Fisher information matrix for Gaussian and categorical distributions

Jakub M. Tomczak

November 28, 2012

1 Notations

Let x be a random variable. Consider a parametric distribution of x with parameters $\boldsymbol{\theta}$, $p(x|\boldsymbol{\theta})$. The contiuous random variable $x \in \mathbb{R}$ can be modelled by *normal distribution* (*Gaussian distribution*):

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{(x-\mu)^2}{2\sigma^2} \Big\} \\ &= \mathcal{N}(x|\mu, \sigma^2), \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} = \big(\mu \ \sigma^2\big)^{\mathrm{T}}$.

A discrete (categorical) variable $x \in \mathcal{X}$, $\mathcal{X}$ is a finite set of K values, can be modelled by *categorical distribution*:[1]

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= \prod_{k=1}^{K} \theta_k^{x_k} \\ &= \mathrm{Cat}(x|\boldsymbol{\theta}), \end{aligned} \tag{2}$$

where $0 \le \theta_k \le 1$, $\sum_k \theta_k = 1$.

For $\mathcal{X} = \{0, 1\}$ we get a special case of the categorical distribution, *Bernoulli distribution*,

$$\begin{aligned} p(x|\theta) &= \theta^x (1-\theta)^{1-x} \\ &= \mathrm{Bern}(x|\theta). \end{aligned} \tag{3}$$

2 Fisher information matrix

2.1 Definition

The *Fisher score* is determined as follows [1]:

$$g(\boldsymbol{\theta}, x) = \nabla_{\boldsymbol{\theta}} \ln p(x|\boldsymbol{\theta}). \tag{4}$$

The Fisher information matrix is defined as follows [1]:

$$\mathbf{F} = \mathbb{E}_x\big[g(\boldsymbol{\theta}, x)\ g(\boldsymbol{\theta}, x)^{\mathrm{T}}\big]. \tag{5}$$

[1] We use the 1-of-K encoding [1].

2.2 Example 1: Bernoulli distribution

Let us calculate the fisher matrix for Bernoulli distribution (3). First, we need to take the logarithm:

$$\ln \mathrm{Bern}(x|\theta) = x \ln \theta + (1-x)\ln(1-\theta). \tag{6}$$

Second, we need to calculate the derivative:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathrm{Bern}(x|\theta) &= \frac{x}{\theta} - \frac{1-x}{1-\theta} \\ &= \frac{x-\theta}{\theta(1-\theta)}.\end{aligned} \tag{7}$$

Hence, we get the following Fisher score for the Bernoulli distribution:

$$g(\theta, x) = \frac{x-\theta}{\theta(1-\theta)}. \tag{8}$$

The Fisher information matrix (here it is a scalar) for the Bernoulli distribution is as follows:

$$\begin{aligned}\mathrm{F} &= \mathbb{E}_x[g(\theta,x)\ g(\theta,x)] \\ &= \mathbb{E}_x\Big[\frac{(x-\theta)^2}{(\theta(1-\theta))^2}\Big] \\ &= \frac{1}{(\theta(1-\theta))^2}\Big\{\mathbb{E}_x[x^2 - 2x\theta + \theta^2]\Big\} \\ &= \frac{1}{(\theta(1-\theta))^2}\Big\{\mathbb{E}_x[x^2] - 2\theta\mathbb{E}_x[x] + \theta^2\Big\} \\ &= \frac{1}{(\theta(1-\theta))^2}\Big\{\theta - 2\theta^2 + \theta^2\Big\} \\ &= \frac{1}{(\theta(1-\theta))^2}\ \theta(1-\theta) \\ &= \frac{1}{\theta(1-\theta)}.\end{aligned} \tag{9}$$

2.3 Example 2: Categorical distribution

Let us calculate the fisher matrix for categorical distribution (2). First, we need to take the logarithm:

$$\ln \mathrm{Cat}(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} x_k \ln \theta_k. \tag{10}$$

Second, we need to calculate partial derivatives:

$$\frac{\partial}{\partial \theta_k} \ln \mathrm{Cat}(x|\boldsymbol{\theta}) = \frac{x_k}{\theta_k}. \tag{11}$$

Hence, we get the following Fisher score for the categorical distribution:

$$g(\theta, x) = \begin{bmatrix} \frac{x_1}{\theta_k} \\ \vdots \\ \frac{x_K}{\theta_K} \end{bmatrix}. \tag{12}$$

Now, let us calculate the product of Fisher score and its transposition:

$$
\begin{bmatrix} \frac{x_1}{\theta_k} \\ \vdots \\ \frac{x_K}{\theta_K} \end{bmatrix} \begin{bmatrix} \frac{x_1}{\theta_k} & \cdots & \frac{x_K}{\theta_K} \end{bmatrix} = \begin{bmatrix} \left(\frac{x_1}{\theta_1}\right)^2 & \frac{x_1 x_2}{\theta_1 \theta_2} & \cdots & \frac{x_1 x_K}{\theta_1 \theta_K} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{x_K x_1}{\theta_K \theta_1} & \frac{x_K x_2}{\theta_K \theta_2} & \cdots & \left(\frac{x_K}{\theta_K}\right)^2 \end{bmatrix}
$$

$$
= \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1K} \\ \vdots & \vdots & \cdots & \vdots \\ g_{K1} & g_{K2} & \cdots & g_{KK} \end{bmatrix}. \tag{13}
$$

Therefore, for g_{kk} we have:

$$
\begin{aligned}
\mathbb{E}_x[g_{kk}] &= \mathbb{E}_x\left[\left(\frac{x_k}{\theta_k}\right)^2\right] \\
&= \frac{1}{\theta_k^2}\mathbb{E}_x[x^2] \\
&= \frac{1}{\theta_k},
\end{aligned} \tag{14}
$$

and for g_{ij}, $i \neq j$:

$$
\begin{aligned}
\mathbb{E}_x[g_{ij}] &= \mathbb{E}_x\left[\frac{x_i x_j}{\theta_i \theta_j}\right] \\
&= \frac{1}{\theta_i \theta_j}\mathbb{E}_x[x_i x_j] \\
&= 0.
\end{aligned} \tag{15}
$$

Finally, we get:

$$
\mathbf{F} = \mathrm{diag}\left\{\frac{1}{\theta_1}, \ldots, \frac{1}{\theta_K}\right\}. \tag{16}
$$

2.4 Example 3: Normal distribution

Let us calculate the Fisher matrix for univariate normal distribution (1). First, we need to take the logarithm:

$$
\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2. \tag{17}
$$

Second, we need to calculate the partial derivatives:

$$
\frac{\partial}{\partial \mu}\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma^2}(x-\mu) \tag{18}
$$

$$
\frac{\partial}{\partial \sigma^2}\mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2. \tag{19}
$$

Hence, we get the following Fisher score for normal distribution:

$$
\begin{aligned}
g(\boldsymbol{\theta}, x) &= \begin{bmatrix} \frac{\partial}{\partial \mu}\mathcal{N}(x|\mu, \sigma^2) \\ \frac{\partial}{\partial \sigma^2}\mathcal{N}(x|\mu, \sigma^2) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2}(x-\mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2 \end{bmatrix}.
\end{aligned} \tag{20}
$$

Now, let us calculate the product of Fisher score and its transposition:

$$\begin{bmatrix} \frac{1}{\sigma^2}(x-\mu) \\ \\ -\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}(x-\mu)^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2}(x-\mu) & -\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}(x-\mu)^2 \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{\sigma^4}(x-\mu)^2 & -\frac{1}{2\sigma^4}(x-\mu)+\frac{1}{2\sigma^6}(x-\mu)^3 \\ \\ -\frac{1}{2\sigma^4}(x-\mu)+\frac{1}{2\sigma^6}(x-\mu)^3 & \frac{1}{4\sigma^4}-\frac{1}{2\sigma^6}(x-\mu)^2+\frac{1}{4\sigma^8}(x-\mu)^4 \end{bmatrix} = \tag{21}$$

$$\begin{bmatrix} g_{11} & g_{12} \\ \\ g_{21} & g_{22} \end{bmatrix},$$

where $g_{12} = g_{21}$.

In order to calculate the Fisher information matrix we need to determine the expected value of all g_{ij}. Hence,[2] for g_{11}:

$$\begin{aligned} \mathbb{E}_x[g_{11}] &= \mathbb{E}_x\Big(\frac{1}{\sigma^4}(x-\mu)^2\Big) \\ &= \frac{1}{\sigma^2}\big(\mathbb{E}_x[x^2]-2\mu^2+\mu^2\big) \\ &= \frac{1}{\sigma^2}\big(\mu^2+\sigma^2-2\mu^2+\mu^2\big) \\ &= \frac{1}{\sigma^2}, \end{aligned} \tag{22}$$

and for g_{12}:

$$\begin{aligned} \mathbb{E}_x[g_{12}] &= \mathbb{E}_x\Big(-\frac{1}{2\sigma^4}(x-\mu)+\frac{1}{2\sigma^6}(x-\mu)^3\Big) \\ &= -\frac{1}{2\sigma^4}(\mathbb{E}_x[x]-\mu)+\mathbb{E}_x\Big(\frac{1}{2\sigma^6}(x^3-3x^2\mu+3x\mu^2-\mu^3)\Big) \\ &= \frac{1}{2\sigma^6}\Big((\mathbb{E}_x[x^3]-3\mu\mathbb{E}_x[x^2]+3\mu^2\mathbb{E}_x[x]-\mu^3)\Big) \\ &= \frac{1}{2\sigma^6}\Big((\mu^3+3\mu\sigma^2-3\mu(\mu^2+\sigma^2)+3\mu^3-\mu^3)\Big) \\ &= 0, \end{aligned} \tag{23}$$

and for g_{22}:

$$\begin{aligned} \mathbb{E}_x[g_{22}] &= \mathbb{E}_x\Big(\frac{1}{4\sigma^4}-\frac{1}{2\sigma^6}(x-\mu)^2+\frac{1}{4\sigma^8}(x-\mu)^4\Big) \\ &= \frac{1}{4\sigma^4}-\frac{1}{2\sigma^6}\mathbb{E}_x[x^2-2x\mu+\mu^2]+\frac{1}{4\sigma^8}\mathbb{E}_x[x^4-4x^3\mu+6x^2\mu^2-4x\mu^3+\mu^4] \\ &= \frac{1}{4\sigma^4}-\frac{1}{2\sigma^6}\Big(\mathbb{E}_x[x^2]-2\mathbb{E}_x[x]\mu+\mu^2\Big)+\frac{1}{4\sigma^8}\Big(\mathbb{E}_x[x^4]-4\mathbb{E}_x[x^3]\mu+6\mathbb{E}_x[x^2]\mu^2-4\mathbb{E}_x[x]\mu^3+\mu^4\Big) \\ &= \frac{1}{4\sigma^4}-\frac{1}{2\sigma^6}\sigma^2+\frac{1}{4\sigma^8}3\sigma^4 \\ &= \frac{1}{2\sigma^4}. \end{aligned} \tag{24}$$

Finally, we get:

$$\mathbf{F} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}. \tag{25}$$

[2]See Section 3 for raw moments of univariate normal distribution.

2.5 Summary

The Fisher information matrix for given distribution:

- Bernoulli distribution:
$$\mathrm{F} = \frac{1}{\theta(1-\theta)},$$
- Categorical distribution:
$$\mathbf{F} = \mathrm{diag}\left\{\frac{1}{\theta_1}, \ldots, \frac{1}{\theta_K}\right\},$$
- Normal distribution:
$$\mathbf{F} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}.$$

3 Appendix: Raw moments

Table 1: The raw moments of univariate normal distribution.

Order	**Expression**	**Raw moment**
1	$\mathbb{E}_x[x]$	μ
2	$\mathbb{E}_x[x^2]$	$\mu^2 + \sigma^2$
3	$\mathbb{E}_x[x^3]$	$\mu^3 + 3\mu\sigma^2$
4	$\mathbb{E}_x[x^4]$	$\mu^4 + 6\mu^2\sigma^2 + 3\mu^4$

References

[1] C. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.