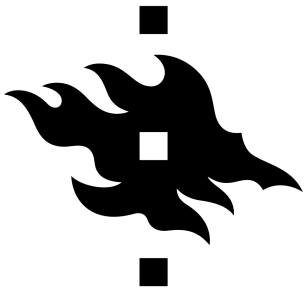
Evaluating Uncertainty with Jensen–Shannon Divergence

Ida Holopainen

A thesis presented for the degree of
Master of Philosophy

Supervisors:
Jukka Corander,
Ulpu Remes

Examiners:
Jukka Corander,
Sangita Kulathinal



UNIVERSITY OF HELSINKI

FACULTY OF SCIENCE

Department of Mathematics and Statistics
University of Helsinki
Finland

May 10, 2021

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA
MATEMATISK-NATURVETENSKAPLIGA FAKULTETEN
FACULTY OF SCIENCE

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildningsprogram – Degree programme
Faculty of Science	Master's programme in Life Science Informatics

Opintosuunta – Studieinrikting – Study track
Biostatistics and bioinformatics

Tekijä – Författare – Author
Ida Holopainen

Työn nimi – Arbetets titel – Title
Evaluating Uncertainty with Jensen–Shannon Divergence

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's thesis	May 2021	49

Tiivistelmä – Referat – Abstract
Traditional parametric statistical inference methods, such as maximum likelihood and Bayesian inference, cannot be used to learn parameter estimates if the likelihood is intractable, for example due to the complexity of the studied phenomenon. This can be overcome by using likelihood-free inference that is used with simulator-based models to learn parameter estimates. Also, traditional methods used in the estimation of uncertainties related to the parameter estimates typically require a likelihood function, and that is why these methods cannot be applied in likelihood-free inference. In this thesis, we present a novel way to compute confidence sets for parameter estimates obtained from likelihood-free inference using Jensen—Shannon divergence.

We consider two test statistics that are based on mean Jensen—Shannon divergence and propose hypothesised asymptotic distributions for them. We test whether these hypothesised distributions can be used in the computation of confidence sets for parameter estimates obtained from likelihood-free inference, and we evaluate the produced confidence sets by studying their frequentist behaviour that is summarised with coverage probabilities. We compare this frequentist behaviour between Jensen—Shannon divergence estimates and confidence sets obtained from grid evaluation of Monte Carlo estimates and from Bayesian optimisation for likelihood-free inference (BOLFI) to the ones obtained from maximum likelihood inference with Wald's and log likelihood-ratio confidence sets using three different models. We also use a simulator-based model with intractable likelihood to study the proposed confidence sets with BOLFI. In order to study the influence of observations on the parameter estimates and their confidence sets, we conducted these experiments with varying the number of observations. We show that Jensen—Shannon divergence based confidence sets meet the expected frequentist behaviour.

Avainsanat – Nyckelord – Keywords
uncertainty, confidence sets, LFI, phi-divergence, Jensen–Shannon divergence

Säilytyspaikka – Förvaringställe – Where deposited
Kumpula Campus Library

Muita tietoja – Övriga uppgifter – Additional information

Preface

The work presented in this master's thesis was performed in Bayesian statistics group in University of Helsinki. I would like to sincerely thank Professor Jukka Corander for being my supervisor for the time I spent in his research group, for giving me this extremely interesting research topic for the master's thesis, and for his guidance throughout this process. I am especially grateful to Doctor Ulpu Remes for advising me in the planning of the experiments, her comments on the thesis, and for her constant availability for guidance. I would also like to extend my gratitude to Professor Timo Koski for his and Jukka's work on the theoretical background of the Jensen–Shannon divergence test statistic as the work presented in this thesis is based on it.

I would also like to thank Professor Sangita Kulathinal for being the second examiner of this thesis, and giving me advice in the examination process. Thanks should also go to Docent Aki Havulinna for giving motivation to finish this thesis on time. Finally, I would like to thank the whole team of the Bayesian statistics group for the great working atmosphere.

In Helsinki, May 10, 2021

Contents

List of Symbols

x	Real valued variable
$\boldsymbol{x}$	Real valued vector
$\boldsymbol{x}^\top$	Transpose of a vector
$\mathbf{I}_t$	Identity matrix with size t
$\mathbf{Y}$	Vector of random variables
$\mathbf{y}$	Vector of observations
n	Number of observations
Ω	Parameter space
θ	Real valued parameter
d	Number of dimensions in the parameter vector $\boldsymbol{\theta}$
$\boldsymbol{\theta_t}$	True value of the parameter $\boldsymbol{\theta}$
$\boldsymbol{\theta_0}$	Value of $\boldsymbol{\theta}$ under null hypothesis
$\hat{\boldsymbol{\theta}}$	Estimate of the parameter $\boldsymbol{\theta}$
k	Number of event classes
$\Gamma(x)$	Gamma function
$\mathbb{E}[X]$	Expectation of random variable X
$P(A)$	Probability of event A
$A(\mathbf{y})$	Confidence set given observations

α	Chosen significance level
χ^2_k	Chi-squared distribution with k degrees of freedom
m	Number of simulations
t	Size of the training set
$\boldsymbol{p}$	Vector containing probabilities of k event classes from observed data
$\boldsymbol{q}$	Vector containing probabilities of k event classes from simulated data
$D_{JS}(\boldsymbol{p} \parallel \boldsymbol{q})$	Jensen–Shannon divergence between densities $\boldsymbol{p}$ and $\boldsymbol{q}$
$D_{KL}(\boldsymbol{p} \parallel \boldsymbol{q})$	Kullback-Leibler divergence between densities $\boldsymbol{p}$ and $\boldsymbol{q}$
$\chi^2(\boldsymbol{p} \parallel \boldsymbol{q})$	Chi-squared divergence between densities $\boldsymbol{p}$ and $\boldsymbol{q}$

Chapter 1

Introduction

1.1 Background and motivation

Statistical models are used to describe and study various phenomena with observed data. However, the model that describes the phenomenon of interest may not be fully known. These unknown properties can be for example parameters used in the statistical model. In this case, what remains unknown regarding the statistical model must be inferred using the observed data. This is called statistical inference. In order to do the inference, we must define a measure of fit between the model and observed data. One approach is to use a likelihood function which is derived from the joint probability distribution between parameter values and the observed data [25] [20].

Parametric statistical inference can be carried out with frequentist or with Bayesian inference. Frequentist approach relies on the idea of frequentist probability that describes the limiting probability of repeated trials, i.e. the statistical model can be seen as an empirical distribution as it is assumed that the observations could be collected independently and repeatedly [25]. In frequentist inference it is also assumed that the true underlying values are fixed [20]. In Bayesian setting, the parameter values are viewed as random variables, and Bayes' formula is used to infer the posterior probability distribution of the estimated parameters [20]. The posterior contains prior beliefs regarding the estimated parameters and the likelihood that the parameters produced the observed data. Even though there are differences in interpreting especially the estimated parameter of the statistical model, both of these approaches use the likelihood function in the inference process.

The inferred parameter estimates of the statistical model contain uncertainty that is caused by observations. Set of observations used in the inference is typically assumed to represent the general population of the studied phenomenon because it is impossible in general to capture all possible observations [12]. The more information we have on the

studied phenomenon, the less uncertainty there is related to the parameter estimates of the statistical model. Especially from the frequentist point of view, there is less variation between the observed estimates in repeated experiments when the number of observations increase. This type of variation between the estimates is called uncertainty.

Evaluation of the uncertainty related to the parameter estimates is crucial part of the parametric statistical inference [20]. The uncertainty of the estimated parameters is often summarised with confidence or credible sets. Confidence sets are used especially in frequentist setting, and credible sets are used in Bayesian setting. While credible sets define a proportion of the posterior probability distribution of the estimated parameters, confidence sets are areas in the parameter space that are assumed to cover the true underlying parameter value with given probability in repeated experiments. Confidence sets can be derived from likelihood based test statistics that are used in hypothesis testing, where a null hypothesis regarding the studied phenomenon is tested against an alternative one. The test statistic summarises observed data, and it has a distribution that describes the behaviour of the test statistic values computed from observations which the parameter values under the null hypothesis can produce. This distribution is usually asymptotic but in some rare cases it can be exact. The distribution can be used in the computation of confidence sets for the maximum likelihood parameter estimates at chosen confidence level that is defined as the probability that the computed confidence sets cover the true underlying parameter value.

Previously introduced methods of parametric statistical inference and estimation of the uncertainties of the parameters typically require a likelihood function. However, the likelihood function can be intractable for example due to the complexity of the studied phenomenon. The intractability of the likelihood function in parametric inference can be overcome with likelihood-free inference that is used with simulator-based statistical models [6]. In likelihood-free inference, the likelihood can be approximated or its evaluation can be bypassed. Bypassing is used especially in Bayesian likelihood-free inference where the posterior distribution for the parameters is inferred. In this thesis, we focus on likelihood-free inference methods that bypass the evaluation of the likelihood function, and that are based on the evaluation of a similarity measure between observed and simulated data [26]. The estimate for the unknown model parameters is based on the parameter values that minimize this similarity measure. This, however, leaves unaddressed a problem that we want to study in this thesis: how to estimate the uncertainties related to these estimates?

An example of a phenomenon that is modelled with simulator-based statistical model is the effect of vaccination and negative frequency-dependent selection on pneumococcal population [4]. The observed data consists of samples of pneumococcal isolates that have been genotyped by sequencing. The isolates are grouped based on the characteristics found in their genotype, and these groups are called sequence clusters. The distribution of the sequence clusters of the samples is used to summarise the observed data. This discrete

distribution can be seen also as a set of event probabilities of an isolate belonging to certain sequence cluster under selection. The similarity between these event probabilities obtained from observed and simulated data can be measured with Jensen–Shannon divergence. Thus, the Jensen–Shannon divergence can be used to infer the parameter estimates in this model. The inference based on this was used successfully and it managed to produce credible parameter estimates in previous work [4].

In this thesis we wanted to study if the Jensen–Shannon divergence can be used to compute confidence sets for the parameter estimates obtained from likelihood-free inference that bypasses the evaluation of the likelihood function. To evaluate the confidence sets, we studied if these confidence sets have the expected frequentist behaviour that can be defined with coverage probabilities. The work focuses on observations that can be summarised with discrete probability distributions, i.e. event probabilities.

1.2 Objectives of the thesis

Aim of this thesis was to study the behaviour of two novel Jensen–Shannon divergence based test statistics, and to see if these test statistics could be used in the computation of confidence sets for parameter estimates that minimize the Jensen–Shannon divergence. The computation of the confidence sets was based on asymptotic distributions that were hypothesised for the two test statistics, and we studied if the confidence sets based on these hypothesised asymptotic distributions have the frequentist property that is expected from them.

Answers to these research questions were obtained by studying the frequentist behaviour of the proposed test statistics. The study consisted of inferring the parameters in three models that had tractable likelihood functions, and in one simulator-based model with intractable likelihood function. These experiments were done for all the models using likelihood-free parameter estimation using Jensen–Shannon divergence, either using Monte Carlo estimates or with Bayesian optimization for likelihood free inference (BOLFI) method [13]. Maximum likelihood inference was also used with the models with the tractable likelihood. Frequentist behaviour of the confidence sets of the proposed test statistics and their hypothetical approximative distribution was compared with log likelihood-ratio and Wald's confidence sets that were computed for models with tractable likelihood. The comparison was based on coverage probabilities that are defined as a frequency that the confidence sets include data producing parameter value. Confidence sets are used to summarise the uncertainty related to the inferred parameter estimates that is caused by the limitation of observed data. That is why we also wanted to study the effect of the observations on the resulting parameter estimates and their confidence sets by varying the number of observations in all of the experiments. In the case of

the simulator-based model, that was the negative frequency-dependent selection model (NFDS), the number of observations was varied by using different model settings.

1.3 Structure of the thesis

The remainder of this thesis is divided into three chapters. In Chapter 2 the theory related to the used methods is presented, and it is divided into two parts. It first introduces the maximum likelihood, Bayesian and likelihood-free inference, from which two estimation methods are presented: inference with Monte Carlo estimates and Bayesian optimization for the likelihood-free inference (BOLFI) (Section 2.1). The Chapter 2 then proceeds to discuss about confidence sets, and presents two likelihood based routines to compute the confidence sets: log likelihood-ratio and Wald's confidence sets (Section 2.2). Finally, the new test statistic and confidence sets studied in this thesis are presented in Section 2.3. A toy model is used in this chapter as a running example to demonstrate inference process with the maximum likelihood estimation, Monte Carlo estimates, and with BOLFI (Section 2.1). Also, it is used to demonstrate the computation of the confidence sets with log likelihood-ratio, Wald's confidence sets and with the new Jensen–Shannon divergence based confidence sets (Section 2.2–2.3). Related work is discussed in Section 2.4.

Chapter 3 shows the results from the repeated experiments that are done for three different models: the toy model, log linear model and for the negative frequency-dependent selection model (NFDS). The first two models have a tractable likelihood, and the frequentist behaviour of the likelihood based confidence sets and Jensen–Shannon based confidence sets are compared. The third model is simulator based and it is used in the repeated experiments to evaluate the proposed method, and finally, the computation of the confidence sets for the paramater estimates obtained from BOLFI for this model with real pneumococcal data is applied.

Chapter 4 the relevant results are summarised and suggestions for the future research are discussed, and concludes the thesis.

Chapter 2

Theory

2.1 Statistical inference

Statistics are used to describe and quantify properties of any phenomena or process of interest with observed data. Data is a collection of quantified features observed from the phenomenon of interest. Observations are assumed to contain random variation and uncertainty, and are considered as realizations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ of random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. In statistical inference, collected data with prior information regarding the phenomenon can be used to construct a statistical model that is used to explain the phenomenon. Especially in parametric statistical inference the inferred properties are unknown parameter values of the statistical model. Henceforth, the inferred properties are considered as unknown parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \Omega$, $d \in \mathbb{N}$. As an example, statistical model can be a joint probability distribution that models the variation of the observations, and the inferred properties can be unknown parameter values of the used joint distribution. However, especially in biological context, the modelled phenomena can be more complex. One example of a complex phenomenon is the negative frequency-dependent selection that has been observed in pneumococcal population after introduction of vaccination, which was studied with a simulator based model introduced by Corander et al. [4].

Next, we will describe the three different approaches of inference: maximum likelihood inference, Bayesian inference, and likelihood-free inference. From the perspective of this thesis, the maximum likelihood and the likelihood-free inference are the most relevant inference methods, and the use of these methods is demonstrated with a toy model.

2.1.1 Maximum likelihood inference

Maximum likelihood inference is often used in frequentist inference. In frequentist inference the true values in parameter vector $\boldsymbol{\theta}$ are considered to be fixed but unknown. Frequentist inference relies on the frequentist probability in which the probability is defined as the limit of many trials. Hence, the statistical model models the empirical probability distribution.

In maximum likelihood (ML) inference the properties of the underlying probability distribution are inferred by maximizing the likelihood function $L : \Omega \to \mathbb{R}$ which is a function of parameter values of the joint probability density function for fixed data, $L(\boldsymbol{\theta}) := f_{\mathbf{Y}}(\boldsymbol{\theta}; \mathbf{y})$ [16]. The likelihood function describes the likelihood that the chosen parameter values explain the observed data, i.e. the fit of the statistical model to the observed data. The estimate $\hat{\boldsymbol{\theta}}$ is the value of the parameter vector $\boldsymbol{\theta}$ that explains the observed data $\mathbf{y}$ best as it maximizes the likelihood function:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} L(\boldsymbol{\theta}). \tag{2.1}$$

The obtained estimate is assumed to be a random variable, and under some regularity condititions it is also assumed to follow asymptotically a d-dimensional multivariate normal distribution where the mean vector is the true value of the parameter vector, and the covariance matrix is the inverse matrix of the Fisher information matrix [16]:

$$\hat{\boldsymbol{\theta}} \underset{as}{\sim} \mathcal{N}\left(\boldsymbol{\theta}_t, \mathbf{i}(\boldsymbol{\theta}_t)^{-1}\right), \tag{2.2}$$

where $\boldsymbol{\theta}_t$ contains the true, data producing values of the parameter vector. The Fisher information matrix for variable vector $\boldsymbol{\theta}$ is defined as

$$\mathbf{i}(\boldsymbol{\theta}) = \begin{pmatrix} i_{1,1}(\boldsymbol{\theta}) & \dots & i_{1,d}(\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ i_{d,1}(\boldsymbol{\theta}) & \dots & i_{d,d}(\boldsymbol{\theta}) \end{pmatrix}, \tag{2.3}$$

which constitutes of the Fisher information for each entry of the parameter vector,

$$i_{i,j}(\boldsymbol{\theta}) = \mathbb{E}\left[-\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_i \partial \theta_j}\right], \quad i, j = 1, \dots, d. \tag{2.4}$$

The ML estimate can be seen as consistent estimate which is described with the limiting behaviour of the probability that the distance between the ML estimate and true value would be greater than any $\epsilon > 0$ [16]

$$P(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t\| > \epsilon) \to 0, \quad \text{when } n \to \infty.$$

This also can be viewed that the number of observations n affects the precision of the ML estimate: the estimate becomes more precise with greater number of observations.

Running example with the toymodel 2.1.1. Here, we present a toy model that is used throughout this chapter to demonstrate maximum likelihood and likelihood-free inference methods, and the computation of the confidence intervals for the resulting parameter estimates. In this example, we present the likelihood function of the toy model that is used to infer the maximum likelihood estimates for parameter $\theta \in \mathbb{R}$. The parameter θ is used to compute the event probabilities for multinomial distribution. Multinomial distribution models observed counts or frequencies in k categories.

Let g be a function $g : \mathbb{R} \times \mathbb{N} \to [0, 1]$, that defines the entries of the probability vector $\boldsymbol{p} = (p_1, \ldots, p_k)$, where

$$g(\theta, i) = \frac{\exp(-\theta|1 - i|)}{\sum_{i=1}^{k} \exp(-\theta|1 - i|)} = p_i, \quad i = 1, \ldots, k. \tag{2.5}$$

Here, the normalizing term guarantees that these p_i probabilities sum up to 1. When $\theta > 0$ the probabilities are sorted in descending order, $p_1 > \cdots > p_k$, when $\theta = 0$ the probabilities are same for each class, and when $\theta < 0$, the probabilities are in increasing order, i.e. $p_1 < \cdots < p_k$. These probabilities can be used as event probabilities $\boldsymbol{p} = (p_1, \ldots, p_k)$ in the multinomial distribution with k classes. The observations $\mathbf{y}$ are observed counts in each k class. The likelihood function of this multinomial distribution becomes:

$$L(\boldsymbol{p}; \mathbf{y}) = \frac{n!}{\prod_{i=1}^{k} y_i!} \prod_{i=1}^{k} p_i^{y_i} = \frac{\Gamma(n+1)}{\prod_{i=1}^{k} \Gamma(y_i + 1)} \prod_{i=1}^{k} p_i^{y_i}. \tag{2.6}$$

The log likelihood function of the multinomial distribution becomes

$$l(\boldsymbol{p}; \mathbf{y}) = \log \left(\frac{\Gamma(n+1)}{\prod_{i=1}^{k} \Gamma(y_i + 1)} \prod_{i=1}^{k} p_i^{y_i} \right) \tag{2.7}$$

$$= \log\left(\Gamma(n+1)\right) - \sum_{i=1}^{k} \log\left(\Gamma(y_i + 1)\right) + \sum_{i=1}^{k} y_i \log(p_i). \tag{2.8}$$

We did thousand repeated experiments where we inferred the parameter estimates for simulated data using maximum likelihood and likelihood-free inference methods. We also wanted to study the effect of the observations on the estimation results, and to study this, we did the repeated experiments by simulating data with four different number of observations. The number of observations, n, was set to 50, 100, 500, and to 1000. The

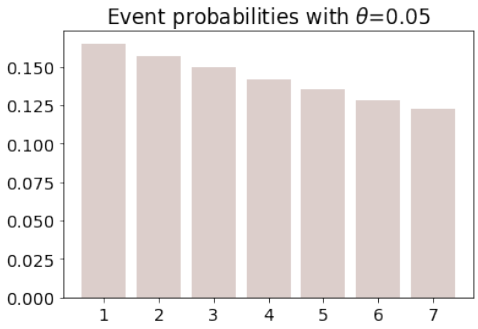


Figure 2.1: The event probabilities for each seven categories of the multinomial distribution with chosen $\theta = 0.05$.

parameter θ was set to 0.05, and the number of categories k was set to 7. The event probabilities for each categories with chosen θ and k can be seen in Figure 2.1. As the parameter θ is positive, it can be seen that the event probabilities of the classes are in decreasing order, and do not contain zero probabilities.

Here, we show the maximum likelihood estimates of parameter θ that were obtained from the repeated experiments. The Figure 2.2 presents the distributions of the maximum likelihood estimates for each number of observations. It can be seen that the distributions are centred at the true value of the parameter θ, and the variance observed in the distribution decreases as the number of observations increase. This can be explained by the effect that the increase in the number of observations has on the estimated event probabilities, which approach to the event probabilities generated by the parameter θ. The resulting distributions of the estimates confirm that the obtained maximum likelihood estimates from the these repeated experiments behave in a way that is expected from the maximum likelihood estimates.

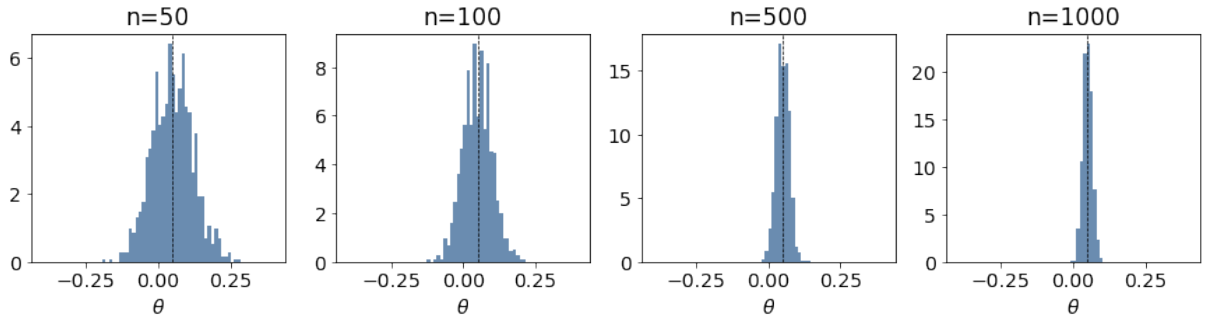


Figure 2.2: The distribution of the maximum likelihood from 1000 repeated experiments with different number of observations used, $n \in (50, 100, 500, 1000)$.

2.1.2 Bayesian inference

In Bayesian inference the Bayes' theorem is applied to infer the posterior distribution for the parameter $\boldsymbol{\theta}$ because the parameter is assumed to be a realisation of a random variable following unknown distribution [11]. In contrast to maximum likelihood inference that produces a point estimate, the Bayesian inference produces a posterior distribution that is the probability distribution of the parameter vector $\boldsymbol{\theta}$ given the prior belief and likelihood that is based on observed data. The prior distribution $p(\boldsymbol{\theta})$ summarises the prior information or belief regarding the parameter $\boldsymbol{\theta}$, and the likelihood function $p(\mathbf{y} \mid \boldsymbol{\theta})$ describes the conditional probability on data $\mathbf{y}$ given parameter vector $\boldsymbol{\theta}$. Applying this information to the Bayes' theorem, the posterior distribution becomes:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{2.9}$$

$$\propto p(\mathbf{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{2.10}$$

Above in Equation (2.9) $p(\mathbf{y})$ is the marginal likelihood that is considered as a normalizing constant to ensure the unit integral over the parameter space Ω:

$$p(\mathbf{y}) = \int p(\boldsymbol{\theta})\, p(\mathbf{y} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{2.11}$$

Sometimes the marginal likelihood is too complex to compute or intractable, and that is why the relative posterior distribution in Equation (2.10) is used in stead [11].

The inferred posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ can be used to deduce properties regarding the estimated parameter $\boldsymbol{\theta}$ and uncertainty related to the inference. For example, one of the deduced properties can be a point estimate of the parameter, $\hat{\boldsymbol{\theta}}$, which can be the mode of the posterior, known as maximum a posterior, or the mean of the posterior distribution [11].

2.1.3 Likelihood-free inference

Likelihood-free inference is a branch of both Bayesian and frequentist inference which is applied especially in the field of biology to model complex systems. Normally, for these types of complex systems the likelihood function cannot be constructed because it is either unknown, too difficult to derive, or computationally too expensive. When the likelihood is intractable, the usual parametric statistical inference methods cannot be applied, such as maximum likelihood inference. However, with likelihood-free inference the parameter estimation can be done without using the true likelihood function, and several methods have been proposed in the literature [6].

Likelihood-free inference is used with simulator-based models to infer the estimates of the parameters. Simulator is a computer program that generates data $\mathbf{y}_\theta$ from given parameter values $\boldsymbol{\theta} \in \Omega$ by random sampling. The simulator defines a statistical model between the sample space and the set of possible probability distributions on the sample space [6]. As the simulated system can be complex, the simulator can contain several internal states or latent variables, z_i, which depend on the given parameter values $\boldsymbol{\theta}$, i.e. $z_i \sim P(z_i \mid \boldsymbol{\theta})$ [6]. This results that the generated data $\mathbf{y}_\theta \sim P(\mathbf{y}_\theta \mid z_i, \boldsymbol{\theta})$.

The workflow in Bayesian likelihood-free inference depends on the chosen method. In this thesis, the focus is on the methods that draw samples θ from the defined prior, generate simulated data $\mathbf{y}_\theta$ based on drawn samples, and finally approximate the posterior distribution by comparing the similarity between simulated and the observed data. The similarity measure is a defined discrepancy that can be computed from chosen summary statistics. In this thesis we focus on categorical data that can be summarised with event probabilities of each class k, $\boldsymbol{p} = (p_1, \ldots, p_k)$. The similarity measure that is used throughout the thesis in the comparison of the event probabilities between the observed and simulated data is Jensen–Shannon divergence $D_{JS}(\boldsymbol{p} \parallel \mathbf{q})$ that is defined as

$$D_{JS}(\boldsymbol{p} \parallel \boldsymbol{q}) = \frac{1}{2} D_{KL}\left(\boldsymbol{p} \parallel \boldsymbol{m}\right) + \frac{1}{2} D_{KL}\left(\boldsymbol{q} \parallel \boldsymbol{m}\right). \tag{2.12}$$

Here D_{KL} refers to the Kullback–Leibler divergence, $\boldsymbol{p}$ and $\boldsymbol{q}$ refer to the event probabilities of observed and simulated data respectively, and $\boldsymbol{m} = \frac{1}{2}(\boldsymbol{p} + \boldsymbol{q})$. Kullback–Leibler divergence in discrete case is defined as

$$D_{KL}(\boldsymbol{p} \parallel \boldsymbol{q}) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right). \tag{2.13}$$

The similarity measure is used with an acceptance threshold to approximate the posterior distribution [26]. So in this case the evaluation of the likelihood function is bypassed but there exists alternative methods where the likelihood function is evaluated. The differences between the two methods are discussed in literature [6]. Example of the methods that bypass the evaluation of the likelihood function is approximated Bayesian computation (ABC) which constitutes of several computational methods [18]. Also, in this thesis we have used grid evaluation with Monte Carlo estimates and Bayesian optimization for likelihood-free inference (BOLFI) that are discussed next.

Likelihood-free inference with Monte Carlo estimates

A simple but possibly computationally demanding method to conduct likelihood-free inference is to compute Monte Carlo estimate over a specified grid G that is a collection

of values of the estimated parameter vector $\boldsymbol{\theta}$, i.e. $G = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_l)$ where $l \in \mathbb{N}$. The computational cost of computing the Monte Carlo estimate depends on the number of evaluated points in the grid and the used sample size m.

The Monte Carlo estimate is based on the Monte Carlo expectation that describes the concentration of the probability mass of the distribution that the realization of the random variables follow [3]. The expectation of an estimated quantity at $\boldsymbol{\theta}$ is

$$\mu(\boldsymbol{\theta}) = \mathbb{E}\left[f(\mathbf{x}_{\boldsymbol{\theta}})\right]. \tag{2.14}$$

Here $f(\mathbf{x}_{\boldsymbol{\theta}})$ is the estimated quantity that is computed from the sampled data at $\boldsymbol{\theta} \in G$, $\mathbf{x}_{\boldsymbol{\theta}} = (x_{\boldsymbol{\theta}}^{(1)}, \ldots, x_{\boldsymbol{\theta}}^{(m)})$, that's distribution is known. The Monte Carlo expectation can be estimated by computing the sample mean of independently and identically sampled data

$$\hat{\mu}(\boldsymbol{\theta}) = \frac{1}{m} \sum_{l=1}^{m} f(\mathbf{x}_{\boldsymbol{\theta}}^{(l)}). \tag{2.15}$$

The sample mean is known to converge in distribution to the previously presented expectation with large enough m by the law of large numbers [3].

For the purposes used in this thesis, the Monte Carlo estimate is used to evaluate the mean of Jensen–Shannon divergences computed from the m simulated data generated at $\boldsymbol{\theta} \in G$. The divergence is formed from the the summary statistics generated at parameter value $\boldsymbol{\theta}$. The summary statistics are the expected event probabilities of each class k computed from observed and simulated data where $\hat{\boldsymbol{p}}$ refers to the expected event probabilities of the observed data, and $\hat{\boldsymbol{q}}_{\boldsymbol{\theta}}^{(l)}$, $l = 1, \ldots, m$, contains the expected event probabilities of each class k from the lth simulated data at the parameter $\boldsymbol{\theta}$. The estimated event probabilities depend on the number of observations. The Monte Carlo estimate of the Jensen–Shannon divergence becomes

$$\hat{D}_{JS}(\boldsymbol{\theta}) = \frac{1}{m} \sum_{l=1}^{m} D_{JS}(\hat{\boldsymbol{p}} \parallel \hat{\boldsymbol{q}}_{\boldsymbol{\theta}}^{(l)}). \tag{2.16}$$

The estimate of the parameter vector $\boldsymbol{\theta}$ is now the value within the grid G that minimizes the Monte Carlo estimate of the Jensen–Shannon divergence:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in G}{\arg\min}\, \hat{D}_{JS}(\boldsymbol{\theta}). \tag{2.17}$$

Bayesian optimization for likelihood-free inference

Here we present a likelihood-free inference method that has better sample and computational efficiency compared for example to grid evaluation of Monte Carlo estimates.

Bayesian optimization for likelihood-free inference (BOLFI) is a inference method used in engine for likelihood-free inference (ELFI) [13], [17]. BOLFI uses Bayesian optimization that is used in finding the maximum or minimum values of functions that form is unknown but which output can be computed. In BOLFI the unknown function is assumed to follow a Gaussian process and it is modelled with surrogate model. Here, the surrogate model is defined as expected discrepancy $\Delta_{\boldsymbol{\theta}}$ that is obtained from the posterior distribution of the discrepancy. In BOLFI the surrogate model can be used to infer the parameter estimate $\hat{\boldsymbol{\theta}}$ that is the minimizing value of the expected discrepancy, or it can be used to construct an approximative posterior distribution for the parameter vector $\boldsymbol{\theta}$.

The surrogate model used in BOLFI is a Gaussian process which models the function of discrepancies between the observed and simulated data, $f : \Omega \to \mathbb{R}$ and $f(\boldsymbol{\theta}) := \Delta_\theta$. The function f is assumed to follow Gaussian process. Gaussian process is defined as a finite collection of random variables that have a joint Gaussian distribution [23]. Gaussian process constitutes of chosen prior mean function, $m : \Omega \to \mathbb{R}$ and chosen kernel, $k : \Omega \times \Omega \to \mathbb{R}$ [23]. The mean function and kernel are defined as

$$\begin{aligned} m(\boldsymbol{\theta}) &= \mathbb{E}\left[f(\boldsymbol{\theta})\right] \\ k(\boldsymbol{\theta}, \boldsymbol{\theta'}) &= \mathbb{E}\left[(f(\boldsymbol{\theta}) - m(\boldsymbol{\theta}))(f(\boldsymbol{\theta'}) - m(\boldsymbol{\theta'}))\right], \end{aligned}$$

and the discrepancy function can be written as

$$f(\boldsymbol{\theta}) \sim \mathcal{GP}(m(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta'})). \tag{2.18}$$

In Bayesian optimization the beliefs about the modelled function are updated by acquiring set of new data $\xi_i := (\boldsymbol{\theta}_i, \Delta_{\boldsymbol{\theta}_i})$, where $i = 1, \ldots, t$ and $t \in \mathbb{N}$ is the total size of the data to be acquired, that maps the sampled parameter values to the computed discrepancy. The new acquired data is applied with the current belief on Bayes' theorem, and the new data is used to update the joint Gaussian distribution which dimensionality increases for every new set of data. Also, the hyperparameters of the chosen mean function and kernel are updated as the new data is acquired. Example of a popular kernel function is the squared exponential (SE) kernel that is defined as

$$k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \sigma^2 \exp\left(-\frac{\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2}{2l^2}\right). \tag{2.19}$$

Here are two hyperparameters that are estimated: first parameter is σ^2 that evaluates the variability of the kernel function, and l is the lengthscale of the parameter $\boldsymbol{\theta}$ that is used to extrapolate the kernel function behaviour over the parameter space. For example, if the variance term σ^2 is large, the Gaussian process model is assumed to have lot of variability from the mean function, and large values of the lengthscale parameter l covering

the parameter range would imply that there is extremely small variability in the values that the kernel function gets over the parameter range. As the variance and lengthscale parameters of the kernel function are inferred in Bayesian manner to explain the observed values of the discrepancy function, prior information regarding the possible values of these hyper parameters can be used to avoid overfitting.

After acquiring training set with the size of t the assumed distribution of the discrepancy function given the data points becomes

$$\mathbf{f}_t \mid \boldsymbol{\theta}_t \sim \mathcal{N}(\mathbf{m}_t, \mathbf{K}_t + \sigma^2 \mathbf{I}_t), \tag{2.20}$$

where $\mathbf{f}_t$ is the vector of discrepancy values at iteratively acquired parameter values $\boldsymbol{\theta}_t$: $\mathbf{f}_t = (f(\boldsymbol{\theta}_1), \ldots, f(\boldsymbol{\theta}_t))^\top$, and $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t)$. The resulting $\mathbf{m}_t$ is a vector constituting of the values of the mean function at $\boldsymbol{\theta}_t$, and the chosen kernel defines the entries of a positive definite covariance matrix $\mathbf{K}_{i,j} := k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$, for $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \boldsymbol{\theta_0}$:

$$\mathbf{m}_t = \begin{pmatrix} m(\boldsymbol{\theta}_1) \\ \vdots \\ m(\boldsymbol{\theta}_t) \end{pmatrix}, \quad \mathbf{K}_t = \begin{pmatrix} k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) & \ldots & k(\boldsymbol{\theta}_1, \boldsymbol{\theta_0}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{\theta}_t, \boldsymbol{\theta}_1) & \ldots & k(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t) \end{pmatrix}. \tag{2.21}$$

During the process of acquisition of new data, the selection of the next set of parameter values is based on the chosen acquisition function. If the currently collected data consists of the set $(\xi_1, \ldots, \xi_i)$, where $i < t$, the information from the current predictive mean $\mu_i(\boldsymbol{\theta})$ and variance $v_i(\boldsymbol{\theta}) + \sigma^2$ of the joint Gaussian distribution is used to find the next parameter value $\boldsymbol{\theta}_{i+1}$ and compute the discrepancy $f(\boldsymbol{\theta}_{i+1})$ which is used to update the joint Gaussian distribution. Because the chosen acquisition method affects how the sampled parameter values are chosen and because the size of training data affects the size of the joint Gaussian, they also affect the accuracy of the approximated discrepancy. However, the acquisition function is used to increase the sample efficiency which reduces the size of required training data.

Predictive function of the discrepancy f at $\boldsymbol{\theta} \in \Omega$, given the training data set $\boldsymbol{\xi}^{(t)} = (\xi_1, \ldots, \xi_t)$, follows d-variate normal distribution [13]:

$$f(\boldsymbol{\theta}) \mid \boldsymbol{\xi}^{(t)} \sim \mathcal{N}(\mu_t(\boldsymbol{\theta}), v_t(\boldsymbol{\theta}) + \sigma^2). \tag{2.22}$$

Here the posterior predictive mean $\mu_t(\boldsymbol{\theta})$ and variance $v_t(\boldsymbol{\theta}) + \sigma^2$ are obtained from the mean function, kernel, and covariance matrix in which $\mathbf{k}_t(\boldsymbol{\theta}) = (k(\boldsymbol{\theta}, \boldsymbol{\theta}_1), \ldots, k(\boldsymbol{\theta}, \boldsymbol{\theta_0}))^\top$ [23]:

$$\mu_t(\boldsymbol{\theta}) = m(\boldsymbol{\theta}) + \mathbf{k}_t(\boldsymbol{\theta}, \boldsymbol{\theta}')^\top (\mathbf{K}_t + \sigma^2 \mathbf{I})^{-1} (\mathbf{f}_t - \mathbf{m}_t), \tag{2.23}$$

$$v_t(\boldsymbol{\theta}) = k(\boldsymbol{\theta}, \boldsymbol{\theta}) - \mathbf{k}_t(\boldsymbol{\theta})^\top (\mathbf{K}_t + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_t(\boldsymbol{\theta}). \tag{2.24}$$

Information obtained from this posterior predictive distribution of the function of discrepancy is used in the approximation of the likelihood function and in the inference of the parameter estimates. The approximate likelihood $\hat{L}_u^{(t)}$ is constructed from a chosen threshold, posterior predictive mean and variance, and from the cumulative distribution of standard normal distribution [13]:

$$\hat{L}_u^{(t)}(\boldsymbol{\theta}) \propto \Phi\left(\frac{h - \mu_t(\boldsymbol{\theta})}{\sqrt{v_t(\boldsymbol{\theta}) + \sigma_n^2}}\right). \tag{2.25}$$

Here h is the chosen threshold, $\mu_t(\boldsymbol{\theta})$ is the posterior mean and $v_t(\boldsymbol{\theta}) + \sigma_t^2$ is the posterior variance of the Gaussian process. Function $\Phi(x)$ is the cumulative distribution of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du. \tag{2.26}$$

The approximated likelihood $\hat{L}_u^{(t)}$ is used with prior $p(\boldsymbol{\theta})$ to compute the approximative posterior distribution for the parameter $\boldsymbol{\theta}$ [13]:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\ &\approx \hat{L}_u^{(t)}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}). \end{aligned}$$

The parameter estimates $\hat{\boldsymbol{\theta}}$ are inferred by finding the minimizing value of the posterior predictive mean of the discrepancy [13]:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Omega} \mu_t(\boldsymbol{\theta}). \tag{2.27}$$

Running example with the toymodel 2.1.2. In this example, the maximum likelihood estimates are compared to the estimates obtained from the likelihood-free inference, i.e. minimum Jensen–Shannon divergence estimates computed with Monte Carlo estimates and with BOLFI. The estimates were obtained from 1000 repeated experiments conducted for varying number of observations in which $n \in (50, 100, 500, 1000)$. The evaluation range for θ was set to $[-0.5, 2]$. The maximum likelihood estimates were obtained by minimizing the negative likelihood function within the parameter range. The Monte Carlo estimates were computed over the specified grid set by the parameter range with number of steps of 750. In BOLFI the parameter range was used to define a uniform prior distribution for the parameter θ, and chosen kernel was the SE kernel presented in Equation (2.19) with added bias. With single estimated parameter this kernel becomes:

$$k(\theta_i, \theta_j) = \sigma^2 \exp\left(-\frac{(\theta_i - \theta_j)^2}{2l^2}\right) + b, \tag{2.28}$$

where b is the bias term. We also set a Gamma prior distribution, with shape and rate parameters of 2 and 5 respectively, for the lengthscale l.

The Figure 2.3 presents the distribution of the estimates obtained from each set of experiments. It can be seen that the distribution of the four set of experiments coincide between the used estimation methods. The width of the distribution gets narrower as the number of observations increases. This is explained by the limiting behaviour of the observations: the number of counts in each class increase and produce values of event probabilities that are closer to the expected values as the number of observations increase.

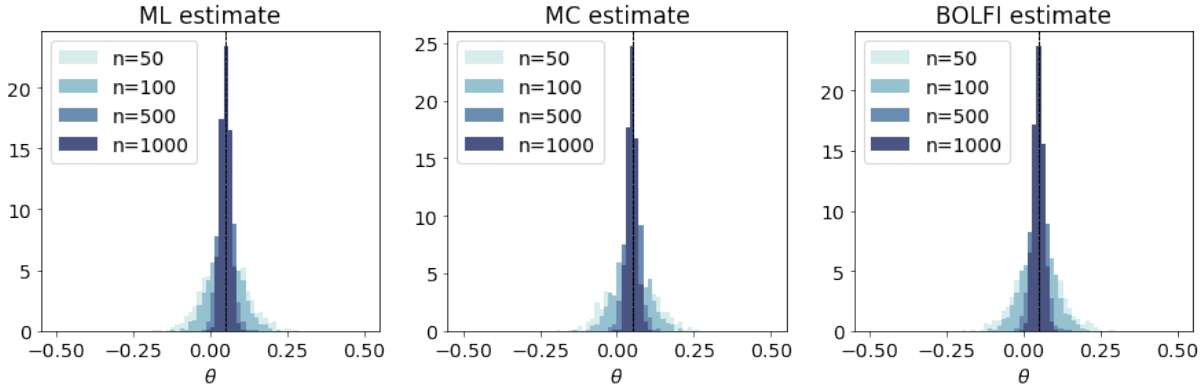


Figure 2.3: The normalized histograms present the distributions of the maximum likelihood estimates, Monte Carlo estimates, and BOLFI estimates of the experiments where the number of observations n was varying. Dashed line represents the true value of the parameter, $\theta = 0.05$.

2.2 Evaluating the uncertainty of the estimates

Earlier different estimation methods were presented and demonstrated with the toy model. This section describes the difference between confidence and credible sets and presents two routines that depend on likelihood function to compute the confidence sets. These are demonstrated with the toy model and later compared with the confidence sets based on Jensen–Shannon divergence. Confidence and credible intervals, also known as Bayesian confidence intervals, are used to summarise and measure the uncertainty related to the estimates of the model parameters. Even though the purpose of using both confidence and credible intervals is the same, the interpretation differs between them.

Confidence sets, or intervals when the estimated parameter is one-dimensional, are used in frequentist inference. The confidence in this context describes the probability that the confidence sets computed from the realizations of random vector $\mathbf{Y}$ cover the true parameter value, and that is why the confidence can be seen as epistemic probability. Confidence sets are defined as a subset of the parameter space based on the observations $\mathbf{y}$, $\mathbf{A}(\mathbf{y}) \subset \Omega$. The probability that the true value of the parameter of interest is within this region is at least the value set by the confidence level that is defined as $1 - \alpha$ where the chosen $\alpha \in [0, 1]$:

$$P(\boldsymbol{\theta} \in A(\mathbf{y})) \geq 1 - \alpha. \tag{2.29}$$

It can be seen from the definition and from Equation (2.29) that confidence sets can be constructed in several ways. One of the ways is to use test statistics since there exists duality between the confidence sets and test statistics. The test statistics together with p-values are also used to summarize the confidence or uncertainty related to the parameter estimates.

Credible sets are used in Bayesian inference and they are defined as a certain portion of the posterior density of the estimated parameter. Therefore the credible sets are based on observations $\mathbf{y}$ but also incorporate prior information regarding the estimated parameter vector $\boldsymbol{\theta}$ in a form of prior distribution. The credible sets summarize that given the observed data and prior information, the true parameter value, that is a realization of the random variable, belongs to this set with certain probability of $1 - \alpha$ set by $\alpha \in [0, 1]$:

$$P(\boldsymbol{\theta} \in A(\mathbf{y})) = \int_{A(\mathbf{y})} p(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta} \geq 1 - \alpha. \tag{2.30}$$

Since the credible sets are defined as certain proportion of the posterior density, they can be computed in several ways as well. Examples of credible sets are equal-tailed set and highest density set.

The emphasis in this thesis is in the computation of confidence sets with the novel suggested method based on Jensen–Shannon divergence. In order to compare the results of this method to the other well known routines to compute the confidence sets, two of them are presented next: likelihood ratio confidence sets, and Wald's confidence sets.

2.2.1 Confidence sets constructed from likelihood ratio test

Likelihood ratio test (LRT) is a test statistic which can be used to compute the confidence intervals for the maximum likelihood estimate. The likelihood ratio test statistic $T(\mathbf{y})$ is used to compare the fit of the statistical model, i.e. chosen likelihood function, between

the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ and null hypothesis $\boldsymbol{\theta}_0 \in \Omega_0 \subset \Omega$ given data $\mathbf{y}$:

$$T(\mathbf{y}) = 2\log\left(\frac{L(\hat{\boldsymbol{\theta}};\mathbf{y})}{L(\boldsymbol{\theta}_0;\mathbf{y})}\right) = 2\left[l(\hat{\boldsymbol{\theta}};\mathbf{y}) - l(\boldsymbol{\theta}_0;\mathbf{y})\right]. \tag{2.31}$$

The LRT statistic follows χ^2_d distribution, where d is the dimensionality of the parameter vector $\boldsymbol{\theta}$, under the null hypothesis. The asymptotic distribution of the test statistic can be used to construct confidence set for the maximum likelihood estimate given the observations $\mathbf{y}$. The confidence set is constructed by evaluating the area in the parameter space that produces values of LRT that are smaller than the value on the x axis of the asymptotic distribution that produces the right side probability of given confidence level α. This value is referred as $\chi^2_d(\alpha)$. The confidence sets become:

$$\begin{aligned} A(\mathbf{y}) &= \{\boldsymbol{\theta} : 2[l(\hat{\boldsymbol{\theta}};\mathbf{y}) - l(\boldsymbol{\theta};\mathbf{y})] < \chi^2_d(\alpha)\} \\ &= \{\boldsymbol{\theta} : l(\hat{\boldsymbol{\theta}};\mathbf{y}) - l(\boldsymbol{\theta};\mathbf{y}) < \frac{1}{2}\chi^2_d(\alpha)\}. \end{aligned} \tag{2.32}$$

The Equation (2.32) is also known as deviance, and it describes the deviance from the null hypothesis [25]. Here, parameter values far from the ML estimate produce large deviance values, and values close to ML estimate produce small values.

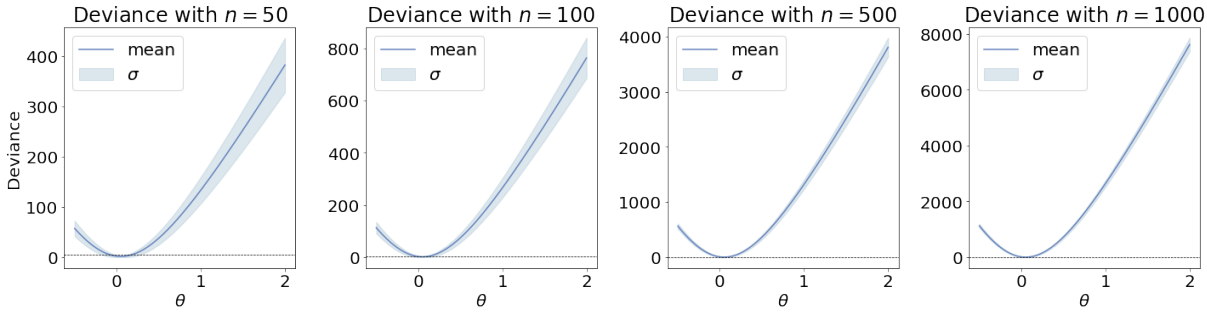


Figure 2.4: Mean and standard deviation σ of the deviance computed from thousand repeated experiments along the parameter range set for θ. The dashed line represents the acceptance threshold for the 95% confidence interval.

Running example with the toymodel 2.2.1. This example presents the confidence intervals computed for the parameter θ in the toy model using likelihood ratio. Several values of confidence levels α were chosen and used in all experiments, $\alpha \in (0.50, 0.10, 0.05, 0.01)$. The likelihood ratio was computed using the deviance presented Equation (2.32). The

Table 2.1: The coverage probabilities of the repeated experiments for the likelihood ratio confidence intervals computed for the toy model.

n	50				100				500				1000			
Expected	0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99
LR	0.49	0.91	0.95	0.99	0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99	0.49	0.90	0.95	0.99

Figure 2.4 shows the mean deviance over the grid with the added standard deviation from the mean. It can be seen that the number of observations affects values of the deviance.

Table 2.1 presents the computed coverage probabilities of the 1000 repeated experiments. Coverage probabilities define the proportion of the confidence intervals in the repeated experiments that cover the true value of the estimated parameter θ, in this case the value of 0.05. The coverage probabilities were computed by comparing whether the value of the test statistic, here the deviance, at the null was within the set acceptance threshold, here $\frac{1}{2}\chi^2_d(\alpha)$. It can be seen from the Table 2.1 that the observed coverage probabilities coincide well with the expected probabilities.

2.2.2 Confidence sets constructed from Wald's test statistic

The $100(1-\alpha)\%$ confidence set constructed from Wald's test statistic is based on the difference between the ML estimate and null hypothesis, and is centred around the parameter estimate $\hat{\boldsymbol{\theta}}$ due to the asymptotic normality of the estimate as seen in Equation (2.2). The Wald's test statistic for the estimate $\hat{\boldsymbol{\theta}}$ given observations $\mathbf{y}$ is defined as:

$$w(\mathbf{y}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \mathbf{i}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{2.33}$$

where $\mathbf{i}(\hat{\boldsymbol{\theta}})$ is the Fisher information matrix at $\hat{\boldsymbol{\theta}}$ presented in Equation 2.4. This test statistic follows asymptotically χ^2_d distribution under the null hypothesis. Now the approximative confidence ellipsoid for parameter $\hat{\boldsymbol{\theta}}$ is the region of the parameter space for which the values of the test statistic are smaller than the value producing right hand probability of α in the χ^2_d distribution, $\chi^2_d(\alpha)$:

$$A(\mathbf{y}) = \{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathbf{i}(\hat{\boldsymbol{\theta}})^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) < \chi^2_d(\alpha)\}. \tag{2.34}$$

Here $\mathbf{i}(\hat{\boldsymbol{\theta}})^{-1}$ is the inverse of the Fisher information matrix computed at the ML estimate $\hat{\boldsymbol{\theta}}$. In practice, the computation of the confidence ellipsoid with a statistical model that meets adequate regularity conditions is based on the computation of the eigenvalues λ and eigenvectors $\boldsymbol{v}$ of the inverse Fisher information matrix:

$$\mathbf{i}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{2.35}$$

The eigenvalues and eigenvectors define the principal axes of the ellipsoid. The eigenvalues affect length of the axes with desired α level of the χ^2_d distribution, $\chi^2_d(\alpha)$, and the vectors affect the possible rotation of the ellipsoid. If the ellipsoid is rotated, it means that the model parameters are correlated [20].

When $d = 1$, i.e. only single parameter needs to be estimated for the statistical model, the Wald's confidence interval can be computed in easier manner. Since Wald's test statistic follows a χ^2_1 distribution with single parameter, the square root of the statistic follows the standard normal distribution, $i(\hat{\theta})^{1/2}(\hat{\theta} - \theta_0) \sim N(0, 1)$, under null hypothesis. Below, $z_{\alpha/2}$ is the value on the x axis of standard normal distribution which produces the right side probability of $\alpha/2$:

$$A(\mathbf{y}) = \{\theta : i(\hat{\theta})^{1/2}|\hat{\theta} - \theta| < z_{\alpha/2}\}. \tag{2.36}$$

This holds for $\theta \in (\hat{\theta} - \frac{z_{\alpha/2}}{i(\hat{\theta})^{1/2}}, \hat{\theta} + \frac{z_{\alpha/2}}{i(\hat{\theta})^{1/2}})$. Hence the $100(1-\alpha)\%$ Wald's confidence interval centered around the ML estimate becomes

$$\mathbf{w}^{1/2} = \left(\hat{\theta} - \frac{z_{\alpha/2}}{i(\hat{\theta})^{1/2}}, \quad \hat{\theta} + \frac{z_{\alpha/2}}{i(\hat{\theta})^{1/2}} \right). \tag{2.37}$$

Running example with the toymodel 2.2.2. In this example the Wald's confidence intervals are derived for the toy model. Since the toy model contains only one estimated parameter θ, the confidence intervals can be computed directly by applying the Equation (2.37).

Fisher information for $\hat{\theta}$ can be computed from the second derivative of the log likelihood function defined in (2.7). By differentiating the log likelihood function $l(g(\theta), \mathbf{y})$ twice we get

$$l''(g(\theta), \mathbf{y}) = -n \left[\frac{\sum_{j=0}^{k-1} j^2 \exp(-\theta j)}{\sum_{j=0}^{k-1} \exp(-\theta j)} - \left(\frac{\sum_{j=0}^{k-1} j \exp(-\theta j)}{\sum_{j=0}^{k-1} \exp(-\theta j)} \right)^2 \right].$$

Fisher information at $\hat{\theta}$ is then

$$i(\hat{\theta}) = \mathbb{E}[-l''(\hat{\theta})] = n \left[\frac{\sum_{j=0}^{k-1} j^2 \exp(-\theta j)}{\sum_{j=0}^{k-1} \exp(-\theta j)} - \left(\frac{\sum_{j=0}^{k-1} j \exp(-\theta j)}{\sum_{j=0}^{k-1} \exp(-\theta j)} \right)^2 \right].$$

The value of the Fisher information at the estimate can be applied to the Equation (2.37) with the desired confidence levels $\alpha \in (0.50, 0.10, 0.05, 0.01)$. Table 2.2 below presents the

coverage probabilities from the thousand repeated experiments for each number of observations n. What can be seen from the table is that the computed coverage probabilities are coincide well with the expected values since the observed maximum deviance from the expected value is 0.01.

Table 2.2: The coverage probabilities of the repeated experiments for the Wald's confidence intervals computed for the toy model.

n	50				100				500				1000			
Expected	0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99
Wald	0.49	0.91	0.95	0.99	0.50	0.90	0.95	0.99	0.50	0.91	0.95	0.99	0.49	0.90	0.95	0.99

2.3 Evaluation of confidence intervals with Jensen–Shannon divergence

Confidence intervals based on likelihood ratio and Wald's test statistics can be used when there is information available about the likelihood function. However, with complex simulator-based model this information is not available. This section presents the Jensen–Shannon divergence based test statistics, their hypothesized distributions, and the computation of the confidence sets for the minimum mean Jensen–Shannon divergence estimate using these distributions.

2.3.1 ϕ-divergence

ϕ-divergence is used to measure the difference between two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. In discrete case the divergence from $\boldsymbol{p} = (p_1, \ldots, p_k)$ to $\boldsymbol{q} = (q_1, \ldots, q_k)$ is defined as

$$D_\phi(\boldsymbol{p} \parallel \boldsymbol{q}) = \sum_{i=1}^{k} q_i \phi\left(\frac{p_i}{q_i}\right), \tag{2.38}$$

where q_i and p_i are entries of the discrete probability distributions $\boldsymbol{p}$, and $\boldsymbol{q}$, and ϕ is a convex function [21]. Here, we assume that if $q_i = 0$, then $p_i = 0$, and $0\phi\left(\frac{0}{0}\right) = 0$.

The choice of function ϕ leads to different divergences. Examples of these divergences are Kullback-Leibler divergence, D_{KL}, Jensen–Shannon divergence, D_{JS}, and χ^2-divergence. The Jensen–Shannon divergence is defined in Equation (2.12) that is used in this chapter with the χ^2 divergence to define the test statistic used in the construction of

the confidence sets. The χ^2 divergence with chosen $\phi(x) = (x-1)^2$ from $\boldsymbol{p}$ to $\boldsymbol{q}$ is defined as

$$\chi^2(\boldsymbol{q} \parallel \boldsymbol{p}) = \sum_{i=1}^{k} \frac{(q_i - p_i)^2}{p_i^2}. \tag{2.39}$$

2.3.2 Construction of the Jensen–Shannon divergence statistic

As mentioned in the Section 2.3.1 the Jensen–Shannon divergence is constructed from the ϕ-divergence, and it is used to compare the fit between two probability distributions. Here, we use the Jensen–Shannon divergence to compare the fit between observed and simulated data that can be summarised with distributions that summarise the event counts in k class. These event probabilities are estimated from the data by dividing the number of observed events in each class with the total number of observations, and hence the estimates depend on the number of observations. The expected event probabilities of the observed and simulated data are referred as $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$ respectively. In this work, we assume that there are no zero probabilities, i.e. $p_i, q_i > 0$ for all $i = 1, \ldots, k$.

The theory of the Jensen–Shannon divergence statistic presented here is based on the theoretical results obtained in the article [5]. The expected event probabilities $\hat{\boldsymbol{p}}$ of the observations define the null hypothesis for the Jensen–Shannon divergence statistic that assumes that the expected event probabilities are same as the true data producing event probabilities, i.e. $\hat{p}_i = p_i^0$, $i = 1, \ldots, k$ where p_i^0 is the ith event probability of the true data producing event probability vector $\boldsymbol{p}^0 = (p_1^0, \ldots, p_k^0)$. However, the estimated event probabilities can be seen as deviations from the true underlying event probabilities that have produced the data: $\hat{p}_i = p_i^0 + \frac{c_i}{\sqrt{n}}$, where $\sum_{i=1}^{k} c_i = 0$ [5]. This defines the set of alternative hypotheses that assume $\hat{p}_i = p_i^0 + \frac{c_i}{\sqrt{n}}$.

The Jensen–Shannon divergence statistic is constructed from the χ^2 divergence that is known to approximate the ϕ-divergences:

$$8nD_{JS}(\hat{\boldsymbol{p}} \parallel \hat{\boldsymbol{q}}) \approx n\chi^2(\hat{\boldsymbol{q}} \parallel \hat{\boldsymbol{p}}), \tag{2.40}$$

where n is the number of observations [5]. Here the $n\chi^2(\hat{\boldsymbol{q}} \parallel \hat{\boldsymbol{p}})$ divergence is defined as

$$n\chi^2(\hat{\boldsymbol{q}} \parallel \hat{\boldsymbol{p}}) = n\sum_{i=1}^{k} \frac{(\hat{q}_i - \hat{p}_i)^2}{\hat{p}_i} = \sum_{i=1}^{k} \frac{(\xi_i - n\hat{p}_i)^2}{n\hat{p}_i} \tag{2.41}$$

where $\xi_i = n\hat{q}_i$. Under the null hypothesis this has the asymptotic non-central χ^2 distribution with $k-1$ degrees of freedom and non-centrality parameter δ:

$$\delta := n\sum_{i=1}^{k} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0}. \tag{2.42}$$

The test statistic can also be constructed from the sum over m Jensen–Shannon divergences generated at $\boldsymbol{\theta}$ [5]. The test statistic of the sums becomes

$$S^{(m)}(\boldsymbol{\theta}) := \sum_{l=1}^{m} 8nD_{JS}\left(\hat{\boldsymbol{p}} \parallel \hat{\boldsymbol{q}}_{\boldsymbol{\theta}}{}^{(l)}\right) \approx \sum_{l=1}^{m} n\chi^2\left(\hat{\boldsymbol{q}}_{\boldsymbol{\theta}}{}^{(l)} \parallel \hat{\boldsymbol{p}}\right). \tag{2.43}$$

For this test statistic, the proposed approximative distribution at the null is a non-central $\chi^2_{m(k-1)}$ distribution with the non-centrality parameter $m\delta$ [5].

Mean of JSD statistics

The aim of the thesis is to test the hypothesis of the proposed test statistic based on the sample mean of the Jensen–Shannon divergences. The sample mean of the Jensen–Shannon divergences can be constructed from the sum statistic $S^{(m)}(\boldsymbol{\theta})$:

$$\overline{D}^{(m)}_{JS}(\boldsymbol{\theta}) = \frac{8n}{m}\sum_{l=1}^{m} D_{JS}\left(\hat{\boldsymbol{p}} \parallel \hat{\boldsymbol{q}}_{\boldsymbol{\theta}}{}^{(l)}\right). \tag{2.44}$$

As the statistic $S^{(m)}(\boldsymbol{\theta})$ is known to follow the non-central $\chi^2_{m(k-1)}$ distribution with the non-central parameter δ, the expectation and the variance of the statistic become

$$\mathbb{E}\left[S^{(m)}(\boldsymbol{\theta})\right] = m(k-1) + m\delta \tag{2.45}$$

$$\operatorname{var}(S^{(m)}(\boldsymbol{\theta})) = 2m(k-1) + 4m\delta. \tag{2.46}$$

The expectation and variance for the mean Jensen–Shannon divergence can be obtained by applying the properties of expectation and variance. After multiplying with the scale parameter $1/m$, the expectation and variance become:

$$\mathbb{E}\left[\frac{1}{m}S^{(m)}(\boldsymbol{\theta})\right] = (k-1) + \delta \tag{2.47}$$

$$\operatorname{var}\left(\frac{1}{m}S^{(m)}(\boldsymbol{\theta})\right) = \frac{2(k-1)}{m} + \frac{4\delta}{m}. \tag{2.48}$$

It can be seen that as the sample size increases, the variance decreases, and especially $\operatorname{var}\left(\frac{1}{m}S^{(m)}\right) = 0$ when $m \to \infty$. However, the increase of m does not affect the expectation, it can be seen as a linear transformation of the random variable $\delta \sim \chi^2_{(k-1)}$ with fixed constant $k-1$. This implies that the sample mean test statistic follows approximatively $\chi^2_{(k-1)}$ distribution with location parameter $k-1$.

2.3.3 Construction of the confidence intervals

Assuming that the test statistic of the mean Jensen–Shannon divergence follows the hypothesized approximative $\chi^2_{(k-1)}$ distribution with location parameter $k-1$ under the null hypothesis, it can be used to construct the confidence sets in such way that proportion of the confidence sets contain the true value $\boldsymbol{\theta_0}$ approaches to $1-\alpha$. The confidence sets become:

$$A(\mathbf{y}) = \{\boldsymbol{\theta} : \frac{8n}{m} \sum_{l=1}^{m} D_{JS}\left(\hat{\boldsymbol{p}} \parallel \hat{\boldsymbol{q}}_{\boldsymbol{\theta}}{}^{(l)}\right) < \chi^2_{k-1}(\alpha)\}, \tag{2.49}$$

where $\mathbf{y}$ refers to the observed data that has been used to compute the observed event probabilities $\boldsymbol{p}$ and $\chi^2_{k-1}(\alpha)$ refers to the value that produces the right side probability α in the shifted χ^2_{k-1} distribution with the location parameter $k-1$. However, there exists special case when the true parameter is not included to the confidence sets which can lead to empty confidence sets: in this case the computed minimum mean Jensen–Shannon divergence estimate exceeds the critical value set by $\chi^2_{k-1}(\alpha)$. This can happen when the observed data does not match to the expected simulated data at any $\boldsymbol{\theta} \in \Omega$.

The normalized version of the mean Jensen–Shannon divergence statistic, $\overline{D}^{(m)}_{NJS}$, is constructed by subtracting the minimum value of the mean Jensen–Shannon divergence, i.e. the mean value observed at the estimate $\hat{\boldsymbol{\theta}}$, and thus the normalized minimum value of the mean Jensen–Shannon divergence becomes 0, similarly with the log likelihood-ratio test statistic:

$$\overline{D}^{(m)}_{NJS}(\boldsymbol{\theta}) = \overline{D}^{(m)}_{JS}(\boldsymbol{\theta}) - \overline{D}^{(m)}_{JS}(\hat{\boldsymbol{\theta}}). \tag{2.50}$$

Normalization guarantees that the minimum Jensen–Shannon divergence estimate is included in the confidence sets, and thus the confidence sets based on normalized values do not contain empty sets even though the true parameter would not be included to them. The hypothesized distribution for the normalized mean Jensen–Shannon divergence statistic is χ^2_d, where d is the dimension of the parameter vector $\boldsymbol{\theta}$. Thus, the hypothesized confidence sets become:

$$A(\mathbf{y}) = \{\boldsymbol{\theta} : \overline{D}^{(m)}_{NJS}(\boldsymbol{\theta}) < \chi^2_d(\alpha)\}. \tag{2.51}$$

Theoretical basis of this approximative distribution of the normalized mean statistic under the null hypothesis is outside of the scope of the thesis.

Running example with the toymodel 2.3.1. In this example, the resulting coverage probabilities of the mean Jensen–Shannon divergence based confidence intervals computed using the assumed hypothesized distributions for the toy model are examined. Figure 2.5

demonstrates the behaviour of the mean JSD statistic over the chosen range of parameter θ in the Monte Carlo and BOLFI experiments with the acceptance threshold of 95% confidence interval. It can be seen that the number of observations n affects the values of Jensen–Shannon divergence. Table 2.3 presents the observed coverage probabilities from Monte Carlo experiments and from BOLFI. The coverage probabilities were computed using both hypothesized approximative χ^2 distributions. It can be seen that the observed coverage probabilities seem to follow the expected values especially with larger confidence levels. However, the observed coverage probabilities for 50% confidence intervals were smaller than the expected value, especially with the normalized Monte Carlo estimates of the Jensen–Shannon divergence. The results of these experiments are presented more thoroughly in the Chapter 3.

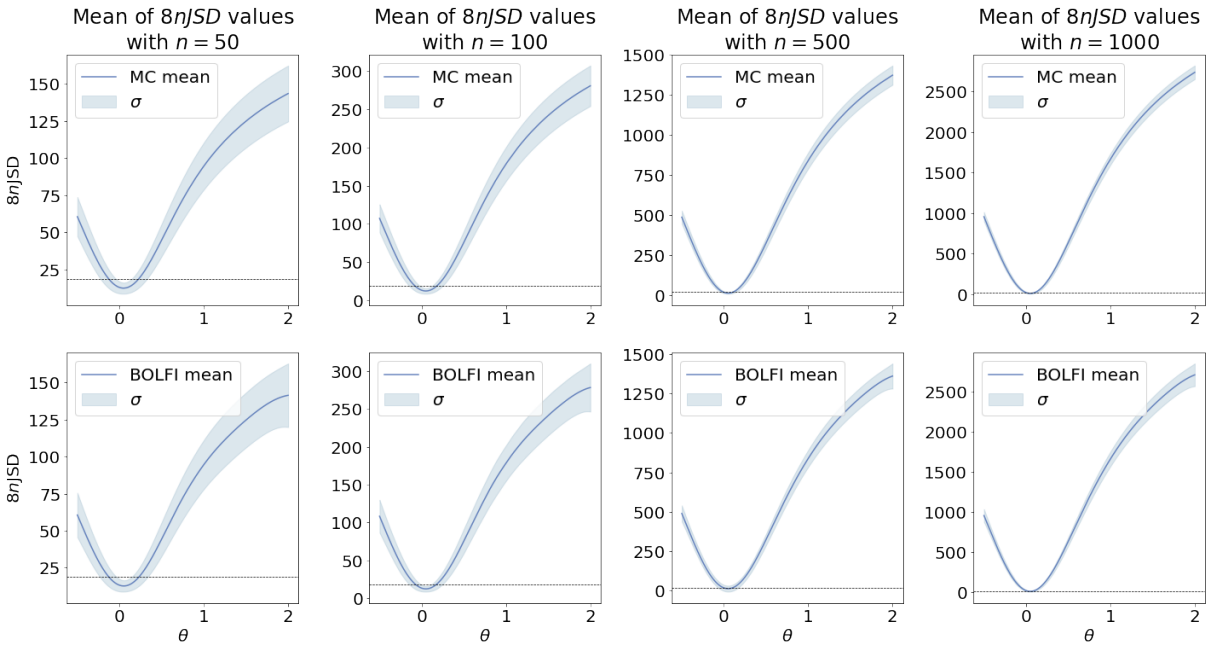


Figure 2.5: Mean curve of the mean Jensen–Shannon divergence and standard deviation σ computed from thousand repeated toy model experiments along the parameter range set for θ. The first row presents the mean curve of Monte Carlo experiments and second row presents similar mean curve obtained from BOLFI experiments. The dashed line represents the acceptance threshold of 95% confidence interval.

Table 2.3: Here is presented the coverage probabilities of the Jensen–Shannon divergence (D_{JS}) based confidence sets from thousand repeated experiments conducted with different number of observations n. The experiments were done using grid evaluation of the Monte Carlo estimates (MC) and BOLFI. The approximative distributions of χ^2_{k-1} with shifting parameter $k-1$ and χ^2_d for normalized values were applied, where $k-1=6$ and $d=1$.

		MC		BOLFI	
n	Expected	D_{JS} χ^2_{k-1}	D_{JS} χ^2_d	D_{JS} χ^2_{k-1}	D_{JS} χ^2_d
50	0.50	0.46	0.30	0.43	0.47
	0.90	0.88	0.89	0.86	0.88
	0.95	0.93	0.94	0.92	0.93
	0.99	0.98	0.99	0.98	0.99
100	0.50	0.49	0.32	0.47	0.47
	0.90	0.89	0.89	0.87	0.88
	0.95	0.95	0.94	0.94	0.94
	0.99	0.98	0.99	0.99	0.99
500	0.50	0.50	0.41	0.46	0.47
	0.90	0.90	0.90	0.88	0.90
	0.95	0.95	0.95	0.93	0.95
	0.99	0.99	0.99	0.99	0.98
1000	0.50	0.47	0.41	0.42	0.47
	0.90	0.88	0.89	0.85	0.90
	0.95	0.94	0.96	0.93	0.95
	0.99	0.99	0.99	0.99	0.99

2.4 Related work

In this thesis, we study the uncertainty related to the minimum Jensen–Shannon divergence estimates by using repeated experiments to study the frequentist behaviour of these estimates and their confidence sets. In this section, we discuss previous research related to this research topic. Even though there is not previous research especially related to Jensen–Shannon divergence based confidence sets, there exists still research related to some ϕ-divergence based test statistics that will be discussed here. We also discuss some other methods that have been proposed or used to compute the confidence sets in likelihood-free inference, especially in approximate Bayesian computation.

There has been lot of research related to ϕ-divergence based test statistics. However, majority of these test statistics are used to test some specific feature of data. For example, one study shows how ϕ-divergence test statistic can be used to test symmetry structure observed in contingency tables [22]. Another study shows how ϕ-divergence statistic can be used to test for likelihood ratio ordering between independent multinomial samples, and how this test statistic can be considered as an extension of the likelihood ratio and χ^2 test statistics [19]. Also, one study presented a family of ϕ-divergence test statistics that can

be used to test goodness of fit. In the same study, the confidence bands for the evaluated distribution function were studied [15]. The research topic that was closest from the work presented in this thesis, presents an empirical ϕ-divergence test statistic that is used to test simple and composite null hypotheses [2]. The same article also presents a way to compute the confidence interval in the case of single parameter. The most interesting part of this article is related to the proposed test statistic and its approximative distribution that are similar to what we have proposed for the normalized version of the Jensen–Shannon divergence test statistic, even though their test statistic is based on Kullback-Leibler divergence.

There has been lot of methods proposed in the literature to compute confidence sets for the likelihood-free estimates. Currently, methods that rely on using approximate Bayesian computation with likelihood ratio odds estimation to construct the approximate likelihood, tend to use likelihood ratio test statistic to compute confidence intervals [6], [7]. This has been proposed also in frequentist setting [9]. Another method using approximate Bayesian computation suggested to compute confidence sets from approximate confidence distributions that were constructed from rejection sampling [27]. Another approach to compute confidence sets is using Monte Carlo samples, even though this method has been criticised to be more computationally expensive. One example of this approach approximated the minimax expected size confidence sets [24]. However, maybe the most well known method to compute the approximative confidence sets is bootstrapping that has been widely used due to its simplicity [10].

Chapter 3

Results

3.1 Used models

This section introduces two multivariable models that were used in the experiments in addition to the toy model that has been introduced and used as running example in Chapter 2. The first of the two models is a log linear model that models cell counts in 2-way table, and the second one is a simulator-based negative frequency-dependent selection (NFDS) model. In all of these models, the observed data can be summarised as event probabilities that can be used to compute the Jensen–Shannon divergence.

3.1.1 Log linear model

Log linear model can be used to model observed counts in contingency tables with n observations [1]. Especially with the 2-way tables, the log linear model becomes:

$$\log(\mu_{i,j}) = \lambda + X_i\lambda^X + Y_j\lambda^Y + X_iY_j\lambda^{XY} \tag{3.1}$$

The model has three real valued parameters λ^X, λ^Y, and λ^{XY} which are used with the overall effect parameter $\lambda \in \mathbb{R}$ to compute estimates of the observed counts, $\mu_{i,j}$, in each cell of the 2-way table. The parameter λ ensures the sum over expected counts in each cell equals to the number of all observations, n, parameters λ^X and λ^Y model the effect of X and Y variables respectively. The parameter λ^{XY} models the possible association between the two variables. If parameter $\lambda^{XY} \neq 0$, then the generated log linear model is known as saturated model.

In the experiments the entries of 2-way table of two effect coded variables $X = (1, -1)$ and $Y = (1, -1)$, presented in Table 3.1, are modelled with the log linear model presented in Equation (3.1). For the contingency table presented in Table 3.1, the expected counts

become

$$\log(\mu_{i,j}) = \begin{cases} \lambda + \lambda^X + \lambda^Y + \lambda^{XY}, & \text{for } (1,1) \\ \lambda + \lambda^X - \lambda^Y - \lambda^{XY}, & \text{for } (1,2) \\ \lambda - \lambda^X + \lambda^Y - \lambda^{XY}, & \text{for } (2,1) \\ \lambda - \lambda^X - \lambda^Y + \lambda^{XY}, & \text{for } (2,2). \end{cases} \tag{3.2}$$

The overall effect λ can be computed using the free parameters λ^X, λ^Y, and λ^{XY} with number of observations:

$$\lambda = \log\left(\frac{n}{S}\right), \tag{3.3}$$

where S is the sum of exponentials of the parameters describing the effect of variables X and Y, and the possible association parameter in each cell of the table

$$S = \exp(\lambda^X+\lambda^Y+\lambda^{XY})+\exp(\lambda^X-\lambda^Y-\lambda^{XY})+\exp(-\lambda^X+\lambda^Y-\lambda^{XY})+\exp(-\lambda^X-\lambda^Y+\lambda^{XY}).$$

Table 3.1: 2×2 contingency table for variables $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ with n observations.

	$Y_1 = 1$	$Y_2 = -1$	
$X_1 = 1$	$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,+}$
$X_2 = -1$	$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,+}$
	$\mu_{+,1}$	$\mu_{+,2}$	n

The parameters λ^X, λ^Y, and λ^{XY} model the expected cell counts $\mu_{i,j}$ in the 2×2 contingency table, i.e. how many of the n observations belong to certain cell of the contingency table. The expected counts can be used to compute the expected frequencies, $\mathbf{p} := (\mu_{1,1}/n, \mu_{1,2}/n, \mu_{2,1}/n, \mu_{2,2}/n)$, and these can be used in the multinomial likelihood function as event probabilities. The multinomial likelihood function, $L(\mathbf{p}; \mathbf{y})$, and the logarithm of it, $l(\mathbf{p}; \mathbf{y})$, become

$$L(\mathbf{p}; \mathbf{y}) = \frac{n!}{\prod_{i=1}^4 y_i!} \prod_{i=1}^4 p_i^{y_i}, \tag{3.4}$$

$$l(\mathbf{p}; \mathbf{y}) = \log\Gamma(n+1) - \sum_{i=1}^4 \log\Gamma(y_i+1) + \sum_{i=1}^4 y_i \log p_i. \tag{3.5}$$

The log linear model was applied in two ways: first the model was applied with two parameters by setting the association parameter λ^{XY} to zero, and then using the saturated version of the model. For the first set of experiments, the parameter values λ^X and λ^Y were set to -0.25 and 0.15 respectively. The saturated log linear model was applied by setting λ^X, λ^Y and λ^{XY} to -0.20, 0.10, and 0.4 respectively.

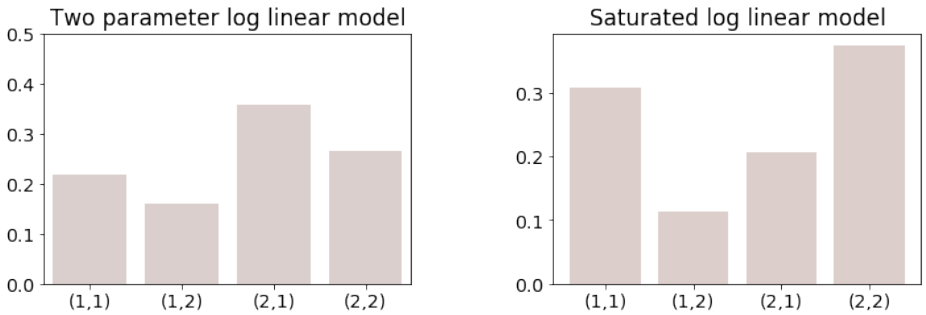


Figure 3.1: Bar plots showing the event probabilities of each class in both models. The parameter values used to compute the event probabilities in the two parameter log linear model were $\left(\lambda^X, \lambda^Y\right) = (-0.25, 0.15)$. For the saturated log linear model the parameters were set to $\left(\lambda^X, \lambda^Y, \lambda^{XY}\right) = (-0.20, 0.10, 0.4)$.

3.1.2 Negative frequency-dependent selection model

Under negative frequency dependent selection (NFDS), rare alleles are associated with a positive selection pressure that affects the population dynamics. This mechanism is used to explain the heterogeneity observed for example in bacterial population located in human body where novel antigen appears among bacteria individuals. First, the antigen causing mutation becomes more frequent in the population as it is not recognized by the antibodies. However, the novel antibody is not beneficial to a bacterium individual anymore once it is recognised more frequently by the immune responses. However, the more this antigen is recognised by the immune response, the less beneficial it becomes to a bacterium individual.

The Figure 3.2 presents the vaccine and non-vaccine type frequencies among the sequence clusters at each time point observed from Massachusetts data that was used in previous work [4]. The number of observations in the Massachusetts data was 616. What can be seen from the Figure 3.2 is that the population dynamics change after the introduction of the vaccine. The effect of the vaccination can be seen as the decrease of the vaccine type isolates at later time points, and as increase of the non-vaccine type isolates especially in the sequence clusters containing the highest portion of the observed isolates.

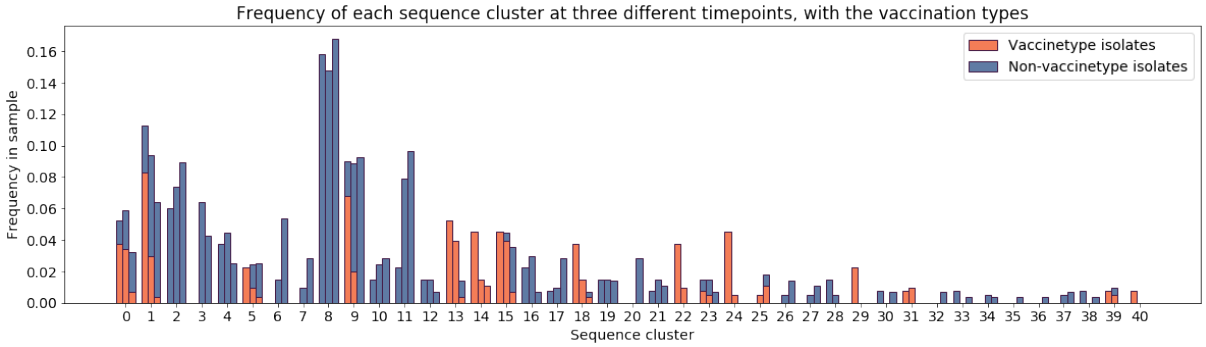


Figure 3.2: The observed frequencies of the vaccine (orange) and non-vaccine (blue) type isolates among the 41 sequence cluster. The three bars at each sequence cluster present the observed frequencies at each three time point when samples were collected. First bar refers to early sample that has been collected just after the introduction of the vaccine, second one is a mid sample and third one is late sample.

The experiments in which the computation of confidence sets based on Jensen–Shannon divergence is applied, use a simulator-based model that describes the effect of NFDS and vaccination on pneumococcal population [4]. The homogenous-rate multilocus model of NFDS used here describes the reproductive success of an individual within a population under NFDS. In homogenous-rate multilocus model the number of offspring arising from individual i at time t is modelled as a Poisson-distributed random variable $X_{i,t}$:

$$X_{i,t} \sim Poi\left(\left(\frac{\kappa}{N_t}\right)(1-m)\,(1-v_i)\,(1+\sigma_f)^{\pi_{i,t}}\right). \tag{3.6}$$

Here, the first term corresponds to the general density-dependent selection in which the carrying capacity κ was assumed to be 10^5, and N_t is the population size at time t. The second term describes the effect of migration into the population that affected the reproductive fitness of an individual, in which the m is the migration rate. The third term describes the vaccine effect on the reproductive fitness of an individual if the isolate has vaccine serotype: here v_i has value of parameter v if it has vaccine serotype, otherwise v_i is set to zero. The last term describes the pressure of NFDS to an isolate, where σ_f is the selection pressure, and the exponent $\pi_{i,t}$ describes the deviation from the equilibrium genotype in isolate i at time t that is computed from the frequencies of the accessory genes of each isolate. When the last term is greater than one, NFDS has positive effect to the isolate, i.e. the isolate contains more rare genes compared to the equilibrium state, and when the term is less than one, it has negative effect to the isolate as it contains genes that are more common. When the last term is one, NFDS does not have effect on

the reproducibility of the isolate.

We tested the Jensen–Shannon divergence based confidence sets on parameter estimates inferred using the simulator-based NFDS model, which did not have tractable likelihood. In the first part of the experiments, we used simulated data that was initialised with the Massachusetts data. The parameters estimated in the experiments were vaccine selection pressure v, selection pressure σ_f, and immigration rate m. The data producing values were set to be the estimates that are inferred earlier with BOLFI for the data collected from Massachusetts [8]: v was set to be 0.07280, σ_f was set to be 0.00743, and m was set to be 0.00548. In order to study the effect of observations on the inferred parameter estimates and on their confidence sets, we used three different settings with the NFDS model by varying either the number of time points when the isolates were collected, or by varying the number of collected isolates at each time point. In the first setting, 250 isolates were collected at three time points, in the second setting 250 isolates were collected at six time points, and in the third setting 1000 isolates were collected at three time points. The first setting is closest to what we have observed with the Massachusetts data, and that is why it is the most interesting to us. In the second part of the experiments, we inferred the parameter estimates for the same parameters using the Massachusetts data, and computed the Jensen–Shannon divergence based confidence intervals for the estimates, and compared the results with credible intervals computed in previous work [8].

3.2 Results

Here, we examine the inference results obtained from each experiment, and evaluate the performance of the used confidence sets. This is done by examining the observed distributions of the Jensen–Shannon divergence test statistics, and by comparing the coverage probabilities of the confidence sets obtained from each method to their expected values. The $100(1-\alpha)\%$ confidence sets of the parameters estimates were computed using the values of $\alpha \in (0.50, 0.10, 0.05, 0.01)$ for each method. The confidence sets based on Wald's and log likelihood ratio test statistics were computed for maximum likelihood estimates. The Jensen–Shannon divergence based confidence sets were computed for minimum Jensen–Shannon divergence estimates obtained from grid evaluation of Monte Carlo estimates or from BOLFI. In each set of experiments, the coverage probabilities were simply computed as proportion of the experiments in which the value of the used test statistic at the null was equal or less than the acceptance value set by α. However, in experiments using multivariable models, the coverage probabilities for Wald's confidence sets were computed as the proportion of the experiments for which the true value of the parameter vector was included to the confidence ellipsoid.

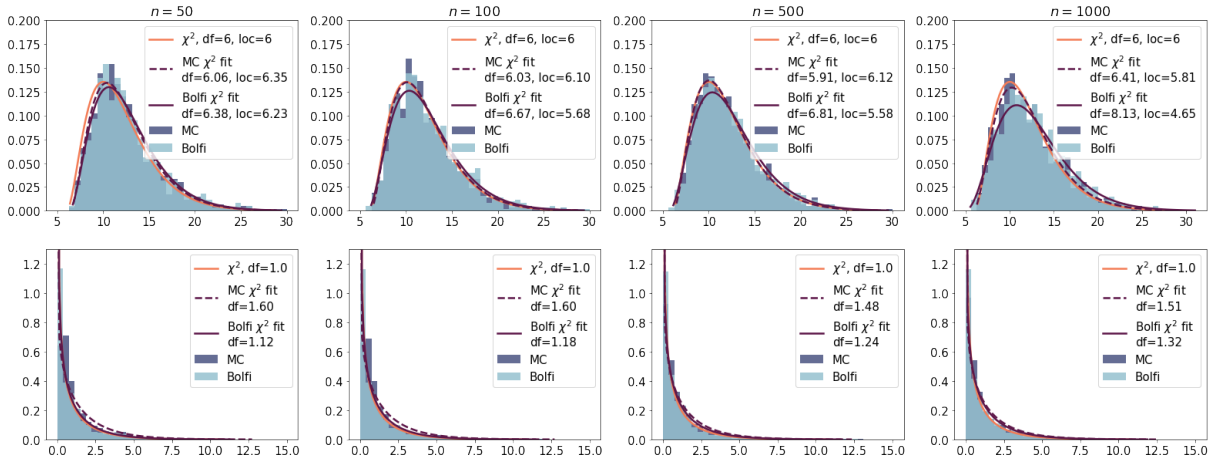


Figure 3.3: Histograms presenting the distribution of the observed mean JSD values at the null, i.e. at $\theta = 0.05$. Upper row presents the approximative distribution with the hypothesis of χ^2_{k-1} fit with location parameter $k-1$, where $k-1 = 6$, and separately fitted χ^2 distributions for results obtained from Monte Carlo experiments and from BOLFI. The lower panel presents the distributions of the normalized values from the experiments with the hypothetical approximative χ^2_d distribution, where $d = 1$, and fits computed separaterly for MC and BOLFI results.

We first examine the results obtained from the toy model that was presented in Chapter 2. Next, we examine the results obtained from the log linear model in which the number of estimated parameters was set to two and three. Finally, the results of the NFDS model with three different settings are examined, and we apply the computation of the normalized Jensen–Shannon divergence confidence intervals for the parameter estimates inferred with BOLFI using the Massachusetts data.

3.2.1 Toy model

The toy model was used in Chapter 2 to demonstrate inference process with maximum likelihood and likelihood-free inference, and to demonstrate the computation of the confidence intervals. The results constitute of the coverage probabilities computed for the ML estimates using the log likelihood-ratio and Wald's confidence interval, and for the estimates that minimize the Jensen–Shannon divergence using confidence sets were computed using the proposed method described in Section 2.3. The distributions of the estimates for each number of observation, $n \in (50, 100, 500, 1000)$, are summarised in Figure 2.3.

The χ^2 distribution with 1 degree of freedom was used in the computation of Wald's

confidence intervals, log likelihood-ratio confidence intervals, and in the confidence intervals based on normalized values of Jensen–Shannon divergence. For the mean values of Jensen–Shannon divergence the χ^2 distribution with 6 degrees of freedom and location parameter of 6 was used. Figure 3.3 presents the observed distribution of the mean values of the Jensen–Shannon divergence at null obtained from Monte Carlo estimates and from BOLFI, and the proposed approximative χ^2 distribution. Also, the hypothesised distribution for the normalized values of the mean Jensen-Shannon divergence is shown. Separately fitted χ^2 distributions for the Jensen–Shannon divergence values obtained from Monte Carlo estimates and from BOLFI are shown as well. It seems that the fitted degrees of freedom and location parameter for the mean Jensen–Shannon test statistic distribution are near the proposed values of $k - 1$ especially with the Monte Carlo distribution. However, the fitted degrees of freedom for the χ^2 distribution in the normalized values are greater for the Monte Carlo estimates than for the BOLFI. Overall, the fitted distributions seem to be similar with the proposed ones.

Table 3.2 presents all the results obtained from maximum likelihood inference, grid evaluation of the Monte Carlo estimates and from BOLFI. It can be seen that the coverage probabilities computed from Wald's and log likelihood ratio confidence interval follow the expected values the best. It can be seen that especially coverage probabilities of the 50% confidence intervals computed from Jensen–Shannon divergence test statistics are lower than the expected value. The rest of the coverage probabilities seem to follow the expected values, even though there exists more deviation from the expected values compared to the coverage probabilities of Wald's and log likelihood ratio.

3.2.2 Log linear model

The log linear model is the other model with tractable likelihood that was used in the repeated experiments. It contains more parameters to be estimated compared to the toy model. We wanted to study if the increase in the number of estimated parameters affect the resulting confidence sets and the coverage probabilities. The model was applied in two ways: first, the model contained only two parameters, λ^X and λ^Y, to be estimated, and then the saturated version of the model was used with three parameters, λ^X, λ^Y and λ^{XY}. We did thousand repeated experiments for different number of observations, $n \in (50, 100, 500, 1000)$, and we used maximum likelihood inference and grid evaluation of the Monte Carlo expectations infer the parameter estimates. The parameter range was set from -1 to 1 for each parameter in the grids that were used to compute the mean Jensen–Shannon divergence in the Monte Carlo estimate experiments. The step size was set to 201 for each parameter in the two parameter log linear model, and in the saturated log linear model, the step size was set to 101.

Figure 3.4 and Figure 3.5 present the distribution of the observed parameter estimates.

Table 3.2: The resulting coverage probabilities with their expected values are gathered from all the confidence interval experiments done for the toymodel including the ones conducted with ML inference, evaluation of Monte Carlo estimates over the grid (MC) and with BOLFI. Wald, LR and D_{JS} stand for Wald's confidence intervals, confidence intervals based on log likelihood ratio, and confidence intervals based on Jensen–Shannon divergence respectively. Confidence intervals based on Jensen–Shannon divergence are computed using χ^2_{k-1} with location parameter $k-1$ and using χ^2_d for normalized values. The parameter values used in the approximative distribution were $k-1=6$ and $d=1$.

		ML		MC		BOLFI	
n	Expected	Wald	LR	D_{JS} χ^2_{k-1}	D_{JS} χ^2_d	D_{JS} χ^2_{k-1}	D_{JS} χ^2_d
50	0.50	0.49	0.49	0.46	0.30	0.43	0.47
	0.90	0.91	0.91	0.88	0.89	0.86	0.88
	0.95	0.95	0.95	0.93	0.94	0.92	0.93
	0.99	0.99	0.99	0.98	0.99	0.98	0.99
100	0.50	0.50	0.50	0.49	0.32	0.47	0.47
	0.90	0.90	0.90	0.89	0.89	0.87	0.88
	0.95	0.95	0.95	0.95	0.94	0.94	0.94
	0.99	0.99	0.99	0.98	0.99	0.99	0.99
500	0.50	0.50	0.50	0.50	0.41	0.46	0.47
	0.90	0.91	0.90	0.90	0.90	0.88	0.90
	0.95	0.95	0.95	0.95	0.95	0.93	0.95
	0.99	0.99	0.99	0.99	0.99	0.99	0.98
1000	0.50	0.49	0.49	0.47	0.41	0.42	0.47
	0.90	0.90	0.90	0.88	0.89	0.85	0.90
	0.95	0.95	0.95	0.94	0.96	0.93	0.95
	0.99	0.99	0.99	0.99	0.99	0.99	0.99

It can be seen in both of the figures that the distribution of the maximum likelihood estimates and the distribution of the minimum Jensen-Shannon divergence estimates look similar. Also, the width of the distribution gets narrower as the number of observations increase as expected.

In log linear model, we also used several values of $\alpha \in (0.5, 0.1, 0.05, 0.01)$ to define the $100\%(1-\alpha)$ confidence sets for the parameter estimates using Wald's and log likelihood methods for maximum likelihood estimates, and mean Jensen–Shannon divergence, and its normalized version for minimum Jensen–Shannon divergence estimates. We also computed the coverage probabilities for each method from the set of experiments. Confidence sets based on Wald's, log likelihood and normalized Jensen–Shannon divergence test statistics used the same asymptotic χ^2_d distribution, where d was 2 in two parameter model and 3 in saturated model. Confidence sets based on mean Jensen–Shannon divergence were computed using the hypothesised asymptotic χ^2_{k-1} distribution with the location parameter $k-1$; in both, two and three parameter case, the hypothesised distribution

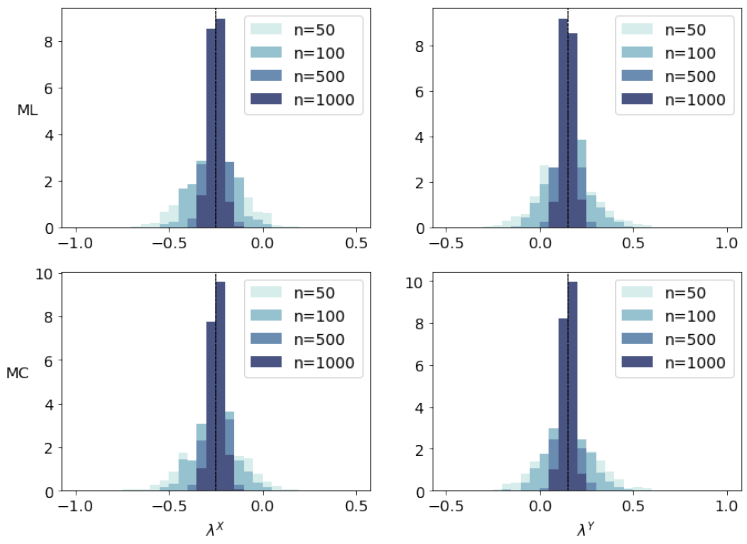


Figure 3.4: Histograms presenting the distribution of the estimates of the log linear model with two parameters obtained from maximum likelihood (ML) inference and from grid evaluation of the Monte Carlo estimates (MC). The first column presents the marginal distribution of the estimates $\hat{\lambda}^X$ and the second column presents the marginal distribution of the estimates for $\hat{\lambda}^Y$. The dashed line presents the true value of the parameters.

became χ^2_3 with location parameter 3 as the 2-way table contains 4 cells.

The Figures 3.6a and 3.6b show the observed distribution of the mean Jensen–Shannon divergences and its normalized values at the null. The figures also show the hypothesised χ^2 distributions, and separately fitted χ^2 distributions. The shape of the hypothesised distributions and fitted distributions seem to coincide, except in the distributions describing the normalized Jensen–Shannon divergence values at the null with the two parameter model. Here, the fitted degrees of freedom are greater than the hypothesised value. The resulting coverage probabilities computed for both log linear models are presented in the Table 3.3. It seems that the coverage probabilities coincide between the computed confidence sets in each model, and those seem to follow the expected values. Even though there exists deviation from the expected coverage probabilities, the trend is similar between the each method in both models. The largest deviance from the expected coverage probability value can be seen with the coverage probabilities computed from normalized version of the Jensen–Shannon divergence based confidence sets, when the number of observations is 50. This can be explained by the difference observed between the shapes of the observed and hypothesised distributions at the null: there exists less observations of small normalized Jensen–Shannon divergence values than the hypothesised distribution expects.

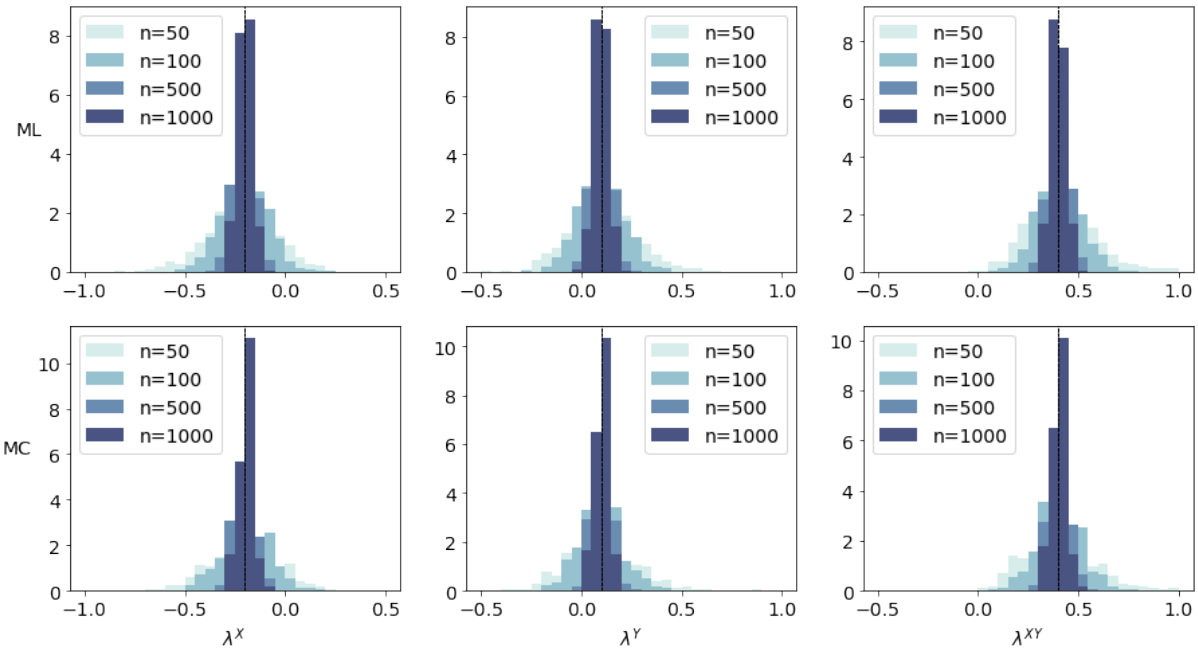


Figure 3.5: Marginal distributions of the parameter estimates obtained from maximum likelihood (ML) and from minimizing Jensen–Shannon divergence using grid evaluation of Monte Carlo estimates (MC). Each column presents the marginal distribution of estimates computed for each parameter, and vertical dashed line presents the true value of the parameters.

3.2.3 NFDS model

In this section, we go through the results obtained from the experiments, where we used the simulated data from NFDS model. We conducted a set of 200 repeated experiments with each three settings: first setting containing 250 collected isolates at three time points, second setting containing 250 collected isolates at six time points, and third setting containing 1000 isolates collected at three time points. We computed coverage probabilities based on these 200 repeated experiments for each setting. We used BOLFI for the inference process, since the grid evaluation of Monte Carlo estimates would have been computationally too expensive with this model.

The parameter range were set for v, σ_f, and for m to be $[10^{-6}, 0.5]$, $[10^{-3}, 0.22]$, and $[10^{-6}, 0.2]$ respectively and the parameters were assumed to follow a uniform prior within the defined ranges in all of the experiments. Due to the small values of the parameters, the experiments were carried in logarithmic space. We used squared exponential kernel

Table 3.3: Table presenting the expected and observed coverage probabilities of the two log linear models. Wald, LR and JSD stand for Wald's confidence set, confidence set based on log likelihood ratio test, and confidence set based on Jensen–Shannon divergence respectively.

		Log linear model with 2 parameters				Saturated log linear model			
n	Expected	Wald	LR	$D_{JS}\ \chi^2_{k-1}$	$D_{JS}\ \chi^2_d$	Wald	LR	$D_{JS}\ \chi^2_{k-1}$	$D_{JS}\ \chi^2_d$
50	0.50	0.48	0.48	0.46	0.42	0.46	0.49	0.46	0.47
	0.90	0.90	0.90	0.92	0.89	0.86	0.91	0.90	0.90
	0.95	0.95	0.95	0.96	0.95	0.93	0.95	0.94	0.95
	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
100	0.50	0.47	0.47	0.47	0.43	0.52	0.51	0.50	0.48
	0.90	0.90	0.90	0.91	0.89	0.87	0.91	0.91	0.91
	0.95	0.95	0.95	0.96	0.95	0.92	0.95	0.95	0.95
	0.99	0.99	0.99	0.99	0.99	0.97	0.99	0.98	0.98
500	0.50	0.47	0.48	0.48	0.44	0.49	0.48	0.49	0.49
	0.90	0.88	0.88	0.88	0.87	0.89	0.90	0.90	0.90
	0.95	0.94	0.94	0.94	0.94	0.94	0.95	0.95	0.95
	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
1000	0.50	0.48	0.48	0.47	0.47	0.48	0.49	0.48	0.49
	0.90	0.91	0.91	0.90	0.91	0.89	0.90	0.90	0.90
	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95
	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

(Equation 2.19) with automatic relevance determination, which means that each parameter dimension is associated with separate lengthscale parameter. For each lengthscale parameter we set a Gamma prior where both rate and shape parameters were set to 2. The number of evidence, i.e. the number of evaluated points, was set to 2000, and the hyper parameters of the kernel and lengthscale were updated on every tenth iteration. The Jensen–Shannon divergence was predicted over a chosen grid using the resulting predictive posterior mean of the Gaussian process. We used step size of 51 for each parameter range used already in the inference in this grid.

The hypothesised asymptotic distribution used in the computation of the mean Jensen–Shannon divergence based confidence sets used information of the possible event classes. However, with the NFDS model we could not use this information because we did not know the actual number of the possible event classes for the isolates to be observed in. Even though we had the theoretical number of event classes that was based on the number of sequence clusters, used time points in the experiments, and on the possibility for the observed isolate being either vaccine type or not, we observed that this theoretical number of the event classes overestimated the true number of possible event classes. That is why in these experiments, we computed only the normalized version of the mean Jensen–Shannon divergence confidence intervals that required only the information about the

number of estimated parameters, and the χ^2_3 distribution was used in the computation of the normalized Jensen–Shannon divergence confidence intervals.

The histogram of the parameter estimates from each set of experiments are shown in Figure 3.7. It can be seen that the obtained estimates are distributed around the true values of the parameters in each set of experiment, and the modes of the histograms are close to the true values of the parameters. The histogram which mode is furthest from the true value is observed in Figure 3.7b which describes the distribution of the immigration rate estimates obtained from set of experiments where the number of time points was six and number of observed isolates at each time point was 250. It can also be seen that the number of observations seem to have an effect on the observed distribution of the estimates. This can be seen by comparing the two resulting histograms of the experiments where the number of time points was same but the number of observed isolates at each time point varied (Figure 3.7a and Figure 3.7c). It seems that the observed variance of the estimates has decreased as the number of observed isolates at each time point have increased, which confirms the expected behaviour of the estimates as the number of observations is varied.

The Figure 3.8 shows the distribution of the mean Jensen–Shannon divergence values at the null for each set of experiments. It seems that the hypothesised distribution fits to the observed one. This can be seen especially with the distribution obtained from experiments that used the largest number of observations, which has also larger proportion of small normalised mean Jensen–Shannon divergence values observed at null compared to the experiment that had less observations (Figures 3.8a and 3.8c). The Table 3.4 contains the coverage probabilities computed for each set of experiment. It can be seen from this table that the coverage probabilities seem to coincide with the expected values. However, the resulting coverage probabilities, observed from the experiments where 250 isolates were collected at six time points, seem to have smaller values compared to expected values and to the models containing smaller number of time points. This could be explained with the observed distribution of the normalised mean Jensen–Shannon divergence values at null (Figure 3.8b) that contains higher proportion of large normalised mean Jensen–Shannon divergence values what the hypothesised distribution would expect.

Computation of the confidence sets

Finally, the Jensen–Shannon divergence based confidence sets were applied in the parameter inference with data collected from Massachusetts. Same settings were used in BOLFI as presented earlier with the repeated experiments, and this experiment was also carried in logarithmic space. Figure 3.9 shows the predicted Jensen–Shannon divergence, estimate and the acceptance value of 95% confidence set. The plots have been fixed to the grid values that minimize the predicted Jensen–Shannon divergence. Table 3.5 presents

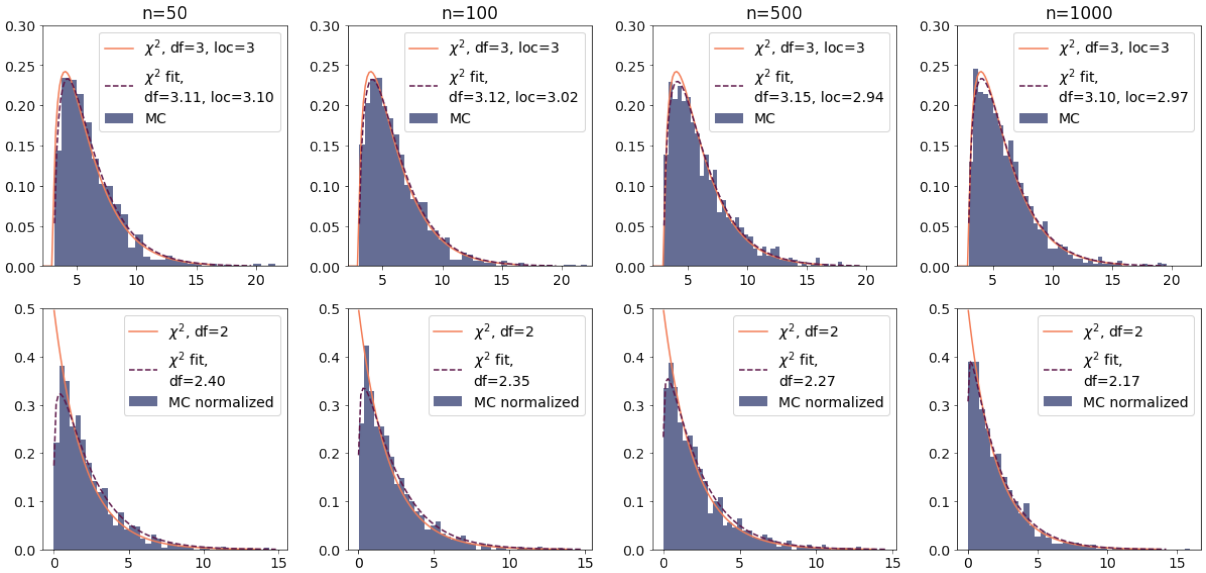
Table 3.4: Coverage probabilities computed for normalized Jensen–Shannon divergence values from 200 repeated NFDS experiments with three different settings. First column contains the experiments where the number of observed isolates was 250 at each three time point; second column contains the coverage probabilities of the experiments where 250 isolates where observed at six different time points; third column contains coverage probabilities from experiments where thousand isolates where observed at three different time points.

Expected	n=250 and t=3	n=250 and t=6	n=1000 and t=3
0.50	0.44	0.45	0.57
0.90	0.83	0.80	0.83
0.95	0.92	0.84	0.88
0.99	0.97	0.92	0.96

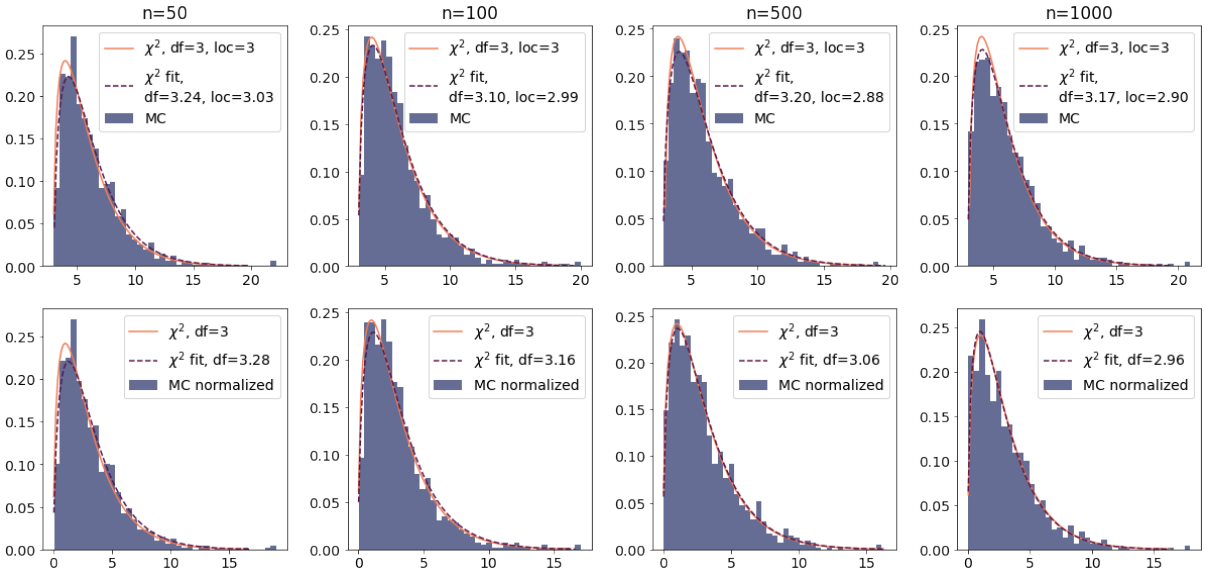
the resulting confidence intervals for the inferred parameters m, σ_f and v after taking exponential of them in order to compare the results with the previously inferred parameter values. It can be seen that the confidence intervals of the parameters tend to become wider as the percentage grows. What can be seen also from the minimum Jensen–Shannon divergence estimates is that they coincide well with the previously obtained estimates for the model [4]. It also seems that the confidence intervals computed here are narrower compared to the credible intervals of the estimates shown in previous work.

Table 3.5: Jensen–Shannon based confidence intervals (CI) for the inferred estimates from the Massachusetts data.

Parameter	m	σ_f	v
Estimate	0.0068	0.0069	0.07533
50% CI	(0.0051, 0.0084)	(0.0053, 0.0091)	(0.0640, 0.0885)
90% CI	(0.0042, 0.0102)	(0.0042, 0.0108)	(0.0640, 0.0987)
95% CI	(0.0042, 0.0102)	(0.0040, 0.0114)	(0.0574, 0.0987)
99% CI	(0.0038, 0.0113)	(0.0036, 0.0127)	(0.0574, 0.0987)

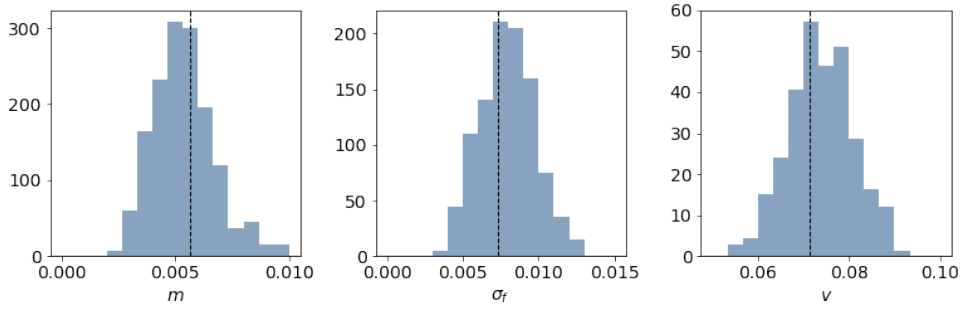


(a) Log linear model with two parameters

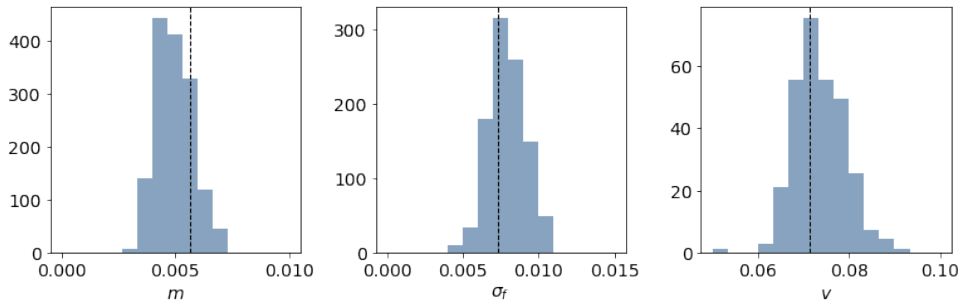


(b) Saturated log linear model

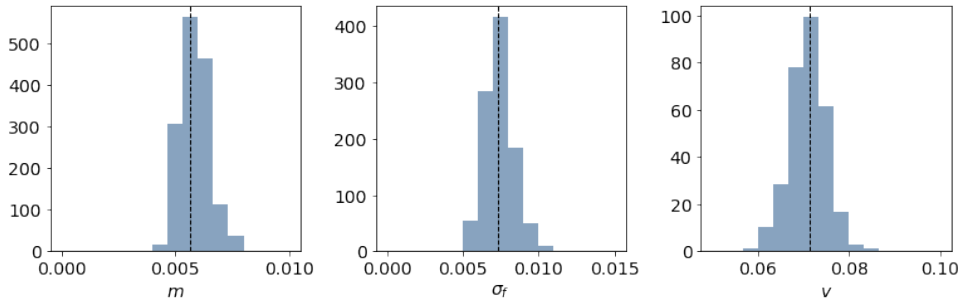
Figure 3.6: Observed distributions at the null from two parameter and saturated log linear models from each set of Monte Carlo experiments. Upper row contains the distribution of the mean values of Jensen–Shannon divergence, and lower row contains normalized values. The proposed approximative χ^2 distribution is plotted as orange line, and for comparison, a separately fitted χ^2 distribution is plotted with dashed line.



(a) Estimates observed from the NFDS model using three time points with 250 observed isolates.

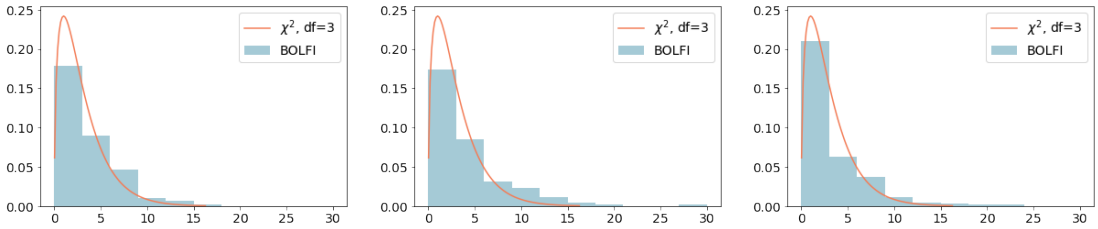


(b) Estimates observed from the NFDS model using six time points with number of 250 observed isolates at each time point.



(c) Estimates observed from the NFDS model using three time points with number of 1000 observed isolates at each time point.

Figure 3.7: Observed distribution of estimates from 200 repeated experiments for each set of experiments. The vertical dashed line is the true value of the parameter.



(a) (b) (c)

Figure 3.8: Observed distribution of the normalized mean divergence values from 200 repeated experiments for each set of experiments. Orange line represents the proposed approximative χ^2 distribution. (a) Distribution observed from the NFDS model using three time points with 250 observed isolates; (b) Distribution observed from the NFDS model using six time points with number of 250 observed isolates at each time point; (c) Distribution observed from the NFDS model using three time points and 1000 observed isolates at each time point.

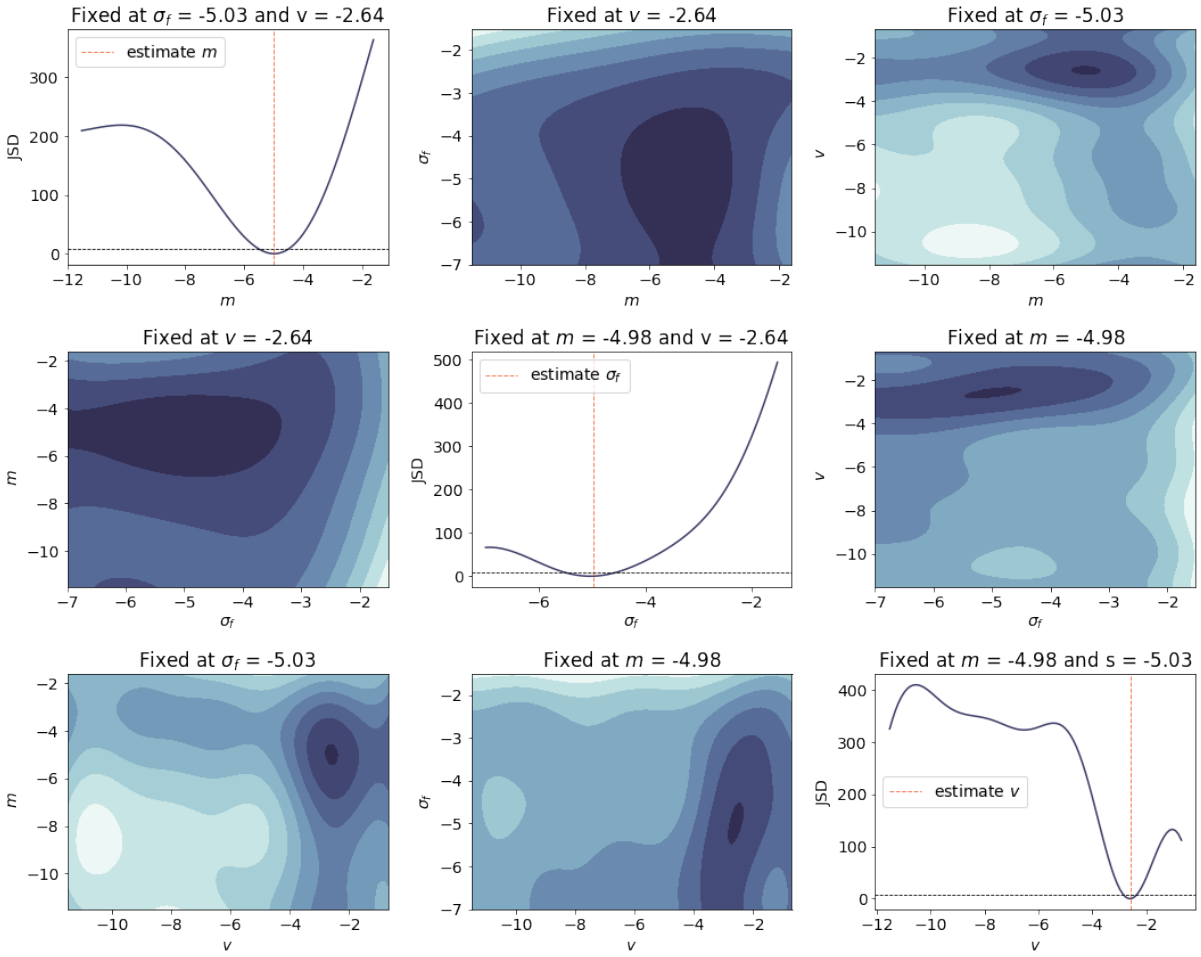


Figure 3.9: Predicted values of the Jensen–Shannon divergence (JSD) from the inference process for the Massachusetts data that have been normalized. Diagonal plots contain the estimate of the parameter, plotted as dashed vertical line, and the acceptance threshold for the 95% confidence set plotted with dashed horizontal line. The acceptance threshold has been computed using χ^2_3 distribution. Deeper blue colour refers to smaller value of divergence.

Chapter 4

Discussion and Conclusions

This thesis focused on studying the whether the mean Jensen–Shannon divergence and its normalised version can be used as a test statistic to compute the confidence sets for estimates that are minimizing values of the mean Jensen–Shannon divergence. We also wanted to see whether the hypothesized distributions approximate the distribution of these test statistics, and if the confidence sets constructed from these test statistics have the frequentist property that is expected from them. The proposed method is suitable especially for likelihood-free inference in which Jensen–Shannon divergence similarity measure is used with observations that can be summarised with discrete probability distributions. The most beneficial feature of this method is that the approximation of the likelihood is not required in the computation of the confidence sets.

4.1 Discussion

The method presented in this thesis to compute the confidence sets in novel way is suitable especially for distributions of event probabilities that do not contain zero probabilities. The approximation of χ^2 distribution for the χ^2 test statistic is known to become asymptotically inconsistent when the number of event classes become too large, or when some of the event probabilities are too small [14]. This inconsistency of the asymptotic distribution can also be seen with the distributions of χ^2 and Jensen–Shannon divergences.

The hypothesized approximative distribution of the mean Jensen–Shannon divergence was used in the computation of the confidence sets. However, the hypothesized approximative distribution describes the distribution of the possible values of the mean Jensen–Shannon divergence computed from m simulated samples against various observations generated at the null, instead of modelling the distribution of sample means of Jensen–Shannon divergences based on single observation. Thus, the right tail of the approximative

distribution consists of greater values of the mean Jensen–Shannon divergence, that can be a result of observing data by chance that is not well explained by the parameter value at the null that generated it. This can lead to a special case of observing empty confidence set for the estimate, which can happen when the value of the mean Jensen–Shannon divergence at the estimate is greater than the critical value set by α in the approximative distribution. The observation of empty confidence sets lead us to study the behaviour of the normalised version of the mean Jensen–Shannon divergence.

This lead us to study the behaviour of the normalized version of the mean Jensen–Shannon divergence, and to try the hypothesized distribution for this test statistic. We observed that the χ^2_d distribution seemed to fit well with the observed values of these normalized test statistics observed at the true value of the parameters (Figure 3.3, and 3.6)

The frequentist behaviour of the confidence sets based on Jensen–Shannon divergence were studied with models containing varying number of parameters. The hypothesized distributions were visually compared with fitted versions of the distributions with the histograms of the observed mean Jensen–Shannon divergence values that were collected from thousand repeated experiments. The hypothetical distributions seemed to fit well with the observed distribution of the divergence obtained from Monte Carlo estimates and from BOLFI (Figure 3.3, and Figure 3.6).

Also, the frequentist behaviour was studied by comparing the observed the coverage probabilities to log likelihood-ratio and Wald's confidence sets. The results confirmed that both of the hypothetical approximative distributions had good frequentist properties as they followed the expected values of the coverage probabilities (Table 3.2, and Table 3.3). The largest deviation of 20% from the expected coverage probability among the models with tractable likelihood was observed from the Monte Carlo experiments of the toy model with the normalized version of the Jensen–Shannon divergence based confidence interval (Table 3.2). This deviation could be explained by the observed distribution of the mean Jensen–Shannon divergences that are located outside of the approximative distribution seen in Figure Figure 3.3. The reason for this behaviour could be related to the random variation in the Monte Carlo estimates of the Jensen–Shannon divergences.

The normalized test statistic introduced in this thesis was also tested with the NFDS-model that has intractable likelihood. The observed distributions of each setting seemed to fit the hypothetical distribution (Figure 3.8, and the coverage probabilities of each of model settings seemed to follow the expected values with largest deviation being 11% (Table 3.4). This is interesting, as the number of observations varied between the model settings, and also the number of event classes. The deviations from the expected coverage probability can be explained with the number of repeated experiments: due to the computational complexity of the model, the number of repeated experiments carried with BOLFI were limited to 200. Thus the results of these repeated experiments can be influenced by

random variation. Finally, the normalised mean Jensen–Shannon divergence confidence set was applied for real data collected from Massachusetts, and the resulting parameter estimates and their confidence intervals were compared with the previously obtained estimates and credible intervals of the same parameters [4]. The obtained estimates and computed confidence intervals seem to coincide with the previously obtained estimates and credible intervals, even though the confidence intervals are narrower compared to the credible sets (Table 3.5, and Table 1 in [4]).

4.2 Conclusions and future work

We can conclude that both of these test statistics, the mean of Jensen–Shannon divergence and the normalized version, can be used to compute the confidence sets for parameter estimates. The confidence sets based on the hypothetical distributions produced coverage probabilities that where comparable with the log likelihood-ratio and Wald's confidence sets, and these confidence sets seemed to follow the expected value of the coverage probability. However, the risk of observing empty confidence sets with the mean of Jensen–Shannon divergence should be acknowledged.

The numerical results show that the proposed method to compute Jensen–Shannon divergence based confidence sets meet the expected frequentist behaviour related to confidence sets, and that it can be used as new method to compute the confidence sets. These confidence sets of the parameter estimates could bring new information to the inference, especially to the likelihood-free inference where the posterior distribution is approximated by bypassing the evaluation of the likelihood. These confidence sets describe what type of influence parameters have on the Jensen–Shannon divergence surface, and this information could be compared with the credible sets computed from the approximated posterior distribution. The method could also be easily applicable to any other likelihood-free inference method that relies on the similarity measure of the observed and simulated data.

This thesis leaves lot of room for future applications and further study. Here, we focused on testing the proposed method with models with maximum of three estimated parameters. It would be interesting to see whether the hypothetical distributions still keep their frequentist properties with larger number of parameters. Also, larger experiments with models with intractable likelihood would be interesting to cover, as in this thesis the number of repeated experiments was limited to 200 in stead of 1000. One major topic for further study includes the study of the theoretical background of the hypothetical distribution of the normalized Jensen–Shannon divergence that was not covered in this thesis. In this thesis we did not study the power of the used test statistics but this could be a potential study topic for the future.

Bibliography

[1] Alan Agresti. *Foundations of linear and generalized linear models.* Wiley series in probability and statistics. John Wiley & Sons Inc, Hoboken, New Jersey, 2015.

[2] N. Balakrishnan, N. Martín, and L. Pardo. Empirical phi-divergence test statistics for testing simple and composite null hypotheses. *Statistics (Berlin, DDR)*, 49(5):951–977, Sep 3, 2015.

[3] Phelim P. Boyle. Options: A Monte Carlo approach. *Journal of Financial Economics*, 4(3):323–338, 1977. ID: 271671.

[4] Jukka Corander, Christophe Fraser, Michael U. Gutmann, Brian Arnold, William P. Hanage, Stephen D. Bentley, Marc Lipsitch, and Nicholas J. Croucher. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nature Ecology & Evolution.*, 2017.

[5] Jukka Corander, Ida Holopainen, Ulpu Remes, and Timo Koski. Asymptotic likelihood-free inference by jensen-shannon divergence and generative adversarial sampling. *Under preparation*, 2021.

[6] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2020.

[7] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. Jun 6, 2015.

[8] Nicholas J. Croucher, Jonathan A. Finkelstein, Stephen I. Pelton, Patrick K. Mitchell, Grace M. Lee, Julian Parkhill, Stephen D. Bentley, William P. Hanage, and Marc Lipsitch. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics*, 45(6):656–663, May 5, 2013.

[9] Niccolò Dalmasso, Rafael Izbicki, and Ann B. Lee. Confidence sets and hypothesis testing in a likelihood-free inference setting., 2020.

[10] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of statistics*, 7(1):1–26, Jan 1, 1979.

[11] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian data analysis*. 3rd edition, 2021. Available at http://www.stat.columbia.edu/ gelman/book/.

[12] Jelle J. Goeman. *Randomness and the Games of Science*, pages 91–109. The Challenge of Chance: A Multidisciplinary Approach from Science and the Humanities. Springer International Publishing, Cham, 2016. ID: Goeman2016.

[13] Michael U. Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. Jan 14, 2015.

[14] Shelby J. Haberman. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*, 83(402):555–560, Jun 1, 1988.

[15] Leah Jager and Jon A. Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of statistics*, 35(5):2018–2053, Oct, 1 2007.

[16] Dirk P. Kroese and Joshua C.C. Chan. *Statistical Modeling and Computation*. Springer New York, New York, NY, 2013.

[17] Jarno Lintusaari, Henri Vuollekoski, Antti Kangasrääsiö, Kusti Skytén, Marko Järvenpää, Pekka Marttinen, Michael U. Gutmann, Aki Vehtari, Jukka Corander, and Samuel Kaski. Elfi: Engine for likelihood-free inference. *Journal of Machine Learning Research*, 19(16):1–7, 2018.

[18] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate bayesian computational methods. *Statistics and computing*, 22(6):1167–1180, Oct 21, 2011.

[19] Nirian Martin, Raquel Mata, and Leandro Pardo. Phi-divergence statistics for the likelihood ratio order: An approach based on log-linear models. *Journal of multivariate analysis*, 130:387–408, Sep 2014.

[20] Pekka Nieminen and Pentti Saikkonen. *Tilastollisen päättelyn kurssi*. Helsingin yliopisto, Matematiikan ja tilastotieteen laitos, 2013.

[21] Leandro Pardo. *Statistical inference based on divergence measures*. Taylor and Francis, 2005.

[22] Leandro Pardo and Nirian Martín. Minimum phi-divergence estimators and phi-divergence test statistics in contingency tables with symmetry structure: An overview. *Symmetry (Basel)*, 2(2):1108–1120, Jun 11, 2010.

[23] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2nd edition, 2006.

[24] Chad M. Schafer and Philip B. Stark. Constructing confidence regions of optimal expected size. *Journal of the American Statistical Association*, 104(487):1080–1089, Sep 2009.

[25] Tore Schweder and Nils L. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2016.

[26] Mikael Sunnåker, Alberto G. Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate Bayesian computation. *PLOS Computational Biology*, 9(1):e1002803, 2013.

[27] Suzanne Thornton, Wentao Li, and Min-ge Xie. An effective likelihood-free approximate computing method with statistical inferential guarantees. May 29, 2017.