entropy

MDPI

Article

Minimum Message Length Inference of the Exponential Distribution with Type I Censoring

Enes Makalic [1,*] and Daniel Francis Schmidt [2]

[1] Melbourne School of Population and Global Health, The University of Melbourne, Parkville, VIC 3010, Australia
[2] Faculty of Information Technology, Monash University, Clayton, VIC 3168, Australia; daniel.schmidt@monash.edu
* Correspondence: emakalic@unimelb.edu.au; Tel.: +61-3-8344-0860

Abstract: Data with censoring is common in many areas of science and the associated statistical models are generally estimated with the method of maximum likelihood combined with a model selection criterion such as Akaike's information criterion. This manuscript demonstrates how the information theoretic minimum message length principle can be used to estimate statistical models in the presence of type I random and fixed censoring data. The exponential distribution with fixed and random censoring is used as an example to demonstrate the process where we observe that the minimum message length estimate of mean survival time has some advantages over the standard maximum likelihood estimate.

Keywords: minimum message length; exponential distribution; maximum likelihood; survival analysis; censoring

Citation: Makalic, E.; Schmidt, D.F. Minimum Message Length Inference of the Exponential Distribution with Type I Censoring. *Entropy* **2021**, *23*, 1439. https://doi.org/10.3390/e23111439

Academic Editors: Raúl Alcaraz, Luca Faes, Leandro Pardo and Boris Ryabko

Received: 13 September 2021
Accepted: 19 October 2021
Published: 30 October 2021

1. Introduction

In Type I random censoring we observe for each item i either the true survival time $T_i = t_i$ ($t_i > 0$) or the censoring time $C_i = c_i$ ($c_i > 0$), where capital letters are used to denote random variables. The data consists of joint realisations of the random variables $(Y_i = y_i, \Delta_i = \delta_i)$ $(i = 1, \ldots, n)$ where

$$Y_i = \min(T_i, C_i), \tag{1}$$

$$\Delta_i = \mathrm{I}(T_i \leq C_i) = \begin{cases} 1, & \text{if } T_i \leq C_i \text{ (observed survival)} \\ 0, & \text{if } T_i > C_i \text{ (observed censoring).} \end{cases} \tag{2}$$

The censoring time C_i may be fixed (i.e., $C_i = c$ for all $i = 1, \ldots, n$) or a random variable that may depend on other factors (e.g., loss to follow-up). The likelihood function of n observed data points $D = \{(y_1, \delta_1), \ldots, (y_n, \delta_n)\}$ is

$$p(D) = \prod_{i=1}^{n} (p_T(y_i)(1 - F_C(y_i)))^{\delta_i} (p_C(y_i)(1 - F_T(y_i)))^{1-\delta_i}$$

where $p_T(t|\theta)$ and $F_T(t|\theta)$ denote the probability density and the cumulative density function of the random variable T, respectively. Inference about the survival times $(t_1, \ldots, t_n)$ is of key interest in many areas of science and is commonly done by maximizing the likelihood and dropping terms relevant to C only.

This manuscript examines inference of models in the presence of censored data under the minimum message length (MML) framework. MML (see Section 3) is Bayesian technique for model selection and parameter estimation that is based on data compression and key principles of information theory. MML is known to possess strong theoretical properties [1–3] and has previously been successfully applied to a wide range of statistical

models [1]. Here, we demonstrate how MML can be used to infer models under fixed censoring as well as type I random censoring. We use the exponential distribution (see Section 2) as a simple example to demonstrate the key steps and compare the MML estimator to the well-known maximum likelihood estimator in this setting (see Section 2.1). Although MML analysis of the exponential distribution is not new (see, for example, [1,4]), the MML principle has not been applied to any kind of survival data with censoring to date.

The main contributions of this manuscript are to: (i) introduce the MML principle of inductive inference and demonstrate how the Wallace–Freeman MML approximation can used to infer exponential models with type I censored data; (ii) show that the MML estimate of the mean lifetime has some advantages over the usual maximum likelihood estimate for small samples and that it converges to the maximum likelihood estimate for large sample sizes, (iii) incorporate the proposed codelengths for censored exponential distributions into MML finite mixture models allowing for inference of all parameters as well as the number of mixture classes; and (iv) compare the MML principle to the closely related minimum description length principle.

2. Exponential Distribution

Consider the case of a randomly censored exponential parameter studied in [5] where the lifetime data and the censoring data are assumed to be exponentially distributed

$$T_i \sim \text{Exp}(\beta), \quad C_i \sim \text{Exp}(\alpha), \quad i = 1, \ldots, n, \tag{3}$$

and $\alpha, \beta > 0$ denote the mean censoring time and survival time, respectively. Under this model, the joint probability distribution of $(Y_i = y_i, \Delta_i = \delta_i)$ is

$$p(Y_i = y_i, \Delta_i = 1) = p_T(y_i)(1 - F_C(y_i)) \tag{4}$$
$$p(Y_i = y_i, \Delta_i = 0) = p_C(y_i)(1 - F_T(y_i)) \tag{5}$$

where (Y_i, Δ_i) are defined in (1) and (2) respectively. In contrast to random censoring, in fixed censoring an item is observed for a period of time, say $c > 0$, and its actual survival time t_i is known if the item fails before time $t_i \leq c$; otherwise, we only know that the item survived past $t_i > c$. In the case of exponentially distributed survival times, the observed data $(Y = y_i, \Delta_i = \delta_i)$ follows

$$T_i \sim \text{Exp}(\theta), \quad C_i = c, \quad i = 1, \ldots, n, \tag{6}$$

where $c > 0$ is a fixed constant (the follow up period) known a priori. Given n data points $D = \{(y_1, \delta_1), \ldots, (y_n, \delta_n)\}$, the aim is to estimate the unknown mean lifetime survival $\beta > 0$ (random censoring) or $\theta > 0$ (fixed censoring).

2.1. Maximum Likelihood Estimation

The method of maximum likelihood is the most common approach used to obtain parameter estimates in parametric models. Under the censored exponential model, maximum likelihood proceeds by setting the parameter estimate $\hat{\beta}(D)$ to the value that maximises the probability of the data. From (4) and (5), the joint probability of the data D is

$$p(D|\alpha, \beta) = \left(\frac{1}{\beta}\right)^k \left(\frac{1}{\alpha}\right)^{n-k} \exp\left(-\left(\frac{1}{\beta} + \frac{1}{\alpha}\right) \sum_{i=1}^{n} y_i\right), \tag{7}$$

where $k = (\sum_i \delta_i)$ is the number of observed uncensored survival times. Maximizing the likelihood function is equivalent to minimizing the negative log-likelihood function

$$-\log p(D|\alpha, \beta) = k \log \beta + (n - k) \log \alpha + \left(\frac{1}{\beta} + \frac{1}{\alpha}\right) \sum_{i=1}^{n} y_i. \tag{8}$$

Maximum likelihood estimates of the mean survival and censoring times are

$$\hat{\beta}(D)_{\mathrm{ML}} = \frac{1}{k}\sum_{i=1}^{n} y_i, \quad \hat{\alpha}(D)_{\mathrm{ML}} = \frac{1}{n-k}\sum_{i=1}^{n} y_i, \tag{9}$$

respectively. Provided the count of observed survival times $k \in (0, n)$, the maximum likelihood estimates $\hat{\alpha}(D)$ and $\hat{\beta}(D)$ are finite; otherwise, if $k = 0$ or $k = n$, one of the maximum likelihood estimates $\hat{\alpha}(D)$ or $\hat{\beta}(D)$ is infinite. Kim [5] showed that maximum likelihood estimates have infinite mean and variance in this setting. However, the expected value of the maximum likelihood estimate $\hat{\beta}(D)$ is finite if we condition on $k > 0$. Kim [5] further showed that, provided $k \in (0, n)$, the maximum likelihood estimates $\hat{\alpha}(D)$ and $\hat{\beta}(D)$ are unbiased, strongly consistent (without any condition on k) and asymptotically normally distributed.

In the case of Type I censored data with a fixed censoring time $c > 0$, the negative log-likelihood function of the data is

$$-\log p(D|\theta; c) = k\log(\theta) + \frac{1}{\theta}\left(\sum_{i=1}^{n} y_i\delta_i\right) + \frac{c(n-k)}{\theta} \tag{10}$$

where $k = \sum_i \delta_i$ as before. Under fixed censoring, the maximum likelihood estimate of the mean survival time $\hat{\theta}(D)$ is (see, for example, [6])

$$\hat{\theta}(D) = \frac{c(n-k) + \sum_{i=1}^{n} \delta_i y_i}{k}. \tag{11}$$

In case of no censoring (i.e., $k = n$ implying complete data), (11) reduces to $(\sum_i y_i)/n$, which is the usual maximum likelihood estimate for the exponential distribution with complete data. The sampling distribution of (11) is asymptotically normal with mean θ and variance

$$\frac{\theta^2}{n(1-\exp(-c/\theta))} = \frac{\theta^2}{nF_T(c|\theta)}. \tag{12}$$

Conditional upon $k > 0$, Mendenhall and Lehman [7] obtained the exact mean and variance of the maximum likelihood estimate (11)

$$\mathbb{E}\{\hat{\theta}\} = \theta - c\left(\frac{q}{p} - n\mathbb{E}\{k^{-1}\} + 1\right), \quad \mathbb{V}\{\hat{\theta}\} = (nc)^2\mathbb{V}\{k^{-1}\} + (\theta^2 - c^2q/p^2)\mathbb{E}\{k^{-1}\},$$

where

$$p = 1 - \exp(-c/\theta), \quad q = 1 - p, \quad \mathbb{E}\{k^{-a}\} = \frac{1}{1-q^n}\sum_{k=1}^{n}\frac{1}{k^a}\binom{n}{k}p^k q^{n-k},$$

for $a = 1, 2, \ldots$ and $\mathbb{V}\{k^{-1}\} = \mathbb{E}\{k^{-2}\} - (\mathbb{E}\{k^{-1}\})^2$. However, the large sample normal approximation of the distribution of the maximum likelihood estimates is inaccurate and not representative of the behaviour of the estimate in the small to moderate sample size regime [7]. Balakrishnan and Davies [8] further show that the maximum likelihood estimate computed based on a censoring time c' will always produce an estimate which is Pitman closer to the data generating model θ than the maximum likelihood estimate computed with a shorter censoring time $c < c'$. In the next section, we introduce the MML principle of inductive inference (see Section 3) and demonstrate how MML can be used to infer exponential models with censoring (see Section 4).

3. Minimum Message Length

Introduced in the late 1960s by Wallace and Boulton [9], the minimum message length (MML) principle [1,9–11] is a framework for inductive inference based on ideas in information theory and data compression. Under the MML framework, the aim is to

transmit a set of data (a message) from a hypothetical sender to a receiver over a noiseless transmission channel. The MML message is designed to consist of two parts:

1. the *assertion*: an encoding of the model structure and the associated model parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \in \mathbb{R}^p$.
2. the *detail*: a description of the data D using the model $p(D|\boldsymbol{\theta})$ that was specified in the assertion.

The length of the assertion measures the complexity of the model, with complex models requiring longer codelengths compared to simpler models, while the detail captures how well a model fits the data. The length of the two-part message, $I(D, \boldsymbol{\theta})$, is the sum of the length of the assertion, $I(\boldsymbol{\theta})$, and the length of detail, $I(D|\boldsymbol{\theta})$; namely,

$$I(D, \boldsymbol{\theta}) = \underbrace{I(\boldsymbol{\theta})}_{\text{assertion}} + \underbrace{I(D|\boldsymbol{\theta})}_{\text{detail}}. \tag{13}$$

Within the MML framework we seek the model

$$\hat{\boldsymbol{\theta}}(D) = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \{I(D, \boldsymbol{\theta})\} \tag{14}$$

that minimises the length of this message. Due to the two-part nature of the message, MML automatically balances the trade-off between model complexity and the goodness of fit of the model to the data. By measuring the quality of a model in (say) bits, MML is a yardstick that can be universally used to compare models with different parameters and structures.

There exist several approches to computing message lengths (13), with the strict MML procedure (SMML) [1,12] and the MML87 approximation [1,10] being the most widely known. In contrast to the SMML procedure whose construction is known to be NP hard [13], the MML87 approximation is computationally tractable and most widely used in practice. The MML87 codelength approximation to (13) is

$$I_{87}(D, \boldsymbol{\theta}) = \underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2} \log |J_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \frac{p}{2} \log \kappa_p}_{\text{assertion}} + \underbrace{\frac{p}{2} - \log p(D|\boldsymbol{\theta})}_{\text{detail}} \tag{15}$$

where $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the prior distribution for the parameters $\boldsymbol{\theta}$, $|J_{\boldsymbol{\theta}}(\boldsymbol{\theta})|$ is the determinant of the expected Fisher information matrix, $p(D|\boldsymbol{\theta})$ is the likelihood function of the model and κ_p is a quantization constant [14,15] that depends on the number of parameters p. Specifically, for small p we have

$$\kappa_1 = \frac{1}{12}, \quad \kappa_2 = \frac{5}{36\sqrt{3}}, \quad \kappa_3 = \frac{19}{192 \times 2^{1/3}}, \tag{16}$$

while κ_p is well-approximated for large p by [1]:

$$\frac{p}{2}(\log \kappa_p + 1) \approx -\frac{p}{2} \log 2\pi + \frac{1}{2} \log p\pi - \gamma, \tag{17}$$

where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant. The MML87 codelength, evaluated at the minimum, is the shortest codelength of a two-part message that encodes both the model parameters $\boldsymbol{\theta} \in \Theta$ and the data D. The MML87 approximation is known to be invariant under smooth one-to-one reparameterizations of the likelihood function and is asymptotically equivalent to the well-known Bayesian information criterion (BIC) [16] as $n \to \infty$ with $p > 0$ fixed; that is,

$$I_{87}(D, \boldsymbol{\theta}) = -\log p(D|\boldsymbol{\theta}) + \frac{p}{2} \log n + O(1) \tag{18}$$

where the $O(1)$ term depends on the prior distribution, the Fisher information and the number of parameters p. Unlike model selection criteria such as Akaike's information

criterion (AIC) and BIC, MML allows for both parameter estimation and model selection within the same unified framework. Furthermore, in models where the number of parameters grows with n or the sample size is relatively small, the difference between the MML87 codelength and BIC can be substantial. Examples include analysis of multiple short time series, where several measurements are collected over a period of time for a large number of study participants [17], learning finite mixture models [18] and discriminating between Poisson and geometric distributions based on observed data [19]. In the latter example, both the Poisson and geometric distribution have the same number of free parameters so that model selection with BIC is equivalent to choosing the model with the higher likelihood. In contrast, MML87 takes into account the complexity of each distribution [20] and not just the number of parameters, resulting in improved model selection performance for small sample sizes [19].

MML has been successful applied to a wide range of problems (e.g., decision trees [21], factor analysis [22], linear causal models [23], mixture modelling [18,24]) demonstrating excellent parameter estimation properties and model selection performance that is on par or better than commonly used techniques such as Akaike's information criterion (AIC) [25] and the Bayesian information criterion (BIC). A brief tutorial overview of minimum message length can be found in [19].

4. Minimum Message Length Inference of Type I Censored Exponential Data

To encode and transmit censored data $D = \{(y_1, \delta_1), \ldots, (y_n, \delta_n)\}$ between the hypothetical sender and receiver within the MML framework, we have two options:

- Transmit the censoring indicators $(\delta_1, \ldots, \delta_n)$ first and then transmit the lifetime survival data $(y_1, \ldots, y_n)$ given the receiver now knows which of the n data points are censored (see Section 4.1);
- Transmit the censoring indicators and the lifetime data simultaneously (see Section 4.2).

We shall now estimate the MML87 codelength (15) for both the joint and the conditional encoding schemes for the censored exponential distribution setting introduced in Section 2.

4.1. Conditional Encoding of the Data

Under the conditional encoding framework, the sender transmits the censoring indicators δ_i first, and then transmits the lifetime data y_i using the conditional distribution of the data given the observed censoring indicators. The total message length of the data $D = \{(y_1, \delta_1), \ldots, (y_n, \delta_n)\}$ and the parameters $\boldsymbol{\theta}$ with the conditional encoding is

$$I_{87}(D, \boldsymbol{\theta}) = I_{87}(\phi, \boldsymbol{\delta}) + I_{87}(\psi, \mathbf{y}|\boldsymbol{\delta}), \tag{19}$$

where $\theta = \{\phi, \psi\}$ are the model parameters defined below, $I(\phi, \boldsymbol{\delta})$ denotes the message length of the censoring indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$, and $I(\psi, \mathbf{y}|\boldsymbol{\delta})$ denotes the codelength of the survival data $\boldsymbol{y} = (y_1, \ldots, y_n)$, given that the censoring indicators are known to the receiver. From (3), the probability of observing an uncensored datum, say $\phi > 0$, is

$$\phi = P(T_i \leq C_i) = \frac{\alpha}{\alpha + \beta}, \quad (i = 1, \ldots, n), \tag{20}$$

implying that the censoring indicators follow a Bernoulli distribution with probability ϕ; that is, $\delta_i \sim$ Bernoulli(ϕ), or equivalently, k follows the binomial distribution $k \sim$ binomial(ϕ, n).

The MML87 codelength of the binomial distribution was previously derived in [1,18] and is included here for completeness. Briefly, to compute the MML87 codelength (15) we require the Fisher information $J_\phi(\phi)$ and the prior distribution $\pi_\phi(\phi)$ for the probability of observing an uncensored datum. The Fisher information is well-known

$$J_\phi(\phi) = \frac{n}{\phi(1-\phi)}. \tag{21}$$

We assume the prior distribution for the censoring probability ϕ to be the beta distribution ($\phi \sim \text{beta}(a, b)$) with probability density function

$$\pi_\phi(\phi|a,b) = \frac{\phi^{a-1}(1-\phi)^{b-1}}{B(a,b)} \tag{22}$$

where $a, b > 0$ are the shape and scale parameters respectively and $B(a, b)$ is the usual beta function. Substituting (21) and (22) into the MML87 codelength (15) and noting that $\kappa_1 = 1/12$, yields

$$\begin{aligned} I_{87}(\phi,\delta) = -\left(k+a-\tfrac{1}{2}\right)\log\phi - \left(n+b-\tfrac{1}{2}-k\right)\log(1-\phi) + \\ \log B(a,b) + \tfrac{1}{2}(1+\log n - \log 12) \end{aligned} \tag{23}$$

where, as before, $k = (\sum_i \delta_i)$. The codelength (23) is minimised at the MML87 estimate

$$\hat{\phi}_{87}(\delta) = \frac{k+a-1/2}{n+a+b-1}. \tag{24}$$

Note that, in the special case of uniform prior distribution ($a = b = 1$), the MML87 estimate simplifies to

$$\hat{\phi}_{87}(\delta) = \frac{k+1/2}{n+1}. \tag{25}$$

The shortest MML87 codelength for the censoring indicators is therefore given by $I_{87}(\hat{\phi}_{87}, \delta)$. It remains to work out the conditional codelength of the surivival times given the censoring indicators, $I_{87}(\psi, \mathbf{y}|\delta)$.

We note that the conditional likelihood of the lifetime datum y_i is

$$p(y_i|\alpha,\beta,\delta=0) = p(y_i|\alpha,\beta,\delta=1) = \left(\frac{1}{\beta}+\frac{1}{\alpha}\right)\exp\left(-y_i\left(\frac{1}{\beta}+\frac{1}{\alpha}\right)\right), \tag{26}$$

which is the exponential distribution with mean $\psi = (1/\beta + 1/\alpha)^{-1}$; that is,

$$y_i|\delta_i \sim \text{Exp}(\psi), \quad i = 1, \ldots, n. \tag{27}$$

The Fisher information of the exponential distribution is

$$J_\psi(\psi) = \frac{n}{\psi^2}. \tag{28}$$

In terms of the prior distribution for ψ, Schmidt and Makalic [4] consider the conjugate exponential distribution with a hyperparameter ψ_0 that controls the prior mean. Here, we would like an objective prior distribution on the mean ψ that is free of hyperparameters and has heavy tails so that large values of ψ are not penalized too severely. Additionally, our choice of the prior distribution should ideally lead to an easy to compute analytic estimate of ψ. A reasonable option is the half-Cauchy distribution which has heavy tails however it leads to MML estimates that are roots of polynomial functions of $s = \sum_i(y_i)$. Instead, we will use the Fréchet (inverse Weibull) distribution with probability density function

$$\pi_\psi(\psi) = \psi^{-2}\exp(-\psi^{-1}), \quad \psi > 0, \tag{29}$$

which is a type of generalized extreme value distribution and has Cauchy-like heavy tails. Substituting (29), (28) into (3), we obtain the MML87 codelength

$$I_{87}(\psi,\mathbf{y}|\delta) = (n+1)\log\psi + \frac{1}{\psi}\left(1+\sum_{i=1}^{n} y_i\right) + \frac{1}{2}(1+\log n - \log 12). \tag{30}$$

The MML87 estimate of the mean ψ is

$$\hat{\psi}_{87}(\mathbf{y}) = \frac{s+1}{n+1} \tag{31}$$

where, as before, $s = \sum_i y_i$. The MML87 estimate corresponds to the usual maximum likelihood estimate $\hat{\psi}_{\mathrm{ML}}(\mathbf{y}) = s/n$ with one additional data point that has a unit contribution to the mean. The expected mean squared error of the MML87 estimate is

$$\mathbb{E}\{(\hat{\psi}_{87}(\mathbf{y}) - \psi)^2\} = \frac{\psi(\psi(n+1)-2)+1}{(n+1)^2} \tag{32}$$

which dominates the maximum likelihood estimate for

$$\psi > \frac{n}{n+\sqrt{n(2n+1)}}, \quad n > 0. \tag{33}$$

As the sample size n increases, we note that

$$\lim_{n\to\infty}\left\{\frac{n}{n+\sqrt{n(2n+1)}}\right\} = \sqrt{2}-1 \approx 0.414 \tag{34}$$

implying the MML87 estimate dominates maximum likelihood for all $\psi > 0.414$ in terms of expected mean squared error for large n. However, we note that, unlike the MML87 estimate with this choice of prior distribution, the maximum likelihood estimate is invariant to scaling of the data.

Substituting $I_{87}(\hat{\phi}, \boldsymbol{\delta})$ and $I_{87}(\hat{\psi}, \mathbf{y}|\boldsymbol{\delta})$ into (19) yields the total (conditional) codelength of the data. The MML87 estimates of the mean lifetime $\hat{\beta}_{87}(D)$ and censoring time $\hat{\alpha}_{87}(D)$ can be recovered from

$$\alpha \to \frac{\psi}{1-\phi}, \quad \beta \to \frac{\psi}{\phi}, \tag{35}$$

for $\phi \in (0,1)$. Next, we examine how the same message can be encoded using joint encoding of lifetime data and censoring indicators.

4.2. Joint Encoding of the Data

Unlike in the conditional encoding, the sender now transmits the survival times and the indicator variables simultaneously. The negative log-likelihood function of the data $D = \{(y_1, \delta_1), \ldots, (y_n, \delta_n)\}$ is given in (8). The Fisher information in this parameterization is

$$J(\alpha, \beta) = \frac{n^2}{\beta\alpha(\alpha+\beta)^2}. \tag{36}$$

We would like to use prior distributions for α and β that are comparable to those in the conditional coding described in Section 4.1. Noting that $\phi \sim \mathrm{beta}(a, b)$, ψ has the standard Frechet distribution and

$$\phi = \frac{\alpha}{\alpha+\beta}, \quad \psi = \left(\frac{1}{\alpha}+\frac{1}{\beta}\right)^{-1}, \tag{37}$$

the Jacobian of the transformation from $(\phi, \psi) \to (\alpha, \beta)$ is

$$\frac{\alpha\beta}{(\alpha+\beta)^3} \tag{38}$$

implying that a commensurate joint prior distribution for α, β is

$$\pi_{\alpha,\beta}(\alpha, \beta) = \frac{\alpha^{a-2} e^{-\frac{\alpha+\beta}{\alpha\beta}} \beta^{b-2} (\alpha+\beta)^{-a-b+1}}{B(a,b)} \tag{39}$$

where $B(\cdot,\cdot)$ is the beta function. In the special case where ϕ is given a uniform prior (i.e., $a = b = 1$), we have

$$\pi_{\alpha,\beta}(\alpha,\beta) = \frac{e^{-\frac{\alpha+\beta}{\alpha\beta}}}{\alpha\beta(\alpha+\beta)} \tag{40}$$

Substituting (36) and (39) into (15), the MML87 codelength is

$$I_{87}(D,\boldsymbol{\theta}) = k\log\beta + (n-k)\log\alpha + \left(\frac{1}{\alpha}+\frac{1}{\beta}\right)\left(\sum_{i=1}^{n} y_i\right) - \log\pi_{\alpha,\beta}(\alpha,\beta)$$
$$+ \log n - \frac{1}{2}\log\Big(\alpha\beta(\alpha+\beta)^2\Big) + \log\kappa_2 + 1 \tag{41}$$

where $\boldsymbol{\theta} = \{\alpha,\beta\}$ and the quantization constant $\kappa_2 = 5/(36\sqrt{3})$. The MML87 estimates that minimize the codelength (41) are

$$\hat{\alpha}_{87}(D) = \frac{2(s+1)(a+b+n-1)}{(n+1)(2b-2k+2n-1)}, \quad \hat{\beta}_{87}(D) = \frac{2(s+1)(a+b+n-1)}{(n+1)(2a+2k-1)}. \tag{42}$$

If required, the corresponding estimates of ϕ and ψ can be obtained from (37).

4.3. Properties

First we show the the conditional (19) and joint (41) MML codelengths are equivalent up to a constant to be specified below. From Sections 4.1 and 4.2, we note the joint density of (Y,Δ) can be expressed as a product of the binomial $p_\Delta(\delta|\phi)$ and exponential densities $p_Y(y|\psi)$

$$p_{Y,\Delta}(y,\delta|\alpha,\beta) = p_\Delta(\delta|\phi)p_Y(y|\psi) = \Big(\phi^\delta(1-\phi)^{1-\delta}\Big)(\exp(-y/\psi)/\psi)$$

where Y and Δ are independent random variables (see Kim [5] (p. 104)). Consequently, as MML87 is invariant under smooth one-to-one reparameterizations of the sampling model, the MML87 joint codelength (41) and the corresponding conditional codelength (19) are identical (except for the minor efficiency gain in the joint codelength discussed below). Specifically, the relationship between the joint codelength, $I_{87}(\alpha,\beta,D)$ and conditional codelength, $I_{87}(\psi,\phi,D)$, can be expressed as

$$I_{87}(\psi,\phi,D) = I_{87}(\alpha,\beta,D) + \log\left(\frac{3\sqrt{3}}{5}\right) \tag{43}$$

where the term $\log(3\sqrt{3}/5) \approx 0.0385$ arises due to the quantization constant being smaller in higher dimensions since its more efficient to encode multiple parameters simultaneously compared to encoding each parameter independently.

Furthermore, as the MML87 estimate of ϕ is $\hat{\phi} \in (0,1)$ (see (24)), MML87 estimates of the mean survival times (α,β) are finite for all $k \in [0,n]$ in contrast to the corresponding maximum likelihood estimates (9) which are finite for $k \in (0,n)$. As $n \to \infty$, it is well-known that the MML87 estimates are equivalent to the maximum likelihood estimates (see (18)) which implies that the MML87 estimates are similarly asymptotically normally distributed and strongly consistent.

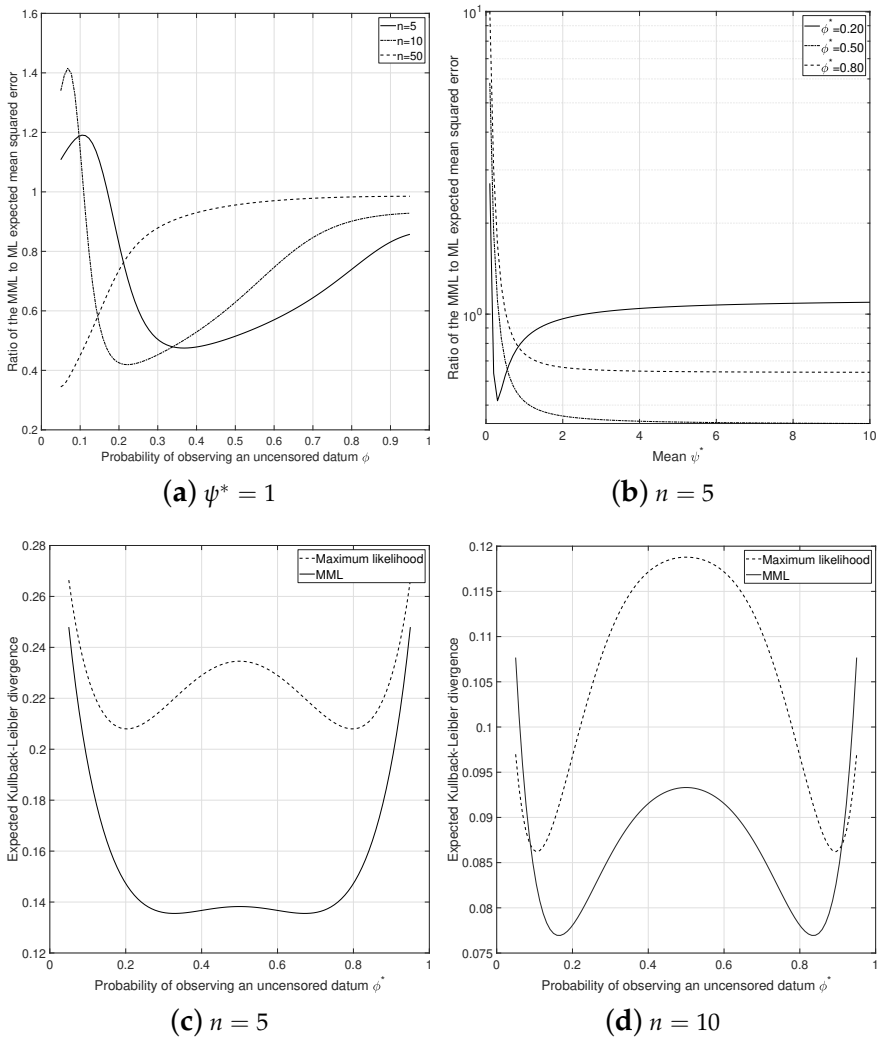
The expected mean square error $\mathbb{E}\{(\hat{\beta}-\beta^*)^2\}$ of the ML and MML87 estimates of β^*, conditional on $k > 0$, is expressible in terms of the generalized hypergeometric function for any $n > 0$. Figure 1 (top) depicts the expected mean squared error between the MML87 and ML estimate of β^*, expressed as a ratio of MML87 to ML with smaller values indicating preference for the MML87 estimate. The expected mean squared error of the MML87 estimate of β^* was generally lower than the corresponding maximum likelihood estimate (except when the true censoring proportion ϕ^* was small) with the biggest difference

observed for small sample sizes, while the two estimates were practically indistinguishable for larger sample sizes $n \geq 100$.

We also compared the MML87 and the ML estimates in terms of the relative entropy or the Kullback–Leibler (KL) divergence, conditional on $k \in (0, n)$. The KL divergence between the true data generating model (α_1, β_1) and the approximating model (α_2, β_2) is

$$D_{\mathrm{KL}}(\alpha_1, \beta_1 || \alpha_2, \beta_2) = \frac{\alpha_1\beta_1(\alpha_2 + \beta_2) + \alpha_2\beta_2\left(\alpha_1 \log\left(\frac{\beta_2}{\beta_1}\right) + \beta_1 \log\left(\frac{\alpha_2}{\alpha_1}\right) - \alpha_1 - \beta_1\right)}{\alpha_2\beta_2(\alpha_1 + \beta_1)},$$

which, as expected, is the sum of the KL divergences between two exponential and two binomial distributions. The KL divergence may be interpreted as the expected amount of extra information required to encode data from (α_1, β_1) using the model (α_1, β_2). The expected KL divergence for both the ML and MML estimators is shown in the bottom of Figure 1, conditional on $k > 0$ and $k < n$. It is clear that for $n = 5$ the MML87 estimate dominates the maximum likelihood estimate in terms of the KL divergence for all $\phi^* \in (0.05, 0.95)$. When the sample size is increased ($n = 10$), the MML87 estimate exhibits smaller KL divergence compared to the ML estimate for all ϕ^* except when $\phi^* \to 0$ or $\phi^* \to 1$ where the maximum likelihood estimate has smaller KL divergence.



(a) $\psi^* = 1$ **(b)** $n = 5$

(c) $n = 5$ **(d)** $n = 10$

Figure 1. Expected mean squared error of β^* and expected KL divergence between the MML87 and ML estimates. Ratio values less than 1 imply that the MML87 estimate has smaller mean squared error in estimating β.

5. Minimum Message Length Inference with Fixed Censoring

Consider now the fixed censoring scenario (6) introduced in Section 2 where the negative log-likelihood function of the data is given in (10). If we wish to encode the data D using the joint MML code (see Section 4.2), we require the negative log-likelihood, the Fisher information and a prior distribution for the mean survival time $\theta > 0$. The negative log-likelihood is given in (10) while the Fisher information for Type I censored data with fixed censoring is:

$$J_\theta(\theta; c) = \frac{n(1 - \exp(-c/\theta))}{\theta^2} = \frac{nF_T(c|\theta)}{\theta^2}, \tag{44}$$

where $F_T(\cdot|\theta)$ is the cumulative distribution function of the survival data T (see Section 2). The reduction in information due to censoring is clearly a function of θ and the cumulative density function of T, with large c resulting in little information loss compared to small c. As expected, as c gets larger

$$\lim_{c\to\infty} J_\theta(\theta; c) = \frac{n}{\theta^2}, \tag{45}$$

which is the usual Fisher information for the exponential distribution with no censoring. The prior distribution for θ is chosen to be the Fréchet ditribution with scale c and probability density function

$$\pi_\theta(\theta; c) = c^{-1}\left(\frac{\theta}{c}\right)^{-2} \exp\left(-\frac{c}{\theta}\right). \tag{46}$$

Substituting (44) and (46) into (15), we obtain the complete MML87 codelength for the joint encoding

$$\begin{aligned} I_{87}(D, \theta) = (k+1)\log(\theta) + \frac{1}{\theta}\left(c((n-k)+1) + \sum_{i=1}^{n} y_i\delta_i\right) \\ + \frac{1}{2}\log\left(\frac{1 - \exp(-c/\theta)}{c^2}\right) + \frac{1}{2}(1 + \log n - \log 12). \end{aligned} \tag{47}$$

Due to the form of the Fisher information, the MML87 estimate of θ that minimizes this codelength is unavailable analytically and must be obtained via numerical optimisation. The maximum likelihood estimate (11) may be used as a starting point for the numerical search.

Consider now the conditional encoding (see Section 4.1) where the probability of observing an uncensored datum, say $\phi > 0$, is

$$\phi = P(T_i \le c) = F_T(c|\theta) = 1 - \exp(-c/\theta), \quad (i = 1, \ldots, n), \tag{48}$$

so that the number of uncensored data points k follows the binomial distribution $k \sim \text{binomial}(\phi, n)$. This implies that the mean survival time can then written as

$$\theta = -c/\log(1 - \phi), \quad \phi \in (0, 1). \tag{49}$$

A naive conditional coding approach proceeds by encoding the censoring indicators following Section 4.1 with codelength (23). To encode $I(\mathbf{y}|\boldsymbol{\delta})$, one would use the conditional probabilities of the lifetime data which are

$$p_{Y|\Delta}(Y_i = y_i|\Delta_i = 1) = \frac{p(T_i = y_i)}{p(T_i \le c)} \quad \text{if } y_i \le c, \tag{50}$$

and

$$p_{Y|\Delta}(Y_i = c|\Delta_i = 0) = 1, \quad p_{Y|\Delta}(Y_i > c|\Delta_i = 0) = 0. \tag{51}$$

for all $i = 1 \ldots, n$. The conditional likelihood of the $k = (\sum_i \delta_i)$ data points is then

$$p(\mathbf{y}|\theta; \boldsymbol{\delta} = 1) = \prod_{i:\delta_i=1} \frac{(1/\theta)\exp(-y_i/\theta)}{1-\exp(-c/\theta)} = -\prod_{i:\delta_i=1} \frac{(1-\phi)^{y_i/c}\log(1-\phi)}{c\phi}. \tag{52}$$

Once the receiver has the censoring data and an estimate of ϕ, they implicitly know θ from (49). The length of the message required to transmit the data $\mathbf{y}$ is

$$\begin{aligned} I_{87}(\theta, \mathbf{y}|\boldsymbol{\delta}) &= k\log\theta + \frac{1}{\theta}\sum_{i:\delta_i=1} y_i + k\log(1-\exp(-c/\theta)) \\ &= -\frac{1}{c}\left(\sum_{i:\delta_i=1} y_i\right)\log(1-\phi) + k\log c\phi - k\log(-\log(1-\phi)) \end{aligned} \tag{53}$$

which is the negative log-likelihood of the data. However, this codelength is inefficient since the probability of censoring $\phi(\theta)$ is not independent of the mean survival time θ. This implies that the precision to which $\phi(\theta)$ is encoded must depend on the lifetime data $\mathbf{y}$, which is not the case in the naive approach where the precision quantum for $\phi(\theta)$ depends on the censoring data $\boldsymbol{\delta}$ only. Consequently, joint MML coding should be used instead of the conditional encoding approach for the fixed censoring setup.

5.1. Example

We observe $n = 20$ items with an exponential life distribution for $c = 150$ h. Out of the 20 items $k = 15$ items fail during the observation period and the sum of their lifetimes (in hours) is $s = \sum_i y_i\delta_i = 835$ [6]. The maximum likelihood estimate of the mean lifetime θ is

$$\hat{\theta}_{\mathrm{ML}}(D) = \frac{150(5)+835}{15} = 105.6 \text{ h}, \tag{54}$$

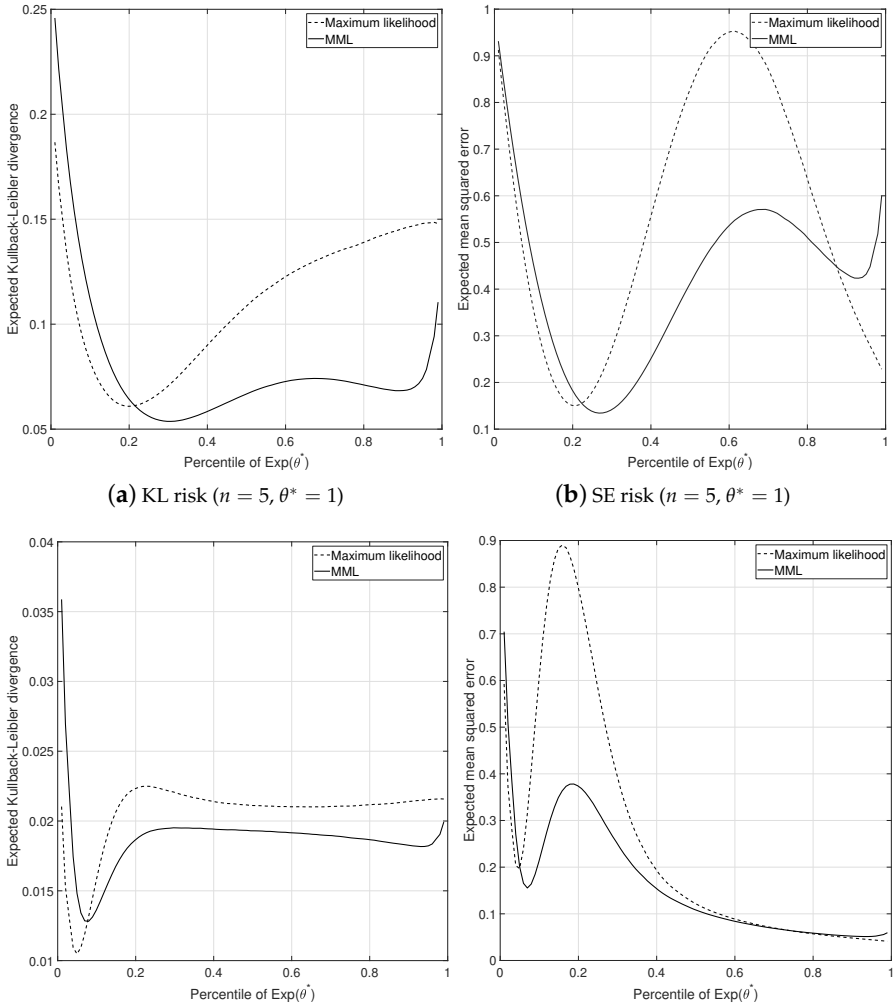
with a negative log-likelihood of 84.9 at the minimum. The MML87 estimate is obtained by a numerical search and is $\hat{\theta}_{87}(D) = 110.1$ h with a codelength of 124.905 bits. The MML87 codelength at the maximum likelihood estimate is 124.924 bits suggesting that there is little difference between the two estimates in this example.

5.2. Properties

To evaluate the performance of the MML87 estimate, we computed the mean squared error risk and the expected Kullback–Leibler (KL) divergence for MML87 and the ML estimates under the data generating model $\theta^* = 1$ and sample sizes $n \in \{5, 25\}$. Since the ML estimate is undefined for $k = 0$, all the results discussed below are conditional on $k > 0$. The KL divergence from the 'true' model θ_1 to the approximating model θ_2 is

$$D_{\mathrm{KL}}(\theta_1||\theta_2) = \left(1 - \frac{\theta_1}{\theta_2} + \log\left(\frac{\theta_1}{\theta_2}\right)\right)\left(\exp\left(-\frac{c}{\theta_1}\right) - 1\right). \tag{55}$$

The results are shown in Figure 2 where the x-axis of each plot is the censoring point c set to the p-th percentile of the data generating model $\mathrm{Exp}(\theta^* = 1)$; for example $c = 0.69$ corresponds to the $p = 0.50$-th percentile of $\mathrm{Exp}(\theta^* = 1)$. It is clear that the MML87 estimate is a reasonable alternative to the maximum likelihood estimate under fixed censoring. For $p \geq 0.20$ (i.e., the 20-th percentile of $\mathrm{Exp}(\theta^* = 1)$) the MML87 estimate dominates the ML estimate in terms of KL risk, while for $p < 0.20$ the estimates are very similar for both sample sizes tested. In terms of the expected mean squared error, the MML87 estimate dominates the ML estimate for all $0.20 \leq p \leq 0.80$ when $n = 5$; for $n = 25$, the estimates are indistinguishable for $p < 0.1$ and $p > 0.69$ and the MML87 estimate again dominates ML for all $0.1 < p < 0.5$.



(**a**) KL risk ($n = 5, \theta^* = 1$) (**b**) SE risk ($n = 5, \theta^* = 1$)

(**c**) KL risk ($n = 25, \theta^* = 1$) (**d**) SE risk ($n = 25, \theta^* = 1$)

Figure 2. Expected Kullback-Leibler (KL) divergence and squared error (SE) risk of the maximum likelihood and MML87 estimates for $n = 5$ (**top**) and $n = 25$ (**bottom**) data points generated from model $\theta^* = 1$. The x-axis on all plots denotes the censoring point c and is set to a percentile of Exp(θ^*).

6. Discussion

This manuscript has demonstrated how minimum message length can be used to infer data with censoring information. Specifically, we have derived MML87 codelengths for the exponential distribution with fixed censoring and random type I censoring. Although information theoretic universal models for the exponential distribution, including those corresponding to MML codes, are known [4], this is the first time MML has been applied to censored data.

The MML87 codelength for the exponential distribution with censoring provides a new means of parameter estimation as well as model selection. In terms of parameter estimation, the MML87 estimate of the mean lifetime θ under type I censoring described in this paper has some advantages over the usual maximum likelihood estimate for small sample sizes. First, the MML87 estimate is defined for all proportions of censoring unlike the maximum likelihood estimate which does not exist when all observations are censored; i.e., $k = (\sum_i \delta_i) = 0$. In addition, the MML87 estimate has on average lower mean squared error risk and lower KL divergence from the data generating model for a wide range of censoring proportions.

In the case of random censoring, the MML87 estimate is available in closed-form while for fixed censoring, it can only be obtained by numerical optimisation. Although

the experiments in the manuscript utilised heavy-tailed prior distributions for the scale parameter as recommended in, for example, [26], the Bayesian nature of MML allows for information prior information to be incorporated directly into the estimation process. The effect of the prior distribution in the examples considered here is expected to be negligible for medium to large sample sizes.

Importantly, the proposed MML87 codelengths can also be used to discriminate between competing models (e.g., exponential vs lognormal) and offer some advantages over the well-known BIC model selection approach. BIC only considers the sample size and the number of parameters when measuring model complexity. In contrast, MML takes into account not just the number of parameters, but also the complexity of the distribution (i.e., the number of random data strings that are fitted well by the distribution). As the sample size $n \to \infty$, the MML87 codelength converges to the BIC and therefore inherits the favourable asymptotic properties of BIC, such as model selection consistency.

The codelengths derived in this manuscript are extendable to MML inference of other censored data types, such as the Weibull and the lognormal distribution, and can be incorporated into more complex models as shown in the next section.

6.1. Clustering Survival Data

To demonstrate the applicability of the codes derived in the manuscript, we implemented the MML87 codes into a Matlab software package for inference of finite mixture models. Our software, called Matlab Snob, features mixture models with categorical data (e.g., multinomial distribution), count data (e.g., geometric, Poisson and negative binomial distributions), continuous data (e.g., normal, Laplace, gamma and Weibull distributions) and survival data (type I fixed and random censored exponential distribution). As a demonstration of Matlab Snob, we used two publicly available survival data sets: (1) Rossi et al.'s criminal recidivism data [27], and (2) survival from malignant melanoma [28].

The crime data was recently analyzed in [29] using variational Bayes estimated finite mixture models. For clustering we used all $n = 432$ observations and the following seven attributes: (1) financial aid (no, yes), (2) full-time work experience before incarceration (no, yes), (3) marital status at time of release (married, not married), (4) released on parole (no, yes), (5) number of convictions prior to current incarceration, (6) age in years at time of release and (7) week of first arrest after release (73.6% censored). This is an example of fixed censoring as all censored observations were censored at 52 weeks. We modelled the categorical attributes using a multinomial distribution, number of convictions was modelled with a negative binomial distribution, while a Gaussian distribution was used for age at time of release. For the week of arrest, we used the exponential distribution model with fixed type I censored data (see Section 5).

The melanoma data set consists of $n = 205$ patients from Denmark who were diagnosed with malignant melanoma. Five attributes were used for clustering: (1) sex (male, female), (2) ulcer (present, absent), (3) age at diagnosis in years, (4) tumour thickness in mm, and (5) censored survival time in years (65.3% censored). Sex and ulcer were modelled via multinomial distributions, while age and tumour thickness were modelled with univariate Gaussian distributions. For the survival time, we used an exponential distribution with random type I censoring (see Section 4) and combined death due to melanoma and death due to other causes as the primary outcome of interest.

Clustering results for the Crime and the Melanoma data sets with Matlab Snob are shown in Table 1. First, since Matlab Snob learns finite mixture models using the MML87 codelength approximation, the same framework is used to estimate all model parameters as well as select the number of classes. For the Crime data, the model with three classes had the smallest codelength, while two classes were selected for the Melanoma data set. We observe that all the classes are relatively well differentiated in terms of average survival time. In the case of the Crime data set, class 1 had the shortest average time to arrest ($\theta = 119$ weeks) and consists of younger individuals (mean age 20.7 years, std. dev. 2.1 years) who are primarily unmarried and have no full-time work experience before

incarceration. In contrast, class 3 comprised older individuals (mean age 36.3 years, std. dev. 4.9 years) 82% of which had full-time work experience, was estimated to have longest average time to arrest ($\theta = 345.9$ weeks). For the melanoma data set, class 1 was estimated to have the shortest average survival time ($\beta = 7.3$ years) and consisted of individuals diagnosed at an older age (mean: 57.3 years, std. dev. 17.1 years) with larger tumours (mean: 5.4 mm, std. dev. 3.4 mm). In contrast, patients assigned to class 2 were diagnosed at a younger age, had smaller tumours and were estimated to have longer survival time on average ($\beta = 37.0$ years).

Table 1. MML finite mixture models for Crime and Melanoma data. The attribute modelling censored survival time is seven in the Crime data set and five in the Melanoma data set.

Data	**Class**	**Attributes**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Crime	1	(50%, 50%)	(73%, 27%)	(3%, 97%)	(42%, 58%)	(r: 2.0, p: 0.4)	(μ: 20.7, σ: 2.1)	(**θ: 119.0**)
	2	(55%, 45%)	(15%, 85%)	(23%, 77%)	(28%, 72%)	(r: 13.4, p: 0.8)	(μ: 24.9, σ: 3.3)	(**θ: 249.3**)
	3	(40%, 60%)	(16%, 84%)	(18%, 82%)	(51%, 49%)	(r: 2.3, p: 0.5)	(μ: 36.3, σ: 4.9)	(**θ: 345.9**)
Melanoma	1	(43%, 57%)	(17%, 83%)	(μ: 57.3, σ: 17.1)	(μ: 5.4, σ: 3.4)	(**α: 12.0**, **β: 7.3**)	–	–
	2	(73%, 27%)	(80%, 20%)	(μ: 49.5, σ: 15.8)	(μ: 1.4, σ: 0.9)	(**α: 8.1**, **β: 37.0**)	–	–

The Matlab Snob clustering software is freely available for download from the Mathworks Filexchange website (ID: 72310) and will be extended to incorporate other survival distributions in the future (eg, Weibull and lognormal distribution). We note that the MML87 codelengths for type I censored exponentially distributed data derived in this paper can also be used in decision tree modelling [21,30,31]. For example, one could represent the leaves of the decision tree with a censored exponential distribution and use MML to infer an optimal decision tree for a data set.

6.2. Minimum Message Length and Minimum Description Length

Minimum message length is closely related to minimum description length (MDL), an inductive inference principle independently developed by Rissanen and colleagues [32–35]. Like MML, the MDL principle is rooted in information theory and, given a data set, seeks a model that would result in the shortest encoding of the data. A recent and popular version of the MDL principle is the normalized maximum likelihood (NML) code which says that the codelength for data $\mathbf{y}$ with respect to model class $\mathcal{M}$ parameterised by models $\boldsymbol{\theta} \in \mathbb{R}^p \in \mathcal{M}$ is

$$-\log p_{\mathrm{NML}}(\mathbf{y}|\mathcal{M}) = -\log p(\mathbf{y}|\hat{\theta}(\mathbf{y}), \mathcal{M}) + \log \sum_{\mathbf{x}} p(\mathbf{x}|\hat{\theta}(\mathbf{x}), \mathcal{M}) \tag{56}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the p parameters and the sum in the second term is taken over the entire data space; we replace the sum with an integral in the case of continuous data. The first term in the NML codelength is the negative log-likelihood of the data evaluated at the maximum likelihood estimate, while the second term represents the parametric complexity of the model class and measures how well models $\boldsymbol{\theta} \in \mathcal{M}$ within the model class $\mathcal{M}$ approximate random data sequences. In particular, a high parametric complexity says that a large number of data sequences can be well-approximated by models within the class. In contrast, the parametric complexity of a simple model that can only well-approximate a few data sequences will tend to be small.

Rissanen [33] derives an asymptotic approximation for the NML codelength which is accurate for medium to large sample sizes:

$$-\log p_{\mathrm{NML}}(\mathbf{y}|\mathcal{M}) = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathcal{M}) + \log \int_{\Theta} \sqrt{|J_1(\boldsymbol{\theta})|} + \frac{p}{2} \log\left(\frac{n}{2\pi}\right) + o(1) \tag{57}$$

where $J_1(\cdot)$ is the per-sample Fisher information matrix. Mera et al. [36] derive a somewhat sharper approximation to the NML codelength using Riemannian geometry tools and apply their new approximation to principal component analysis. Rissanen further shows that, like the MML87 codelength, the NML codelength reduces to the well-known Bayesian information criterion (BIC) in the limit as the sample size $n \to \infty$. Unfortunately, in the case of the exponential distribution with or without type I censoring, the parametric complexity is infinite for both the exact NML codelength and the asymptotic approximation. To circumvent the problem of infinite parametric complexity, one may consider the restricted approximate normalized maximum likelihood (ANML), the two-part ANML or the objective Bayesian code, among others [37].

Although there exist many similarities in the approaches to inference between MML and MDL, there are some important differences which we summmarize below:

- MDL relies on the maximum likelihood estimator and does not offer new means for parameter estimation;
- MDL is decidedly non-Bayesian avoiding the use of any (subjective or objective) prior information;
- MDL nominates the *model class* $\mathcal{M}$ that would result in the shortest encoding of the data and does not does not infer a fully specified model;
- while MML minimises the expected (average) codelength of the data with respect to the marginal data distribution, MDL minimizes the worst-case codelength relative to the ideal code.

In addition to the NML code, other MDL codes exist including the sequential NML code [38] and the conditional NML distribution [34], among others. Clearly, both MML and MDL approaches to inductive inference have merit, and if used correctly, will result in excellent model selection performance as shown in a wide range of applications. A more detailed discussion of MML and MDL similarities and differences can be found in [39] and [1] (pp. 413–415).

Author Contributions: Methodology, E.M. and D.F.S.; Software, E.M. and D.F.S.; Writing—original draft, E.M.; Writing—review & editing, D.F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
BIC	Bayesian information criterion
KL	Kullback–Leibler
MDL	Minimum description length
MML	Minimum message length
ML	Maximum likelihood
NML	Normalized maximum likelihood

References

1. Wallace, C.S. *Statistical and Inductive Inference by Minimum Message Length*, 1st ed.; Information Science and Statistics; Springer: Berlin/Heidelberg, Germany, 2005.
2. Wallace, C.S. False oracles and SMML estimators. In *Proceedings of the International Conference on Information, Statistics and Induction in Science*; World Scientific: Singapore, 1996; pp. 304–316.
3. Wallace, C.S.; Dowe, D.L. Minimum Message Length and Kolmogorov Complexity. *Comput. J.* **1999**, *42*, 270–283. [CrossRef]
4. Schmidt, D.F.; Makalic, E. Universal Models for the Exponential Distribution. *IEEE Trans. Inf. Theory* **2009**, *55*, 3087–3090. [CrossRef]

5. Kim, J.S. Asymptotic properties of the maximum likelihood estimator of a randomly censored exponential parameter. *Commun. Stat. Theory Methods* **1986**, *15*, 3637–3646. [CrossRef]
6. Bartholomew, D.J. The Sampling Distribution of an Estimate Arising in Life Testing. *Technometrics* **1963**, *5*, 3. [CrossRef]
7. Mendenhall, W.; Lehman, E.H. An Approximation to the Negative Moments of the Positive Binomial Useful in Life Testing. *Technometrics* **1960**, *2*, 227–242. [CrossRef]
8. Balakrishnan, N.; Davies, K.F. Pitman closeness results for Type-I censored data from exponential distribution. *Stat. Probab. Lett.* **2013**, *83*, 2693–2698. [CrossRef]
9. Wallace, C.S.; Boulton, D.M. An information measure for classification. *Comput. J.* **1968**, *11*, 185–194. [CrossRef]
10. Wallace, C.S.; Freeman, P.R. Estimation and inference by compact coding. *J. R. Stat. Soc. (Ser. B)* **1987**, *49*, 240–252. [CrossRef]
11. Wallace, C.S.; Dowe, D.L. Refinements of MDL and MML Coding. *Comput. J.* **1999**, *42*, 330–337. [CrossRef]
12. Wallace, C.; Boulton, D. An invariant Bayes method for point estimation. *Classif. Soc. Bull.* **1975**, *3*, 11–34.
13. Farr, G.E.; Wallace, C.S. The complexity of Strict Minimum Message Length inference. *Comput. J.* **2002**, *45*, 285–292. [CrossRef]
14. Conway, J.H.; Sloane, N.J.A. *Sphere Packing, Lattices and Groups*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 1998; p. 703.
15. Agrell, E.; Eriksson, T. Optimization of lattices for quantization. *IEEE Trans. Inf. Theory* **1998**, *44*, 1814–1828. [CrossRef]
16. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
17. Schmidt, D.F.; Makalic, E. Minimum message length analysis of multiple short time series. *Stat. Probab. Lett.* **2016**, *110*, 318–328. [CrossRef]
18. Wallace, C.S.; Dowe, D.L. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Stat. Comput.* **2000**, *10*, 73–83. [CrossRef]
19. Wong, C.K.; Makalic, E.; Schmidt, D.F. Minimum message length inference of the Poisson and geometric models using heavy-tailed prior distributions. *J. Math. Psychol.* **2018**, *83*, 1–11. [CrossRef]
20. Balasubramanian, V. MDL, Bayesian inference, and the geometry of the space of probability distributions. In *Advances in Minimum Description Length: Theory and Applications*; Grünwald, I.J.M., Pitt, M.A., Eds.; MIT Press: Cambridge, MA, USA, 2005; pp. 81–99.
21. Wallace, C.S.; Patrick, J.D. Coding Decision Trees. *Mach. Learn.* **1993**, *11*, 7–22. [CrossRef]
22. Wallace, C.S.; Freeman, P.R. Single-Factor Analysis by Minimum Message Length Estimation. *J. R. Stat. Soc. (Ser. B)* **1992**, *54*, 195–209. [CrossRef]
23. Wallace, C.S.; Korb, K.B. Learning linear causal models by MML sampling. In *Causal Models and Intelligent Data Management*; Gammerman, A., Ed.; Springer: Berlin/Heidelberg, Germany, 1999; pp. 89–111.
24. Schmidt, D.F.; Makalic, E. Minimum Message Length Inference and Mixture Modelling of Inverse Gaussian Distributions. In *AI 2012: Advances in Artificial Intelligence*; Lecture Notes in Computer Science; Thielscher, M., Zhang, D., Eds.; Springer: Berlin/Heidelberg, Germany; Sydney, Australia, 2012; Volume 7691, pp. 672–682.
25. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control.* **1974**, *19*, 716–723. [CrossRef]
26. Polson, N.G.; Scott, J.G. On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Anal.* **2012**, *7*, 887–902. [CrossRef]
27. Rossi, P.; Berk, R.A.; Lenihan, K.J. *Money, Work, and Crime: Some Experimental Results*; Academic Press: Cambridge, MA, USA, 1980.
28. Andersen, P.K.; Borgan, Ø.; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes*; Springer: Berlin, Germany, 2012. [CrossRef]
29. Kohjima, M.; Matsubayashi, T.; Toda, H. Variational Bayes for Mixture Models with Censored Data. In *Machine Learning and Knowledge Discovery in Databases*; Springer International Publishing: Berlin, Germany, 2019; pp. 605–620. [CrossRef]
30. Bou-Hamad, I.; Larocque, D.; Ben-Ameur, H. A review of survival trees. *Stat. Surv.* **2011**, *5*, 44–71. [CrossRef]
31. Dauda, K.A.; Pradhan, B.; Shankar, B.U.; Mitra, S. Decision tree for modeling survival data with competing risks. *Biocybern. Biomed. Eng.* **2019**, *39*, 697–708. [CrossRef]
32. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
33. Rissanen, J. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **1996**, *42*, 40–47. [CrossRef]
34. Rissanen, J.; Roos, T. Conditional NML Universal Models. In Proceedings of the 2007 Information Theory and Applications Workshop (ITA-07), San Diego, CA, USA, 29 January–2 February 2007; IEEE Press: Piscataway, NJ, USA, 2007; pp. 337–341. (Invited Paper).
35. Rissanen, J. Optimal Estimation. *Inf. Theory Newsl.* **2009**, *59*, 1–20
36. Mera, B.; Mateus, P.; Carvalho, A.M. On the minmax regret for statistical manifolds: The role of curvature. *arXiv* **2020**, arXiv:2007.02904.
37. de Rooij, S.; Grünwald, P. An empirical study of minimum description length model selection with infinite parametric complexity. *J. Math. Psychol.* **2006**, *50*, 180–192. [CrossRef]
38. Roos, T.; Rissanen, J. On sequentially normalized maximum likelihood models. In Proceedings of the 1st Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08), Tampere, Finland, 18–20 August 2008; (Invited Paper).
39. Baxter, R.A.; Oliver, J. *MDL and MML: Similarities and Differences*; Technical Report TR 207; Department of Computer Science, Monash University: Clayton, Australia 1994.