The Fisher information matrix: A tutorial for calculation for decision making models

Kazuya Fujita, Kensuke Okada, Kentaro Katahira

Abstract

Measuring trait parameters with high estimation precision is important in psychology. In terms of estimation precision, there is a fundamental quantity, the Fisher information, in statistics. This paper introduces the Fisher information from the perspective of estimation precision and experimental stimulus selection. This paper explained the asymptotic efficiency of the maximum likelihood estimator to explain the importance of the Fisher information in statistics. Then, this paper introduced Computerized adaptive testing (CAT), which uses the Fisher information for stimulus selection as an example of application in psychology. In addition, we recommended CAT for cognitive models due to the extent of the effect and low cost of CAT. In section 3, this paper explained how to calculate the Fisher information of decision making models. In section 4, this paper showed simulation example to explain how to use the Fisher information for selecting stimuli in cognitive experiment.

1 Introduction

There is a fundamental quantity called the Fisher information matrix (Ly et al., 2017). The Fisher information matrix determines the characteristics of statistical estimation and hypothetical testing. The derivatives of log-likelihood are called the score function. The Fisher information matrix is defined as the variance of the score function. For now, it can be said that the Fisher information is something like the amount of information obtained from observations. When a parameter is unidimensional (multi-dimensional), we call the Fisher information (matrix).

The most important example which implies the importance of the Fisher information matrix for the present paper is that the inverse of Fisher information (matrix) can predict variance (covariance matrix) of the Maximum Likelihood estimator (MLE; Chang, 2015). Higher Fisher information leads to higher estimation precision (lower variance of estimation). Further, the example which uses the Fisher information matrix directly in psychological field is Computerized adaptive testing (CAT; Weiss & Kingsbury, 1984; Segall, 2004; Meijer & Nering, 1999; van der Linden, 2018; van der Linden, 1998). In CAT, the Fisher information (matrix) is used for selecting stimuli (e.g., quiz of an educa-

tional test, an item for questionnaire, reward and reward probability for decision making task) adaptively and optimally.

Although the Fisher information matrix is a fundamental and important quantity in statistics, there are few cases where it is used effectively in psyc hology or psychological statistics. It will be useful for determining the experimental design, especially selecting stimuli, with the optimal and objective procedures.

In our understanding, there are some reasons why the Fisher information matrix is not often used in psychology. The first reason is that researchers do not recognize the value of the Fisher information matrix. The second reason is that researchers will have some hurdles in deriving its analytical solution. Therefore, this paper explains why the Fisher information matrix is important, especially for statistical inference in section 2. Then this paper explains how to calculate the Fisher information matrix with some calculation examples in section 3. In addition, this paper provides R code example in section 4.

2 The importance of the Fisher information matrix

2.1 Maximum likelihood estimation

This section check prerequisite knowledge, including probability density function, independent and identically distribution, likelihood function, and maximum likelihood (ML) estimation. If you are not familiar with them, you can refer Myung (2003). This section is based on his tutorial article.

Let $\boldsymbol{y} = (y_1, ..., y_N)$ be N participants (or trials) observations. Probability density function (pdf) provides probability of observation y_i, $p(y_i|\boldsymbol{\xi})$, where $\boldsymbol{\xi}$ are parameters. For instance, normal distribution with mean μ and variance σ^2 is defined with below pdf:

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

Observation $\boldsymbol{y}$ is usually assumed independent and identically distributed, which is represented by i.i.d. The assumption that $\boldsymbol{y}$ is generated independent and identically distribution means

$$p(y_1, ..., y_N|\boldsymbol{\xi}) = p(y_1|\boldsymbol{\xi}) \cdots p(y_N|\boldsymbol{\xi}), \tag{2}$$

which indicates joint distribution, $p(y_1, ..., y_N|\boldsymbol{\xi})$, can be calculated with the products of marginal distribution, $p(y_i|\boldsymbol{\xi})$.

Let $L(\boldsymbol{\xi}|\boldsymbol{y})$ be likelihood function which is calculated with pdf (e.g., equation (1)). Maximum likelihood estimator (MLE) is defined by

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} L(\boldsymbol{\xi}|\boldsymbol{y}) \tag{3}$$

where $\arg\max_{\boldsymbol{\xi}}$ provides argument of maximum likelihood. Therefore, MLE can be estimated by solving

$$\frac{\partial}{\partial \boldsymbol{\xi}} \log L(\boldsymbol{\xi}|\boldsymbol{y}) = 0 \tag{4}$$

where $\frac{\partial}{\partial \boldsymbol{\xi}}$ is differential w.r.t. $\boldsymbol{\xi}$.

2.2 Asymptotic characteristic of ML

To understand the importance of the Fisher information, this section explains the basic fact in statistics, unbiased estimator, information inequality (Cramer-Rao inequality), and asymptotic normality of MLE.

Estimation is to map some (real) values from observed data from statistical perspective. There are many candidate as estimation method. Let $\xi(\chi)$ be estimator where χ is observation pattern. For extreme example, constant function which map a specific value regardless of observed data can be treated as a estimator. We have to evaluate the goodness of the estimation method.

One of the evaluation criteria is mean square error (MSE). Let us assume random variable x_i is generated by distribution having pdf, $p(x|\xi^*)$, with true parameter value ξ^*. Although for simplicity parameter ξ is assumed unidimensional, the case parameters are multidimensional is the same. The estimator is better when it is closer to the true parameter value. We consider the mean square error defined by

$$\text{MSE} = E\Big[(\xi(\chi) - \xi^*)^2\Big] \tag{5}$$

as criterion. Note that, when continuous data x is generated by $p(x)$, $x \sim p(x)$, expectation $E[x]$ and variance $Var[x]$ of x are defined by

$$\begin{aligned} E[X] &= \int_{\chi} xp(x)dx, \\ Var[X] &= \int_{\chi} (x - E[X])^2 p(x)dx. \end{aligned} \tag{6}$$

MSE can be decomposed as

$$MSE = Var[\gamma(\chi)] + \Big(E[\gamma(\chi)] - \xi^*\Big)^2 \tag{7}$$

where the first term is the variance of the estimator, and the second term is square of bias between the expectation of estimator and true parameter value. They are all positive values. It is better to make MSE smaller. There are two ways to make MSE smaller. One is to restrict estimator to unbiased estimator which has zero bias, i.e., $E[\xi(\chi)] - \xi^* = 0$. The second is to reduce the variance of the estimator while permitting bias a bit.

We are focusing on the former method. If we restrict the unbiased estimator, the second term of equation (7) becomes 0, therefore MSE becomes variance of estimator exactly. In this sense, the estimator with the smallest variance becomes the best estimator if we restrict the unbiased estimator. Furthermore, the estimator is represented by just ξ hereafter for brevity.

Although lower variance of estimator means better estimation method, there is a lower bound of variance of the estimator. The below formula, which is called information inequality or Cramer-Rao inequality, provides this lower bound. Let $\hat{\xi}$ be an unbiased estimator of true parameter ξ^*, and let $F(\xi^*)$ be the Fisher information. Then, (in some appropriate conditions) the variance of the estimator satisfies

$$Var[\hat{\xi}] \geq \frac{1}{F(\xi^*)}. \tag{8}$$

Although the lower variance of the estimator is better, it can not be smaller than the inverse of the Fisher information when the parameter is unidimensional. That is, the inverse of the Fisher information is bound. In other words, an estimator which can reduce its variance up to the inverse of the Fisher information is the best estimation method in the class of unbiased estimators. The estimator which holds $Var[\hat{\xi}] = \frac{1}{F(\xi^*)}$ is called efficient estimator. In this paper, we omit conditions, proof, and multidimensional results of the equation (8).

Let us go back to the ML estimator, although the above discussion can be applied to various estimators. ML has various desirable aspects. One of them is asymptotic normality and characteristic of asymptotic variance. ML estimator $\hat{\boldsymbol{\xi}}$ holds

$$\hat{\xi} \to N\Big(\xi^*, \frac{1}{F(\xi^*)}\Big) \tag{9}$$

in regularity conditions (Chang, 2015). This means that the estimator asymptotically follows the normal distribution, which is called asymptotic normality. The mean of normal distribution becomes true parameter value $\boldsymbol{\xi^*}$, and the variance (covariance matrix) of the estimator becomes the inverse of the Fisher information (matrix) asymptotically. Although we omit the regularity conditions and proof of equation (9), most psychological models hold that equation.

In summary of this section, the estimator, the variance of which is the inverse of the Fisher information, is the best from the MSE perspective if we restrict the unbiased estimator. From equation (9), MLE holds this characteristic asymptotically. This characteristic is called asymptotic efficiency. Furthermore, psychology researchers can optimize not only estimation method (i.e., ML) but also experimental design and stimuli from equation (9). That is, there is room to reduce variance (i.e., increase the Fisher information) in terms of experimental design and stimuli (e.g., Lindley, 1956; Chaloner & Verdinelli, 1995) even if we restrict the estimator to MLE.

2.3 Definition of the Fisher information matrix

Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_k)$ be k-dimensional parameters. The Fisher information matrix of i-th participant (or trial) for paramter $\boldsymbol{\xi}$ is defined as

$$F_i(\boldsymbol{\xi}) = E\left[\Big(\frac{\partial}{\partial \boldsymbol{\xi}} \log L(\boldsymbol{\xi}|y_i)\Big)\Big(\frac{\partial}{\partial \boldsymbol{\xi}} \log L(\boldsymbol{\xi}|y_i)\Big)^t\right] \tag{10}$$

where $\frac{\partial}{\partial \boldsymbol{\xi}} \log L(\boldsymbol{\xi}|y_i)$ is $k \times 1$ column vector, ^{t} is transpose symbol. That is, $F_i(\boldsymbol{\xi})$ is $k \times k$ matrix. The (m, n) element of the Fisher information matrix is $F_i(\boldsymbol{\xi})_{m,n} = E\left[\left(\frac{\partial}{\partial \xi_m} \log L(\boldsymbol{\xi}|y_i)\right)\left(\frac{\partial}{\partial \xi_n} \log L(\boldsymbol{\xi}|y_i)\right)\right]$. The expectation is over dependent variables $\boldsymbol{y}$, assuming model $p(y_i)$ is true. The Fisher information is dependent on parameter values $\boldsymbol{\xi}$ and stimuli (and, of course, model). Conventionally, the stimuli symbol is omitted in the Fisher information matrix as above representation, however, note that the Fisher information is dependent on experimental design and stimuli.

In addition, when true model is known, below equation

$$E\left[\left(\frac{\partial}{\partial \boldsymbol{\xi}} \log L(\boldsymbol{\xi}|y_i)\right)\left(\frac{\partial}{\partial \boldsymbol{\xi}} \log L(\boldsymbol{\xi}|y_i)\right)^t\right] = -E\left[\frac{\partial^2}{\partial \xi^2} \log L(\boldsymbol{\xi}|y_i)\right] \tag{11}$$

holds. That is, researchers can calculate the Fisher information matrix with either square of score function or second derivatives of the log-likelihood function. Which method is easier is dependent on the models. In second derivative's definition, the (m, n) element of the Fisher information matrix is $F_i(\boldsymbol{\xi})_{m,n} = -E\left[\frac{\partial^2}{\partial \xi_m \partial \xi_n} \log L(\boldsymbol{\xi}|y_i)\right]$.

2.4 The Fisher information for stimulus selection

We have explained the background theory of computerized adaptive testing (CAT). When the parameter ξ is unidimensional, equation (9) indicates that larger Fisher information means higher estimation precision (lower variance of estimator) of $\hat{\xi}$. That is, from equation (8) and equation (9), ML estimation and stimulus selection based on the Fisher information is the optimal method in terms of estimation precision, though we should meet some conditions such as unbiased characteristic, etc.

When the parameters $\boldsymbol{\xi}$ are multidimensional, researchers have to map the Fisher information matrix to a scalar for maximization. One of the most famous methods is to maximize the determinant of the Fisher information matrix (Segall, 1996), which is called D-optimal (Mulder & van der Linden, 2009) stimulus selection.

Previous researchers have been studying how to select stimuli for parameter estimation (Ahn et al., 2020; DiMattina, 2015; Kontsevich & Tyler, 1999; Toubia, Johnson, Evgeniou, & Delquié, 2013; Watson & Pelli, 1983) or model selection (Heck & Erdfelder, 2019; Myung & Pitt, 2009; Cavagnaro, Pitt, Gonzalez, & Myung, 2013). CAT is stimulus selection method for parameter estimation. One of the merits of CAT (i.e., the Fisher information-based method) is low computational cost. This is because the Fisher information is defined as expectation over observations, which means we do not calculate expectation (integral) numerically.

2.4.1 Some comments

Although CAT is applied in questionnaire and educational test situations (Weiss & Kingsbury, 1984; McGlohen & Chang, 2008; Gibbons et al., 2012; van der Linden, 1998), it can be applied psychological, especially cognitive, experiments. It is obvious considering the statistical background. In addition, there are two merits to apply to psychological experiments compared to questionnaires or education tests. One is the cost to apply it. In educational tests, researchers have to collect over 100 participant data to quantify test items as difficulty parameter etc. In contrast, some cognitive experiments which use objective stimulus (e.g., physical stimulation such as sound, length, and so on, or reward, reward probability) do not need to do it because the scale of stimuli is fixed beforehand. The Second is the extent of the stimulus effect. In general, the effect of CAT will be larger when the number of candidate stimuli patterns is large. The number of stimuli patterns in cognitive experiments tends to be larger than that of educational tests. For instance, suppose two alternatives decision making tasks consisting of reward and reward probability. If the number of reward and reward probability candidates are 10, then researchers have to choose stimuli from $10^4 = 10,000$ patterns.

CAT is the method for adaptive stimulus selection. The adaptive means to select stimuli after collecting participant's observations. This is because the appropriate stimuli with higher information tend to be different among participants (i.e., trait parameters). For example, suppose that researchers have already collected first to j-th trials data. Let $F_{1:j}(\hat{\boldsymbol{\xi}})$ denote the Fisher information matrix of a participant's responses to the first to j-th trials and let $F_{j+1}(\hat{\boldsymbol{\xi}})$ denote the Fisher information matrix based on the candidate stimuli that are presented at $(j+1)$-th trial. In the D-optimal selection, stimuli that maximize

$$\det(F_{1:j}(\hat{\boldsymbol{\xi}}) + F_{j+1}(\hat{\boldsymbol{\xi}})) \tag{12}$$

are selected as the stimuli of the $(j+1)$-th trial to be presented (Segall, 1996). Further, $\hat{\boldsymbol{\xi}}$ is estimates using first to j-th trial observation.

The Fisher information matrix can be used for selecting stimuli before conducting experiments (i.e., fixed stimuli design). This is because the Fisher information matrix is calculated with the expectation of a dependent variable, therefore researchers do not need to collect dependent variable observations. Researchers can use the Fisher information matrix as reference information in addition to a previous study's design and researcher's experience. From equation (9), the Fisher information can be used for calculating confidence interval and sample size design for reducing its variance up to specific values (Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017) . Furthermore, Ly et al. (2017) introduce application of the Fisher information matrix in Jeffreys's prior (Jeffreys, 1946; Jeffreys, 1961) and minimum description length (MDL; Myung, Navarro, & Pitt, 2006) which is one of model selection criteria.

3 Calculation of the Fisher information matrix

This section explains how to derive the Fisher information matrix step-by-step way. First, this paper introduces the Fisher information matrix of normal distribution because it is the most famous. Second, this paper introduces the Fisher information matrix of the logistic regression model because it is the most basic and useful in terms of much application.

3.1 Normal distribution

3.1.1 Log function and exponential function

Let us check prerequisite knowledge for below calculation.

$$\sum_{i=1}^{N} a_i = a_1 + a_2 + \cdots + a_N$$

$$\sum_{i=1}^{N} a = N * a$$

$$\Pi_{i=1}^{N} a_i = a_1 * a_2 * \cdots * a_N$$

$$\log a^b = b * \log a !!! FIXME !!! \tag{A1}$$

$$\log ab = \log a + \log b, \quad \log \Pi_{i=1}^{N} a_i = \sum_{i=1}^{N} \log a_i \tag{A2}$$

$$\log(e^x) = x, \quad \exp(\log(x)) = x \tag{A3}$$

$$\sqrt{a} = a^{\frac{1}{2}}, \frac{1}{a} = a^{-1}, \frac{1}{\sqrt{a}} = (a^{\frac{1}{2}})^{-1} = a^{-\frac{1}{2}}, (a^b)^c = a^{bc} \tag{A4}$$

3.1.2 Calculation of log likelihood function

Suppose data y_i is generated by $N(\mu, \sigma^2)$, and we assume i.i.d. That is, suppose

$$y_i \sim N(\mu, \sigma^2), \quad i = 1, ..., N. \tag{14}$$

The pdf of normal distribution is defined by

$$p(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mu)^2 \right\}. \tag{15}$$

Calculating log function and using equations (A1) to (A4), this is represented by

$$\log p(y_i|\mu, \sigma^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y_i - \mu)^2 \tag{16}$$

The entire data $\boldsymbol{y} = (y_1, ..., y_N)$ is represented by

$$\begin{aligned}\log p(\mathbf{y}|\mu,\sigma^{\mathbf{2}}) &= \log \Pi_{i=1}^{N} p(y_i|\mu,\sigma^2) = \sum_{i=1}^{N} \log p(y_i|\mu,\sigma^2) \\ &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \sum_{i=1}^{N}\frac{1}{2\sigma^2}(y_i-\mu)^2\end{aligned} \tag{17}$$

using i.i.d assumption and equation (A2).

3.1.3 Differential operation

Let us check prerequisite knowledge about differential. The differential of the function $f(x)$ over x is represented by $\frac{\partial}{\partial x}f(x)$ or $f'(x)$. The equations

$$\frac{\partial}{\partial x}(f_1(x)+f_2(x)) = f_1'(x)+f_2'(x), \quad \frac{\partial}{\partial x}\Big(a*f(x)\Big) = a*f'(x) \tag{B1}$$

$$\{f(g(x))\}' = f'(g(x))*g'(x) \tag{B2}$$

$$\frac{\partial}{\partial x}x^{\alpha} = \alpha x^{\alpha-1}, \frac{\partial}{\partial x}c = 0 \tag{B3}$$

$$\frac{\partial}{\partial x}e^{x} = e^{x}, \quad \frac{\partial}{\partial x}e^{f(x)} = e^{f(x)}\Big(\frac{\partial}{\partial x}f(x)\Big) \tag{B4}$$

$$\frac{\partial}{\partial x}\log x = \frac{1}{x}, \quad \frac{\partial}{\partial x}\log f(x) = \frac{1}{f(x)}\Big(\frac{\partial}{\partial x}f(x)\Big) \tag{B5}$$

$$\left\{\frac{1}{g(x)}\right\}' = -\frac{g'(x)}{g(x)}, \quad \left\{\frac{f(x)}{g(x)}\right\}' = \frac{f'(x)g(x)-f(x)g'(x)}{g(x)^2} \tag{B6}$$

are used for below calculation.

3.1.4 Calculation of score function

The differential with respect to μ is represented by

$$\begin{aligned}\frac{\partial}{\partial\mu}\log p(\mathbf{y}|\mu,\sigma^{\mathbf{2}}) &= 0-0-\frac{1}{2\sigma^2}\sum_{i=1}^{N}2(y_i-\mu)(-1) \qquad (\because B1,B2,B3) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\mu)\end{aligned} \tag{19}$$

and, the differential with respect to σ^2 is represented by

$$\frac{\partial}{\partial\sigma^2}\log p(\boldsymbol{y}|\mu,\sigma^2) = -\frac{N}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2. \qquad (\because B1,B2,B3,B5) \tag{20}$$

3.1.5 The characteristic of expectation and variance

Let us check prerequisite knowledge about expectation and variance as below:

$$Var[X] = E[(x - E[X])^2] \tag{D1}$$
$$E[X + Y] = E[X] + E[Y], \quad E[aX] = aE[X] \tag{D2}$$
$$\text{When X and Y are independent,} \quad E[XY] = E[X]E[Y]. \tag{D3}$$

3.1.6 Calculation of the Fisher information matrix

The Fisher information matrix is 2×2 matrix because there are two parameters, $\boldsymbol{\xi} = (\mu, \sigma^2)$. (1, 1) element is the Fisher information of μ and (2, 2) element is that of σ^2.

For the (1, 1) element, we should calculate expectation on the square of the score function. From equation (19), the square of the score function is

$$\begin{aligned}
&\left(\frac{\partial}{\partial \mu} \log p(\boldsymbol{y}|\mu, \sigma^2)\right)^2 \\
&= \frac{1}{\sigma^4}\left(\sum_{i=1}^{N}(y_i - \mu)\right)\left(\sum_{j=1}^{N}(y_j - \mu)\right) \\
&= \frac{1}{\sigma^4}\Big\{(y_1 - \mu) + \cdots + (y_N - \mu)\Big\}\Big\{(y_1 - \mu) + \cdots + (y_N - \mu)\Big\}.
\end{aligned} \tag{22}$$

In this way, the calculation for summation of the product of "observation y_i - mean μ" is often needed in the Fisher information derivation. Let us calculate the part of expectation of equation (22). When $i = j$, the expectation is

$$E_y[(y_i - \mu)^2] = \sigma^2 \tag{23}$$

because y_i is generated by i.i.d N(μ, σ^2). In addition, from i.i.d and N(μ, σ^2) assumption, when $i \neq j$, the expectation is

$$\begin{aligned}
E_y[(y_i - \mu)(y_j - \mu)] &= E_y[y_i y_j - y_i\mu - y_j\mu + \mu^2] \\
&= E_y[y_i y_j] - \mu^2 - \mu^2 + \mu^2 \qquad (\because D2, E[y_i] = \mu) \\
&= \mu^2 - \mu^2 - \mu^2 + \mu^2 \qquad (\because D3) \\
&= 0.
\end{aligned} \tag{24}$$

That is, when observation is independent of each other, the expectation on the product of "observation - mean" equals 0. In Item response theory, the assumption that observation is independent of each other (conditional on parameters) is called (local) independent.

This calculation appears in many models. Assuming local independence, the equation

$$\begin{aligned} & E_y\left[\left(\sum_{i=1}^{N}(\text{Observation}_i - \text{Mean})\right)\left(\sum_{j=1}^{N}(\text{Observation}_j - \text{Mean})\right)\right] \\ &= \sum_{i=1}^{N} E_y\left[(\text{Observation}_i - \text{Mean})^2\right] \\ &= N \times \text{Variance} \end{aligned} \tag{25}$$

holds. Eventually, the expectation of equation (22), (1, 1) element of the Fisher information matrix, is represented by

$$\begin{aligned} E\left[\left(\frac{\partial}{\partial\mu}\log p(\boldsymbol{y}|\mu,\sigma^2)\right)^2\right] &= E\left[\frac{1}{\sigma^4}\left(\sum_{i=1}^{N}(y_i-\mu)\right)\left(\sum_{j=1}^{N}(y_j-\mu)\right)\right] \\ &= \frac{1}{\sigma^4}E\left[\left(\sum_{i=1}^{N}(y_i-\mu)^2\right)\right] \quad (\because \text{local independence}) \\ &= \frac{1}{\sigma^4} * N * \sigma^2 \qquad (\because D2, Var[y_i] = \sigma^2) \\ &= \frac{N}{\sigma^2}. \end{aligned} \tag{26}$$

Further, the variance of MLE becomes $\frac{\sigma^2}{N}$.

Although researchers can calculate (1, 2) and (2, 2) elements in the same way, we omit them because they are not used below. In this case, using second derivatives is easier for these derivations.

3.2 Logistic regression model

The second example is the logistic regression model. As we explain below, many decision making models have the common structure with the logistic model. Therefore, this section's derivation is basic and important for below models.

3.2.1 Log likelihood function

Without loss of generality, assume the logistic model having only one independent variable. Let $y_j, j = 1, ..., J$ be 0-1 observation (e.g., Yes/No, correct/incorrect). Suppose observation y_j is generated by bernoulli distribution (i.i.d) having probability $\eta_j(\boldsymbol{\xi})$. The parameters are intercept α and slope β, namely $\boldsymbol{\xi} = (\alpha, \beta)$. The logistic model assumes this probability $\eta_j(\boldsymbol{\xi})$ is calculated with the logistic function. That is, the logistic model is defined by

$$
\begin{aligned}
y_j &\sim Bernoulli(\eta_j(\boldsymbol{\xi})) \\
\eta_j(\boldsymbol{\xi}) &= logit^{-1}\{\alpha + \beta x_j\} = \frac{1}{1+\exp\{-(\alpha+\beta x_j)\}}.
\end{aligned} \tag{27}
$$

The pdf of this model is

$$
p(y_j|\boldsymbol{\xi}) = \eta_j(\boldsymbol{\xi})^{y_j}(1-\eta_j(\boldsymbol{\xi}))^{1-y_j} \tag{28}
$$

because it assumes bernoulli distribution. Using equations (A1, A2) and i.i.d assumption likewise the example of normal distribution, log likelihood can be represented

$$
\log p(\boldsymbol{y}|\boldsymbol{\xi}) = \sum_{j=1}^{J}\Big\{ y_j\eta_j(\boldsymbol{\xi}) + (1-y_j)\log(1-\eta_j(\boldsymbol{\xi}))\Big\} \tag{29}
$$

where $\boldsymbol{y} = (y_1, ..., y_J)$.

3.2.2 Score function

The Fisher information matrix of the logistic model is easy to derive. One reason is that the differential of the sigmoid function, like a logistic function, is easy to calculate. In general, the derivative of the sigmoid function is represented by

$$
\frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1-\sigma(x)) \tag{30}
$$

where $\sigma(x)$ is sigmoid function. For example, the derivative of the logistic function can be represented by

$$
\begin{aligned}
\frac{\partial}{\partial x} logit^{-1}(x) &= \frac{\partial}{\partial x}\frac{1}{1+e^{-x}} = -\frac{1}{\{1+e^{-x}\}^2}e^{-x}(-1) \qquad (\because \text{B6 etc.}) \\
&= \frac{1}{1+e^{-x}}\frac{e^{-x}}{1+e^{-x}} \\
&= logit^{-1}(x)(1-logit^{-1}(x)) \qquad (\because \frac{e^{-x}}{1+e^{-x}} = \frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}).
\end{aligned} \tag{31}
$$

In Bernoulli distribution case, we have to derive the derivative of choice probability to calculate the differential of log likelihood. In the same way with equation (31), the derivatives of probability $\eta_j(\boldsymbol{\xi})$ with respect to α and β are represented by

$$
\begin{aligned}
\frac{\partial}{\partial\alpha}\eta_j(\boldsymbol{\xi}) &= -\frac{\exp\{-\alpha-\beta x_j\}}{\Big(1+\exp\{-\alpha-\beta x_j\}\Big)^2}(-1) \qquad (\because \text{B6}) \\
&= \eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi})) \qquad (\because \text{Third line of 31}) \\
\frac{\partial}{\partial\beta}\eta_j(\boldsymbol{\xi}) &= -\frac{\exp\{-\alpha-\beta x_j\}}{\Big(1+\exp\{-\alpha-\beta x_j\}\Big)^2}(-x_j) \\
&= \eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))x_j.
\end{aligned} \tag{32}
$$

Using this equation, the derivatives of the log likelihood are represented by

$$
\begin{aligned}
\frac{\partial}{\partial\beta}\log p(\boldsymbol{y}|\boldsymbol{\xi}) &= \sum_{j=1}^{J}\Big\{ y_j \frac{1}{\eta_j(\boldsymbol{\xi})}\frac{\partial}{\partial\beta}\eta_j(\boldsymbol{\xi}) + (1-y_j)\frac{1}{1-\eta_j(\boldsymbol{\xi})}\frac{\partial}{\partial\beta}(1-\eta_j(\boldsymbol{\xi}))\Big\} \qquad (\because \text{B5}) \\
&= \sum_{j=1}^{J}\Big\{ y_j(1-\eta_j(\boldsymbol{\xi}))x_j - (1-y_j)\eta_j(\boldsymbol{\xi})x_j \Big\} \qquad (\because \text{equation 32}) \\
&= \sum_{j=1}^{J}\Big\{ y_j x_j - y_j\eta_j(\boldsymbol{\xi})x_j - \eta_j(\boldsymbol{\xi})x_j + y_j\eta_j(\boldsymbol{\xi})x_j \Big\} \\
&= \sum_{j=1}^{J}\Big\{ x_j(y_j-\eta_j(\boldsymbol{\xi})) \Big\} \\
\frac{\partial}{\partial\alpha}\log p(\boldsymbol{y}|\boldsymbol{\xi}) &= \cdots = \sum_{j=1}^{J}\Big\{ (y_j-\eta_j(\boldsymbol{\xi})) \Big\}.
\end{aligned} \tag{33}
$$

We abbreviate the derivation of the alpha because it is the same way as the beta case.

3.2.3 The Fisher information matrix

The expectation of bernoulli distribution is $E[y_j] = \eta_j(\boldsymbol{\xi})$, and the variance of that is $Var[y_j] = \eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))$. Looking at equation (33), the summation of the product of "observation - mean" appears likewise normal distribution example. Therefore, using the local independence assumption, we can use the formula of equation (25) to calculate the Fisher information matrix.

The (1, 1) element of the Fisher information matrix (i.e., the Fisher information of α) becomes

$$
\begin{aligned}
E\Big[\Big(\frac{\partial}{\partial\alpha}\log p(\boldsymbol{y}|\boldsymbol{\xi})\Big)^2\Big] &= E\Big[\Big(\sum_{j=1}^{J}(y_j-\eta_j(\boldsymbol{\xi}))\Big)^2\Big(\sum_{l=1}^{J}(y_l-\eta_l(\boldsymbol{\xi}))\Big)^2\Big] \qquad (\because \ (33)) \\
&= E\Big[\sum_{j=1}^{J}\Big(y_j-\eta_j(\boldsymbol{\xi})\Big)^2\Big] \qquad (\because \text{First line of (25)}) \\
&= \sum_{j=1}^{J}\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi})) \qquad (\because \text{(D2)}, Var[y_j] = \eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))).
\end{aligned} \tag{34}
$$

We can derive (1, 2) and (2, 2) elements in the same way because the difference in score function is only whether it includes x_j. The (1, 2) and (2, 2) elements are represented by

$$\begin{aligned} E\left[\Big(\frac{\partial}{\partial\alpha}\log p(\boldsymbol{y}|\boldsymbol{\xi})\Big)\Big(\frac{\partial}{\partial\beta}\log p(\boldsymbol{y}|\boldsymbol{\xi})\Big)\right] &= E\left[\sum_{j=1}^{J} x_j\Big(y_j-\eta_j(\boldsymbol{\xi})\Big)^2\right] \\ &= \sum_{j=1}^{J} x_j\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi})) \\ E\left[\Big(\frac{\partial}{\partial\beta}\log p(\boldsymbol{y}|\boldsymbol{\xi})\Big)^2\right] &= E\left[\sum_{j=1}^{J} x_j^2\Big(y_j-\eta_j(\boldsymbol{\xi})\Big)^2\right] \\ &= \sum_{j=1}^{J} x_j^2\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi})) \end{aligned} \tag{35}$$

respectively. In summary, The Fisher information matrix of the logistic model can be represented by

$$F(\boldsymbol{\xi}) = \sum_{j=1}^{J}\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))\begin{pmatrix}1 & x_j\\ x_j & x_j^2.\end{pmatrix} \tag{36}$$

Derivation of a model having multiple independent variables is the same way. For instance, suppose the model which has P independent variables, x_{jp}, and no slope. The (p, p') element of the Fisher information matrix for this logistic model is $\sum_{j=1}^{J}\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))x_{jp}x_{jp'}$.

3.3 Decision making model: Delay discounting model

The next example is the delay discounting (DD; Madden, Begotka, Raiff, & Kastern, 2003; Odum, 2011) model, which is one of the decision making models. People tend to prefer immediate but small rewards to delayed but large rewards. This is explained by discounting of delayed reward with some ratio (delay discounting rate). In experiments, researchers usually use the task which is consists alternative A (A^s) providing $\$x_j^s$ with time $t_j = 0$ (i.e., immediate reward alternative), and alternative B (A^d) providing $\$x_j^d$ with time $t_j = d_j$ (i.e., delayed reward alternative). The research object is often to determine the discounting function's shape or estimate the parameter determining discounting function (i.e., discounting rate).

3.3.1 Log likelihood function

Some decision making models can be represented by a kind of logistic regression model with elaborated structure in the exponential function of the logistic

function. It looks like a logistic regression, the exploratory variables of which are the difference in the alternative's value.

The DD model assumes

$$\begin{aligned} y_j &\sim Bernoulli(\eta_j(\boldsymbol{\xi})), \quad j = 1, ..., J \\ \eta_j(\boldsymbol{\xi}) &= logit^{-1}\{\beta(u(A^d) - u(A^s))\}, \quad \boldsymbol{\xi} = (k, \beta) \\ u(A^s) &= x_j^s \\ u(A^d) &= \frac{x_j^d}{1 + k * t_j^d} \end{aligned} \tag{37}$$

where $u(\cdot)$ is utility of alternative, k is discounting rate, and β is inverse temperature parameter. The first two lines represent the "general" two alternative choice model, which looks like a logistic regression, the exploratory variables of which are the difference of alternative's value. The calculation of the alternative's utility is different among decision making models. In the DD model, we assumes the immediate reward's utility is as it is, and delayed reward is discounted by hyperbolic function like a denominator. The reward x_j^s, x_j^d and time delay t_j^d are stimuli to be set by experimenter, and $\boldsymbol{\xi} = (k, \beta)$ is parameter to be estimated.

The pdf of DD model (decision making model) is

$$p(y_j|\boldsymbol{\xi}) = \eta_j(\boldsymbol{\xi})^{y_j}(1 - \eta_j(\boldsymbol{\xi}))^{1-y_j} \tag{38}$$

in the same way as before. Therefore, the log likelihood function is represented by

$$\log p(\boldsymbol{y}|\boldsymbol{\xi}) = \sum_{j=1}^{J}\{y_j \log \eta_j(\boldsymbol{\xi}) + (1 - y_j)\log(1 - \eta_j(\boldsymbol{\xi}))\}. \tag{39}$$

3.3.2 Score function

The derivation of the score function is the same as before. That is, we have to derive the differential of the log likelihood function after derivation of the differential of choice probability $\eta_j(\boldsymbol{\xi})$. The differential of $\eta_j(\boldsymbol{\xi})$ can be represented by

$$\begin{aligned} \frac{\partial}{\partial k}\eta_j(\boldsymbol{\xi}) &= -\frac{1}{\Big(1 + \exp\{-\beta(u(A^d) - u(A^s))\}\Big)^2} \\ &\qquad \exp\{-\beta(u(A^d) - u(A^s))\}\Big[-\beta\Big(\frac{\partial}{\partial k}u(A^d)\Big)\Big] \qquad (\because (B6)) \\ &= \eta_j(\boldsymbol{\xi})(1 - \eta_j(\boldsymbol{\xi}))\Big[\beta\Big(\frac{\partial}{\partial k}u(A^d)\Big)\Big] \qquad (\because \text{Third line of (31)}) \\ \frac{\partial}{\partial \beta}\eta_j(\boldsymbol{\xi}) &= \eta_j(\boldsymbol{\xi})(1 - \eta_j(\boldsymbol{\xi}))\Big[u(A^d) - u(A^s)\Big] \end{aligned} \tag{40}$$

in the same way as equations (30) to (32). We abbreviate the derivation about β because it is the same as the k case. Because most decision making models just add elaborated structure in the exponential function, this has the structure of "choice probaility$\times$(1$-$choice probability)$\times$the differential of argument of exponential function".

The score function can be represented by

$$\begin{aligned}
&\frac{\partial}{\partial k}\log p(\boldsymbol{y}|\boldsymbol{\xi}) \\
&= \sum_{j=1}^{J}\Big\{ y_j(1-\eta_j(\boldsymbol{\xi}))\Big[\beta\frac{\partial}{\partial k}u(A^d)\Big] - (1-y_j)\eta_j(\boldsymbol{\xi})\Big[\beta\frac{\partial}{\partial k}u(A^d)\Big]\Big\} \qquad (\because (40)) \\
&= \sum_{j=1}^{J}\Big\{ \Big[\beta\frac{\partial}{\partial k}u(A^d)\Big](y_j-\eta_j(\boldsymbol{\xi}))\Big\} \\
\frac{\partial}{\partial \beta}\log p(\boldsymbol{y}|\boldsymbol{\xi}) &= \sum_{j=1}^{J}\Big\{ \Big[u(A^d)-u(A^s)\Big](y_j-\eta_j(\boldsymbol{\xi}))\Big\}
\end{aligned} \tag{41}$$

in the same way as equation (33). This has also a structure of "observation - mean" as well as the logistic regression model example.

3.3.3 The Fisher information matrix

Using the local independence assumption, we can use the formula of equation (25) in the same way as before. The (1, 1) element of the Fisher information matrix (i.e., the Fisher information of k) can be derived as

$$\begin{aligned}
E\left[\left(\frac{\partial}{\partial k}\log p(\boldsymbol{y}|\boldsymbol{\xi})\right)^2\right] &= E\left[\sum_{j=1}^{J}\Big[\beta\frac{\partial}{\partial k}u(A^d)\Big]^2(y_j-\eta_j(\boldsymbol{\xi}))^2\right] \\
&= \sum_{j=1}^{J}\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))\Big[\beta\frac{\partial}{\partial k}u(A^d)\Big]^2
\end{aligned} \tag{42}$$

in the same way as equation (34). Further, (1, 2) and (2, 2) elements of that can be derived as

$$\begin{aligned}
E\left[\left(\frac{\partial}{\partial k}\log p(\boldsymbol{y}|\boldsymbol{\xi})\right)\left(\frac{\partial}{\partial \beta}\log p(\boldsymbol{y}|\boldsymbol{\xi})\right)\right] &= \sum_{j=1}^{J}\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))\Big[\beta\frac{\partial}{\partial k}u(A^d)\Big]\Big[u(A^d)-u(A^s)\Big] \\
E\left[\left(\frac{\partial}{\partial \beta}\log p(\boldsymbol{y}|\boldsymbol{\xi})\right)^2\right] &= \sum_{j=1}^{J}\eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))\Big[u(A^d)-u(A^s)\Big]^2
\end{aligned} \tag{43}$$

respectively.

Finally, we have to calculate $\frac{\partial}{\partial k}u(A^d)$. Using model assumption (equation (37)) and equation (B6), $\frac{\partial}{\partial k}u(A^d)$ can be derived as

$$\frac{\partial}{\partial k}u(A^d) = -x_j^d\frac{1}{\{1+kt_j^d\}^2}t_j^d. \tag{44}$$

We can calculate the Fisher information using above equations.

3.4 General formula for two-alternative decision making models

As we described above, some decision making models look like the logistic regression, the exploratory variables of which are the difference of alternative's values after evaluating each alternative's value separately. In this class, models have the same structure as the Fisher information matrix. Therefore, we derive the "general" formula of the Fisher information matrix for two-alternative decision making models in this section.

3.4.1 Model

Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_P)$ be P parameters. Suppose the 0-1 data $\boldsymbol{y} = (y_1, ..., y_J), j = 1, ..., J$ is generaged by

$$\begin{aligned} y_j &\sim Bernoulli(\eta_j(\boldsymbol{\xi})), \quad j = 1, ..., J \\ \eta_j(\boldsymbol{\xi}) &= logit^{-1}\{g(A_j; \boldsymbol{\xi}) - g(B_j; \boldsymbol{\xi})\} \\ &= \frac{1}{1 + \exp\left\{-\{g(A_j; \boldsymbol{\xi}) - g(B_j; \boldsymbol{\xi})\}\right\}} \end{aligned} \tag{45}$$

where $g(\cdot)$ is function for calculating alternative's value. For now, we do not need to assume a specific function shape to derive the formula. The $g(A_j; \boldsymbol{\xi}), g(B_j; \boldsymbol{\xi})$ are the value of A and value of B, respectively. The $g(\cdot)$ depends on parameters and stimuli (and obviously, task structure and models). Let $D(\boldsymbol{\xi}) = g(A_j; \boldsymbol{\xi}) - g(B_j; \boldsymbol{\xi})$ be for simple notation.

Delay discounting (DD) model For example, DD model assumes $g(A_j; \boldsymbol{\xi}) = \beta u(A^s)$, $g(B_j; \boldsymbol{\xi}) = \beta u(A^d)$. The definitions of each symbol are defined by equation (37).

Expected utility theory (EUT) model Suppose that alternative A_j provide $\$x_{jA}$ with probability p_{jA} and $\$y_{jA}$ with probability $1 - p_{jA}$. The alternative B_j is defined by the same way. The EUT assumes $g(A_j; \boldsymbol{\xi}) = \beta * V(A_j) = \beta * \{v(x_{jA})p_{jA} + v(y_{jA})(1 - p_{jA})\}$. $V(A_j)$ is total value of alternative A. $v(x_{jA})$ is utility of reward $\$x_{jA}$, which is often defined by $v(x_{jA}) = x_{jA}^{\alpha}$, where α is a parameter.

Cumulative prospect theory (CPT) model CPT (Tversky & Kahneman, 1992) assumes $V(A_j) = v(x_{jA})w(p_{jA}) + v(y_{jA})(1 - w(p_{jA})) \quad if \quad x_{jA} > y_{jA}$. The function $w(\cdot)$ is the probability weighting function. One of the most famous definition is $w(p) = \frac{p^\gamma}{\{p^\gamma + (1-p)^\gamma\}^{1/\gamma}}$ (Tversky & Kahneman, 1992).

Furthermore, the logistic regression can be represented with $D(\boldsymbol{\xi}) = \alpha + \beta \boldsymbol{x_j}$, which means the logistic model is one of the special case of above general model.

3.4.2 Log likelihood function

As we will describe hereafter, the derivation procedure is almost the same as before. The pdf can be represented by

$$p(y_j|\boldsymbol{\xi}) = \eta_j(\boldsymbol{\xi})^{y_j}(1-\eta_j(\boldsymbol{\xi}))^{1-y_j} \tag{46}$$

because we assumes bernoulli distribution. Therefore, the log likelihood becomes

$$\log p(\boldsymbol{y}|\boldsymbol{\xi}) = \sum_{j=1}^{J}\{y_j \log \eta_j(\boldsymbol{\xi}) + (1-y_j)\log(1-\eta_j(\boldsymbol{\xi}))\} \tag{47}$$

in the same way as before.

3.4.3 Score function

The differential of choice probability $\eta_j(\boldsymbol{\xi})$ with respect to m-th parameter ξ_m can be derived as

$$\begin{aligned}\frac{\partial}{\partial \xi_m}\eta_j(\boldsymbol{\xi}) &= -\frac{1}{\Big\{1+\exp(-D(\boldsymbol{\xi}))\Big\}^2}\exp(-D(\boldsymbol{\xi}))\left[-\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right] \\ &= \eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))\left[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right]\end{aligned} \tag{48}$$

in the same way as equation (31). Note that the differential of the difference of each alternative, $\left[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right] = \left[\frac{\partial}{\partial \xi_m}\Big(g(A_j;\boldsymbol{\xi}) - g(B_j;\boldsymbol{\xi})\Big)\right] = \left[\frac{\partial}{\partial \xi_m}g(A_j;\boldsymbol{\xi}) - \frac{\partial}{\partial \xi_m}g(B_j;\boldsymbol{\xi})\right]$, has to be calculated. This part is different among models. Therefore, we derive the Fisher information matrix's general formula without derive a specific equation for now.

The differential of the log likelihood function can be represented by

$$\begin{aligned}\frac{\partial}{\partial \xi_m}\log p(\boldsymbol{y}|\boldsymbol{\xi}) &= \sum_{j=1}^{J}\Big\{y_j(1-\eta_j(\boldsymbol{\xi}))\left[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right] - (1-y_j)\eta_j(\boldsymbol{\xi})\left[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right]\Big\} \\ &= \sum_{j=1}^{J}\Big\{\Big[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\Big](y_j - \eta_j(\boldsymbol{\xi}))\Big\}\end{aligned} \tag{49}$$

in the same way as before.

3.4.4 The Fisher information matrix

Using equation (25), the Fisher information matrix can be derived in the same way as before. The $(m, n), m, n = 1, ..., P$, element of the Fisher information matrix, $F(\boldsymbol{\xi})_{m,n}$ can be derived as

$$\begin{aligned} F(\boldsymbol{\xi})_{m,n} &= E\left[\left(\frac{\partial}{\partial \xi_m}\log p(\boldsymbol{y}|\boldsymbol{\xi})\right)\left(\frac{\partial}{\partial \xi_n}\log p(\boldsymbol{y}|\boldsymbol{\xi})\right)\right] \qquad (\because \text{def}) \\ &= \sum_{j=1}^{J} E\left[(y_j - \eta_j(\boldsymbol{\xi}))^2\right]\left[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right]\left[\frac{\partial}{\partial \xi_n}D(\boldsymbol{\xi})\right] \qquad (\because (25)) \\ &= \sum_{j=1}^{J} \eta_j(\boldsymbol{\xi})(1-\eta_j(\boldsymbol{\xi}))\left[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right]\left[\frac{\partial}{\partial \xi_n}D(\boldsymbol{\xi})\right]. \end{aligned} \tag{50}$$

This is the Fisher information matrix of general decision making models represented by equation (45).

Finally, we have to derive $\left[\frac{\partial}{\partial \xi_m}D(\boldsymbol{\xi})\right]$. This is a different part among models. Although choice probability is dependent on models, it is calculated with the equation (45) directly. Let us check some examples.

In logistic regression model of above example, parameters are $\boldsymbol{\xi} = (\alpha, \beta)$. Because the differential is $\frac{\partial}{\partial \alpha}D(\boldsymbol{\xi}) = 1$, $\frac{\partial}{\partial \beta}D(\boldsymbol{\xi}) = x_i$, equation (36) is the same as the equation (50). In DD model, the parameters are $\boldsymbol{\xi} = (k, \beta)$ and $D(\boldsymbol{\xi}) = \beta\{u(A^d) - u(A^s)\}$. Because the differential is $\frac{\partial}{\partial k}D(\boldsymbol{\xi}) = \beta\left[\frac{\partial}{\partial k}u(A^d)\right]$, $\frac{\partial}{\partial \beta}D(\boldsymbol{\xi}) = u(A^d) - u(A^s)$, equations (42) and (43) are the same as the equation (50). In CPT model, Fujita and Okada (preprint) derived the Fisher information matrix of CPT model added another parameter. Although it is not easy compared to above models, they derived $\frac{\partial}{\partial \boldsymbol{\xi}}D(\boldsymbol{\xi})$ in the same way.

4 The example code

This section provides R code example for Delay Discounting (DD) model. This will help you re-learn how to calculate the Fisher information and use the Fisher information for designing experimental stimuli.

4.1 The functions for calculating the Fisher information

4.1.1 The basic function

Let us define the functions for calculating the alternative's value and choice probability for j-th trial. Let Para be (true) parameter values which is (k, β) vector. Let Stimulus_j be stimuli set of j-th trial which is 2 (alternative A, B) $\times$ 2 (stimulus, reward and delayed time) matrix. Stimulus_j[1, 2] is always 0 because alternative A is the immediate reward alternative. CalcValue_j is the function for calculating the alternative's value, and CalcAllValue_j is the function for

calculating all alternative's values. CalcProb_j is the function for calculating choice probability. These functions are just representing model assumptions.

Listing 1: Basic function

```
#####
# function for calculating alternative values
#####
CalcValue_j <- function(Para, Stimulus_j, cate) {
  Value <- Stimulus_j[cate, 1] / (1 + Para[1] * Stimulus_j[cate, 2])
      # hyperbolic model
  return(Value)
}
CalcAllValue_j <- function(Para, Stimulus_j) {
  Value <- numeric(length = 2)
  for(cate in 1:2) {
    Value[cate] <- CalcValue_j(Para = Para, Stimulus_j = Stimulus_j,
        cate=cate)
  }
  return(Value)
}
#####
# function for calculating choice probability
#####
CalcProb_j <- function(Para, Stimulus_j) {
  Value <- CalcAllValue_j(Para = Para,Stimulus_j = Stimulus_j)
  prob <- logistic( Para[2] * (Value[2] - Value[1]) )
}
```

4.1.2 The function for calcula ting the Fisher information

Let us define the function for calculating the Fisher information. PartValue is the function for calculating the derivatives of the difference of each alternative value. The first (second) element is derivatives with respect to $\alpha(\beta)$. As we derive them in section 3.4, we can use $\frac{\partial}{\partial k}D(\boldsymbol{\xi}) = \beta\Big[\frac{\partial}{\partial k}u(A^d)\Big]$, $\frac{\partial}{\partial \beta}D(\boldsymbol{\xi}) = u(A^d) - u(A^s)$.

Listing 2: Second derivatives

```
PartValue <- list()
PartValue[[1]] <- function(Para, Stimulus_j) {
  res <- - Stimulus_j[2, 1] * ( Stimulus_j[2, 2] / (1 + Para[1] *
      Stimulus_j[2, 2])^2 )
  res <- res * Para[2]
  return(res)
}
PartValue[[2]] <- function(Para, Stimulus_j) {
  Value <- CalcAllValue_j(Para = Para,Stimulus_j = Stimulus_j)
  res <- Value[2] - Value[1]
  return(res)
}
```

Let us define the CalcFisher function for calculating the Fisher information matrix of j-th trial. In addition, the TestIF function calculates the Fisher information matrix of all trials. The variable Stimulus is all trial's stimuli which is the number of trials $\times$ 2(alternative) $\times$ 2(reward, delayed time) array, namely, it is the collection of Stimulus_j. Note that the Fisher information matrix of all trials is a simple summation of the Fisher information matrix of a single trial. When other model is used, you can use this code example by modifying the CalcValue_j function and the PartValue function corresponding to a new model.

Listing 3: The Fisher information

```
CalcFisher <- function(Para, Stimulus_j) {
  PNum <- length(Para) # the number of parameters
  Mat <- matrix(NA, ncol = PNum, nrow = PNum) # Fihser information
      matrix
  probj <- CalcProb_j(Para = Para,Stimulus_j = Stimulus_j) # choice
      probability
  for(irow in 1:PNum) {
    for(jcol in 1:PNum) {
      Mat[irow, jcol] <- probj * (1 - probj) *
                         PartValue[[irow]](Para=Para, Stimulus_j =
                             Stimulus_j) *
                         PartValue[[jcol]](Para = Para,Stimulus_j =
                             Stimulus_j)
    }
  }
  return(Mat)
}

TestIF <- function(Para, Stimulus) {
  J <- dim(Stimulus)[1] # the number of trials
  PNum <- length(Para)
  Mat <- matrix(data = 0, nrow = PNum, ncol = PNum)
  for(t in 1:J) {
    Mat <- Mat + CalcFisher(Para = Para,Stimulus_j = Stimulus[t, , ])
  }
  return(Mat)
}
```

4.2 The simulation for evaluating the Fisher information of each stimuli pattern

Using above functions, researchers can calculate the Fisher information matrix with specific (true) parameter values and stimuli (i.e., Stimulus_j, Stimulus). As we explained before, the calculation of the Fisher information matrix in code is not needed a dependent variable (choice data). Therefore, researchers can quickly evaluate the goodness of stimuli and design before conducting experiments. Note that the calculation is dependent on assumed parameter distribution and values. Researchers can determine this distribution with a vague distribution like uniform distribution covering almost parameter values, their

own experience, the theoretical background of the model, or posterior distribution obtained from previous studies. In addition, it is important to check that selected stimuli are good enough for specific parameter values, as we elaborate on later.

In the below simulation, the parameter is generated by distribution we assumed multiple times (50 times in this simulation), and the expectation of the Fisher information over these parameters is calculated. In the below code, for the later explanation, we assumed delta distribution for k. However, researchers usually better to assume a more dispersed distribution (e.g., normal, uniform, beta distribution, etc.). As for candidate stimuli which will be evaluated in the Fisher information, the number of reward candidate values (rewardCand) is 20, and the number of delayed time values (timeCand) is 10. Therefore, we will evaluate the goodness of stimuli for parameter estimation at 4000 ($20^2 \times 10$) stimuli points.

Listing 4: Stimuli analysis by the Fisher information

```
#--- settings
Iter <- 50 # the number of calculation repetition

#--- parameter values
k <- rep(0.2, Iter) # delta distiribution
beta <- runif(n = Iter, min = 0.5, max = 3)
Para <- cbind(k, beta)

#--- stimulus candidate
rewardCand <- seq(1, 20, by=1); NumRew <- length(rewardCand) # reward
timeCand <- seq(1, 10, by=1); NumTime <- length(timeCand) # time delay

#--- box
ResFisherMat_iter <- array(NA, dim = c(Iter, NumRew, NumRew, NumTime,
    2, 2))
ResFisherMat <- array(0, dim = c(NumRew, NumRew, NumTime, 2, 2)) # the
     Fisher information matrix

for(iRewA in 1:NumRew) {
  for(jRewB in 1:NumRew) {
    for(ktimeB in 1:NumTime) {
      # set stimuli set
      StimTemp <- array(NA, dim = c(2, 2))
      StimTemp[1, 1] <- rewardCand[iRewA]; StimTemp[1, 2] <- 0
      StimTemp[2, 1] <- rewardCand[jRewB]; StimTemp[2, 2] <- timeCand[
          ktimeB]
      for(iter in 1:Iter) {
        # calculate the Fisher information matrix
        ResFisherMat_iter[iter, iRewA, jRewB, ktimeB, , ] <- CalcFisher
            (Para = Para[iter, ],Stimulus_j = StimTemp)
        ResFisherMat[iRewA, jRewB, ktimeB, , ] <- ResFisherMat[iRewA,
            jRewB, ktimeB, , ] + ResFisherMat_iter[iter, iRewA, jRewB,
             ktimeB, , ]
      } # iter loop
    } # time of B loop
  } # reward of B loop
```

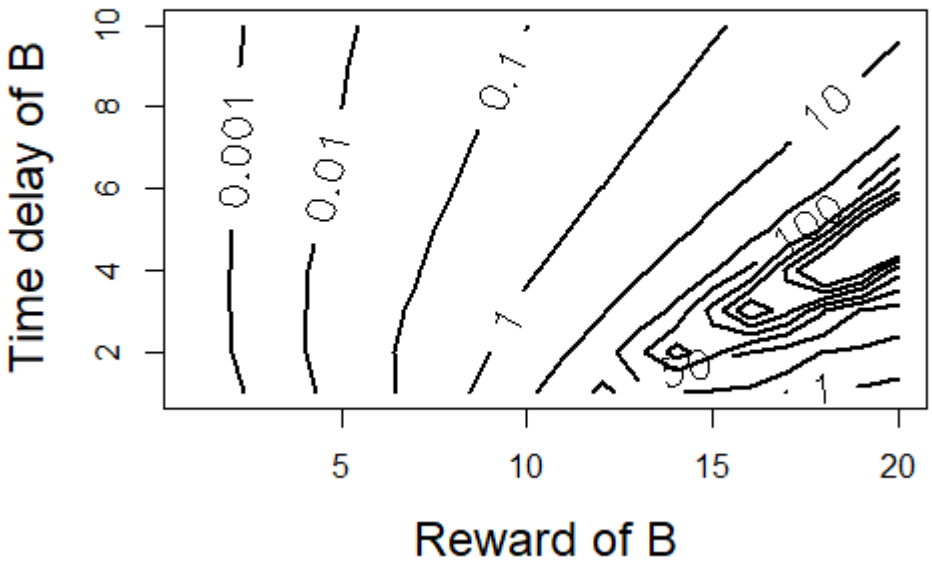

Figure 1: The Fisher information of k for each stimulus. The x-axis represents reward of B, and the y-axis represents delayed time of B. Higher fisher information means better stimuli pattern.

```
31 } # reward of A loop
32 ResFisherMat <- ResFisherMat / Iter # mean
```

The (expectation of) Fisher information matrix of each stimuli is saved in the ResFisherMat variable. Let us focus on the Fisher information for k. Figure 1 represents the Fisher information for k of each stimuli pattern. In this figure, we fixed the reward value of A as 10 because the 3-d plot is difficult to figure out.

These results are accordant with our intuition. When the reward value of B is smaller than the reward value of A, this trial will not contribute to estimation because the reward of B is discounted, which means most participants will choose alternative A regardless of their trait parameters. Further, when the reward value is too large (plus small time delay) or the time delay is too large, the trial will not contribute to estimation because most participants will choose one alternative regardless of their trait parameters. In contrast, some stimuli patterns provide over 100 Fisher information. In short, there are "sweet" spots that will provide higher estimation precision (higher Fisher information) with

reflecting individual differences well. The merit of using the Fisher information is that researchers can detect these sweet spots with well-organized and objective calculations.

This result is based on 2-d plot. Therefore, researchers may be able to detect good stimuli patterns without the above calculation if the reward value of A is fixed. However, how about if the reward value of A is not determined? In addition, there are more elaborated tasks and models that may have more dimensions of stimulus, which will make designing stimuli be more difficult. Further, whether a specific stimuli pattern is good for parameter estimation dependents on the participant's trait parameter values. Using the Fisher information, researchers can evaluate the goodness of stimuli with any dimensions and parameter values in the same framework.

4.3 Application to fixed design: selecting all trial's stimuli

The previous simulation is for evaluating stimuli of a single trial. In psychological experiments, researchers usually estimate parameters with multiple trials. Therefore, we have to select multiple trial stimuli. Fortunately, each trial of most tasks does not influence each other, at least theoretically. Therefore, we can use a greedy way, which selects optimal stimuli in each trial. In this example, to generate parameter values from assumed distribution and select optimal design based on the Fisher information at above parameter values is used to determine fixed design. We selected optimal stimuli for the 20 trials in the below simulation. We have calculated the Fisher information matrix at specific parameter values, which is saved on ResFisherMat_iter, therefore we reuse it.

Listing 5: Selecting optimal design

```
Trial <- 20 # the number of trials
Stimulus <- array(NA, dim = c(Trial, 2, 2)) # Trial * alternative 2 *
     stimuli 2 (reward, time delay)

iter_sa <- sample(1:Iter, size = Trial, replace = FALSE)
for(trial in 1:Trial) {
  #--- calculate determinant ---
  ResDetTemp <- array(NA, dim = c(NumRew, NumRew, NumTime))
  for(iRewA in 1:NumRew) {
    for(jRewB in 1:NumRew) {
      for(ktimeB in 1:NumTime) {
        ResDetTemp[iRewA, jRewB, ktimeB] <- det( ResFisherMat_iter[
            iter_sa[trial], iRewA, jRewB, ktimeB, , ] )
      }
    }
  } # stimuli loop

  #--- maximizing stimuli ---
  ind <- which( ResDetTemp == max(ResDetTemp), arr.ind = TRUE ) #
      index of stimuli maximizing det(F)
  # set stimuli
  Stimulus[trial, 1, 1] = rewardCand[ind[1, 1]]; Stimulus[trial, 1,
      2] <- 0 # stimuli of A
```

```
  Stimulus[trial, 2, 1] = rewardCand[ind[1, 2]]; Stimulus[trial, 2,
      2] <- timeCand[ind[1, 3]] # stimuli of B
} # trial loop
```

What we did in the above code is just to calculate the determinant of Fisher information matrix and find out stimuli maximizing this value. We iterated 20 times (i.e., the number of trials). Although we have selected optimal stimuli already, we better to check some characteristics of the selected stimuli. One checkpoint is that researchers better to run MCMC based simulation to check estimation precision more precisely because the Fisher information is asymptotic prediction. Note that MCMC based simulation will take much time, therefore it should be done after fixing (optimal) design. Second, researchers better to check predicted estimation precision for each parameter value.

In the simulation below, we confirm how much estimation precision this selected stimuli provide for participants with specific parameter values. Namely, the code below changes (true) parameter values while fixing the stimuli.

Listing 6: Asymptotic SD of each parameter

```
#--- parameter values ---
kseq <- seq(0.01, 0.99, length.out = 20)
betaseq <- seq(0.1, 5, length.out = 20)

#--- box
ResFisher_para <- ResFisher_inv <- array(NA, dim = c(length(kseq),
    length(betaseq), 2, 2))

for(ik in 1:length(kseq)) {
  for(jbeta in 1:length(betaseq)) {
    ResFisher_para[ik, jbeta, , ] <- TestIF(Para = c(kseq[ik], betaseq
        [jbeta]), Stimulus = Stimulus)
    ResFisher_inv[ik, jbeta, , ] <- solve( ResFisher_para[ik, jbeta, ,
         ] ) # inverse matrix of the Fisher information matrix
  }
}
```

Let us focus on the parameter k. Figure 2 represents the predicted posterior standard deviation (PSD) for each parameter value. From this figure, we can infer that parameter k around 0.2 will be estimated with high estimation precision, however, parameter k above 0.6 will not be. This is because we selected stimuli assuming the true parameter k value is 0.2, which means selected stimuli are for this values. It is unrealistic in real experiments to know true parameter values as a delta distribution. This setting was for an explanation. Therefore, researchers can assume some distribution for parameter values as a vague distribution like a uniform distribution, distribution based on theoretical background or researcher's experience, or posterior distribution obtained from previous experiments. In this case, if you change from "k <- rep(0.2, Iter)" to "k <-(some distribution)", the results will be changed.

In general, if researchers set vague distribution, they can estimate most parameters well. However, vague distribution means it includs ineffective stimuli

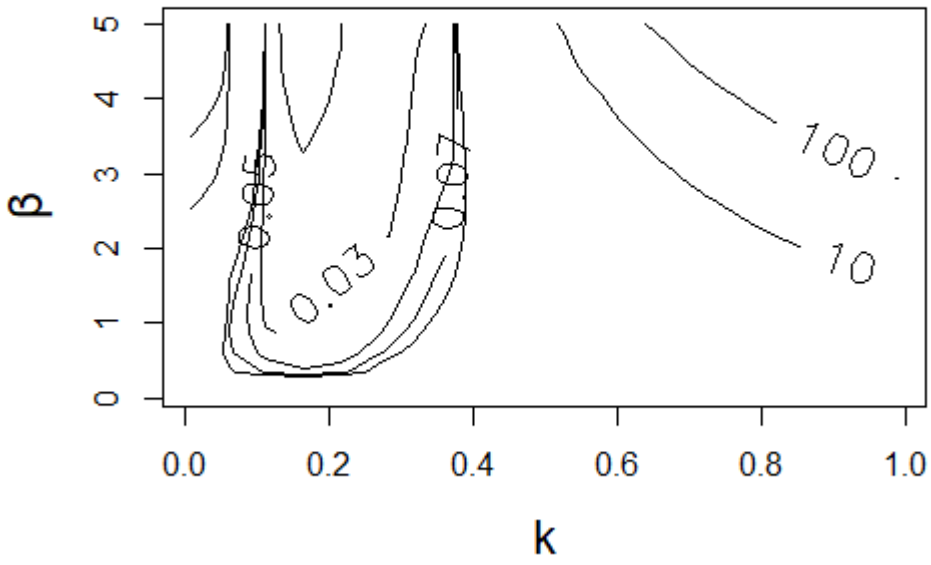

Figure 2: The aysmptotic PSD of each parameter value. The x-axis represents parameter k, and the y-axis represents parameter β.

for specific participants. For example, in the above example, stimuli for parameter k around 0.2 is not good for parameter k above 0.6. Therefore, it is important to check the above figure. In other words, researchers better to check whether stimuli will provide enough estimation precision for each parameter value. Further, in the CAT framework, researchers can select stimuli based on the Fisher information after obtaining response data (i.e., temporal parameter estimates). Therefore, it will be more efficient than a fixed design.

In summary, the Fisher information matrix is important and fundamental quantity from some statistical aspects. This paper focused on the fact that the Fisher information can predict variance (covariance matrix) of ML estimator and stimulus selection method. Maximizing the (determinant of) Fisher information matrix means maximizing estimation precision. This paper provided some examples calculating of the Fisher information matrix from the simple regression models to more elaborated decision making models. In addition, this paper provides example code to explain how to use the Fisher information as a stimulus selection criteria.

References

Ahn, W. Y., Gu, H., Shen, Y., Haines, N., Hahn, H. A., Teater, J. E., ... Pitt, M. A. (2020). Rapid , precise , and reliable measurement of delay discounting using a bayesian learning algorithm. *Scientific Reports*, 1-10. Retrieved from `https://doi.org/10.1038/s41598-020-68587-x` doi: 10.1038/s41598-020-68587-x

Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., & Myung, J. I. (2013). Discriminating among probability weighting functions using adaptive design optimization. *Journal of Risk and Uncertainty*, *47*, 255-289. doi: 10.1007/s11166-013-9179-3

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*, 273-304. doi: 10.1214/ss/1177009939

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*, 1-20. doi: 10.1007/s11336-014-9401-5

DiMattina, C. (2015). Fast adaptive estimation of multidimensional psychometric functions. *Journal of Vision*, *15*, 1-20. doi: 10.1167/15.9.5

Fujita, K., & Okada, K. (preprint). Adaptive optimal stimulus selection in cognitive models using a model averaging approach. *preprint*.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of the cat-anx: A computerized adaptive test for depression. *Arch Gen Psychiatry*, *69*, 1104-1112. doi: 10.1001/archgenpsychiatry.2012.14

Heck, D. W., & Erdfelder, E. (2019). Maximizing the expected information gain of cognitive modeling via design optimization. *Computational Brain and Behavior*, *2*, 202-209. doi: 10.1007/s42113-019-00035-0

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *1007*, 453-461.

Jeffreys, H. (1961). *Theory of probability. (third ed).* Oxford, UK: Oxford University Press.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*, 2729-2737. doi: 10.1016/S0042-6989(98)00285-5

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*, 986-1005. doi: 10.1214/aoms/1177728069

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E. J. (2017). A tutorial on fisher information. *Journal of Mathematical Psychology*, *80*, 40-55. Retrieved from `http://dx.doi.org/10.1016/j.jmp.2017.05.006` doi: 10.1016/j.jmp.2017.05.006

Madden, G. J., Begotka, A. M., Raiff, B. R., & Kastern, L. L. (2003). Delay discounting of real and hypothetical rewards. *Experimental and Clinical Psychopharmacology*, *11*, 139-145. doi: 10.1037/1064-1297.11.2.139

McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808-821. doi: 10.3758/BRM.40.3.808

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and an example. *Applied Psychological Measurement*, *23*, 187-194. doi: https://doi.org/10.1177/01466219922031310

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273-296. doi: 10.1007/s11336-008-9097-5

Myung, J. I. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90-100. doi: 10.1016/S0022-2496(02)00028-7

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167-179. doi: 10.1016/j.jmp.2005.06.008

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499-518. doi: 10.1037/a0016104

Odum, A. L. (2011). Delay discounting: Trait variable? *Behavioural Processes*, *87*, 1-9. doi: 10.1016/j.beproc.2011.02.007

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331-354. doi: 10.1007/BF02294343

Segall, D. O. (2004). Computerized adaptive testing. *Encyclopedia of Social Measurement*, 429-438. doi: 10.1016/B0-12-369398-5/00444-8

Toubia, O., Johnson, E., Evgeniou, T., & Delquié, P. (2013). Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Science*, *59*, 613-640. Retrieved from

http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1120.1570 doi: 10.1287/mnsc.1120.1570

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297-323.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201-216. doi: 10.1007/BF02294775

van der Linden, W. J. (2018). Adaptive testing. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume three: application* (p. 197-228). Boca Raton, FL: CRC Press.

Watson, A. B., & Pelli, D. G. (1983). Quest: A bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*, 113-120. doi: https://doi.org/10.3758/BF03202828

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361-375. doi: 10.1111/j.1745-3984.1984.tb01040.x