Interdisciplinary Information Sciences Vol. 17, No. 3 (2011) 155–174

ISSN 1340-9050 print/1347-6157 online
DOI 10.4036/iis.2011.155

An Introduction to Maximum Likelihood Estimation and Information Geometry

Keiji MIURA[1,2,*]

[1]*Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan*
[2]*PRESTO, JST, c/o GSIS, Tohoku University, Sendai 980-8579, Japan*

Received August 18, 2011; final version accepted September 29, 2011

In this paper, we review the maximum likelihood method for estimating the statistical parameters which specify a probabilistic model and show that it generally gives an optimal estimator with minimum mean square error asymptotically. Thus, for most applications in information sciences, the maximum likelihood estimation suffices. Fisher information matrix, which defines the orthogonality between parameters in a probabilistic model, naturally arises from the maximum likelihood estimation. As the inverse of the Fisher information matrix gives the covariance matrix for the estimation errors of the parameters, the orthogonalization of the parameters guarantees that the estimates of the parameters distribute independently from each other. The theory of information geometry provides procedures to diagonalize parameters globally or at all parameter values at least for the exponential and mixture families of distributions. The global orthogonalization gives a simplified and better view for statistical inference and, for example, makes it possible to perform a statistical test for each unknown parameter separately. Therefore, for practical applications, a good start is to examine if the probabilistic model under study belongs to these families.

KEYWORDS: maximum likelihood estimation, information geometry, Fisher information matrix

1. Introduction

Mathematical modeling provides a powerful way to understand and forecast natural phenomena. While differential equations work for modeling time evolutions of the systems whose mechanisms are clear and deterministic [17, 25, 29], probabilistic models are needed to treat unclear and stochastic phenomena appropriately [8–11, 15, 26].

When we want to get some useful information for stochastic events, it is quite helpful to have statistical distributions of the events. Although drawing histograms helps to find a tendency in data, it requires a lot of observations to obtain a smooth and reliable estimate of the distribution. Thus, it is in general more convenient to use a parametric model to estimate the distribution when any constraints on the shape of the distribution are known. In this case, a distribution is estimated by specifying a set of statistical parameters based on the observed data.

The most natural and popular way to estimate the parameters is the maximum likelihood estimation where the parameter values that are most likely to generate the observed data [8, 28] are chosen. As far as applications to information sciences are concerned, the maximum likelihood estimation gives an optimal estimator for most problems.

A metric, Fisher information matrix, naturally arises in the maximum likelihood estimation as a measure of independency between estimated parameters [2, 3, 6, 23]. As the inverse of the Fisher information matrix gives the covariance matrix for the estimation errors of the parameters, the orthogonalization of the parameters guarantees that the estimates of the parameters distribute independently from each other. However, diagonalizing Fisher matrix by linear algebra "locally" or at specific parameter values does not make sense because the parameter values are unknown and to be estimated.

The theory of information geometry tells us how to find an orthogonal parametrization for many probabilistic models including the exponential and mixture families, where the new parameters are orthogonal to each other "globally" or at all parameter values. This global orthogonalization gives a simplified and better view for statistical inference and, for example, makes it possible to perform a statistical test for each unknown parameter independently.

In these lecture notes, after we review some basic properties of the maximum likelihood estimation, we will discuss Fisher information matrix and demonstrate how to diagonalize it for specific examples.

2010 Mathematics Subject Classification: Primary 62F12, Secondary 62F05.
* Corresponding author. E-mail: miura@ecei.tohoku.ac.jp

2. Maximum Likelihood Estimation

The maximum likelihood method is the most popular way to estimate the parameter θ which specifies a probability function $P(X = x|\theta)$ of a discrete stochastic variable X (or a probability density function $p(x|\theta)$ of a continuous stochastic variable X) based on the observations $x_1, x_2, \ldots, x_n$ which were independently sampled from the distribution. Note that for a continuous stochastic variable X, the probability density function $p(x|\theta)$ satisfies

$$P(X < r|\theta) = \int_{-\infty}^{r} p(x|\theta)dx. \tag{1}$$

The maximum likelihood estimate is the value $\hat{\theta}$ which maximize the likelihood function which is defined by

$$L(\theta) = \Pi_{i=1}^{n} P(X = x_i|\theta) = P(X = x_1|\theta)P(X = x_2|\theta) \cdots P(X = x_n|\theta), \tag{2}$$

when X is a discrete stochastic variable and

$$L(\theta) = \Pi_{i=1}^{n} p(x_i|\theta) = p(x_1|\theta)p(x_2|\theta) \cdots p(x_n|\theta), \tag{3}$$

when X is a continuous stochastic variable. That is, the maximum likelihood estimation chooses the model parameter $\hat{\theta}$ which is the most likely to generate the observed data.

The maximum likelihood estimation is a heart of mathematical statistics and many beautiful theorems prove its optimality rigorously under certain regularity conditions [8, 28] as we will see in the next chapter. Therefore, as far as the applications to information sciences are concerned, the maximum likelihood estimation works and suffices for most problems.

2.1 Example: Hypergeometric distribution

Let us try to estimate the total number N of raccoons in Mt. Aoba by the capture-recapture method [10].

First, N_1 raccoons are captured, labeled and then released. Next, r raccoons are randomly captured. Then the probability that x out of r captured raccoons are labeled is given by

$$P(X = x|N, N_1, r) = \frac{\binom{N - N_1}{r - x}\binom{N_1}{x}}{\binom{N}{r}}. \tag{4}$$

Note that there are $\binom{N}{r}$ different patterns of capturing r racoons out of the total N racoons. Similarly, there are $\binom{N_1}{x}$ different patterns of capturing x labeled raccoons out of the total N_1 labeled raccoons and $\binom{N-N_1}{r-x}$ different patterns of capturing $r - x$ unlabeled raccoons out of the total $N - N_1$ unlabeled raccoons.

Table 1. Meaning of symbols in hypergeometric distribution.

$N - N_1$	Number of total unlabeled raccoons
$r - x$	Number of unlabeled raccoons captured for observation
N_1	Number of total labeled raccoons
x	Number of labeled raccoons recaptured for observation
N	Number of total raccoons
r	Number of raccoons captured for observation

Suppose that $N_1 = 25$, $r = 25$, and $x = 7$. Then the maximum likelihood estimate, $\hat{N}$ is the value which maximize the likelihood

$$L(N) = \frac{\binom{N - 25}{25 - 7}\binom{25}{7}}{\binom{N}{25}}. \tag{5}$$

Figure 1 shows that $L(N)$ has the maximum at $\hat{N} = 89$. Actually, this result can be also derived analytically by considering if $\frac{L(N)}{L(N-1)}$ is larger than 1. In fact, the ratio,

$$\frac{L(N)}{L(N-1)} = \frac{(N - N_1)(N - r)}{(N - N_1 - r + x)N}, \tag{6}$$

gets smaller than 1 when $N > \frac{N_1 r}{x} = 89.28571$. This leads to $\hat{N} = 89$ because N must be an integer.

Finally, note that this example has all discrete variables and only one observation for X.

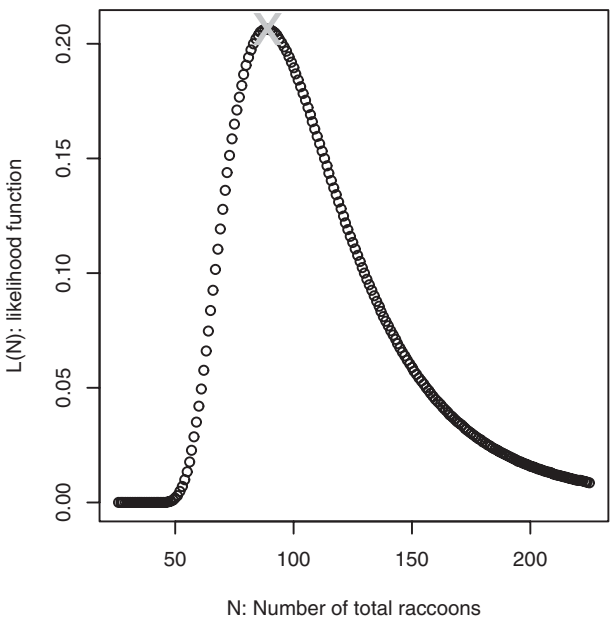

Fig. 1. Likelihood function $L(N)$ in (5) as a function of number of total raccoons in Mt. Aoba. The hypergeometric distribution with $N_1 = 25$, $r = 25$, and $x = 7$ was used. The gray cross denotes the peak whose location is $\hat{N} = 89$.

R code for Fig. 1

```
X=7;
N1=25;
r=25;
L=1:200*0;
for (N in r:(200+r)) L[N-r] = dhyper(X,N1,N-N1,r);
N=r+1:200;
# postscript("Fig1.eps",width=3.3,height=3.7,horizontal=F,onefile=F,paper="special");
# par(oma=c(0,0,0,0), mar=c(5.4,4,1.5,0),cex=2/3,cex.main=1);
plot(N,L,ylab='L(N): likelihood function',xlab='N: Number of total raccoons');
points(which.max(L)+r,L[which.max(L)], pch='x',cex=3,col='gray');
# dev.off()
```

2.2 (Counter)example: Normal distribution

Let us try to estimate the mean μ and the variance σ of a normal distribution:

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. \tag{7}$$

When $X_1, X_2, \cdots, X_n$ are sampled from the distribution independently, the likelihood for each observation is given by

$$\begin{aligned} p(X_1|\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_1-\mu)^2}, \\ p(X_2|\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_2-\mu)^2}, \\ &\vdots \\ p(X_n|\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_n-\mu)^2}. \end{aligned} \tag{8}$$

Sometimes this is rewritten as $X_i \sim \mathcal{N}(\mu,\sigma)$ for simplicity. The likelihood L is defined by

$$L(\mu,\sigma) = p(X_1|\mu,\sigma)p(X_2|\mu,\sigma)\cdots p(X_n|\mu,\sigma). \tag{9}$$

Instead of maximizing the likelihood L itself, one can maximize the log likelihood without loss of generality:

$$\log L(\mu,\sigma) = \sum_{i=1}^{n} \log p(X_i|\mu,\sigma). \tag{10}$$

Note that taking log simplifies the calculation as we will see because the product turns to a sum. At the parameter values $(\hat{\mu},\hat{\sigma})$ which maximize the log likelihood, the derivative of the log likelihood with respect to μ and σ should be zero:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\mu}\log L(\mu,\sigma) = \sum_{i=1}^{n}\frac{\partial}{\partial\mu}\log p(X_i|\mu,\sigma), \quad \text{and} \\ 0 &= \frac{\partial}{\partial\sigma}\log L(\mu,\sigma) = \sum_{i=1}^{n}\frac{\partial}{\partial\sigma}\log p(X_i|\mu,\sigma), \end{aligned} \tag{11}$$

where

$$\begin{aligned} \frac{\partial}{\partial\mu}\log p(x|\mu,\sigma) &= \frac{\partial}{\partial\mu}\left(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{x-\mu}{\sigma^2}, \quad \text{and} \\ \frac{\partial}{\partial\sigma}\log p(x|\mu,\sigma) &= \frac{\partial}{\partial\sigma}\left(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}\right) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}. \end{aligned} \tag{12}$$

Thus, we have

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\frac{X_i-\hat{\mu}}{\hat{\sigma}^2}, \quad \text{and} \\ 0 &= \sum_{i=1}^{n}\left(-\frac{1}{\hat{\sigma}} + \frac{(X_i-\hat{\mu})^2}{\hat{\sigma}^3}\right). \end{aligned} \tag{13}$$

By solving the equations for $\hat{\mu}$ and $\hat{\sigma}$, we obtain

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{and} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2. \end{aligned} \tag{14}$$

Note that $\hat{\sigma}$ is biased, that is, not correct on average. It is widely known that

$$\hat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 \tag{15}$$

works as an unbiased estimator of σ [8]. The difference between the two estimators is the normalization factors, n and $n-1$. We discuss the optimality of the maximum likelihood estimator in the large limit of n in the next section. However, for finite n, no theorem guarantees the optimality of the maximum likelihood estimator. Note that the ratio of the above two estimators gets negligible in the large n limit:

$$\lim_{n\to\infty}\frac{n-1}{n} \to 1, \tag{16}$$

suggesting that the maximum likelihood estimator for σ can be optimal only asymptotically (that is, in the large n limit).

3. Optimality of Maximum Likelihood Estimation

In the last chapter we introduced the maximum likelihood estimator as a natural way for parameter estimation. However, we did not discuss the performance of the maximum likelihood estimator there. Here we briefly review three good properties of the maximum likelihood estimation [8, 28] which are summarized in Table 2. Although these theorems assume certain regularity conditions, here we only mention that the conditions are mild and satisfied in most cases.

Table 2. Theorems for maximum likelihood estimation.

Theorem	Sketch of Proof
consistency (asymptotic correctness)	positivity of Kullback–Leibler divergence
asymptotic normality: $\hat{\theta} \sim \mathcal{N}(\theta, \frac{1}{n}I^{-1})$	central limit theorem for delta method
efficiency (minimum variance)	Cauchy–Schwarz inequality

3.1 Consistency (asymptotic correctness)

The first property is that in the large limit of the number of observation n, the estimate $\hat{\theta}$ goes to the true value θ. This property is essential for an estimator. As there is a counterexample for finite n as we saw in the last example, here we discuss the consistency in the large n limit.

Instead of the maximization of the log-likelihood, let us consider the maximization of the log-likelihood subtracted by a constant without loss of generality:

$$\begin{aligned}
&\frac{1}{n}\log L(\hat{\theta}) - \text{constant} \\
&= \frac{1}{n}\sum_{i=1}^{n}\log p(X_i|\hat{\theta}) - \int p(x|\theta)\log p(x|\theta)dx \\
&\xrightarrow[n\to\infty]{} \int p(x|\theta)\log p(x|\hat{\theta})dx - \int p(x|\theta)\log p(x|\theta)dx \\
&= \int p(x|\theta)\log\frac{p(x|\hat{\theta})}{p(x|\theta)} \\
&= -D(p(x|\theta)||p(x|\hat{\theta})) \\
&\leq 0,
\end{aligned} \tag{17}$$

where the arrow means the limit of large number of observations n and the summation was replaced by the integral with $p(x|\theta)$ as a weight there: $\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \int p(x)f(x)dx$. Here we use the fact that the Kullback–Leibler divergence

$$D(p_1(x)|p_2(x)) = \int p_1(x)\log\frac{p_1(x)}{p_2(x)} \tag{18}$$

takes the minimum value 0 only when $p_1(x) = p_2(x)$. Therefore the maximum of the log-likelihood is obtained only at $\hat{\theta} = \theta$.

3.2 Asymptotic normality: $\hat{\theta} \sim \mathcal{N}(\theta, \frac{1}{n}I^{-1})$

The second property is that the maximum likelihood estimate $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is normal-distributed when X's are random samples from $p(x|\theta)$. For clarity, imagine that $\hat{\theta}$ is obtained for EACH realization of a set of observations $X_1, X_2, \ldots, X_n$. Thus, say, if a set of n observations of X is sampled repeatedly 100 times, 100 estimates $\hat{\theta}$'s are obtained. The theorem says that the distribution of $\hat{\theta}$ is a normal distribution. The variance of the normal distribution or the estimation error is actually minimum among all possible estimators as we will see in the third theorem. This second theorem can be used, for example, when statistical tests are performed on $\hat{\theta}$ by assuming the normal shape. In addition, in information geometry, this theorem guarantees that only the covariance or orthogonality matters for the distribution of $\hat{\theta}$.

Here we only prove for the case when θ is one dimensional or a scalar for educational purposes, although it is easy to extend the proof to general cases.

When the number of observation n is large and the estimate $\hat{\theta}$ is close to the true value θ, the definitional equation of $\hat{\theta}$ can be Taylor-expanded by $(\hat{\theta} - \theta)$ to the first order:

$$\begin{aligned}
0 &= \frac{d}{d\theta}\log L(\hat{\theta}) \\
&= \sum_{i=1}^{n}\frac{d}{d\theta}\log p(X_i|\hat{\theta}) \\
&= \sum_{i=1}^{n}\frac{d}{d\theta}\log p(X_i|\theta) + (\hat{\theta}-\theta)\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log p(X_i|\theta) + O((\theta-\hat{\theta})^2) \\
&= \sum_{i=1}^{n}\frac{d}{d\theta}\log p(X_i|\theta) + (\hat{\theta}-\theta)n\int p(x|\theta)\frac{d^2}{d\theta^2}\log p(x|\theta)dx + O((\theta-\hat{\theta})^2) \\
&= \sum_{i=1}^{n}\frac{d}{d\theta}\log p(X_i|\theta) - (\hat{\theta}-\theta)nI + O((\theta-\hat{\theta})^2),
\end{aligned} \tag{19}$$

where the Fisher information is defined as

$$\begin{aligned}
I &= \int p(x|\theta)\left(\frac{d}{d\theta}\log p(x|\theta)\right)^2 dx \\
&= -\int p(x|\theta)\frac{d^2}{d\theta^2}\log p(x|\theta)dx,
\end{aligned} \tag{20}$$

where partial integral was used to derive another form of the definition. Note that although a sum was replaced by an integral in (19), the difference was assumed to be $O(\hat{\theta} - \theta)$ and negligible. Therefore, we have

$$\hat{\theta} - \theta = \frac{1}{nI}\sum_{i=1}^{n}\frac{d}{d\theta}\log p(X_i|\theta) + \text{negligible terms}. \tag{21}$$

The righthand side turns out to be normal-distributed, $\mathcal{N}(0, \frac{1}{n} I^{-1})$ by the central limit theorem for $\frac{1}{n}\sum Y_i$ with $Y_i = \frac{d}{d\theta}\log p(X_i|\theta)$. The central limit guarantees that the mean of a sufficiently large number of independent random variables will be (approximately) normally distributed [8, 10]. Note that the mean is zero because of the consistency of the estimation, or, by using

$$\overline{Y_i} = \int p(x|\theta)\left(\frac{d}{d\theta}\log p(x|\theta)\right)dx = \int \frac{d}{d\theta} p(x|\theta)dx = \frac{d}{d\theta}\int p(x|\theta)dx = \frac{d}{d\theta}1 = 0. \tag{22}$$

The variance is derived as

$$\left(\frac{1}{nI}\right)^2 n\,\mathrm{Var}\left[\frac{d}{d\theta}\log p(X|\theta)\right] = \left(\frac{1}{nI}\right)^2 n\int p(x|\theta)\left(\frac{d}{d\theta}\log p(x|\theta)\right)^2 dx = \left(\frac{1}{nI}\right)^2 nI = \frac{1}{nI}. \tag{23}$$

Thus, $\hat{\theta} \sim \mathcal{N}(\theta, \frac{1}{n} I^{-1})$. Note that the variance decreases with increasing n. This result can also be represented as $\sqrt{n}(\hat{\theta} - \theta) \sim \mathcal{N}(0, I^{-1})$ in the large n limit. The Fisher information I denotes the precision of the estimation because its inverse is the variance of the estimation error.

Two technical notes: (1) I is called Fisher information or Fisher information matrix for multiple dimensional cases. It is essentially the inner product defined on a space of functions with $p(x|\boldsymbol{\theta})$ as a weight:

$$\begin{aligned} I_{ij} &= \left\langle \frac{d}{d\theta_i}\log p(x|\boldsymbol{\theta}), \frac{d}{d\theta_j}\log p(x|\boldsymbol{\theta})\right\rangle \\ &= \int p(x|\boldsymbol{\theta})\left(\frac{d}{d\theta_i}\log p(x|\boldsymbol{\theta})\right)\left(\frac{d}{d\theta_j}\log p(x|\boldsymbol{\theta})\right)dx. \end{aligned} \tag{24}$$

I^{-1} means the inverse matrix of I. (2) The "Taylor expansion" used above is called delta method and a very basic technique in statistics. Many important results for the large limit of the number of observations are "asymptotically" derived by using the delta method [8, 28].

3.3 Efficiency (minimum variance)

The third property is that the maximum likelihood estimator has the minimum mean variance of estimation error among any other estimator which are unbiased (correct on average).

This is derived by the Cauchy–Schwarz inequality

$$\mathrm{Var}[f] \geq \frac{(\mathrm{Cov}[f, g])^2}{\mathrm{Var}[g]} \tag{25}$$

with

$$f = \hat{\theta}(x_1, x_2, \ldots, x_n) \quad (\text{ANY estimator of } \theta \text{ with } \mathrm{Ex}[\hat{\theta}] = \theta) \tag{26}$$

and

$$g = \frac{d}{d\theta}\log L(x|\theta) \tag{27}$$

where

$$L(x|\theta) = p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta). \tag{28}$$

For simplicity, let us consider the one dimensional case where θ is a scalar for educational purposes. Then we have

$$\begin{aligned} \mathrm{Cov}[f, g] &= \int L(x|\theta)(\hat{\theta} - \theta)\frac{d}{d\theta}\log L(x|\theta)dx_1 dx_2 \cdots dx_n \\ &= \int L(x|\theta)(\hat{\theta} - \theta)\frac{\frac{d}{d\theta}L(x|\theta)}{L(x|\theta)}dx_1 dx_2 \cdots dx_n \\ &= \int (\hat{\theta} - \theta)\frac{d}{d\theta}L(x|\theta)dx_1 dx_2 \cdots dx_n \\ &= -\int \frac{d(\hat{\theta} - \theta)}{d\theta}L(x|\theta)dx_1 dx_2 \cdots dx_n \quad (\text{partial integral}) \\ &= \int L(x|\theta)dx_1 dx_2 \cdots dx_n \quad (\hat{\theta} \text{ does not depend on } \theta) \\ &= \int p(x_1|\theta)dx_1 \int p(x_2|\theta)dx_2 \cdots \int p(x_n|\theta)dx_n \\ &= 1, \end{aligned} \tag{29}$$

and

$$
\begin{aligned}
&\mathrm{Var}[g] \\
&= \int L(x|\theta)\left(\frac{d}{d\theta}\log L(x|\theta)\right)^2 dx_1 dx_2 \cdots dx_n \\
&= \int L(x|\theta)\left(\frac{d}{d\theta}\sum_{i=1}^{n}\log p(x_i|\theta)\right)^2 dx_1 dx_2 \cdots dx_n \\
&= \int L(x|\theta)\left(\sum_{i=1}^{n}\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 dx_1 dx_2 \cdots dx_n \\
&= \int L(x|\theta)\left(\sum_{i=1}^{n}\frac{d}{d\theta}\log p(x_i|\theta)\right)\left(\sum_{j=1}^{n}\frac{d}{d\theta}\log p(x_j|\theta)\right) dx_1 dx_2 \cdots dx_n \\
&= \int L(x|\theta)\left[\sum_{i=1}^{n}\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 + \sum_{i\neq j}\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)\left(\frac{d}{d\theta}\log p(x_j|\theta)\right)\right] dx_1 dx_2 \cdots dx_n \\
&= \int p(x_1|\theta)\cdots p(x_n|\theta)\left[\sum_{i=1}^{n}\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 + \sum_{i\neq j}\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)\left(\frac{d}{d\theta}\log p(x_j|\theta)\right)\right] dx_1 dx_2 \cdots dx_n \\
&= \sum_{i=1}^{n}\int p(x_i|\theta)\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 dx_i + \sum_{i\neq j}\int\int p(x_i|\theta)p(x_j|\theta)\frac{d}{d\theta}\log p(x_i|\theta)\frac{d}{d\theta}\log p(x_j|\theta) dx_i dx_j \\
&= \sum_{i=1}^{n}\int p(x_i|\theta)\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 dx_i + \sum_{i\neq j}\left(\int p(x_i|\theta)\frac{d}{d\theta}\log p(x_i|\theta) dx_i\right)\left(\int p(x_j|\theta)\frac{d}{d\theta}\log p(x_j|\theta) dx_j\right) \\
&= \sum_{i=1}^{n}\int p(x_i|\theta)\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 dx_i + \sum_{i\neq j}(0)(0) \qquad (30) \\
&= \sum_{i=1}^{n}\int p(x_i|\theta)\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 dx_i \\
&= n\int p(x|\theta)\left(\frac{d}{d\theta}\log p(x|\theta)\right)^2 dx \\
&= nI. \qquad (31)
\end{aligned}
$$

Then, by the Cauchy–Schwarz inequality, we have

$$\mathrm{Var}[\hat{\theta}] \geq \frac{1}{n} I^{-1}. \qquad (32)$$

According to this Cramer–Rao theorem, the variance of ANY unbiased estimator is at least $(nI)^{-1}$. Meamwhile, the maximum likelihood estimator attains this lower bound (the equality holds!). Thus, the maximum likelihood estimator is one of the best estimator among all the estimators.

Finally, note that for multidimensional cases of θ, I^{-1} means the inverse of the Fisher information matrix. And the inequality means the positive-(semi)definiteness of the matrix.

3.4 Example: Binomial distribution

The probability of having X heads out of n trials of coin tosses is given by the binomial distribution:

$$P(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad (33)$$

where the parameter θ denotes the probability of having a head in a single trial. The maximum likelihood estimator is given by solving

$$
\begin{aligned}
0 &= \frac{d}{d\theta}\log L(\theta) \\
&= \frac{d}{d\theta}\log p(X|\theta) \\
&= \frac{d}{d\theta}\log\left[\binom{n}{X}\theta^X(1-\theta)^{n-X}\right] \\
&= \frac{d}{d\theta}\left[\log\binom{n}{X} + X\log\theta + (n-X)\log(1-\theta)\right]
\end{aligned}
$$

$$= \frac{d}{d\theta} X \log\theta - \frac{d}{d\theta}(n-X)\log(1-\theta)$$
$$= \frac{X}{\theta} - \frac{n-X}{1-\theta}. \tag{34}$$

Thus we have,

$$\hat{\theta} = \frac{X}{n}. \tag{35}$$

Naturally, the estimate $\hat{\theta}$ is the sample frequency of the head.

Next, let us consider how reliable this estimate is. By similar calculations as above, the Fisher information is given by

$$\begin{aligned}
I &= \mathrm{E}\left[\left(\frac{d}{d\theta}\log p(X|\theta)\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{X}{\theta} - \frac{n-X}{1-\theta}\right)^2\right] \\
&= \mathrm{E}\left[\frac{1}{\theta^2(1-\theta)^2}((1-\theta)X - \theta(n-X))^2\right] \\
&= \frac{1}{\theta^2(1-\theta)^2}\mathrm{E}[(X-\theta X-\theta n+\theta X)^2] \\
&= \frac{1}{\theta^2(1-\theta)^2}\mathrm{E}[(X-\theta n)^2] \\
&= \frac{1}{\theta^2(1-\theta)^2}\mathrm{E}[(X-\mathrm{E}[X])^2] \\
&= \frac{1}{\theta^2(1-\theta)^2}\mathrm{Var[X]} \\
&= \frac{1}{\theta^2(1-\theta)^2} n\theta(1-\theta) \\
&= \frac{n}{\theta(1-\theta)},
\end{aligned} \tag{36}$$

where $\mathrm{E}[X] = \theta n$ and $\mathrm{Var}[X] = n\theta(1-\theta)$ were used. Thus, the mean $\pm$ the standard deviation can be written as

$$\hat{\theta} \pm \sqrt{I^{-1}} = \frac{X}{n} \pm \sqrt{\frac{\theta(1-\theta)}{n}}. \tag{37}$$

In Fig. 3, the histogram of $\hat{\theta}$ for repeated realizations of X when $n = 100$ and $\theta = 0.6$ is plotted. That is, $(X_1, X_2, \ldots, X_{10000})$ are independently sampled from the binomial distribution with $n = 100$ and $\theta = 0.6$. Then estimates are $\hat{\theta}_i = \frac{X_i}{n} = \frac{X_i}{100}$. Note that in this example, we have only one observation in each estimation. The histogram of $\hat{\theta}$ was plotted in Fig. 3.

Another way to utilize the Fisher information I is to perform a statistical test. If a null hypothesis is $\theta = \theta_0$, then, $\hat{\theta}$ should be distributed around θ_0:

$$\hat{\theta} \sim \mathcal{N}\left(\theta_0, \frac{1}{n} I(\theta_0)^{-1}\right). \tag{38}$$

Note that the variance shrinks with the number of observations n, indicating that the more observations the more reliable the estimate should be. For example, when the null hypothesis is $\theta_0 = 0.5$, the (two-sigma) confidence interval (almost $p = 0.05$) is given by

$$\theta_0 \pm 2\sqrt{I(\theta_0)^{-1}} = 0.5 \pm 2\sqrt{\frac{0.5^2}{100}} = 0.5 \pm 0.1. \tag{39}$$

If the true value is $\theta = 0.6$ as in Fig. 3, for half cases $\hat{\theta}$ exceeds 0.6 and, then, the null hypothesis is rejected.

3.5 Example: Gamma distribution with conventional coordinates

The gamma distribution is often used for the distribution of the inter-event time intervals [18–21] and other unimodal distributions of positive variables ($x > 0$):

$$p(x|\lambda, \alpha) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}. \tag{40}$$

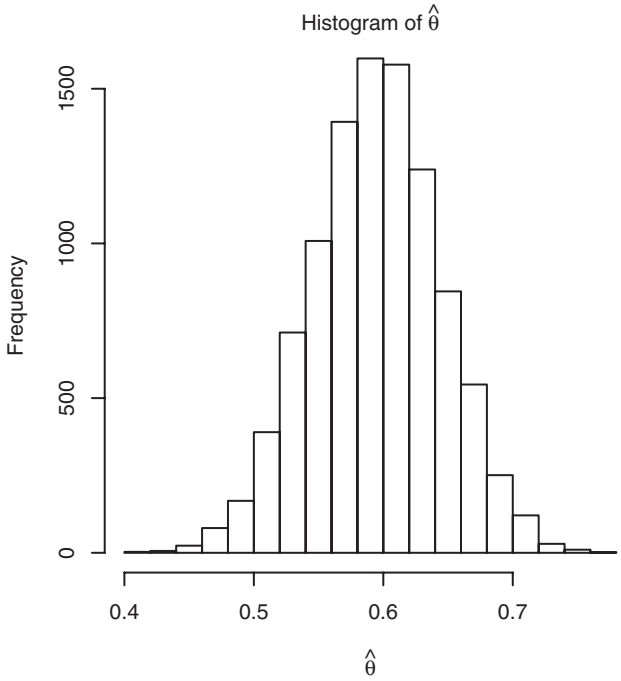

Fig. 2. Histogram of estimated θ. Although the true value is $\theta = 0.6$, there is a trial-by-trial estimation error as seen as the standard deviation of the histogram ($= 0.04911697$). The theoretical prediction was $\sqrt{I(\theta = 0.6)^{-1}} = \sqrt{\frac{0.6 \cdot 0.4}{100}} = 0.04898979$.

R code for Fig. 2

```
# postscript("Fig2.eps",width=3.3,height=3.7,horizontal=F,onefile=F,paper="special");
# par(oma=c(0,0,0,0), mar=c(5.4,4,1.5,0),cex=2/3,cex.main=1);
theta = rbinom(10000, 100, 0.6)/100;
sd(theta); # 0.04911697
hist(theta, xlab=expression(hat(theta)), main=expression(paste("Histogram of", hat(theta))));
# dev.off()
```

When $X_1, X_2, \cdots, X_n$ are observed independently, the likelihood for each observation is given by

$$\begin{aligned} p(X_1|\lambda,\alpha) &= \frac{\alpha^\lambda}{\Gamma(\lambda)} X_1^{\lambda-1} e^{-\alpha X_1}, \\ p(X_2|\lambda,\alpha) &= \frac{\alpha^\lambda}{\Gamma(\lambda)} X_2^{\lambda-1} e^{-\alpha X_2}, \\ &\vdots \\ p(X_n|\lambda,\alpha) &= \frac{\alpha^\lambda}{\Gamma(\lambda)} X_n^{\lambda-1} e^{-\alpha X_n}. \end{aligned} \tag{41}$$

The likelihood L is defined by

$$L(\lambda,\alpha) = p(X_1|\lambda,\alpha)p(X_2|\lambda,\alpha)\cdots p(X_n|\lambda,\alpha). \tag{42}$$

At the parameter value $(\hat{\lambda}, \hat{\alpha})$ which maximize the log likelihood, the derivative of the log likelihood with respect to λ and α should be zero:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\lambda} \log L(\lambda,\alpha) = \sum_{i=1}^n \frac{\partial}{\partial\lambda} \log p(X_i|\lambda,\alpha), \quad \text{and,} \\ 0 &= \frac{\partial}{\partial\alpha} \log L(\lambda,\alpha) = \sum_{i=1}^n \frac{\partial}{\partial\alpha} \log p(X_i|\lambda,\alpha), \end{aligned} \tag{43}$$

where

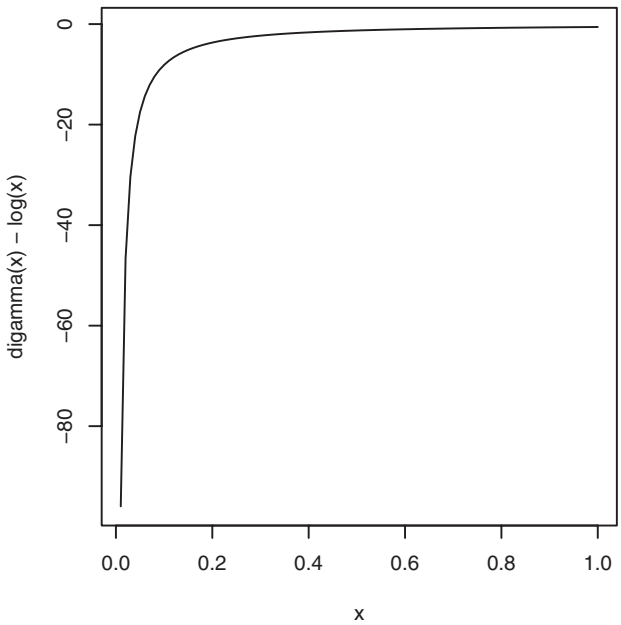

Fig. 3. $\frac{\Gamma'(x)}{\Gamma(x)} - \log x$ is a monotonic function.

R code for Fig. 3

```
# postscript("Fig3.eps",width=3.3,height=3.7,horizontal=F,onefile=F,paper="special");
# par(oma=c(0,0,0,0), mar=c(5.4,4,1.5,0),cex=2/3,cex.main=1);
x=1:100/100; plot(x, digamma(x)-log(x), type='l');
# dev.off()
```

$$\frac{\partial}{\partial\lambda}\log p(x|\lambda,\alpha) = \log\alpha - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)} + \log x, \quad \text{and,}$$
$$\frac{\partial}{\partial\alpha}\log p(x|\lambda,\alpha) = \frac{\lambda}{\alpha} - x. \tag{44}$$

Thus,

$$0 = n\log\alpha - n\frac{\Gamma'(\lambda)}{\Gamma(\lambda)} + \sum_{i=1}^{n}\log X_i, \quad \text{and,}$$
$$0 = \frac{n\lambda}{\alpha} - \sum_{i=1}^{n} X_i. \tag{45}$$

That is, the maximum likelihood estimates are obtained by solving the following equations numerically:

$$\frac{\Gamma'(\hat{\lambda})}{\Gamma(\hat{\lambda})} - \log\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\log X_i - \log\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right), \quad \text{and,}$$
$$\hat{\alpha} = \frac{\hat{\lambda}}{\frac{1}{n}\sum_{i=1}^{n} X_i}. \tag{46}$$

The first equation can be solved for $\hat{\lambda}$ uniquely as the left hand side of the equation is the monotonic function of $\hat{\lambda}$ as shown in Fig. 3. Then, $\hat{\alpha}$ is obtained from the second equation.

The following formulae help to compute the Fisher information matrix:

$$\mathrm{E}[X] = \int_0^\infty p(x|\lambda,\alpha)x\,dx = \frac{\lambda}{\alpha}, \tag{47}$$
$$\mathrm{Var}[X] = \int_0^\infty p(x|\lambda,\alpha)(x-\mathrm{E}[X])^2 dx = \frac{\lambda}{\alpha^2}, \tag{48}$$

$$\mathrm{E}[\log(X)] = -\log\alpha + \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}, \tag{49}$$

$$\mathrm{E}[X\log(X)] = -\frac{\lambda}{\alpha}\log\alpha + \frac{1}{\alpha} + \frac{\lambda}{\alpha}\frac{\Gamma'(\lambda)}{\Gamma(\lambda)}, \quad \text{and,} \tag{50}$$

$$\mathrm{E}[\log(X)\log(X)] = (\log\alpha)^2 - 2(\log\alpha)\frac{\Gamma'(\lambda)}{\Gamma(\lambda)} + \frac{\Gamma''(\lambda)}{\Gamma(\lambda)}. \tag{51}$$

The Fisher information matrix is given by

$$\begin{aligned}
&\int p(x|\lambda,\alpha)\frac{\partial\log p(x|\lambda,\alpha)}{\partial\lambda}\frac{\partial\log p(x|\lambda,\alpha)}{\partial\lambda}dx \\
&= \mathrm{E}\left[\left(\log\alpha - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)} + \log X\right)^2\right] \\
&= \mathrm{E}\left[\left(\log\alpha - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right)^2 + 2\left(\log\alpha - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right)\log X + (\log X)^2\right] \\
&= \frac{\Gamma''(\lambda)}{\Gamma(\lambda)} - \frac{\Gamma'(\lambda)^2}{\Gamma(\lambda)^2} \\
&= \frac{\Gamma''(\lambda)\Gamma(\lambda) - \Gamma'(\lambda)^2}{\Gamma(\lambda)^2} \\
&= \left(\frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right)' \\
&= \frac{\partial^2}{\partial\lambda^2}\log\Gamma(\lambda) \\
&= \mathrm{Trigamma}(\lambda),
\end{aligned} \tag{52}$$

$$\begin{aligned}
&\int p(x|\lambda,\alpha)\frac{\partial\log p(x|\lambda,\alpha)}{\partial\lambda}\frac{\partial\log p(x|\lambda,\alpha)}{\partial\alpha}dx \\
&= \mathrm{E}\left[\left(\log\alpha - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)} + \log X\right)\left(\frac{\lambda}{\alpha} - X\right)\right] \\
&= \mathrm{E}\left[\left(\log\alpha - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right)\frac{\lambda}{\alpha} + \frac{\lambda}{\alpha}\log X - \left(\log\alpha - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right)X - X\log X\right] \\
&= -\frac{1}{\alpha},
\end{aligned} \tag{53}$$

and

$$\begin{aligned}
&\int p(x|\lambda,\alpha)\frac{\partial\log p(x|\lambda,\alpha)}{\partial\alpha}\frac{\partial\log p(x|\lambda,\alpha)}{\partial\alpha}dx \\
&= \mathrm{E}\left[\left(\frac{\lambda}{\alpha} - X\right)^2\right] \\
&= \mathrm{Var}[X] \\
&= \frac{\lambda}{\alpha^2}.
\end{aligned} \tag{54}$$

That is

$$I = \begin{pmatrix} \mathrm{Trigamma}(\lambda) & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & \frac{\lambda}{\alpha^2} \end{pmatrix}. \tag{55}$$

In Fig. 4, the maximum likelihood estimates $(\hat{\lambda}, \hat{\alpha})$ for repeated realizations of a set of $100X$'s are plotted. That is, $(X_1, X_2, \ldots, X_{100})$ are independently sampled from the gamma distributions with $\alpha = 2$ and $\lambda = 2$ and the estimates were computed. For each estimation, a point was plotted in $(\hat{\lambda}, \hat{\alpha})$-plane. The estimation was repeated many times. The covariance matrix estimated from the numerical simulation in Fig. 4 was

$$\hat{\Sigma} = \begin{pmatrix} 0.07218597 & 0.07252140 \\ 0.07252140 & 0.09499623 \end{pmatrix}, \tag{56}$$

and its inverse was

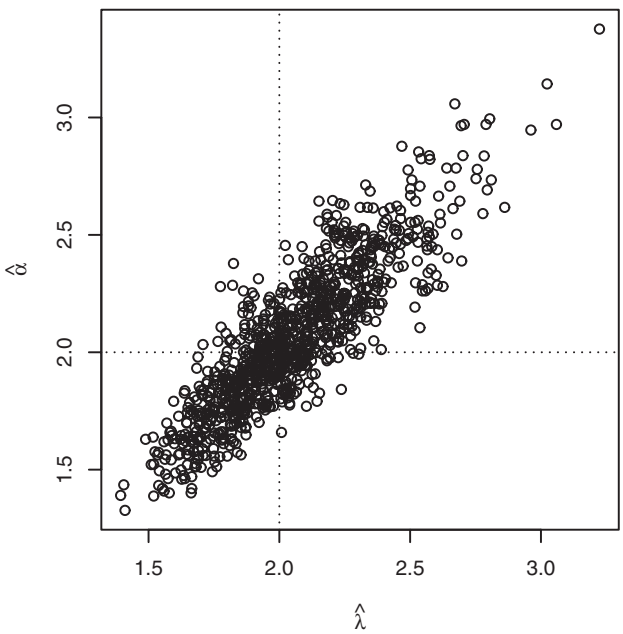

Fig. 4. Maximum likelihood estimation of λ and α repeated many times for different realizations of $100X$'s. Note the negative correlations as predicted by the Fisher information matrix.

R code for Fig. 4

```
# postscript("Fig4.eps",width=3.3,height=3.7,horizontal=F,onefile=F,paper="special");
# par(oma=c(0,0,0,0), mar=c(5.4,4,1.5,0),cex=2/3,cex.main=1);
lambda = (1:1000) * 0;
alpha = lambda;
for (i in 1:1000){
   X = rgamma(100, 2, 2);
   m = mean(X);
   l = mean(log(X));
   fr = function(x){(l - log(m) + log(x) - digamma(x))**2};
   lambda[i] = optimize(fr, interval=c(0,100))$minimum;
   alpha[i]=lambda[i]/m;
};
plot(lambda,alpha,xlab=expression(hat(lambda)),ylab=expression(hat(alpha)));
lines(c(1,3.5),c(2,2),lty=3); lines(c(2,2),c(1,3.5),lty=3);
# dev.off()

Sigma = cov(cbind(lambda,alpha))
# 0.07218597 0.07252140
# 0.07252140 0.09499623
solve(Sigma)
# 59.44541 -45.38143
# -45.38143 45.17152
100 * trigamma(2)
# 64.49341
```

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 59.44541 & -45.38143 \\ -45.38143 & 45.17152 \end{pmatrix}, \tag{57}$$

whose theoretical values are

$$100I = \begin{pmatrix} 64.49341 & -50 \\ -50 & 50 \end{pmatrix}. \tag{58}$$

There are small discrepancies between the numerical and theoretical values, probably, because the number of samples (= 100) is not large enough.

3.6 Example: Gamma distribution with mixed coordinates

Let us transform the parameters as

$$\mu = \frac{\lambda}{\alpha}, \quad \text{and,}$$
$$\kappa = \lambda. \tag{59}$$

We will discuss how this transformation was found later in Sec. 4. The distriution with the new parameters is given by

$$p(x|\mu,\kappa) = \left(\frac{\kappa}{\mu}\right)^{\kappa} \frac{1}{\Gamma(\kappa)} x^{\kappa-1} e^{-\frac{\kappa}{\mu}x}. \tag{60}$$

Then the maximum likelihood estimation is given by

$$\frac{\Gamma'(\hat{\kappa})}{\Gamma(\hat{\kappa})} - \log\hat{\kappa} = \frac{1}{n}\sum_{i=1}^{n}\log X_i - \log\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right), \quad \text{and,}$$
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{61}$$

Actually, the maximum likelihood estimator is invariant under parameter transformations in general and, thus, the following relations hold for the maximum likelihood estimators of the gamma distribution:

$$\hat{\mu} = \frac{\hat{\lambda}}{\hat{\alpha}}, \quad \text{and,}$$
$$\hat{\kappa} = \hat{\lambda}. \tag{62}$$

Notice that

$$\int p(x|\mu,\kappa)\frac{\partial}{\partial\mu}\log p(x|\mu,\kappa)\frac{\partial}{\partial\mu}\log p(x|\mu,\kappa)dx = \frac{\kappa}{\mu^2},$$
$$\int p(x|\mu,\kappa)\frac{\partial}{\partial\mu}\log p(x|\mu,\kappa)\frac{\partial}{\partial\kappa}\log p(x|\mu,\kappa)dx = 0, \quad \text{and,}$$
$$\int p(x|\mu,\kappa)\frac{\partial}{\partial\kappa}\log p(x|\mu,\kappa)\frac{\partial}{\partial\kappa}\log p(x|\mu,\kappa)dx = \text{Trigamma}(\kappa) - \frac{1}{\kappa}. \tag{63}$$

That is, the Fisher information matrix is diagonal for ANY parameter values:

$$I = \begin{pmatrix} \frac{\kappa}{\mu^2} & 0 \\ 0 & \text{Trigamma}(\kappa) - \frac{1}{\kappa} \end{pmatrix}. \tag{64}$$

The result that $\text{Var}[\hat{\mu}] = \frac{1}{n}\frac{\mu^2}{\kappa}$ is natural, because μ just scales the entire distribution uniformly and κ is a regularity parameter. In fact, the larger κ, the similar X's are. When κ is large enough the distribution looks normal, and, in the large limit of κ, the distribution becomes almost a delta function. Thus, by estimating κ, one can know how reliable the estimate of μ is: $\text{Var}[\hat{\mu}] = \frac{1}{n}\frac{\mu^2}{\kappa}$.

The covariance matrix estimated from the numerical simulation in Fig. 5 was

$$\hat{\Sigma} = \begin{pmatrix} 0.0050553689 & -0.0001849231 \\ -0.0001849231 & 0.0721859667 \end{pmatrix}, \tag{65}$$

and its inverse was

$$\hat{\Sigma}^{-1} = 100\begin{pmatrix} 1.97828040 & 0.00506788 \\ 0.00506788 & 0.13854406 \end{pmatrix}, \tag{66}$$

whose theoretical values are

$$100I = 100\begin{pmatrix} 2 & 0 \\ 0 & 0.1449341 \end{pmatrix}. \tag{67}$$

There are small discrepancies between the numerical and theoretical values, probably, because the number of samples (= 100) is not large enough.

Note that $\hat{\mu}$ and $\hat{\kappa}$ are independent from each other. Therefore, error bars can be plotted separately for $\hat{\mu}$ and $\hat{\kappa}$. That is, a statistical test can be performed for each variable independently.

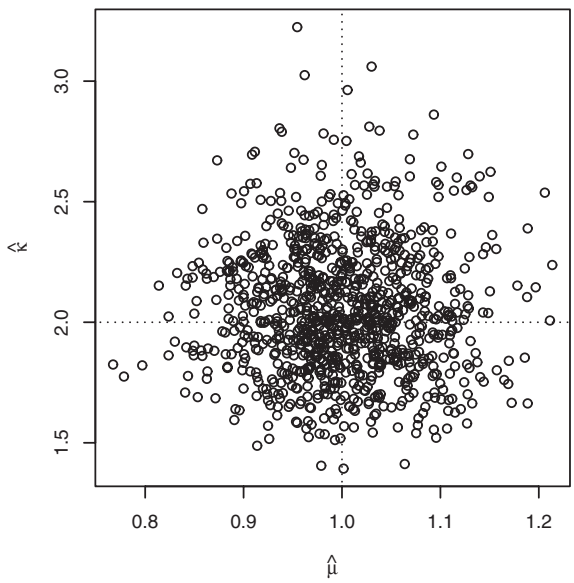

Fig. 5. Maximum likelihood estimation of μ and κ repeated many times for different realizations of $100X$'s. Note that there is no correlation as predicted by the diagonal Fisher information matrix.

R code for Fig. 5

```
# postscript("Fig5.eps",width=3.3,height=3.7,horizontal=F,onefile=F,paper="special");
# par(oma=c(0,0,0,0), mar=c(5.4,4,1.5,0),cex=2/3,cex.main=1);
lambda = (1:1000) * 0;
alpha = lambda;
for (i in 1:1000){
   X = rgamma(100, 2, 2);
   m = mean(X);
   l = mean(log(X));
   fr = function(x){(l - log(m) + log(x) - digamma(x))**2};
   lambda[i] = optimize(fr, interval=c(0,100))$minimum;
   alpha[i]=lambda[i]/m;
};
mu = lambda/alpha;
kappa = lambda;
plot(mu,kappa,xlab=expression(hat(mu)),ylab=expression(hat(kappa)))
lines(c(0.5,1.5),c(2,2),lty=3); lines(c(1,1),c(1,4),lty=3);
# dev.off()

Sigma = cov(cbind(mu,kappa))
# 0.0050553689 -0.0001849231
# -0.0001849231 0.0721859667
solve(Sigma)/100
# 1.97828040 0.00506788
# 0.00506788 0.13854406
```

3.7 Example: Log linear model with naive coordinates

Let X and Y be stochastic variables which are binary such as codes, spins, or neuronal activities [24]. The two variables are not necessarily independent and, thus, can be correlated. There are four types of events:

$$\begin{aligned}(X,Y) &= (0,0),\\ (X,Y) &= (0,1),\\ (X,Y) &= (1,0), \quad \text{and},\\ (X,Y) &= (1,1).\end{aligned} \tag{68}$$

Then, the statistical model can be specified by P_{00}, P_{01}, P_{10}, and $P_{11} = 1 - P_{00} - P_{01} - P_{10}$ and written as

$$P(X=x,Y=y|P_{00},P_{01},P_{10}) = P_{00}(1-x)(1-y) + P_{01}(1-x)y + P_{10}x(1-y) + P_{11}xy. \tag{69}$$

Then,

$$\begin{aligned}
\frac{\partial \log P(x,y)}{\partial P_{00}} &= \frac{(1-x)(1-y)-xy}{P(x,y)} = \frac{1-x-y}{P(x,y)}, \\
\frac{\partial \log P(x,y)}{\partial P_{01}} &= \frac{(1-x)y-xy}{P(x,y)} = \frac{y-2xy}{P(x,y)}, \quad \text{and,} \\
\frac{\partial \log P(x,y)}{\partial P_{10}} &= \frac{x(1-y)-xy}{P(x,y)} = \frac{x-2xy}{P(x,y)}.
\end{aligned} \tag{70}$$

The maximum likelihood estimator should satisfy

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log P(X_i,Y_i)}{\partial P_{00}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1-X_i-Y_i}{P(X_i,Y_i)} = \frac{P_{00}^{(obs)}}{P_{00}} - \frac{P_{11}^{(obs)}}{P_{11}}, \\
0 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log P(X_i,Y_i)}{\partial P_{01}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i-2X_iY_i}{P(X_i,Y_i)} = \frac{P_{01}^{(obs)}}{P_{01}} - \frac{P_{11}^{(obs)}}{P_{11}}, \quad \text{and,} \\
0 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log P(X_i,Y_i)}{\partial P_{10}} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i-2X_iY_i}{P(X_i,Y_i)} = \frac{P_{10}^{(obs)}}{P_{10}} - \frac{P_{11}^{(obs)}}{P_{11}},
\end{aligned} \tag{71}$$

where $p_{XY}^{(obs)}$ denotes the observed frequency. That is

$$\begin{aligned}
\hat{P}_{00} &= P_{00}^{(obs)}, \\
\hat{P}_{01} &= P_{01}^{(obs)}, \quad \text{and,} \\
\hat{P}_{10} &= P_{10}^{(obs)}.
\end{aligned} \tag{72}$$

Thus, the naive parameters are estimated naturally based on frequencies.

By using

$$\begin{aligned}
X^2 &= X, \quad \text{and,} \\
Y^2 &= Y,
\end{aligned} \tag{73}$$

the Fisher information matrix is obtained as

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial P_{00}}\frac{\partial \log P(X,Y)}{\partial P_{00}}\right] &= \mathrm{E}\left[\frac{1}{P^2}\right] - \mathrm{E}\left[\frac{X}{P^2}\right] - \mathrm{E}\left[\frac{Y}{P^2}\right] + 2\mathrm{E}\left[\frac{XY}{P^2}\right] = \frac{1}{P_{00}} + \frac{1}{P_{11}}, \\
\mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial P_{00}}\frac{\partial \log P(X,Y)}{\partial P_{01}}\right] &= \mathrm{E}\left[\frac{XY}{P^2}\right] = \frac{1}{P_{11}}, \\
\mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial P_{00}}\frac{\partial \log P(X,Y)}{\partial P_{10}}\right] &= \mathrm{E}\left[\frac{XY}{P^2}\right] = \frac{1}{P_{11}}, \\
\mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial P_{01}}\frac{\partial \log P(X,Y)}{\partial P_{01}}\right] &= \mathrm{E}\left[\frac{Y}{P^2}\right] = \frac{1}{P_{01}} + \frac{1}{P_{11}}, \\
\mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial P_{10}}\frac{\partial \log P(X,Y)}{\partial P_{10}}\right] &= \mathrm{E}\left[\frac{X}{P^2}\right] = \frac{1}{P_{10}} + \frac{1}{P_{11}}, \quad \text{and,} \\
\mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial P_{01}}\frac{\partial \log P(X,Y)}{\partial P_{10}}\right] &= \mathrm{E}\left[\frac{XY}{P^2}\right] = \frac{1}{P_{11}}.
\end{aligned} \tag{74}$$

That is,

$$I = \begin{pmatrix} \frac{1}{P_{00}} + \frac{1}{P_{11}} & \frac{1}{P_{11}} & \frac{1}{P_{11}} \\ \frac{1}{P_{11}} & \frac{1}{P_{01}} + \frac{1}{P_{11}} & \frac{1}{P_{11}} \\ \frac{1}{P_{11}} & \frac{1}{P_{11}} & \frac{1}{P_{10}} + \frac{1}{P_{11}} \end{pmatrix}. \tag{75}$$

Note that the Fisher information matrix is not (block) diagonal.

3.8 Example: Log linear model with η-coordinates

For the log linear model, instead of (P_{00}, P_{01}, P_{10}), one can use

$$\begin{aligned}
\eta_1 &= \mathrm{Prob}\{X=1\} = \mathrm{E}[X] = P_{10} + P_{11} = 1 - P_{00} - P_{01}, \\
\eta_2 &= \mathrm{Prob}\{Y=1\} = \mathrm{E}[Y] = P_{01} + P_{11} = 1 - P_{00} - P_{10}, \quad \text{and,} \\
\eta_{12} &= \mathrm{E}[XY] = P_{11} = 1 - P_{00} - P_{01} - P_{10},
\end{aligned} \tag{76}$$

as new coordinates or parameters. Note that there are only three degree of freedom for parameters and the parameter transformation from (P_{00}, P_{01}, P_{10}) to $(\eta_1, \eta_2, \eta_{12})$ is one-to-one:

$$\begin{aligned} P_{00} &= 1 - \eta_1 - \eta_2 + \eta_{12}, \\ P_{01} &= \eta_2 - \eta_{12}, \quad \text{and,} \\ P_{10} &= \eta_1 - \eta_{12}. \end{aligned} \tag{77}$$

Then, the statistical model can be written as

$$\begin{aligned} P(X = x, Y = y|\eta_1, \eta_2, \eta_{12}) &= P_{00}(1-x)(1-y) + P_{01}(1-x)y + P_{10}x(1-y) + P_{11}xy \\ &= (1 - \eta_1 - \eta_2 + \eta_{12})(1-x)(1-y) + (\eta_2 - \eta_{12})(1-x)y \\ &\quad + (\eta_1 - \eta_{12})x(1-y) + \eta_{12}xy. \end{aligned} \tag{78}$$

The maximum likelihood estimator is obtained by the invariance as

$$\begin{aligned} \hat{\eta}_1 &= P_{10}^{(obs)} + P_{11}^{(obs)}, \\ \hat{\eta}_2 &= P_{01}^{(obs)} + P_{11}^{(obs)}, \quad \text{and,} \\ \hat{\eta}_{12} &= P_{12}^{(obs)}. \end{aligned} \tag{79}$$

By using

$$\begin{aligned} \frac{\partial \log P(x,y)}{\partial \eta_1} &= \frac{-1 + 2x + y - 2xy}{P(x,y)}, \\ \frac{\partial \log P(x,y)}{\partial \eta_2} &= \frac{-1 + x + 2y - 2xy}{P(x,y)}, \quad \text{and,} \\ \frac{\partial \log P(x,y)}{\partial \eta_{12}} &= \frac{1 - 2x - 2y + 4xy}{P(x,y)}, \end{aligned} \tag{80}$$

the Fisher information matrix is obtained as

$$\begin{aligned} \mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial \eta_1}\frac{\partial \log P(X,Y)}{\partial \eta_1}\right] &= \mathrm{E}\left[\frac{1}{P^2}\right] - \mathrm{E}\left[\frac{Y}{P^2}\right] = \frac{1}{P_{00}} + \frac{1}{P_{10}}, \\ \mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial \eta_2}\frac{\partial \log P(X,Y)}{\partial \eta_2}\right] &= \mathrm{E}\left[\frac{1}{P^2}\right] - \mathrm{E}\left[\frac{X}{P^2}\right] = \frac{1}{P_{00}} + \frac{1}{P_{01}}, \\ \mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial \eta_1}\frac{\partial \log P(X,Y)}{\partial \eta_2}\right] &= \mathrm{E}\left[\frac{1}{P^2}\right] - \mathrm{E}\left[\frac{X}{P^2}\right] - \mathrm{E}\left[\frac{Y}{P^2}\right] + \mathrm{E}\left[\frac{XY}{P^2}\right] = \frac{1}{P_{00}}, \\ \mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial \eta_1}\frac{\partial \log P(X,Y)}{\partial \eta_{12}}\right] &= -\mathrm{E}\left[\frac{1}{P^2}\right] + \mathrm{E}\left[\frac{Y}{P^2}\right] = -\frac{1}{P_{00}} - \frac{1}{P_{10}}, \\ \mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial \eta_2}\frac{\partial \log P(X,Y)}{\partial \eta_{12}}\right] &= -\mathrm{E}\left[\frac{1}{P^2}\right] + \mathrm{E}\left[\frac{X}{P^2}\right] = -\frac{1}{P_{00}} - \frac{1}{P_{01}}, \quad \text{and,} \\ \mathrm{E}\left[\frac{\partial \log P(X,Y)}{\partial \eta_{12}}\frac{\partial \log P(X,Y)}{\partial \eta_{12}}\right] &= -\mathrm{E}\left[\frac{1}{P^2}\right] = \frac{1}{P_{00}} + \frac{1}{P_{01}} + \frac{1}{P_{10}} + \frac{1}{P_{11}}. \end{aligned} \tag{81}$$

That is,

$$I = \begin{pmatrix} \frac{1}{P_{00}} + \frac{1}{P_{10}} & \frac{1}{P_{00}} & -\frac{1}{P_{00}} - \frac{1}{P_{10}} \\ \frac{1}{P_{00}} & \frac{1}{P_{00}} + \frac{1}{P_{01}} & -\frac{1}{P_{00}} - \frac{1}{P_{01}} \\ -\frac{1}{P_{00}} - \frac{1}{P_{10}} & -\frac{1}{P_{00}} - \frac{1}{P_{01}} & \frac{1}{P_{00}} + \frac{1}{P_{01}} + \frac{1}{P_{10}} + \frac{1}{P_{11}} \end{pmatrix}. \tag{82}$$

The Fisher information matrix is not (block) diagonal, again.

3.9 Example: Log linear model with θ-coordinates

Consider another coordinate system:

$$P(X = x, Y = y|\theta) = \exp\{\theta_1 x + \theta_2 y + \theta_3 xy - \Psi(\theta)\}, \tag{83}$$

where $\Psi(\theta) = \log(1 + e^{\theta_1} + e^{\theta_2} + e^{\theta_1+\theta_2+\theta_3}) = -\log P_{00}$ is a normalization factor. The name "log linear model" originates from this form. θ's can be explicitly writen by the naive coordinates P_{ij}.

$$\begin{aligned} \theta_1 &= \log\frac{P_{10}}{P_{00}}, \\ \theta_2 &= \log\frac{P_{01}}{P_{00}}, \quad \text{and,} \end{aligned}$$

$$\theta_3 = \log \frac{P_{11}P_{00}}{P_{01}P_{10}}. \tag{84}$$

By using,

$$\begin{aligned}
\frac{\partial \log P}{\partial \theta_1} &= X - \frac{\partial \Psi}{\partial \theta_1} = X - (P_{10} + P_{11}), \\
\frac{\partial \log P}{\partial \theta_2} &= Y - \frac{\partial \Psi}{\partial \theta_2} = Y - (P_{10} + P_{11}), \quad \text{and,} \\
\frac{\partial \log P}{\partial \theta_3} &= XY - \frac{\partial \Psi}{\partial \theta_3} = XY - P_{11},
\end{aligned} \tag{85}$$

and $E[X] = P_{10} + P_{11}$ etc., the Fisher information matrix is obtained as

$$I = \begin{pmatrix} (P_{11} + P_{10})(P_{00} + P_{01}) & P_{11} - (P_{11} + P_{01})(P_{11} + P_{10}) & P_{11}(P_{00} + P_{01}) \\ P_{11} - (P_{11} + P_{01})(P_{11} + P_{10}) & (P_{11} + P_{01})(P_{00} + P_{10}) & P_{11}(P_{00} + P_{10}) \\ P_{11}(P_{00} + P_{01}) & P_{11}(P_{00} + P_{10}) & P_{11}(1 - P_{11}) \end{pmatrix}. \tag{86}$$

The Fisher information matrix is not (block) diagonal.

3.10 Counter example: Uniform distribution

Let us consider the uniform distribution [8]:

$$p(x|\theta) = \frac{1}{\theta}, \tag{87}$$

where

$$0 < X < \theta. \tag{88}$$

When $X_1, X_2, \ldots, X_n$ are sampled from $p(x|\theta)$, the maximum likelihood estimator Y is

$$Y = \max\{X_1, X_2, \ldots, X_n\}. \tag{89}$$

The Cramer–Rao theorem says

$$\mathrm{Var}[\hat{\theta}] \geq \frac{1}{n} \mathrm{E}\left[\left(\frac{\partial}{\partial \theta} \log p(x|\theta) \right)^2 \right] = \frac{\theta^2}{n}, \tag{90}$$

for any unbiased estimator $\hat{\theta}$.

Because the density function of Y is

$$f(y|\theta) = \frac{ny^{n-1}}{\theta^n}, \tag{91}$$

and

$$\mathrm{E}[Y] = \int_0^\theta y f(y) dy = \frac{n}{n+1}\theta, \tag{92}$$

let us use an unbiased estimator

$$\frac{n+1}{n} Y, \tag{93}$$

instead of Y itself. Then,

$$\begin{aligned}
\mathrm{Var}\left[\frac{n+1}{n} Y\right] &= \left(\frac{n+1}{n}\right)^2 \mathrm{Var}[Y] \\
&= \left(\frac{n+1}{n}\right)^2 (\mathrm{E}[Y^2] - \mathrm{E}[Y]^2) \\
&= \left(\frac{n+1}{n}\right)^2 \left(\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 \right) \\
&= \frac{\theta^2}{n(n+2)}.
\end{aligned} \tag{94}$$

Apparently, this VIOLATES the Cramer–Rao inequality.

Actually, in the derivation of the Cramer–Rao inequality, the exchangeability of derivative and integral were assumed and used. That is, when a parameter specifies the region of integral, the Cramer–Rao inequality does not necessarily holds.

3.11 Appendix. Invariance of maximum likelihood estimation under parameter transformation

Another merit of the maximum likelihood estimation is that the result does not depend on the coordinates or parameter sets used.

For example, the parameters of the gamma distribution can be transformed from (α, λ) to (μ, κ). When

$$\begin{aligned} \mu &= \frac{\lambda}{\alpha} \quad \text{and} \\ \kappa &= \lambda \end{aligned} \tag{95}$$

are satisfied, the two distribution $p(x|\alpha, \lambda)$ and $p(x|\mu, \kappa)$ have exactly the same function form $p(x)$.

As the maximum likelihood estimation tries to estimate the form of the density function $p(x)$ from observations, the estimates using different parameter sets give the same result. That is, the following "invariance" holds for the estimators:

$$\begin{aligned} \hat{\mu} &= \frac{\hat{\lambda}}{\hat{\alpha}} \quad \text{and} \\ \hat{\kappa} &= \hat{\lambda}. \end{aligned} \tag{96}$$

Thus the result is independent from parameter sets. Although there are choices for parametrization, the reality or the function form of the distribution is invariant. This allows us to change parameters or coordinates and play around. In this way, the maximum likelihood estimation is compatible with the information geometry, where parameters are transformed so that the Fisher matrix for the new parameter set is diagonal.

4. Exponential Family and Parameter Orthogonalization

For the exponential family,

$$p(x|\boldsymbol{\theta}) = \exp\left\{\sum \theta_i k_i(x) - \Psi(\boldsymbol{\theta})\right\}, \tag{97}$$

a parameter transform which diagonalize the Fisher information matrix exists [5, 6]. Thus, for practical applications, it is worth testing if the probabilistic model one considers belongs to this family. The exponential family includes normal, gamma, beta, binomial, Poisson, and negative binomial distributions and shares many nice statistical properties [8].

New coordinates (or parameters) η's can be defined as

$$\eta_i = \mathrm{E}[k_i(X)] = \frac{\partial}{\partial \theta_i} \Psi(\boldsymbol{\theta}). \tag{98}$$

Because it can be easily proved that

$$\mathrm{E}\left[\frac{\partial}{\partial \theta_i} \log p(X|\boldsymbol{\theta}) \frac{\partial}{\partial \eta_j} \log p(X|\boldsymbol{\eta})\right] = \delta_{ij}, \tag{99}$$

it is a good idea to use mixed coordinates such as (θ_1, η_2) or (η_1, θ_2) instead of θ-coordinates (θ_1, θ_2) or η-coordinates (η_1, η_2).

4.1 Example: Gamma distribution

Consider the gamma distribution with the conventional coordinates:

$$\begin{aligned} p(x|\lambda, \alpha) &= \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x} \\ &= \exp\{(\lambda - 1)\log x - \alpha x - \log \Gamma(\lambda) + \lambda \log \alpha\}. \end{aligned} \tag{100}$$

This belongs to the exponential family because one can take

$$\begin{aligned} \theta_1 &= \lambda - 1, \\ k_1(x) &= \log x, \\ \theta_2 &= -\alpha, \quad \text{and,} \\ k_2(x) &= x. \end{aligned} \tag{101}$$

Then, η's are defined by

$$\eta_1 = \frac{\partial \Psi(\boldsymbol{\theta})}{\partial \theta_1} = \frac{\partial \Psi(\boldsymbol{\theta})}{\partial \lambda} = \frac{\Gamma'(\lambda)}{\Gamma(\lambda)} - \log \alpha, \quad \text{and,}$$
$$\eta_2 = \frac{\partial \Psi(\boldsymbol{\theta})}{\partial \theta_2} = -\frac{\partial \Psi(\boldsymbol{\theta})}{\partial \alpha} = \frac{\lambda}{\alpha}. \tag{102}$$

Thus, the mixture coordinates $(\theta_1 + 1, \eta_2) = (\lambda, \mu) = (\kappa, \mu)$ give a diagonal Fisher information matrix. Note that the matrix is diagonal at any point of the coordinates (μ, κ). That is, the Fisher information matrix was globally diagonalized by the parameter transformation.

5. Mixture Family and Parameter Orthogonalization

For the mixture family,

$$p(x|\boldsymbol{\eta}) = \sum \eta_i q_i(x) + \left(1 - \sum \eta_i\right) q_0(x), \tag{103}$$

a parameter transform which diagonalize the Fisher information matrix exists [5, 6]. Thus, for practical applications, it is worth testing if the probabilistic model one considers belongs to this family. Note that a mixture family appears in a doubly stochastic process where a (hidden) index i is chosen with probability η_i and then x is generated by $q_i(x)$.

New coordinates (or parameters) θ's can be defined as

$$\theta_i = \frac{\partial}{\partial \eta_i} \phi(\boldsymbol{\eta}) = \int (q_i(x) - q_0(x)) \log p(x|\boldsymbol{\eta}) dx, \tag{104}$$

where $\phi(\eta)$ is the negative entropy

$$\phi(\boldsymbol{\eta}) = \mathrm{E}[\log p(X, \boldsymbol{\eta})]. \tag{105}$$

Because it can be easily proved that

$$\mathrm{E}\left[\frac{\partial}{\partial \theta_i} \log p(X, \boldsymbol{\theta}) \frac{\partial}{\partial \eta_j} \log p(X, \boldsymbol{\eta})\right] = \delta_{ij}, \tag{106}$$

it is a good idea to use mixed coordinates such as (θ_1, η_2) or (η_1, θ_2) instead of a θ-coordinates (θ_1, θ_2) or η-coordinates (η_1, η_2).

5.1 Example: Log linear model

Consider the log linear model with the η coordinates:

$$P(X = x, Y = y|\boldsymbol{\eta}) = (1 - \eta_1 - \eta_2 + \eta_{12})(1 - x)(1 - y) + \eta_1 x(1 - y) + \eta_2 y(1 - x) + \eta_{12}(x + y - 3xy). \tag{107}$$

This belongs to the mixture family because one can take

$$\begin{aligned}
\eta_0 &= 1 - \eta_1 - \eta_2 + \eta_{12}, \\
q_0(x, y) &= (1 - x)(1 - y), \\
\eta_1 &= \eta_1, \\
q_1(x, y) &= x(1 - y), \\
\eta_2 &= \eta_2, \\
q_2(x, y) &= y(1 - x), \\
\eta_3 &= -\eta_{12}, \quad \text{and,} \\
q_3(x, y) &= x + y - 3xy.
\end{aligned} \tag{108}$$

Then, θ_3 is defined by

$$\theta_3 = \mathrm{E}[(q_3(X, Y) - q_0(X, Y)) \log P(X, Y)] = \log \frac{P_{10} P_{01}}{P_{00} P_{11}}. \tag{109}$$

Thus, the mixture coordinates $(\eta_1, \eta_2, \theta_3)$ give a (block-)diagonal Fisher information matrix. Again,

$$\begin{aligned}
\eta_1 &= P_{10} + P_{11}, \\
\eta_2 &= P_{01} + P_{11}, \quad \text{and,} \\
\theta_3 &= \log \frac{P_{10} P_{01}}{P_{00} P_{11}}.
\end{aligned} \tag{110}$$

The Fisher information matrix is given by

$$I = \left(\frac{1}{\frac{1}{P_{00}}+\frac{1}{P_{10}}+\frac{1}{P_{01}}+\frac{1}{P_{11}}}\right)\begin{pmatrix} (\frac{1}{P_{00}}+\frac{1}{P_{10}})(\frac{1}{P_{01}}+\frac{1}{P_{11}}) & \frac{1}{P_{00}}-(\frac{1}{P_{00}}+\frac{1}{P_{01}})(\frac{1}{P_{00}}+\frac{1}{P_{10}}) & 0 \\ \frac{1}{P_{00}}-(\frac{1}{P_{00}}+\frac{1}{P_{01}})(\frac{1}{P_{00}}+\frac{1}{P_{10}}) & (\frac{1}{P_{00}}+\frac{1}{P_{01}})(\frac{1}{P_{10}}+\frac{1}{P_{11}}) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{111}$$

Note that the Fisher information matrix is block diagonal.

6. Remarks

In this paper, we mostly focused on the local metrics of the information geometry. One thing which is very important, but we did not discuss is Pythagoras theorem for Kullback–Leibler divergence. The Pythagoras theorem provides another benefit of information geometry from a more global viewpoint [4, 5, 12–14, 16, 22].

Here we did not discuss model selection [1, 7, 27]. But it is important to choose the statistical model which fits the data best among different parametric statistical models. It can be performed by using the Akaike information criteria in combination with the maximum likelihood estimation.

REFERENCES

[1] Akaike, H., "A new look at the statistical model identification," *IEEE Trans. Auto. Control*, **19**: 716–723 (1974).
[2] Amari, S., "Differential geometry of curved exponential families — curvatures and information loss," *Ann. Statist.*, **10**: 357–385 (1982).
[3] Amari, S., Differential-geometrical methods in statistics, Lecture notes in statistics 28, Springer-Verlag, New York (1985).
[4] Amari, S., "Information geometry of the em and em algorithms for neural networks," *Neural Networks*, **8**: 1379–1408 (1995).
[5] Amari, S., "Information geometry on hierarchy of probability distributions," *IEEE Trans. on Information Theory*, **47**: 1701–1711 (2001).
[6] Amari, S., and Nagaoka, H., Methods of Information Geometry, American Mathematical Society, Providence, RI (2001).
[7] Bishop, C. M., Pattern Recognition and Machine Learning, Springer-Verlag, Berlin (2006).
[8] Casella, G., and Berger, R. L., Statistical Inference, Duxbury, Pacific Grove, CA (2002).
[9] Durrett, R., Essentials of Stochastic Processes, Springer-Verlag, Berlin (1999).
[10] Feller, W., An Introduction to Probability Theory and Its Applications Vol. 1, Wiley, Hoboken (1968).
[11] Gardiner, C. W., Handbook of Stochastic Methods, Springer-Verlag, Berlin (1985).
[12] Ikeda, S., Amari, S., and Nakahara, H., "Convergence of the wake-sleep algorithm," *Advances in Neural Information Processing Systems*, **11**: 239–245 (1999).
[13] Ikeda, S., Tanaka, T., and Amari, S., "Stochastic reasoning, free energy, and information geometry," *Neural Comput.*, **16**: 1779–1810 (2004).
[14] Ikeda, S., Tanaka, T., and Amari, S., "Information geometry for turbo decoding," *Systems and Computers in Japan*, **36**: 758–765 (2005).
[15] Kanji, G. K., 100 Statistical Tests, Sage, London (2006).
[16] Komaki, F., "Bayesian predictive densities based on latent information priors," *J. Stat. Plan Inference*, **141**: 3705–3715 (2011).
[17] Kuznetsov, I. A., and Kuznetsov, Y. A., Elements of Applied Bifurcation Theory, Springer, Berlin (1998).
[18] Miura, K., Okada, M., and Amari, S., "Estimating spiking irregularities under changing environments," *Neural Comput.*, **18**: 2359–86 (2006).
[19] Miura, K., Okada, M., and Amari, S., "Unbiased estimator of shape parameter for spiking irregularities under changing environments," *Advances in Neural Information Processing Systems*, **18**: 891–8 (2006).
[20] Miura, K., Tsubo, Y., Okada, M., and Fukai, T., "Balanced excitatory and inhibitory inputs to cortical neurons decouple firing irregularity from rate modulations," *J. Neurosci.*, **27**: 13802–12 (2007).
[21] Miura, K., and Uchida, N., "A rate-independent measure of irregularity for event series and its application to neural spiking activity," *47th IEEE Conference on Decision and Control*, pages 2006–11 (2008).
[22] Murata, N., Takenouchi, T., Kanamori, T., and Eguchi, S., "Information geometry of u-boost and bregman divergence," *Neural Comput.*, **16**: 1437–1481 (2004).
[23] Murray, M. K., and Rice, J. W., Differential Geometry and Statistics, Chapman and Hall, New York (1993).
[24] Nakahara, H., and Amari, S., "Information-geometric measure for neural spikes," *Neural Comput.*, **14**: 2269–2316 (2002).
[25] Ott, E., Chaos in Dynamical Systems, Cambridge University Press, Cambridge (2002).
[26] Risken, H., The Fokker-Planck Equation, Springer-Verlag, Berlin (1989).
[27] Tibshirani, R., Hastie, T., and Friedman, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, Berlin (2009).
[28] van der Vaart, A. W., Asymptotic Statistics, Cambridge University Press, Cambridge (1998).
[29] Wiggins, S., Introduction to Applied Nonlinear Dynamical Systems and Chaos, Springer, Berlin (1990).