$.

$$\begin{aligned}
\operatorname{cov}[y_i, y_j] &= \mathbb{E}[(y_i - \mathbb{E}[y_i]) \cdot (y_j - \mathbb{E}[y_j])] \\
&= \mathbb{E}[((A_i\mathbf{x} + b_i) - (A_i\mathbb{E}[\mathbf{x}] + b_i)) \cdot \left((A_j\mathbf{x} + b_j) - (A_j\mathbb{E}[\mathbf{x}] + b_j)\right)] \\
&= \mathbb{E}[(A_i(\mathbf{x} - \mathbb{E}[\mathbf{x}])) \cdot \left(A_j(\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)] \\
&= \mathbb{E}[(A_i(\mathbf{x} - \mathbb{E}[\mathbf{x}])) \left(A_j(\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)^\top] \\
&= \mathbb{E}[A_i(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top A_j^\top] \\
&= A_i\operatorname{cov}[\mathbf{x}]A_j^\top
\end{aligned}$$

Thus we see that

$$\operatorname{cov}[\mathbf{y}] = A\operatorname{cov}[\mathbf{x}]A^\top$$

as desired. ∎

6 Given the dataset $\mathcal{D} = \{(x, y)\} = \{(0,1), (2,3), (3,6), (4,8)\}$
(a) Find the least squares estimate $y = \boldsymbol{\theta}^\top \mathbf{x}$ by hand using Cramer's Rule.
(b) Use the normal equations to find the same solution and verify it is the same as part (a).
(c) Plot the data and the optimal linear fit you found.
(d) Find randomly generate 100 points near the line with white Gaussian noise and then compute the least squares estimate (using a computer). Verify that this new line is close to the original and plot the new dataset, the old line, and the new line.

(a) Let $y = a_1 x + a_2$ be our least squares estimate, let $n = 4$ be the number of points, and let

$$D = \det \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = 35.$$

Then

$$\begin{aligned} a_1 &= \frac{\det \begin{bmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{bmatrix}}{D} \\ &= \frac{62}{35} \end{aligned}$$

and

$$\begin{aligned} a_2 &= \frac{\det \begin{bmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{bmatrix}}{D} \\ &= \frac{18}{35}. \end{aligned}$$

So our least squares estimate is $y = \frac{62}{35} x + \frac{18}{35}$.

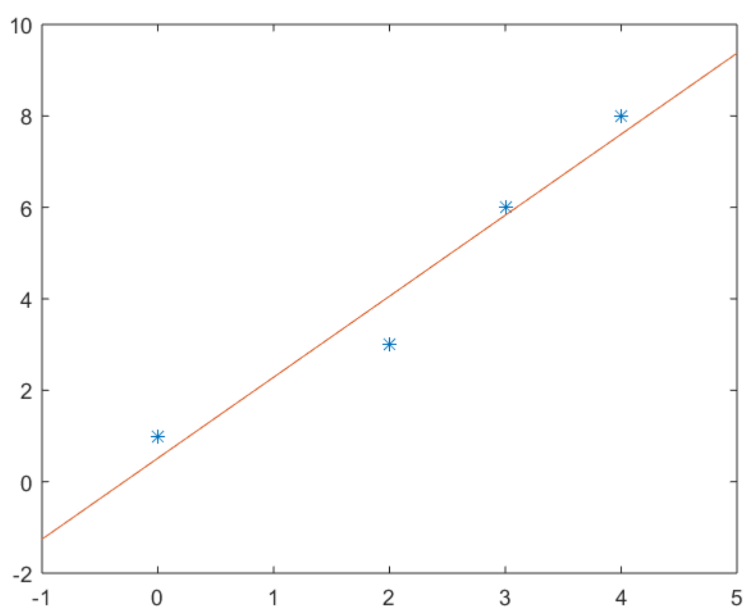
(b) Recall that

$$\vec{\theta} = (X^\top X)^{-1} X^T \vec{y}$$

where $X = [\vec{x}_0 \ \vec{x}_1]$, $\vec{x}_0 = \mathbf{1}$, and $\vec{x}_1 = \vec{x}$. Let $n = 4$ denote the number of points. It follows that

$$
\begin{aligned}
D &= \det X^\top X \\
&= \det \begin{bmatrix} \mathbf{1}\cdot\mathbf{1} & \mathbf{1}\cdot\vec{x} \\ \mathbf{1}\cdot\vec{x} & \vec{x}\cdot\vec{x} \end{bmatrix} \\
&= \det \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \\
(X^\top X)^{-1} &= \frac{1}{D}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \\
X^\top \vec{y} &= \begin{bmatrix} \mathbf{1}\cdot\vec{y} \\ \vec{x}\cdot\vec{y} \end{bmatrix} \\
&= \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\
\vec{\theta} &= (X^\top X)^{-1} X^\top \vec{y} \\
&= \frac{1}{D}\begin{bmatrix} \sum x_i^2 \cdot \sum y_i - \sum x_i \cdot \sum x_i y_i \\ -\sum x_i \cdot \sum y_i + n \cdot \sum x_i y_i \end{bmatrix} \\
&= \begin{bmatrix} \frac{18}{35} \\ \frac{62}{35} \end{bmatrix}
\end{aligned}
$$

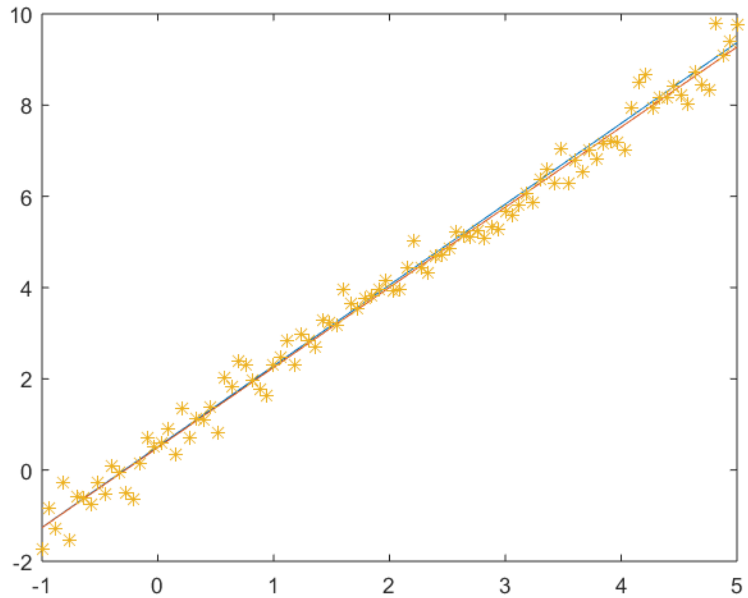
So our linear fit is $y = \dfrac{62}{35}x + \dfrac{18}{35}$, the same as in part (a).



(c)

The graph above shows the dataset of 4 points and the optimal linear fit

$$y = 1.7714x + 0.5143$$



(d)

The graph above shows the new dataset (100 points near the old line with white Gaussian noise), the new line, and the old line.
The new line is

$$y = 1.7560x + 0.4938$$

which is visibly close to the old line

$$y = 1.7714x + 0.5143$$

in the graph.

Matlab code:

```
x = [0 2 3 4];
y = [1 3 6 8];
linearCoefficients = polyfit(x,y,1);
xFit = linspace(0, 10, 100);
yFit = polyval(linearCoefficients, xFit);
yNoise = awgn(yFit, 10);
noiseCoeffs = polyfit(xFit,yNoise,1);
yNoiseFit = polyval(noiseCoeffs, xFit);

% Plot the old dataset and optimal linear fit
% plot(x,y, '*', xFit, yFit)

% Plot the new dataset, new fit, and old fit
% plot(xFit,yFit, xFit, yNoiseFit, xFit, yNoise, '*')
```

∎