Math 534 Homework 3 Problem 1

Problem 1: Exercise J-2.2

Derive the Gradient, Hessian and Fisher Information Matrix

Solution

Differentials, Derivatives, Gradient, and Hessian
Here we calculate both the gradient and the hessian. To derive these values we begin by calculating differentials. We have:

$$\begin{aligned}
d\ell(d\mu) =& trace\left(\Sigma^{-1} \sum^{n} (x_i - \mu) d\mu^T \right) \\
dd\ell(d\mu, d\mu) =& -nd\mu^T \Sigma^{-1} d\mu \\
d\ell(d\Sigma) =& -(1/2) trace\left\{ \left(n\Sigma^{-1} - \Sigma^{-1} \sum^{n} (x_i - \mu)(x_i - \mu)^t \Sigma^{-1} \right) d\Sigma \right\} = -(1/2) trace(Ad\Sigma) \\
dd\ell(d\mu, d\Sigma) =& -trace\left\{ \Sigma^{-1} d\Sigma \Sigma^{-1} \sum^{n} (x_i - \mu) d\mu^T \right\} = -trace\left\{ \Sigma^{-1} d\Sigma C d\mu^T \right\} \\
dd\ell(d\Sigma, d\Sigma) =& (-1/2) trace\left\{ \left(-nI + 2\Sigma^{-1} \sum^{n} (x_i - \mu)(x_i - \mu)^t \right) \Sigma^{-1} d\Sigma \Sigma^{-1} d\Sigma \right\} = (-1/2) trace\left\{ Z d\Sigma \Sigma^{-1} d\Sigma \right\}
\end{aligned}$$

Now that we have all our differentials, lets begin by deriving the values of the gradient. Recall that our gradient has p many parameters for μ and $p(p+1)/2$ parameters for σ. We have:

$$\begin{aligned}
\frac{\partial}{\partial \mu_i} \ell =& \left\{ \Sigma^{-1} \sum (x_i - \mu) \right\}_i \\
\frac{\partial}{\partial \sigma_{ii}} =& -\frac{1}{2} \left\{ A \right\}_{ii} \\
\frac{\partial}{\partial \sigma_{ij}} =& -\frac{1}{2} \left\{ A \right\}_{ij} - \frac{1}{2} \left\{ A \right\}_{ji} = -\left\{ A \right\}_{ij}
\end{aligned}$$

The elements of the hessian are derived from the following:

$$\frac{\partial^2}{\partial \mu_i \partial \mu_j} = \left\{ -n\Sigma^{-1} \right\}_{ij},$$

$$\frac{\partial^2}{\partial \mu_i \partial \sigma_{kl}} = -\begin{cases}
\text{Case 1} : k = l & \left\{ \Sigma^{-1} \right\}_{ik} \left\{ C \right\}_k \\
\text{Case 2} : i = k, i \neq l & \left\{ \Sigma^{-1} \right\}_{ik} \left\{ C \right\}_l + \left\{ \Sigma^{-1} \right\}_{il} \left\{ C \right\}_k \\
\text{Case 3} : i \neq k, i = l & \left\{ \Sigma^{-1} \right\}_{ik} \left\{ C \right\}_l + \left\{ \Sigma^{-1} \right\}_{il} \left\{ C \right\}_k \\
\text{Case 4} : i \neq k \neq l & \left\{ \Sigma^{-1} \right\}_{ik} \left\{ C \right\}_l + \left\{ \Sigma^{-1} \right\}_{il} \left\{ C \right\}_k
\end{cases}$$

Note that Cases 2, 3, 4 have the same formula, and the main distinction between Case 1 and the other three cases is that in case 1, $k = l$, but in the remaining cases *k is not equal to l.* So when programming you will need to only consider two cases.

$$\frac{\partial^2}{\partial\sigma_{ij}\partial\sigma_{kl}} = (-1/2)\times \begin{cases} \text{Case } 1 : i=j, l=k & \left\{Z\right\}_{ki}\left\{\Sigma^{-1}\right\}_{ik} \\ \text{Case } 2 : i\neq j, k\neq l & \left\{Z\right\}_{ki}\left\{\Sigma^{-1}\right\}_{jl} + \left\{Z\right\}_{lj}\left\{\Sigma^{-1}\right\}_{ik} + \left\{Z\right\}_{kj}\left\{\Sigma^{-1}\right\}_{il} + \left\{Z\right\}_{li}\left\{\Sigma^{-1}\right\}_{jk} \\ \text{Case } 3 : i\neq j, k=l & \left\{Z\right\}_{ki}\left\{\Sigma^{-1}\right\}_{jk} + \left\{Z\right\}_{kj}\left\{\Sigma^{-1}\right\}_{ik} \\ \text{Case } 4 : i=j, k\neq l & \left\{Z\right\}_{li}\left\{\Sigma^{-1}\right\}_{ik} + \left\{Z\right\}_{ki}\left\{\Sigma^{-1}\right\}_{il} \end{cases}$$

Fisher Information Matrix

The Fisher information matrix is calculated by the quantity $E[-\nabla^2\ell(\theta)]$. In order to calculate this we take the expectation of the differentials from above. Using the fact that X_i's are randomly sampled from a p-variate normal distribution, we have $E[X_i-\mu]=0$ and

$$E[\sum^n (X_i-\mu)(X_i-\mu)^T] = E[\sum^n \begin{bmatrix} (X_1-\mu_1)^2 & ... & (X_1-\mu_1)(X_p-\mu_{1p}) \\ ... & ... & ... \\ (X_p-\mu_p)(X_1-\mu_1) & ... & (X_p-\mu_p)^2 \end{bmatrix} = n\Sigma$$

Hence,

$$\begin{aligned} E[dd\ell(d\mu, d\mu)] &= -nd\mu^T\Sigma^{-1}d\mu \\ E[dd\ell(d\mu, d\Sigma)] &= 0 \\ E[dd\ell(d\Sigma, d\Sigma)] &= (-n/2)trace\left\{\Sigma^{-1}d\Sigma\Sigma^{-1}d\Sigma\right\} = (-n/2)trace\left\{Ud\Sigma\Sigma^{-1}d\Sigma\right\} \end{aligned}$$

which implies

$$\begin{aligned} E[\frac{\partial^2}{\partial\mu_i\partial\mu_j}] &= \left\{-n\Sigma^{-1}\right\}_{ij} \\ E[\frac{\partial^2}{\partial\sigma_{ij}\partial_{\mu j}}] &= 0 \end{aligned}$$

and

$$\frac{\partial^2}{\partial\sigma_{ij}\partial\sigma_{lk}} = (-n/2)\times \begin{cases} \text{Case } 1 : i=j, l=k & \left\{U\right\}_{ki}\left\{\Sigma^{-1}\right\}_{ik} \\ \text{Case } 2 : i\neq j, k\neq l & \left\{U\right\}_{ki}\left\{\Sigma^{-1}\right\}_{jl} + \left\{U\right\}_{lj}\left\{\Sigma^{-1}\right\}_{ik} + \left\{U\right\}_{kj}\left\{\Sigma^{-1}\right\}_{il} + \left\{U\right\}_{li}\left\{\Sigma^{-1}\right\}_{jk} \\ \text{Case } 3 : i\neq j, k=l & \left\{U\right\}_{ki}\left\{\Sigma^{-1}\right\}_{jk} + \left\{U\right\}_{kj}\left\{\Sigma^{-1}\right\}_{ik} \\ \text{Case } 4 : i=j, k\neq l & \left\{U\right\}_{li}\left\{\Sigma^{-1}\right\}_{ik} + \left\{U\right\}_{ki}\left\{\Sigma^{-1}\right\}_{il} \end{cases}$$