1. (a) We are given the geometric distribution with parameter p, $P(X = k) = (1-p)^{k-1}p$. With $\bar{k}$ denoting the mean of k, the log-likelihood can be written as follows

$$\begin{aligned}
l(X,p) &= \sum_{i=1}^{n} \log f(X_i;p) \\
&= \sum_{i=1}^{n} \log{(1-p)^{k_i-1}p} \\
&= \sum_{i=1}^{n} \log(1-p)^{k_i-1} + \log p \\
&= n(\bar{k}-1)\log(1-p) + n\log p
\end{aligned}$$

We then solve the score equation:

$$\begin{aligned}
0 &= \frac{\partial l(p)}{\partial p} \\
&= -\frac{n(\bar{k}-1)}{1-p} + \frac{n}{p} \\
&= -pn\bar{k} + pn + n - np \\
&= n - pn\bar{k}
\end{aligned}$$

We therefore see that

$$n = pn\bar{k} \to p = \frac{1}{\bar{k}}$$

is our maximum likelihood estimate.

(b) Each draw from the normal $N(\mu, \Sigma) = \frac{1}{\sqrt{2\pi\Sigma^2}}e^{-\frac{(x-\mu)^2}{2\Sigma^2}}$ has likelihood $\Pi_{i=1}^{n}\frac{1}{\sqrt{2\pi\Sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\Sigma^2}} = \frac{1}{(2\pi\Sigma^2)^{n/2}}e^{-\frac{1}{(2\Sigma^2)^n)}\sum_i^n(x_i-\mu)^2}$. The log likelihood is therefore of the form

$$\begin{aligned}
l(X,\mu) &= \sum_{i=1}^{n} \log \frac{1}{(2\pi\Sigma^2)^{n/2}}e^{-\frac{1}{(2\Sigma^2)^n)}\sum_j^n(x_j-\mu)^2} \\
&= n\log\frac{1}{(2\pi\Sigma^2)^{n/2}} - \frac{n}{(2\Sigma^2)^n}\sum_j^n(x_j-\mu)^2
\end{aligned}$$

We then take the derivative and solve for the MLE of μ.

$$\begin{aligned}
0 &= \frac{\partial l}{\partial \mu} \\
&= \frac{n}{(2\Sigma^2)^n} \sum_j^n 2(x_j - \mu) \\
&= -n\mu + \sum_j^n x_j
\end{aligned}$$

We can then rewrite this as

$$\mu = \frac{1}{n} \sum_{j=1}^n x_j$$

(c) We now assume Σ is $\mathrm{diag}(\sigma_1, ..., \sigma_d)$. Over n samples, the log-likelihood would then be

$$l(\mu, \sigma_1, ..., \sigma_d) = -\frac{nd}{2}\log(2\pi) - n\sum_j^d \log \sigma_j - \frac{1}{2}\sum_i^n \sum_j^d \frac{(x_{ij} - \mu_j)^2}{\sigma_j^2}$$

From the previous part, we already have $\hat{\mu}$, so we now differentiate with respect to each σ_k,

$$\frac{\partial l}{\partial \sigma_k} = -\frac{n}{\sigma_k} + \sum_i^n \frac{(x_{ik} - \mu_k)^2}{\sigma_k^3} = 0$$

Which gives us

$$\sigma_k^2 = \frac{1}{n}\sum_i^n (x_{ik} - \hat{\mu}_k)^2 \rightarrow \hat{\sigma_j} = \sqrt{\frac{1}{n}\sum_i^n (x_{ij} - \hat{\mu}_j)^2}$$

(d) We are given that for one sample, the density is

$$\frac{1}{(2\pi)^{d/2}(\alpha)^{d/2}|\sigma_0|^{1/2}} \exp\left(-\frac{1}{2\alpha}(x-\mu)^T \sigma_0^{-1}(x-\mu)\right)$$

Therefore, the log-likelihood over n samples is

$$-\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\Sigma_0 - \frac{nd}{2}\log(\alpha) - \frac{1}{2\alpha}\sum_i^n (x_i - \mu)^T \Sigma_0^{-1}(x_i - \mu)$$

We then get the score equation for α and solve for $\hat{\alpha}$

$$\frac{\partial l}{\partial \alpha} = -\frac{nd}{2}\frac{1}{\alpha} + \frac{1}{2\alpha^2}\sum_i^n (x_i - \mu)^T \Sigma_0^{-1}(x_i - \mu) = 0$$

$$-nd\alpha + \sum_i^n (x_i - \mu)^T \Sigma_0^{-1} (x_i - \mu) = 0$$

$$\hat{\alpha} = \frac{1}{nd} \sum_i^n (x_i - \mu)^T \Sigma_0^{-1} (x_i - \mu)$$

2. (a) The likelihood function here is a function of (ignoring some constants)

$$\exp(-\frac{1}{2\sigma^2} \sum_i^n (Y_i - X_i\beta)^2)$$

The log likelihood is therefore of the form, with C as a constant

$$l = -\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) + C$$

Taking the partial with respect to β and setting it to 0 yields

$$-2X^T(Y - X\beta) = 0 \rightarrow X^T X\beta = X^T Y$$

Therefore, $\hat{\beta} = (X^T X)^{-1} X^T Y$

(b) i. $\hat{y} = HY$ is the fitted value vector from least squares. We know by construction that because $\hat{\beta}$ minimizes the sum of squared residuals and thus $\hat{y} = X\hat{\beta}$ provides a projection of Y onto the column space of X, making it the least squares estimate.

ii. We know $HX = (X(X^T X)^{-1} X^T)X = X((X^T X)^{-1}(X^T X)) = X$ since $X^T X$ is invertible

iii. Consider the transpose of H,

$$H^T = (X(X^T X)^{-1} X^T)^T = X(X^T X)^{-1} X^T$$

since $(X^T)^T = X$ and $(X^T X)^{-1}$ is symmetric. Therefore, $H^T = T$.

iv. Consider

$$H^2 = H \cdot H = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X I_d (X^T X)^{-1} X^T = H$$

v. The column space of X is by definition all vectors that can be written as linear combinations of the column vectors of X. We constructed H such that for any y, $\hat{y} = Hy$ minimizes the least squares distance. In particular, we picked H to be the vector of coefficients of the orthogonal projection of y onto L. This can be verified by calculating Hz for z that is either a scalar multiple of a vector in X ($Hz = z$) or for z that is orthogonal to L ($Hz = X(X^T X)^{-1} X^T z = 0$).

vi. Since X is a $n \times d$ matrix with full column rank, it has rank d. H is a projection onto the d-dimensional subspace L, so it also has rank d. We also know that H is idempotent though, and so the trace is simply the sum of its eigenvalues (which are either 0 or 1 for a projection matrix like H) and we know d of the eigenvalues are 1, so the trace of H is d.

vii.
$$e = Y - \hat{y} = Y - HY = (I - H)Y$$

1^T is in L, so $1^T H = 1$. Therefore,

$$1^T Y - 1^T (HY) = Y - Y = 0$$

3. (a) In the case that $d \leq n$, we have that the first d rows contain the singular values and the other $n - d$ rows are all 0. In the case that $d > n$, the first n columns contain the nonzero singular values and the remaining $d - n$ columns are all 0.

(b) Given the SVD, consider

$$XX^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma V^T V \Sigma^T U^T = U\Sigma\Sigma^T U^T$$

Since $\Sigma\Sigma^T$ is simply a diagonal matrix with entries $\sigma_1^2, ..., \sigma_k^2$, the columns of U are eigenvectors of XX^T with corresponding eigenvalues σ_i^2.

Similarly, consider

$$X^T X = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^T U^T U \Sigma V^T = V\Sigma^T \Sigma V^T$$

$\Sigma^T \Sigma$ is $d \times d$ (instead of $n \times n$ as above) diagonal matrix with entries $\sigma_1^2, ..., \sigma_k^2$ and so the columns of V are eigenvectors of $X^T X$ with eigenvalues σ_i^2.

(c) The SVD gives us $X = \sum_i^k \sigma_i u_i v_i^T$. Consider for arbitrary i,

$$Xv_i = \sum_j^k \sigma_j u_j v_j^T v_i = \sigma_i u_i$$

This is because only the $j = i$ term survives.

We also have that

$$X^T u_i = \sum_j^k \sigma_j v_j u_j^T u_i = \sigma_i v_i$$

(d)

$$||X||_F^2 = \sum_i^n \sum_j^d X_{ij}^2 = \sum_i (\sum_j \sigma_j v_j u_j^T)(\sum_j \sigma_j u_j v_j^T) = \sum_i (\sum_j \sigma_j^2 v_j^T v_j) = \sum_j \sigma_j^2$$

(e) Since U is unitary,

$$|X| = \max_{v:|v|_2=1} |Xv|_2 = \max |U\Sigma V^T v| = \max |\Sigma V^T v|$$

Then let $y = V^T v$, we then have that because V is unitary that $|y|_2 = 1$ and so $\max |\Sigma V^T v| = \max |\Sigma y|$. And since Σ is a diagonal matrix with σ_1 as the largest singular value, the max is attained when $y = (1, 0, 0, ...)^T$.

(f) Given SVD, we know the eigenvalues in absolute value are exactly the singular values, so we have

$$|\det(X)| = \prod_i^n \sigma_i$$

(g) With $H = X(X^T X)^{-1} X^T$, we consider

$$X^T X = V(\Sigma^T \Sigma) V^T$$

$$(X^T X)^{-1} = V(\Sigma^T \Sigma)^{-1} V^T$$

by invertibility. We then have that

$$H = U\Sigma V^T (V(\Sigma^T \Sigma)^{-1} V^T) V \Sigma^T U^T = U\Sigma(\Sigma^T \Sigma)^{-1} \Sigma^T U^T$$

4. See coding submission

5. See coding submission