STA414 HW #1 Solutions

Due: N/A

Problem 1 (Variance and covariance)

Let $\mu_A = \mathbb{E}(A)$ and $\mu_B = \mathbb{E}(B)$.

(a) We note that since μ_A and μ_B are fixed, then $A \perp B$ implies that $(A - \mu_A) \perp (B - \mu_B)$. Therefore

$$\begin{aligned}\text{cov}(A, B) &= \mathbb{E}((A - \mu_A)(B - \mu_B)) \\ &= \mathbb{E}(A - \mu_A)\mathbb{E}(B - \mu_B) \\ &= 0\end{aligned}$$

where the second line is by the independence stated above and the final line is because linearity of expectation gives that $E(A - \mu_A) = E(A) - E(\mu_A) = \mu_A - \mu_A = 0$.

(b) We use the fact that for a random variable Z, $\text{var}(Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2$. Setting $Z = A + aB$, we obtain

$$\begin{aligned}\text{var}(A + aB) &= \mathbb{E}((A + aB)^2) - \mathbb{E}(A + aB)^2 \\ &= \mathbb{E}(A^2 + 2aAB + a^2B^2) - (\mu_A + a\mu_B)^2 \\ &= (\mathbb{E}(A^2) - \mu_A^2) + a^2(\mathbb{E}(B^2) - \mu_B^2) + 2a(\mathbb{E}(AB) - \mu_A\mu_B) \\ &= \text{var}(A) + a^2\text{var}(B) + 0\end{aligned}$$

where the last line uses the independence of A and B to obtain that $\mathbb{E}(AB) = \mu_A\mu_B$.

Note: this is a special case of a more general result, which is that for two random variables A, B that may not be independent, $\text{var}(aA + bB) = a^2\text{var}(A) + b^2\text{var}(B) + 2ab\text{cov}(A, B)$.

Problem 2 (Densities)

(a) Yes.

(b) The density is $f_X(x) = \frac{4}{\sqrt{2\pi}}e^{-8x^2}$.

(c) Setting $x = 0$ above yields $f_X(0) = 4/\sqrt{2\pi} \approx 1.6 > 1$.

(d) Since X is a continuous random variable, the probability that it takes on any fixed value is 0.

(e) The definition of a probability density function is that it is the derivative of the cumulative distribution function. The key property that this implies is that the integral of f_X over a set A equals the probability that $X \in A$. Therefore, the pdf can take on arbitrary values, including values greater than 1, provided its integral over any set is never greater than 1. In particular, the integral of the pdf over the entire support of X equals 1, which can be verified by integrating the density f_X over the entire real line.

Problem 3 (Calculus)

(a) $\partial(z^T y)/\partial z_i = \partial(\sum_i z_i y_i)/\partial z_i = y_i$, so $\nabla(z^T y) = y$.
y^T is also acceptable, depending on convention. Similarly for the following questions.

(b) $\partial(z^T z)/\partial z_i = \partial(\sum_i z_i^2)/\partial z_i = 2z_i$, so $\nabla(z^T z) = 2z$.

(c) A methodical solution is to write

$$\begin{aligned}
\frac{\partial}{\partial z_i}(z^T A z) &= \frac{\partial}{\partial z_i}\sum_{j,k} A_{jk} z_j z_k \\
&= \sum_{j,k\neq i} A_{jk} z_j z_k + \sum_{k\neq i} A_{ik} z_i z_k + \sum_{j\neq i} A_{ji} z_j z_i + A_{ii} z_i^2 \\
&= \sum_{k\neq i} A_{ik} z_k + \sum_{j\neq i} A_{ji} z_j + 2A_{ii} z_i \\
&= \sum_{k} A_{ik} z_k + \sum_{j} A_{ji} z_j \\
&= (Az)_i + (z^T A)_i
\end{aligned}$$

2Az

which means that $\nabla(z^T A z) = (A + A^T)z$. Alternatively, glib equations such as (C.20) in Bishop 2006 can be used provided you transpose as required to achieve addition in the product rule:

$$\begin{aligned}
\frac{\partial}{\partial z_i}(z^T A z) &= \left(\frac{\partial}{\partial z_i} z^T\right)(Az) + \left[z^T \frac{\partial}{\partial z_i}(Az)\right]^T \\
&= I_n A z + (z^T A)^T \quad \text{making use of 3(d), below} \\
&= Az + A^T z \\
&= (A + A^T)z
\end{aligned}$$

(d) Since Az is vector-valued, the "gradient" of Az will be a vector of vectors, i.e., a matrix. This is actually a generalization of the traditional concept of a gradient; it's better referred to as a Jacobian. Let us determine the answer one component at a time, by computing the gradient of the j-th component of Az for each j. We have

$$\frac{\partial}{\partial z_i}(Az)_j = \frac{\partial}{\partial z_i}\sum_k A_{jk} z_k = A_{ji}$$

which means that $\partial(Az)/\partial z_i = A_{:,i}$, the i-th column of A. Combining over all i, we see that $\nabla(Az) = A$, which is analogous to the scalar case. (It must be, for what if A is a 1×1 matrix?)

See also (C.18) in Bishop 2006.

Problem 4 (Regression)

(a) The reason for the assumption that $n \geq m$ is that we need the matrix $X^T X$ to be invertible. The matrix is $m \times m$, so this means that its rank must be at least m. But as we will show, its rank is at most n, and so if $n < m$, there is no hope of inverting $X^T X$.

To upper-bound the rank of $X^T X$, we let the column-vector x_i denote the i-th observation (i.e., x_i is m-by-1), so that

$$X^T X = \begin{pmatrix} x_1 & \cdots & x_m \end{pmatrix} \cdot \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix} = \sum_{i=1}^{n} x_i \cdot x_i^T.$$

The result then follows easily, for each matrix $X_i \cdot X_i^T$ is a rank-1 n-by-n matrix and $X^T X$, beings a sum of n of these, cannot have its rank exceed n.

(b) Representing Y as $Y = X\beta + \varepsilon$ for $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, we have

$$\hat{\beta} = (X^T X)^{-1} X^T X \beta + (X^T X)^{-1} X^T \varepsilon = \beta + (X^T X)^{-1} X^T \varepsilon.$$

Since $\mathbb{E}(\varepsilon) = 0$, this easily gives that $\mathbb{E}(\hat{\beta}) = \beta$. To compute the variance, we write

$$\begin{aligned} \operatorname{var}(\hat{\beta}) &= \mathbb{E}\left((X^T X)^{-1} X^T \varepsilon ((X^T X)^{-1} X^T \varepsilon)^T\right) \\ &= \mathbb{E}\left((X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}\right) \\ &= (X^T X)^{-1} X^T \mathbb{E}\left(\varepsilon \varepsilon^T\right) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

(This is a special case of the general fact that if $Z \sim \mathcal{N}(\mu, \Sigma)$ and A is a matrix, then $AZ \sim \mathcal{N}(A\mu, A\Sigma A^T)$.)

(c) The density of Y given X is

$$\begin{aligned} f(y|X, \beta) &= (2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{1}{2}\left((y - X\beta)^T (\sigma^2 I)^{-1} (y - X\beta)\right)\right) \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{(y - X\beta)^T (y - X\beta)}{2\sigma^2}\right). \end{aligned}$$

The log-likelihood is therefore

$$\ell(\beta) = -\frac{n}{2}\left(\ln\left(\frac{2}{\pi}\right) + 2\ln\sigma\right) - \frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)$$

Since the first term is free of β, it suffices to take the gradient of the second term only. To do so, we write

$$(y - X\beta)^T (y - X\beta) = y^T y + \beta^T (X^T X) \beta - 2 y^T X \beta$$ scalar so transpose to itself

from which application of the identities proven in question 3 quickly gives

$$\nabla\left((y - X\beta)^T (y - X\beta)\right) = 2(\beta^T X^T X - y^T X)$$

Note output is a row vector …

can freely take transpose of row vector as long as dimension matches

Or depending on convention, you can also write $-2X(y - X\beta)$. Putting everything togther, we get that the gradient of the log-likelihood is as follows.

$$\nabla \ell(\beta) = -\frac{1}{\sigma^2}(\beta^T X^T X - y^T X) \quad \text{or} \quad \frac{X(y - X\beta)}{\sigma^2}$$

Notice that the solution to $\nabla \ell(\beta) = 0$ is indeed the MLE of β. It's also free of σ, which means we do not need to know the magnitude of our noise in order to estimate β. The latter fact is fortunate, but is a special feature of this model and not a foregone conclusion.

Problem 5 (Ridge regression)

(a) We do not require $n \geq m$ for ridge regression. This is because the matrix $X^T X + \lambda I$ is always invertible for positive λ, for the following reason: $X^T X$, being a sample covariance matrix, is positive definite. You can prove this by considering the singular value decomposition (SVD) of X. Therefore, its eigenvalues are non-negative. Adding λI to $X^T X$ adds λ to each eigenvalue, and therefore renders all the eigenvalues strictly positive. Therefore, $X^T X + \lambda I$ is invertible regardless of the original rank of $X^T X$.

(b) Let X' and Y' denote the modified X and Y respectively. It is easy to see that $X'^T Y' = X^T Y$: the reason is that the last m entries of Y' are 0. It remains then only to show that $(X')^T X' = X^T X + \lambda I$. But this is also simple, because the i,j-th entry of $X'^T X'$ is the inner product of the i-th and j-th columns of X'. If $i \neq j$, this will equal the inner product of the corresponding columns of X', i.e., the i,j-th entry of $X^T X$. But when $i = j$, it will equal the squared norm of the i-th column of X plus $(\sqrt{\lambda})^2 = \lambda$.

Problem 6 (High dimensions)

(a) R code:

```
for (d in c(2,10,1000)) {
  print(1-.99^d)
}

[1] 0.0199
[1] 0.09561792
[1] 0.9999568
```

(b) R code:

```
for (d in c(2,10,1000)) {
  print(.5^d)
}

[1] 0.25
[1] 0.0009765625
[1] 9.332636e-302
```

(c) The answer is (c), after considering the trends in (a) and (b).