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 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,

$$\mathbf{Y_n} = \boldsymbol{\psi}\left(\mathbf{M_n}\right) . \tag{2.1}$$

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 $\mathrm{V}[\mathbf{Y_n}]$ (see eq. (4.16)) and a bias given by the following definition.

Definition 2.1. *The bias of a point estimator $\mathbf{Y_n}$ is given by*

$$\operatorname{bias}(\mathbf{Y_n}) = \mathbb{E}\left[\mathbf{Y_n}\right] - \boldsymbol{\theta}^\star . \tag{2.2}$$

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

$$\operatorname{MSE}(\mathbf{Y_n}) = \mathbb{E}\left[\left(\mathbf{Y_n} - \boldsymbol{\theta}^\star\right)^\intercal \left(\mathbf{Y_n} - \boldsymbol{\theta}^\star\right)\right] \tag{2.3}$$

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

Postulate 2.1. *The mean squared error of the sampling distribution of $\bar{\boldsymbol{\theta}}_{\mathbf{n}}$ as given by* **??** *is equivalent to*

$$\operatorname{MSE}(\mathbf{Y_n}) = \operatorname{tr}\left((\mathbf{Y_n})\right) + \operatorname{bias}^2\Big(\mathbf{Y_n}\Big) , \tag{2.4}$$

where $\operatorname{tr}(\boldsymbol{A})$ *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

$$\boldsymbol{CRLB_n} = \frac{1}{n} \boldsymbol{\mathcal{I}}^{-1}\left(\boldsymbol{\theta}^\star\right) . \tag{2.5}$$

By **??**, if $\bar{\boldsymbol{\theta}}_{\mathbf{n}}$ is an unbiased estimator of $\boldsymbol{\theta}^\star$, then the mean squared error is given by

$$\operatorname{MSE}(\mathbf{Y_n}) = \operatorname{tr}\big((\mathbf{Y_n})\big) . \tag{2.6}$$

By **????**, 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 $\boldsymbol{CRLB_n}$,

$$\operatorname{MSE}(\mathbf{Y_n}) \geq \operatorname{tr}(\boldsymbol{CRLB_n}) . \tag{2.7}$$

2.1 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 2.3. *A parameter index $\hat{\boldsymbol{\theta}}(n \mid w)$ that maximizes the* likelihood *of observing some realization of* $\mathbf{M_n}$ *given* $\mathrm{W} = w$ *and* A *(see assumption 8) is a maximum likelihood estimator of* $\boldsymbol{\theta}^\star$*. That is,*

$$\hat{\boldsymbol{\theta}}(n \mid w) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Omega}}{\arg\max}\, \mathcal{L}_n(\boldsymbol{\theta} | w, \alpha)\,, \tag{2.8}$$

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.,

$$\hat{\boldsymbol{\theta}}(n \mid w) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Omega}}{\arg\max}\, \ell(\boldsymbol{\theta} | w, \mathbf{m_n})\,. \tag{2.9}$$

According to Bickel [**?**], 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

$$\mathrm{s}\left(\hat{\boldsymbol{\theta}}(n \mid w) | w, \mathbf{m_n}\right) = \mathbf{0}\,, \tag{2.10}$$

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

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 **??**. Since $\hat{\boldsymbol{\theta}}(n \mid w)$ is a function of the random sample $\mathbf{M}(w,n)$, it has a *sampling distribution*.

Definition 2.4. *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 **??**. This generative model depends on `generate_msft`, the generative model for the *masked system failure time* as described by **??**. Note that in **??**, **??** may be approximated with `find_mle`, a function that numerically solves the stationary points of the maximum likelihood equations as described by **??**.

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 $\hat{\boldsymbol{\theta}}(n \mid w)$.

Input:
$\boldsymbol{\theta}^\star$, the true parameter index.
w, the cardinality of the candidate set.
n, the number of masked system failure times.

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

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

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

Postulate 2.2. *As $n \to \infty$, the sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$ converges in probability to $\boldsymbol{\theta}^\star$, written*

$$\mathbf{Y}(n \mid w) \xrightarrow{P} \boldsymbol{\theta}^\star \,, \tag{2.11}$$

since

$$\lim_{n \to \infty} \mathrm{P}\Big[\mathrm{MSE}\left(\mathbf{Y}(n \mid w)\right) < \epsilon\Big] = 1 \tag{2.12}$$

for every $\epsilon > 0$.

By **??** and eq. (4.16), the variance-covariance of the asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$ is given by

$$[\mathbf{Y}(n \mid w)] = \mathbb{E}\left[\left(\mathbf{Y}(n \mid w) - \boldsymbol{\theta}^\star\right)\left(\mathbf{Y}(n \mid w) - \boldsymbol{\theta}^\star\right)^{\mathsf{T}}\right] . \tag{2.13}$$

Postulate 2.3. *The variance-covariance of the asymptotic sampling distribution of $\hat{\boldsymbol{\theta}}(n \mid w)$ 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*

$$\boldsymbol{\Sigma_n}(\boldsymbol{\theta}^\star | w, \alpha) \equiv \frac{1}{n} \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}^\star | w, \alpha) \,. \tag{2.14}$$

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

Postulate 2.4. *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*

$$\mathbf{Y}(n \mid w) \xrightarrow{d} \operatorname{MVN}\Big(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma_n}(\boldsymbol{\theta}^\star | w, \alpha)\Big) \,. \tag{2.15}$$

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

$$\operatorname{MSE}\left(\mathbf{Y}(n \mid w)\right) = \frac{1}{n} \operatorname{tr}\left(\boldsymbol{\mathcal{I}}^{-1}\Big(\boldsymbol{\theta}^\star | w, \alpha\Big)\right) = \operatorname{tr}\left(\mathrm{CRLB}_{\mathrm{n\,w}}\right) \,. \tag{2.16}$$

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 $\mathbf{Y}(n \mid w) \xrightarrow{d} \operatorname{MVN}\left(\boldsymbol{\theta}^\star, \Sigma_{\mathrm{n}}(\boldsymbol{\theta}^\star | w, \alpha)\right)$, it follows that

$$\mathbf{Y}(n \mid w) \xrightarrow{d} \operatorname{MVN}\left(\hat{\boldsymbol{\theta}}(n \mid w), \boldsymbol{\Sigma_n}(\hat{\boldsymbol{\theta}}(n \mid w) | w, \alpha)\right) \,. \tag{2.17}$$

TODO: talk about $\mathrm{A} = \alpha$. 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 2.1. *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*

$$\mathbf{Y}(n \mid w) \overset{approx.}{\sim} \operatorname{MVN}\left(\hat{\boldsymbol{\theta}}(n \mid w), \boldsymbol{\Sigma_n}(\hat{\boldsymbol{\theta}}(n \mid w) | w, \alpha)\right) \,. \tag{2.18}$$

2.2 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

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{r} \boldsymbol{A_i}\right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \bar{\boldsymbol{\theta}}^{(i)}\right) \,, \tag{2.19}$$

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 2.5. *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*

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{r} \boldsymbol{A_i}\right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \hat{\boldsymbol{\theta}}^{(i)}\right) , \tag{2.20}$$

has a sampling distribution given by

$$\mathbf{Y_n} = \left(\sum_{i=1}^{r} \boldsymbol{A_i}\right)^{-1} \left(\sum_{i=1}^{r} \boldsymbol{A_i} \mathbf{Y}^{(i)}(n_i \,|\, w_i, \alpha_i)\right) , \tag{2.21}$$

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

$$\boldsymbol{A_i} = n_i \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star | w_i, \alpha_i) \tag{2.22}$$

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

Theorem 2.2. *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 2.3. *The variance-covariance of $\mathbf{Y_n}$ is given by*

$$[\mathbf{Y_n}] = \left(\sum_{i=1}^{r} n_i \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^\star | w_i, \alpha_i)\right)^{-1} . \tag{2.23}$$

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

$$
\begin{aligned}
[\mathbf{Y_n}] &= \left[\boldsymbol{B}\left(\sum_{i=1}^{r}\boldsymbol{A_i}\mathbf{Y}^{(i)}(n_i \mid w_i, \alpha_i)\right)\right] && \text{(b)}\\
&= \boldsymbol{B}\left[\sum_{i=1}^{r}\boldsymbol{A_i}\mathbf{Y}^{(i)}(n_i \mid w_i, \alpha_i)\right]\boldsymbol{B}^{\mathsf{T}} && \text{(c)}\\
&= \boldsymbol{B}\left(\sum_{i=1}^{r}\left[\boldsymbol{A_i}\mathbf{Y}^{(i)}(n_i \mid w_i, \alpha_i)\right]\right)\boldsymbol{B}^{\mathsf{T}} && \text{(d)}\\
&= \boldsymbol{B}\left(\sum_{i=1}^{r}\boldsymbol{A_i}\left[\mathbf{Y}^{(i)}(n_i \mid w_i, \alpha_i)\right]\boldsymbol{A_i}^{\mathsf{T}}\right)\boldsymbol{B}^{\mathsf{T}}\,. && \text{(e)}
\end{aligned}
$$

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

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

By **??**, this is equivalent to $\boldsymbol{A_i^{-1}}$. Performing this substitution results in

$$
\begin{aligned}
[\mathbf{Y_n}] &= \boldsymbol{B}\left(\sum_{i=1}^{r}\boldsymbol{A_i}\boldsymbol{A_i^{-1}}\boldsymbol{A_i}^{\mathsf{T}}\right)\boldsymbol{B}^{\mathsf{T}} && \text{(g)}\\
&= \boldsymbol{B}\left(\sum_{i=1}^{r}\boldsymbol{A_i}^{\mathsf{T}}\right)\boldsymbol{B}^{\mathsf{T}}\,. && \text{(h)}
\end{aligned}
$$

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

$$
\begin{aligned}
[\mathbf{Y_n}] &= \boldsymbol{B}\left(\boldsymbol{B}^{\mathsf{T}}\right)^{\boldsymbol{-1}}\boldsymbol{B}^{\mathsf{T}} && \text{(i)}\\
&= \boldsymbol{B} && \text{(j)}
\end{aligned}
$$

By **??**, $\boldsymbol{B}$ is given by

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

and by **??**, $\boldsymbol{A_i}$ is given by

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

Performing these substitution results in

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

□

The weighted estimator asymptotically achieves the Cramér-Rao lower-bound as given by **??** 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 2.5. *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* **??**, *written*

$$\mathbf{Y_n} \xrightarrow{d} \text{MVN}\left(\boldsymbol{\theta}^\star, \left(\sum_{i=1}^{r} n_i \boldsymbol{\mathcal{I}}\left(\boldsymbol{\theta}^\star | w_i\right)\right)^{-1}\right). \tag{2.24}$$

The generative model for $\mathbf{Y_n}$ is given by **??**.

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 $\mathbf{Y_n}$.

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

Bibliography

[] 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

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

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

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

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

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

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

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

Substituting **??** with its definition gives

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

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

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

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

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

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

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

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

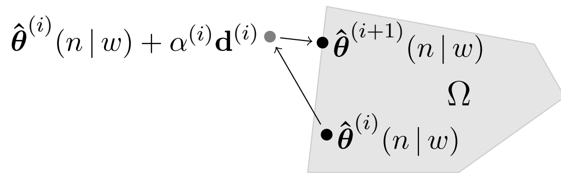
$$p_{\mathrm{K|C,S,W,A}}(k|\mathcal{C}, t, w, \alpha, \boldsymbol{\Theta}^{\star}) = \frac{\frac{R_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star})}{\binom{m-1}{w}}\left((1-\alpha)h_{\mathrm{S}}(t|\boldsymbol{\Theta}^{\star}) - (1-\alpha\frac{m}{w})\sum_{j\in\mathcal{C}} h_j(t|\boldsymbol{\Theta}_j^{\star})\right)h_k(t|\boldsymbol{\Theta}_k^{\star})}{\frac{R_{\mathrm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

The above equation may be simplified to

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

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 1: 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 **??**, $\hat{\boldsymbol{\theta}}(n \mid w)$ is the point on this surface that is at a maximum,

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

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 **??**, subject to the constraint $\boldsymbol{\theta}^\star \in \Omega$, is shown in **??**. 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**
  - $\hat{\boldsymbol{\theta}}^{(i+1)}(n \mid w) \leftarrow \texttt{project}\left(\hat{\boldsymbol{\theta}}^{(i)}(n \mid w) + \alpha^{(i)}\mathbf{d}^{(\mathbf{i})}, \Omega\right)$
  - $i \leftarrow i + 1$
- **end**
- **return** $\hat{\boldsymbol{\theta}}^{(i)}(n \mid w)$

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

In **??**, $\mathbf{d}^{(i)}$ is a unit vector denoting a *promising* direction in which to search for a better solution than $\hat{\boldsymbol{\theta}}^{(i)}(n \,|\, w)$. The Fisher scoring algorithm is a slight variation of Newton-Raphson[1] in 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

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

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. □

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

The sampling distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ is a function of $\mathbf{M_n}$ (as opposed to a particular realization), and since $\mathbf{M_n}$ 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

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

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 $\mathrm{H}(\mathrm{K}|\mathrm{S}=t)$.

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 **??**, $\mathbf{Y_n}$ is a random vector drawn from a multivariate normal distribution,

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

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 $[\mathbf{g}(\mathbf{Y_n})]$, written

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

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 $[\mathbf{g}(\mathbf{Y_n}]$ is given by the sample covariance

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

Thus,

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

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

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

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,

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

it follows that

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

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.,

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

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

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*

  $(\hat{\boldsymbol{\theta}}_\mathbf{n})$

for $i = 1$ *to* r **do**

  $\hat{\boldsymbol{\theta}}_\mathbf{n}^{(\mathbf{i})} \leftarrow \texttt{generate\_mle}(\hat{\boldsymbol{\theta}}_\mathbf{n})$

end

return *sample covariance of* $\hat{\boldsymbol{\theta}}_\mathbf{n}^{(\mathbf{1})}, \ldots, \hat{\boldsymbol{\theta}}_\mathbf{n}^{(\mathbf{r})}$