Optimal cutpoints for random observations

Marius Schmidt Rainer Schwabe

April 9, 2014

Abstract

For the discretisation of a continuous random variable into different categories the choice of cutpoints is necessary. A popular application is the contingent valuation method. As the choice of cutpoints directly effects the quality of the parameter estimates optimal cutpoints are desirable in order to obtain the most accurate estimates. We consider an arbitrary number of cutpoints and determine optimal cutpoints for the exponential and Gumbel distribution and prove that the c-optimal cutpoints for the location parameter of the logistic distribution have corresponding equal category probabilities. Furthermore, we show that in the limiting case for infinitely many cutpoints there is no loss of information.

Keywords: c-optimality, cutpoints, exponential distribution, Gumbel distribution, logistic distribution, multinomial distribution

1 Introduction

In an experiment it can be of interest to partition a continuous random variable into different categories. This requires the choice of cutpoints, where k cutpoints lead to $k+1$ categories. Such a discretisation of a continuous random variable is useful if the random variable is not directly observable.

M. Schmidt (✉) · R. Schwabe
Institut für Mathematische Stochastik
Otto-von-Guericke-Universität Magdeburg
PF 4120, 39016 Magdeburg, Germany
E-Mail: marius.schmidt@ovgu.de
E-Mail: rainer.schwabe@ovgu.de

There are various areas of application, for example in medicine, engineering and economics. For determining the size of approximately spherical grains of granular material several sieves with different hole diameters can be used. The grain size is assumed to follow a continuous probability distribution. Then a good choice for the hole diameters of the sieves are the optimal cutpoints.
To determine the economical value of non-market goods the contingent valuation method is used, where people are asked how much they are willing to pay for a new good. To avoid biases no open-ended questions should be asked. The decision of the interviewed people should be based on prespecified values. In the case of two categories the people are asked if they are willing to pay a certain value for the new good. In order to achieve more precise results the number of cutpoints can be increased.

The quality of the parameter estimates depends on the choice of cutpoints. Thus optimal cutpoints allow for most accurate estimation of the parameters. Bickel et al. (2001) showed that in the case of two categories the median is the optimal cutpoint for many symmetric distributions which are location families. Nguyen and Torsney (2007) determined numerically optimal cutpoints for the logistic distribution. Gunduz and Torsney (2006) and Rabie and Flournoy (2013) computed optimal designs in the context of contingent valuation and contingent response studies.
We consider the exponential, Gumbel and logistic distribution and determine analytically optimal cutpoints for an arbitrary number of categories. Moreover, we show that the Fisher information for the discretised random variable converges to the Fisher information of the continuous random variable for an appropriate choice of cutpoints. In Sections 2 to 4 we introduce some theory of cutpoints and optimality criteria. Sections 5 and 6 contain the results concerning the optimal cutpoints for the exponential and logistic distribution. Extensive proofs are deferred to an appendix.

2 Model description

Let $X : \Omega \to X(\Omega) = \mathscr{X} \subseteq \mathbb{R}$ be a continuous random variable with cumulative distribution function (CDF) F_X and probability distribution function (pdf) f_X, both depending on a parameter vector $\boldsymbol{\theta}$. We note that the support $\mathscr{X}$ of X is independent of the value $\boldsymbol{\theta}$ of the parameter vector.
The continuous random variable is partitioned into $k + 1$ categories by the choice of k cutpoints $x_1 < \ldots < x_k \in \mathscr{X}$. To simplify notation we introduce the cutpoint vector $\boldsymbol{x} = (x_1, \ldots, x_k)$ and denote by Ξ_k the set of all k-dimensional cutpoint vectors. For $i = 1, \ldots, k + 1$, category i contains all $x \in \mathscr{X}$ with $x_{i-1} < x \leq x_i$, where $x_0 := \inf(\mathscr{X})$ and $x_{k+1} := \sup(\mathscr{X})$. In Figure 2.1 this is schematically illustrated.
We note that $\mathscr{X}$ need not to be bounded. For $\mathscr{X} = \mathbb{R}$ we have $x_0 = -\infty$ and $x_{k+1} = \infty$. The discretisation of the continuous random variable X is modelled mathematically by

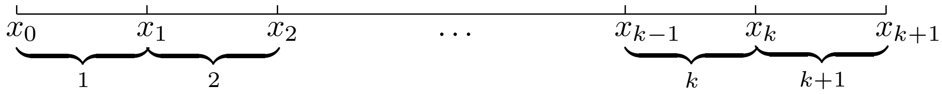

Figure 2.1: Partitioning of the domain $\mathscr{X}$ into $k+1$ categories

the multinomial distribution. Let $N, k \in \mathbb{N}$ and $p_1, \ldots, p_{k+1} \in [0,1]$ with $\sum_{i=1}^{k+1} p_i = 1$. Then the multinomially distributed random variable $\boldsymbol{Y} \sim M\big(N, (p_1, \ldots, p_{k+1})\big)$ has the discrete probability function

$$f_{\boldsymbol{Y}}(n_1, \ldots, n_{k+1}) = \binom{N}{n_1, \ldots, n_{k+1}} \cdot p_1^{n_1} \cdot \ldots \cdot p_{k+1}^{n_{k+1}}, \tag{2.1}$$

where $\binom{N}{n_1,\ldots,n_{k+1}} = \frac{N!}{n_1! \cdot \ldots \cdot n_{k+1}!}$ is the multinomial coefficient and $n_1, \ldots, n_{k+1} \in \mathbb{N}_0$ with $\sum_{i=1}^{k+1} n_i = N$. The multinomial distribution is a multivariate generalisation of the binomial distribution and in the case of two categories they are equivalent.
For the choice of k cutpoints $x_1 < \ldots < x_k \in \mathscr{X}$ we define the random variables Y_i for the categories:

$$Y_i = \begin{cases} 1 & \text{if } x_{i-1} < X \leq x_i \\ 0 & \text{else} \end{cases}, \quad i = 1, \ldots, k+1. \tag{2.2}$$

Since the cutpoints are a null set ($P(X = x_i) = 0$) and it is negligible to which category a cutpoint belongs. Thus, for $x_0 > -\infty$ the value x_0 will be assumed to belong to the first category. The category probabilities are given by

$$p_{i,\boldsymbol{\theta}} = F_X(x_i, \boldsymbol{\theta}) - F_X(x_{i-1}, \boldsymbol{\theta}), \quad i = 1, \ldots, k+1. \tag{2.3}$$

We note that $F_X(x_0, \boldsymbol{\theta}) := \lim_{x \to x_0} F_X(x, \boldsymbol{\theta}) = 0$ and $F_X(x_{k+1}, \boldsymbol{\theta}) := \lim_{x \to x_{k+1}} F_X(x, \boldsymbol{\theta}) = 1$ hold by the definition of x_0 and x_{k+1}. The vector of the random variables Y_i

$$\boldsymbol{Y} = (Y_1, \ldots, Y_{k+1})^T \tag{2.4}$$

has the multinomial distribution $\boldsymbol{Y} \sim M\big(1, (p_{1,\boldsymbol{\theta}}, \ldots, p_{k+1,\boldsymbol{\theta}})\big)$.
Choosing the cutpoints $x_1, \ldots, x_k$ for i.i.d. random variables $X_1, \ldots, X_N$ leads to the i.i.d. random variables $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_N} \sim M\big(1, (p_{1,\boldsymbol{\theta}}, \ldots, p_{k+1,\boldsymbol{\theta}})\big)$, and the distribution of the sum $\sum_{j=1}^{N} \boldsymbol{Y_j}$ is multinomial $M\big(N, (p_{1,\boldsymbol{\theta}}, \ldots, p_{k+1,\boldsymbol{\theta}})\big)$.

3 Information for discretisation

In this section we consider the information for the discretised random variable. Discretising a continuous random variable leads to a loss of information. A desirable property would be that in the limiting case for $k \to \infty$ cutpoints the Fisher information of the multinomially distributed random variable converges to the Fisher information of the continuous random variable $\boldsymbol{I}_X(\boldsymbol{\theta})$. The next theorem shows that this can be realized for a suitable choice of cutpoints.

Theorem 3.1
Let X be an absolutely continuous random variable satisfying suitable regularity conditions and let the sequence of multinomially distributed random variables $\mathbf{Y}_k$ be defined by discretising X with k cutpoints $x_{1,k}, \ldots, x_{k,k}$. If $\max_{i=1,\ldots,k+1} \{F_X(x_{i,k}) - F_X(x_{i-1,k})\} \to 0$ for $k \to \infty$ then the limit of the Fisher information of $\mathbf{Y}_k$ is given by

$$\lim_{k\to\infty} \boldsymbol{I}_{\mathbf{Y}_k}(x_{1,k}, \ldots, x_{k,k}; \boldsymbol{\theta}) = \boldsymbol{I}_X(\boldsymbol{\theta}). \tag{3.1}$$

Hence, in the limiting case there is no loss of information as long as the category probabilities converge to zero.

In what follows we derive the information for the discretised random variable for some location and/or scale families. As a first example we consider the exponential distribution, which constitutes a scale family with CDF of the form

$$P(X \leq x) = F(\lambda x) \tag{3.2}$$

for $x \in \mathscr{X} = [0, \infty)$, where $F(z) = 1 - e^{-z}$ is the standard CDF and $\theta = \lambda > 0$ the scale parameter.
The second example is the logistic distribution, which is a location-scale family with CDF of the form

$$P(X \leq x) = F\left(\frac{x - \mu}{\sigma}\right), \tag{3.3}$$

where $F(z) = 1/(1 + e^{-z})$ is the standard CDF, μ and $\sigma > 0$ are the location and scale parameter and $\mathscr{X} = \mathbb{R}$. An equivalent formulation for the logistic distribution is obtained with $a = -\frac{\mu}{\sigma}$ and $b = \frac{1}{\sigma} > 0$:

$$P(X \leq x) = F(a + bx). \tag{3.4}$$

We also consider the Gumbel distribution with standard CDF $F(z) = \exp(-\exp(-z))$, which is also a location-scale family.
The quality of the chosen cutpoints will be measured in terms of the Fisher information $\boldsymbol{I}_{\boldsymbol{Y}}(x_1, \ldots, x_k; \theta)$. Since the Fisher information depends on the cutpoints, they are mentioned explicitly in the notation. For a scale family we have $\theta = \lambda$ and for a location-scale family we use $\boldsymbol{\theta} = (a, b)$. Let $z_{(a,b),i} = a + bx_i$ and $z_{\lambda,i} = \lambda x_i$ for $i = 1, \ldots, k$ and let $f(z) = F'(z)$ be the standard pdf. Note that $f(z) = e^{-z}$ for the exponential and

$f(z) = e^{-z}/(1+e^{-z})^2$ for the logistic distribution. We define the following matrices

$$\boldsymbol{D}_{f,\boldsymbol{\theta}} = \operatorname{diag}(f(z_{\boldsymbol{\theta},1}), \dots, f(z_{\boldsymbol{\theta},k})),$$

$$\boldsymbol{D}_{p,\boldsymbol{\theta}} = \operatorname{diag}(p_{1,\boldsymbol{\theta}}, \dots, p_{k+1,\boldsymbol{\theta}}),$$

$$\boldsymbol{H} = (\boldsymbol{I}_k \quad \mathbf{0}_k) - (\mathbf{0}_k \quad \boldsymbol{I}_k),$$

where $\boldsymbol{I}_k$ is the k-dimensional identity matrix and $\mathbf{0}_k = (0, \dots, 0)^T$ is the k-dimensional zero vector. For an absolutely continuous random variable X, whose probability distribution is a scale family or a location-scale family, and a k-dimensional cutpoint vector $\boldsymbol{x} = (x_1, \dots, x_k) \in \Xi_k$ the Fisher information for $\boldsymbol{Y}$ is given by Nguyen and Torsney (2007)

$$I_{\boldsymbol{Y}}(\boldsymbol{x}; \lambda) = \boldsymbol{x} \cdot \boldsymbol{Q}_\lambda \cdot \boldsymbol{x}^T, \tag{3.5}$$

$$\boldsymbol{I}_{\boldsymbol{Y}}(\boldsymbol{x}; a, b) = \boldsymbol{X}_D^T \cdot \boldsymbol{Q}_{a,b} \cdot \boldsymbol{X}_D \tag{3.6}$$

with intensity matrix $\boldsymbol{Q}_{\boldsymbol{\theta}} = \boldsymbol{D}_{f,\boldsymbol{\theta}} \cdot \boldsymbol{H} \cdot \boldsymbol{D}_{p,\boldsymbol{\theta}}^{-1} \cdot \boldsymbol{H}^T \cdot \boldsymbol{D}_{f,\boldsymbol{\theta}}$ and design matrix $\boldsymbol{X}_D = (\mathbf{1}_k, \boldsymbol{x}^T)$, where $\mathbf{1}_k$ is the k-dimensional unit vector.

Discretising N i.i.d. random variables $X_1, \dots, X_N$ with cutpoints $x_1, \dots, x_k$ leads to i.i.d. random variables $\boldsymbol{Y_1}, \dots, \boldsymbol{Y_N}$. The Fisher information for $\boldsymbol{Y_1}, \dots, \boldsymbol{Y_N}$ is given by $\boldsymbol{I}_{\boldsymbol{Y_1},\dots,\boldsymbol{Y_N}}(x_1, \dots, x_k; \boldsymbol{\theta}) = N \cdot \boldsymbol{I}_{\boldsymbol{Y_1}}(x_1, \dots, x_k; \boldsymbol{\theta})$.
In design theory the normalised information matrix M is considered. For N observations it is given by (cf. Silvey, 1980, p. 14):

$$\boldsymbol{M}_{\boldsymbol{Y_1},\dots,\boldsymbol{Y_N}}(x_1, \dots, x_k; \boldsymbol{\theta}) = \frac{1}{N} \cdot \boldsymbol{I}_{\boldsymbol{Y_1},\dots,\boldsymbol{Y_N}}(x_1, \dots, x_k; \boldsymbol{\theta}) = \boldsymbol{I}_{\boldsymbol{Y_1}}(x_1, \dots, x_k; \boldsymbol{\theta}). \tag{3.7}$$

Hence, it suffices to consider $N = 1$ and $\boldsymbol{Y} = \boldsymbol{Y_1}$.

4 Optimality criteria

Usually the variance of an unbiased estimator is bounded below by the inverse of the Fisher information and the maximum likelihood estimator is asymptotically efficient with asymptotical covariance matrix equal to the inverse of the Fisher information. Thus the information should be maximised in a certain sense (cf. Silvey, 1980, p. 3-5). We adopt the terminology from the theory of optimal designs (cf. Silvey, 1980, p. 10-14).
For a one-dimensional parameter θ the cutpoint vector $\boldsymbol{x}^*$ is optimal if the inequality $M_{\boldsymbol{Y}}(\boldsymbol{x}^*; \theta) \geq M_{\boldsymbol{Y}}(\boldsymbol{x}; \theta)$ is satisfied for all cutpoint vectors $\boldsymbol{x} \in \Xi_k$.
For multi-dimensional parameter vector $\boldsymbol{\theta}$ we consider the c-optimality criterion, which aims at minimising the variance of the estimator for the linear aspect $\boldsymbol{c}^T\boldsymbol{\theta}$ (cf. Atkinson

and Donev, 1996, p. 113). Such a linear aspect $\boldsymbol{c}^T\boldsymbol{\theta}$ is identifiable if there exists a vector $\boldsymbol{v}$ such that $\boldsymbol{c} = \boldsymbol{M}_{\boldsymbol{Y}}(\boldsymbol{x};\boldsymbol{\theta})\ \boldsymbol{v}$. The cutpoint vector $\boldsymbol{x}^*$ is c-optimal if the linear aspect $\boldsymbol{c}^T\boldsymbol{\theta}$ is identifiable and the inequality $\boldsymbol{c}^T\,\boldsymbol{M}_{\boldsymbol{Y}}^{-}(\boldsymbol{x}^*;\boldsymbol{\theta})\ \boldsymbol{c} \leq \boldsymbol{c}^T\,\boldsymbol{M}_{\boldsymbol{Y}}^{-}(\boldsymbol{x};\boldsymbol{\theta})\ \boldsymbol{c}$ holds for all cutpoint vectors $\boldsymbol{x} \in \Xi_k$ for which $\boldsymbol{c}^T\boldsymbol{\theta}$ is identifiable. Here $\boldsymbol{M}_{\boldsymbol{Y}}^{-}(\boldsymbol{x};\boldsymbol{\theta})$ is a generalized inverse of $\boldsymbol{M}_{\boldsymbol{Y}}(\boldsymbol{x};\boldsymbol{\theta})$.

The quality of a cutpoint vector $\boldsymbol{x} = (x_1,\ldots,x_k)$ can be expressed by the efficiency. We consider the relative efficiency with respect to the optimal cutpoint vector $\boldsymbol{x}^*$ and the relative efficiency with respect to the continuous random variable X. For a one-dimensional parameter θ these efficiencies are defined by

$$\text{eff}(\boldsymbol{x},\boldsymbol{\theta}) = \frac{M_{\boldsymbol{Y}}(\boldsymbol{x};\boldsymbol{\theta})}{M_{\boldsymbol{Y}}(\boldsymbol{x}^*;\boldsymbol{\theta})}, \qquad\qquad \text{eff}_X(\boldsymbol{x},\boldsymbol{\theta}) = \frac{M_{\boldsymbol{Y}}(\boldsymbol{x};\boldsymbol{\theta})}{M_X(\boldsymbol{\theta})}.$$

For c-optimality the relative c-efficiency with respect to the continuous random variable X is defined by

$$\text{eff}_{c,X}(\boldsymbol{x},\boldsymbol{\theta}) = \frac{\boldsymbol{c}^T\,\boldsymbol{M}_X^{-}(\boldsymbol{\theta})\ \boldsymbol{c}}{\boldsymbol{c}^T\,\boldsymbol{M}_{\boldsymbol{Y}}^{-}(\boldsymbol{x};\boldsymbol{\theta})\ \boldsymbol{c}}.$$

Since the model is non-linear the optimal cutpoints depend on the unknown parameters and they are called locally optimal (Chernoff, 1953). For a genereal solution the continuous random variable X is standardised by a canonical transformation (Ford et al., 1992). For the location-scale family the standardised random variable $Z : \Omega \to Z(\Omega) = \mathscr{Z}$ is defined by

$$Z = \frac{X-\mu}{\sigma} = a + bX, \tag{4.1}$$

while $Z = \lambda X$ is the standardised random variable in the scale family. We have $F_Z = F$ and $\mathscr{Z} = \mathscr{X}$. The advantage of the transformation is that the optimal cutpoint vector $\boldsymbol{z}^* = (z_1^*,\ldots,z_k^*)$ for Z does not depend on the unknown parameters. We get the optimal cutpoints for X via the inverse transformation $x_i^* = \mu + \sigma z_i^* = \frac{z_i^*-a}{b}$ for $i = 1,\ldots,k$. Hence, the optimisation problem is solved for arbitrary parameter vector $\boldsymbol{\theta}$ (Ford et al., 1992).

Let $\boldsymbol{Y}(X)$ and $\boldsymbol{Y}(Z)$ be the multinomially distributed random variables defined by discretising X and Z with k cutpoints x_i and $z_i = a + bx_i$ respectively. We give a relation between the information for $\boldsymbol{Y}(X)$ and $\boldsymbol{Y}(Z)$ (cf. Nguyen and Torsney, 2007). If the distribution of X is a location-scale family, according to equation (3.6) the information for Z is given by $\boldsymbol{M}_{\boldsymbol{Y}(Z)}(\boldsymbol{z};\boldsymbol{\theta}_0) = \boldsymbol{Z}_D^T \cdot \boldsymbol{Q}_{\boldsymbol{\theta}_0} \cdot \boldsymbol{Z}_D$, where $\boldsymbol{Z}_D = (\boldsymbol{1}_k, \boldsymbol{z}^T)$ and $\boldsymbol{\theta}_0 = (0,1)$. Let $\boldsymbol{B} = \left(\begin{smallmatrix}1 & a\\ 0 & b\end{smallmatrix}\right)$. Then we have $\boldsymbol{Z}_D = \boldsymbol{X}_D \cdot \boldsymbol{B}$. Since $b \neq 0$ the transformation matrix $\boldsymbol{B}$

is non-singular and we conclude:

$$\boldsymbol{M}_{\boldsymbol{Y}(X)}(\boldsymbol{x};\boldsymbol{\theta}) = \left(\boldsymbol{B}^{-1}\right)^T \cdot \boldsymbol{M}_{\boldsymbol{Y}(Z)}(\boldsymbol{z};\boldsymbol{\theta}_0) \cdot \boldsymbol{B}^{-1}. \tag{4.2}$$

For c-optimality the criterion function is given by

$$\boldsymbol{c}^T \cdot \boldsymbol{M}^-_{\boldsymbol{Y}(X)}(\boldsymbol{x};\boldsymbol{\theta}) \cdot \boldsymbol{c} = \boldsymbol{c}^T_{\boldsymbol{B}} \cdot \boldsymbol{M}^-_{\boldsymbol{Y}(Z)}(\boldsymbol{z};\boldsymbol{\theta}_0) \cdot \boldsymbol{c}_{\boldsymbol{B}}, \tag{4.3}$$

where $\boldsymbol{c}_{\boldsymbol{B}} = \boldsymbol{B}^T\boldsymbol{c}$. In order to estimate the location parameter μ as effectively as possible we have to minimise the criterion function $(1,0) \cdot \boldsymbol{M}^-_{\boldsymbol{Y}(Z)}(\boldsymbol{z};\boldsymbol{\theta}_0) \cdot (1,0)^T$, which can be shown by the delta method (cf. Nguyen and Torsney, 2007).
If the distribution of X is a scale-family, we analogously obtain the relation

$$M_{\boldsymbol{Y}(X)}(\boldsymbol{x};\lambda) = \frac{1}{\lambda^2} \cdot M_{\boldsymbol{Y}(Z)}(\boldsymbol{z};\lambda_0), \tag{4.4}$$

where $\lambda_0 = 1$. So we can maximise the standardised information $M_{\boldsymbol{Y}(Z)}(\boldsymbol{z};\lambda_0)$ instead of $M_{\boldsymbol{Y}(X)}(\boldsymbol{x};\lambda)$.

5 Optimal cutpoints for the exponential distribution

An exponentially distributed random variable $X \sim \mathrm{Exp}(\lambda)$ with $0 < \lambda \in \mathbb{R}$ has the CDF $F_X(x) = 1 - e^{-\lambda x}$ and the pdf $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. The exponential distribution is used for modelling life times and it is the only continuous memoryless distribution, which means that $P(X > s + t \,|\, X > s) = P(X > t)$ for all $s, t \geq 0$. This property enables the derivation of a recursion for the information and the optimal cutpoints as we will show in this section.
The information for the multinomially distributed random variable $\boldsymbol{Y}$ in the case of the standard exponential distribution follows from equation (3.5):

$$M_{\boldsymbol{Y}}(z_1,\ldots,z_k;\lambda_0) = \frac{z_1^2}{e^{2z_1} - e^{z_1}} + \sum_{i=2}^{k} \frac{\left(e^{-z_i} \cdot z_i - e^{-z_{i-1}} \cdot z_{i-1}\right)^2}{e^{-z_{i-1}} - e^{-z_i}} + \frac{z_k^2}{e^{z_k}}. \tag{5.1}$$

In the following, the Lambert W-function is used, which is defined as the inverse $W = g^{-1}$ of the function $g(w) = w \cdot e^w$ for $w \geq -1$. Hence, $W(x)$ is defined for $x \geq -1/e$ and it satisfies the equation $W(x) \cdot e^{W(x)} = x$. For more information about the Lambert W-function see Corless et al. (1996). The next theorem gives the optimal cutpoint for $k = 1$.

Theorem 5.1
Let X be an exponentially distributed random variable with scale parameter λ. The optimal cutpoint is given by $x^* = \frac{1}{\lambda} \cdot (W(-2 \cdot e^{-2}) + 2) \approx 1.594/\lambda$.

Proof:
The standardised random variable $Z = \lambda X$ is considered. The multinomially distributed random variable $\boldsymbol{Y}$ is defined by the choice of a cutpoint $z \in (0, \infty)$. By equation (5.1) the information is given by

$$M_{\boldsymbol{Y}}(z; \lambda_0) = \frac{z^2}{e^z - 1}. \tag{5.2}$$

The derivative of the information with respect to z is given by

$$\frac{\mathrm{d}M_{\boldsymbol{Y}}(z; \lambda_0)}{\mathrm{d}z} = \frac{z \cdot (2e^z - 2 - ze^z)}{(e^z - 1)^2}.$$

The derivative equals zero for $z \in (0, \infty)$ if and only if $2e^z - 2 - ze^z = 0$ holds. The equation can be rewritten as $(z - 2) \cdot e^{z-2} = -2e^{-2}$. With the notation of the Lambert W-function one solution can be written as $z^* = W\big(-2e^{-2}\big) + 2$. The other solution is $z = 0$. It can be checked analytically that $M_{\boldsymbol{Y}}(z; \lambda_0)$ is maximised at $z = z^*$ and z^* is unique. It follows that the optimal cutpoint for the continuous random variable X is given by $x^* = \frac{1}{\lambda} \cdot (W(-2 \cdot e^{-2}) + 2)$. □

The optimal cutpoint x^* with $F(x^*) = 0.797$ is much larger than the expectation $1/\lambda$ and more than twice the median $x_{med} = \frac{1}{\lambda} \cdot \ln(2) \approx 0.693/\lambda$. The next lemma gives a representation of the Fisher information which will be needed for further results.

Lemma 5.2
Let X be an absolutely continuous random variable, whose distribution is a scale family with scale parameter λ, and let $\boldsymbol{Y}$ be the multinomially distributed random variable defined by discretising X with k cutpoints $x_1, \ldots, x_k$. Let $\tilde{\boldsymbol{Y}}$ be the multinomially distributed random variable defined by discretising a random variable with conditional distribution of $X - x_1$, given $X > x_1$, with $k - 1$ cutpoints $\tilde{x}_i = x_{i+1} - x_1$, $i = 1, \ldots, k - 1$. Then we have the following relation for the Fisher information:

$$\begin{aligned} \boldsymbol{I}_{\boldsymbol{Y}}(x_1, \ldots, x_k; \lambda) &= P(X > x_1) \cdot \boldsymbol{I}_{\tilde{\boldsymbol{Y}}}(\tilde{x}_1, \ldots, \tilde{x}_{k-1}; \lambda) \\ &\quad - P(X \le x_1) \cdot \left(\frac{\partial}{\partial \lambda} \left(\frac{1}{p_{1,\lambda}} \right) \cdot \frac{\partial p_{1,\lambda}}{\partial \lambda} + \frac{1}{p_{1,\lambda}} \cdot \frac{\partial^2 p_{1,\lambda}}{\partial \lambda^2} \right). \end{aligned} \tag{5.3}$$

Remark 5.3
Lemma 5.2 can be analogously formulated for an arbitrary absolutely continuous random variable satisfying some regularity conditions and for a multidimensional parameter vector $\boldsymbol{\theta}$.

The following lemma based on Lemma 5.2 and the memorylessness gives an alternative representation of the information for the exponential distribution.

Lemma 5.4
Let X be an exponentially distributed random variable with scale parameter λ. Let $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$ be defined by discretising X with k cutpoints $x_1, \ldots, x_k$ and with $k-1$ cutpoints $\tilde{x}_i = x_{i+1} - x_1$, $i = 1, \ldots, k-1$, respectively. Then the following relation for the information holds:

$$M_{\boldsymbol{Y}}(x_1, \ldots, x_k; \lambda) = \frac{x_1^2}{e^{\lambda x_1} - 1} + e^{-\lambda x_1} \cdot M_{\tilde{\boldsymbol{Y}}}(\tilde{x}_1, \ldots, \tilde{x}_{k-1}; \lambda). \tag{5.4}$$

Since the exponential distribution is memoryless, the assertion follows directly from Lemma 5.2.
With the representation of the information in Lemma 5.4 a relation between the $k-1$ optimal cutpoints and differences of the k optimal cutpoints can be established.

Theorem 5.5
Let X be an exponentially distributed random variable with scale parameter λ. For $k-1$ cutpoints let $\tilde{x}_1^*, \ldots, \tilde{x}_{k-1}^*$ be optimal and for k cutpoints let $x_1^*, \ldots, x_k^*$ be optimal. Then $\tilde{x}_i^* = x_{i+1}^* - x_1^*$ holds for $i = 1, \ldots, k-1$.

Proof:
The theorem follows from the representation of the information in equation (5.4). Since $e^{-\lambda x_1} > 0$ for all $x_1 \in \mathbb{R}$, the information $M_{\tilde{\boldsymbol{Y}}}(\tilde{x}_1, \ldots, \tilde{x}_{k-1}; \lambda)$ can be maximised independently from x_1 and it is maximal for the $k-1$ optimal cutpoints. □

Application of Theorem 5.5 leads to a recursive formula for the optimal cutpoints and the corresponding optimal information.

Theorem 5.6
Let X be an exponentially distributed random variable with scale parameter λ and $(r_l)_{l \geq 0}$ the following recursively defined sequence:

$$r_0 := 0, \tag{5.5}$$

$$r_\ell := W\big((-r_{\ell-1} - 2) \cdot e^{-r_{\ell-1}-2}\big), \quad \ell \geq 1. \tag{5.6}$$

The $k \geq 1$ optimal cutpoints are given by

$$x_{1,k}^* = \frac{1}{\lambda} \cdot (r_k + r_{k-1} + 2), \tag{5.7}$$

$$x_{i,k}^* = \sum_{\ell=k-i+1}^{k} x_{1,\ell}^* = \sum_{\ell=k-i+1}^{k} \frac{1}{\lambda} \cdot (r_\ell + r_{\ell-1} + 2), \quad 1 \leq i \leq k. \tag{5.8}$$

The information for the optimal cutpoints is given by

$$M_{\mathbf{Y}}(x_{1,k}^*, \ldots, x_{k,k}^*; \lambda) = \frac{1}{\lambda^2} \cdot (-r_k \cdot (r_k + 2)). \tag{5.9}$$

For $k = 2$ the two optimal cutpoints are given explicitly in the following corollary, where the equation

$$e^{W(-2\cdot e^{-2})} = \frac{-2 \cdot e^{-2}}{W(-2 \cdot e^{-2})} \tag{5.10}$$

is applied. The difference of the two optimal cutpoints equals the value of the optimal cutpoint for $k = 1$ and thus we have $x_{2,2}^* - x_{1,2}^* = x_{1,1}^* = \frac{1}{\lambda} \cdot (W(-2 \cdot e^{-2}) + 2)$.

Corollary 5.7
Let X be an exponentially distributed random variable with scale parameter λ. For $k = 2$ the two optimal cutpoints are given by

$$x_{1,2}^* = \frac{1}{\lambda} \cdot \Big[W\big(-2e^{-2}\big) + W\Big(\frac{1}{2} \cdot W\big(-2e^{-2}\big) \cdot \big(2 + W\big(-2e^{-2}\big)\big)\Big) + 2 \Big] \approx \frac{1}{\lambda} \cdot 1.018,$$

$$x_{2,2}^* = \frac{1}{\lambda} \cdot \Big[2W\big(-2e^{-2}\big) + W\Big(\frac{1}{2} \cdot W\big(-2e^{-2}\big) \cdot \big(2 + W\big(-2e^{-2}\big)\big)\Big) + 4 \Big] \approx \frac{1}{\lambda} \cdot 2.611.$$

The CDF and the location of the optimal cutpoints are visualised in Figure 5.1. In Table 5.1 the optimal cutpoints for the exponential distribution with parameter $\lambda = 1$, the value of the CDF, which is independent of the parameter λ, and the information are listed. The relative efficiency with respect to the continuous exponential distribution is also independent of λ and is equal to the information of the multinomial distributed random variable, because the information of the exponential distribution is 1 for $\lambda = 1$.

The k optimal cutpoints interlace the $k - 1$ optimal ones. As shown in Table 5.1 the information is increasing with the number of cutpoints, but the gain in information becomes smaller. By Theorem 3.1 the Fisher information converges to the Fisher information $I_X(\lambda) = \frac{1}{\lambda^2}$ of an exponentially distributed random variable.

Example 5.8
We compare the $k + 1$ optimal cutpoints given in Theorem 5.6 with equally spaced quantiles $x_i = F_X^{-1}(\frac{i}{k+1}) = -\frac{1}{\lambda} \cdot \ln\big(1 - \frac{i}{k+1}\big)$. For these cutpoints the probabilities of the categories are equal and for $k = 1$ the cutpoint is the median. As shown in Theorem 3.1 the information converges in both cases for $k \to \infty$ to the information of the exponentially distributed random variable. However, for a fixed number of cutpoints the information for the optimal ones is much greater as shown in Table 5.2, where the

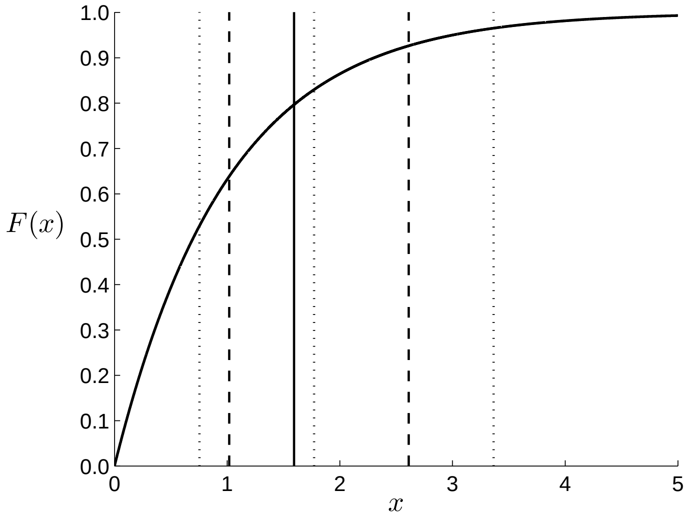

Figure 5.1: CDF and optimal cutpoints for the exponential distribution with parameter $\lambda_0 = 1$ (solid line: 1 cutpoint, dashed line: 2 cutpoints, dotted line: 3 cutpoints)

Number	Cutpoints	$F(z^*)$	$M_{\boldsymbol{Y}}(\boldsymbol{z}^*; \lambda_0)$	$\text{eff}_Z(\boldsymbol{z}^*, \lambda_0)$
1	1.594	0.797	0.648	0.648
2	1.018	0.639	0.820	0.820
	2.611	0.927		
3	0.754	0.530	0.891	0.891
	1.772	0.830		
	3.365	0.965		
4	0.600	0.451	0.927	0.927
	1.354	0.742		
	2.372	0.907		
	3.966	0.981		

Table 5.1: Optimal cutpoints for the exponential distribution with parameter $\lambda = \lambda_0 = 1$

relative efficiencies are calculated with respect to the optimal cutpoints and with respect to the exponentially distributed random variable. Note again that the relative efficieny does not depend on λ.

Using the results for the exponential distribution, we obtain optimal cutpoints for the Gumbel distribution, which is an extreme value distribution. For $x \in \mathbb{R}$ its CDF is given by $F_G(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right)$. If $X \sim \text{Exp}(\lambda)$ then the random variable $G = -\sigma \cdot \ln(X)$ has the Gumbel distribution with location parameter $\mu = \ln(\lambda)$ and scale parameter σ.

Number	$M_{\boldsymbol{Y}}(\boldsymbol{z};\lambda_0)$	$\text{eff}(\boldsymbol{z},\lambda_0)$	$\text{eff}_Z(\boldsymbol{z},\lambda_0)$
1	0.480	0.741	0.480
2	0.649	0.792	0.649
3	0.735	0.825	0.735
4	0.787	0.849	0.787
5	0.822	0.867	0.822
10	0.902	0.918	0.902
20	0.949	0.953	0.949

Table 5.2: Information and relative efficiencies for cutpoints with corresponding equal category probabilities for the exponential distribution with parameter $\lambda = \lambda_0 = 1$

Corollary 5.9
Let G have the Gumbel distribution with location parameter μ and known scale parameter σ_0. The optimal cutpoints are given by $g_i^* = \mu - \sigma_0 \cdot \ln(z_{k+1-i}^*)$, where $z_1^*, \ldots, z_k^*$ are the optimal cutpoints for $Z \sim \text{Exp}(1)$.

Since we have a bijective and monotonic mapping between G and $X \sim \text{Exp}(\lambda)$ with $\lambda = e^{\mu}$ the optimal cutpoints for the Gumbel distribution follow directly.

6 Optimal cutpoints for the logistic distribution

The logistic distribution is used in contingent valuation studies (cf. Gunduz and Torsney, 2006) and has many other applications, such as in psychophysical experiments. A logistically distributed random variable $X \sim \text{Lo}(\mu, \sigma^2)$ has the CDF F_X and pdf f_X:

$$F_X(x) = \frac{1}{1 + e^{-\frac{x-\mu}{\sigma}}}, \qquad f_X(x) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma \cdot \left(1 + e^{-\frac{x-\mu}{\sigma}}\right)^2}.$$

Our aim is to estimate the parameter μ as effectively as possible and therefore we determine the c-optimal cutpoints for μ. Nguyen and Torsney (2007) have numerically calculated up to five optimal cutpoints. The next theorem gives for an arbitrary number of cutpoints the optimal ones.

Theorem 6.1
Let $X \sim \text{Lo}(\mu, \sigma^2)$. The $k \geq 2$ c-optimal cutpoints for μ are the $i/(k+1)$-quantiles $x_i^* = \mu + \sigma \cdot \ln\left(\frac{i}{k+1-i}\right)$ for $i = 1, \ldots, k$. Thus the category probabilities are equal and are given by $p_{i,\theta}^* = 1/(k+1)$.

The value of the criterion function and the relative efficiency with respect to the standard logistic distribution are shown in Table 6.1. The efficiency $\text{eff}_{c,Z}(\boldsymbol{z}^*, \boldsymbol{\theta}_0)$ is increasing with the number of cutpoints and converges to 1 as the number of cutpoints tends to infinity.

Number	$(1,0) \cdot M_{\boldsymbol{Y}}^{-1}(\boldsymbol{z}^*; \boldsymbol{\theta}_0) \cdot (1,0)^T$	$\text{eff}_{c,Z}(\boldsymbol{z}^*, \boldsymbol{\theta}_0)$
2	3.375	0.889
3	3.200	0.937
4	3.125	0.960
5	3.086	0.972
6	3.063	0.979
10	3.025	0.992

Table 6.1: Information and relative efficiency for c-optimal cutpoints for μ for the logistic distribution with parameters $\mu = 0$ and $\sigma = 1$

The cutpoints $x_i^* = \mu + \sigma \cdot \ln\left(\frac{i}{k+1-i}\right)$ for $i = 1, \ldots, k,\ k \geq 1$, are also optimal for the logistic distribution with known scale parameter σ and thus one-dimensional parameter $\theta = \mu$ since then the Fisher information is given by the term S_1 defined in the proof of Theorem 6.1 given in the appendix.
For other optimality criteria like D-optimality or c-optimality for the scale parameter σ there Nguyen and Torsney (2007) proposed numeric solutions for the optimal cutpoints. In contrast to their conjecture the c-optimal cutpoints for σ turn out not to be symmetric in the case of an odd number of cutpoints. Then there are two optimal cutpoint vectors which are symmetric to each other. In particular, for $k = 3$ cutpoints and thus four categories the optimal cutpoint vectors for the standard logistic distribution are $\boldsymbol{z}^* = (-3.323, -1.752, 2.039)$ or $\boldsymbol{z}^* = (-2.039, 1.752, 3.323)$. The value of the criterion function is given by 0.923 for the optimal cutpoints and the relative efficiency with respect to the continuous random variable is given by $\text{eff}_{c,Z}(\boldsymbol{z}^*, \boldsymbol{\theta}) = 0.758$.
If different cutpoint vectors of equal length can be chosen for different observations and one is interested in estimating σ as effectively as possible it can be numerically shown that it is optimal to chose two different cutpoint vectors of length 3 instead of using the same cutpoint vector for each observation. The optimal cutpoint vectors for the standard logistic distribution are $\boldsymbol{z}_1^* = (-3.242, -1.676, 2.105)$ and $\boldsymbol{z}_2^* = -\boldsymbol{z}_1^*$, where each cutpoint vector should be used for half of the observations. The value of the criterion function is given by 0.919 for these optimal cutpoint vectors and the relative efficiency with respect to the continuous random variable is given by $\text{eff}_{c,Z}(\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, \boldsymbol{\theta}) = 0.761$. So there is a little improvement choosing two different cutpoint vectors. For an even number of cutpoints it is optimal to use the same cutpoint vector.

7 Discussion

We have presented some theory of cutpoints and have derived a formula for the corresponding Fisher information. Furthermore, we have shown that the Fisher information of the multinomially distributed random variable converges to the Fisher information of the continuous random variable if the maximal category probability converges to zero. Thus there is no loss of information in the limiting case.

Due to the memorylessness of the exponential distribution we could establish a relation between the information for k cutpoints and the information for the $k-1$ cutpoint differences. This has led to a recursive relation between the $k-1$ optimal cutpoints and the differences of the k optimal cutpoints. Based on these results we gave for an arbitrary number of cutpoints a recursive formula for the optimal ones.

For the logistic distribution we have proved that the c-optimal cutpoints for the location parameter have corresponding equal category probabilities. A possible extension could be the investigation of other optimality criteria or using different cutpoint vectors for different observations. It can be optimal to use the same cutpoint vector for all observations, but as pointed out there are also examples where usage of different cutpoint vectors improves the value of the criterion function.

Another topic for further research is the consideration of other probability distributions as non-symmetric distributions with multidimensional parameter vector. Probably analytical solutions will rarely exist and numerical optimisation techniques will have often to be used.

A Appendix

Proof of Theorem 3.1:

The log-likelihood function follows from equation (2.1):

$$l(\boldsymbol{\theta}; y_1, \ldots, y_{k+1}) = \ln(L(\boldsymbol{\theta}; y_1, \ldots, y_{k+1})) = \ln(f_{Y,\boldsymbol{\theta}}(y_1, \ldots, y_{k+1})) = \sum_{i=1}^{k+1} y_i \cdot \ln(p_{i,\boldsymbol{\theta}}).$$

We note that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a row vector. The Fisher information is given by

$$\begin{aligned}\boldsymbol{I}_{\boldsymbol{Y}}(x_{1,k}, \ldots, x_{k,k}; \boldsymbol{\theta}) &= E\left(\frac{\partial}{\partial \boldsymbol{\theta}^T}\Big(l(\boldsymbol{\theta}; Y_1, \ldots, Y_{k+1})\Big) \cdot \frac{\partial}{\partial \boldsymbol{\theta}}\Big(l(\boldsymbol{\theta}; Y_1, \ldots, Y_{k+1})\Big)\right) \\ &= -\sum_{i=1}^{k+1} \frac{\partial}{\partial \boldsymbol{\theta}^T}\Big(\ln(p_{i,\boldsymbol{\theta}})\Big) \cdot \frac{\partial}{\partial \boldsymbol{\theta}}\Big(\ln(p_{i,\boldsymbol{\theta}})\Big) \cdot p_{i,\boldsymbol{\theta}}.\end{aligned}$$

By the mean value theorem for integrals there exist $\xi_i \in [x_{i-1,k}, x_{i,k}]$, $i = 1, \ldots, k+1$, such that $p_{i,\boldsymbol{\theta}} = F_X(x_{i,k}, \boldsymbol{\theta}) - F_X(x_{i-1,k}, \boldsymbol{\theta}) = f_X(\xi_i, \boldsymbol{\theta}) \cdot (x_{i,k} - x_{i-1,k})$. We note that:

$$\frac{\partial}{\partial \boldsymbol{\theta}}\Big(\ln\big(f_X(\xi_i, \boldsymbol{\theta}) \cdot (x_{i,k} - x_{i-1,k})\big)\Big) = \frac{\partial}{\partial \boldsymbol{\theta}}\Big(\ln\big(f_X(\xi_i, \boldsymbol{\theta})\big)\Big).$$

The sum converges to the Stieltjes integral for $k \to \infty$ cutpoints:

$$\begin{aligned}
&\lim_{k\to\infty} \boldsymbol{I}_{\boldsymbol{Y}}(x_{1,k}, \ldots, x_{k,k}; \boldsymbol{\theta}) \\
&= \lim_{k\to\infty} \sum_{i=1}^{k+1} \frac{\partial}{\partial \boldsymbol{\theta}^T}\Big(\ln\big(f_X(\xi_i, \boldsymbol{\theta})\big)\Big) \cdot \frac{\partial}{\partial \boldsymbol{\theta}}\Big(\ln\big(f_X(\xi_i, \boldsymbol{\theta})\big)\Big) \cdot \big(F_X(x_{i,k}, \boldsymbol{\theta}) - F_X(x_{i-1,k}, \boldsymbol{\theta})\big) \\
&= \int_{\mathscr{X}} \left[\frac{\partial}{\partial \boldsymbol{\theta}^T}\Big(\ln\big(f_X(x, \boldsymbol{\theta})\big)\Big) \cdot \frac{\partial}{\partial \boldsymbol{\theta}}\Big(\ln\big(f_X(x, \boldsymbol{\theta})\big)\Big)\right] \mathrm{d}F_X(x) \\
&= E\left(\frac{\partial l(\boldsymbol{\theta}; X)}{\partial \boldsymbol{\theta}^T} \cdot \frac{\partial l(\boldsymbol{\theta}; X)}{\partial \boldsymbol{\theta}}\right) = \boldsymbol{I}_X(\boldsymbol{\theta}),
\end{aligned}$$

where $l(\boldsymbol{\theta}; X)$ is the log-likelihood function for the continuous random variable X. □

Proof of Lemma 5.2:
The log-likelihood function is given by $l(\lambda; y_1, \ldots, y_{k+1}) = \sum_{i=1}^{k+1} y_i \cdot \ln(p_{i,\lambda})$. The first and second derivative of the log-likelihood function with respect to λ can be calculated as:

$$\begin{aligned}
\frac{\partial l(\lambda; y_1, \ldots, y_{k+1})}{\partial \lambda} &= \sum_{i=1}^{k+1} \frac{\partial l(\lambda; y_1, \ldots, y_{k+1})}{\partial p_{i,\lambda}} \cdot \frac{\partial p_{i,\lambda}}{\partial \lambda} = \sum_{i=1}^{k+1} \frac{y_i}{p_{i,\lambda}} \cdot \frac{\partial p_{i,\lambda}}{\partial \lambda}, \\
\frac{\partial^2 l(\lambda; y_1, \ldots, y_{k+1})}{\partial \lambda^2} &= \sum_{i=1}^{k+1} y_i \cdot \left(\frac{\partial}{\partial \lambda}\left(\frac{1}{p_{i,\lambda}}\right) \cdot \frac{\partial p_{i,\lambda}}{\partial \lambda} + \frac{1}{p_{i,\lambda}} \cdot \frac{\partial^2 p_{i,\lambda}}{\partial \lambda^2}\right).
\end{aligned} \tag{A.1}$$

The Fisher information is given by

$$\begin{aligned}
I_{\boldsymbol{Y}}(x_1, \ldots, x_k; \lambda) = -E&\left(\frac{\partial^2 l(\lambda; Y_1, \ldots, Y_{k+1})}{\partial \lambda^2}\right) \\
&= -P(X > x_1) \cdot E\left(\frac{\partial^2 l(\lambda; Y_1, \ldots, Y_{k+1})}{\partial \lambda^2}\middle| Y_1 = 0\right) \\
&\quad - P(X \leq x_1) \cdot E\left(\frac{\partial^2 l(\lambda; Y_1, \ldots, Y_{k+1})}{\partial \lambda^2}\middle| Y_1 = 1\right),
\end{aligned} \tag{A.2}$$

The condition $Y_1 = 1$ implies $Y_2 = \ldots = Y_{k+1} = 0$. Thus the conditional expectation is given by

$$E\left(\frac{\partial^2 l(\lambda;Y_1,\ldots,Y_{k+1})}{\partial\lambda^2}\middle| Y_1=1\right)=\frac{\partial}{\partial\lambda}\left(\frac{1}{p_{1,\lambda}}\right)\cdot\frac{\partial p_{1,\lambda}}{\partial\lambda}+\frac{1}{p_{1,\lambda}}\cdot\frac{\partial^2 p_{1,\lambda}}{\partial\lambda^2}.$$

We note that the conditional distribution of $(Y_2,\ldots,Y_{k+1})^T$, given $Y_1=0$, is multinomial $M\big(1,(q_{1,\lambda},\ldots,q_{k,\lambda})\big)$, where $q_{i,\lambda}=\frac{p_{i+1,\lambda}}{1-p_{1,\lambda}}$ for $i=1,\ldots,k$ (see Balakrishnan and Nevzorov, 2003, p. 251-253).
The distribution of $\tilde{Y}$ is multinomial $M\big(1,(\tilde{p}_{1,\lambda},\ldots,\tilde{p}_{k,\lambda})\big)$, where the category probabilities are given by

$$\tilde{p}_{i,\lambda}=P(x_i-x_1\le X-x_1\le x_{i+1}-x_1\,|\,X>x_1)=\frac{P(x_i\le X\le x_{i+1})}{P(X>x_1)}=q_{i,\lambda}.$$

Thus the equation

$$E\left(\frac{\partial^2 l(\lambda;Y_1,\ldots,Y_{k+1})}{\partial\lambda^2}\middle| Y_1=0\right)=E\left(\frac{\partial^2 l(\lambda;\tilde{Y}_1,\ldots,\tilde{Y}_k)}{\partial\lambda^2}\right)$$

holds and the representation of the information is shown. □

Proof of Theorem 5.6:
The standardised random variable $Z=\lambda X$ is considered. First we prove the equations (5.7) and (5.9) by induction on the number of cutpoints.
Let $k=1$. By Theorem 5.1 we have $z_{1,1}^*=r_0+2+r_1=2+W\big(-2\cdot e^{-2}\big)$. Hence, by equations (5.2) and (5.10) the information is given by

$$\begin{aligned}M_{\boldsymbol{Y}}(z_{1,1}^*;\lambda_0)&=\frac{\big(W\big(-2\cdot e^{-2}\big)+2\big)^2}{e^{W(-2\cdot e^{-2})+2}-1}=-W\big(-2\cdot e^{-2}\big)\cdot\big(W\big(-2\cdot e^{-2}\big)+2\big)\\&=-r_1\cdot(r_1+2).\end{aligned}$$

We make the induction step $k\to k+1$. Assume that equations (5.7) and (5.9) hold for k cutpoints. Since by Theorem 5.5 the optimal differences are equal to the optimal cutpoints $z_{1,k}^*,\ldots,z_{k,k}^*$, we insert them into the information from equation (5.4):

$$\begin{aligned}M(z_{1,k+1})&:=\frac{z_{1,k+1}^2}{e^{z_{1,k+1}}-1}+e^{-z_{1,k+1}}\cdot M_{\boldsymbol{Y}}(z_{1,k}^*,\ldots,z_{k,k}^*;\lambda)\\&=\frac{z_{1,k+1}^2}{e^{z_{1,k+1}}-1}-e^{-z_{1,k+1}}\cdot r_k\cdot(r_k+2).\end{aligned}$$

The first derivative with respect to $z_{1,k+1}$ is given by

$$\frac{\mathrm{d}M(z_{1,k+1})}{\mathrm{d}z_{1,k+1}}=\frac{2z_{1,k+1}\cdot e^{z_{1,k+1}}-2z_{1,k+1}-z_{1,k+1}^2\cdot e^{z_{1,k+1}}}{\left(e^{z_{1,k+1}}-1\right)^2}+e^{-z_{1,k+1}}\cdot r_k\cdot(r_k+2).$$

Setting the derivative equal to zero leads to the equation

$$0 = 2z_{1,k+1} \cdot e^{2z_{1,k+1}} - 2z_{1,k+1} \cdot e^{z_{1,k+1}} - z_{1,k+1}^2 \cdot e^{2z_{1,k+1}} + \left(e^{2z_{1,k+1}} - 2e^{z_{1,k+1}} + 1\right) \cdot r_k \cdot (r_k + 2),$$

which can be rewritten as the equivalent equation:

$$(-z_{1,k+1} \cdot e^{z_{1,k+1}} - 2 + 2e^{z_{1,k+1}} + r_k \cdot e^{z_{1,k+1}} - r_k) \cdot (z_{1,k+1} \cdot e^{z_{1,k+1}} + r_k \cdot e^{z_{1,k+1}} - r_k) = 0.$$

The equation is satisfied if and only if at least one of the two factors is equal to zero. We look for the zeros of $z_{1,k+1} \cdot e^{z_{1,k+1}} + r_k \cdot e^{z_{1,k+1}} - r_k$. This leads to the equation:

$$\frac{z_{1,k+1} + r_k}{r_k} = e^{-z_{1,k+1}}.$$

The left-hand side of the equation is a linear function in $z_{1,k+1}$, the right-hand side is a strictly convex function. Thus, there can be only two solutions of the equation. One solution is $z_{1,k+1} = 0$. We have $-1 < r_k < 0$ for $k \geq 1$ and thus the slope of the linear function is less than -1. Hence, the linear function can only have a second intersection point with the function $e^{-z_{1,k+1}}$ in the negative range because the latter function has a larger slope for $z_{1,k+1} > 0$.
We look for the solutions of the equation:

$$-z_{1,k+1} \cdot e^{z_{1,k+1}} - 2 + 2e^{z_{1,k+1}} + r_k \cdot e^{z_{1,k+1}} - r_k = 0.$$

Rewriting and multiplication with e^{-r_k-2} leads to the equation:

$$e^{z_{1,k+1}-r_k-2} \cdot \left(z_{1,k+1} - r_k - 2\right) = (-r_k - 2) \cdot e^{-r_k-2}.$$

Again one solution is $z_{1,k+1} = 0$, which is not valid. With the notation of the Lambert W-function the second solution can be written as

$$z_{1,k+1}^* = r_k + 2 + W\left((-r_k - 2) \cdot e^{-r_k-2}\right) = r_k + 2 + r_{k+1}.$$

We have to verify that there is a maximum at $z_{1,k+1}^*$. The function $M(z_{1,k+1})$ converges to zero for $z_{1,k+1} \to \infty$ and to $-r_k \cdot (r_k + 2)$ for $z_{1,k+1} \to 0$, because by L'Hospital's rule we have:

$$\lim_{z_{1,k+1} \to 0} \frac{z_{1,k+1}^2}{e^{z_{1,k+1}} - 1} = \lim_{z_{1,k+1} \to 0} \frac{2z_{1,k+1}}{e^{z_{1,k+1}}} = 0.$$

By two times application of L'Hospital's rule it can be shown that the derivative of $M(z_{1,k+1})$ converges to $1 + r_k \cdot (r_k + 2) > 0$ for $z_{1,k+1} \to 0$. Hence, there must exist a maximum in the positive range which is located at $z^*_{1,k+1}$. We have the following relation:

$$e^{r_{k+1}} = e^{W\left((-r_k-2)\cdot e^{-r_k-2}\right)} = \frac{(-r_k-2)\cdot e^{-r_k-2}}{W\left((-r_k-2)\cdot e^{-r_k-2}\right)} = \frac{(-r_k-2)\cdot e^{-r_k-2}}{r_{k+1}}$$

Inserting the optimal cutpoints into the information and using the previous equation leads to:

$$\begin{aligned} M_{\boldsymbol{Y}}(z^*_{1,k+1},\dots,z^*_{k+1,k+1};\lambda_0) &= \frac{(r_k+2+r_{k+1})^2}{e^{r_k+2+r_{k+1}}-1} - e^{-r_k-2-r_{k+1}}\cdot r_k\cdot(r_k+2) \\ &= \frac{(r_k+2+r_{k+1})^2}{\dfrac{-r_k-2}{r_{k+1}}-1} - \frac{r_{k+1}}{-r_k-2}\cdot r_k\cdot(r_k+2) \\ &= -r_{k+1}\cdot(r_{k+1}+2). \end{aligned}$$

The assertion is proved and with transformation we get the optimal cutpoints $x^*_{1,k}$ for the random variable X and the corresponding optimal information given in the theorem.

By induction on i we show equation (5.8) for the optimal cutpoints $x^*_{i,k}$. For $i = 1$ we have $x^*_{1,k} = x^*_{1,k}$ for all $k \geq 1$. We make the inductive step $i \to i+1$. Using the recursion in Theorem 5.5 and the induction hypothesis we have

$$x^*_{i+1,k} = x^*_{1,k} + x^*_{i,k-1} = x^*_{1,k} + \sum_{l=k-i}^{k-1} x^*_{1,l} = \sum_{l=k-(i+1)+1}^{k} x^*_{1,l}$$

and the theorem is proven. □

Proof of Theorem 6.1:
The standardised random variable $Z = a + bX$ is considered. Instead of minimising the criterion function $(1,0)\cdot \boldsymbol{M}^{-1}_{\boldsymbol{Y}}(z_1,\dots,z_k;\boldsymbol{\theta}_0)\cdot(1,0)^T$ we want to maximise the inverse. The information matrix is given in equation (3.6) and its determinant is given by $\det\left(\boldsymbol{M}_{\boldsymbol{Y}}(z_1,\dots,z_k;\boldsymbol{\theta}_0)\right) = S_1\cdot S_2 - S_3^2$ with

$$S_1 = \sum_{i=1}^{k+1}\frac{\left[1_{\{i<k+1\}}f(z_i) - 1_{\{i>1\}}f(z_{i-1})\right]^2}{F(z_i)-F(z_{i-1})},\; S_2 = \sum_{i=1}^{k+1}\frac{\left[1_{\{i<k+1\}}f(z_i)\cdot z_i - 1_{\{i>1\}}f(z_{i-1})\cdot z_{i-1}\right]^2}{F(z_i)-F(z_{i-1})},$$

$$S_3 = \sum_{i=1}^{k+1}\frac{\left[1_{\{i<k+1\}}f(z_i) - 1_{\{i>1\}}f(z_{i-1})\right]\cdot\left[1_{\{i<k+1\}}f(z_i)\cdot z_i - 1_{\{i>1\}}f(z_{i-1})\cdot z_{i-1}\right]}{F(z_i)-F(z_{i-1})}.$$

Here $1_{\{i<k+1\}}$ and $1_{\{i>1\}}$ are indicator functions, the former being equal to 1 if $i < k+1$

and 0 else. Hence, we have to maximise the function

$$\left[(1,0)\cdot \boldsymbol{M}_{\boldsymbol{Y}}^{-1}(z_1,\ldots,z_k;\boldsymbol{\theta}_0)\cdot(1,0)^T\right]^{-1} = \frac{S_1\cdot S_2 - S_3^2}{S_2} = S_1 - \frac{S_3^2}{S_2}. \tag{A.3}$$

We consider the sum S_1 and note that $f(z_i) = F(z_i)\cdot(1-F(z_i))$ and $F(z_i) = \sum_{j=1}^{i} p_{j,\boldsymbol{\theta}}$:

$$S_1 = \sum_{i=1}^{k+1} \frac{\left[\sum_{j=1}^{i} p_{j,\boldsymbol{\theta}}\cdot\left(1-\sum_{j=1}^{i} p_{j,\boldsymbol{\theta}}\right) - \sum_{j=1}^{i-1} p_{j,\boldsymbol{\theta}}\cdot\left(1-\sum_{j=1}^{i-1} p_{j,\boldsymbol{\theta}}\right)\right]^2}{p_{i,\boldsymbol{\theta}}}. \tag{A.4}$$

The sum in equation (A.4) is considered as a function $g(p_{1,\boldsymbol{\theta}},\ldots,p_{k+1,\boldsymbol{\theta}})$. We show with induction on $k\geq 1$ that the function g is maximised for $p_{1,\boldsymbol{\theta}} = \ldots = p_{k+1,\boldsymbol{\theta}} = \frac{r}{k+1}$ subject to the constraints $p_{1,\boldsymbol{\theta}},\ldots,p_{k+1,\boldsymbol{\theta}} \geq 0$ and $\sum_{i=1}^{k+1} p_{i,\boldsymbol{\theta}} = r$ for arbitrary $0 < r \in \mathbb{R}$. For $k=1$ we have

$$g(p_{1,\boldsymbol{\theta}},p_{2,\boldsymbol{\theta}}) = p_{1,\boldsymbol{\theta}}\cdot(1-p_{1,\boldsymbol{\theta}})^2 + \frac{1}{r-p_{1,\boldsymbol{\theta}}}\cdot\left(r\cdot(1-r) - p_{1,\boldsymbol{\theta}}\cdot(1-p_{1,\boldsymbol{\theta}})\right)^2$$

and the derivative of g with respect to $p_{1,\boldsymbol{\theta}}$ is given by

$$\frac{\partial g(p_{1,\boldsymbol{\theta}},p_{2,\boldsymbol{\theta}})}{\partial p_{1,\boldsymbol{\theta}}} = r\cdot(r-2p_{1,\boldsymbol{\theta}}),$$

which is zero for $p_{1,\boldsymbol{\theta}} = r/2$. Since the second derivative is negative, this is a maximum. We make the inductive step $k\to k+1$. Now we separate the $(k+2)$th summand of the function g:

$$\begin{aligned} g(p_{1,\boldsymbol{\theta}},\ldots,p_{k+2,\boldsymbol{\theta}}) = &\sum_{i=1}^{k+1} \frac{\left[\sum_{j=1}^{i} p_{j,\boldsymbol{\theta}}\cdot\left(1-\sum_{j=1}^{i} p_{j,\boldsymbol{\theta}}\right) - \sum_{j=1}^{i-1} p_{j,\boldsymbol{\theta}}\cdot\left(1-\sum_{j=1}^{i-1} p_{j,\boldsymbol{\theta}}\right)\right]^2}{p_{i,\boldsymbol{\theta}}} \\ &+ \frac{\left[r\cdot(1-r) - (r-p_{k+2,\boldsymbol{\theta}})\cdot(1-r+p_{k+2,\boldsymbol{\theta}})\right]^2}{p_{k+2,\boldsymbol{\theta}}}. \end{aligned}$$

By induction hypothesis the first sum is maximised for $p_{1,\boldsymbol{\theta}} = \ldots = p_{k+1,\boldsymbol{\theta}} = \frac{r-p_{k+2,\boldsymbol{\theta}}}{k+1} =: \tilde{p}$. Plugging in these values we obtain for the first sum:

$$\begin{aligned} &\sum_{i=1}^{k+1} \frac{\left[i\cdot\tilde{p}\cdot(1-i\cdot\tilde{p}) - (i-1)\cdot\tilde{p}\cdot(1-(i-1)\cdot\tilde{p})\right]^2}{\tilde{p}} \\ &= (k+1)\tilde{p} - \frac{1}{3}(k+1)\tilde{p}^3 - 2(k+1)^2\tilde{p}^2 + \frac{4}{3}(k+1)^3\tilde{p}^3. \end{aligned}$$

The first sum is maximised for arbitrary $p_{k+2,\boldsymbol{\theta}}$ and thus the multidimensional optimisation problem is reduced to a one-dimensional one. We have to maximise the function

$\tilde{g}(p_{k+2,\boldsymbol{\theta}}) := g(\tilde{p}, \dots, \tilde{p}, p_{k+2,\boldsymbol{\theta}})$, which is given by

$$\tilde{g}(p_{k+2,\boldsymbol{\theta}}) = \frac{-r^3 + 3r^2 p_{k+2,\boldsymbol{\theta}} - 3rp_{k+2,\boldsymbol{\theta}}^2 + p_{k+2,\boldsymbol{\theta}}^3 + (4r^3 - 6r^2 + 3r - p_{k+2,\boldsymbol{\theta}}^3) \cdot (k+1)^2}{3(k+1)^2}.$$

The derivative is a quadratic polynomial with zeros $\frac{r}{k+2}$ and $-\frac{r}{k} < 0$. Inserting $\frac{r}{k+2}$ in the second derivative we have $\tilde{g}''(\frac{r}{k+1}) = -\frac{6r}{k+1} < 0$ and thus the function $\tilde{g}$ is maximised for $p_{k+2,\boldsymbol{\theta}} = \frac{r}{k+2}$. Hence, the function g is maximised for $p_{i,\boldsymbol{\theta}} = \frac{r}{k+2}$, $i = 1, \dots, k+2$.

With $r = 1$ we can conclude that the sum S_1 is maximised for $p_{1,\boldsymbol{\theta}}^* = \dots = p_{k+1,\boldsymbol{\theta}}^* = \frac{1}{k+1}$. The corresponding cutpoints are given by $z_i^* = \ln\left(\frac{i}{k+1-i}\right)$, $i = 1, \dots, k$. Now we show that the sum S_3 vanishes for these cutpoints:

$$S_3 = (k+1) \cdot \sum_{i=1}^{k+1} \left[1_{\{i<k+1\}} f(z_i^*) - 1_{\{i>1\}} f(z_{i-1}^*)\right] \cdot \left[1_{\{i<k+1\}} f(z_i^*) \cdot z_i^* - 1_{\{i>1\}} f(z_{i-1}^*) \cdot z_{i-1}^*\right]$$

$$= (k+1) \cdot \left(\sum_{i=1}^{k} f(z_i^*)^2 \cdot z_i^* - \sum_{i=2}^{k} f(z_i^*) f(z_{i-1}^*) \cdot (z_i^* + z_{i-1}^*) + \sum_{i=2}^{k+1} f(z_{i-1}^*)^2 \cdot z_{i-1}^* \right) = 0.$$

Since $z_i^* = -z_{k-i}^*$ and f is symmetric, the three sums are equal to zero. We have $-S_3^2/S_2 \leq 0$ and thus this term is maximised for $z_1^*, \dots, z_k^*$. Hence, these cutpoints maximise $S_1 - S_3^2/S_2$. By transformation we obtain the optimal cutpoints $x_1^*, \dots, x_k^*$. □

References

Atkinson, A. C. and Donev, A. N. (1996). *Optimum Experimental Designs.* Clarendon Press, Oxford.

Balakrishnan, N. and Nevzorov, V. B. (2003). *A Primer on Statistical Distributions.* Wiley, Hoboken.

Bickel, P., Buyske, S., Chang, H. and Ying, Z. (2001). On maximizing item information and matching difficulty with ability. *Psychometrika*, Vol. 66, 69-77.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, **24**, 586-602.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey und D. J., Knuth, D. E. (1996). On The Lambert W Function. *Advances in Computational Mathematics*, **5**, 329-359.

Ford, I., Torsney, B. and Wu, C. F. J. (1992). The Use of a Canonical Form in the Construction of Locally Optimal Designs for Non-Linear Problems. *Journal of the Royal Statistical Society, Series B*, **54**, 569-583.

Gunduz, N. and Torsney, B. (2006). Some advances in optimal designs in contingent valuation studies. *Journal of Statistical Planning and Inference*, **136**, 1153-1165.

Nguyen, T. and Torsney, B. (2007). Optimal Cutpoint Determination: The Case of One Point Design. In *mODa 8 - Advances in Model-Oriented Design and Analysis.* López-Fidalgo, J., Rodríguez-Díaz, J., Torsney, B. (eds.), Physica-Verlag Heidelberg, 131-138.

Rabie, H. and Flournoy, N. (2013). Optimal designs for contingent response models with application to toxicity-efficacy studies. *Journal of Statistical Planning and Inference*, **143**, 1371-1379.

Silvey, S. D. (1980). *Optimal Design.* Chapman and Hall, London.