CMSC 25025 / STAT 37601

HW1 - DUE: April 9, 2019

Student: Dimitrios Tanoglidis

In collaboration with: Ying Cai

Problem 2: Maximum Likelihood

(a) The log-likelihood is:

$$\begin{aligned}
\ell(\vec{X};\lambda) &= \sum_{i=1}^{n}\log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) \\
&= \sum_{i=1}^{n}\left(\log e^{-\lambda}+\log\lambda^{X_i}-\log X_i!\right) \\
&= \sum_{i=1}^{n}(-\lambda)+\log\lambda\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\log X_i! \\
&= \log\lambda\sum_{i=1}^{n}X_i-\lambda n-\sum_{i=1}^{n}\log X_i! \qquad (1)
\end{aligned}$$

The score equation is the derivative of (1) with respect to λ. We set this derivative equal to zero and we have:

$$\frac{\partial\ell(\vec{X};\lambda)}{\partial\lambda}=\frac{1}{\lambda}\sum_{i=1}^{n}X_i-n=0\Rightarrow \qquad (2)$$

$$\boxed{\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}X_i} \qquad (3)$$

(b) The Normal distribution in d dimensions is:

$$\mathcal{N}(X_i;\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\vec{X}_i-\vec{\mu})^T\Sigma^{-1}(\vec{X}_i-\vec{\mu})\right] \qquad (4)$$

So, the log-likelihood in this case is $\ell(\vec{X};\mu,\Sigma)=\sum_{i=1}^{n}\log\mathcal{N}(X_i;\mu,\Sigma)$, or:

$$\boxed{\ell(\vec{X};\mu,\Sigma)=-\frac{n}{2}\log|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}(\vec{X}_i-\vec{\mu})^T\Sigma^{-1}(\vec{X}_i-\vec{\mu})+\text{const.}} \qquad (5)$$

Now, we have to solve the score equation:

$$\frac{\partial\ell(\vec{X};\mu,\Sigma)}{\partial\mu}=0 \qquad (6)$$

Here we will use the following matrix calculus identity: Let y, x vectors and A matrix. Then, if $\alpha=y^TAx$, we have that:

$$\frac{\partial\alpha}{\partial z}=x^TA^T\frac{\partial y}{\partial z}+y^TA\frac{\partial x}{\partial z} \qquad (7)$$

Working for $x=y=X_i-\mu$ and $A=\Sigma^{-1}=(\Sigma^{-1})^T$, we have using the above equation:

$$
\begin{aligned}
\frac{\partial \ell(\vec{X};\mu,\Sigma)}{\partial \mu} &= -\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}\frac{\partial}{\partial \mu}(X_i-\mu) - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}\frac{\partial}{\partial \mu}(X_i-\mu) \\
&= \sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1} \qquad (8)
\end{aligned}
$$

And setting this equal to zero, we have:

$$
\Sigma^{-1}\sum_{i=1}^{n}(X_i-\mu) = 0 \Rightarrow \boxed{\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i} \qquad (9)
$$

Using that Σ is positive definite matrix.

(c) Here $\Sigma^{-1} = diag(\sigma_1^{-1}, \ldots, \sigma_d^{-1})$.
Also:

$$
|\Sigma| = \det \Sigma = \prod_{i=1}^{d} \sigma_i \qquad (10)
$$

So, the first term of the log-likelihood (5), can be written as:

$$
\frac{n}{2}\log|\Sigma| = \frac{n}{2}\log\prod_{i=1}^{d}\sigma_i = \frac{n}{2}\sum_{i=1}^{d}\log\sigma_i \qquad (11)
$$

Also, we can now write the second term of the log-likelihood (5) as a double sum:

$$
-\frac{1}{2}\sum_{i=1}^{n}(\vec{X}_i-\vec{\mu})^T\Sigma^{-1}(\vec{X}_i-\vec{\mu}) = -\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\sigma_j^{-1}(X_i-\mu)_i^2 \qquad (12)
$$

So, we can now write the total log-likelihood as:

$$
\ell(\vec{X};\mu,\Sigma) = -\frac{n}{2}\sum_{i=1}^{d}\log\sigma_i - \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\sigma_j^{-1}(X_i-\mu)_j^2 \qquad (13)
$$

Now, we can derive and solve the score equation:

$$
\begin{aligned}
\frac{\partial \ell(\vec{X};\mu,\Sigma)}{\partial \sigma_k} &= -\frac{n}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i}\delta_{ik} + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{1}{\sigma_j^2}\delta_{jk}(X_i-\mu)_j^2 \\
&= -\frac{n}{2}\frac{1}{\sigma_k} + \frac{1}{\sigma_k^2}\sum_{i=1}^{d}(X_i-\mu)_k^2 \qquad (14)
\end{aligned}
$$

Setting this equal to zero, we have:

$$
-\frac{n}{2}\frac{1}{\sigma_k} + \frac{1}{\sigma_k^2}\sum_{i=1}^{d}(X_j-\mu)_k^2 = 0 \Rightarrow \hat{\sigma}_k = \frac{1}{n}\sum_{i=1}^{d}(X_i-\hat{\mu})_k^2 \qquad (15)
$$

or:

$$
\boxed{\hat{\sigma}_i = \frac{1}{n}\sum_{j=1}^{d}(X_j-\hat{\mu})_i^2} \qquad (16)
$$

(d) In this case $|\Sigma| = |\alpha\Sigma_0| = \alpha^d|\Sigma_0|$ and $\Sigma^{-1} = \alpha^{-1}\Sigma_0^{-1}$, so, the log-likelihood is:

$$\ell = -\frac{nd}{2}\log\alpha - \frac{n}{2}\log|\Sigma_0| - \frac{1}{2}\sum_{i=1}^{d}\alpha^{-1}(X_i-\mu)^T\Sigma_0^{-1}(X_i-\mu) + \text{const} \tag{17}$$

Obtaining and solving the score equation, we have:

$$\frac{\partial\ell}{\partial\alpha} = -\frac{nd}{2}\frac{1}{\alpha} + \frac{1}{2}\frac{1}{\alpha^2}\sum_{i=1}^{d}(X_i-\mu)^T\Sigma_0^{-1}(X_i-\mu) \tag{18}$$

So:

$$\boxed{\hat{\alpha} = \frac{1}{nd}\sum_{i=1}^{d}(X_i-\mu)^T\Sigma_0^{-1}(X_i-\mu)} \tag{19}$$

Problem 3: Regression

(a) To get the least square estimate of y, we have to minimize $||y-\hat{y}||^2$, which is :

$$(y-X\hat{\beta})^T(y-X\hat{\beta}) = y^Ty - 2\hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta} \tag{20}$$

Setting the derivative with respect to $\hat{\beta}$ equal to zero, gives:

$$-2X^Ty + 2X^TX\hat{\beta} = 0 \tag{21}$$

i.e.

$$X^Ty = X^TX\hat{\beta} \tag{22}$$

multiply both sides of the equation with $X(X^TX)^{-1}$:

$$X(X^TX)^{-1}X^Ty = X(X^TX)^{-1}X^TX\hat{\beta} \tag{23}$$

Which is exactly:

$$Hy = X\hat{\beta} = \hat{y} \tag{24}$$

(b)

$$\begin{aligned} HX &= X(X^TX)^{-1}X^TX \\ &= X[(X^TX)^{-1}(X^TX)] \\ &= X \end{aligned} \tag{25}$$

(c)

$$\begin{aligned} H^T &= [X(X^TX)^{-1}X^T]^T \\ &= X[(X^TX)^{-1}]^TX^T \\ &= X[(X^TX)^T]^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned} \tag{26}$$

(d)

$$\begin{aligned} H^2 &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= X[(X^TX)^{-1}(X^TX)](X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned} \tag{27}$$

(e) Since

$$\hat{y} = Hy = X(X^TX)^{-1}X^Ty = Xv, \tag{28}$$

where $v = (X^TX)^{-1}X^Ty$ is a vector in R^d.
Therefore, $\hat{y}$ is a linear combination of the columns of X,
i.e.

$$\hat{y} \in \mathcal{L} \tag{29}$$

and $\hat{y} = Hy$ is the projection of y onto $\mathcal{L}$.

(f) Since (X^TX) is nonsingular, $(X^TX)^{-1}$ exists, i.e. X has full rank:

$$\text{rank}(X) = \min(n, d) = d \tag{30}$$

Also:

$$\begin{aligned}
\text{tr}(H) &= \text{tr}(X(X^TX)^{-1}X^T) \\
&= \text{tr}(X^TX(X^TX)^{-1}) \\
&= \text{tr}(I_d) \\
&= d
\end{aligned} \tag{31}$$

Therefore,

$$\text{rank}(X) = d = \text{tr}(H) \tag{32}$$

Problem 4: Singular Value Decomposition

(a) It is enough to show that:

$$(XX^T)U = U\hat{\Lambda}, \tag{33}$$

Where $\hat{\Lambda} = \lambda_i\delta_{ij}$, $m \times m$ diagonal matrix with the eigenvalues λ_i across the diagonal. Indeed, from the decomposition of X as $X = U\Sigma V^T$, we have:

$$\begin{aligned}
(XX^T)U &= U\Sigma V^TV\Sigma^TU^TU \\
&= U(\Sigma\Sigma^T),
\end{aligned} \tag{34}$$

where we used the property of the orthogonal matrices, that $U^TU = I = V^TV$. Now, we can identify the $m \times m$ matrix $\Sigma\Sigma^T$ with $\hat{\Lambda}$, since:

$$(\Sigma\Sigma^T) = diag(\sigma_i^2, 0, \ldots, 0) \tag{35}$$

where the zero eigenvalues are there only if $r < m$. So, eigenvalues are σ_i^2 for $r < m$ and 0 for $r \leq i \leq m$.
Similarly, we get that:

$$(X^TX)V = V(\Sigma^T\Sigma). \tag{36}$$

Now, the matrix of the eigenvalues is the diagonal $n \times n$ matrix $\Sigma^T\Sigma$ with entries σ_i^2 for $i \leq r < n$ and 0 otherwise.

(b) Working in a similar way as in (a), we get:

$$XV = U\Sigma V^TV = U\Sigma, \tag{37}$$

So, now the matrix of the eigenvalues is the matrix Σ (which has the singular values σ_i along the diagonal), so, equivalently $Xv_i = \sigma_i u_i$. Similarly, $X^T U = V\Sigma^T U^T U = V\Sigma T$, again a diagonal matrix $\hat{\Lambda} = \Sigma^T$, so it easily follows that $X^T u_i = \sigma_i v_i$.

(c) We start by expressing the elements X_{ij} in terms of the decomposition. Using the fact that for a matrix $C = AB$, $C_{ij} = \sum_\ell A_{i\ell}B_{\ell j}$, we have:

$$X_{ij} = \sum_\ell U_{i\ell}(\Sigma V^T)_{\ell j} = \sum_\ell \sum_k U_{i\ell}\Sigma_{\ell k}V^T_{kj} = \sum_\ell \sum_k U_{i\ell}\Sigma_{\ell k}V_{jk}. \tag{38}$$

Now, we use that $\Sigma_{\ell k} = \sigma_\ell \delta_{\ell k}$ and thus, the above double sum reduces to a simple sum:

$$X_{ij} = \sum_\ell \sigma_\ell U_{i\ell}V_{j\ell} \tag{39}$$

Now, we can write:

$$\sum_{ij} X^2_{ij} = \sum_{ij} \sum_{\ell m} \sigma_\ell \sigma_m U_{i\ell}V_{j\ell}U_{im}V_{jm}, \tag{40}$$

where in expressing the second X_{ij} as a sum, I changed ℓ to m. Now, rearranging the terms appropriately, we have:

$$\sum_{ij} X^2_{ij} = \sum_{\ell m} \sigma_\ell \sigma_m \sum_i U_{i\ell}U_{im} \sum_j V_{j\ell}V_{jm} \tag{41}$$

It is easy to see that: $\sum_i U_{i\ell}U_{im} = (U^TU)_{\ell m} = \delta_{\ell m}$, and similarly $\sum_j V_{j\ell}V_{jm} = (V^TV)_{\ell m} = \delta_{\ell m}$ (from the orthogonality of U and V). Thus the sum reduces to:

$$\sum_{ij} X^2_{ij} = \sum_\ell \sigma^2_\ell \tag{42}$$

So, for the Frobenius norm we have:

$$\boxed{\|X\|_F = \sqrt{\sum_{ij} X^2_{ij}} = \sqrt{\sum_{i=1}^r \sigma^2_i}} \tag{43}$$

(d) Here I will assume that X is a square matrix, in order to be able to define the determinant X. Then using the properties of the determinants, we have:

$$|\det(X)| = |\det(U\Sigma V^T)| = |\det(U)||\det(\Sigma)||\det(V^T)| \tag{44}$$

Now $|\det(V^T)| = |\det(V)|$. Also:

$$|\det(U^TU)| = |\det(I)| = 1 = |\det(U^T)||\det(U)| = |\det(U)|^2 \Rightarrow |\det(U)| = \pm 1 \tag{45}$$

and similarly for $|\det(V)|$. Furthermore, since now we assume that σ is a diagonal $m \times m$ matrix:

$$|\det(\Sigma)| = \prod_{i=1}^m \sigma_i, \tag{46}$$

and thus:

$$\boxed{|\det(X)| = \prod_{i=1}^m \sigma_i} \tag{47}$$

(e) We have in a straightforward way:

$$\begin{aligned}
H &= U\Sigma V^T(V\Sigma^T U^T U\Sigma V^T)^{-1}V\Sigma^T U^T \\
&= U\Sigma V^T(V^T)^{-1}(\Sigma^T\Sigma)^{-1}V^{-1}V\Sigma^T U^T \\
&= U\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T U^T
\end{aligned} \tag{48}$$

So, the hat matrix is

$$\boxed{H = U\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T U^T} \tag{49}$$

Note, that in the case where Σ is square and invertible, this further reduces to $H = UU^T$.

(f) Least square regression estimate is:

$$\hat{\beta} = (X^T X)^{-1}X^T y. \tag{50}$$

Write now $X = U\Sigma_{(k)}V^T$. (Here I put (k) as a subscript, for convenience to write $\Sigma^T_{(k)}$, its transpose).

Then, we have:

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1}X^T\,y \\
&= (V\Sigma^T_{(k)}U^T U\Sigma_{(k)}V^T)^{-1}V\Sigma^T_{(k)}U^T\,y \\
&= V^T(\Sigma^T_{(k)}\Sigma_{(k)})^{-1}V^{-1}V\Sigma^T_{(k)}U^T\,y \\
&= V^T(\Sigma^T_{(k)}\Sigma_{(k)})^{-1}\Sigma^T_{(k)}U^T\,y
\end{aligned} \tag{51}$$

Again, if the matrix $\Sigma_{(k)}$ is square and invertible, this gives the nice result: $\hat{\beta} = V^T\Sigma^{-1}U^T\,y$