Derivation of fisher information of a multivariate Gaussian density

Kevin W Li

March 7, 2017

We are interested to find the fisher information of this density with respect to parameter s upon which μ and Σ depends. This notes assumes fluency in matrix calculus (denominator layout) in the matrix cookbook or wikipedia.

Preliminaries

For scalar x, scalar function f, vectors a, b, c, and matrices A, B, C:

$$\begin{aligned}
d\frac{a^\intercal Ab}{da} &= Ab \\
d\frac{a^\intercal Ab}{dA} &= d\frac{Tr(a^\intercal Ab)}{dA} = d\frac{Tr(Aa^\intercal b)}{dA} = ab^\intercal \\
d\frac{f(b(x))}{dx} &= \left[\frac{df(b)}{db}\right]^\intercal \nabla_x b \\
d\frac{f(A(x))}{dx} &= Tr\left[\left(\frac{dx}{dA}\right)^\intercal \nabla_x A\right] \\
d\frac{a(c)^\intercal b}{dc} &= \frac{da^\intercal}{dc} b \\
d\frac{\log A}{dx} &= A^{-T}\nabla_x A \\
dA^{-1} &= -A^{-1}dAA^{-1} \\
d\frac{A^{-1}}{dx} &= -A^{-1}\nabla AA^{-1} \\
d\frac{Tr(AB^{-1}C)}{dB} &= -B^{-T}CAB^{-T}
\end{aligned}$$

In addition, the first derivative of a vector w.r.t a scalar is a vector, first derivative of a marix w.r.t a scalar is a matrix.

Derivation

The multivariate Gaussian density is written as:

$$p(x) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right) = \frac{|\Lambda|^{1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \Lambda(x-\mu)\right)$$

where Λ is the precision matrix and is symmetric. Define

$$\nabla\mu = \frac{d\mu}{ds}$$
$$\nabla\Sigma = \frac{d\Sigma}{s}$$
$$\nabla\Lambda = \frac{d\Lambda}{ds} = \frac{d\Sigma^{-1}}{ds} = -\Sigma^{-1}\nabla\Sigma\Sigma^{-1}$$

In the following, we will use the fact that $\Lambda = \Lambda^\intercal$

And write the log density as:

$$\log p(x) = \frac{1}{2}\log|\Lambda| - \frac{1}{2}(x-\mu)^\intercal\Lambda(x-\mu) + constant$$

There two ways ahead. One can write out the second derivative of the log density using the chain rule. It will be composed of four terms in which two of them will involve a score term $\frac{d\log p(x)}{d\mu}$ and $\frac{d\log p(x)}{d\Lambda}$. The expectation of these two terms will be zeros, so we ended up having only two terms. However, $\frac{d^2\log p(x)}{d\mu^2}$ is a matrix and $\frac{d^2\log p(x)}{d\Lambda^2}$ is a four-dimentional tensor which we should avoid... It is still possible to work, but less clear.

The second way is to compute the first derivative first in terms of the parameters, and then take derivative again.

$$\begin{aligned}\frac{d\log p(x)}{ds} &= \left[\frac{d\log p(s)}{d\mu}\right]^\intercal\nabla\mu + Tr\left[\frac{d\log p(s)}{d\Lambda}^\intercal\nabla\Lambda\right] \\ &= \underbrace{(x-\mu)^\intercal\Lambda\nabla\mu}_{M_1} + \frac{1}{2}\underbrace{Tr\left[\left[\Lambda^{-1} - (x-\mu)(x-\mu)^\intercal\right]\nabla\Lambda\right]}_{M_2}\end{aligned}$$

Taking the square yields something hard to proceed. Let's continue with the other equivavlent definition of FI that uses second derivatives.

Take the derivative of M_1 and then evaluate the negative expectation

$$\begin{aligned}\frac{dM_1}{ds} &= \left[\frac{dM_1}{d\mu}\right]^\intercal\nabla\mu + Tr\left[\left[\frac{dM_1}{d\Lambda}\right]^\intercal\nabla\Lambda\right] \\ &= \left[-\Lambda\nabla\mu + \nabla^2\mu\Lambda(x-\mu)\right]^\intercal\nabla\mu + Tr\left[[(x-\mu)\nabla\mu^\intercal]^\intercal\nabla\Lambda\right] \\ -\left\langle\frac{dM_1}{ds}\right\rangle &= \nabla\mu^\intercal\Lambda\nabla\mu = \nabla\mu^\intercal\Sigma^{-1}\nabla\mu\end{aligned}$$

Do the same for M_2

$$\begin{aligned}\frac{dM_2}{ds} &= \left[\frac{dM_2}{d\mu}\right]^\intercal\nabla\mu + Tr\left[\left[\frac{dM_2}{d\Lambda}\right]^\intercal\nabla\Lambda\right] \\ &= -\left[\frac{d}{d\mu}Tr\left[(x-\mu)(x-\mu)^\intercal\nabla\Lambda\right]\right]^\intercal\nabla\mu + Tr\left[\left[\frac{d}{d\Lambda}Tr\left[\Lambda^{-1}\nabla\Lambda\right] - \frac{d}{d\Lambda}Tr\left[(x-\mu)(x-\mu)^\intercal\nabla\Lambda\right]\right]^\intercal\nabla\Lambda\right] \\ &= -2(x-\mu)^\intercal\nabla\Lambda\nabla\mu + Tr\left[\left[-\Lambda^{-1}\nabla\Lambda\Lambda^{-1} + \Lambda^{-1}\nabla^2\Lambda - (x-\mu)(x-\mu)^\intercal\nabla^2\Lambda\right]^\intercal\nabla\Lambda\right] \\ &= -2(x-\mu)^\intercal\nabla\Lambda\nabla\mu + Tr\left[-\Lambda^{-1}\nabla\Lambda\Lambda^{-1}\nabla\Lambda\right] + Tr\left[\left[\Lambda^{-1} - (x-\mu)(x-\mu)^\intercal\right]\nabla^2\Lambda\nabla\Lambda\right] \\ -\left\langle\frac{dM_2}{ds}\right\rangle &= 0 + Tr\left[-\Lambda^{-1}\nabla\Lambda\Lambda^{-1}\nabla\Lambda\right] + Tr\left[\left[\Lambda^{-1}\nabla^2\Lambda - \Sigma\nabla^2\Lambda\right]\nabla\Lambda\right] \\ -\left\langle\frac{dM_2}{ds}\right\rangle &= Tr\left[\Lambda^{-1}\nabla\Lambda\Lambda^{-1}\nabla\Lambda\right] = Tr\left[\Lambda^{-1}\Sigma^{-1}\nabla\Sigma\Sigma^{-1}\Lambda^{-1}\Sigma^{-1}\nabla\Sigma\Sigma^{-1}\right] = Tr\left[\nabla\Sigma\Sigma^{-1}\nabla\Sigma\Sigma^{-1}\right]\end{aligned}$$

The last equality is because the first contains first moment, and the third term is zero even.

So the fisher information is then

$$J(s) = \nabla\mu^{\intercal}\Sigma^{-1}\nabla\mu + \frac{1}{2}Tr\left[\nabla\Sigma\Sigma^{-1}\nabla\Sigma\Sigma^{-1}\right]$$