Estimating the sampling distribution of the maximum likelihood estimator of the parameters of a series system given a sample of masked system failures

Alexander Towell
atowell@siue.edu

Abstract

First, the *parametric family* of an abstract *series system* is derived. Then, the *asymptotic sampling distribution* (*multivariate normal*) of a *maximum likelihood estimator* of the true parameter is derived conditioned on the *information* given by a random sample of *masked system failure times*. The asymptotic sampling distributions of statistics that are functions of the parameter follow as a result. Finally, these results are applied to series systems in which component lifetimes are *exponentially* distributed, which lead to closed formulas on a set of *minimally sufficient statistics*.

Contents

1 Introduction

Empirical modeling involves a blending of substantive subject matter and statistical information. In our case, the substantive information includes assumptions like a series system and the α-masked candidate set models described later. Such information suggests the relevant variables and simplifies learning from data (in our case, maximum likelihood estimation).

In contrast, statistical information stems from the chance regularities in data, which are statistical model attempts to "adequately" capture.

Discuss latent or hidden variables.

Draw a graph in prob. model section. What do we know, how we know it, etc.

We consider series systems consisting of some fixed number of components with uncertain lifetimes. We desire a mathematical model of the system that may be used to predict the failure time of the system and its component cause. However, since the system lifetime and the component cause of failure is uncertain, we are interested in probabilistically modeling the system lifetime, e.g., there is a 75% chance that component 3 will cause a system failure in the next 3 years.

We assume the system belongs to some parametric family but whose true parameter index is unknown. Furthermore, we also assume there is a sample of *masked system failure times* that carry information about the true parameter index.

Using the information in the sample, we employ the frequentist approach of repeated experiments which induces a sampling distribution on the maximum likelihood point estimator of the fixed but unknown true parameter index where *a priori* any value in the parameter space is equally likely. The point estimator converges to a multivariate normal sampling distribution with a mean given by the true parameter index and a variance that is inversely proportional to sample size.

2 Probabilistic model

A probabilistic model describes how observable data about the system is generated. In what follows, we describe our model and observable data.

A *set* is an unordered collection of distinct elements. The set of all subsets of a set is called the *powerset*. The powerset of $\mathcal{X}$ is denoted by $2^{\mathcal{X}}$. An interval in $2^{\mathbb{R}}$ is a set of all the real numbers lying between two indicated real numbers, e.g.,

$$(a, b] \coloneqq \{\, x \in \mathbb{R} \mid x \geq a \wedge x \leq b \,\} \in 2^{\mathbb{R}} \,. \tag{2.1}$$

Definition 2.1. *A probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*, where* $\mathrm{P}\colon 2^{\Omega} \mapsto [0,1]$ *is the probability set function,* Ω *is the* sample space*, and* $\mathcal{F}$ *is the set of measurable events (subsets of* Ω*).*

A random variable is a measurable function defined on a probability space.

Definition 2.2. *Given a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*, a random variable* X *is a function of type* $\Omega \mapsto \texttt{codom}(\mathrm{X})$*.*

The range (or image) of X is also called the space of X.

For a random variable X, we use $\mathrm{X} = x$ to denote the subset containing all elements $e \in \Omega$ that X maps to the value of x.

2.1 Series system

The system consists of m components. We make the following assumption about the state of the components.

Assumption 1. *The* j^{th} *component is initially in a non-failed state and permanently enters a failed state at some uncertain time* $t_j > 0$ *for* $j = 1, \ldots, m$*.*

Components have uncertain lifetimes as given by the following definition.

Definition 2.3. *The lifetime of the* j^{th} *component is given by the continuous random variable* $\mathrm{T}_j \in (0, \infty)$ *for* $j = 1, \ldots, m$*.*

We make the following assumption about the joint distribution of the component lifetimes.

Assumption 2. *The random variables* $\mathrm{T}_1, \ldots, \mathrm{T}_m$ *are mutually independent.*

The system is a series system.

Assumption 3. *The system is in a non-failed state if all* m *components are in a non-failed state, otherwise the system is in a failed state.*

A system in which at least k of the components must be in a non-failed state for the system to be in a non-failed system is denoted a k-out-of-m system, thus a series system is an m-out-of-m system.

The uncertain lifetime of the system is given by the following definition.

Theorem 2.1. *The lifetime of the system is given by the random variable*

$$\mathrm{S} = \min(\mathrm{T}_1, \ldots, \mathrm{T}_m) \tag{2.2}$$

with a sample space $(0, \infty)$*, where* $\mathrm{T}_1, \ldots, \mathrm{T}_m$ *are the component lifetimes.*

Proof. A series system fails whenever any component fails, therefore its lifetime is equal to the lifetime of the component with the minimum lifetime. □

2.2 Probability distribution functions and parametric families

As random variables, it is not certain which values $\mathrm{S}, \mathrm{T}_1, \ldots, \mathrm{T}_m$ will realize. However, their respective probability distribution functions quantify the uncertainty. The cumulative distribution function is given by the following definition.

Definition 2.4. *Let* X *be some random variable. The probability that* $\mathrm{X} \leq x$ *is given by its cumulative distribution function, denoted by*

$$F_{\mathrm{X}}(x) = \mathrm{P}[\mathrm{X} \leq x]\,. \tag{2.3}$$

The probability density function is given by the following definition.

Definition 2.5. *Let* X *be some continuous random variable. The* relative likelihood *that* $\mathrm{X} = x$ *is given by its probability density function, denoted by*

$$f_{\mathrm{X}}(x) = \frac{\mathrm{d}}{\mathrm{d}t} F_{\mathrm{X}}(t)\,. \tag{2.4}$$

By the Second Fundamental Theorem of Calculus, the probability that a continuous random variable $\mathrm{X} \in (a, b]$, $a \leq b$, is given by

$$\mathrm{P}[a < \mathrm{X} \leq b] = F_{\mathrm{X}}(b) - F_{\mathrm{X}}(a) = \int_a^b f_{\mathrm{X}}(s)\mathrm{d}s\,. \tag{2.5}$$

By these definitions, the relative likelihood that the system will fail at time t is $f_{\mathrm{S}}(t)$ and probability the j^{th} component will before time t is $F_{\mathrm{T}_j}(t)$.

Notation. *We simplify the notation and let* F_j *denote* F_{T_j} *and* f_j *denote* f_{T_j}*. Column vectors are denoted by boldface lower case letters, e.g.,* $\mathbf{x}$*, and row vectors are denoted by the transpose of column vectors, e.g.,* $\mathbf{x}^{\mathsf{T}}$*. Matrices are denoted by boldface upper case greek letters, e.g.,* $\boldsymbol{A}$*. The* j^{th} *column of* $\boldsymbol{A}$ *is denoted by* $\boldsymbol{A}_j$ *(or* $[\boldsymbol{A}]_j$*) and the* (j,k)*-th element of* $\boldsymbol{A}$ *is denoted by* $\boldsymbol{A}_{j\,k}$ *(or* $[\boldsymbol{A}]_{j\,k}$*).*

We restrict our attention to parametrized families of probability distribution functions. The system lifetime has a probability density function that is a member of the following parametric family.

Definition 2.6. *The system lifetime has a probability density function that is a member of the parametric family of probability density functions denoted by*

$$\mathcal{F}_{\boldsymbol{\Theta}} = \left\{ f_{\mathrm{S}}\left(\,\cdot \,|\, \boldsymbol{\Theta}\right) : \boldsymbol{\Theta} \in \mathbb{R}^{q\times m} \right\}\,, \tag{2.6}$$

where $\boldsymbol{\Theta}_j \in \boldsymbol{\Omega}$ *corresponds to the parameter index of the* j^{th} *component lifetime.*

Particular values of $\boldsymbol{\Theta}$ index particular probability density functions within the $\mathcal{F}_{\boldsymbol{\Theta}}$-family. The true probability density function of the system lifetime is given by the following definition.

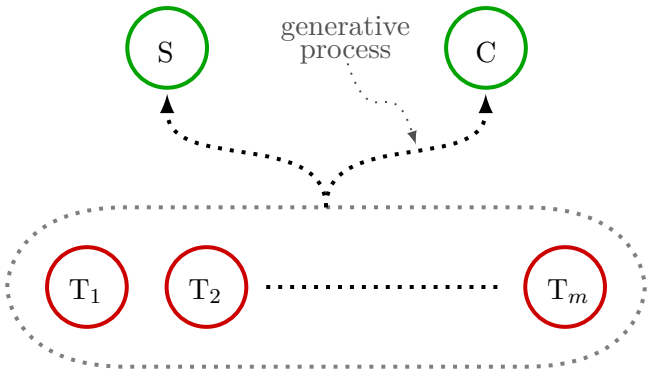
Definition 2.7. *The distribution of the system lifetime has a true parameter index denoted by* $\boldsymbol{\Theta}^{\star}$*, and we say that*

$$\mathrm{S} \sim f_{\mathrm{S}}(\,\cdot \,|\, \boldsymbol{\Theta}^{\star})\,. \tag{2.7}$$

By these definitions, it follows that

$$\mathrm{T}_j \sim f_j\Big(\,\cdot \,|\, \boldsymbol{\Theta}_j^{\star}\Big)\,. \tag{2.8}$$

Figure 1: Graphical model of random variables where $\mathrm{S} = \min\{\mathrm{T}_1, \mathrm{T}_2 \ldots, \mathrm{T}_m\}$ and masked component failure C is generated by some process that is a function of $\boldsymbol{\Theta}^\star$.



2.3 α-masked component failure

By theorem 2.1, if a system failure occurs at time t, then one of the components failed at time t.[1]

Definition 2.8. *The discrete random variable indicating the component responsible for a system failure is denoted by* K *with a sample space* $\mathcal{K} = \{1, \ldots, m\}$. *That is,* $\mathrm{K} = j$ *indicates that* $\mathrm{T}_j < \mathrm{T}_k$ *for* $k \in \mathcal{K} \setminus \{j\}$.

We may not be certain about which component caused a system failure, but we may be able to narrow the component cause to some *subset*. We denote these a *masked* component failure as given by the following definition.

Definition 2.9. *Let* C *denote the random masked component failure with a sample space* $2^{\mathcal{K}}$.

We suppose there is some process that, informed by some degree of knowledge about the components, generates a masked component failure which *plausibly*[2] contains the failed component as depicted in fig. 1.

A masked component failure represents the notion of a diagnostician who is able to, with varying degrees of success, isolate the failed component to some subset of the system. We consider a particular generative model given by the following definition.

Definition 2.10. *An* α*-masked component failure* C *contains the failed component with probability* $\alpha \cdot 100\%$, *where* α *is denoted its* accuracy. *Let* A *denote the random accuracy of a masked component failure with a sample space* $\mathcal{A} \subseteq [0, 1]$. *The distribution of a random masked component failure* C *given* $\mathrm{A} = \alpha$ *is denoted a random* α*-masked component failure.*

The accuracy of a masked component failure could theoretically convey information about the system. As a simplification, we make the following assumption.

[1] Only a single component can cause a system failure since two or more continuous random variables cannot realize the same value.

[2] The random masked component failures carry information about the true parameter index $\boldsymbol{\Theta}^\star$ by being dependent in some way on the component lifetimes.

Assumption 4. *The random accuracy* A *of masked component failures is independent of the random system lifetime* S *and the random component cause* K *and therefore does not carry information about the true parameter index* $\boldsymbol{\Theta}^\star$*.*

A random masked component failure may have an uncertain cardinality. We model this uncertainty with the random variable given by the following definition.

Definition 2.11. *Let* $\mathrm{W} = |\mathrm{C}|$ *denote the random cardinality of a random masked component failure with a sample space* $\{1, \ldots, m-1\}$*.* W *may never realize* m *since that does not isolate a proper subset of the components and it may never realize* 0 *since that does not estimate the component cause of failure.*

Note that if $\mathrm{W} = m$ were to be realized, contrary to the above definition, then a consistent point estimator for the lifetime distribution S is possible but no consistent estimator of the component lifetimes $\mathrm{T}_1, \ldots, \mathrm{T}_m$ is possible. In **??**, we provide a point estimator in such a situation that is inconsistent but minimizes the *mean squared error*.

The cardinality of a masked component failure could theoretically convey information about the system. As a simplification, we make the following assumption.

Assumption 5. *The random cardinality* W *of masked component failures is independent of the random system lifetime* S *and the random component cause* K *and therefore does not carry information about the true parameter index* $\boldsymbol{\Theta}^\star$*.*

We make the following simplifying assumption.

Assumption 6. *The random masked component failure* C *conditioned* K *is independent of* S*.*

The α-masked component model is somewhat naive since C given K is conditionally independent of S. By assumption 6, an α-masked component failure carries information about $\boldsymbol{\Theta}^\star$ through its dependency on K.

The indicator function and the k-subset functions are needed for subsequent material and are given by the following definitions.

Definition 2.12 (Indicator function)**.** *Given a subset* $\mathcal{A}$ *of a set* $\mathcal{B}$*, the indicator function* $\mathbb{1}_{\mathcal{A}} \colon \mathcal{B} \mapsto \{0, 1\}$ *is equal to* 1 *if* $x \in \mathcal{A}$ *and* 0 *otherwise.*

Definition 2.13. *The subsets of* $\mathcal{A}$ *that are of cardinality* k *is defined as*

$$[\mathcal{A}]^k \coloneqq \{\, \mathcal{X} \in \mathcal{A} \mid |\mathcal{A}| = k \,\}, \tag{2.9}$$

which has a cardinality

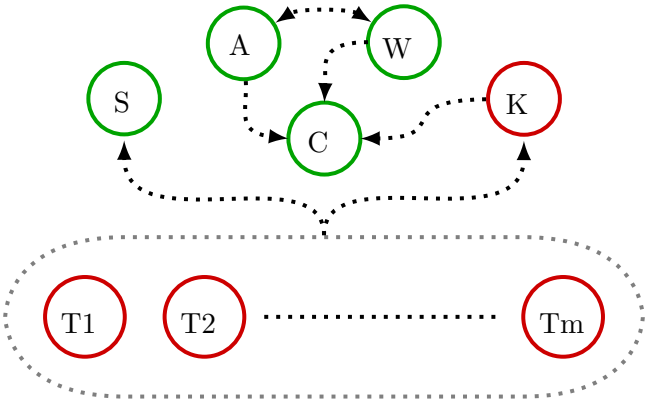
$$\binom{|\mathcal{A}|}{k}. \tag{2.10}$$

The distribution of candidate sets conditioned on a specific cardinality is given by the following definition.

Definition 2.14. *The random candidate set* C *conditioned on* $\mathrm{W} = w$ *has a sample space given by*

$$\mathcal{W}_w = \left[2^{\mathrm{K}}\right]^w. \tag{2.11}$$

Figure 2: Graphical model of random variables where $\mathrm{S} = \min\{\mathrm{T}_1, \mathrm{T}_2 \ldots, \mathrm{T}_m\}$, C is a random α-masked candidate set of cardinality w, K is the random component failure, W is the cardinality of random candidate sets, and A is the random α.



By assumptions 4 and 5, the distributions of W and A is not dependent on the true parameter index $\boldsymbol{\Theta}^\star$. Any inferences about the parametric model will be *independent* of W and A except in the sense that conditioning on larger α-masked component failures or less accurate accuracy α generates estimators with greater variance. This, of course, is quite important, but the fact that a particular W or A is realized says nothing by itself about $\boldsymbol{\Theta}^\star$.

In general, a sample of masked system failures may have variable sized candidate sets. In **??**, we derive the sampling distribution of the maximum likelihood estimator of this kind of sample using the inverse-variance weighted mean.

The *primary* data point is the *masked system failure* as given by the following definition.

Definition 2.15. *An* α-masked system failure *consists of a system failure time* t*, a corresponding set of failed candidate components* $\mathcal{C}$*, and an expected accuracy* α *of the* $\mathcal{C}$*, denoted by* $\langle \mathrm{t}, \mathcal{C}, \alpha \rangle$*.*

Assumption 7. *A random* α-masked system failure *is a jointly distributed random system lifetime* S*, random failed candidate set* C*, and random* α*-accuracy* A*.*

Assumption 8. *A random sample of* n masked system failures*, denoted by*

$$\mathbf{M_n} = \left(\langle \mathrm{S}_1, \mathrm{C}_1, \mathrm{A}_1 \rangle, \ldots, \langle \mathrm{S}_n, \mathrm{C}_n, \mathrm{A}_n \rangle\right) , \tag{2.12}$$

consists of n *independent and identically jointly distributed pairs of system failure times and component sets.*

Assumption 9. *The only information we have about the system and its true parameter index* $\boldsymbol{\Theta}^\star$ *is given by observing a random sample of* n *masked component failures,*

$$\left(\langle \mathrm{S}_1 = \mathrm{t}_1, \mathrm{C}_1 = \mathcal{C}_1, \mathrm{A}_1 = \alpha_1 \rangle, \ldots, \langle \mathrm{S}_n = \mathrm{t}_n, \mathrm{C}_n = \mathcal{C}_n, \mathrm{A}_n = \alpha_n \rangle\right) . \tag{2.13}$$

Given the assumed model, the object of statistical interest is the (unknown) true parameter index $\boldsymbol{\Theta}^\star$. Provided an estimate of $\boldsymbol{\Theta}^\star$, any characteristic that is a function of $\boldsymbol{\Theta}^\star$ can be estimated

as a result, thus the primary objective is to use the *information* in a sample of n masked system failures to estimate $\boldsymbol{\Theta}^{\star}$ with some quantifiable uncertainty.

3 Derivation of parametric probability distribution functions

Chapter 2 described the probabilistic model. In what follows, we derive the parametric functions that are entailed by the model.

3.1 System lifetime

The complement of the cumulative distribution function is the reliability function and is given by the following definition.

Definition 3.1. *Let* X *be some random variable. The probability that* $\mathrm{X} > x$ *is given by its reliability function, denoted by*

$$R_{\mathrm{X}}(x) = \mathrm{P}[\mathrm{X} > x] = 1 - F_{\mathrm{X}}(x) \ . \tag{3.1}$$

The reliability function of the system lifetime is given by the following theorem.

Theorem 3.1. *The system lifetime has a reliability function given by*

$$R_{\mathrm{S}}\left(\mathrm{t}|\Theta^{\star}\right) = \prod_{j=1}^{m} R_j\left(\mathrm{t}|\Theta_{\mathrm{j}}^{\star}\right) \ , \tag{3.2}$$

where R_j *is the reliability function of the* j^{th} *component.*

Proof. By definition 3.1, the reliability function is given by

$$R_{\mathrm{S}}\left(\mathrm{t}|\Theta^{\star}\right) = \mathrm{P}[\mathrm{S} > \mathrm{t}] \, . \tag{a}$$

By eq. (2.2), $\mathrm{S} = \min(\mathrm{T}_1, , \ldots , \mathrm{T}_m)$. Performing this substitution yields

$$R_{\mathrm{S}}(\mathrm{t}|\Theta^{\star}) = \mathrm{P}\big[\min(\mathrm{T}_1, , \ldots , \mathrm{T}_{\mathrm{m}}) > \mathrm{t}\big] \, . \tag{b}$$

For the minimum to be larger than t, every component must be larger than t, leading to the equivalent equation

$$R_{\mathrm{S}}(\mathrm{t}|\Theta^{\star}) = \mathrm{P}[\mathrm{T}_1 > \mathrm{t}, \ldots , \mathrm{T}_{\mathrm{m}} > \mathrm{t}] \, . \tag{c}$$

By assumption 2, the component lifetimes are independent, thus by the axioms of probability

$$R_{\mathrm{S}}\left(\mathrm{t}|\Theta^{\star}\right) = \mathrm{P}[\mathrm{T}_1 > \mathrm{t}] \cdots \mathrm{P}[\mathrm{T}_m > \mathrm{t}] \, . \tag{d}$$

By **??**, the probability that $\mathrm{T}_j > \mathrm{t}$ is equivalent to the reliability function of the j^{th} component lifetime evaluated at t. Performing these substitutions results in

$$R_{\mathrm{S}}\left(t|\boldsymbol{\Theta}^{\star}\right) = \prod_{j=1}^{m} R_j(t|\boldsymbol{\Theta}_j^{\star}) \, . \tag{e}$$

□

The probability density function of the system's lifetime is given by the following theorem.

Theorem 3.2. *The lifetime of the m-out-of-m system has a probability density function given by*

$$f_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) = \sum_{j=1}^{m} \left(f_j(t|\boldsymbol{\Theta}_j^{\star}) \prod_{\substack{k=1 \\ k\neq j}}^{m} R_k(t|\boldsymbol{\Theta}_k^{\star}) \right) , \tag{3.3}$$

where R_k and f_k are respectively the reliability and probability density functions of the k^{th} component.

Proof. By **??**,

$$f_{\mathrm{S}}\left(t|\boldsymbol{\Theta}^{\star}\right) = -\frac{\mathrm{d}}{\mathrm{d}t} R_{\mathrm{S}}\left(t|\boldsymbol{\Theta}^{\star}\right) . \tag{a}$$

By theorem 3.1, this is equivalent to

$$f_{\mathrm{S}}\left(t|\boldsymbol{\Theta}^{\star}\right) = -\frac{\mathrm{d}}{\mathrm{d}t} \prod_{j=1}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) . \tag{b}$$

By the product rule, this is equivalent to

$$\begin{aligned} f_{\mathrm{S}}(t) = \frac{\mathrm{d}}{\mathrm{d}t} R_1\left(t|\boldsymbol{\Theta}_1^{\star}\right) \prod_{j=2}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) \\ - R_1\left(t|\boldsymbol{\Theta}_1^{\star}\right) \frac{\mathrm{d}}{\mathrm{d}t} \prod_{j=2}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) . \end{aligned} \tag{c}$$

By **??**, we can substitute $-\frac{\mathrm{d}}{\mathrm{d}t} R_1\left(t|\boldsymbol{\Theta}_1^{\star}\right)$ with $f_1\left(t|\boldsymbol{\Theta}_1^{\star}\right)$, resulting in

$$\begin{aligned} f_{\mathrm{S}}(t) = f_1(t) \prod_{j=2}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) \\ - R_1\left(t|\boldsymbol{\Theta}_1^{\star}\right) \frac{\mathrm{d}}{\mathrm{d}t} \prod_{j=2}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) . \end{aligned} \tag{d}$$

Applying the product rule again results in

$$\begin{aligned} f_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) = f_1\left(t|\boldsymbol{\Theta}_1^{\star}\right) \prod_{j=2}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) \\ + f_2\left(t|\boldsymbol{\Theta}_2^{\star}\right) \prod_{\substack{j=1 \\ j\neq 2}}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) \\ - R_1\left(t|\boldsymbol{\Theta}_1^{\star}\right) R_2\left(t|\boldsymbol{\Theta}_2^{\star}\right) \frac{\mathrm{d}}{\mathrm{d}t} \prod_{j=3}^{m} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) . \end{aligned} \tag{e}$$

We see a pattern emerge. Applying the product rule $m-1$ times results in

$$\begin{aligned} f_{\mathrm{S}}\left(t|\boldsymbol{\Theta}^{\star}\right) = \sum_{j=1}^{m-1} f_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) \prod_{\substack{k=1 \\ k\neq j}}^{m} R_k\left(t|\boldsymbol{\Theta}_k^{\star}\right) \\ - \prod_{j=1}^{m-1} R_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) \frac{\mathrm{d}}{\mathrm{d}t} R_m\left(t|\boldsymbol{\Theta}_m^{\star}\right) \\ = \sum_{j=1}^{m} f_j\left(t|\boldsymbol{\Theta}_j^{\star}\right) \prod_{\substack{k=1 \\ k\neq j}}^{m} R_k\left(t|\boldsymbol{\Theta}_k^{\star}\right) . \end{aligned} \tag{f}$$

□

The failure rate function simplifies some of the subsequent material and is given by the following definition.

Definition 3.2. *A random variable* X *with a probability density function* f_{X} *and reliability function* R_{X} *has a failure rate function given by*

$$h_{\mathrm{X}}(t) = \frac{f_{\mathrm{X}}(t)}{R_{\mathrm{X}}(t)}\,. \tag{3.4}$$

The failure rate function of the system lifetime is given by the following theorem.

Theorem 3.3. *The system lifetime has a failure rate function given by*

$$h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) = \sum_{j=1}^{m} h_j(t|\boldsymbol{\Theta}_j^{\star}) \tag{3.5}$$

where h_j *is the failure rate function of the* j^{th} *component lifetime.*

Proof. By eq. (3.4), the failure rate function for the system is given by

$$h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) = \frac{f_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star})}{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star})}\,. \tag{a}$$

Plugging in these parametric functions results in

$$h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) = \frac{\sum_{j=1}^{m}\left(f_j(t|\boldsymbol{\Theta}_j^{\star})\prod_{\substack{k=1\\k\neq j}}^{m} R_k(t|\boldsymbol{\Theta}_k^{\star})\right)}{\prod_{j=1}^{m} R_j(t|\boldsymbol{\Theta}_j^{\star})}\,, \tag{b}$$

which can be simplified to

$$h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) = \sum_{j=1}^{m} f_j(t|\boldsymbol{\Theta}_j^{\star})/R_j(t|\boldsymbol{\Theta}_j^{\star}) \tag{c}$$

$$= \sum_{j=1}^{m} h_j(t|\boldsymbol{\Theta}_j^{\star})\,. \tag{d}$$

□

3.2 Component cause of system failure

According to definition 2.8, the component responsible for a system failure is given by the discrete random variable K. The joint likelihood that component k is the cause of a system failure occurring at time t is given by the following theorem.

Theorem 3.4. *The joint probability density function of random variable* K *and random system lifetime* S *is given by*

$$f_{\mathrm{K,S}}(k,t|\boldsymbol{\Theta}^{\star}) = f_k(t|\boldsymbol{\Theta}^{\star})\prod_{\substack{j=1\\j\neq k}}^{m} R_j(t|\boldsymbol{\Theta}^{\star})\,, \tag{3.6}$$

where f_k *and* R_k *are respectively the probability density and reliability functions of the* k^{th} *component.*

Proof. See appendix A for a detailed proof. However, a quick justification follows. The likelihood that component k is the cause of a system failure at time t is equivalent to the likelihood that component k fails at time t and the other components do not fail before time t. According to assumption 2, the random variables $\mathrm{T}_1, \ldots, \mathrm{T}_m$ are mutually independent, therefore the likelihood is the product of the probability density function $f_k(t|\boldsymbol{\Theta}_k^\star)$ and the reliability functions $R_j(t|\boldsymbol{\Theta}_j^\star)$ for $j \in \mathcal{K} \setminus \{k\}$. □

By definition 2.8, if a system failure occurs at time t, then each component has a particular probability of being the cause as given by the following theorem.

Theorem 3.5. *The conditional probability mass function $p_{\mathrm{K|S}}$ maps each component $k \in \mathcal{K}$ to the probability that component k is the cause of the system failure at time t and is given by*

$$p_{\mathrm{K|S}}\left(k|t, \boldsymbol{\Theta}^\star\right) = \frac{h_k\left(t|\boldsymbol{\Theta}_k^\star\right)}{h_{\mathrm{S}}\left(t|\boldsymbol{\Theta}^\star\right)}\,, \tag{3.7}$$

where h_j and h_{S} are respectively the failure rate functions of the j^{th} component and the system.

Proof. By the axioms of probability,

$$p_{\mathrm{K|S}}(k|t, \boldsymbol{\Theta}^\star) = \frac{f_{\mathrm{K,S}}(k, t|\boldsymbol{\Theta}^\star)}{f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}\,, \tag{a}$$

thus we wish to show that

$$\frac{h_k(t|\boldsymbol{\Theta}_k^\star)}{h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} = \frac{f_{\mathrm{K,S}}(k, t|\boldsymbol{\Theta}^\star)}{f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}\,. \tag{b}$$

By theorem 3.3, the system's failure rate function is given by

$$h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) = \sum_{j=1}^{m} h_j(t|\boldsymbol{\Theta}_j^\star)\,. \tag{3.5 revisited}$$

Performing this substitution results in

$$\frac{h_k(t|\boldsymbol{\Theta}_k^\star)}{h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} = \frac{h_k(t|\boldsymbol{\Theta}_k^\star)}{\sum_{j=1}^{m} h_j(t|\boldsymbol{\Theta}_j^\star)}\,. \tag{c}$$

By definition 3.2, the failure rate function of the p^{th} component is given by

$$h_p(t|\boldsymbol{\Theta}_p^\star) = \frac{f_p(t|\boldsymbol{\Theta}_p^\star)}{R_p(t|\boldsymbol{\Theta}_p^\star)}\,. \tag{d}$$

Performing this substitution results in

$$\frac{h_k(t|\boldsymbol{\Theta}_k^\star)}{h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} = \frac{f_k(t|\boldsymbol{\Theta}_k^\star)}{R_k(t|\boldsymbol{\Theta}_k^\star)} \cdot \left[\underbrace{\frac{f_1(t|\boldsymbol{\Theta}_1^\star)}{R_1(t|\boldsymbol{\Theta}_1^\star)} + \cdots + \frac{f_m(t|\boldsymbol{\Theta}_m^\star)}{R_m(t|\boldsymbol{\Theta}_m^\star)}}_{A}\right]^{-1}\,. \tag{e}$$

To combine the fractions labeled A in the above equation, we make their respective denominators the same, resulting in

$$\frac{h_k(t|\boldsymbol{\Theta}_k^\star)}{h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} = \underbrace{\frac{f_k(t|\boldsymbol{\Theta}_k^\star)}{R_k(t|\boldsymbol{\Theta}_k^\star)}}_{B} \cdot \frac{\overbrace{\prod_{j=1}^{m} R_j(t|\boldsymbol{\Theta}^\star)}^{C}}{\sum_{j=1}^{m}\left[\prod_{\substack{p=1\\p\neq j}}^{m} R_p(t|\boldsymbol{\Theta}_p^\star)\right] f_j(t|\boldsymbol{\Theta}_j^\star)} \,. \tag{f}$$

Dividing the part of the above equation labeled C in the numerator by the part labeled B in the denominator results in

$$\frac{h_k(t|\boldsymbol{\Theta}_k^\star)}{h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} = \frac{f_k(t|\boldsymbol{\Theta}_k^\star)\prod_{\substack{j=1\\j\neq k}}^{m} R_j(t|\boldsymbol{\Theta}_j^\star)}{\sum_{j=1}^{m} f_j(t|\boldsymbol{\Theta}_j^\star)\prod_{\substack{p=1\\p\neq j}}^{m} R_p(t|\boldsymbol{\Theta}_p^\star)} \,. \tag{g}$$

By eq. (3.3), the denominator in the above equation is the density $f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)$ and, by eq. (3.6), the numerator is the joint density $f_{\mathrm{K,S}}(k,t|\boldsymbol{\Theta}^\star)$. Performing these substitutions results in

$$\frac{h_k(t|\boldsymbol{\Theta}_k^\star)}{h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} = \frac{f_{\mathrm{K,S}}(k,t|\boldsymbol{\Theta}^\star)}{f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} \,. \tag{h}$$

□

Corollary 3.5.1. *The joint probability density function of random variable* K *and random system lifetime* S *is given by*

$$f_{\mathrm{K,S}}(k,t|\boldsymbol{\Theta}^\star) = h_k(t|\boldsymbol{\Theta}_k^\star) R_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) \,. \tag{3.8}$$

The proof of eq. (3.8) immediately follows from eqs. (3.6) and (3.7).

3.3 α-masked component failures

The series system consists of m components with random lifetimes $\mathrm{T}_1, \ldots, \mathrm{T}_m$. These random variables are *latent* (unobservable) and thus, for instance, we cannot know with certainty from a sample of masked system failures which component failed.

The α-candidate sets have a conditional probability distribution given by the following theorem.

Theorem 3.6. *The random α-candidate set* C *conditioned on* W, K, *and* A *has a probability distribution given by*

$$p_{\mathrm{C|K,W,A}}(\mathcal{C}|k,w,\alpha) = \frac{\alpha}{\binom{m-1}{w-1}} \mathbb{1}_{\mathcal{C}}(k) + \frac{1-\alpha}{\binom{m-1}{w}} \mathbb{1}_{\overline{\mathcal{C}}}(k) \tag{3.9}$$

with a sample space $\mathcal{W}_w$.

Proof. Suppose we condition on $\mathrm{K} = k$, i.e., the k^{th} -component failed. By **??**, a random α-candidate set C contains k with probability α and does *not* contain k with probability $1-\alpha$.

We partition the sample space $\mathcal{W}_w \coloneqq [\mathcal{K}]^w$ into the set of sets containing k and its complement. There are ${}^{m-1}C_{w-1}$ candidate sets that contain k since we may place k in the "first" position and of the remaining $w-1$ positions uniformly choosing any of the remaining $m-1$ components. There are ${}^{m-1}C_w$ candidate sets that do not contain k since, by removing k from the set of m choices, uniformly choosing w of the remaining $m-1$ components.

By definition, the outcomes in the partition containing k occur with probability α and we assign uniform probability $({}^{m-1}C_{w-1})^{-1}$ to any such outcome. By the product rule the joint probability that a random α-candidate set contains k is

$$\frac{\alpha}{\binom{m-1}{w-1}}\,. \tag{a}$$

Similarly, the outcomes in the partition not containing k occur with probability $1-\alpha$ and we assign uniform probability $({}^{m-1}C_{w})^{-1}$ to any such outcome. By the product rule the joint probability that a random α-candidate set does not contain k is

$$\frac{1-\alpha}{\binom{m-1}{w}}\,. \tag{b}$$

□

Remark. *The α-candidate set model is somewhat naive since if $\mathbb{1}_{\mathcal{C}_1}(k) = \mathbb{1}_{\mathcal{C}_2}(k)$ then*

$$p_{\mathrm{C|K,S,W,A}}(\mathcal{C}_1|k,t_1,w,\alpha) = p_{\mathrm{C|K,S,W,A}}(\mathcal{C}_2|k,t_2,w,\alpha)\,. \tag{3.10}$$

That is, α-masked component failures that include the failed component are equally likely to be chosen and C *given* K *is conditionally independent of the system failure time* S*. A more informative generative model of masked component failures may preferentially include components that are more likely to fail at the given system failure time.* △

The random variable K is *latent*. In a masked system failure sample, the only observable is the candidate set, the system failure time, and the expected accuracy α of the α-masked candidate set. The joint probability density function of C and S conditioned on W and A is given by the following theorem.

Theorem 3.7. *The conditional joint probability density function of random variable* C *and random system lifetime* S *given* $\mathrm{W} = w$ *is given by*

$$f_{\mathrm{C,S|W,A}}(\mathcal{C},t|w,\alpha,\boldsymbol{\Theta}^\star) = \sum_{k=1}^{m} p_{\mathrm{C|K,W,A}}(\mathcal{C}|k,w,\alpha) f_{\mathrm{K,S}}(k,t|\boldsymbol{\Theta}^\star) \tag{3.11}$$

with a sample space $\mathcal{W}_w \times (0,\infty)$. Equation (3.11) may be rewritten as

$$f_{\mathrm{C,S|W,A}}(\mathcal{C},t|w,\alpha,\boldsymbol{\Theta}^\star) = \frac{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}{\binom{m-1}{w}}\left((1-\alpha)h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) - \left(1-\alpha\tfrac{m}{w}\right)\textstyle\sum_{k\in\mathcal{C}} h_k(t|\boldsymbol{\Theta}^\star_k)\right). \tag{3.12}$$

Proof. By the laws of probability, the joint distribution of C, K and S conditioned on W and A may be written as

$$f_{\mathrm{C,K,S|W,A}}(\mathcal{C},k,t|w,\alpha,\boldsymbol{\Theta}^\star) = f_{\mathrm{C|K,S,W,A}}(\mathcal{C}|k,t,w,\alpha,\boldsymbol{\Theta}^\star) f_{\mathrm{K,S|W,A}}(k,t|w,\alpha,\boldsymbol{\Theta}^\star)\,. \tag{a}$$

We observe that C conditioned on K, W, and A is independent of S and the true parameter index $\boldsymbol{\Theta}^\star$ and the joint distribution of S and K is independent of W and A. Thus, we simplify the joint probability density function to

$$f_{\mathrm{C,K,S|W,A}}(\mathcal{C},k,t|w,\alpha,\boldsymbol{\Theta}^\star) = p_{\mathrm{C|K,W,A}}(\mathcal{C}|k,w,\alpha) f_{\mathrm{K,S}}(k,t|\boldsymbol{\Theta}^\star)\,. \tag{b}$$

The marginal joint distribution of C and S given W is calculated by summing the joint probability distribution over K which has a sample space $\{1,\ldots,m\}$. □

The *optimal* case is provided by the 1-masked candidate set as given by the following corollary.

Corollary 3.7.1. *The joint probability density of* C *and* S *conditioned on* $\mathrm{W} = w$ *and* $\alpha = 1$ *is given by*

$$f_{\mathrm{C,S|W,A}}(\mathcal{C}, t|w, \alpha = 1, \boldsymbol{\Theta}^\star) = \frac{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}{\binom{m-1}{w-1}} \sum_{k \in \mathcal{C}} h_k(t|\boldsymbol{\Theta}^\star) \tag{3.13}$$

with a sample space $\mathcal{W}_w \times (0, \infty)$.

Suppose the candidate sets are chosen *randomly*, which is equivalent to the $\frac{w}{m}$-masked candidate set model.

Corollary 3.7.2. *The joint probability density of* C *and* S *conditioned on* $\mathrm{W} = w$ *and* $\alpha = \frac{w}{m}$ *is given by*

$$f_{\mathrm{C,S|W,A}}(\mathcal{C}, t|w, \alpha = w/m, \boldsymbol{\Theta}^\star) = \frac{f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}{\binom{m}{w}} \tag{3.14}$$

with a sample space $\mathcal{W}_w \times (0, \infty)$.

Proof. Intuitively, since random candidate sets of w components (out of m) are randomly chosen, any of the $\binom{m}{w}$ subsets of size w are equally likely and thus the probability mass of each candidate set is just $\binom{m}{w}^{-1}$. Taking the product of the density for the system lifetime S and $\binom{m}{w}^{-1}$ yields the result. However, a more rigorous proof is given by substituting α with $\frac{w}{m}$ in eq. (3.12) and simplying, yielding the intermediate result

$$f_{\mathrm{C,S|W,A}}(\mathcal{C}, t|w, \alpha, \boldsymbol{\Theta}^\star) = \frac{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}{\binom{m-1}{w}} \frac{m-w}{m} h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) \tag{a}$$

$$= \frac{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}{\binom{m}{w}} h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)\,. \tag{b}$$

By **??**, $f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) = R_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)$. Making this substitution yields the result. □

As intuition suggests, candidate sets that are randomly chosen carry no information about the true parameter index. Therefore, the set of informative α-masked candidate set models for estimating $\boldsymbol{\Theta}^\star$ is indexed by the interval $\frac{w}{m} < \alpha \leq 1$.

Remark. *Conversely, suppose we have an* adversarial *diagnostician who generates* α*-masked candidate sets where* $\alpha < \frac{w}{m}$*. Then, the random candidate sets serve as* misinformation*. If it is known or estimated that the candidate sets are misinforming, then we may estimate the answer to a related but different question, "What is* $\boldsymbol{\Theta}^\star$ *not likely to be?" While not nearly as informative, the only case in which no information about* $\boldsymbol{\Theta}^\star$ *is conveyed is when candidate sets are randomly selected, i.e.,* $\alpha = \frac{w}{m}$*.* △

Corollary 3.7.3. *The conditional probability mass function* $p_{\mathrm{C|S,W,A}}$ *maps each candidate set* $\mathcal{C}$ *to the probability that the given candidate set contains the component responsible for the system failure at time* t*. By the axioms of probability,*

$$p_{\mathrm{C|S,W,A}}(\mathcal{C}|t, w, \alpha, \boldsymbol{\Theta}^\star) = \frac{(1-\alpha) h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) - (1 - \alpha\frac{m}{w}) \sum_{j \in \mathcal{C}} h_j(t|\boldsymbol{\Theta}_j^\star)}{\binom{m-1}{w} h_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)} \mathbb{1}_{\mathcal{W}_w}(\mathcal{C})\,, \tag{3.15}$$

where h_{S} *and* h_j *are the failure rate functions of the system and the* j^{th} *component respectively.*

Proof. Suppose it is given that $\mathrm{W} = w$ and $\mathrm{A} = \alpha$. By assumption 5 and **??** and the axioms of probability, we may rewrite the joint distribution of C and S as

$$f_{\mathrm{C,S|W,A}}(\mathcal{C}, t|w, \alpha, \boldsymbol{\Theta}^\star) = p_{\mathrm{C|S,W,A}}(\mathcal{C}|t, w, \alpha, \boldsymbol{\Theta}^\star) f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)\,, \tag{a}$$

which may be rearranged to

$$p_{\mathrm{C|S,W,A}}(\mathcal{C}|t, w, \alpha, \boldsymbol{\Theta}^\star) = \frac{f_{\mathrm{C,S|W,A}}(\mathcal{C}, t|w, \alpha, \boldsymbol{\Theta}^\star)}{f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)}\,, \tag{b}$$

where $f_{\mathrm{C,S|W,A}}$ is the joint probability density function given by corollary 3.7.1. Making this substitution and further simplifying yields the result. □

By assumption 5, the random lifetime S and the cardinality of the random candidate set $\mathrm{W} = |\mathrm{C}|$ are independent, thus by the axioms of probability the joint density of C, S, and W is given by

$$f_{\mathrm{C,S,W|A}}(\mathcal{C}, t, w|\alpha, \boldsymbol{\Theta}^\star) = p_{\mathrm{C|S,W,A}}(\mathcal{C}|t, w, \alpha, \boldsymbol{\Theta}^\star) p_{\mathrm{W|A}}(w|\alpha) f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star)\,. \tag{3.16}$$

Corollary 3.7.4. *When $\alpha = 1$, $p_{\mathrm{C|S,W,A}}$ simplifies to*

$$p_{\mathrm{C|S,W,A}}(\mathcal{C}|t, w, \alpha = 1, \boldsymbol{\Theta}^\star) = \frac{\mathbb{1}_{\mathcal{W}_w}(\mathcal{C})}{\binom{m-1}{w-1}} \frac{\sum_{j \in \mathcal{C}} h_j\left(t|\boldsymbol{\Theta}_j^\star\right)}{h_{\mathrm{S}}\left(t|\boldsymbol{\Theta}^\star\right)}\,, \tag{3.17}$$

where h_{S} and h_j are the failure rate functions of the system and the j^{th} component respectively.

Corollary 3.7.5. *The random component failure* K *given* S *is conditionally independent of* C, W, *and* A, *i.e.,*

$$p_{\mathrm{K|C,S,W,A}}(k|\mathcal{C}, t, w, \alpha, \boldsymbol{\Theta}^\star) = p_{\mathrm{K|S}}(k|t, \boldsymbol{\Theta}^\star)\,. \tag{3.18}$$

See appendix C for a proof.

4 Likelihood and Fisher information

Given a random variable X, the *information content* of observing a particular realization $\mathrm{X} = x$ is defined as

$$\mathrm{I}_{\mathrm{X}}(x) \coloneqq -\log p_{\mathrm{X}}(x)\,. \tag{4.1}$$

Intuitively, the more likely the outcome x, the less information content of observing x, e.g., if x is certain then observing x conveys no information.

The *expected value* of the information content is given by the following definition.

Definition 4.1. *The* entropy *of discrete random* X *is a measure of uncertainty defined as*

$$\mathrm{H}(\mathrm{X}) \coloneqq \mathrm{E}[\mathrm{I}_{\mathrm{X}}(\mathrm{X})] = \mathrm{E}[-\log p_{\mathrm{X}}(\mathrm{X})]\,. \tag{4.2}$$

The following example may provide some intuition about the basis of the entropy metric.

Example 1 *Assuming the only information diagnosticians have about a series system is provided by the distribution of* K, *the lower-bound on average number of "yes" or "no" questions needed to determine the component cause of a failed system is given by* H(K) *(where the* logarithm *is base-*2*). In the worst-case,* K *is uniformly distributed and* $\mathrm{H}(\mathrm{K}) = \log_2 m$.

The expectation $\mathrm{E}\Big[-\log f_{\mathrm{X}(x)}\Big]$ of a *continuous* random variable X is called *differential* entropy. Since *negative* values are possible in this case, the previous example no longer applies but the metric is still useful.

Suppose we are interested in the entropy of X given $\mathrm{Y} = y$,

$$\mathrm{H}(\mathrm{X}|\mathrm{Y} = y) \coloneqq \mathrm{E}\Big[-\log f_{\mathrm{X}\,|\,\mathrm{Y}}(\mathrm{X}\,|\,y)\Big]\,. \tag{4.3}$$

More information is never expected to increase the uncertainty, i.e., $\mathrm{H}(\mathrm{X}|\mathrm{Y} = y) \leq \mathrm{H}(\mathrm{X})$, and if X and Y are statistically dependent then $\mathrm{H}(\mathrm{X}|\mathrm{Y} = y) < \mathrm{H}(\mathrm{X})$.

If the diagnosticians described in example 1 are provided with the additional information that the system failed at time t, then the lower-bound on the average number of questions required to isolate the component cause is given by $\mathrm{H}(\mathrm{K}|\mathrm{S} = t) \leq \mathrm{H}(\mathrm{K})$.

If we *average* over all of the values Y may realize, we get the conditional entropy

$$\mathrm{H}(\mathrm{X}|\mathrm{Y}) \coloneqq \mathrm{E}\Big[-\log f_{\mathrm{X}\,|\,\mathrm{Y}}(\mathrm{X}\,|\,\mathrm{Y})\Big]\,. \tag{4.4}$$

The joint entropy of X and Y is defined as

$$\mathrm{H}(\mathrm{X},\mathrm{Y}) \coloneqq \mathrm{E}\Big[-\log f_{\mathrm{X},\mathrm{Y}}(\mathrm{X},\mathrm{Y})\Big] \tag{4.5}$$

which may also be computed by

$$\mathrm{H}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{X}|\mathrm{Y}) + \mathrm{H}(\mathrm{Y})\,. \tag{4.6}$$

If X and Y are *statistically independent*, then $\mathrm{H}(\mathrm{X}|\mathrm{Y}) = \mathrm{H}(\mathrm{X})$ and $\mathrm{H}(\mathrm{Y}|\mathrm{X}) = \mathrm{H}(\mathrm{Y})$.

When we make an *observation* of a random variable, the entropy is a measure of the *expected information content* provided by the observation. The expected information provided by a sample of n masked system failures is given by the following theorem.

Theorem 4.1. *Since* S *is statistically independent of* W *and* A*,*

$$\mathrm{H}(\mathrm{C},\mathrm{S}|\mathrm{W} = w, \mathrm{A} = \alpha) = \mathrm{H}(\mathrm{C}|\mathrm{S}, \mathrm{W} = w, \mathrm{A} = \alpha) + \mathrm{H}(\mathrm{S})\,. \tag{4.7}$$

Given some distribution S, the *maximum* entropy is when the failed component is in the candidate set by random chance, $\mathrm{A} = \frac{w}{m}$, as given by

$$\mathrm{H}(\mathrm{C},\mathrm{S}|\mathrm{W} = w, \mathrm{A} = w/m) = \mathrm{H}(\mathrm{S}) + \log_2 \binom{m}{w} \tag{4.8}$$

and the *minimum* entropy is when the failed component is *certainly* in the candidate set, $\mathrm{A} = 1$, as given by

$$\mathrm{H}(\mathrm{C},\mathrm{S}|\mathrm{W} = w, \mathrm{A} = 1) = \mathrm{H}(\mathrm{S}) + \mathrm{H}(\mathrm{C}|\mathrm{S}, \mathrm{W} = w, \mathrm{A} = 1) \tag{4.9}$$

Proof. The ... □

The rate of change of the entropy with respect to α is given by the following theorem.

Theorem 4.2. *test*

In what follows, we frequently prefer to work with vectors rather than matrices.

Definition 4.2. *Let* $\text{vec}\colon \mathbb{R}^{m\times q} \mapsto \mathbb{R}^{m\cdot q}$ *denote the linear transformation of a matrix of* m *rows and* q *columns into a column vector of* $m \cdot q$ *elements, where consecutive matrix rows are placed side by side. Additionally, the* j^{th} *component of a vector* $\mathbf{a}$ *is denoted by* $[\mathbf{a}]_j$ *or* a_j*.*

Notation. *By definition 4.2,* $\boldsymbol{\Theta}^\star$ *has an equivalent representation denoted by* $\boldsymbol{\theta}^\star \equiv \text{vec}(\boldsymbol{\Theta}^\star)$ *and the* j^{th} *component of* $\boldsymbol{\theta}^\star$ *is denoted by* $\theta_j^\star$*. We use matrix and vector representations interchangeably.*

The likelihood of observing a particular realization $\mathbf{M_n} = \langle \mathcal{C}_1, t_1\rangle, \ldots, \langle \mathcal{C}_n, t_n\rangle$ conditioned on $\mathrm{W} = w$ and $\mathrm{A} = \alpha$ is given by the following definition.

Definition 4.3. *A random sample of* n *masked system failure times conditioned on candidate sets of cardinality* w *has a joint probability density given by*

$$\mathrm{f}_{\mathbf{M_n}|\mathrm{W,A}}(\langle \mathcal{C}_1, t_1\rangle, \ldots, \langle \mathcal{C}_n, t_n\rangle \mid w, \alpha, \boldsymbol{\Theta}^\star) = \prod_{i=1}^{n} f_{\mathrm{S}}(t_i|\boldsymbol{\Theta}^\star) \cdot \mathrm{f}_{\mathrm{C|S,W,A}}(\mathcal{C}_i \mid t_i, w, \alpha, \boldsymbol{\Theta}^\star)\,, \tag{4.10}$$

where f_{S} *is the marginal density given by theorem 3.2 and* $f_{\mathrm{C|S,W,A}}$ *is the conditional probability mass given by corollary 3.7.4.*

A random variable is a measurable function from a probability space to a measurable space of values that the variable can take on. Such a space of values are is called a random variable.

The subsequent material makes a few assumptions, denoted *regularity conditions*, given by the following.

Assumption 10. *The following regularity conditions hold for the* $\mathcal{F}_{\boldsymbol{\Theta}}$*-family of the series system:*

1. $\boldsymbol{\theta}^\star$ *is interior to the parameter support* Ω*.*
2. *The log-likelihood* ℓ *is continuous, thrice differentiable, and bounded.*
3. *The true parameter value* $\boldsymbol{\theta}^\star$ *is identified, i.e.,*

$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}\in\Omega} \mathrm{E}_{\boldsymbol{\theta}^\star}\left[\ln f_{\mathrm{C,S}\,|\,\mathrm{W,A}}(\mathrm{C}, \mathrm{S}\,|\,w, \alpha, \boldsymbol{\theta})\right]. \tag{4.11}$$

By definition 4.3, the probability density that $\mathbf{M_n} = \langle \mathcal{C}_1, t_1, \alpha\rangle, \ldots, \langle \mathcal{C}_n, t_n, \alpha\rangle$ conditioned on $\mathrm{W} = w$ and $\mathrm{A} = \alpha$ is given by

$$f_{\mathbf{M_n}|\mathrm{W,A}}(\langle \mathcal{C}_1, t_1\rangle, \ldots, \langle \mathcal{C}_n, t_n\rangle|w, \alpha, \boldsymbol{\theta}^\star) = \prod_{i=1}^{n} f_{\mathrm{C,S|W,A}}(\mathcal{C}_i, t_i|w, \alpha, \boldsymbol{\theta}^\star)\,. \tag{4.10 revisited}$$

If we fix the sample of masked system failures and allow the parameter $\boldsymbol{\theta}$ to change, we have the likelihood function.

Definition 4.4. *The likelihood function represents the* likelihood *of parameter index* $\boldsymbol{\theta}$ *with* $\mathbf{M_n}$ *fixed at* $\langle \mathcal{C}_1, t_1\rangle, \ldots, \langle \mathcal{C}_n, t_n\rangle$ *is given by*

$$\mathcal{L}_n(\boldsymbol{\theta}|w, \alpha) \coloneqq \prod_{i=1}^{n} f_{\mathrm{C,S|W,A}}(\mathcal{C}_i, t_i|w, \alpha, \boldsymbol{\theta})\,. \tag{4.12}$$

The likelihood is a measure of the extent to which the given sample provides support for particular values of a parameter in $\mathcal{F}_{\boldsymbol{\Theta}}$; its surface captures all the information there is about $\boldsymbol{\theta}^\star$ from the given sample.

The log-likelihood plays a more central role than the likelihood for reasons that will be made clear later on.

Definition 4.5. *The log-likelihood with respect to $\boldsymbol{\theta}$ given a particular $\mathbf{M_n} = \langle \mathcal{C}_1, t_1, \alpha\rangle, \ldots, \langle \mathcal{C}_n, t_n, \alpha\rangle$ conditioned on $\mathrm{W} = w$ and $\mathrm{A} = \alpha$ is given by*

$$\ell_{\mathrm{n}}(\boldsymbol{\theta}|w,\alpha) = \sum_{i=1}^{n} \ln f_{\mathrm{C,S|W,A}}(\mathcal{C}_i, t_i|w,\alpha,\boldsymbol{\theta})\,. \tag{4.13}$$

In subsequent material, we need the gradient operator given by the following definition.

Definition 4.6 (Gradient operator)**.** *The gradient operator $\nabla\colon (\mathbb{R}^q \mapsto \mathbb{R}) \mapsto (\mathbb{R}^q \mapsto \mathbb{R}^q)$ is defined as*

$$\nabla\, \mathrm{g} \coloneqq \left(\frac{\partial\, \mathrm{g}}{\partial x_1}, \ldots, \frac{\partial\, \mathrm{g}}{\partial x_q}\right)^{\mathsf{T}}. \tag{4.14}$$

The score function is given by the following definition.

Definition 4.7. *The Fisher's score function, denoted by $\mathrm{s_n}\colon \mathbb{R}^{m\cdot q} \mapsto \mathbb{R}^{m\cdot q}$, is the gradient of the log-likelihood function given a sample of n masked failure times,*

$$\mathrm{s_n}(\boldsymbol{\theta}|w,\alpha) \coloneqq \nabla\, \ell_{\mathrm{n}}(\boldsymbol{\theta}|w,\alpha)\,. \tag{4.15}$$

The variance-covariance is given by the following definition.

Definition 4.8. *A random vector $\mathbf{X} \in \mathbb{R}^p$ has a variance-covariance given by*

$$\mathrm{var}[\mathbf{X}] \coloneqq \mathrm{E}\Big[(\mathbf{X} - \mathrm{E}[\mathbf{X}])\,(\mathbf{X} - \mathrm{E}[\mathbf{X}])^{\mathsf{T}}\Big]\,, \tag{4.16}$$

which is a p-by-p semi-positive definite matrix.

The score function given a realization $\mathbf{M_n} = \langle \mathcal{C}_1, \mathrm{t}_1, \alpha\rangle, \ldots, \langle \mathcal{C}_n, \mathrm{t}_\mathrm{n}, \alpha\rangle$ given $\mathrm{W} = w$ and $\mathrm{A} = \alpha$ is a statistic. When we consider it as a function of a random sample, it is a random vector. When evaluated at the true parameter index $\boldsymbol{\theta}^\star$, it has a variance-covariance given by the following definition.

Definition 4.9 (Fisher information matrix)**.** *The score as a function of a random sample $\mathbf{M_n}$ given $\mathrm{W} = w$ and $\mathrm{A} = \alpha$ has a variance-covariance known as the Fisher information matrix, a $(m\cdot q)$-by-$(m\cdot q)$ matrix denoted by*

$$\boldsymbol{\mathcal{I}_n}(\boldsymbol{\theta}^\star|w,\alpha) \coloneqq \mathrm{var}_{\boldsymbol{\theta}^\star}\left[\mathrm{s_n}(\boldsymbol{\theta}^\star \mid w,\alpha)\right]. \tag{4.17}$$

The Hessian is given by the following definition.

Definition 4.10. *The Hessian operator $\mathcal{H}\colon (\mathbb{R}^q \mapsto \mathbb{R}) \mapsto (\mathbb{R}^q \mapsto \mathbb{R}^{q\times q}$ applied to $\mathrm{g}\colon \mathbb{R}^q \mapsto \mathbb{R}$ is a q-by-q matrix where the (j,k)-th element is defined as $\frac{\partial^2 \mathrm{g}}{\partial x_j \partial x_k}$.*

This is somewhat related to the *entropy*, except instead of the logarithm of the density we take the logarithm of the gradient of the density, i.e.,

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha) = \mathrm{var}_{\boldsymbol{\theta}^\star}\left[\Big(\nabla_{\boldsymbol{\theta}} \ln f_{\mathrm{C,S\,|\,W,A}}(\mathrm{C},\mathrm{S}\,|\,w,\alpha,\boldsymbol{\theta})\Big)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right]. \tag{4.18}$$

Under the regularity conditions, the information matrix may be computed from the Hessian and is given by the following definition. For clarity, the (i,j)-th element of the information matrix is given by the following definition.

Definition 4.11. *The information matrix of $\boldsymbol{\Theta}^\star$ given by the joint distribution of* C *and* S *conditoned on* $\mathrm{W} = w$ *and* $\mathrm{A} = \alpha$ *is gyven by the expectation of the negative of the Hessian of the log-likelihood function ℓ evaluated at $\boldsymbol{\theta}^\star$. The (i,j)-th element of $\boldsymbol{\mathcal{I}}$ is given by*

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha)_{i\,j} = -\sum_{\mathcal{C}\in\mathcal{W}_w}\int_0^\infty \left[\tfrac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f_{\mathrm{C,S|W,A}}(\mathcal{C},t|w,\alpha,\boldsymbol{\theta})\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} \times f_{\mathrm{C,S|W,A}}(\mathcal{C},t|w,\alpha,\boldsymbol{\theta}^\star)\,\mathrm{d}t\,, \tag{4.19}$$

where $\mathcal{W}_w \times (0,\infty)$ is the sample space.

Observe that
?? may be difficult to compute.

Definition 4.12. *The information matrix is given by the expectation of the negative of the Hessian of the log-likelihood function ℓ evaluated at $\boldsymbol{\theta}^\star$. That is,*

$$\boldsymbol{\mathcal{I}}_{\mathrm{n}}\left(\boldsymbol{\theta}^\star|w\right) = -\,\mathbb{E}_{\boldsymbol{\theta}^\star}\left[\mathcal{H}\left(\ell_n\left(\boldsymbol{\theta}|w\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right], \tag{4.20}$$

where the expectation is taken with respect to $\boldsymbol{\theta}^\star$ and the Hessian is taken with respect to $\boldsymbol{\theta}$ and then evaluated at $\boldsymbol{\theta}^\star$.

Notation. *We denote the Fisher information matrix of $\boldsymbol{\theta}^\star$ given $\mathbf{M}(w,1)$ by*

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w) \coloneqq \boldsymbol{\mathcal{I}}_{\mathbf{1}}(\boldsymbol{\theta}^\star|w)\ . \tag{4.21}$$

The information matrix is a key quantity in large sample theory; it quantifies the *expected* information (as a reduction of uncertainty) $\mathbf{M}(w,n)$ carries about $\boldsymbol{\theta}^\star$. For sufficiently large samples, the log-likelihood has a Taylor series approximation given by

$$\ell_n\left(\boldsymbol{\theta}|w\right) \approx -\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}^\star\right)^{\intercal}\boldsymbol{\mathcal{I}}_{\mathrm{n}}\left(\boldsymbol{\theta}^\star|w\right)\left(\boldsymbol{\theta}-\boldsymbol{\theta}^\star\right)\ , \tag{4.22}$$

which is maximized at $\boldsymbol{\theta}^\star$. Intuitively, the more "peaked" the log-likelihood is expected to be at $\boldsymbol{\theta}^\star$, the more information a random sample is expected to carry about $\boldsymbol{\theta}^\star$.

The information independent samples have about about $\boldsymbol{\theta}^\star$ is *additive* as given by the following postulate.

Postulate 4.1. *Algebraically, the Fisher information independent samples have about the true parameter index $\boldsymbol{\theta}^\star$ is additive such that*

$$\boldsymbol{\mathcal{I}}_{\boldsymbol{n_1}+\boldsymbol{n_2}}\left(\boldsymbol{\theta}^\star|\cdot\ \right) = \boldsymbol{\mathcal{I}}_{\boldsymbol{n_1}}(\boldsymbol{\theta}^\star|\cdot\ ) + \boldsymbol{\mathcal{I}}_{\boldsymbol{n_2}}(\boldsymbol{\theta}^\star|\cdot\ )\,. \tag{4.23}$$

By postulate 4.1, the Fisher information of a random sample of n masked system failure times drawn from $\mathbf{M_n}$ given $\mathrm{W} = w$ and $\mathrm{A} = \alpha$ is given by

$$\boldsymbol{\mathcal{I}_n}(\boldsymbol{\theta}^\star|w,\alpha) = n\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha)\,. \tag{4.24}$$

A *masked system failure time* in which the cardinality of the candidate set is distributed according to W, as described by **??**, has a Fisher information matrix given by the following theorem.

Theorem 4.3. *The* Fisher information *matrix of true parameter index* $\boldsymbol{\theta}^\star$ *with respect to a realization of jointly distributed random system lifetime* S *and candidate set* C *is given by*

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star) = \sum_{\langle w,\alpha\rangle\in\mathcal{W}\times\mathcal{A}} p_{\mathrm{W,A}}(w,\alpha)\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha)\,, \tag{4.25}$$

where $p_{\mathrm{W,A}}$ *is the joint distribution of* W *and* A *with a support* $\mathcal{W}\times\mathcal{A}$.

Proof. Suppose we have a random sample of n masked system failure times. By **??**, the expected number of sample points that have w α-masked candidate sets is given by

$$q \coloneqq n\cdot p_{\mathrm{W,A}}(w,\alpha)\,. \tag{a}$$

Thus, by the additive property of *Fisher information* given by postulate 4.1, the information matrix of $\boldsymbol{\Theta}^\star$ conditioned on q such observations is given by

$$\boldsymbol{\mathcal{I}_q}(\boldsymbol{\theta}^\star|w,\alpha) = n\cdot p_{\mathrm{W,A}}(w,\alpha)\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha)\,. \tag{b}$$

When we sum over the entire support of W and A, we arrive at the *information matrix* of $\boldsymbol{\Theta}^\star$ conditioned on n masked candidate sets as given by

$$\boldsymbol{\mathcal{I}_n}(\boldsymbol{\theta}^\star) = \sum_{\langle w,\alpha\rangle\in\mathcal{W}\times\mathcal{A}} n\cdot p_{\mathrm{W,A}}(w,\alpha)\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha)\,. \tag{c}$$

When we set n to 1, we get the result

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star) = \sum_{\langle w,\alpha\rangle\in\mathcal{W}\times\mathcal{A}} p_{\mathrm{W,A}}(w,\alpha)\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha)\,. \tag{d}$$

□

5 Sampling distribution of parameter estimators

By definitions 2.6 and 2.7, the random system lifetime S has a probability density function that is a member of the $\mathcal{F}_{\boldsymbol{\Theta}}$-family with a true parameter index $\boldsymbol{\theta}^\star$. Let some estimator of $\boldsymbol{\theta}^\star$ be a function of the information in a random sample of n masked system failure times. The estimator is a function of a random sample, thus it has *sampling distribution*, a random vector denoted by $\mathbf{Y_n}$. That is,

$$\mathbf{Y_n} = \boldsymbol{\psi}\left(\mathbf{M_n}\right)\,. \tag{5.1}$$

A particular realization of the estimator is denoted by $\bar{\boldsymbol{\theta}}_\mathbf{n}$, i.e., $\mathbf{Y_n} = \bar{\boldsymbol{\theta}}_\mathbf{n}$. All else being equal, we prefer estimators which vary only *slightly* from sample to sample with a *central tendency* around $\boldsymbol{\theta}^\star$. That is, we prefer unbiased estimators in which each component has small variance.

The sampling distribution of $\bar{\boldsymbol{\theta}}_\mathbf{n}$ has a variance-covariance given by $\mathrm{Var}\left(\mathbf{Y_n}\right)$ (see eq. (4.16)) and a bias given by the following definition.

Definition 5.1. *The bias of a point estimator* $\mathbf{Y_n}$ *is given by*

$$\text{bias}\left(\mathbf{Y_n}\right) = \mathbb{E}\left[\mathbf{Y_n}\right] - \boldsymbol{\theta}^\star \,. \tag{5.2}$$

We are often faced with a trade-off between bias and variance. A measure of estimator accuracy that is a function of both the bias and the variance is the mean squared error as given by the following definition.

Definition 5.2. *The mean squared error (MSE) of the sampling distribution of* $\bar{\boldsymbol{\theta}}_\mathbf{n}$ *is given by*

$$\text{MSE}\left(\mathbf{Y_n}\right) = \mathbb{E}\left[\left(\mathbf{Y_n} - \boldsymbol{\theta}^\star\right)^\mathsf{T}\left(\mathbf{Y_n} - \boldsymbol{\theta}^\star\right)\right] \tag{5.3}$$

An equivalent way to compute the mean squared error is given by the following postulate.

Postulate 5.1. *The mean squared error of the sampling distribution of* $\bar{\boldsymbol{\theta}}_\mathbf{n}$ *as given by definition 5.2 is equivalent to*

$$\text{MSE}\left(\mathbf{Y_n}\right) = \text{tr}\left(\text{Var}\left(\mathbf{Y_n}\right)\right) + \text{bias}^2\Big(\mathbf{Y_n}\Big)\,, \tag{5.4}$$

where $\text{tr}\left(\boldsymbol{A}\right)$ *computes the sum of the diagonal elements of square matrix* $\boldsymbol{A}$*.*

Fisher information reduces the uncertainty about $\boldsymbol{\theta}^\star$. The *minimum-variance unbiased estimator* (UMVUE) has a lower-bound given by the inverse of the *Fisher information matrix*, denoted the *Cramér-Rao lower-bound*. The minimum variance obtainable from a random sample of n masked system failure times drawn from $\mathbf{M_n}$ is given by

$$\boldsymbol{CRLB_n} = \frac{1}{n}\boldsymbol{\mathcal{I}}^{-1}\left(\boldsymbol{\theta}^\star\right)\,. \tag{5.5}$$

By eq. (5.4), if $\bar{\boldsymbol{\theta}}_\mathbf{n}$ is an unbiased estimator of $\boldsymbol{\theta}^\star$, then the mean squared error is given by

$$\text{MSE}(\mathbf{Y_n}) = \text{tr}\big(\text{Var}\left(\mathbf{Y_n}\right)\big)\,. \tag{5.6}$$

By eqs. (5.5) and (5.6), the mean squared error of any unbiased point estimator of $\boldsymbol{\theta}^\star$ in which the only *a priori* information is given by a random sample $\mathbf{M_n}$ has a lower-bound given by the trace of $\boldsymbol{CRLB_n}$,

$$\text{MSE}(\mathbf{Y_n}) \geq \text{tr}(\boldsymbol{CRLB_n})\,. \tag{5.7}$$

5.1 Maximum likelihood estimator for α-masked system failure times each consisting of w components

Suppose the only information about the true parameter index $\boldsymbol{\theta}^\star$ is given by a sample of n α-masked system failure times in which only those system failures with w components with accuracy α are observed. The maximum likelihood estimator based on this conditional sample is given in the following definition.

Definition 5.3. *A parameter index* $\hat{\boldsymbol{\theta}}(n \,|\, w)$ *that maximizes the* likelihood *of observing some realization of* $\mathbf{M_n}$ *given* $\text{W} = w$ *and* A *(see assumption 8) is a maximum likelihood estimator of* $\boldsymbol{\theta}^\star$*. That is,*

$$\hat{\boldsymbol{\theta}}(n \,|\, w) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \mathcal{L}_n(\boldsymbol{\theta}|w,\alpha)\,, \tag{5.8}$$

where $\boldsymbol{\Omega}$ *is the feasible parameter space of the parametric family.*

As further justification of assumption 7, the maximum likelihood estimator maximizes the likelihood of generating the given $\mathbf{m_n}$ such that for each system failure time a component in the corresponding candidate set is the cause.

The accuracy of the maximum likelihood estimator is a function of the agreement between the model and reality. In our model, the primary point of disagreement is given by the assumption of a series system under a parametric model of independent component lifetimes, i.e., $\mathcal{F}_{\boldsymbol{\Theta}}$. However, if the model is a reasonable abstraction of the objects of interest,

The distribution of candidate sets, i.e, C, disagrees with the way the actual candidate sets are generated. Even if this is the case, the logic that components that are more likely to have caused a system failure *with respect to the parameters* (other ways in which a component might fail are outside the scope of the parametric model)

The accuracy of the maximum likelihood estimator is a function of the agreement between the model and reality. In our model, the primary point of disagreement is given by the assumption of a series system under a parametric model of independent component lifetimes, i.e., $\mathcal{F}_{\boldsymbol{\Theta}}$.

Another point of disagreement is the assumption that candidates are generated from the joint probability densities of the component lifetimes $\mathrm{T}_1, \ldots, \mathrm{T}_m$ and random candidate sets C. However, recall the assumption is that, under the assumed model $\mathrm{S} = \min(\mathrm{T}_1, \ldots, \mathrm{T}_m)$, components that are more likely to have failed at observable time $\mathrm{S} = t$ are more likely to be in the observable candidate set.

Note that *any other* distribution of candidate sets generates a smaller likelihood given the parametric model. We claim this is the *optimal* candidate set since it provides the most information about $\boldsymbol{\theta}$. If the parametric family $\mathcal{F}_{\boldsymbol{\Theta}}$ is a very accurate model of the reality, then whether the failed component is in the candidate set or not is irrelevant. What matters more is that the most likely components are in the candidate set, as that will provide the most information about $\boldsymbol{\theta}$.

If this is not a very accurate model, the maximum likelihood estimator may not be a consistent estimator even if . Rather, the maximum likelihood estimator will generate an estimate that makes the provided sample the most likely to be generated by the assumed model, i.e., the component lifetimes will not be accurately modeled since they must be modeled in such a way as to make the biased sample more likely.

The logarithm is a monotonically increasing function, thus the parameter value $\hat{\boldsymbol{\theta}}(n \mid w)$ that maximizes $\mathcal{L}$ also maximizes the log-likelihood ℓ (see definition 4.5), i.e.,

$$\hat{\boldsymbol{\theta}}(n \mid w) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} \ell(\boldsymbol{\theta} | w, \mathbf{m_n}) \,. \tag{5.9}$$

According to Bickel [1], if the parameter support $\boldsymbol{\Omega}$ of the parametric family is open, the log-likelihood ℓ is differentiable with respect to $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}(n \mid w)$ exists, then $\hat{\boldsymbol{\theta}}(n \mid w)$ must be a stationary point of ℓ as given by

$$\mathrm{s}\left(\hat{\boldsymbol{\theta}}(n \mid w) | w, \mathbf{m_n}\right) = \mathbf{0} \,, \tag{5.10}$$

where s is the score function given by eq. (4.15). In cases where there are no closed-form solutions to eq. (5.10), iterative methods may be used as described in appendix D.

An estimator of $\boldsymbol{\theta}^\star$ given a sample $\mathbf{M}(w, n) = \mathbf{m_n}$ is the maximum likelihood estimator $\hat{\boldsymbol{\theta}}(n \mid w)$ described in definition 5.3. Since $\hat{\boldsymbol{\theta}}(n \mid w)$ is a function of the random sample $\mathbf{M}(w, n)$, it has a *sampling distribution.*

Definition 5.4. *The* sampling distribution *of* $\hat{\boldsymbol{\theta}}(n \mid w)$ *is a random vector denoted by* $\mathbf{Y}(n \mid w) \in \mathrm{R}^{m \cdot q}$.

The generative model for the maximum likelihood estimator, `generate_mle`$(w|\cdot\,)$, is described by algorithm 1. This generative model depends on `generate_msft`, the generative model for the *masked system failure time* as described by **??**. Note that in algorithm 1, line 7 may be approximated with `find_mle`, a function that numerically solves the stationary points of the maximum likelihood equations as described by algorithm 3.

Algorithm 1: Generative model of the maximum likelihood estimator conditioned on w candidates

Result: a realization of a maximum likelihood estimate from the sampling distribution of $\hat{\boldsymbol{\theta}}(n \,|\, w)$.

Input:

$\boldsymbol{\theta}^\star$, the true parameter index.

w, the cardinality of the candidate set.

n, the number of masked system failure times.

Output:

$\hat{\boldsymbol{\theta}}(n \,|\, w)$, a realization of $\mathbf{Y}(n \,|\, w)$.

1 **Model** `generate_mle`$(n, w, \boldsymbol{\theta}^\star)$
2 $\quad \mathbf{m_n} \leftarrow \emptyset$
3 $\quad$ **for** $i \leftarrow 1$ **to** n **do**
4 $\quad\quad \mathbf{M} \leftarrow$ `generate_msft` $(w, \boldsymbol{\theta}^\star)$
5 $\quad\quad \mathbf{m_n} \leftarrow \mathbf{m_n} \cup \{\mathbf{M}\}$
6 $\quad$ **end**
7 $\quad \hat{\boldsymbol{\theta}}(n \,|\, w) \leftarrow \arg\max_{\boldsymbol{\theta} \in \Omega} \ell\left(\boldsymbol{\theta} | w, \mathbf{m_n}\right)$
8 $\quad$ **return** $\hat{\boldsymbol{\theta}}(n \,|\, w)$

The asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}(n \,|\, w)$ is a consistent estimator of $\boldsymbol{\theta}^\star$.

Postulate 5.2. *As $n \to \infty$, the sampling distribution of $\hat{\boldsymbol{\theta}}(n \,|\, w)$ converges in probability to $\boldsymbol{\theta}^\star$, written*

$$\mathbf{Y}(n \,|\, w) \xrightarrow{P} \boldsymbol{\theta}^\star \,, \tag{5.11}$$

since

$$\lim_{n \to \infty} \mathrm{P}\Big[\mathrm{MSE}\left(\mathbf{Y}(n \,|\, w)\right) < \epsilon\Big] = 1 \tag{5.12}$$

for every $\epsilon > 0$.

By eqs. (4.16) and (5.11), the variance-covariance of the asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}(n \,|\, w)$ is given by

$$\mathrm{Var}\big[\mathbf{Y}(n \,|\, w)\big] = \mathbb{E}\left[\left(\mathbf{Y}(n \,|\, w) - \boldsymbol{\theta}^\star\right)\left(\mathbf{Y}(n \,|\, w) - \boldsymbol{\theta}^\star\right)^{\mathsf{T}}\right] . \tag{5.13}$$

Postulate 5.3. *The variance-covariance of the asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}(n \,|\, w)$ obtains the* Cramér-Rao lower-bound *for point estimators that are strictly a function of n masked system failure times in which candidate sets are of cardinality w. We denote this variance-covariance by*

$$\boldsymbol{\Sigma}_{\boldsymbol{n}}(\boldsymbol{\theta}^\star | w, \alpha) \equiv \frac{1}{n} \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}^\star | w, \alpha) \,. \tag{5.14}$$

The asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}(n \,|\, w)$ is normally distributed.

Postulate 5.4. *As $n \to \infty$, the sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$ converges in distribution to a multivariate normal distribution with a mean $\boldsymbol{\theta}^\star$ and a variance-covariance $\Sigma_{\mathrm{n}}(\boldsymbol{\theta}^\star|w,\alpha)$, written*

$$\mathbf{Y}(n \mid w) \xrightarrow{d} \mathrm{MVN}\Big(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma}_{\boldsymbol{n}}(\boldsymbol{\theta}^\star|w,\alpha)\Big) \; . \tag{5.15}$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}(n \mid w)$ is an asymptotically *efficient* estimator since it obtains the Cramér-Rao lower-bound as given by **??**. Thus, for a sufficiently large sample of masked system failure times, the sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$ varies only *slightly* from sample to sample with a *central tendency* around $\boldsymbol{\theta}^\star$.

By eqs. (5.5), (5.6) and (5.11) and **??**, the asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$ has the minimum mean squared error of any unbiased estimator,

$$\mathrm{MSE}\left(\mathbf{Y}(n \mid w)\right) = \frac{1}{n} \operatorname{tr}\left(\boldsymbol{\mathcal{I}}^{-1}\Big(\boldsymbol{\theta}^\star|w,\alpha\Big)\right) = \operatorname{tr}\left(\mathrm{CRLB}_{\mathrm{n\,w}}\right) \; . \tag{5.16}$$

Remark. *Look up Slverty's theorem to justify the next paragraph.* △

To estimate the sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$, we may assume the sample size is sufficiently large such that the asymptotic distribution becomes a reasonable approximation. Since $\mathbf{Y}(n \mid w) \xrightarrow{P} \boldsymbol{\theta}^\star$ and $\mathbf{Y}(n \mid w) \xrightarrow{d} \mathrm{MVN}\left(\boldsymbol{\theta}^\star, \Sigma_{\mathrm{n}}(\boldsymbol{\theta}^\star|w,\alpha)\right)$, it follows that

$$\mathbf{Y}(n \mid w) \xrightarrow{d} \mathrm{MVN}\left(\hat{\boldsymbol{\theta}}(n \mid w), \boldsymbol{\Sigma}_{\boldsymbol{n}}(\hat{\boldsymbol{\theta}}(n \mid w)|w,\alpha)\right) \; . \tag{5.17}$$

TODO: talk about A = α. We don't really want to think about modeling it, only saying that it is given in a sample. Say, for instance, the masked system failure time also has a label for how accurate they determine the candidate set to be, i.e., α-candidate set model where α can vary over the sample. Group by w and α, compute the information matrix, then combine them as described in this section to get the final estimator.

Thus, we can approximate $\boldsymbol{\theta}^\star$ and $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w,\alpha)$ and obtain the following result. The proof of this theorem is beyond the scope of this paper.

Theorem 5.1. *For sufficiently large sample size n, the sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$ is approximately normally distributed with a mean $\hat{\boldsymbol{\theta}}(n \mid w)$ and a variance-covariance matrix $\Sigma_{\mathrm{n}}(\hat{\boldsymbol{\theta}}(n \mid w)|w,\alpha)$, written*

$$\mathbf{Y}(n \mid w) \overset{approx.}{\sim} \mathrm{MVN}\left(\hat{\boldsymbol{\theta}}(n \mid w), \boldsymbol{\Sigma}_{\boldsymbol{n}}(\hat{\boldsymbol{\theta}}(n \mid w)|w,\alpha)\right) \; . \tag{5.18}$$

5.2 Maximum likelihood estimator

Consider an i.i.d. sample of r asymptotically unbiased estimates of $\boldsymbol{\theta}^\star$ denoted by $\bar{\boldsymbol{\theta}}^{(i)}$ which have sampling distributions with variance-covariances given respectively by $\boldsymbol{\Sigma}^{(i)}$ for $i = 1, \ldots, r$. The maximum likelihood estimator of $\boldsymbol{\theta}^\star$ given these point estimates is the inverse-variance weighted mean and is given by

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{r} \boldsymbol{A_i}\right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i}\bar{\boldsymbol{\theta}}^{(i)}\right) \; , \tag{5.19}$$

where $\boldsymbol{A_i}$ is the inverse of $\boldsymbol{\Sigma}^{(i)}$.

Suppose that the estimates are given by the maximum likelihood estimator described in **??**. The maximum likelihood estimator given these estimates has a sampling distribution given by the following definition.

Definition 5.5. *Let* $\mathbf{M_n}$ *be a random sample of* n *masked system failure times in which* n_i *realizations have* w_i α_i*-masked component failures for* $i = 1, \ldots, r$*. The maximum likelihood estimator given by*

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{r} \boldsymbol{A_i} \right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \hat{\boldsymbol{\theta}}^{(i)} \right) , \tag{5.20}$$

has a sampling distribution given by

$$\mathbf{Y_n} = \left(\sum_{i=1}^{r} \boldsymbol{A_i} \right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i) \right) , \tag{5.21}$$

where $\mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i)$ *is the sampling distribution of* $\hat{\boldsymbol{\theta}}(n_i \,|\, w_i, \alpha_i)$ *for* $i = 1, \ldots, r$ *and*

$$\boldsymbol{A_i} = n_i \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\star} | w_i, \alpha_i) \tag{5.22}$$

is the information matrix for $\boldsymbol{\theta}^{\star}$ *with respect to* $\mathbf{M}(w_i, \alpha_i, n_i)$*.*

Theorem 5.2. *The random vector* $\mathbf{Y_n}$ *is an asymptotically unbiased estimator of* $\boldsymbol{\theta}^{\star}$*.*

Proof. The expectation of $\mathbf{Y_n}$ is given by

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Y_n}\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{r} \boldsymbol{A_i} \right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i) \right) \right] && \text{(a)} \\
&= \left(\sum_{i=1}^{r} \boldsymbol{A_i} \right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \, \mathbb{E}\left[\mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i) \right] \right) && \text{(b)} \\
&= \left(\sum_{i=1}^{r} \boldsymbol{A_i} \right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \right) \boldsymbol{\theta}^{\star} && \text{(c)} \\
&= \boldsymbol{\theta}^{\star} . && \text{(d)}
\end{aligned}$$

□

Theorem 5.3. *The variance-covariance of* $\mathbf{Y_n}$ *is given by*

$$\mathrm{Var}\left[\mathbf{Y_n}\right] = \left(\sum_{i=1}^{r} n_i \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\star} | w_i, \alpha_i) \right)^{-1} . \tag{5.23}$$

Proof. Let

$$\boldsymbol{B} = \left(\sum_{i=1}^{r} \boldsymbol{A_i} \right)^{-1} . \tag{a}$$

The variance-covariance of $\mathbf{Y_n}$ is given by

$$\operatorname{Var}[\mathbf{Y_n}] = \operatorname{Var}\left[\boldsymbol{B}\left(\sum_{i=1}^{r} \boldsymbol{A_i}\mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i)\right)\right] \tag{b}$$

$$= \boldsymbol{B}\operatorname{Var}\left[\sum_{i=1}^{r} \boldsymbol{A_i}\mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i)\right]\boldsymbol{B^\intercal} \tag{c}$$

$$= \boldsymbol{B}\left(\sum_{i=1}^{r} \operatorname{Var}\left[\boldsymbol{A_i}\mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i)\right]\right)\boldsymbol{B^\intercal} \tag{d}$$

$$= \boldsymbol{B}\left(\sum_{i=1}^{r} \boldsymbol{A_i}\operatorname{Var}\left[\mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i)\right]\boldsymbol{A_i^\intercal}\right)\boldsymbol{B^\intercal}. \tag{e}$$

By **??**, the variance-covariance of $\mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i)$ is given by

$$\frac{1}{n_i}\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}^\star|w_i, \alpha_i)\,. \tag{f}$$

By eq. (5.22), this is equivalent to $\boldsymbol{A_i^{-1}}$. Performing this substitution results in

$$\operatorname{Var}[\mathbf{Y_n}] = \boldsymbol{B}\left(\sum_{i=1}^{r} \boldsymbol{A_i}\boldsymbol{A_i^{-1}}\boldsymbol{A_i^\intercal}\right)\boldsymbol{B^\intercal} \tag{g}$$

$$= \boldsymbol{B}\left(\sum_{i=1}^{r} \boldsymbol{A_i^\intercal}\right)\boldsymbol{B^\intercal}. \tag{h}$$

The summation is equivalent to $\left(\boldsymbol{B^\intercal}\right)^{\boldsymbol{-1}}$. Performing this substitution results in

$$\operatorname{Var}[\mathbf{Y_n}] = \boldsymbol{B}\left(\boldsymbol{B^\intercal}\right)^{\boldsymbol{-1}}\boldsymbol{B^\intercal} \tag{i}$$

$$= \boldsymbol{B} \tag{j}$$

By eq. (a), $\boldsymbol{B}$ is given by

$$\left(\sum_{i=1}^{r} \boldsymbol{A_i}\right)^{-1} \tag{k}$$

and by eq. (5.22), $\boldsymbol{A_i}$ is given by

$$n_i\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w_i, \alpha_i)\,. \tag{l}$$

Performing these substitution results in

$$\operatorname{Var}[\mathbf{Y_n}] = \left(\sum_{i=1}^{r} n_i\,\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star|w_i, \alpha_i)\right)^{-1}. \tag{m}$$

□

The weighted estimator asymptotically achieves the Cramér-Rao lower-bound as given by eq. (5.5) thus it is the asymptotic UMVUE estimator of $\boldsymbol{\theta}^\star$ given a random sample of masked samples failures in which n_i realizations have w_i α_i-masked component failures for $i = 1, \ldots, r$.

A linear combination of multivariate normal distributions is a multivariate normal distribution, thus the asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ is normally distributed.

Postulate 5.5. *As $n \to \infty$, the sampling distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ converges in distribution to a multivariate normal with a mean $\boldsymbol{\theta}^\star$ and a variance-covariance given by eq. (5.23), written*

$$\mathbf{Y}_{\mathbf{n}} \xrightarrow{d} \text{MVN}\left(\boldsymbol{\theta}^\star, \left(\sum_{i=1}^{r} n_i \boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^\star | w_i\right)\right)^{-1}\right). \tag{5.24}$$

The generative model for $\mathbf{Y}_{\mathbf{n}}$ is given by algorithm 4.

Algorithm 2: Generative model of maximum likelihood estimator

Input:
$\boldsymbol{\Theta}^\star$, the true parameter value of the series system.

Output:
$\hat{\boldsymbol{\theta}}_{\mathbf{n}}$, a realization of $\mathbf{Y}_{\mathbf{n}}$.

Model `generate_mle`($\boldsymbol{\Theta}^\star$)
  draw accuracy $\alpha \sim p_{\text{A}}(\cdot)$
  draw α-masked failure cardinality $w \sim p_{\text{W|A}}(\cdot\,|\alpha)$
  $\hat{\boldsymbol{\theta}}_{\mathbf{n}} \leftarrow$ `generate_mle`($\boldsymbol{\Theta}^\star, w, \alpha$)
  return $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$

Bibliography

[1] Peter Bickel and Kjell Doksum. *Mathematical Statistics*, volume 1, chapter 2, page 117. Prentice Hall, 2 edition, 2000.

A Alternative proof of equation 3.6

Equation (3.6) on page 10 asserts that

$$f_{\text{K,S}}(k, t \,|\, \boldsymbol{\Theta}^\star) = f_k(t \,|\, \boldsymbol{\Theta}_k^\star) \prod_{\substack{j=1 \\ j \neq k}}^{m} R_j(t \,|\, \boldsymbol{\Theta}_j^\star).$$

Proof. First, note that $f_{\text{K,S}}(k, t \,|\, \boldsymbol{\Theta}^\star)$ is a proper density since

$$\int_0^\infty \sum_{j=1}^{m} f_{\text{K,S}}(j, t \,|\, \boldsymbol{\Theta}^\star)\, \mathrm{d}t = \int_0^\infty f_{\text{S}}(t \,|\, \boldsymbol{\Theta}^\star)\, \mathrm{d}t = 1\,. \tag{a}$$

Consider a 3-out-of-3 system. By assumption 2, T_1, T_2, and T_3 are mutually independent. Thus, the joint density that $\mathrm{T}_1 = t_1$, $\mathrm{T}_2 = t_2$, and $\mathrm{T}_3 = t_3$ is given by

$$f_{\mathrm{T}_1,\mathrm{T}_2,\mathrm{T}_3}(t_1, t_2, t_3 \,|\, \boldsymbol{\Theta}^\star) = \prod_{p=1}^{3} f_p(t_p \,|\, \boldsymbol{\Theta}_p^\star)\,. \tag{b}$$

Component 1 causes a system failure during the interval $(0, t)$, $t > 0$, if component 1 fails during the given interval and components 2 and 3 survive longer than component 1. That is, $0 < \mathrm{T}_1 < t$, $\mathrm{T}_1 < \mathrm{T}_2$, and $\mathrm{T}_1 < \mathrm{T}_3$.

Let $p(t)$ denote the probability that component 1 causes a system failure during the interval $(0, t)$,

$$p(t) = \mathrm{P}[0 < \mathrm{T}_1 < t \cap \mathrm{T}_1 < \mathrm{T}_2 \cap \mathrm{T}_1 < \mathrm{T}_3]\,. \tag{c}$$

This probability is given by

$$p(t) = \int_{t_1=0}^{t_1=t} \int_{t_2=t_1}^{t_2=\infty} \int_{t_3=t_1}^{t_3=\infty} \prod_{p=1}^{3} f_p(t_p \,|\, \boldsymbol{\Theta}_p^\star)\,\mathrm{d}t_3\,\mathrm{d}t_2\,\mathrm{d}t_1\,. \tag{d}$$

Performing the integration over t_3 results in

$$p(t) = \int_{t_1=0}^{t_1=t} \int_{t_2=t_1}^{t_2=\infty} \prod_{p=1}^{2} f_p(t_p \,|\, \boldsymbol{\Theta}_p^\star) R_3(t_1 \,|\, \boldsymbol{\Theta}_3^\star)\,\mathrm{d}t_2\,\mathrm{d}t_1\,. \tag{e}$$

Performing the integration over t_2 results in

$$p(t) = \int_{t_1=0}^{t_1=t} f_1(t_1 \,|\, \boldsymbol{\Theta}_1^\star) \prod_{p=2}^{3} R_p(t_1 \,|\, \boldsymbol{\Theta}_p^\star)\,\mathrm{d}t_1\,. \tag{f}$$

Taking the derivative of $p(t)$ with respect to t produces the density of probability near t. By the Second Fundamental Theorem of Calculus,

$$\frac{\mathrm{d}p}{\mathrm{d}t} = f_1(t \,|\, \boldsymbol{\Theta}_1^\star) \prod_{p=2}^{3} R_p(t \,|\, \boldsymbol{\Theta}_p^\star)\,. \tag{g}$$

The probability density $\frac{\mathrm{d}p}{\mathrm{d}t}$ is equivalent to the joint density that $\mathrm{K} = 1$ and $\mathrm{S} = t$ as given by theorem 3.4, i.e., $\frac{\mathrm{d}p}{\mathrm{d}t} = f_{\mathrm{K},\mathrm{S}}(1, t \,|\, \boldsymbol{\Theta}^\star)$. Generalizing from this completes the proof. □

B Alternative proof of equation ??

By **??**, S and W are independent and the marginal distribution of W is independent of $\boldsymbol{\theta}^\star$, thus the likelihood of observing a particular realization, $\mathbf{M_n} = \mathbf{m_n}$, is given by

$$\begin{aligned} f_{\mathbf{M_n}}(\mathbf{m_n} \,|\, \boldsymbol{\theta}^\star) &= \prod_{i=1}^{n} f_{\mathrm{C,S}\,|\,\mathrm{W}}\Big(\mathcal{C}_i, \mathrm{t_i} \,\Big|\, |\mathcal{C}_\mathrm{i}|, \boldsymbol{\theta}^\star\Big)\, p_\mathrm{W}(|\mathcal{C}_i|) \\ &= \left[\prod_{i=1}^{n} f_{\mathrm{C,S}\,|\,\mathrm{W}}\Big(\mathcal{C}_i, \mathrm{t_i} \,\Big|\, |\mathcal{C}_\mathrm{i}|, \boldsymbol{\theta}^\star\Big)\right] \left[\prod_{i=1}^{n} p_\mathrm{W}(|\mathcal{C}_i|)\right]\,. \end{aligned} \tag{B.1}$$

If we fix $\mathbf{m_n}$ and allow the parameter $\boldsymbol{\theta}$ to change, we have the likelihood function

$$\mathcal{L}\left(\boldsymbol{\theta} \,|\, \mathbf{m_n}\right) = c \prod_{i=1}^{n} f_{\mathrm{C,S}\,|\,\mathrm{W}}\Big(\mathcal{C}_i, \mathrm{t_i} \,\Big|\, |\mathcal{C}_\mathrm{i}|, \boldsymbol{\theta}^\star\Big)\ , \tag{B.2}$$

where the product over $p_\mathrm{W}(\,\cdot\,)$ is constant with respect to $\boldsymbol{\theta}$ and has been relabeled as c. The log-likelihood with respect to $\boldsymbol{\theta}$ then is given by

$$\begin{aligned} \ell(\boldsymbol{\theta} \,|\, \mathbf{m_n}) &= \ln c + \sum_{i=1}^{n} \ln f_{\mathrm{C,S}\,|\,\mathrm{W}}(\mathcal{C}_i, \mathrm{t_i} \,|\, |\mathcal{C}_\mathrm{i}|, \boldsymbol{\theta}) \\ &= c' + \sum_{w=1}^{m-1} \left[\sum_{i \in \mathbb{A}(w)} \ln f_{\mathrm{C,S}\,|\,\mathrm{W}}(\mathcal{C}_i, \mathrm{t_i} \,|\, \mathrm{w}, \boldsymbol{\theta}) \right] , \end{aligned} \tag{B.3}$$

where $\mathbb{A}(w) = \{i \in \{1, \ldots, n\} \colon |\mathcal{C}_i| = w\}$. Let $\mathbf{m_{n_w}} = \{(\mathcal{C}_i, \mathrm{t_i}) \colon \mathrm{i} \in \mathbb{A}(\mathrm{w})\}$. The part in brackets is the log-likelihood $\ell(\boldsymbol{\theta} \,|\, \mathbf{m_{n_w}})$. Performing this substitution yields

$$\ell(\boldsymbol{\theta} \,|\, \mathbf{m_n}) = \ln c + \sum_{w=1}^{m-1} \ell(\boldsymbol{\theta} \,|\, \mathbf{m_{n_w}})\,. \tag{B.4}$$

The score function is given by

$$\mathrm{s}\left(\boldsymbol{\theta} \,|\, \mathbf{m_n}\right) = \sum_{w=1}^{m-1} \mathrm{s}\left(\boldsymbol{\theta} \,|\, w, \mathbf{m_{n_w}}\right)\ , \tag{B.5}$$

The information matrix may be computed by taking the negative of the expectation of the Jacobian of the score over the true joint distribution of S, C, and W, giving

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^\star\right) = -\,\mathbb{E}_{\boldsymbol{\theta}^\star} \left[\mathcal{J}\left(\mathrm{s}\left(\boldsymbol{\theta} \,|\, \mathbf{M_1}\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} \right] . \tag{B.6}$$

Substituting eq. (B.5) with its definition gives

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^\star\right) = -\,\mathbb{E}_{\boldsymbol{\theta}^\star} \left[\mathcal{J}\left(\sum_{w=1}^{m-1} \mathrm{s}\left(\boldsymbol{\theta} \,|\, w, \mathbf{M}(w,1)\right) \right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} \right] . \tag{B.7}$$

The Jacobian is a linear operator so it can be moved inside the summation, giving

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^\star\right) = -\,\mathbb{E}_{\boldsymbol{\theta}^\star} \left[\sum_{w=1}^{m-1} \mathcal{J}\left(\mathrm{s}\left(\boldsymbol{\theta} \,|\, w, \mathbf{M}(w,1)\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} \right] . \tag{B.8}$$

The Jacobian of s is the Hessian of ℓ. Performing this substitution gives

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^\star\right) = -\,\mathbb{E}_{\boldsymbol{\theta}^\star} \left[\sum_{w=1}^{m-1} \mathcal{H}\left(\ell\left(\boldsymbol{\theta} \,|\, w, \mathbf{M}(w,1)\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} \right] . \tag{B.9}$$

The expectation is a linear operator so it may be moved inside the summation, giving

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^{\star}\right) = -\sum_{w=1}^{m-1} \mathbb{E}_{\boldsymbol{\theta}^{\star}}\left[\mathcal{H}\left(\ell\left(\boldsymbol{\theta} \,|\, w, \mathbf{M}(w,1)\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\star}}\right] . \tag{B.10}$$

The expectation sums over the probability mass function $\mathrm{W} \sim p_{\mathrm{W}}(\,\cdot\,)$, giving

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^{\star}\right) = -\sum_{w=1}^{m-1} p_{\mathrm{W}}(w)\, \mathbb{E}_{\boldsymbol{\theta}^{\star}}\left[\mathcal{H}\left(\ell\left(\boldsymbol{\theta} \,|\, w, \mathbf{M}(w,1)\right)\right)\right] , \tag{B.11}$$

where the expectation is now over the true marginal joint distribution of C and S given $\mathrm{W} = w$. The expectation of the Hessian of ℓ is equivalent to the negative of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^{\star} \,|\, w, \mathbf{M}(w,1))$. Performing this substitution gives

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^{\star}\right) = \sum_{w=1}^{m-1} p_{\mathrm{W}}(w)\, \boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^{\star} \,|\, w\right) . \tag{B.12}$$

□

C Proof of corollary 3.7.5

Corollary 3.7.5 on page 15 asserts that K given S is conditionally independent of C, W, and A.

Proof. By the laws of probability,

$$p_{\mathrm{K|C,S,W,A}}(k|\mathcal{C},t,w,\alpha,\boldsymbol{\Theta}^{\star}) = \frac{f_{\mathrm{K,C,S|W,A}}(k,\mathcal{C},t|w,\alpha,\boldsymbol{\Theta}^{\star})}{f_{\mathrm{C,S|W,A}}(\mathcal{C},t|w,\alpha,\boldsymbol{\Theta}^{\star})} \tag{a}$$

which may be rewritten as

$$p_{\mathrm{K|C,S,W,A}}(k|\mathcal{C},t,w,\alpha,\boldsymbol{\Theta}^{\star}) = \frac{p_{\mathrm{C|K,S,W,A}}(\mathcal{C}|k,t,w,\alpha,\boldsymbol{\Theta}^{\star}) f_{\mathrm{K,S|W,A}}(k,t|w,\alpha,\boldsymbol{\Theta}^{\star})}{f_{\mathrm{C,S|W,A}}(\mathcal{C},t|w,\alpha,\boldsymbol{\Theta}^{\star})} . \tag{b}$$

By assumptions 5 and 6, we may simplify the above equation to

$$p_{\mathrm{K|C,S,W,A}}(k|\mathcal{C},t,w,\alpha,\boldsymbol{\Theta}^{\star}) = \frac{p_{\mathrm{C|K,W,A}}(k|\mathcal{C},w,\alpha) f_{\mathrm{K,S}}(k,t|\boldsymbol{\Theta}^{\star})}{f_{\mathrm{C,S|W,A}}(\mathcal{C},t|w,\alpha,\boldsymbol{\Theta}^{\star})} . \tag{c}$$

By eqs. (3.8) and (3.9), the above equation may be rewritten as

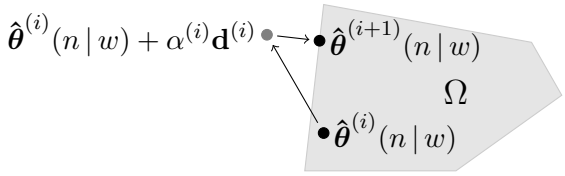
$$p_{\mathrm{K|C,S,W,A}}(k|\mathcal{C},t,w,\alpha,\boldsymbol{\Theta}^{\star}) = \frac{\frac{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star})}{\binom{m-1}{w}}\Big((1-\alpha)h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) - (1-\alpha\frac{m}{w})\sum_{j\in\mathcal{C}} h_j(t|\boldsymbol{\Theta}^{\star}_j)\Big) h_k(t|\boldsymbol{\Theta}^{\star}_k)}{\frac{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star})}{\binom{m-1}{w}}\Big((1-\alpha)h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) - \big(1-\alpha\frac{m}{w}\big)\sum_{k\in\mathcal{C}} h_k(t|\boldsymbol{\Theta}^{\star}_k)\Big) h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star})} . \tag{d}$$

The above equation may be simplified to

$$p_{\mathrm{K|C,S,W,A}}(k|\mathcal{C},t,w,\alpha,\boldsymbol{\Theta}^{\star}) = \frac{h_k(t|\boldsymbol{\Theta}^{\star}_k)}{h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star})} \tag{e}$$

which is the same as $p_{\mathrm{K|S}}(k|t,\boldsymbol{\Theta}^{\star})$ and thus K given S is conditionally independent of C, W, and A. □

Figure 3: Projection onto convex set Ω.



D Numerical solutions to the MLE

The function $\ell(\boldsymbol{\theta} \mid w, \mathbf{m_n})$ is the log-likelihood with respect to $\boldsymbol{\theta} \in \Omega$ where $\Omega \subset \mathbb{R}^{m \cdot q}$. This function has a surface in an $(m \cdot q + 1)$ dimensional space, where a particular point on this surface represents the log-likelihood of observing $\mathbf{m_n}$ with respect to $\boldsymbol{\theta}$. By definition 5.3, $\hat{\boldsymbol{\theta}}(n \mid w)$ is the point on this surface that is at a maximum,

$$\hat{\boldsymbol{\theta}}(n \mid w) = \arg\max_{\boldsymbol{\theta} \in \Omega} \ell(\boldsymbol{\theta} \mid w, \mathbf{m_n}) \,. \qquad \text{(5.9 revisited)}$$

Generally there is no closed-form solution that solves $\hat{\boldsymbol{\theta}}(n \mid w)$, in which case iterative search methods may be used to numerically approximate a solution.

The general version of iterative search that numerically approximates a solution to eq. (5.10), subject to the constraint $\boldsymbol{\theta}^\star \in \Omega$, is shown in algorithm 3. Since iterative search is a local search method, it may fail to converge to a global maximum.

Algorithm 3: Iterative maximum likelihood search

Result: an approximate solution to the stationary points of the maximum likelihood equation

Input:

$\boldsymbol{\Theta}^\star$, the true parameter index.

Output:

$\hat{\boldsymbol{\theta}}(n \mid w)$, an approximation of the maximum likelihood estimate.

Model `find_mle`$(\mathbf{m_n})$
  $\hat{\boldsymbol{\theta}}^{(0)}(n \mid w) \leftarrow$ an initial starting point in Ω
  $i \leftarrow 0$
  while stopping criteria *not satisfied* **do**
    $\hat{\boldsymbol{\theta}}^{(i+1)}(n \mid w) \leftarrow$ `project` $\left(\hat{\boldsymbol{\theta}}^{(i)}(n \mid w) + \alpha^{(i)}\mathbf{d^{(i)}}, \Omega\right)$
    $i \leftarrow i + 1$
  end
  return $\hat{\boldsymbol{\theta}}^{(i)}(n \mid w)$

Assume parameter space Ω is convex. Then, the function `project`$(\boldsymbol{\theta}, \Omega)$ in algorithm 3 projects any point $\boldsymbol{\theta}$ to the nearest point in the parameter space Ω as depicted by fig. 3. Thus, the search method is restricted to searching over the feasible parameter space.

In algorithm 3, $\mathbf{d}^{(i)}$ is a unit vector denoting a *promising* direction in which to search for a better solution than $\hat{\boldsymbol{\theta}}^{(i)}(n \mid w)$. The Fisher scoring algorithm is a slight variation of Newton-Raphson[3] in

[3] If the *observed* information matrix is used instead of the *expected* information matrix, the Fisher scoring algortihm is equivalent to Newton-Raphson.

which

$$\mathbf{d}^{(i)} = \boldsymbol{\mathcal{I}}^{-1}\left(\hat{\boldsymbol{\theta}}^{(i)}(n \,|\, w)\right) \mathrm{s}\left(\hat{\boldsymbol{\theta}}^{(i)}(n \,|\, w)\right) . \tag{D.1}$$

The scalar $\alpha^{(i)} > 0$ is the distance to move from point $\hat{\boldsymbol{\theta}}^{(i)}(n \,|\, w)$ to generate the next point, $\hat{\boldsymbol{\theta}}^{(i+1)}(n \,|\, w)$. Most naturally, the solution to

$$\alpha^{(i)} = \arg\max_{\alpha > 0} \ell(\hat{\boldsymbol{\theta}}^{(i)}(n \,|\, w) + \alpha \cdot \mathbf{d}^{(i)}) \tag{D.2}$$

is desired, e.g., using *Golden Section* line search to find an approximate solution. Finally, the iterations are repeated until some *stopping criteria* is satisfied. Most naturally, this is given by

$$\| \mathrm{s}(\hat{\boldsymbol{\theta}}^{(i)}(n \,|\, w)) \| < \epsilon \,, \tag{D.3}$$

signifying a stationary point has been reached.

E General joint distribution

By the axioms of probability, the joint distribution of C, S, W, and A may be decomposed into the chain

$$\begin{aligned} f_{\mathrm{C,S,W,A}}(\mathcal{C}, t, w, \alpha | \boldsymbol{\theta}^\star) = \\ p_{\mathrm{C|S,W,A}}(\mathcal{C}|t, w, \alpha, \boldsymbol{\theta}^\star) p_{\mathrm{W,A|S}}(w, \alpha | t, \boldsymbol{\Theta}^\star) f_{\mathrm{S}}(t|\boldsymbol{\Theta}^\star) \,. \end{aligned} \tag{E.1}$$

If we remove **??**, the distribution of W and A may be dependent on S but we still they do not carry information about the true parameter index $\boldsymbol{\theta}^\star$, thus we may rewrite the above equation as

$$f_{\mathrm{C,S,W,A}}(\mathcal{C}, t, w, \alpha | \boldsymbol{\theta}^\star) = p_{\mathrm{C|S,W,A}}(\mathcal{C}|t, w, \alpha, \boldsymbol{\theta}^\star) p_{\mathrm{W,A|S}}(w, \alpha | t) f_{\mathrm{S}}(t|\boldsymbol{\theta}^\star) \,. \tag{E.2}$$

Since W and A carry no information about the true parameter index $\boldsymbol{\theta}^\star$, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ is independent of the distribution of W and A.

Proof.

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathbf{n}} &= \arg\max_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \ln p_{\mathrm{C|S,W,A}}(\mathcal{C}_i|t_i, w_i, \alpha_i, \boldsymbol{\theta}) \right. \\ &\qquad \left. + \sum_{i=1}^{n} \ln f_{\mathrm{S}}(t_i|\boldsymbol{\theta}) + \underbrace{\sum_{i=1}^{n} \ln p_{\mathrm{W,A|S}}(w_i, \alpha_i | t_i)}_{\text{constant}} \right] , \end{aligned} \tag{a}$$

where $w_i = |\mathcal{C}_i|$ and the last term is constant with respect to $\boldsymbol{\theta}$ and thus may be dropped from the maximum likelihood equation without changing the point that maximizes the likelihood. □

The sampling distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ is a function of $\mathbf{M}_{\mathbf{n}}$ (as opposed to a particular realization), and since $\mathbf{M}_{\mathbf{n}}$ is a function of W and A, the sampling distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ is also a function of W and A.

Given a statistical model, the asymptotic sampling distribution of the maximum likelihood estimator is nearly automatic. The most taxing (and uncertain) part is the *model selection*. Since we may not have much confidence in any particular model of the joint distribution of W and S, the *expected* information matrix is sketchy. In general, the *observed* information matrix is preferred. The sampling distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ converges in distribution to a multivariate normal with a mean given by $\boldsymbol{\theta}^{\star}$ and a variance-covariance given by the inverse of the *observed* information matrix, written

$$\mathbf{Y_n} \xrightarrow{d} \text{MVN}\left(\boldsymbol{\theta}^{\star}, \boldsymbol{J}_{\boldsymbol{n}}^{-1}(\boldsymbol{\theta}^{\star})\right) . \tag{E.3}$$

If an accurate distribution of W and A given S could be constructed such that it is dependent on $\boldsymbol{\theta}^{\star}$, then the distribution of W and A in a sample carries extra information about $\boldsymbol{\theta}^{\star}$. A plausible model may depend upon $\text{H}(\text{K}|\text{S} = t)$.

F Sampling distribution of functions of parameters

Suppose we have a characteristic of interest $\mathbf{g}\colon \mathbb{R}^{m\cdot q} \mapsto \mathbb{R}^{p}$ that is a function of the parametric model. The *true* value of the characteristic is given by $\mathbf{g}(\boldsymbol{\theta}^{\star})$. By the invariance property of maximum likelihood estimators, if $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}^{\star}$ then $\mathbf{g}(\hat{\boldsymbol{\theta}}_{\mathbf{n}})$ is the maximum likelihood estimator of $\mathbf{g}(\boldsymbol{\theta}^{\star})$.

By postulate 5.4, $\mathbf{Y_n}$ is a random vector drawn from a multivariate normal distribution,

$$\mathbf{Y_n} \sim \text{MVN}\left(\boldsymbol{\theta}^{\star}, \text{Var}\left[\mathbf{Y_n}\right]\right) , \tag{5.24 revisited}$$

therefore $\mathbf{g}\left(\mathbf{Y_n}\right)$ is a random vector. Under the regularity conditions (see assumption 10), $\mathbf{g}\left(\mathbf{Y_n}\right)$ is asymptotically normally distributed with a mean given by $\mathbf{g}(\boldsymbol{\theta}^{\star})$ and a variance-covariance given by $\text{Var}\left[\mathbf{g}(\mathbf{Y_n})\right]$, written

$$\mathbf{g}\left(\mathbf{Y_n}\right) \xrightarrow{d} \text{MVN}\left(\mathbf{g}\left(\boldsymbol{\theta}^{\star}\right), \text{Var}\left[\mathbf{g}(\mathbf{Y_n})\right]\right) . \tag{F.1}$$

The generative model may be used to generate samples of statistics that are functions of the true parameter index $\boldsymbol{\theta}^{\star}$. A sample drawn from $\mathbf{g}(\mathbf{Y_n})$ may be generated by drawing r maximum likelihood estimates from the sampling distribution

$$\langle \hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{1})}, \ldots, \hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{r})} \rangle , \tag{F.2}$$

and applying $\mathbf{g}$ to each, resulting in the sample

$$\langle \mathbf{g}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{1})}\right), \ldots, \mathbf{g}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{r})}\right) \rangle . \tag{F.3}$$

An estimate of the variance-covariance $\text{Var}\left[\mathbf{g}(\mathbf{Y_n}\right]$ is given by the sample covariance

$$\hat{\text{Var}}\left[\mathbf{g}\left(\mathbf{Y_n}\right)\right] = \frac{1}{r}\sum_{i=1}^{r}\left(\mathbf{g}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{i})}\right) - \mathbf{g}(\hat{\boldsymbol{\theta}}_{\mathbf{n}})\right)\left(\mathbf{g}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{i})}\right) - \mathbf{g}(\hat{\boldsymbol{\theta}}_{\mathbf{n}})\right)^{\intercal} . \tag{F.4}$$

Thus,

$$\mathbf{g}\left(\mathbf{Y_n}\right) \sim \text{MVN}\left(\mathbf{g}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}\right), \hat{\text{Var}}\left[\mathbf{Y_n}\right]\right) . \tag{F.5}$$

G Bootstrap of sample covariance

Another estimator of the variance-covariance of $\mathbf{Y_n}$ is given by the sample covariance.

Definition G.1. *Given a sample of r maximum likelihood estimates,* $\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(r)}$*, the maximum likelihood estimator of the variance-covariance of the sampling distribution of* $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ *is given by the sample covariance,*

$$\hat{\mathrm{Var}}\left[\mathbf{Y_n}\right] = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(i)} - \boldsymbol{\theta}^{\star}\right)\cdot\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(i)} - \boldsymbol{\theta}^{\star}\right)^{\mathsf{T}} . \tag{G.1}$$

The sample covariance depends upon a random sample of maximum likelihood estimates, and thus the sample covariance is a random matrix. However, by the property that maximum likelihood estimators converge in probability to the true value,

$$\hat{\mathrm{Var}}\left[\mathbf{Y_n}\right] \xrightarrow{P} \mathrm{Var}\left[\mathbf{Y_n}\right] , \tag{G.2}$$

it follows that

$$\mathbf{Y_n} \xrightarrow{d} \mathrm{MVN}\left(\boldsymbol{\theta}^{\star}, \hat{\mathrm{Var}}\left[\mathbf{Y_n}\right]\right) . \tag{G.3}$$

If only one maximum likelihood estimate is realized, $\mathbf{Y_n} = \hat{\boldsymbol{\theta}}_{\mathbf{n}}$, then the sample covariance given by **??** cannot be computed. In the *parametric Bootstrap*, the sample covariance is approximated by using $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ as an estimate of $\boldsymbol{\theta}^{\star}$ in the generative model described by **??**, i.e.,

$$\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(r)} \leftarrow \texttt{generate\_mle}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}}\right) , \tag{G.4}$$

and the sample covariance of these approximate maximum likelihood estimates is an asymptotically unbiased estimator of $\mathrm{Var}\left[\mathbf{Y_n}\right]$.

Algorithm 4: Sample covariance of a bootstrapped sample of maximum likelihood estimates.

Input:
$\hat{\boldsymbol{\theta}}_{\mathbf{n}}$, an estimate of $\boldsymbol{\theta}^{\star}$.

Output:
sample covariance of a bootstrapped sample of maximum likelihood estimates.

```
Model BoostrapCovariance
  (θ̂_n)
for i = 1 to r do
  θ̂_n^(i) ← generate_mle(θ̂_n)
end
return sample covariance of θ̂_n^(1), ..., θ̂_n^(r)
```