Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Olav Helland-Moe

Objective inference for correlation

Master's thesis in Industrial mathematics
Supervisor: Gunnar Taraldsen
June 2021

NTNU
Norwegian University of
Science and Technology

Olav Helland-Moe

Objective inference for correlation

Master's thesis in Industrial mathematics
Supervisor: Gunnar Taraldsen
June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Sammendrag

Denne masteroppgaven tar for seg problemstillingen om å hente mest mulig informasjon om korrelasjonen i en binormal fordeling basert på observerte punkter i planet. Forventningsverdiene og variansene er antatt kjente. Til tross for denne tilsynelatende forenklingen, er den kjent for å gi komplikasjoner.

Oppgaven er en videreføring av prosjektoppgaven som gikk ut på å finne de beste metodene for å estimere korrelasjonen. Masteroppgaven utvider dette temaet ved å i tillegg se på metoder for å beregne usikkerheten. Usikkerheten blir først og fremst uttrykt ved hjelp av såkalte konfidensfordelinger. Bayesianske, og frequentistiske metoder blir brukt til både estimeringen av korrelasjonen og usikkerheten. Fiduse metoder blir også brukt til å utrykke usikkerhet. For å sammenligne metoder og estimatorer vil tap- og riskfunksjoner bli brukt, deriblant kvadrattap, Fisher information metric og Kullback-Leibler divergens. De Bayesianske metodene er basert på objektive priorfordelinger som uniform, Jeffreys, penalized complexity (PC) og arcsine prior. For usikkerhetsberegning er en ekstra prior tatt i bruk, her navngitt som arctanh prior. De fiduse metodene er basert på generalisert fidus inferens.

Analysen for å estimere korrelasjonen viser at Bayesianske estimatorer kan gi bedre resultater enn frekventististe estimatorer som MLE og empirisk korrelasjon ved små datasett. Ut av dem, står posterior gjennomsnitt med uniform prior som en god kandidat.

For usikkerhetsmål vil ikke posteriorfordelingene har tilstrekkelig frekventistisk dekning og kan ikke brukes for små datasett. Den eksakte konfidensfordelingen gitt i Teorem 3.2 er den mest konsistente fordelingen og er derfor anbefalt.

Flere av nøkkelresultatene i oppgaven er som følger:

1. Posterior median minimerer forventet Fisher information metric, og Posterior gjennomsnitt minimerer forventet Kullback-Leibler divergens

2. En rekke konfidensfordelinger kan bli beregnet ved hjelp av pivoter gitt i ligning (32) og mer generelt ligning (38)

3. Fremgangsmåter for å lage konfidensfordelinger ved hjelp av pivotaler og data genererende funksjoner i tilfeller hvor en direkte invertering ikke er mulig.

Abstract

This master thesis considers inference of the correlation in a binormal distribution based on observed points in the plane. The means and variances are assumed known. Despite this seemingly simplification, it is well known to give complications.

The thesis is an continuation of the project report which focused on finding the best methods for estimating the correlation. It expands this topic by introducing methods for quantifying the uncertainty. The uncertainty will first and foremost be expressed in terms of so called confidence distributions. Bayesian and frequentist methods will be used in both estimation of the correlation and uncertainty. Fiducial methods will be used for expressing the uncertainty. To compare methods and estimators, loss and risk functions will be used, including squared error, Fisher information metric and Kullback-Leibler divergence. The Bayesian methods are based on objective prior distributoins such as uniform, Jeffreys, penalized complexity (PC) and arcsine prior. For uncertainty, an additional prior is used, which will be referred to as the arctanh prior. The fiducial methods are based on generalized fiducial inference.

The analysis of the point estimators shows that the Bayesian estimators can outperform the frequentist estimators such as the MLE and empirical correlation, for small data sizes. Out of the estimators, the posterior mean using the uniform prior stands out as a good candidate.

For the uncertainty quantification, none of the posterior distributions will have sufficient frequentistic coverage. The exact confidence distribution here given in theorem 3.2, performs the most consistently and is therefore recommended. The thesis introduces a general and a specific methods for creating confidence distributions using pivotal quantities.

Multiple key results in this thesis is as follows

1. Posterior median minimizes expected Fisher Information Metric and posterior mean minimizes expected Kullback-Liebler divergence.

2. A collection of confidence distributions can be calculated using the pivots in equation (32) and more generally equation (38).

3. A Procedure for creating confidence distributions using pivots and data generating function in cases where a direct inversion is not possible.

Preface

This thesis was written for the course *TMA4900 - Industrial Mathematics, Master's Thesis* and marks the end of my studies at the Norwegian University of Science and Technology (NTNU). I would like to thank my supervisor Gunnar Taraldsen for all our meetings and the guidance he has provided. It has been a pleasure to receive help from such an enthusiastic supervisor. I would also like to thank my family for the support along the way. Finally, I thank my wife for being a light in these challenging times. My time in Trondheim would not have been the same without her.

June 2021
Trondheim
Olav Helland-Moe

Contents

List of Figures

1 Introduction

Parameter inference is an essential part of statistics and is the link between statistical models and physical processes. By using well defined models that are tailored to a physical process, it is possible to make further inference about it, including predictions. The validity of such inference is dependent on how well the model fits the process. While the choice of model is important, it is just as important to find methods that can give as much information about the model as possible.

The baseline for making statistical models is the probability space. It consists of the triplet $(\Omega, \mathcal{F}, P)$. Ω is a sample space with all the possible outcomes ω. $\mathcal{F}$ is a family of events in Ω. P is a probability measure on $(\Omega, \mathcal{F})$ (Karr 1993, p. 23-24). A random variable on the probability space X is a function that maps from the sample space Ω to the real line, where every set $(X \leq x) = \{\omega \in \Omega | X \leq x\}$ is an event in $\mathcal{F}$. By having these criteria in addition to Kolmogorovs axioms of probability, a statistical model for the random variable can be denoted $P(X \leq x) = F_X(x)$ (Karr 1993, p. 52). The function $F_X(x)$ is known as the cumulative distribution function (CDF) which states the behaviour of X. If X is said to be absolutely continuous, then the probability density function (PDF)

$$f_X(x) = \frac{d}{dx} F_X(x)$$

exists (Karr 1993, p. 52). If either the CDF or the PDF of a random variable is known, it is possible to make inference about X. This can be useful in terms of making predictions, confidence intervals and other means of describing a physical process around X.

Practically, it is usually not possible to fully know the distribution or model of a random variable. In order to analyse these processes statistically, the distribution of the process is assumed. Accurately finding a specific distribution for a process is not realistic. The compromise is to assume a family of distributions given by a set of parameters $\theta \in \Omega_\Theta$. Ω_Θ is known as the model parameter space and is the set of all possible parameters θ. We can define a family of distributions of θ with the CDF $F_X(x|\theta)$ and the PDF $f_X(x|\theta)$. A goal can then be to gain as much information about the parameter θ.

If the results of an experiment is given by a set of n independent data points on the form (x_i, y_i), it can be possible to model the data using the binormal distribution. Each data point is then assumed to be a realization of the vector (X, Y) where both X and Y are normally distributed. The parameter in the distribution consists of the mean of X and Y, the variance of X and Y and the correlation between X and Y. Out of all the parameters, only the correlation states the relation and dependency between X and Y. The relation is such that if $X = x$ is known with mean equal to 0 and variance equal to 1 and the correlation is ρ, then Y is normally distributed around the line $y = \rho x$ with variance $(1 - \rho^2)$ (Taraldsen 2020). This is visualized for both 100 data points in figure 1 and for 10 data points in figure 2. The black lines display the true line $y = \rho x$. The goal in many cases is to predict y when x is known. If ρ is known, then the best predictor for y, under some conditions, is along the line $y = \rho x$ (Taraldsen 2020). If ρ is unknown, methods for estimating the correlation is useful for estimating y. The red lines show the best predictions when the correlation

is estimated using the empirical correlation, see (40). The two figures shows two different scenarios in a model problem. One with a large data size and one with a small data size. As seen, estimation of both ρ and y can be less accurate with less information. The correlation can also give information about the binormal variables. If the correlation is 0, then X and Y are independent (Shao 2003, Example 1.17). Finding methods for reliably testing if there is no correlation can therefore be very useful.

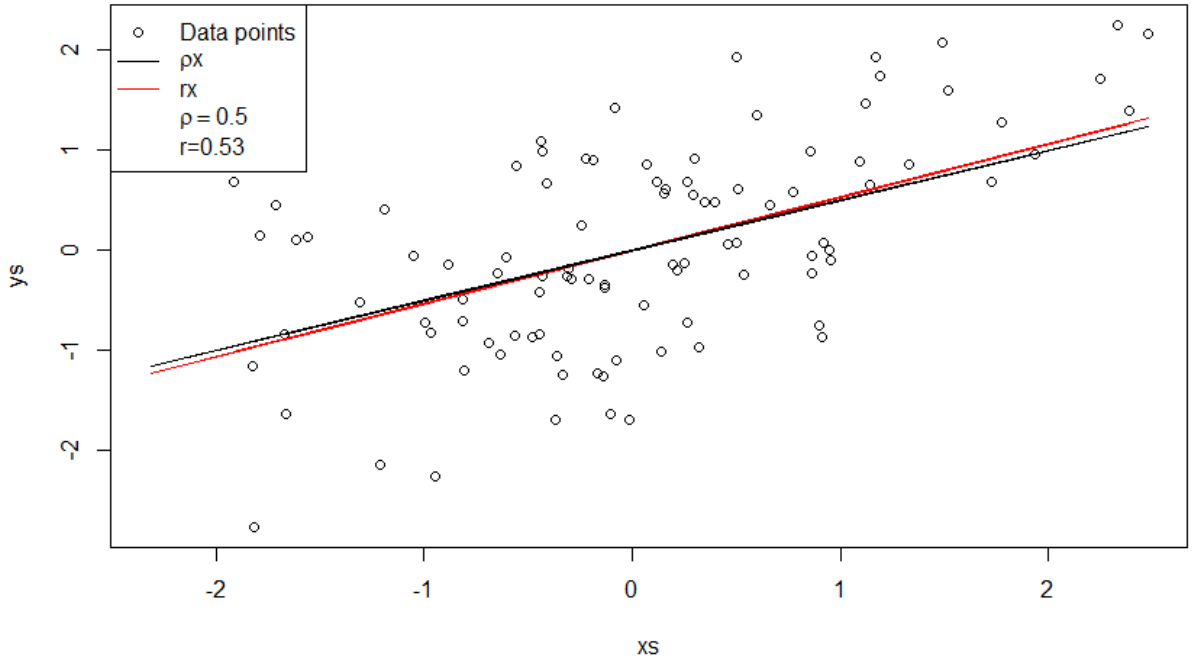


Figure 1: Figure of 100 independent binormal data points (x, y) with known means 0 and known variances 1. The black line is the line $y = \rho x$ and the red line is the line $y = rx$, where r is the empirical correlation, see (40).

Work on the correlation of a binormal distribution is not new. The specific problem where the means and variances are known does also occur in different articles. As late as 2012, Fosdick and Raftery tested multiple point estimators using both frequentist and Bayesian estimators. A similar report focusing on uncertainty for the correlation of the binormal was written by Fosdick and Perlman 2016. This thesis will add to these findings with various approaches to both point estimation and uncertainty quantification.

A small comment on notation is that random variables or vectors will be denoted using capital letters such as X. Constant variables or vectors will be written in lowercase such as x.

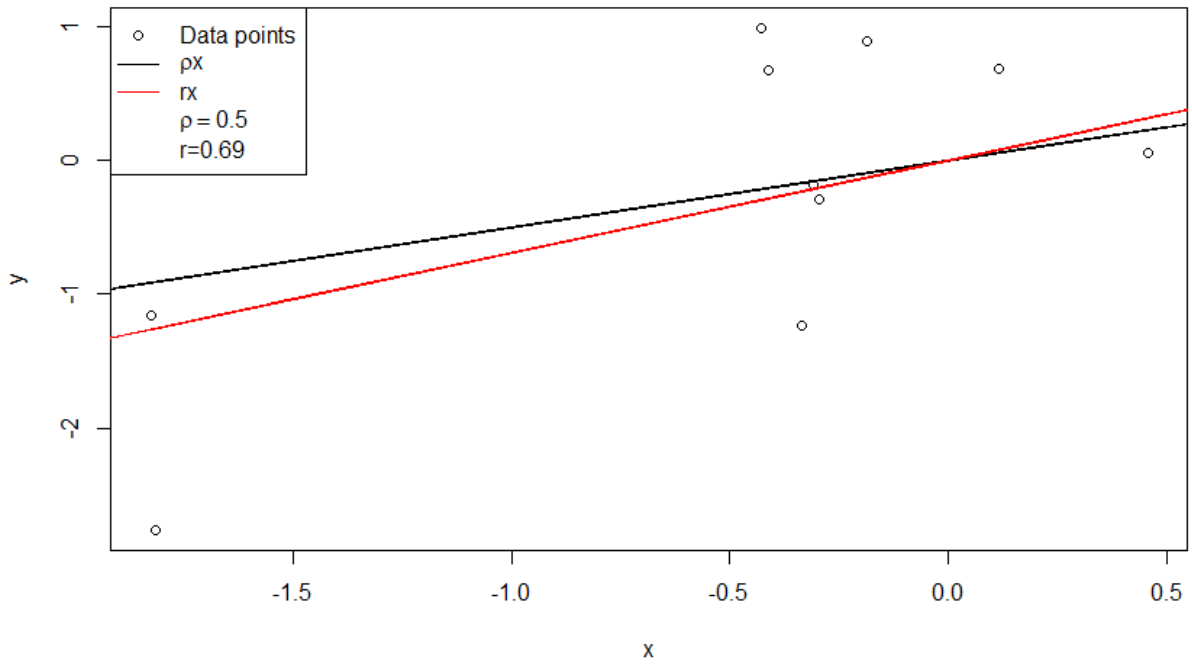


Figure 2: Figure of 10 independent binormal data points (x, y) with known means 0 and known variances 1. The black line is the line $y = \rho x$ and the red line is the line $y = rx$, where r is the empirical correlation, see (40).

2 General theory

2.1 Statistical model

Studying data sets using statistical models is a cornerstone of the field of statistics. They give a framework that allows for both inference and validity testing of the inference.

A statistical model can be defined using a cumulative distribution function (CDF) also known as the distribution function. Given the probability measure P and the parameter θ, the CDF F is defined as $F(x|\theta) = P(X \leq x|\theta)$. If X is a random vector, then the event $(X \leq x)$ is replaced with $(X_1 \leq x_1, ..., X_n \leq x_n)$ (Shao 2003, p. 4). When denoting that X follows the statistical model F with parameter θ, the notation $X \sim F(x|\theta)$ is used.

For simplicity, from here on we will only consider cases where the data X is continuous. A statistical model can be described using what is known as the probability density function (PDF) or the density function. The PDF is defined using the CDF as

$$f(x|\theta) = \frac{d}{dx}F(x|\theta).$$

The density function is commonly used as it is very useful in both visualization of the model and in analytical and numerical inference of the data and parameters.

When studying data in the context of a statistical model, it is assumed to be sampled given some true value for the parameter θ. Usually, the parameters are unknown and not directly observable. The goal of a model problem is to gain as much information as possible about the unknown parameters. One aspect is to find the best guess for the parameter given the data, known as a point estimator. Another aspect is to expand on the point estimator by quantifying the uncertainty of the location of a parameter. There are multiple methods developed for parameter inference. In fact, mulitple statistical fields have been created based on different aproaches. Two of the larger fields are frequentist and Bayesian statistics which will be mentioned more later. The field of fiducial inference will also be visited.

2.1.1 Data generating function

A useful way of studying a model problem is to denote the model using a relation between the observed data X, the parameters θ and some random variable U that is independent of the parameters. By introducing such a relation, it is possible to study either the data or the parameters using the known distribution of U. There are a total of three types of relations, where one of the components are expressed as a function of the others.

The first is the data generating function. It describes how the data is obtained by some underlying process U which is transformed to the data Y using the true parameter θ. The definition is as follows:

Definition 2.1. Let θ be a parameter in Ω_Θ, X be random data mapping to Ω_X and G be a function $G : \Omega_\Theta \times \Omega_U \to \Omega_X$. G is a data generating function if

$$X = G(\theta, U),$$

where $U \in \Omega_U$ is a random variable with distribution independent of θ.

Data generating functions can be created directly from the observed data or from some sufficient statistics of the data. There might be multiple ways of generating data, which means that a data generating function does not need to be unique. An example of non-uniqueness can be seen in both section 3.7.2 and 3.7.4

A second relation is a pivot or pivotal quantity. A pivot is a function of the data and the parameter that has distribution independent of the parameter. The definition of a pivot is as follows

Definition 2.2. A random vector $U = Q(X, \theta)$ is a pivotal quantity if the distribution of U is independent of the parameter θ. That is if $X \sim F(x|\theta)$, then $U = Q(X, \theta)$ has the same distribution for all values of θ (Casella and Berger 2002, Definition 9.2.6).

Out of the three, the pivot is a more common term. That is due to its usefulness in model testing and uncertainty quantification. These topics will be discussed further later.

The third and final representation is the model generating function. This function describes the parameter θ as a relation between the observable data and the random variable U. The definition is as follows:

Definition 2.3. Let Θ be a random parameter in Ω_Θ, U be a random variable in Ω_U distributed independently of θ, $X = x$ be a observed data in Ω_X and M be a function $M : \Omega_X \times \Omega_U \to \Omega_\Theta$. M is a model generating function if

$$\Theta = M(x, U).$$

The distribution of a model generating function gives a distributions estimator for the parameter θ. The data is no longer treated as a random variable, but rather observed data x. Additionally, the model generating function applies a distribution to the parameter given by U. If no distribution was assumed for the parameter prior to the creation of a model generating function, M and U can be chosen almost arbitrarily. However, a well constructed model generating function can prove to give useful inference for the parameter.

There are various advantageous properties to each of these relations. When estimating parameters, a model generating function is generally the goal. The form $\Theta = M(x, U)$ gives a relation to the data as well as some random process U. This allows for a distribution of the parameter which is adjusted by the data. However, a well constructed model generating function is not easily available. Generally they can be obtained using either a pivot or a data generating function using some form of inversion with respect to the parameter. The challenge is that they might not be invertible. For example if the dimension of the data is larger than the dimension of the parameters, solutions to the data generating functions might not exist. A model generating function can be obtained in other ways as well. It is for instance possible to create a model generating function based on a distribution of the parameter, see Bayesian statistics. In some of these situations, finding a density function for the distribution of Θ might be more fruitful. One of the advantages of model generating

functions is that they can be used for sampling from the distribution which can be very efficient.

When inverting either a data generating function or a pivot into a model generating function, it should be noted that the latter will treat the data as constant and the parameter as a random variable.

An example of all three relations can be found for a normal distribution with unknown mean and known variance 1. In that case, a data generating function is

$$X_i = \theta + U_i,\ U_i \sim N(0,1),\ i = 1, ..., n.$$

By using the sufficient statistics $\bar{X} = \sum X_i/n$ (Casella and Berger 2002, Example 6.2.4) the data generating function can be reduced to

$$\bar{X} = \theta + U,\ U = \frac{1}{n}\sum_{i=1}^{n} U_i \sim N(0, 1/n).$$

The dimension of the data and the parameter is equal such that inversions wrt. θ is possible. A pivot can be calculated as

$$U = Q(X, \theta) = \bar{X} - \theta \sim N(0, 1/n).$$

Finally, a model generating function is

$$\Theta = M(x, U) = \bar{x} - U \sim N(\bar{x}, 1/n),$$

from either the inversion of the data generating function or the pivot. As these two are one-to-one, the inversions are the same.

In this case, the inversion from the data generating function into a pivot and a model generating function is trivial. If the original data generating function was used instead of the reduced one, then neither the inversion of the pivot nor the model generating function would be trivial. As there are more data points than parameters, a solution for θ will not exist for all sets (Y, U). If such a dimension reduction is not available, other methods have to be used in order to find the model generating function. This is the case in the main problem of this thesis.

2.1.2 Sufficient statistics

Sufficient statistics is an important quantity in the field of parameter estimation. They can be interpreted as sufficient amount of information about the observed data that can be used to estimate an unknown parameter θ (Casella and Berger 2002, p. 272). A more precise mathematical definition of sufficient statistics is

Definition 2.4 (Sufficient statistics)**.** Statistic T(X) is a sufficient statistic for θ if the conditional distribution of the sample X given the value T(X) does not depend on θ.

(Casella and Berger 2002, Definition 6.2.1) A method of determining a sufficient statistic is based on Fisher's factorization theorem.

Theorem 2.1 (Fisher-Neyman Factorization theorem)**.** *If the density of X given the parameter θ is $f(x|\theta)$, $T(X)$ is a sufficient statistic if and only if*

$$f(x|\theta) = h(x)g(T(x), \theta).$$

(Casella and Berger 2002, Theorem 6.2.6)

Sufficient statistics can be anything from scalars to vectors (Casella and Berger 2002, p. 278). However, the dimension of the sufficient statistic does not have an upper bound, only a lower bound. Adding more information does not limit the sufficiency, however removing information can. In order to deal with the smallest possible amount of sufficient information, the term minimal sufficient statistics was introduced. A sufficient statistic is minimal if it can be written as a function of any other sufficient statistics (Casella and Berger 2002, Definition 6.2.11). In that regard, it cannot be reduced any further. Another characteristic of the minimal sufficient statistic is given in the following theorem.

Theorem 2.2. *Let X be distributed with density $f(x|\theta)$. The statistics $S(X)$ is a minimal sufficient statistic for the parameter θ if and only if*

$$\frac{f(x|\theta)}{f(y|\theta)} \text{ independent of } \theta \iff S(x) = S(y).$$

(Casella and Berger 2002, Theorem 6.2.13)

2.1.3 Frequentist and Bayesian statistics

When studying a model problem from a frequentist perspective, the parameter is treated as a fixed and unobservable quantity. The inference about the parameter is therefore only given by the assumed knowledge about the stochastic behaviour of the data. Methods for working on point estimation and uncertainty will therefore base itself solely on the likelihood function of the data and other formulations of the model as seen in subchapter 2.1.1. The definition of a likelihood function is a follows

Definition 2.5 (Likelihood function)**.** Let $X = (X_1, ..., X_n)$ be random observable data with joint distribution $f_X(x|\theta)$ with parameter θ. The likelihood function of θ is

$$L(\theta) = f_X(x|\theta).$$

(Casella and Berger 2002, Definition 6.3.1)
The natural logarithm of the likelihood function is known as the log-likelihood function

$$l(\theta) = \ln L(\theta).$$

Unlike in frequentist statistics, the Bayesian statistician assumes that the parameters can be described as a random variable or vector. The marginal distribution of these parameters $\pi(\theta)$ is known as the prior distribution of θ. An interpretation of the prior distribution represents the prior knowledge about the parameter. The basis of Bayesian statistics is Bayes rule. It states that for the random variables X and Y, the conditional density

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f(y|x)f_X(x)}{f_Y(y)}.$$

$f_X(x)$ and $f_Y(y)$ is the marginal densities of X and Y, $f(x,y)$ is the joint density and $f(y|x)$ is the conditional density of y given x (Casella and Berger 2002, Theorem 1.3.5). If one assumes the prior distributions of θ, given by the PDF $\pi(\theta)$, the distribution of θ, given the observed data $X = x$, can be calculated as

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} \propto f(x|\theta)\pi(\theta).$$

$\pi(\theta|x)$ is known as the posterior distribution of θ. (Schervish 1995, p. 4) More on posterior distributions in section 2.3.2

2.2 Point estimators

Assume a statistical model F for the data X with parameter θ. θ can be a scalar or a vector of parameters. Given the model, a function of the parameters $g(\theta)$ can be of interest. Point estimation denotes the methods for estimating the value of $g(\theta)$ using the observed data $X = x$. The goal will then be to estimate $g(\theta)$ as accurately as possible.

2.2.1 Decision theory

The following is from the book Theory of Point Estimation (Lehmann and Casella 1998, p. 4-7).

One may view the choice of estimator $\delta(X)$ for the parameter $g(\theta)$ as a decision problem. The goal of point estimation is to be as close as possible to the true value on average. It is therefore necessary with a measure of "closeness" which can be used to order the point estimators. There are many ways of measuring "closeness" or distance such as the squared distance $(\delta(X) - g(\theta))^2$. A collective term is loss functions, which measures the loss of choosing an estimator. The loss of estimating $g(\theta)$ by $\delta(X)$ is $L(\theta, \delta(X))$, where θ is the "true" paramter value. Loss functions are designed such that they are non-negative and 0 at $\delta(x) = g(\theta)$. In other words the loss is zero under the correct estimation. The average (expected) loss is known as the risk. The risk is denoted as

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))]. \tag{1}$$

An optimal estimator would be one that minimizes the risk. An issue is that there exists no estimator that minimizes the risk for all θ, unless $g(\theta)$ is constant. This can be proven

by the estimator $\delta(X) = g(\theta_0)$ which will have 0 risk whenever $g(\theta) = g(\theta_0)$. The decision of estimators is therefore not trivial and is dependent on prioritization. There are ways of dealing with this problem. In a more vague formulation, estimators that performs well for some parameter values but performs significantly worse in other areas can be viewed as worse. A more precise formulations is to minimize a measure that combines risk for all parameter values. Two examples are the supremum of the risk over all parameters,

$$\sup_{\theta \in \Omega_\Theta} R(\theta, \delta), \tag{2}$$

and the weighted average risk over all parameters,

$$\int_{\Omega_\Theta} R(\theta, \delta) w(\theta) d\theta. \tag{3}$$

Both are arbitrary, but can be interesting in each of their own regard. Choosing the estimator that minimizes the maximum risk is known as minimaxing, where the estimator is then a minimax estimator. This is a method that is widely used in many fields such as numerical mathematics to minimize the worst case scenario. A minimax estimator ensures that the worst case scenario is restricted as much as possible. However, such a choice can be at the expense of the risk in general. Minimizing the average weighted risk will handle the downside of the minimax estimator, but the choice of weighting $w(\theta)$ is often arbitrary especially if there is no information about θ. A Bayesian statistician might use a prior distribution of θ as the weighting. Such an estimator is known as a Bayesian estimator and will be explored more.

A second problem is that there are multiple choices of loss and risk which can give different minimizers. The choice can often be arbitrary. Squared error is a common choice, however that is mostly due to its simplicity. A philosophy behind the choice of loss function is that they will represent some actual loss of the choice. An example is to minimize the loss of the best prediction following an estimated parameter. This is for instance used for point estimation parameters of an ARMA model in time series modelling (Brockwell and Davis 2016, Burg's Algorithm). However, this can open another box of similar problems with respect to evaluating the best prediction.

2.2.2 Frequentist and Bayesian approach to point estimation

A common choice of parameter estimator for a frequentist, is the maximum likelihood estimator, or MLE for short. The MLE can be seen as the choice of parameters that maximizes the likelihood of the observed data.

Definition 2.6 (Maximum likelihood estimator)**.** Let $L(\theta)$ be the likelihood function of X. $h(X)$ is the maximum likelihood estimator, MLE, if

$$h(X) = \arg\max_{\theta} L(\theta).$$

There are numerous characteristics that have been uncovered for the MLE. Among these are many asymptotic properties (Lehmann and Casella 1998, p. 444).

In a Bayesian approach to point estimation, more information is available. Under the assumption of a prior distribution, the distribution of the parameter is expressed by the posterior distribution. A further development of risk similar to the weighted average risk in equation (3) is now clearly defined by using the prior distribution $\pi(\theta)$ as weighting. This will in turn be the total risk of the estimator δ, $R(\delta)$. An estimator that minimizes such a risk is known as a Bayesian estimator.

The total risk of an estimator is not needed when creating a Bayesian estimator. An estimator that minimizes the Bayes risk for all data points will satisfy the definition of a Bayesian estimator. The Bayes risk is the expected loss over the posterior distribution. That is

$$R(X, \delta(X)) = \int_{\Theta} L(\theta, \delta(X))\pi(\theta|x)d\theta.$$

(Lehmann and Casella 1998, p. 225-228) Unlike the total risk, the Bayes risk of an estimator is given as a function of the data. Minimizing the risk with respect to the function δ is a much more direct procedure and is a natural setting for finding $\delta(X)$.

2.3 Distribution estimators

Uncertainty is given a formal definition by BIPM in their Guide to the expression of uncertainty in measurement, GUM. BIPM is the international organisation which handles both the International System of Units (SI) and the international reference time scale (UTC)(*BIPM homepage*). The guide defines two types of uncertainty

1. Type A: those which are evaluated by statistical methods
2. Type B: those which are evaluated by other means

(JCGM 2008a, p. IX). The classification of type A and B does not make any statement of the origin of the error or the nature of the origin. The purpose of the classification is to separate two approaches for evaluating the uncertainty. Both uses probability distributions in their evaluation, but the methods used to obtain the distribution can differ. Type A is based on series of observations and is therefore more of a frequentist view on uncertainty. Type B is on the other hand based on available information and can therefore be included in a Bayesian perspective of uncertainty. This interpretation is more clearly stated in 5.1.2 in the Supplement 1 of the GUM (JCGM 2008b). As both type A and type B uncertainty can occur for the same measurand, they can be combined. GUM represents this combination as combined standard uncertainty. In this thesis Type A uncertainty is the main focus. Type B uncertainty is introduced as well in the form of Bayesian priors, however none of the priors are based on prior information.

GUM focuses on two ways of reporting uncertainty of a measurement U, standard uncertainty and expanded uncertainty. Standard uncertainty is the uncertainty expressed as a standard deviation s and expanded uncertainty is an interval given by $[U - k \cdot s, U + k \cdot s]$,

where k is the cover factor. This thesis will focus mostly on expanded uncertainty and an expansion of that term into distributions.

Expanded uncertainty can be described as interval estimators. An interval estimator is an interval of some quantity as a function of the data X. These intervals are given by some upper and lower bound on the form

$$I = [a(X), b(X)],$$

where X is some observable data (Casella and Berger 2002, p. 414). In terms of uncertainty, these interval estimators are used to give an expanded estimate of where the true value for the quantity is located. The cover factor k in expanded uncertainty decides either how often or the likelihood of the interval covering the true value. Keep in mind that the frequency and the likelihood of coverage are not necessarily the same. The difference is the characteristic that separates so called confidence intervals from credibility intervals. Confidence intervals are the frequentist choice of uncertainty while the credibility interval is for Bayesian statisticians.

The simplest of the two interval estimators are the credible intervals. Given some prior distribution for the parameter θ and some data $X = x$, the posterior distribution for θ is available. By using the posterior, it is possible to assign any interval with a likelihood of θ being contained by the interval. A credible interval is designed such that the likelihood is at least some level α. The interval is then known as a $\alpha\%$ credible interval (Casella and Berger 2002, p. 435-436).

A credible interval is not possible in a frequentist view. A frequentist cannot assign a probability distribution to the parameter θ and is unable to give a probability for θ to be inside any interval. The alternative is a confidence interval. An $\alpha\%$ confidence interval is designed such that the interval will cover the true value in at least $\alpha\%$ of the cases. That is, if the experiment is repeated m times, then this interval estimator will cover the true parameter value in at least $\frac{\alpha}{100} \cdot m$ of the cases (Casella and Berger 2002, p. 418-419).

A further expansion on interval estimators are distribution estimators. Instead of describing the location of the parameter using intervals, distributions can give a much richer information of where the parameter might be located. Additionally, they can be used to create interval estimators. One can define a distribution estimator using a distribution function

$$C(\theta|X),$$

where x is the observed data. It is necessary that C satisfies the criteria for a distribution function, see (Schervish 1995, Definition B.7). The posterior distribution is an example of a commonly used distribution estimator. Another example is the confidence distribution.

2.3.1 Confidence distribution

The idea of confidence intervals can be expanded further into what is known as confidence distributions, or CDs. They can be defined using both hypothesis tests and confidence intervals. We will here focus on the relationship to confidence intervals for a one-dimensional parameter, as this is the most relevant. In Definition 3.1, Schweder and Hjort 2016 defines a confidence distribution as

Definition 2.7 (Confidence distribution)**.** A non-decreasing right-continuous function of the one-dimensional θ, depending on the data X , say $C(\theta|X)$, is the cumulative distribution function for a confidence distribution for θ provided it has a uniform distribution as function of X.

A confidence distribution, like a posterior distribution, is a distribution estimator dependent on some observed data. With respect to the data, a CDF $C(\theta|X)$ is a stochastic variable for each parameter value θ. As the definition states, at the true parameter value for θ the CD is a uniform distribution for the data. This is a necessary condition that allows the quantiles of the CD to be confidence intervals. The reason is that the α quantile of the CD is given by the inverse $C^{-1}(\alpha|x)$ such that

$$P(\theta_0 \leq C^{-1}(\alpha|X)) = P(C(\theta_0|X) \leq \alpha) = \alpha.$$

This only holds if $C(\theta_0|X)$ is uniform with respect to the data X. As a result, the CD is a cumulative distribution for the confidence of a scalar parameter. Quantiles of the CD is equivalent to one-sided confidence intervals for the parameter. Any two-sided confidence interval can also be created by combining one-sided confidence intervals. Similarly, p-values of any test of the parameter can be calculated using the CD.

As for confidence intervals, confidence distributions can be calculated using pivots. Given the pivot $Q(X, \theta)$ with distribution function G independent of θ, a CD for Θ at $X = x$ is then $C(\theta|x) = G(Q(x, \theta))$. This CD will satisfy the definition 2.7 (Schweder and Hjort 2016, p. 59). It can also be represented using a model generating function, by inverting the pivot $U = Q(X, \theta)$ at $X = x$ wrt. θ. A confidence distribution can be created by inverting certain data generating function wrt. θ as described in proposition 1 by Taraldsen 2021.

2.3.2 Posterior distribution

The Bayesian posterior is also a distribution estimator which could be used for uncertainty quantification. In a Bayesian context, it can be used to assign probabilities to sets of the parameter. It will therefore give a more direct picture of the location of the parameter. Unlike the confidence distribution, the posterior distribution is given by some choice of prior distribution.

The choice of prior distribution is an important part of Bayesian statistics. Different choices can at worst give significantly different results. In certain cases where prior knowledge about a parameter is available, a prior distribution can be used to represent that knowledge. The prior can in that case improve the analysis by using additional information outside of the base model for the data. An issue is the objectivity of the inference using a prior distribution. Especially when no information of the parameter is known.

The alternative is to choose priors that do not represent any prior knowledge. These are known as objective priors and are the essential components in objective Bayesian. As the priors do not represent any prior knowledge, a bigger question about their legitimacy arises. Additionally, capturing objectivity in a prior is not a trivial task. That is due to the fact that formally defining objectivity with regards to priors has proven challenging. (Consonni et al.

2018). As a result, there are many different approaches to an objective prior. Alternatives will be studied in 2.5.

A problem that arises in objective Bayesian is improper priors. A prior is said to be improper if the integral of the prior is not finite. As a result, the prior is not a density function. Despite an improper prior, it is possible that the posterior is proper. The question is the validity of the inference made using such a posterior. There are differing views on how to deal with improper priors. Some argue that they should never be used. Others try to create frameworks which allows for improper priors. A typical approach is to use limiting distributions of posteriors with proper priors to define posteriors with improper priors (Bioche and Druilhet 2016)(Taraldsen, Tufto, and Lindqvist 2018). A more detailed view of the validity of improper priors will not be the focus of this report.

2.3.3 Generalised fiducial distribution

In the early 20th century, R.A. Fisher proposed a method for creating probability distributions for a parameter θ using the likelihood function of the data. This distribution could be used to create interval estimators for the parameter similar to Bayesian methods using posteriors. The difference was that the fiducial distribution would not be based on a choice of prior. In the one-dimensional parameter case the fiducial distribution is what we now name confidence distribution. Fisher did not like the term confidence and argued that the fiducial distribution was a probability distribution similar to a Bayesian posterior. When studying the fiducial methods in multi-parameter problems arose such as nonuniquenes of the distributions and nonexactness of the interval estimators. After a loss of interest during the late 20th century, there was a resurgence in the early 21th century with different approaches to the fiducial argument. Amongst them were the confidence distribution, but also the generalised fiducial inference.(Schweder and Hjort 2016, Chapter 6)

In 2009, Hannig proposed an expansion of the fiducial argument, which was expanded on further in 2016(Hannig et al. 2016). The focus was on problems where a model generating function might not be easily available. Let

$$Y = G(\theta, U)$$

be a data generation function. If there exists a unique solution for θ for all U and Y, then a model generating function can be constructed using the inversion. If not, other approaches are necessary. Hannig et al. 2016 mentions two possible scenarios. Either there exists multiple solutions or there exists no solution for a given set Y and U. For the former problem, Dempster-Shafer calculus is mentioned as a solution. The latter problem is the focus of the generalised fiducial inference.

Given the observed data Y, if there exists no inversion of the data generating function wrt. θ for some $U = u$ then u is removed from the possible sample space of U. In order to avoid what is known as the Borel paradox, the set of admissible U are defined as

$$\mathbb{U}_{y,\epsilon} = \{U : ||y - G(\theta, U)|| \leq \epsilon \text{ for some } \theta\}.$$

The inversion of the data generating function can in turn be based on the set $U|U \in \mathbb{U}_{y,\epsilon}$. Let A be the event s.t. $A_{y,\epsilon} = (U \in \mathbb{U}_{y,\epsilon})$. Using this condition the GFD is defined by the model generating function

$$\Theta = \lim_{\epsilon \to 0} \left[\arg\min_{\theta} ||y - G(U, \theta)|| \middle| A_{y,\epsilon} \right]$$

(Hannig et al. 2016). The random variable on the right hand side converges in distribution.

An explicit formula for solving this problem with various choices for norm $||\cdot||$ is also presented by Hannig et. al.. The formula is on the form similar to a transformation from the likelihood function to a distribution of θ. That is

$$r(\theta|y) \propto f(y|\theta)J(y, \theta),$$

where J is similar to a Jacobian. $J(y, \theta)$ is defined as

$$J(y, \theta) = D\left(\frac{d}{d\theta} G(u, \theta) \Big|_{u=G^{-1}(y,\theta)} \right).$$

If one compared to the Bayesian approach to distribution estimators, $J(y, \theta)$ can be viewed as a data dependent prior. The function $D(A)$ takes in a matrix and return a scalar. Different choices of norm $||\cdot||$ will give different functions $D(A)$. If θ is a scalar parameter and the dimension of y is a vector, then $\frac{d}{d\theta} G(u, \theta)$ is also a vector. If A is a vector, $D(A)$ takes the following expressions under different norms:

1. l_2 norm: $D(A) = \sqrt{\sum_i A_i^2}$
2. l_∞ norm. $D(A) = \sum_i |A_i|$.

(Hannig et al. 2016).

Like posteriors under Jeffreys prior, the GFD is invariant with respect to smooth reparametrization. An important note about the GFD is that it is not unique. As shown, different choices of norm can result in different distribution functions. Similarly, different data generating functions can give different GFD.(Hannig et al. 2016)

2.3.4 Decision theory for distribution estimators

Similarly to point estimator, there is a goal to find the best method to quantify the uncertainty. Loss and risk can be used to evaluate distribution estimators, similarly to point estimators. Similar definitions to loss and risk from Bayesian statistics can be applied to confidence distributions in order to create confidence loss and confidence risk. The following definitions are given the parameter θ, confidence distribution $C(\theta|x)$ and data $X = x$. Confidence loss is

$$L(\theta, C(x|\theta)) = \int_{-\infty}^{\infty} L(\theta, s) dC(s|x)$$

and the confidence risk is

$$R_F(\theta, C(\cdot|\theta)) = E_\theta L(\theta, C(X|\theta)), \tag{4}$$

where E_θ is the expectation over the data X given the true parameter θ. As for point estimators, the confidence distribution with smallest risk is preferred. A confidence distribution is denoted as uniformly better than another confidence distribution if the risk is not greater for all true parameter values. (Schweder and Hjort 2016, p. 161-163)

2.4 Alternatives for loss functions

The following loss functions are examples that will be used further in this report.

- Mean squared error: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$,
- Mean absolute error: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$,
- Fisher information metric: $L(\theta, \hat{\theta}) = \left| \int_\theta^{\hat{\theta}} \sqrt{I(s)} ds \right|$,
- Kullback-Leibler divergence: $L(\theta, \hat{\theta}) = \kappa(f(\cdot|\theta)||f(\cdot|\hat{\theta}))$.

$I(\theta)$ is the Fisher information of the distribution and $\kappa(f(\cdot|\theta)||f(\cdot|\hat{\theta}))$ is the Kullback-Leibler divergence of the model. Both of these will be studied further. One minor comment is that all of these loss functions are symmetric with the exception of the Kullback-Leibler divergence. It is therefore important to use θ and $\hat{\theta}$ appropriately.

2.4.1 MAE and MSE

The mean absolute error (MAE) and mean squared error (MSE) are two direct distance measures of the parameters. Unlike the Fisher information metric and the Kullback-Leibler divergence, they are independent of the model in question. Of the two, the mean squared error is the more common, as it is used in many types of regression. The expectation of the mean squared error, $E_\theta(\theta - \hat{\theta}(X))^2$, is the variance of the estimator $\hat{\theta}$ if $E_\theta(\hat{\theta}) = \theta$. As variance (or the standard deviation) is often used as a measure of uncertainty, optimizing a parameter based on minimal variance is a natural choice.

2.4.2 Fisher information and Fisher information metric

The Fisher information metric is a distance in Fisher information between two different parameter choices. It is based in the field of information geometry, which is an overlap between differential geometry and statistics. The purpose of Fisher information metric is to measure the shortest distance between two parameters and can therefore be used as a similarity between the respective two models. (Taylor 2019)

The Fisher information, or more generally the Fisher information matrix plays a major role in the asymptotic behaviour of the maximum likelihood estimate. It is also generally used

in information theory. It measures the expected information that is given by the observed data about the parameter θ. Some of the useful properties of the Fisher information is that it is invariant to reparametrization and is positive semi-definite. Lehmann and Casella 1998, p. 115-116 defines the Fisher information $I(\theta)$ of a one-dimensional parameter θ as

Definition 2.8 (Fisher information)**.** Let X be distributed with density $f(x|\theta)$, where θ is a one-dimensional parameter. The Fisher information is

$$I(\theta) = E_\theta\left[\left(\frac{\partial l(\theta|x)}{\partial\theta}\right)^2\right],$$

where $l(\theta|x) = \ln(f(x|\theta))$ is the log-likelihood function of the data x.

With the Fisher information defined, the Fisher information metric is expressed in the following definition.

Definition 2.9 (Fisher information metric)**.** Let $I(\theta)$ be the Fisher information of parameter θ. The Fisher information metric is

$$L(\theta,\hat{\theta}) = \left|\int_\theta^{\hat{\theta}} \sqrt{I(s)}ds\right|.$$

(Taylor 2019)

The following are some properties of the Fisher information that are useful.

Corollary 2.2.1. *The Fisher information $I(\theta)$ is non-negative.*

Proof. For any non-negative random variable Y, the expectation Y is also non-negative. If $E(Y) = 0$, then $Y = 0$ almost surely. (Karr 1993, Proposition 4.11)

This implies that

$$I(\theta) = E\left[\left(\frac{\partial l(\theta|x)}{\partial\theta}\right)^2\right] \geq 0,$$

with equality if

$$\frac{\partial}{\partial\theta} l(\theta|x)\Big|_{\theta=\theta_0} = 0$$

for all x. □

Corollary 2.2.2 (Reparametrization of the Fisher information)**.** *Let the $I_\Theta(\theta)$ be the Fisher information of $f(x|\theta)$ for the parameter θ. For every parameter ϕ such that $\theta = \theta(\phi)$, where $\theta(\phi)$ is continuously differentiable, the Fisher information of ϕ is*

$$I_\Phi(\phi) = I_\Theta\big(\theta(\phi)\big)\left(\frac{\partial\theta(\phi)}{\partial\phi}\right)^2.$$

Proof. The proof follows from the chain rule. For the log-likelihood functions $l(\phi) = \ln f(x|\phi)$ and $l(\theta) = \ln f(x|\theta)$

$$\frac{\partial l(\phi)}{\partial \phi} = \frac{\partial l(\theta)}{\partial \theta}\frac{\partial \theta(\phi)}{\partial \phi}.$$

Inserting this expression into the definition of Fisher information gives

$$I(\phi) = E\left[\left(\frac{\partial l(\phi)}{\partial \phi}\right)^2\right] = E\left[\left(\frac{\partial l(\theta)}{\partial \theta}\right)^2\right]\left(\frac{\partial \theta}{\partial \phi}\right)^2$$

□

Another property of Fisher information is that if X and Y are independent random variables with respective Fisher information I_X and I_Y, then the joint (X, Y) has Fisher information $I_{X,Y} = I_X + I_Y$.

Theorem 2.3. *Let X and Y be independent with probability densities $f_X(x|\theta)$ and $f_Y(x|\theta)$ and their respective Fisher informations $I_X(\theta)$ and $I_Y(\theta)$ exists. Then the joint Fisher information $I_{X,Y}(\theta)$ is given by*

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta)$$

(Lehmann and Casella 1998, Theorem 5.8).

Theorem 2.3 is based on Theorem 5.8 of Lehmann and Casella 1998. They present extra conditions which will be satisfied in this thesis. For simplicity, those conditions are left out, but can be found in the book.

2.4.3 Kullback-Leibler divergence

The Kullback-Leibler divergence, also known as Relative Entropy, is the divergence or "difference" of one distribution to a reference distribution. The mathematical definition is as given

Definition 2.10 (Kullback-Leibler divergence)**.** For the two models with probability density functions $f(x)$ and $g(x)$, the Kullback-Leibler divergence of g with respect to f is

$$\kappa(f||g) = -\int_{\Omega_X} f(x) \ln \frac{g(x)}{f(x)} dx = E_f[\ln f - \ln g].$$

The divergence can be interpreted as the information loss that follows from approximating the distribution f with the distribution g (Simpson et al. 2017). The loss function should therefore bring a different type of loss than MSE and MAE. It is crucial to note that this measure is not a distance measure as it does not uphold the criteria of symmetry, that is $\kappa(g||f) \neq \kappa(f||g)$. It does however hold the property of non-negativity, with $\kappa(f||g) = 0$ only for $f = g$, like the other loss functions. Additionally, if X and Y are independent with respective Kullback-Leibler divergence κ_X and κ_Y, then the Kullback-Leibler divergence of the joint (X, Y) is $\kappa_{X,Y} = \kappa_X + \kappa_Y$.

Theorem 2.4. *The Kullback-Leibler divergence of any two models with probability density function f and g is non-negative and is only zero when $f = g$ almost everywhere. If X and Y are independent with respective Kullback-Leibler divergence κ_X and κ_Y, then the Kullback-Leibler divergence of the joint (X, Y) is*

$$\kappa_{X,Y} = \kappa_X + \kappa_Y.$$

(Schervish 1995, Theorem 2.93)

2.5 Alternatives for objective priors

This section will go through the different choices of priors that will be used in this report.

2.5.1 Jeffreys prior

Jeffreys prior is defined as

Definition 2.11 (Jeffreys prior)**.** Let $I(\theta)$ be the Fisher information of the one-dimensional parameter θ. Then Jeffreys prior is defined as

$$\pi_j(\theta) = \sqrt{I(\theta)}$$

(Schervish 1995, p. 122). In addition to its simplisity, Jeffreys prior holds the property of invariance under injective transformation. This property means that calculating the Jeffreys prior for a parameter $\phi(\theta)$, where ϕ is injective, is equivalent to conduct the transformation $\phi(\theta)$ on the prior $\pi_j(\theta)$, i.e. $\pi_j(\phi) = \pi_j(\theta(\phi))|\frac{\partial\theta}{\partial\phi}|$.

Corollary 2.4.1. *Jeffreys prior is invariant to injective transformations.*

Proof. The corollary follows directly from corollary 2.2.2.

$$\pi_j(\phi) = \sqrt{I(\phi)} = \sqrt{I(\theta)}\left|\frac{\partial\theta}{\partial\phi}\right| = \pi_j(\theta)\left|\frac{\partial\theta}{\partial\phi}\right|$$

□

2.5.2 Penalised complexity prior

Another choice of prior is the Penalised complexity prior or PC prior for short. This is a more recent prior which was introduced in the mid to late 2010s. The goal of the prior is to contrast priors that are too flexible or complex which can lead to over-fitting. It is inspired by the principle of Occam's razor which states that a simple model should be chosen as long as there is not a sufficient reason to choose a more complex one (Simpson et al. 2017). The principle is applied by penalizing the divergence from the standard model/simplest model.

The PC prior uses the Kullback-Leibler divergence to measure the loss of a simple model compared to a more complex and flexible one. The Kullback-Leibler divergence is not a

distance as it does not uphold symmetry, however Simpson et al. 2017 argues that the asymmetry is benifitial as Occam's razor is also an asymmetric principle. The complexity measure of the parameter θ is denoted as

$$d(\theta) = \sqrt{2\kappa(f(\,\cdot\,,\theta)||f(\,\cdot\,,\theta_0))},$$

where $f(x|\theta_0)$ is the least complex variation of the family $f(x|\theta)$, and $kappa(f(\,\cdot\,,\theta||f(\,\cdot\,,\theta_0))$ is the Kullback-Leibler divergence.

The penalization on the complexity d is defined such that if a prior of d is $\pi(d)$ and $r \in (0,1)$

$$\frac{\pi(d+\delta)}{\pi(d)} = r^{\delta},$$

for every $\delta > 0$. For smaller values of r, the faster the prior will go to zero as the complexity increases. This means that smaller values of r will give a larger penalization. Simpson et al. 2017 states that the solution is that d has to be exponentially distributed with rate-paremeter $\lambda = -\ln r$. Here, large values for λ will imply larger penalization on complexity. The PC prior of the parameter θ is then defined using the transformation $d(\theta)$.

Definition 2.12 (Penalized complexity prior)**.** For the penalization parameter $\lambda \in (0,\infty)$ and distribution $f(x|\theta)$ with θ_0 being the least complex choice of θ, the penalized complexity prior is

$$\pi(\theta) = \lambda e^{-\lambda d(\theta)} \left| \frac{\partial d(\theta)}{\partial \theta} \right|,$$

for $d(\theta) = \sqrt{2\kappa(f(\,\cdot\,|\theta_0)||f(\,\cdot\,|\theta))}$

(Simpson et al. 2017).

There are some variations of the PC prior that occurs for different scenarios. If d has an upper bound, the exponential distribution is not an accurate fit. In these situations, the truncated exponential is a more suited solution. This will not alter the shape of the prior, but rather the normalization. Another case is when $d(\theta)$ is not monotone, but piecewise monotone. In that situation, the equality of the prior in Definition 2.12 is exchanged with a proportionality. The normalization of that prior can be calculated by studying the disjoint intervals where $d(\theta)$ is monotone (Simpson et al. 2017).

2.5.3 Uniform prior

A natural choice of objective prior is the uniform prior, which gives the same amount of weight to each possible outcome. This prior will not add any information about the parameter and is independent of the model of the observable data. The uniform prior is defined as

Definition 2.13 (Uniform prior)**.** The uniform prior of parameter $\theta \in \Omega_\Theta$ is

$$\pi(\theta) = \frac{1}{|\Omega_\Theta|} I_{\{\theta \in \Omega_\Theta\}},$$

where $|\Omega_\Theta|$ is the Lebesque measure of the Ω_Θ.

The uniform prior is proper as long as $|\Omega_\Theta| < \infty$, that is if Ω_Θ is bounded. In the continuous univariate case the set Ω_Θ is on the form (a, b). The Lebesque measure of Ω_Θ is then $b - a$. The interval Ω_Θ can also be closed.

An issue with the uniform prior is that it is not invariant under most bijective transformations. The choice of focus parameter can therefore significantly determine the analysis.

2.5.4 Reference prior

The idea behind the reference prior is to find a prior that maximizing the information gained from the observed data. An alternative perspective is that the prior is the prior with least amount of information. The expected information gain is calculated using the Kullback-Leibler divergence. Bernardo 2005 argues that the reference prior appears to be the only known objective prior distribution which satisfy four "reasonably" necessary conditions of objectivity. Those are Generality, Invariance, Consistent marginalization and Consistent sampling properties.

In 2009, Berger, Bernardo, and Sun presented a formal definition of the reference prior including a formula for constructing them. A prior is defined as a reference prior for a model if it is a permissible prior and has the MMI property (Berger, Bernardo, and Sun 2009, Definition 8). Both permissible priors and the MMI property is defined in the article. As the reference prior is not calculated directly, details will be spared. However a short description of the terms will be presented.

The term permissible priors includes all proper priors, but can also be used for certaint improper priors. The idea is to justify the use of proper posterior distributions with improper prior using a convergence of a sequence of posteriors with proper priors. The convergence is based on expected Kullback-Leibler divergence as "distance" measure. Each proper prior in the sequence of posteriors are defined as the improper prior restricted to a compact set of the parameter space. See Definition 4 and 5 for further details (Berger, Bernardo, and Sun 2009).

The MMI, or Maximizing Missing Information, property is related to the potential information gain for a prior relative to others in the class of prior functions $\mathcal{P}$. The information gain, or expected information, is measured in terms of the expected Kullback-Leibler divergence between a prior and a posterior after k experiments. Finally, a prior $p \in \mathcal{P}$ satisfies the MMI property if, given the class of priors $\mathcal{P}$, the difference in expected information between p and any other in prior in $\mathcal{P}$ on any compact subset of the parameter space is non-negative as the number of experiments k goes to infininty. See definition 6 and 7 for more details (Berger, Bernardo, and Sun 2009).

Berger, Bernardo, and Sun 2009, p. 905 argues that the reference prior is reduced to Jeffreys prior in the continous one-dimensional case, under asymptotic posterior normality.

2.5.5 Invariant prior

Let X be the random observable data of some distribution $F(x|\theta)$. The family F is a group family if under some group of transformation G, the distribution of $Y = gX$ is still in the

family F for all $g \in G$. Let $\bar{G}$ be the group of corresponding transformation for G over the parameter space. The prior distribution Π of θ is invariant with respect to $\bar{G}$ if the distribution of $\bar{g}\theta$ is also distributed by Π for all $\bar{g} \in \bar{G}$.

An invariant prior is a natural choice for prior if a group G exists as it will reflect the invariance of the model in general. Of course, if no such group exists, then an invariant prior of this kind is not possible. (Lehmann and Casella 1998, p. 245-246)

3 Binormal distribution with known mean and variance

3.1 The base model

The two-dimensional vector (X, Y) follows a binormal distribution with mean μ and covariance matrix Σ

$$(X, Y) \sim N(\mu, \Sigma)$$

if the density function is

$$f_{X,Y}(x, y|\mu, \Sigma) = \frac{1}{2\pi}|\Sigma|^{-1/2} e^{-\frac{1}{2}((x,y)-\mu)^{\mathrm{T}}\Sigma^{-1}((x,y)-\mu)}$$

(Casella and Berger 2002, Definition 4.5.10). Assuming known mean equal to zero and known variance equal to 1 for both X and Y, the density function is given as

$$f_{X,Y}(x, y|\rho) = \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right), \tag{5}$$

where ρ is the correlation between X and Y. In certain cases, the correlation is treated as a random variable and will be denoted as P. This can be confused with the probability measure P, however no situation from here on will occur where P is used as anything but a random correlation.

It should be noted that any binormal distribution with known means and variances can be transformed into the case with mean 0 and variance 1.

3.2 Change of variables

It is possible to make a linear transformation of the random variables in a binormal distribution such that the new variables are uncorrelated and therefore also independent. For the binormal with known means 0 and known variances 1, the transformations $U = X + Y$, $V = X - Y$ satisfies that property. That is

Corollary 3.0.1. *Let (X, Y) be a binormally distributed vector with means 0, variances 1 and correlation ρ. Define $U = X + Y$ and $V = X - Y$. Then U and V are independent and $U \sim N(0, 2(1+\rho))$, $V \sim N(0, 2(1-\rho))$.*

Proof. In order to prove the relation, it is sufficient to show that

$$E(U) = E(X + Y) = 0,\ E(V) = E(X - Y) = 0,$$

$$Var(U) = Var(X) + Var(Y) + 2E(XY) = 2 + 2\rho = 2(1+\rho),$$
$$Var(V) = Var(X) + Var(Y) - 2E(XY) = 2(1-\rho).$$
$$Cov(U, V) = E(UV) = E((X + Y)(X - Y)) = EX^2 - EY^2 = 0.$$

As any linear combination of jointly normally distributed random variables are also normally distributed (Casella and Berger 2002, Definition 4.5.10), the corollary is proven. □

Corollary 3.0.1 implies that it is possible to study the properties of the distribution of X and Y through the transformation $U = X + Y$ and $V = X - Y$. The independence of U and V in addition to their very familiar univariate normal distribution can be used further in analysis. This is both with regards to studying properties and analysis as well as simulation of the data. Some of the beneficial properties that will be used is that the Fisher information and Kullback-Leibler divergence of the joint of two independent random variables is the sum of the separate information as seen in 2.3 and 2.4.

3.3 Sufficient statistics

Let U, V be the transformation of X and Y given in 3.0.1. We can find the minimal sufficient statistics of both the X, Y system and the U, V system. For the U, V system, one can find sufficient statistic using knowledge about the normal distribution. A minimal sufficient statistic of a normally distributed variable Z with mean 0 and unknown variance σ^2 is $\sum_i Z_i^2$. The proof follows from the density of the vector $(Z_1, ..., Z_n)$

$$f_{(Z_1,...,Z_n)}(z_1, ..., z_n|\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum z_i^2\right).$$

Using Fisher-Neymanns factorization theorem, it is clear that the statistic $\sum_i Z_i^2$ is sufficient. It will also be minimal as the ratio of the densities between the data $(z_1, ..., z_n)$ and $(\tilde{z}_1, ..., \tilde{z}_n)$ boils down to $\exp(-(\sum_i z_i^2 - \sum_i \tilde{z}_i^2)/(2\sigma^2))$.

As the correlation occur in the variance of U and V separately, the same minimal sufficient statistic apply separately to U and V. Let $(X^{\{n\}}, Y^{\{n\}})$ be a vector of n independent data points of the binormal in (5) on the form $(X_1, Y_1, ..., X_n, Y_n)$, and $U^{\{n\}} = X^{\{n\}} + Y^{\{n\}}$ and $V^{\{n\}} = X^{\{n\}} - Y^{\{n\}}$. Then sufficient statistics of the data can be expressed as

$$\begin{aligned} S_1 &= \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} (X_i + Y_i)^2 \\ S_2 &= \sum_{i=1}^{n} V_i^2 = \sum_{i=1}^{n} (X_i - Y_i)^2. \end{aligned} \tag{6}$$

It is possible to test if these sufficient statistics are in fact minimal using theorem 2.2.

The joint density function of $(U^{\{n\}}, V^{\{n\}})$, with realization $(u, v) = (u_1, v_1, ..., u_n, v_n)$ can be written in terms of the respective sufficient statistics $s_1(u, v) = \sum_{i=1}^{n} u_i^2$ and $s_2(u, v) = \sum_{i=1}^{n} v_i^2$. The density is then

$$f_{U^{\{n\}},V^{\{n\}}}(u, v|\rho) = \left(\frac{1}{2\pi\sqrt{1-\rho^2}}\right)^n \exp\left(-\frac{1}{4}\sum_{i=1}^{n}\left(\frac{u_i^2}{1+\rho} + \frac{v_i^2}{1-\rho}\right)\right).$$

The ratio of the density under two realizations of the data (u, v) and $(\tilde{u}, \tilde{v}) = (\tilde{u_1}, \tilde{v_1}, ..., \tilde{u_n}, \tilde{v_n})$ is

$$\frac{f_{U^{\{n\}},V^{\{n\}}}(u, v|\rho)}{f_{U^{\{n\}},V^{\{n\}}}(\tilde{u}, \tilde{v}|\rho)} = \exp\left(-\frac{1}{4}\left(\frac{1}{1+\rho}\left(\sum_{i=1}^{n} u_i^2 - \sum_{i=1}^{n} \tilde{u_i}^2\right) + \frac{1}{1-\rho}\left(\sum_{i=1}^{n} v_i^2 - \sum_{i=1}^{n} \tilde{v_i}^2\right)\right)\right).$$

The ratio is independent of ρ if and only if the sufficient statistics of (u, v) equals the sufficient statistics of $(\tilde{u}, \tilde{v})$, that is $s_1(u, v) = s_1(\tilde{u}, \tilde{v})$ and $s_2(u, v) = s_2(\tilde{u}, \tilde{v})$. S_1 and S_2 are therefore minimal sufficient statistics.

It is also possible to find minimal sufficient statistics directly from the (X, Y) using Fisher-Neyman factorization theorem, 2.1. The joint density of n binormal vectors (X_i, Y_i), denoted $(X^{\{n\}}, Y^{\{n\}})$, with realization $(x, y) = (x_1, y_1, ..., x_n, y_n)$ is

$$f_{X^{\{n\}},Y^{\{n\}}}(x, y|\rho) = \left(\frac{1}{2\pi\sqrt{1-\rho^2}}\right)^n \exp\left(-\frac{1}{2(1-\rho^2)}\sum_{i=1}^{n}\left(x_i^2 + y_i^2 + \rho x_i y_i\right)\right).$$

It is possible to express the data in the density as the two functions $\sum_{i=1}^{n}(x_i^2 + y_i^2)$ and $\sum_{i=1}^{n} x_i y_i$. If the pair of statistics for (X, Y) are defined as $T_1 = \sum_{i=1}^{n}(X_i^2 + Y_i^2)$, $T_2 = \sum_{i=1}^{n} X_i Y_i$,then it is possible to rewrite the density of the data into $f_{X^{\{n\}},Y^{\{n\}}}(x, y|\rho) = 1 \cdot h(T_1(x, y), T_2(x, y), \rho)$. Under Fisher-Neymann factorization theorem, T_1 and T_2 are sufficient statistics. It is possible to write T_1 and T_2 as linear combinations of S_1 and S_2. That is

$$T_1 = \frac{S_1 + S_2}{2}, \; T_2 = \frac{S_1 - S_2}{4}.$$

As there is a one-to-one relation between the two pairs of statistics, T_1 and T_2 are necessarily minimal as well.

A benefit of using S_1 and S_2 is that they are both independently gamma distributed. The distribution is given by the standardization of U and V,

$$\frac{1}{\sqrt{2(1+\rho)}}U \sim N(0, 1) \implies \frac{1}{2(1+\rho)}\sum_{i=1}^{n} U_i^2 \sim \chi_n^2 = \Gamma\left(\frac{n}{2}, 2\right),$$

where $\Gamma(\alpha, \beta)$ is the gamma distribution with shape parameter α and scale parameter β. Finally, the distribution of S_1 and S_2 are given by

$$S_1(U) = \sum_{i=1}^{n} U_i^2 \sim 2(1+\rho)\Gamma\left(\frac{n}{2}, 2\right) = \Gamma\left(\frac{n}{2}, 4(1+\rho)\right)$$

and

$$S_2(V) = \sum_{i=1}^{n} V^2 \sim \Gamma\left(\frac{n}{2}, 4(1-\rho)\right).$$

A benefit of working with the gamma distribution is that it is a familiar distribution with significant work behind it. Methods for simulating the gamma distribution is well formulated and implemented in various programming languages. Some transformations of the gamma distribution is also known, and can be of help (*Gamma distribution*). Additionally, the sufficient statistics can be used to make a data generating function. The data generating function is in that case

$$S_1 = 2(1+\rho)Z_1, \quad S_2 = 2(1-\rho)Z_2,$$

where both Z_1 and Z_2 are chi-square distributed with n degrees of freedom.

3.4 Symmetry conditions for estimators

When studying estimators for the correlation, it is important that they hold properties that one would expect given the data. The base model for our data is that the vector (X, Y) is binormally distributed with unknown correlation ρ and known mean 0 and variance 1. The definition of the correlation ρ is

$$\rho = \frac{E[XY]}{\sqrt{E[X^2]E[Y^2]}} \tag{7}$$

(Casella and Berger 2002, p. 169). Assume that ρ is the correlation of the vector (X, Y). It is of interest to study the correlation of the vector (Y, X), $(-X, -Y)$, $(-X, Y)$ and $(X, -Y)$. All of these are different variations of the vector (X, Y) which will follow the same model however with potentially different correlations. Using equation (7), it is apparent that (Y, X) and $(-X, -Y)$ will have correlation ρ. For $(X, -Y)$ or $(-X, Y)$ the correlation can be calculated as

$$\tilde{\rho} = \frac{E[X(-Y)]}{\sqrt{E[X^2]E[(-Y)^2]}} = -\frac{E[XY]}{\sqrt{E[X^2]E[Y^2]}} = -\rho.$$

In other words, if either X or Y changes sign, the sign of the correlation will change. Estimators of the correlation should keep these properties.

Before expressing these properties more precisely for estimators, the multiple ways of expressing estimators will be described. Firstly, the point estimator is denoted as $\rho = h(X)$. For the distribution estimators, one could either define it using the density $f(\rho|x)$ or the model generating function $P = M(x, U)$.

The properties will first be expressed for the original data, that is for $(x_1, y_1), ..., (x_n, y_n)$ which are n realizations of (X, Y). If the data was transformed to the vectors $(y_1, x_1), ..., (y_n, x_n)$, it is equivalent to assuming the data is n realizations of the vector (Y, X). Here the correlations should be assumed to act similarly. The same holds for the data $(-x_1, -y_1), ..., (-x_n, y_n)$. However, if the data is changed to $(-x_1, y_1)..., (-x_n, y_n)$, that is if all the X_i changes sign, the data can be assumed to be realizations of $(-X, Y)$. This model has opposite sign for the correlation compared to (X, Y). All estimators should therefore make sure that if all the x_i changes sign, so should the correlation. The same applies to the data $(x_1, -y_1), ..., (x_n, -y_n)$.

All of these properties can be expressed for each type of estimator as shows in definition 3.1.

Definition 3.1 (Symmetry conditions for estimators of (X, Y))**.** Let (x, y) be realization from the density (5) on the form $((x_1, y_1), ..., (x_n, y_n))$. If $h(X, Y)$ is a point estimator for the correlation ρ, it satisfies the symmetry conditions if

$$h(x, y) = h(y, x) = h(-x, -y) = -h(-x, y) = -h(x, -y).$$

If $f(\rho|x, y)$ is a density function of a distribution estimator for the correlation ρ then it satisfies the symmetry conditions if

$$f(\rho|x, y) = f(\rho|y, x) = f(\rho| - x, -y) = f(-\rho| - x, y) = f(-\rho|x, -y).$$

If $P = M([x, y], U_1)$ is a model generating function for the correlation, then it satisfies the symmetry conditions if

$$P = M([x, y], U_1) \iff P = M([y, x], U_2) \iff P = M([-x, -y], U_3) \iff$$

$$P = -M([-x, y], U_4) \iff P = -M([x, -y], U_5),$$

where all U_i are identically distributed.

In terms of the minimal sufficient statistics S_1 and S_2 in (6), the properties are expressed more effectively. Expressed in terms of $(X, Y) = ((X_1, Y_1), ..., (X_n, Y_n))$, S_1 and S_2 are

$$S_1(X, Y) = \sum_{i=1}^{n}(X_i + Y_i)^2, \; S_2(X, Y) = \sum_{i=1}^{n}(X_i - Y_i)^2.$$

It is apparent that $S_1(X, Y) = S_1(Y, X) = S_1(-X, -Y)$ and similarly for S_2. This means that some of the symmetry conditions in definition 3.1 are necessarily satisfied using S_1 and S_2. If either X or Y changes sign then S_1 and S_2 are swapped. That is $S_1(X, Y) = S_2(-X, Y)) = S_2(X, -Y)$ and vice versa. This implies that any estimator based on S_1 and S_2 should satisfy the property where swapping S_1 and S_2 should change the sign of the correlation. The properties for S_1 and S_2 are expressed in definition 3.2.

Definition 3.2 (Symmetry conditions for estimators of (S_1, S_2))**.** Let (X, Y) be realization from the density (5) on the form $((x_1, y_1), ..., (x_n, y_n))$. Also let $s_1(x, y) = \sum_i (x_i + y_i)^2$ and $s_2(x, y) = \sum_i (x_i - y_i)^2$. If $h(S_1, S_2)$ is a point estimator for the correlation ρ then it satisfies the symmetry conditions if

$$h(s_1, s_2) = -h(s_2, s_1).$$

If $f(\rho|s_1, s_2)$ is a density function of a distribution estimator for the correlation ρ then it satisfies the symmetry conditions if

$$f(\rho|s_1, s_2) = f(-\rho|s_2, s_1).$$

If $P = M([s_1, s_2], U_1)$ is a model generating function for the correlation, then it satisfies the symmetry conditions if

$$P = M([s_1, s_2], U_1) \iff P = -M([s_2, s_1], U_2),$$

where U_1 and U_2 are identically distributed.

Both of the corollaries expresses a kind of symmetry conditions for the estimators which are necessary for each estimator. There are other possible conditions that could be studied. For instance, scaling either X and/or Y with a constant a. The issue in that case is that the distribution of the vector will change. If X is scaled with a, then the variance of X will change. This will not change the correlation, unless a is negative, but the model will change. If the model was different, say there were assumed to be a unknown common variance, then

scaling is possible. If both X and Y are scaled with a, then the common variance and the correlation will remain unchanged. Expressing this in terms of S_1 and S_2 is also possible as this will imply that the scaling a^2S_1 and a^2S_2 should not change the correlation.

All the symmetry conditions are actually invariance properties of the model under a group of transformations. These transformations are expressed earlier for both the original data (X, Y) and the suffucient statistics (S_1, S_2). Because of lack of time, these will not be expressed formally. However, estimators that keep the symmetry conditions should be so-called equivariance estimator (Casella and Berger 2002, p. 333). The symmetry conditions of the estimators can be seen in section 3.8.3 and 3.10.6. Additionally, any prior that is invariant under the transformations $\tilde{g}\rho = -\rho$, that is $\pi(\rho) = \pi(-\rho)$, should also satisfy the definition of an invariant prior in section 2.5.5. The posteriors of these priors will then satisfy the symmetry conditions in definition 3.1 and 3.2. This can be seen in section 3.6.6.

3.5 Loss functions

The loss functions that will be used are the same as mentioned in 2.4. For the case with binormal with known mean and variance, the loss funcions are as follows:

Mean squared error:

$$L(\rho, \hat{\rho}) = (\rho - \hat{\rho})^2.$$

Mean absolute error:

$$L(\rho, \hat{\rho}) = |\rho - \hat{\rho}|.$$

Fisher information Metric:

$$L(\rho, \hat{\rho}) = |f(\rho) - f(\hat{\rho})|,$$

where $f(x) = \sqrt{2}\text{arctanh}(\sqrt{2}x/\sqrt{1+x^2}) - \text{arcsinh}x$ is the anti-derivative of $\sqrt{1+s^2}/(1-s^2)$ at x.

Kullback-Leibler divergence:

$$L(\rho, \hat{\rho}) = \kappa(\rho||\hat{\rho}) = -\frac{1}{2}\ln\frac{1-\rho^2}{1-\hat{\rho}^2} + \frac{1-\rho\hat{\rho}}{1-\hat{\rho}^2} - 1.$$

It should be noted that as the Kullback-Leibler divergence is not symmetric, it is important to note that ρ is the "true" correlation and $\hat{\rho}$ is an approximation or estimator.

There are also other possibilities for loss function. An issue with using squared errors or absolute errors is that they do not account for bounded parameter spaces. For instance, a distance of 0.1 might be more consequential along the boundary of the parameter space rather than near the center. If the parameter space is transformed to the real line, this is no longer an issue. There are multiple options for transformation. A common choice of transformation used to approximate confidence intervals for the correlation is Fisher's z-transform $z(\rho) = \text{arctanh}(\rho)$. One of the properties of the z-transform is that given the empirical correlation r, $z(r)$ is approximately normally distributed around $z(\rho)$ (Taraldsen 2020). By applying the squared error on the parameter space of $z = z(\rho)$, the issue of a

bounded parameter is no longer there. The resulting loss function is

z-Mean squared error:

$$L(\rho, \hat{\rho}) = (\operatorname{arctanh}(\rho) - \operatorname{arctanh}(\hat{\rho}))^2.$$

Graphs of the loss functions can be seen in figure 3 for $\rho = 0.0$ and in figure 4 for $\rho = 0.9$.

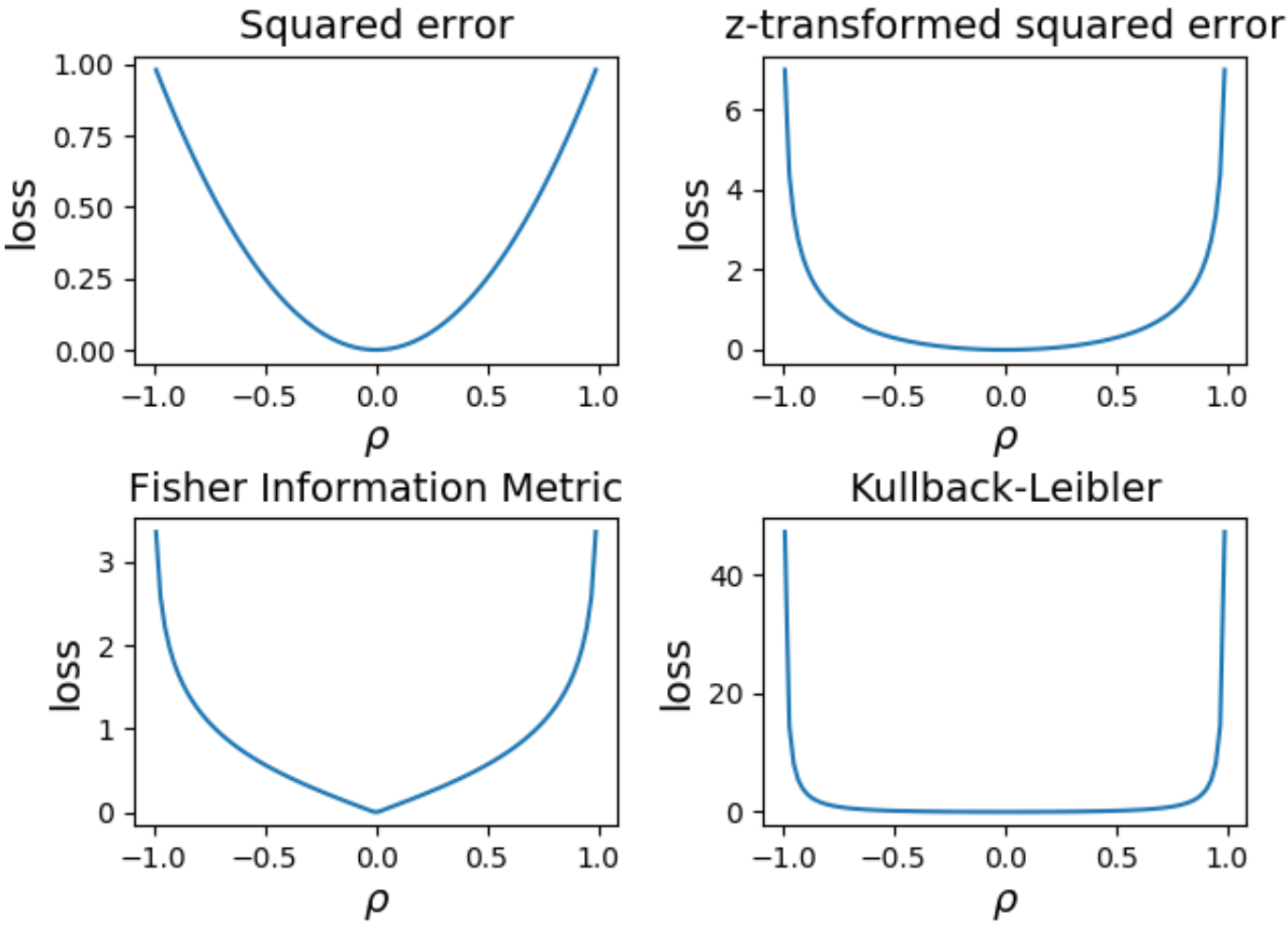


Figure 3: Loss functions for $\rho = 0.0$

3.5.1 Calculating Kullback-Leibler divergence

The Kullback-Leibler divergence (KLD) can be calculated using Corollary 3.0.1 and Theorem 2.4. Firtly, the KLD of V can be calculated using the KLD of U. If we view the distribution of the densities of U and V as $f_U(u|\rho)$ and $f_V(v|\rho)$ respectively, the relation $f_V(v|\rho) = f_U(v|-\rho)$

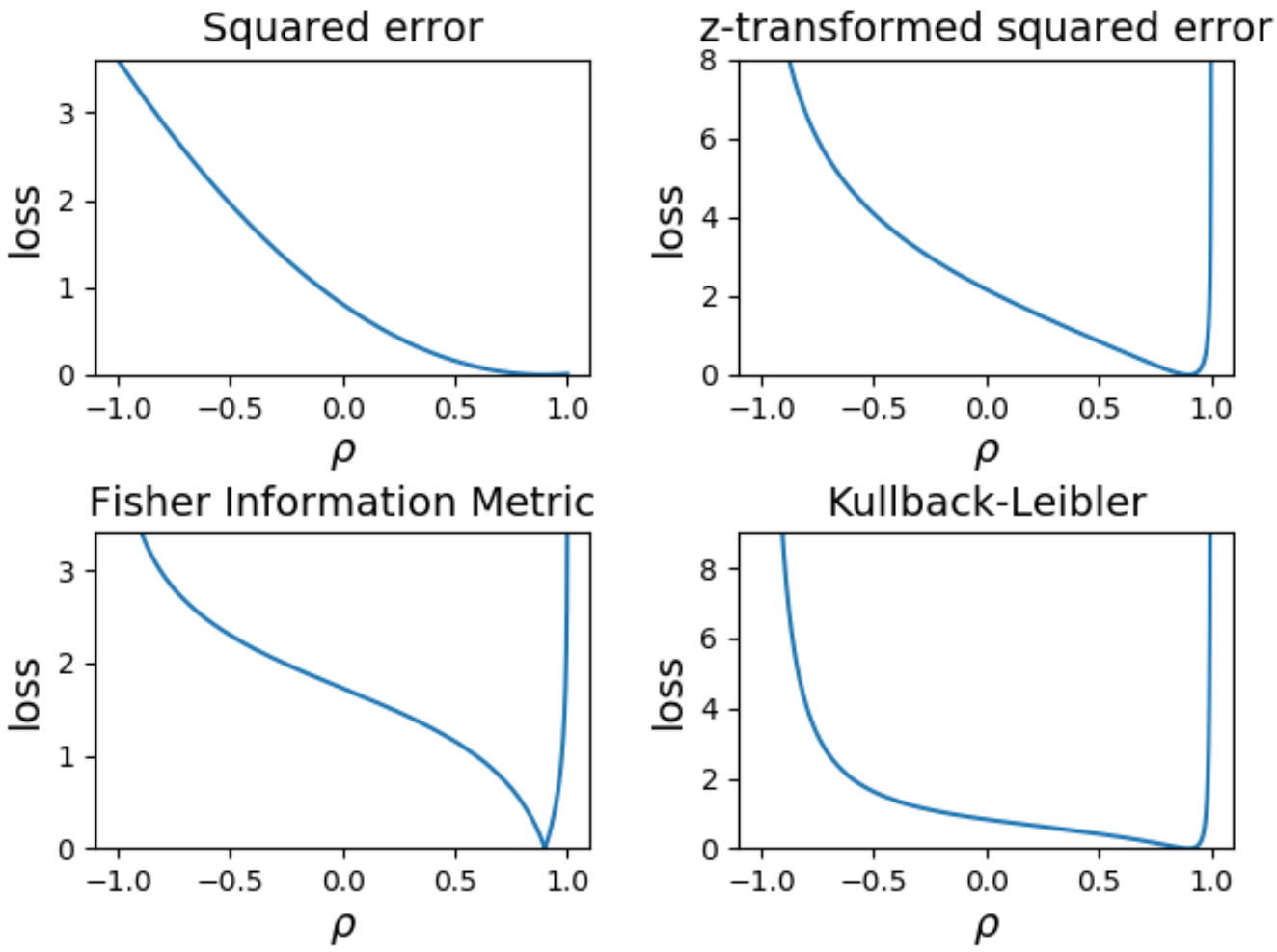


Figure 4: Loss functions for $\rho = 0.9$

holds. The KLD of V is then

$$\kappa(f_V(\,\cdot\,|\rho)||f_V(\,\cdot\,|\hat{\rho})) = \kappa(f_U(\,\cdot\,|-\rho)||f_U(\,\cdot\,|-\hat{\rho})).$$

The first step is to calculate the KLD of U. The Kullback-Leibler divergence of a univariate normal distribution with variance σ^2 is

$$\kappa(f(\,\cdot\,|\sigma_1)||f(\,\cdot\,|\sigma_2)) = \frac{1}{2}\Big(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \ln\frac{\sigma_1^2}{\sigma_2^2}\Big).$$

(Robert 1996, p. 203). By inserting $\sigma_1 = \rho$ and $\sigma_2 = \hat{\rho}$, the Kullback-Leibler divergence of U is

$$\kappa(f_U(\,\cdot\,|\rho)||f_U(\,\cdot\,|\hat{\rho})) = \frac{1}{2}\Big(\frac{1+\rho}{1+\hat{\rho}} - 1 - \ln\frac{1+\rho}{1+\hat{\rho}}\Big)$$

The Kullback-Leibler divergence of V is then

$$\kappa(f_V(\,\cdot\,|\rho)||f_V(\,\cdot\,|\hat{\rho})) = \kappa(f_U(\,\cdot\,|-\rho)||f_U(\,\cdot\,|-\hat{\rho})) = \frac{1}{2}\Big(\frac{1-\rho}{1-\hat{\rho}} - 1 - \ln\frac{1-\rho}{1-\hat{\rho}}\Big).$$

The final expression for the join Kullback-Leibler divergence can be calculated using theorem 2.4. The joint divergence is

$$\kappa(f(\,\cdot\,|\rho)||f(\,\cdot\,|\rho)) = \frac{1-\rho\hat{\rho}}{1-\hat{\rho}^2} - 1 - \frac{1}{2}\ln\frac{1-\rho^2}{1-\hat{\rho}^2}. \tag{8}$$

To be clear, ρ will represent the "true" model while $\hat{\rho}$ represents the approximated model.

3.5.2 Calculating Fisher information and Fisher information metric

It is possible to calculate the Fisher information using the transformation in 3.0.1 and Theorem 2.3.

For univariate normal distributions with mean 0 and variance σ^2, the Fisher information is $I(\sigma^2) = 1/(2\sigma^4)$ (Schervish 1995, Example 2.82). The Fisher information is invariant under reparametrization, see corollary 2.2.2. The Fisher information of parameter ρ under the transformation $\sigma^2 = 2(1+\rho)$ is

$$I_U(\rho) = \frac{1}{2\big(2(1+\rho)\big)^2}\Big(\frac{\partial\ 2(1+\rho)}{\partial\rho}\Big)^2 = \frac{1}{2(1+\rho)^2}.$$

Similarly,

$$I_V(\rho) = \frac{1}{2(1-\rho)^2}$$

The Fisher information for the joint distribution is then the sum of $I_U(\rho)$ and $I_V(\rho)$. That is

$$I(\rho) = \frac{1+\rho^2}{(1-\rho^2)^2}. \tag{9}$$

In order to find the Fisher information metric, it is necessary to find the anti-derivative of the square root of the Fisher information. That is

$$\int \sqrt{I(\rho)}d\rho = \int \frac{\sqrt{1+\rho^2}}{1-\rho^2}d\rho.$$

First, one can rewrite the expression above.

$$\frac{\sqrt{1+\rho^2}}{1-\rho^2} = \frac{1+\rho^2}{(1-\rho^2)\sqrt{1+\rho^2}} = \frac{2}{(1-\rho^2)\sqrt{1+\rho^2}} - \frac{1-\rho^2}{(1-\rho^2)\sqrt{1+\rho^2}}$$

The integral of the second term is known as arcsinh.

$$\int \frac{1}{\sqrt{1+\rho^2}}d\rho = \operatorname{arcsinh}(\rho) + c$$

(Gradshteyn and Ryzhik 2007, p. 2.261). The challenge is then the first term. The solution is to use the transformation $u = \sqrt{2}\rho/\sqrt{1+\rho^2}$.

$$\int \frac{2}{(1-\rho^2)\sqrt{1+\rho^2}} d\rho = \sqrt{2} \int \frac{1 + \frac{u^2}{2-u^2}}{1 - \frac{u^2}{2-u^2}} du = \sqrt{2} \int \frac{1}{1-u^2} du = \sqrt{2}\text{arctanh}\Big(\frac{\sqrt{2}\rho}{\sqrt{1+\rho^2}}\Big) + c$$

(Gradshteyn and Ryzhik 2007, p. 2.172).

Finally, the **Fisher information metric** is

$$\begin{aligned} L(\rho, \hat{\rho}) &= |f(\rho) - f(\hat{\rho})|, \text{ where} \\ f(x) &= \sqrt{2}\text{arctanh}\Big(\frac{\sqrt{2}x}{\sqrt{1+x^2}}\Big) - \text{arcsinh}(x). \end{aligned} \tag{10}$$

3.6 Choice of priors for the correlation

This subsection will go through the different priors that will be discussed in the report. Firstly, all the priors will be introduced before the calculations of the priors are presented. $S_1 = s_1$ and $S_2 = s_2$ are the sufficient statistics defined in (6).

The first prior is the **uniform prior**

$$\pi_u(\rho) = \frac{1}{2} \propto 1,$$

with the following posterior

$$\pi_u(\rho|s_1, s_2) \propto \frac{1}{(1-\rho^2)^{n/2}} \exp\left(-\frac{1}{4}\Big(\frac{s_1}{1+\rho} + \frac{s_2}{1-\rho}\Big)\right).$$

The second prior is **Jeffreys prior**

$$\pi_j(\rho) \propto \frac{\sqrt{1+\rho^2}}{1-\rho^2}, \tag{11}$$

with the following posterior

$$\pi_j(\rho|s_1, s_2) \propto \frac{\sqrt{1+\rho^2}}{(1-\rho^2)^{n/2+1}} \exp\left(-\frac{1}{4}\Big(\frac{s_1}{1+\rho} + \frac{s_2}{1-\rho}\Big)\right). \tag{12}$$

The third prior is the **PC prior**

$$\pi_{pc}(\rho) = \frac{1}{2} \frac{\lambda|\rho|}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} \exp(-\lambda\sqrt{-\ln(1-\rho^2)}), \tag{13}$$

where λ is a parameter that penalizes complexity of the prior. The following posterior is

$$\pi_{pc}(\rho|s_1, s_2) \propto \frac{\lambda|\rho|}{\sqrt{-\ln(1-\rho^2)}} \frac{1}{(1-\rho^2)^{n/2+1}} \exp\left(-\frac{1}{4}\left(\frac{s_1}{1+\rho} + \frac{s_2}{1-\rho}\right) - \lambda\sqrt{-\ln(1-\rho^2)}\right). \tag{14}$$

The choice of penalization in this report will be $\lambda = 10^{-4}$. Section 3.6.2 discusses why this choice is approximately equivalent to the limit of the PC prior as $\lambda \to 0$. The expression of the limit is given in (20).

The fourth prior is the **arcsine prior**

$$\pi_{as}(\rho) = \frac{1}{\pi}\frac{1}{\sqrt{1-\rho^2}}, \tag{15}$$

with the following posterior

$$\pi_{as}(\rho|s_1, s_2) \propto \frac{1}{(1-\rho^2)^{n/2+1/2}} \exp\left(-\frac{1}{4}\left(\frac{s_1}{1+\rho} + \frac{s_2}{1-\rho}\right)\right). \tag{16}$$

The final and fifth prior is the **arctanh prior**

$$\pi_{at}(\rho) \propto \frac{1}{1-\rho^2} \tag{17}$$

with the following posterior

$$\pi_{at}(\rho|s_1, s_2) \propto \frac{1}{(1-\rho^2)^{n/2+1}} \exp\left(-\frac{1}{4}\left(\frac{s_1}{1+\rho} + \frac{s_2}{1-\rho}\right)\right). \tag{18}$$

All the priors are plotted in figure 5. In order to compare their shapes, the improper priors are multiplied by a constant. The graphs shows that the different priors weigh larger values of the correlation differently.

All of the posteriors are visualized in figure 6. The figure shows that all the posteriors will have similar behaviour. In the posteriors, it is possible to see the effect of priors shapes. Jeffrey is the prior that prefer largest correlations, then arctanh, PC, arcsine and uniform follow in the given order.

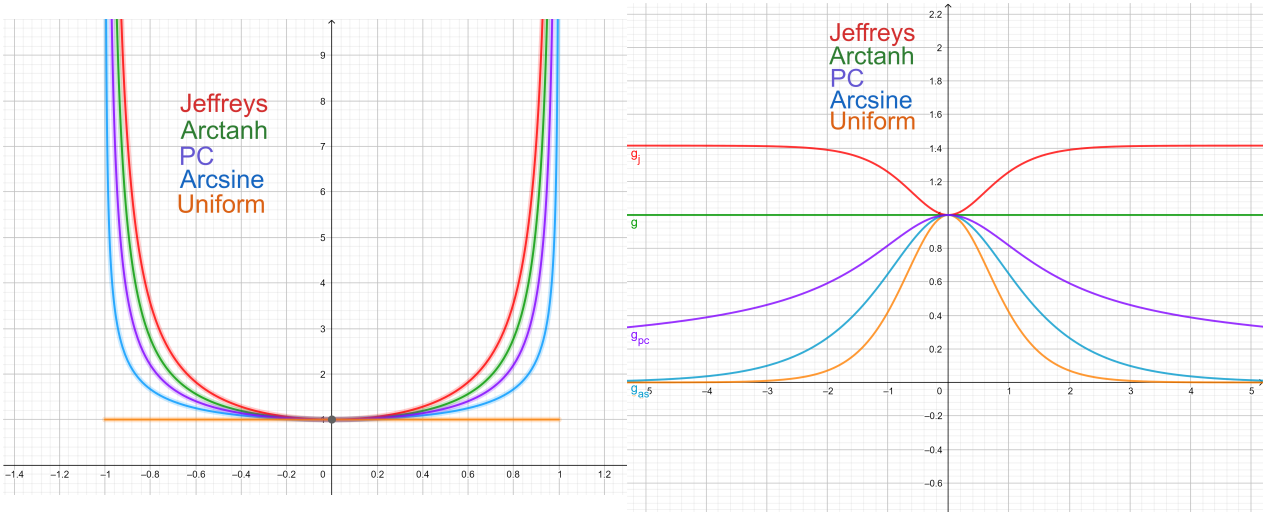


Figure 5: Figure of all the priors as functions of ρ (left) and as functions of $z(\rho) = \text{arctanh}(z)$ (right). All priors are scaled such that they equal 1 at $\rho = z(\rho) = 0$.

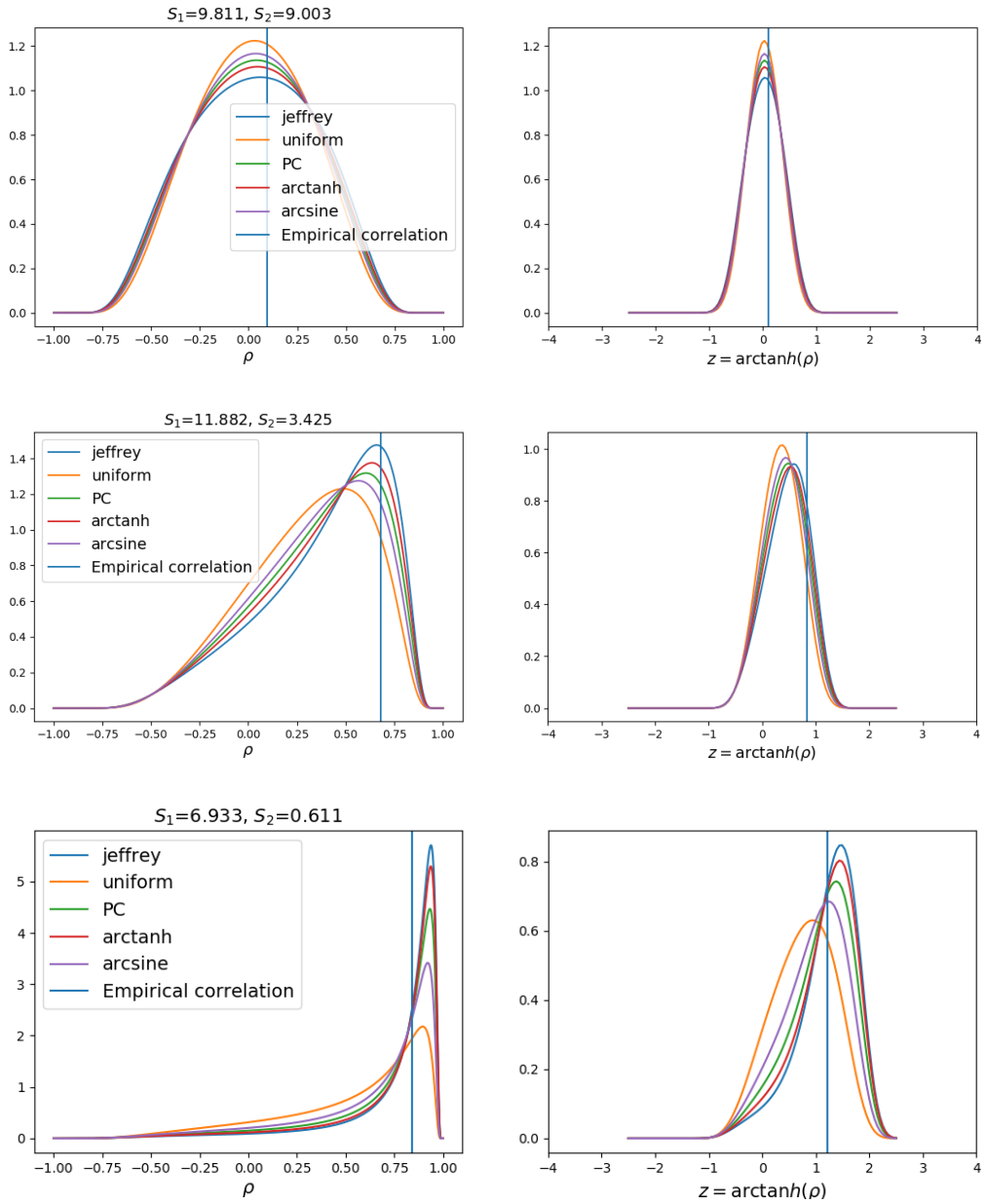


Figure 6: The posteriors for different data samples as function of ρ (left) and $z(\rho) = \text{arctanh}(\rho)$ (right). Each row are based on the data sets given in appendix A.1

3.6.1 Jeffreys prior

As the Fisher information of the model is known, see (9), Jeffreys prior can be calculated.

$$\pi_j(\rho) \propto \frac{\sqrt{1+\rho^2}}{1-\rho^2}.$$

The expression for the prior is only a proportional expression as it is improper. This can be seen in the following calculations.

$$\int_{-1}^{1} \frac{\sqrt{1+\rho^2}}{1-\rho^2} d\rho \geq 2\int_{0}^{1} \frac{1}{1-\rho^2} d\rho = \int_{0}^{1} \left(\frac{1}{1+\rho} + \frac{1}{1-\rho}\right) d\rho$$

$$\geq \int_{0}^{1} \frac{1}{1-\rho} d\rho = -\lim_{\rho \to 1} \ln(1-\rho) = \infty.$$

In fact, the Fisher information metric from end to end is the area of Jeffreys prior. If $L(\rho, \hat{\rho})$ is the Fisher information, then

$$\int_{-1}^{1} \pi_j(\rho) d\rho = L(-1, 1) = \infty.$$

3.6.2 Penalized complexity prior

In the binormal case, the parameter in questions can appear as both the correlation in the original case and in the variances of U and V under the transformation in 3.0.1. There are therefore two ways of arguing for the least complex choice of ρ.

The least complex choice will however be the same for both, which is when X and Y are independent and identically distributed. If that is the case, $U = X + Y$ and $V = X - Y$ will also be independent and identically distributed as the variances will be equal.

By inserting $\hat{\rho} = 0$ into (8), we get the complexity

$$d(\rho) = \sqrt{2\kappa(f(\cdot|\rho)|f(\cdot|0))} = \sqrt{-\ln(1-\rho^2)}.$$

The absolute derivative of d with respect to the correlation is

$$\left|\frac{\partial d(\rho)}{\partial \rho}\right| = \left|\frac{2\rho}{2(1-\rho^2)\sqrt{-\ln(1-\rho^2)}}\right| = \frac{|\rho|}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}}.$$

The PC prior is therefore given as

$$\pi_{PC}(\rho) \propto \frac{\lambda|\rho|}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} e^{-\lambda\sqrt{-\ln(1-\rho^2)}}. \quad (19)$$

It is important to note that the prior is only proportional to the expression above and not equal. The reason is that the transformation $d(\rho)$ is not injective. This means that the

transformation does not satisfy the change of variables formula. It is therefore necessary to study if the prior is proper.

The PC prior is proper with normalization constant 1/2. This is apparent as the transformation $d(\rho)$ is bijective for $\rho \geq 0$ and $\rho < 0$ separately. The area of the expression given in (19) for $\rho \geq 0$ and for $\rho < 0$ are both 1. That implies that the total area of the expression is 2. The normalization constant will therefore be $1/2$.

The exact expression for the PC prior is therefore

$$\pi_{PC}(\rho) = \frac{\lambda|\rho|}{2(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} e^{-\lambda\sqrt{-\ln(1-\rho^2)}}.$$

One problem of the PC prior in (13) is that the denominator is zero as $\rho = 0$. It is therefore necessary to study the limit of the distribution as $\rho \to 0$. Neither $\exp(-\lambda\sqrt{-\ln(1-\rho^2)})$ nor $(1-\rho^2)$ converges to zero, i.e. they are not of interest. It is therefore only necessary to study the limit $\lim_{\rho\to 0} |\rho|/\sqrt{-\ln(1-\rho^2)}$.

$$\lim_{\rho\to 0} \pi_{PC}(\rho) = \lim_{\rho\to 0} \frac{\lambda}{2} \frac{|\rho|}{\sqrt{-\ln(1-\rho^2)}}$$

As both the numerator and the denominator converges to 0, it is possible to use L'Hôpital's rule (Marsden and Weinstein 1985b, p. 522).

$$= \lim_{\rho\to 0} \frac{\lambda}{2} \frac{\rho}{|\rho|} \frac{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}}{\rho} = \frac{\lambda^2}{4\lim_{\rho\to 0} \pi_{PC}(\rho)}.$$

Here, the rule that if $\lim_{x\to x_0} f(x)$ exists and is non-zero, then

$$\lim_{x\to x_0} \frac{1}{f(x)} = \frac{1}{\lim_{x\to x_0} f(x)}$$

(Marsden and Weinstein 1985a, p. 60). The solution of the limit is then

$$\lim_{\rho\to 0} \pi_{PC}(\rho) = \frac{\lambda}{2}.$$

Existence of the PC prior is therefore ensured for $\rho \in (-1, 1)$.

The PC prior creates a family of prior distributions where the penalisation parameter defines it's shape. For large values of λ, the smaller correlations will be weighted. For small values of λ, the correlation around the boundaries will be weighted.

Two scenarios that are interesting in theory, are the two limits of the PC prior. That is for $\lambda \to 0$ or $\lambda \to \infty$. For $\lambda \to \infty$, the prior will converge pointwise to zero everywhere except at $\rho = 0$. This is a result of $e^{-a\lambda}$ being the dominating factor of the sequence for any $a > 0$. The prior will however diverge to infinity at $\rho = 0$ as $\pi_{PC}(0) = \lambda/2$. As long as the distribution of the observed data is non-zero at $\rho = 0$, the posterior distribution will also diverge to infinity at $\rho = 0$ and will therefore not exist there. For that reason, it is not

viable to study that case where $\lambda \to \infty$. An additional comment is that the situation with infinite penalization would necessarily result in a point mass.

For $\lambda \to 0$, the prior will converge pointwise to 0 everywhere except at -1 and 1. The prior will diverge to infinity at both -1 and 1 as $\lim_{\rho \to \pm 1} \pi_{PC}(\rho) = \infty$ for any λ. The PDF of the observed data converges to zero as $\rho \to \pm 1$ faster than the PC prior diverges, $\exp(-s_1/(1+\rho) - s_2/(1-\rho))$ will dominate. This implies that the posterior will also converge to zero at these points. The posterior is therefore bounded and proper. By using a first order Taylor expansion of $\exp(-\lambda\sqrt{-\ln(1-\rho^2)})$ with respect to $\lambda = 0$, the PC prior equals

$$\pi_\lambda(\rho) = \frac{\lambda|\rho|}{2(1-\rho^2)\sqrt{-\ln(1-\rho^2)}}(1 - \lambda\sqrt{-\ln(1-\rho^2)} + o(\lambda)),$$

where o is the little o notation (*Little-O Notation*). Although $\pi_\lambda(\rho)$ converges to 0 almost everywhere as $\lambda \to 0$, the sequence $\frac{1}{\lambda}\pi_\lambda(\rho)$ will converge point wise to

$$\frac{|\rho|}{2(1-\rho^2)\sqrt{-\ln(1-\rho^2)}}.$$

This will holds for any value $\rho \in (-1, 1)$. The pointwise convergence can be seen in figure 7. Using this sequence it is possible to show that the PC prior converges to

$$\pi_{PC}(\rho) \propto \frac{|\rho|}{(1-\rho^2)\sqrt{-\ln(1-\rho^2)}} \tag{20}$$

q-vaguely. The q-vague convergence is often used with regards to prior distribution. Under the conditions of the binormal distribution, the posterior distribution converges q-vaguely if the prior converges q-vaguely. This will in turn ensure beneficial properties for the convergence of estimators. The q-vague convergence is also used to define improper priors as a limit of proper priors (Bioche and Druilhet 2016). The properties of the q-vague convergence will not be discussed further. By proposition 2.10 of (Bioche and Druilhet 2016), the sequence $\pi_n(\rho)$ converges q-vaguely to (20) if

1. there exists a sequence of positive real numbers $\{a_n\}_n$ such that the sequence $\{a_n\pi_n\}_n$ converges pointwise to (20)

2. for any compact set K, there exists a scalar M and some $N \in \mathbb{N}$ such that for $n > N$, $\sup_{\theta \in K} a_n\pi_n(\theta) < M$.

We will define set $\lambda = 1/n$. As $n \to \infty$, $\lambda \to 0$. The first criteria is already satisfied under the sequence $\{1/\lambda(n)\}_n = \{n\}_n$. For the second, it is possible to say that along any compact set $K \in (-1, 1)$, the supremum of $\pi_\lambda(\rho)$ is given by the largest correlation, $\tilde{\rho} = \max_{\rho \in K} |\rho|$. The supremum is bounded if

$$\frac{|\tilde{\rho}|}{2(1-\tilde{\rho}^2)\sqrt{-\ln(1-\tilde{\rho}^2)}} \exp\left(-\lambda\sqrt{-\ln(1-\tilde{\rho})}\right)$$

is bounded for all $\lambda > 0$. As the exponential is always less than 1 for $\lambda \in (0, \infty)$, then the supremum is bounded by

$$M = \frac{|\tilde{\rho}|}{2(1-\tilde{\rho}^2)\sqrt{-\ln(1-\tilde{\rho}^2)}}.$$

The PC prior will therefore converge to (20) q-vaguely as $\lambda \to 0$.

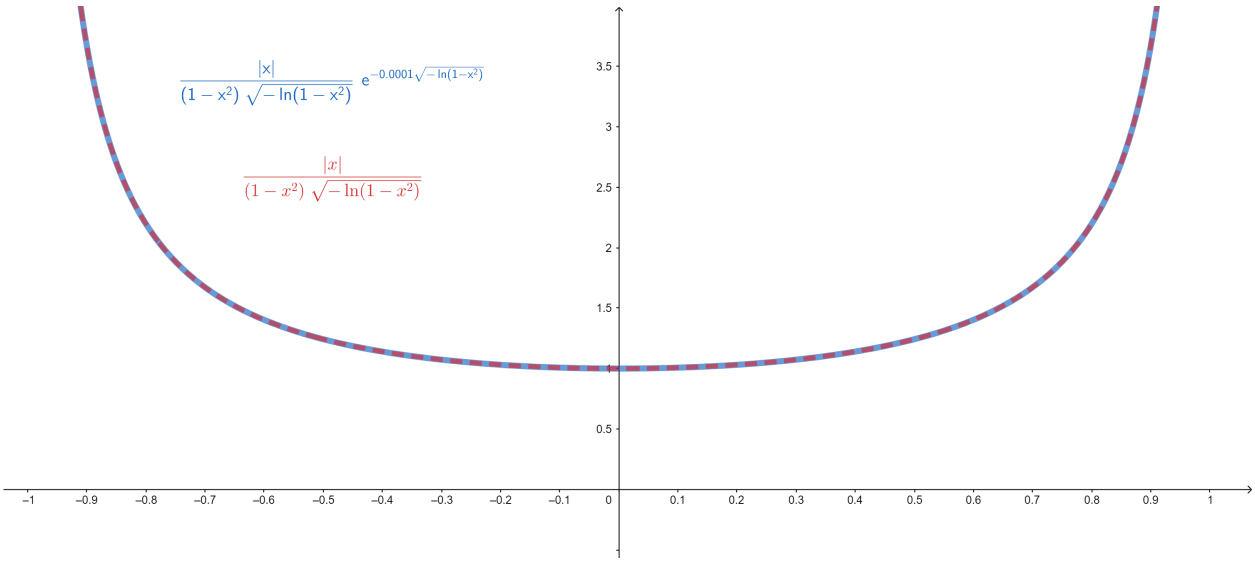


Figure 7: Plot showing the PC prior (in blue) from (13) for $\lambda = 10^{-4}$ and the asymptotic PC prior (in red) from (20).

From the asymptotic expression in (20), the PC prior looks similar to Jeffreys prior. They are not the same, which is especially apparent when studying the priors around $\rho = \pm 1$. Figure 8 compares the two priors and shows the convergence of the PC prior. In this figure both priors are fixed such that they equal 1 at $\rho = 0$.

A small note: There is reason to believe that all PC prior of a one-dimensional parameter will converge q-vaguely to the prior

$$\pi(\theta) \propto \left| \frac{\partial d(\theta)}{\partial \theta} \right|, \tag{21}$$

as $\lambda \to 0$ if $\partial d(\theta)/\partial\theta$ is bounded. The proof is similar to the case for the specific prior for the correlation. The Taylor exansion for the exponential function in the PC prior is similar and the sequence

$$\frac{\pi_\lambda(\theta)}{\lambda} = \left| \frac{\partial d(\theta)}{\partial \theta} \right| (1 - d(\theta)\lambda + o(\lambda))$$

converges point wise to (21). Additionally, the exponential is bounded by 1 as both $d(\theta)$ and λ are non-negative. As a result, the supremum of the sequence $\pi_\lambda(\theta)/\lambda$ along K is bounded if $|\partial d(\theta)/\partial\theta|$ is bounded along K for any compact set K. If the derivative is bounded as well, then the PC prior will converge q-vaguely.

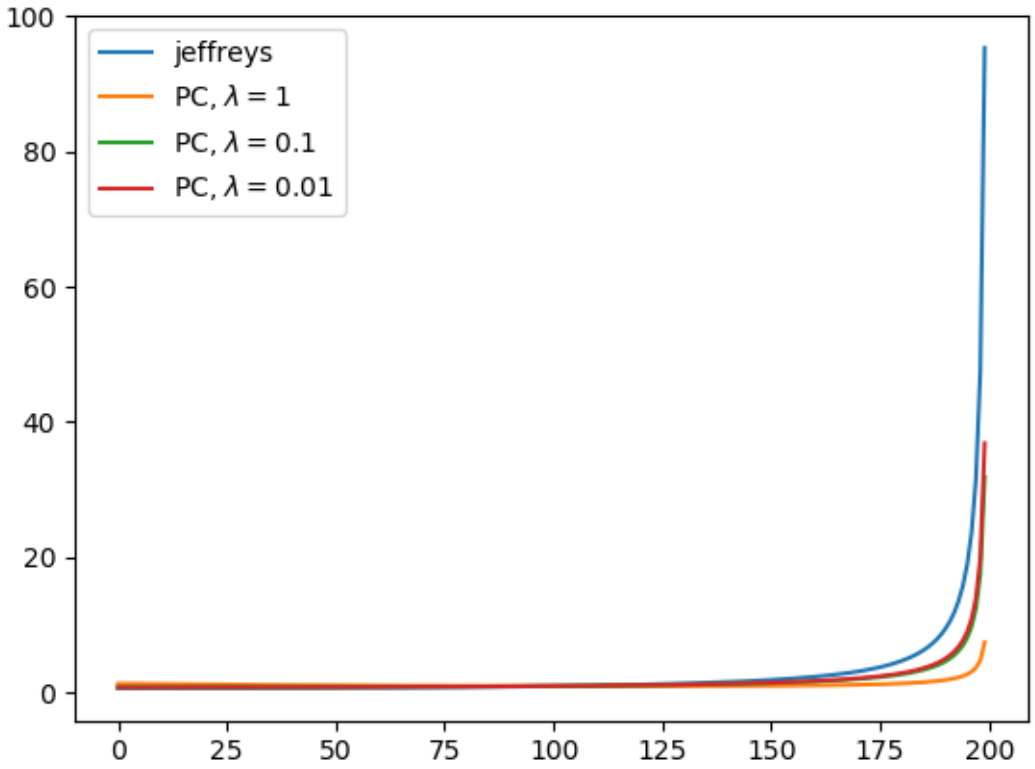


Figure 8: Comparison between Jeffreys prior and PC prios as λ equals 1, 0.1 and 0.01.

3.6.3 Uniform prior

As the parameter space $\Omega_\Theta = (-1, 1)$, the size of the set is $\int_{-1}^{1} d\rho = 2$. This implies that the uniform prior is

$$\pi_U(\rho) = \frac{1}{2} I_{\rho \in (-1,1)},$$

where $I_{\rho \in (-1,1)}$ is the indicator function for the set $\rho \in (-1, 1)$. As the normalization constant is finite, the prior is also proper. It should be noted that the posterior function with the uniform prior will be equal to the pdf of the observed data, however the normalization may differ.

3.6.4 Arcsine prior

The arcsine prior is a prior that is specified for this case and was proposed by Jeffrey. The prior is defined as

$$\pi_{AS} = \frac{1}{\pi\sqrt{1-\rho^2}}$$

(Fosdick and Raftery 2012). One of the benefits of the arcsine is that it is a proper prior with similar shape to Jeffreys prior.

3.6.5 Arctanh prior

Similarly to the arcsine prior, the arctanh prior is specified for this problem. It is the uniform prior under the transformation $z(\rho) = \operatorname{arctanh}(\rho)$. The arctanh prior is then calculated to be

$$\pi_{at}(\rho) \propto \frac{1}{1-\rho^2}.$$

The prior is not proper as the anti-derivative of the prior is naturally the $\operatorname{arctanh}(x)$ function, which diverges at $x = -1$ and $x = 1$.

One of the reasons the prior was chosen, is that an expansion of it will create a confidence distribution for the correlation as a marginal posterior distribution. The expansion is

$$\frac{1}{\sigma^2(1-\rho^2)},$$

under the binormal model with known means, unknown common variance and unknown correlation. The marginal posterior of the correlation can be shown to be the CVCD, see theorem 3.2 and page 47.

3.6.6 Conjugate priors

One could state that all the priors are part of a conjugate family. If a prior distribution is of a conjugate family for the model, then the posterior is also part of that conjugate family (Casella and Berger 2002, Definition 7.2.15). This implies that the posterior distribution of one experiment, can be the prior for a new experiment. In many ways, objective conjugate priors create a true objective prior as it will keep information to the next set of observations while the only biased information in the prior comes from the model choice. There are multiple ways of formulating conjugate prior families. For instance, one could generalize with a family of distributions covering all priors. Instead, an expanding family of conjugate priors will be introduced iteratively.

The first conjugate family to be introduced is on the form

$$\pi(\rho|\alpha,\sigma_1,\sigma_2) \propto (1-\rho^2)^{-\alpha} \exp\left(-\frac{1}{4}\left(\frac{\sigma_1}{1+\rho}+\frac{\sigma_2}{1-\rho}\right)\right).$$

This prior will cover the uniform, the arcsine and the arctanh priors with parameters $\alpha = 0$, $\alpha = 1/2$ and $\alpha = 1$ respectively, and $\sigma_1 = \sigma_2 = 0$.

The parameters of the prior will determine whether it is proper of not. The exponential term of the density will dominate for $\rho \to \pm 1$ as long as σ_1 or σ_2 are non-zero. If they are both positive the prior is proper for all α as it is bounded. If at least one is zero, the parameter α will determine if the prior is proper. For $\alpha < 1$, the prior is proper and for $\alpha \geq 1$, the prior is improper. The argument is based on the conjugate prior's relation to the

beta distribution. Without loss of generality, we assume $\sigma_1 > 0$ and $\sigma_2 = 0$. A lower bound of the integral of the prior is

$$\int_{-1}^{1}(1-\rho^2)^{-\alpha}\exp\left(-\frac{1}{4}\frac{\sigma_1}{1+\rho}\right)d\rho \geq \int_{0}^{1}(1-\rho^2)^{-\alpha}\exp\left(-\frac{1}{4}\frac{\sigma_1}{1+\rho}\right)d\rho \geq e^{-\frac{\sigma_1}{8}}\int_{0}^{1}(1-\rho^2)^{-\alpha}d\rho$$

$$= \frac{1}{2}e^{-\frac{s_1}{8}}\int_{-1}^{1}(1-\rho^2)^{-\alpha}d\rho = e^{-\frac{s_1}{8}}\int_{0}^{1}x^{-\alpha}(1-x)^{-\alpha}dx.$$

The last step used was the transformation $\rho = 2x-1$. The result is a proportional expression to the density of the beta distribution with both parameters equal $-\alpha+1$ (*Beta Distribution*). It is known that the beta distribution is only proper when both parameters are positive, which implies that $\alpha \geq 1$ gives an improper conjugate prior. A upper bound can also be found using

$$\int_{-1}^{1}(1-\rho^2)^{-\alpha}\exp\left(-\frac{1}{4}\frac{\sigma_1}{1+\rho}\right)d\rho \leq \int_{-1}^{1}(1-\rho^2)^{-\alpha}d\rho = 2\int_{0}^{1}x^{-\alpha}(1-x)^{-\alpha}dx.$$

The conjugate prior is therefore also proper if $\alpha < 1$. This result can be compared to the uniform, arcsine and arctanh prior. Both uniform and arcsine are proper as they use $\alpha < 1$, and arctanh is improper as it uses $\alpha = 1$.

An interesting comment about the conjugate prior, is that it is closely related to the beta distribution, as seen in the calculations above. In fact, if $\sigma_1 = \sigma_2 = 0$ they are equivalent. This implies that both the uniform and the arcsine priors are part of the beta distribution family and the arctanh prior is the limit of the beta distribution density as both parameters goes to 0.

The posterior distribution after n data points with sufficient statistics $S_1 = s_1$ and $S_2 = s_2$ on the form (6), is

$$\pi(\rho|s_1, s_1, \alpha, \sigma_1, \sigma_2) \propto (1-\rho^2)^{-\alpha-n/2}\exp\left(-\frac{1}{4}\left(\frac{\sigma_1+s_1}{1+\rho}+\frac{\sigma_2+s_2}{1-\rho}\right)\right).$$

This conjugate family is consistent with Theorem 2.25 presented by Schervish 1995. As seen in the expression above, the posterior is equivalent to the prior, with parameter updates as follows: $\alpha \to \alpha+n/2$, $\sigma_1 \to \sigma_1+s_1$ and $\sigma_2 \to \sigma_2+s_2$. It will therefore represent a conjugate family. It should also be noted that if σ_1 and σ_2 are non-negative, the posterior is always proper as s_1 and s_2 are positive. Only situations with positive σ_1 and σ_2 will therefore be considered.

An expansion of the family above can be made to include either Jeffrey's prior or the PC prior. In order to include Jeffreys prior, some factor $(1+\rho^2)^{\beta}$ can be included in the prior, where Jeffreys prior is given $\beta = 1/2$. For the PC prior one can include $(-\ln(1-\rho^2)/\rho^2)^{-\gamma}$ into the prior. A collective conjugate family of distribution is then given by the density

$$\pi(\rho|\alpha, \beta, \gamma, \sigma_1, \sigma_2) \propto (1-\rho^2)^{-\alpha}(1+\rho^2)^{\beta}\left(-\frac{\ln(1-\rho^2)}{\rho^2}\right)^{-\gamma}\exp\left(-\frac{1}{4}\left(\frac{\sigma_1}{1+\rho}+\frac{\sigma_2}{1-\rho}\right)\right). \tag{22}$$

The conditions for a proper prior is similar to the smaller family described first. It is proper if σ_1 and σ_2 are positive. If either or both equals zero, then the other factors will decide if the prior is proper. The factor $(1+\rho^2)^\beta$ will not affect the existence of a normalization constant. For the parameter γ, it is more difficult to asses which choices that gives proper or improper priors. It is possible to show that $(1-\rho^2)^{-1} \geq (-\ln(1-\rho^2)/\rho^2)^{-1}$ for all $\rho \in (-1,1)$. This would imply that the prior is proper at least for $\alpha+\gamma<1$. However, it is not unlikely that the prior is proper for a larger values for $\alpha+\gamma$ as the logarithm increases relatively slowly. The prior is at least improper if $\alpha \geq 1$ and $\gamma > 0$.

Given n observed data points with sufficient statistics $S_1 = s_1$ and $S_2 = s_2$ on the form (6), the posterior distribution of (22) is

$$\pi(\rho|s_1,s_1) \propto (1-\rho^2)^{-\alpha-n/2}(1+\rho^2)^\beta \left(-\frac{\ln(1-\rho^2)}{\rho^2}\right)^{-\gamma} \exp\left(-\frac{1}{4}\left(\frac{\sigma_1+s_1}{1+\rho}+\frac{\sigma_2+s_2}{1-\rho}\right)\right).$$

These are the same parameter updates as for the first conjugate prior family. In other words, only α, σ_1 and σ_2 are updated. β and γ, on the other hand, will remain constant. This means that they will not reflect the data, but rather the choices of the user. As more data is introduced, α, σ_1 and σ_2 will be updated and dilute the users initial choices for those parameters. A comment is that as data is added, each of these three parameters increase in value. As α increases the factor $(1-\rho^2)^{-\alpha}$ will be more heavily weighted for larger ρ. The posterior will remain bounded because the exponential will always dominate.

The parameter β will not have a large impact on the shape of the prior as $(1+\rho^2) \in (1,2)$. γ on the other hand will have a greater affect as $-\ln(1-\rho^2)$ will diverge to infinity as $\rho \to \pm 1$. The rate of divergence is slower than for $(1-\rho^2)$. As a result, the effect of γ will be diminished as α increases. Another comment about $-\ln(1-\rho^2)/\rho^2$ is that it converges to 1 as $\rho \to 0$. This is shown in the subchapter about PC prior, see 3.6.2. The general argument is that using the Taylor expansion around $\rho = 0$, $-\ln(1-\rho^2) = \rho^2 + o(\rho) \approx \rho^2$ for sufficiently small ρ.

The shape of the family of priors are primarily given by σ_1 and σ_2, and their relative size to α. Firstly, if $\sigma_1 = \sigma_2 = 0$, then the prior is convex. If σ_1 and σ_2 are positive, then the prior changes shape and is bounded at $\rho = \pm 1$. Secondly, the number of modes can vary between one or two depending on the size of σ_1 and σ_2 relative to α.

It is in place to comment on the symmetry conditions in definition 3.2 of posterior distributions generated using the conjugate prior family. As mentioned, the posterior distribution given n data points with sufficient statistics s_1 and s_2 is

$$\pi(\rho|s_1,s_2) \propto (1-\rho^2)^{-\alpha-n/2}(1+\rho^2)^\beta \left(-\frac{\ln(1-\rho^2)}{\rho^2}\right)^{-\gamma} \exp\left(-\frac{1}{4}\left(\frac{\sigma_1+s_1}{1+\rho}+\frac{\sigma_2+s_2}{1-\rho}\right)\right).$$

If the symmetry is to be satisfied, then $\pi(\rho|s_1,s_2) = \pi(-\rho|s_2,s_1)$. As all the factors using α, β and γ includes only ρ^2, the sign will not matter there. Only the exponential will affect the symmetry properties. That is

$$-\frac{1}{4}\left(\frac{\sigma_1+s_1}{1+\rho}+\frac{\sigma_2+s_2}{1-\rho}\right) = -\frac{1}{4}\left(\frac{\sigma_1+s_2}{1-\rho}+\frac{\sigma_2+s_1}{1+\rho}\right).$$

The equality holds only if $\sigma_1 = \sigma_2$. In other words, the symmetry condition in definition 3.2 is satisfied for all of the conjugate priors as long as $\sigma_1 = \sigma_2$. More specifically, all the objective priors introduced earlier satisfy the symmetry conditions as they use $\sigma_1 = \sigma_2 = 0$. These conditions will also correspond to the conjugate prior being an invariant prior as described in section 2.5.5 and 3.4.

3.7 Confidence distributions

A great number of confidence distributions are available through calculations. Only a few of them will be studied. The following confidence distributions will be used in the thesis. Each distribution will be presented using a theorem. Calculations for the distribution can be found in the various subchapters and will be referred to in each theorem. A final overview of the distribution can be seen in section 3.7.6. A greater family of confidence distributions that satisfies the symmetry conditions in definition 3.2 or definition 3.1 is available and described in both section 3.7.2 and 3.7.3.

The first four distributions are exact confidence distributions, that is, all confidence intervals created using them will have exact coverage.

Theorem 3.1 (Unknown Variance Confidence Distribution (UVCD))**.** *If r is the empirical correlation on the form (40) of n data points from the binormal with known mean 0, then a confidence distribution for the correlation is given by the density*

$$f(\rho) = \frac{1}{\sqrt{2}B(n+\frac{1}{2},\frac{1}{2})}(1-r^2)^{\frac{n-1}{2}}(1-\rho^2)^{\frac{n-2}{2}}(1-r\rho)^{\frac{1-2n}{2}}F(\frac{3}{2},-\frac{1}{2},n+\frac{1}{2};\frac{1+r\rho}{2}),$$

and model generating function

$$P = \frac{\gamma}{\sqrt{\gamma^2+U}},$$

where

$$\gamma = \sqrt{V}\frac{r}{\sqrt{1-r^2}} + Z,$$

where $U \sim \chi^2_n$, $V \sim \chi^2_{n-1}$, $Z \sim N(0,1)$, B is the beta function and F is the Gaussian hypergeometric function.

Proof. The proof can be seen in section 3.7.1 from page 45. The density is provided by Taraldsen 2020, which is the reason for the naming. □

Theorem 3.2 (Common Variance Confidence Distribution (CVCD))**.** *If s_1 and s_2 are minimal sufficients statistics on the form (6) of n data points, then a confidence distribution for the correlation is given by the density*

$$f(\rho) = \frac{2}{B(n/2,n/2)}\frac{1}{1-\rho^2}\left(1+\frac{s_2}{s_1}\frac{1+\rho}{1-\rho}\right)^{-\frac{n}{2}}\left(1+\frac{s_1}{s_2}\frac{1-\rho}{1+\rho}\right)^{-\frac{n}{2}} \tag{23}$$

and the model generating function

$$P = \frac{s_1 U_1 - s_2 U_2}{s_1 U_1 + s_2 U_2},$$

where $U_1 \sim \chi^2_n$, $U_2 \sim \chi^2_n$ *and* B *is the beta function.*

Proof. The proof can be seen in section 3.7.1 from page 46. □

A novel CD is presented next.

Theorem 3.3 (Difference Confidence Distribution (DiffCD))**.** *If* s_1 *and* s_2 *are minimal sufficients statistics on the form* (6) *of* n *data points, then a confidence distribution for the correlation is given by the density*

$$f(\rho) = \frac{2^{-n}}{\sqrt{\pi}\Gamma(n/2)} \left| \frac{s_1}{2(1+\rho)} - \frac{s_2}{2(1-\rho)} \right|^{n/2-1/2} K_{1/2-n/2}\left(\frac{1}{2} \left| \frac{s_1}{2(1+\rho)} - \frac{s_2}{2(1-\rho)} \right| \right) \cdot \left(\frac{s_1}{2(1+\rho)^2} + \frac{s_2}{2(1-\rho)^2} \right)$$

and model generating function

$$P = -\frac{1}{4}\frac{s_1 + s_2}{U_2 - U_1} + sgn(U_2 - U_1)\sqrt{\left(\frac{1}{4}\frac{s_1 + s_2}{U_2 - U_1}\right)^2 + 1 - \frac{1}{2}\frac{s_2 - s_1}{U_2 - U_1}}$$

where $U_1 \sim \chi^2_n$, $U_2 \sim \chi^2_n$, Γ *is the gamma function and* K *is the modified Bessel function of the second kind*

Proof. The proof can be found in section 3.7.2 from page 55. □

Theorem 3.4 (CD1)**.** *If* $(x_1, y_1), ..., (x_n, y_n)$ *are* n *data points given the model* (5)*, then a confidence distribution for the correlation is given by the model generating function*

$$P = \frac{1 - \xi}{1 + \xi},$$

where

$$\xi = \prod_{i=1}^{n} \left(\frac{Z_{1,i}}{Z_{2,i}} \frac{x_i - y_i}{x_i + y_i} \right)^{\frac{2}{n}}$$

and all $Z_{1,i}$ *and* $Z_{2,i}$ *are standard normally distributed.*

Proof. Proof can be seen i section 3.7.2 from page 62. □

The following two distributions are generalized fiducial distributions derived in section 3.7.4.

Theorem 3.5 (2-norm Fiducial Distribution)**.** *If s_1 and s_2 are minimal sufficients statistics on the form (6) of n data points, then a generalized fiducial distribution for the correlation is given by the density*

$$r(\rho|s_1,s_2) \propto \sqrt{\left(\frac{s_1}{2(1+\rho)}\right)^2 + \left(\frac{s_2}{2(1-\rho)}\right)^2}(1-\rho^2)^{-n/2}\exp\left(\frac{1}{4}\left(\frac{s_1}{1+\rho}+\frac{s_2}{1-\rho}\right)\right). \quad (24)$$

Proof. The proof can be seen in section 3.7.4. □

Theorem 3.6 (Infininty-norm Fiducial Distribution)**.** *If s_1 and s_2 are minimal sufficients statistics on the form (6) of n data points, then a generalized fiducial distribution for the correlation is given by the density*

$$r(\rho|s_1,s_2) \propto \left(\frac{s_1}{2(1+\rho)}+\frac{s_2}{2(1-\rho)}\right)(1-\rho^2)^{-n/2}\exp\left(-\frac{1}{4}\left(\frac{s_1}{1+\rho}+\frac{s_2}{1-\rho}\right)\right). \quad (25)$$

Proof. The proof can be seen in section 3.7.4. □

Two other fiducial distributions were also created. Their performance was so poor that they will not be used any further. They are however calculated in section 3.7.4.

3.7.1 CD for expanded models

It is possible to expand on the binormal model with known means and variances. The expansions consist of models with additional unknown parameters to the correlation. Two expanded models that will be used: 1) the binormal with known means and unknown variances and 2) the binormal with known means and unknown common variance. Model 1) will in that case have three unknown parameters and model 2) have two. The confidence distribution of both model 1) and 2) are independent on the true values of the other parameters. They should therefore also be confidence distribution when only the correlation is unknown. As the second model uses less information about the data than the first, it is not unreasonable that the distribution of the second model is more precise.

The distribution of the first model is obtained by Taraldsen 2020. The distribution follows from the known relation

$$\sqrt{U}\frac{\rho}{\sqrt{1-\rho^2}} - \sqrt{V}\frac{r}{\sqrt{1-r^2}} = Z \sim N(0,1), \quad (26)$$

where $U \sim \chi^2_n$, $V \sim \chi^2_{n-1}$ and r is the empirical correlation, see section 3.8. The relation above is given by the empirical correlation, which cannot be expressed in terms of the minimal sufficient statistics of our problem. This is due to the fact that the expanded model will have three minimal sufficient statistics and not two. The confidence distribution can either be defined using the model generating function or the corresponding density function. The first can be found by solving (26) with respect to ρ. The solution is then

$$P = \frac{\gamma}{\sqrt{\gamma^2+U}},$$

where

$$\gamma = \sqrt{V}\frac{r}{\sqrt{1-r^2}} + Z.$$

Taraldsen 2020 provides the density of the CD as

$$f(\rho) = \frac{n(n-1)\Gamma(n-1)}{\sqrt{2\pi}\Gamma(n+\frac{1}{2})}(1-r^2)^{\frac{n-1}{2}}(1-\rho^2)^{\frac{n-2}{2}}(1-r\rho)^{\frac{1-2n}{2}}F(\frac{3}{2},-\frac{1}{2},n+\frac{1}{2};\frac{1+r\rho}{2}),$$

where $\Gamma(x)$ is the gamma function and F is the Gaussian hypergeometric function. A small comment about the normalization constant is that it can be rewritten slightly using the beta function. Firstly, $\Gamma(n-1)(n-1)n = \Gamma(n+1)$ and $\sqrt{\pi} = \Gamma(1/2)$ (*Gamma function Calculator*). Secondly, the fraction $\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta) = B(\alpha,\beta)$, where B is the beta function (*Beta function Calculator*). It is therefore possible to rewrite the normalization constant into

$$\frac{n(n-1)\Gamma(n-1)}{\sqrt{2\pi}\Gamma(n+\frac{1}{2})} = \frac{\Gamma(n+1)}{\sqrt{2}\Gamma(\frac{1}{2})\Gamma(n+\frac{1}{2})} = \frac{1}{\sqrt{2}}\frac{1}{B(n+\frac{1}{2},\frac{1}{2})}.$$

The distribution will satisfy the symmetry condition in definition 3.1 as the only cases where the correlation is not squared, is when it is multiplied with r. The empirical correlation satisfies the symmetry conditions of an estimator, see 3.8.3. If r changes sign due to the symmetry, ρ has to do the same if the density should not change. This will then imply that $f(\rho|x,y) = f(-\rho|-x,y)$ and so on.

The second model has the same sufficient statistics as the nested model, which are similarly distributed. That is

$$\frac{S_1}{2\sigma^2(1+\rho)} = U_1 \sim \chi^2_n, \quad \frac{S_2}{2\sigma^2(1-\rho)} = U_2 \sim \chi^2_n.$$

By dividing the two pivots on each other, the following expression is the pivot

$$\frac{S_2}{S_1}\frac{1+\rho}{1-\rho} = \frac{U_2}{U_1} \sim \beta'(n/2,n/2), \tag{27}$$

where $\beta'(n/2,n/2)$ is the beta prime distribution with both parameter as $n/2$ (*Beta prime distribution*). By solving equation (27) with respect to ρ under data $S_1 = s_1$ and $S_2 = s_2$, the model generating function is

$$P = \frac{s_1U_2 - s_2U_1}{s_1U_2 + s_2U_1}.$$

The density function for ρ is

$$f(\rho) = \frac{2}{B(n/2,n/2)}\frac{1}{1-\rho^2}\left(1+\frac{s_2}{s_1}\frac{1+\rho}{1-\rho}\right)^{-\frac{n}{2}}\left(1+\frac{s_1}{s_2}\frac{1-\rho}{1+\rho}\right)^{-\frac{n}{2}}. \tag{28}$$

One comment about the above CD is that under the transformation

$$z(\rho) = \text{arctanh}(\rho) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho},$$

the distribution is symmetric around the point $\ln(s_1/s_2)$. This is apparent by taking the logarithm of (27) as U_1 and U_2 are identically distributed. The transformation $z(x)$ is a commonly used transformation when studying the correlation in a binormal model. It is known as Fisher's z-transformation (Fosdick and Perlman 2016). The CD is also a marginal posterior distribution for the correlation under model 2) with prior $\frac{1}{\sigma^2(1-\rho^2)}$. The join posterior for ρ and the common variance σ^2 is

$$\pi(\rho,\sigma^2|s_1,s_2) \propto \frac{1}{\sigma^2(1-\rho^2)}\frac{1}{(2\pi)^n}\frac{1}{\sigma^{2n}}\frac{1}{(1-\rho^2)^{n/2}}\exp\left(-\frac{1}{4\sigma^2}(\frac{s_1}{1+\rho}+\frac{s_2}{1-\rho})\right)$$

given n data points (x_i, y_i) with sufficient statistics s_1 and s_2 on the form

$$s_1(x_1,y_1,...,x_n,y_n) = \sum_{i=1}^{n}(x_i+y_i)^2,\ \ s_2(x_1,y_1,...,x_n,y_n) = \sum_{i=1}^{n}(x_i-y_i)^2.$$

The posterior can be rewritten into

$$\pi(\rho,\sigma^2|s_1,s_2) \propto \frac{1}{(2\pi)^n}\frac{1}{\sigma^{2n+2}}\frac{1}{(1-\rho^2)^{n/2+1}}\exp\left(-\frac{1}{4\sigma^2}(\frac{s_1}{1+\rho}+\frac{s_2}{1-\rho})\right).$$

With respect to σ^2, the posterior is a inverse-gamma distribution with parameter $\alpha = n$ and $\beta = \frac{1}{4}(s_1/(1+\rho)+s_2/(1-\rho))$. By integrating out σ^2, the marginal posterior for the correlation is

$$\pi(\rho|s_1,s_2) \propto \frac{1}{(2\pi)^n}\frac{1}{(1-\rho^2)^{n/2+1}}\left(\frac{1}{4}\left(\frac{s_1}{1+\rho}+\frac{s_2}{1-\rho}\right)\right)^{-n}\Gamma(n).$$

By rewriting the expression into

$$\begin{aligned}\pi(\rho|s_1,s_2) &\propto \frac{1}{(1-\rho^2)^{n/2+1}}\left(\frac{s_1}{1+\rho}+\frac{s_2}{1-\rho}\right)^{-n/2}\left(\frac{s_1}{1+\rho}+\frac{s_2}{1-\rho}\right)^{-n/2}\\ &= \frac{1}{1-\rho^2}(s_1s_2)^{n/2}\left(1+\frac{s_2}{s_1}\frac{1+\rho}{1-\rho}\right)^{-n/2}\left(1+\frac{s_1}{s_2}\frac{1-\rho}{1+\rho}\right)^{-n/2},\end{aligned}$$

one can see that the marginal posterior for the correlation is the CVCD.

3.7.2 CD from pivots

As mentioned in section 2.1.1, pivots and data generating functions can be used to create model generating functions. Similarly as mentioned in 2.3, pivots are a common method for

creating confidence distributions. Pivots are therefore a natural base for creating CDs either through the approach of finding a density or creating a model generating function.

One could create pivots in multiple ways, for instance using either a data generating function of the original data or data generating function of the minimal sufficient statistics S_1 and S_2. For simplicity, the sufficient statistics will be used as basis. Using the original data or some other transformation of the data is possible. However, the difficulties in the following analysis will only be exacerbated as the dimension of the data is significantly larger. A pivot for S_1 and S_2 can be written as

$$\frac{S_1}{2(1+\rho)} = U_1 \sim \chi^2_n, \quad \frac{S_2}{2(1-\rho)} = U_2 \sim \chi^2_n.$$

In terms of model generating function, it is necessary to have a solution for ρ for any given set $S_1 = s_1$, $S_2 = s_2$, $U_1 = u_1$ and $U_2 = u_2$. An issue is that there are more sufficient statistics than parameters. In addition, if either $u_1 \leq s_1/4$ or $u_2 \leq s_2/4$ no solution can exist. As a result, most sets will not give a solution. By reducing the two equations above down to one through an appropriate function $\phi : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}$, solutions might exist. Applying ϕ to both sides, the following expression is then also a pivot.

$$\phi\left(\frac{S_1}{2(1+\rho)}, \frac{S_2}{2(1-\rho)}\right) = \phi\left(U_1, U_2\right). \tag{29}$$

Not all choices for ϕ are admissible. An example is the function $\phi(x, y) = x$. A framework for ϕ such that the pivot in (29) is invertible is needed. The definition 3.3 expresses conditions for ϕ that, along with theorem 3.7, implicitly defines a confidence distributions.

Definition 3.3. A function $\phi : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto (a, b) \subseteq \mathbb{R}$, is an admissible pivot function or APF if it satisfies the following properties:

1. $\phi(x, y)$ is increasing wrt. x
2. $\phi(x, y)$ is decreasing wrt. y
3. $\lim_{x\to\infty} \phi(x, y) = b$ independent of y
4. $\lim_{y\to\infty} \phi(x, y) = a$ independent of x.

Theorem 3.7. *Let S_1, S_2 be defined as in (6) and U_1 and U_2 are identically and independently chi-square distributed with n degrees of freedom. If $\phi(x, y)$ satisfies 3.3 then the inversion of*

$$\phi\left(\frac{S_1}{2(1+\rho)}, \frac{S_2}{2(1-\rho)}\right) = \phi(U_1, U_2)$$

wrt. ρ at the data points $S_1 = s_1$ and $S_2 = s_2$ is a model generating function for a confidence distribution of ρ.

Proof. As both $\frac{S_1}{2(1+\rho)} = U_1$ and $\frac{S_2}{2(1-\rho)} = U_2$ are pivots, then

$$Q(S_1, S_2, \rho) = \phi\left(\frac{S_1}{2(1+\rho)}, \frac{S_2}{2(1-\rho)}\right) = \phi(U_1, U_2)$$

is also a pivot. An inversion of Q wrt. ρ exists if Q it is a bijective transformation of ρ and it spans the space of $\phi(U_1, U_2)$ for any $S_1 = s_1$ and $S_2 = s_2$. If ϕ satisfies 3.3, then Q is a decreasing function for ρ and therefore a bijection. Additionally, Q will span the space of $\phi(U_1, U_2)$. □

Definition 3.3 and theorem 3.7 gives a framework for creating confidence distribution for the correlation. Even though the theorem only mentions the creation of a model generating function, it is possible to create a density and distribution function as described in section 2.3.1. Finding the density of $\phi(U_1, U_2)$ is necessary for finding the density of the CD.

Not all CDs created using theorem 3.7 will satisfy the symmetry conditions in definition 3.2. An example is the function $\phi(x, y) = x/y^2$. It is an APF, however the corresponding CD will not satisfy the symmetry conditions. The respective pivot function is

$$\frac{S_1}{S_2^2}\frac{(1-\rho)^2}{1+\rho} = \frac{U_1}{U_2^2}.$$

Without showing a proof, the model generating function is

$$P = \frac{U_1}{2U_2^2}\frac{s_2^2}{s_1} + 1 - \sqrt{\left(\frac{U_1}{2U_2^2}\frac{s_2^2}{s_1} + 1\right)^2 + 1 - \frac{U_1}{U_2^2}\frac{s_2^2}{s_1}}. \tag{30}$$

If one would like a proof, it follows the same line of arguments as for the diff CD, as one has to solve a quadratic equation in both situations. Figure 9 shows the distribution of P under $S_1 \approx 2.43$ and $S_2 \approx 0.73$ and $-P$ under $S_1 \approx 0.73$ and $S_2 \approx 2.43$. These distributions do not match, which they should do under the symmetry conditions.

The following corollary poses a reformulation of the symmetry conditions in terms of the pivotal equation.

Corollary 3.7.1. *Let $P = M([s_1, s_2], Z)$ be a model generating function for the correlation such that*

$$\phi\left(\frac{s_1}{2(1+P)}, \frac{s_2}{2(1-P)}\right) \sim \phi(U_1, U_2),$$

where U_1 and U_2 are independently chi-square distributed with n degrees of freedom. The symmetry conditions in definition 3.2 are satisfied if and only if

$$\phi\left(\frac{s_1}{2(1+P)}, \frac{s_2}{2(1-P)}\right) \sim \phi\left(\frac{s_2}{2(1-P)}, \frac{s_1}{2(1+P)}\right).$$

Proof. The proof follows closely to definition 3.2. If the definition is satisfied, then the following two model generating functions satisfy

$$M([s_1, s_2], Z) \sim -M([s_2, s_1], Z),$$

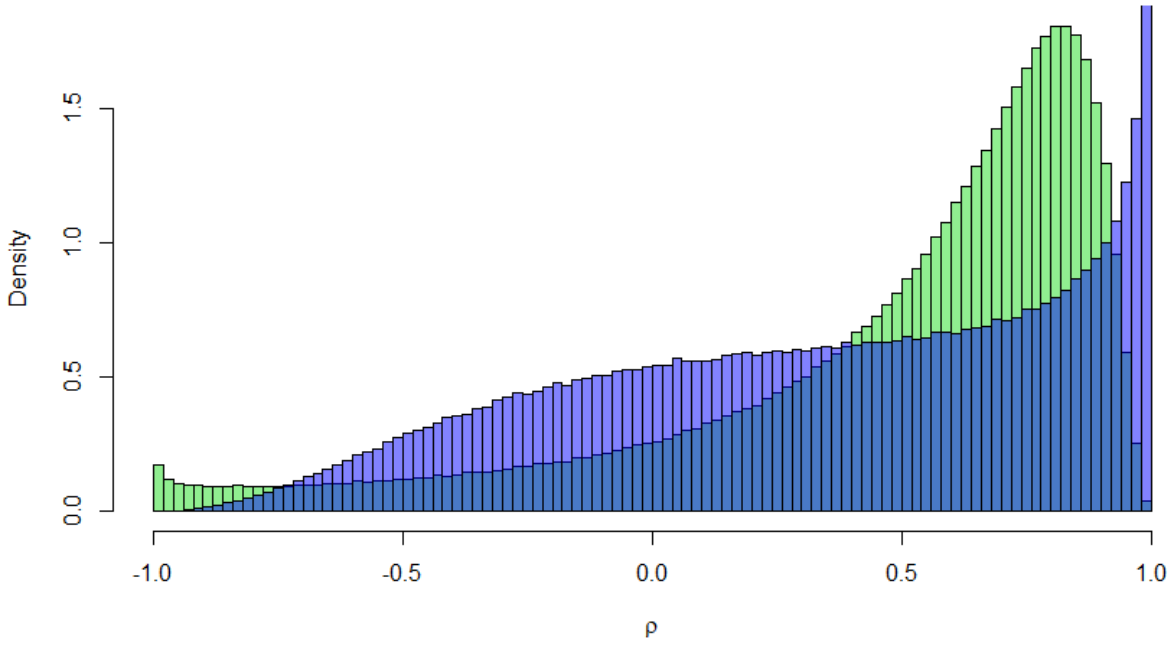


Figure 9: Histogram of the model generating function (30). Green histogram is for the model generating function with data $S_1 \approx 2.43$ and $S_2 \approx 0.73$ and the blue histogram is for the negative model generating function with data $S_1 \approx 0.73$ and $S_2 \approx 2.43$.

where $M([s_1, s_2], Z)$ is the model generating function of (29) given observed data $S_1 = s_1$ and $S_2 = s_2$. If this relation holds true, $-M([s_1, s_2], Z)$ is a model generating function for (29) under data $S_1 = s_2$ and $S_2 = s_1$. As the right hand side of the pivotal equation remains the same between the two scenarios, the distributions of the pivot should not change. In other words, it is necessary that

$$\phi\left(\frac{s_1}{2(1+M([s_1,s_2],Z))}, \frac{s_2}{2(1-M([s_1,s_2],Z))}\right) \sim \phi\left(\frac{s_2}{2(1-M([s_1,s_2],Z))}, \frac{s_1}{2(1+M([s_1,s_2],Z))}\right).$$

If instead, the expression above is true, and $M([s_1, s_2], Z)$ is the model generating function of (29) under $S_1 = s_1$ and $S_2 = s_2$, then $-M([s_1, s_2], Z)$ is a model generating function under $S_1 = s_2$ and $S_2 = s_2$. In other words, it is necessary that

$$M([s_1, s_2], Z) \sim -M([s_2, s_1], Z)$$

and definition 3.2 is satisfied. □

One can compare the conditions in corollary 3.7.1 to the pivot

$$\frac{s_1}{s_2^2}\frac{(1-P)^2}{1+P} \tag{31}$$

with model generating function P as given in (30). The distribution of the pivot is visualized in figure 10 alongside the distribution of U_1/U_2^2. If $\phi(x, y) = x/y^2$ was such that 3.7.1 is satisfied, then the two histograms should be equal. That is clearly not the case.

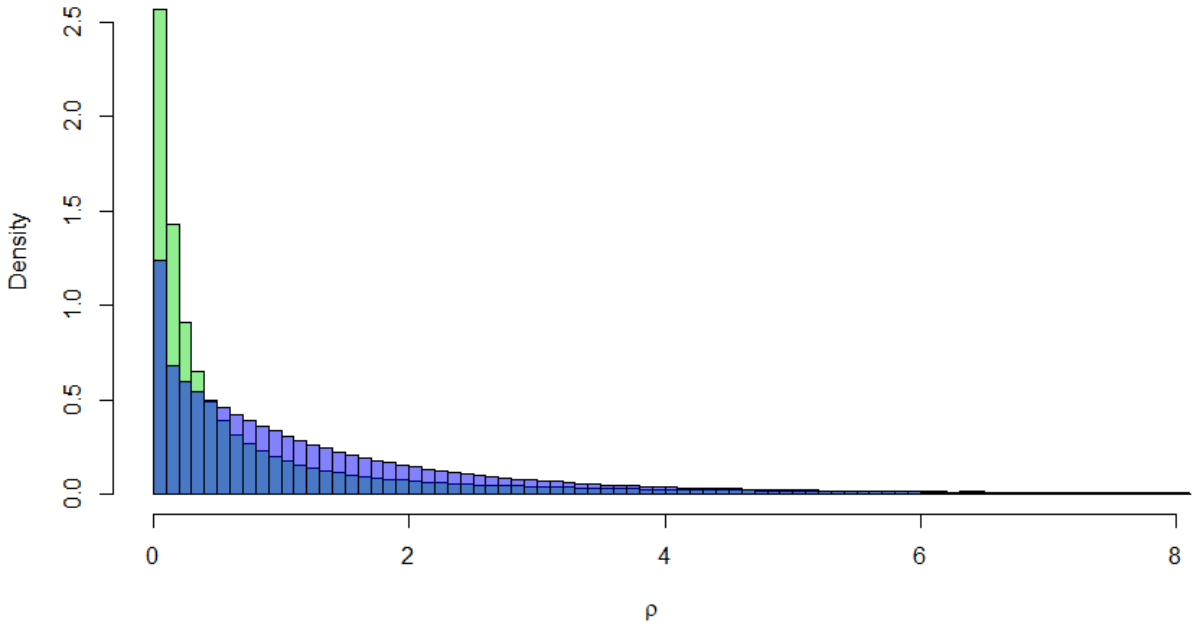


Figure 10: Two histograms of (31). $\tilde{P}$ is the model generating function (30) for $s_1 = 2.43$ and $s_2 = 0.73$. Green histogram is for (31) with data $s_1 \approx 2.43$ and $S_2 \approx 0.73$ and $P = \tilde{P}$. The blue histogram is for (31) with data $s_1 \approx 0.73$ and $s_2 \approx 2.43$ and $P = -\tilde{P}$.

Corollary 3.7.1 does not provide any clear conditions with respect to the APF ϕ. It is however possible to expand on 3.7.1 with conditions on $\phi(x, y)$.

Theorem 3.8. *Let ϕ be an APF as defined in 3.3. The model generating function defined as the inversion of*

$$\phi\left(\frac{S_1}{2(1+\rho)}, \frac{S_2}{2(1-\rho)}\right) = \phi(U_1, U_2)$$

wrt. ρ satisfies the symmetry conditions of definition 3.2 if $\phi(x, y) = h(\phi(y, x))$. h is a decreasing involution. That is, h is it's own inverse.

Proof. It is necessary that h is an involution as $\phi(x, y) = h(\phi(y, x)) = h(h(\phi(x, y))$. Secondly, h is decreasing to make sure that $\phi(x, y)$ and $\phi(y, x)$ both satisfy 3.3.

Let $P = M([s_1, s_2], \phi(U_1, U_2))$ be the model generating function of (29) with observed data $S_1 = s_1$ and $S_2 = s_2$. If ϕ satisfies $\phi(x, y) = h(\phi(y, x))$, then

$$\phi\left(\frac{s_1}{2(1+P)}, \frac{s_2}{2(1-P)}\right) \sim \phi(U_1, U_2)$$

$$\iff h\left(\phi\left(\frac{s_2}{2(1-P)}, \frac{s_1}{2(1+P)}\right)\right) = h(\phi(U_2, U_1)).$$

By applying the function h to the expression above, the equation is

$$\phi\left(\frac{s_2}{2(1-P)}, \frac{s_1}{2(1+P)}\right) \sim \phi(U_2, U_1) \sim \phi(U_1, U_2).$$

Corollary 3.7.1 is therefore satisfied. □

Theorem 3.8 provides an implicit definition of a group of pivots. Each group is given by the involution h. There are many possible choices for h, which can complicate the process of defining "every" confidence distribution given by theorem 3.7. However it is possible to show that any APF $\psi(x, y) = h(\psi(y, x))$ can be reduced to $\phi(x, y) = t(\psi(x, y))$ such that $\phi(x, y) = -\phi(y, x)$. Lemma 3.9 and theorem 3.10 demonstrates the relation and the consequences..

Lemma 3.9. *For any decreasing involution $h(x)$ there exists an increasing function $t(x)$ such that $t(x) = -t(h(x))$.*

Proof. The function $t(x) = x - h(x)$ satisfies the proof for any h. That is

$$t(h(x)) = h(x) - h(h(x)) = h(x) - x = -t(x).$$

t is also increasing as $h(x)$ is decreasing. At least one solution will therefore exist. □

Theorem 3.10. *Let $P = M([s_1, s_2], Z)$ be a model generating function created using theorem 3.7 with any APF ψ satisfying theorem 3.8. There exists an APF ϕ satisfying $\phi(x, y) = -\phi(y, x)$ with model generating function $\tilde{P} = M([s_1, s_2], Z')$ created using theorem 3.7 such that $\tilde{P}$ also is a model generating function under ψ. That is $P \sim \tilde{P}$.*

Proof. For any APF $\psi(x, y)$ such that $\psi(y, x) = h(\psi(x, y))$, there exists an APF ϕ such that $\phi(x, y) = -\phi(y, x)$ and $t(\phi(x, y))$. The proof follows from lemma 3.9. The lemma states that for any involution h there exists a t such that $t(x) = -t(h(x))$. $\phi(y, x) = t(\psi(y, x)) = t(h(\psi(x, y))) = -\phi(x, y)$. The equation

$$\psi\left(\frac{S_1}{2(1+\rho)}, \frac{S_2}{2(1-\rho)}\right) = \psi(U_1, U_2)$$

can be applied t to both sides which results in the pivotal equation

$$\phi\left(\frac{S_1}{2(1+\rho)}, \frac{S_2}{2(1-\rho)}\right) = \phi(U_1, U_2).$$

Both pivotal equations will necessarily have the same solutions as t is monotone and therefore also bijective. □

An example of, and the inspiration for, theorem 3.10 is APFs $\psi(x,y)$ such that $\psi(x,y) = 1/\psi(y,x)$. It can be applied the logarithm, such that $\phi(x,y) = \ln\psi(x,y) = -\ln\psi(y,x) = -\phi(x,y)$. This example will also show that the function $t(x)$ in lemma 3.9 does not have to be unique.

Given 3.10, it is sufficient to only use the APF $\phi(x,y) = -\phi(y,x)$. A simple and explicit functions ϕ that satisfies theorem 3.8 under $h(x) = -x$ is given in the following corollary.

Corollary 3.10.1. *$\phi(x,y) = g(x) - g(y)$, where $g(x)$ is increasing and $\lim_{x\to\infty} g(x) = \infty$, satisfies definition 3.3 and satisfies the conditions in theorem 3.8.*

Proof. The function $\phi(x,y) = g(x) - g(y)$ is a increasing function of x and decreasing function of y as g is increasing. Because $\lim_{x\to\infty} g(x) = \infty$, the limit of $\phi(x,y)$ as either x or y goes to infinity, will be ∞ or $-\infty$ respectively. $\phi(x,y)$ will therefore be an APF. It also satisfies the involution $\phi(x,y) = h(\phi(y,x))$ with $h(x) = -x$. □

An example of a CD which can be created using the function in corollary 3.10.1 is the CVCD. It can be defind under the APF $\phi(x,y) = \ln x - \ln y$. The proof is tied to the one in section 3.7.1. As shown previously, $\phi(x,y) = \ln x - \ln y$ is equivalent to using $\psi(x,y) = \frac{x}{y}$. The resulting pivot equation is then

$$\frac{S_1}{S_2}\frac{1-\rho}{1+\rho} = \frac{U_1}{U_2}.$$

This is equivalent to (27) which is the source for the CVCD.

It would be useful if it is possible to expand theorem 3.10 to state if all ϕ symmetric over $h(x) = -x$ could be reduced to $\phi(x,y) = g(x) - g(y)$. No counter examples have been made, however it is not unlikely that one will exist. The theorem will in either case reduce the scope of the problem significantly. An interesting conjecture would be that theorem 3.8 is not only sufficient, but also necessary. If both conjectures are true, all CDs satisfying the symmetry conditions which originates in theorem 3.7, can be created using the simpler formulation in corollary 3.10.1.

A final question is whether all CDs for the correlation using S_1 and S_2 can be created using the pivotal function in theorem 3.7. Unfortunately there are only one other CD created for this problem using S_1 and S_2 without using theorem 3.7. That is the CVCD. The CVCD can be created using theorem 3.7 with APF $\phi(x,y) = x/y$. The resulting pivotal is equivalent to (27).

Although there is no proof of uniqueness, the function $\phi(x,y) = g(x) - g(y)$ opens up to great possibilities. The respective pivotal equation is

$$g\left(\frac{S_1}{2(1+\rho)}\right) - g\left(\frac{S_2}{2(1-\rho)}\right) = g(U_1) - g(U_2). \tag{32}$$

The issue is to express any model generating function or density function calculated using the inversion of (32). The density of ρ can be calculated by first calculating the distribution of $Z = g(U_1) - g(U_2)$ and conducting the transformation from Z to ρ using (32). The issue

is that calculating the density of Z can be challenging as one needs to integrate out one variable. An expression for the density of Z is available given by an intergral, as seen in corollary 3.10.2. The density of the corresponding confidence distribution is also expressed.

Corollary 3.10.2. *The density of $Z = g(U_1) - g(U_2)$, where U_1 and U_2 are chi-square distributed with n degrees of freedom, can be expressed as*

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\Gamma^2(n/2)} 2^{-n} \left(g^{-1}(|z|+t)\right)^{n/2-1} \left(g^{-1}(t)\right)^{n/2-1} \cdot e^{-\frac{1}{2}\left(g^{-1}(|z|+t)-g^{-1}(t)\right)} \left|g^{-1\prime}(|z|+t)g^{-1\prime}(t)\right| dt. \tag{33}$$

The density of ρ under the transformation

$$Z = g\left(\frac{s_1}{2(1+\rho)}\right) - g\left(\frac{s_2}{2(1-\rho)}\right)$$

is

$$f(\rho) = \int_{-\infty}^{\infty} \frac{1}{\Gamma^2(n/2)} 2^{-n} \left(g^{-1}\left(\left|g\left(\frac{s_1}{2(1+\rho)}\right) - g\left(\frac{s_2}{2(1-\rho)}\right)\right| + t\right)\right)^{n/2-1} \left(g^{-1}(t)\right)^{n/2-1} \exp\left(-\frac{1}{2}\left(g^{-1}\left(\left|g\left(\frac{s_1}{2(1+\rho)}\right) - g\left(\frac{s_2}{2(1-\rho)}\right)\right| + t\right) - g^{-1}(t)\right)\right) \left|g^{-1\prime}\left(\left|g\left(\frac{s_1}{2(1+\rho)}\right) - g\left(\frac{s_2}{2(1-\rho)}\right)\right| + t\right) g^{-1\prime}(t)\right| dt \left(g'\left(\frac{s_1}{2(1+\rho)}\right)\frac{s_1}{2(1+\rho)^2} + g'\left(\frac{s_2}{2(1-\rho)}\right)\frac{s_2}{2(1-\rho)^2}\right). \tag{34}$$

Proof. The density of Z is calculated using the formula which states that the density of the sum of two independent random variables is a convolution. Z can be expressed as the sum of X and Y where $X = g(U_1)$ and $Y = -g(U_2)$. As Z necessarily is symmetric as U_1 and U_2 are identically distributed, it is only necessary to do the convolution for positive z and state the distribution by replacing z with $|z|$. The distribution of $g(U_1)$, and $g(U_2)$, is

$$f(x) = \frac{1}{\Gamma(n/2)} 2^{-n/2} g^{-1}(x) e^{-\frac{1}{2}g^{-1}(x)} g^{-1\prime}(x).$$

The convolution is then as given in (33). The expression of (34) follows directly using the transformation from Z to ρ. □

As the corollary shows, the general expressions for both Z and ρ are not simple. Using them is therefore impractical. Similarly, calculating the model generating function can be challenging. In fact, even simple situations like $g(x) = x^2$ leads to a polynomial equation of fourth degree. These two issues hinders a more exact expression for the corresponding CDs for ρ. Making any analytical evaluation of the CD is therefore at best challenging. Hence, finding an optimal choice of g might not be fully possible. Even comparing two CDs can

be time consuming as all tests will require some use of numerical simulation. Secondly, the choice of the function g is in many ways arbitrary.

A choice of g that gives both a density function and model generating function is identity function $g(x) = x$. The pivotal equation (29) can be rewritten into

$$\frac{s_1}{2(1+\rho)} - \frac{s_2}{2(1-\rho)} = U_1 - U_2. \tag{35}$$

By using the formula in 3.10.2, the density of $Z = U_1 - U_2$ is given by

$$\int_0^\infty \frac{1}{\Gamma^2(n/2)} 2^{-n} \left(|z|+t\right)^{n/2-1} t^{n/2-1} e^{-\frac{1}{2}(|z|+t+t)} dt = \frac{1}{\Gamma^2(n/2)} 2^{-n} e^{-\frac{1}{2}|z|} \int_0^\infty (|z|+t)^{n/2-1} t^{n/2-1} e^{-t} dt.$$

By using solution 8. from 3.383 in (Gradshteyn and Ryzhik 2007), the solution to the intergal above is

$$f_Z(z) = \frac{2^{-n}}{\sqrt{\pi}\Gamma(n/2)} |z|^{n/2-1/2} K_{1/2-n/2}\left(\frac{1}{2}|z|\right),$$

where K is the modified Bessel function of the second kind. Transformed to ρ by using (35), the density function of the corresponding CD is

$$f(\rho) = \frac{2^{-n}}{\sqrt{\pi}\Gamma(n/2)} \left|\frac{s_1}{2(1+\rho)} - \frac{s_2}{2(1-\rho)}\right|^{n/2-1/2} K_{1/2-n/2}\left(\frac{1}{2}\left|\frac{s_1}{2(1+\rho)} - \frac{s_2}{2(1-\rho)}\right|\right) \cdot \left(\frac{s_1}{2(1+\rho)^2} + \frac{s_2}{2(1-\rho)^2}\right).$$

The model generating function can also be solved. First step is to rewrite (35) by multiplying with $2(1-\rho^2)$ and ordering the expressions into

$$2\rho^2(U_2-U_1)+\rho(s_1+s_2)+s_2-s_1+2(U_1-U_2) = 0 \iff \rho^2+\rho\frac{1}{2}\frac{s_1+s_2}{U_2-U_1}+\frac{1}{2}\frac{s_2-s_1}{U_2-U_1}-1 = 0,$$

as long as $U_1 \neq U_2$. For $U_1 = U_2$, the solution is given by

$$\rho(s_1+s_2)+s_2-s_1 = 0 \iff \rho = \frac{s_1-s_2}{s_1+s_2}.$$

For $U_1 \neq U_2$, there are two possible solutions

$$P = -\frac{1}{4}\frac{s_1+s_2}{U_2-U_1} \pm \sqrt{\left(\frac{1}{4}\frac{s_1+s_2}{U_2-U_1}\right)^2 + 1 - \frac{1}{2}\frac{s_2-s_1}{U_2-U_1}}. \tag{36}$$

Figure 11 shows the two solutions given in (36) as functions of $x = U_2 - U_1$. s_1 and s_2 are set to 3.4 and 2.6 respectively, however the same result can be seen for any value of s_1 and s_2. The choice of which solution to use is dependent on the sign of $U_2 - U_1$.

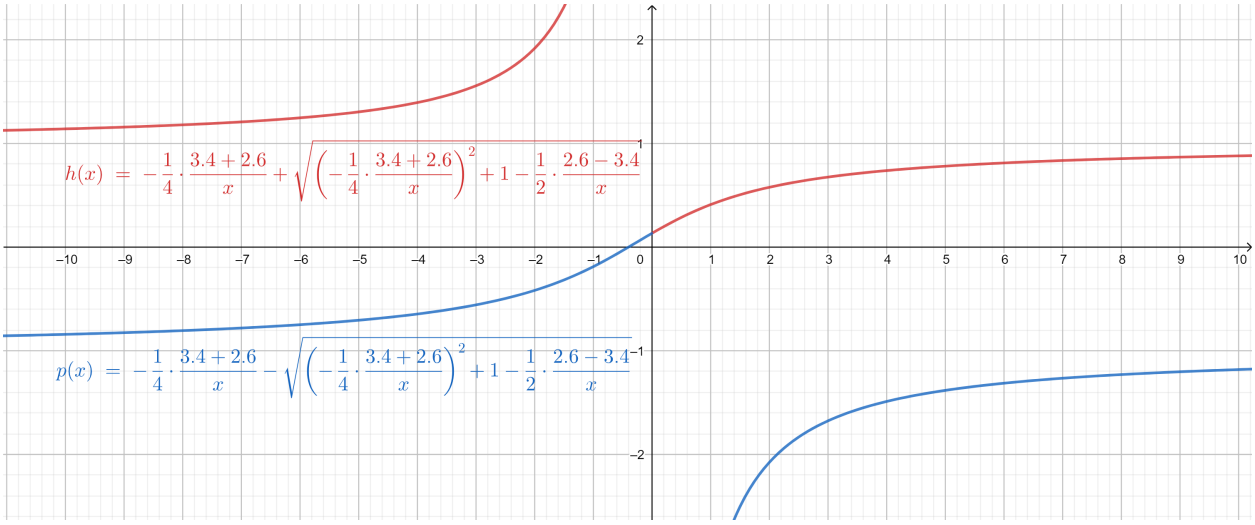


Figure 11: Visualization of the two solution from (36). Red graph represents solution using positive term and blue graph is for negative term. x is the value of $U_2 - U_1$. Observed data for curve is $s_1 = 3.4$ and $s_2 = 2.6$

Using that knowledge, the model generating function of ρ can be rewritten into

$$P = -\frac{1}{4}\frac{s_1+s_2}{U_2-U_1} + \operatorname{sgn}(U_2-U_1)\sqrt{\left(\frac{1}{4}\frac{s_1+s_2}{U_2-U_1}\right)^2 + 1 - \frac{1}{2}\frac{s_2-s_1}{U_2-U_1}}. \tag{37}$$

There are many other possible choices for g. One general comment about each of these possible CDs, is that the median is common for all of them. To calculate the median, the following lemma is necessary.

Lemma 3.11. *Let $X \sim F(x)$ be a continous random variable with median m_X, and $t(x)$ is a continuous monotone transformation. The median of $Y = t(X)$ is then*

$$m_Y = t(m_X).$$

The inverse expression

$$m_X = t^{-1}(m_Y)$$

also holds.

Proof. The relation of the Lemma was presented by Gunnar Taraldsen during a meeting. As I was unable to find any references, I constructed the following proof.

The median of X can be defined as

$$F_X(m) = \frac{1}{2} \iff x_{\text{med}} = F_X^{-1}\left(\frac{1}{2}\right).$$

The CDF of $Y = t(X)$ is given by the transformation

$$F_Y(y) = F_X(t^{-1}(y)).$$

As t is monotone, the inverse exists. The median of Y is then defined as

$$F_Y(m_Y) = F_X(t^{-1}(m_Y)) = \frac{1}{2}.$$

The median of Y can then be expressed as

$$t^{-1}(m_Y) = F_X^{-1}(\frac{1}{2}) = m_X.$$

Similarly, $m_Y = t(m_X)$. □

The common median can then be expressed in the following theorem.

Theorem 3.12. *Let $P = M([s_1, s_2], Z)$ be the model generating function defined in theorem 3.7 with APF defined in corollary 3.10.1. The median of P is then*

$$Median(P) = \frac{s_1 - s_2}{s_1 + s_2} = \frac{2\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i^2 + y_i^2)}.$$

Proof. The proof is based on Lemma 3.11. Assume the conditions of this theorem. Let $Z = g(U_1) - g(U_2)$. Then P is given by the transformation

$$Z(P) = g\left(\frac{s_1}{2(1+P)}\right) - g\left(\frac{s_2}{2(1-P)}\right).$$

As discussed earlier, this transformation is monotone. The median of Z is naturally 0 as $U_1 \sim U_2$. The median for P, m_P, will therefore satisfy the equation

$$g\left(\frac{s_1}{2(1+m_P)}\right) - g\left(\frac{s_2}{2(1-m_P)}\right) = 0.$$

The solution is for the point where

$$\frac{s_1}{2(1+m_P)} = \frac{s_2}{2(1-m_P)} \iff m_p = \frac{s_1 - s_2}{s_1 + s_2}.$$

□

The median expressed in Theorem 3.12 is the empirical correlation when the variances are assumed to be equal but unknown. Other characteristics of the CDs such as the mean and mode are not generally shared. The fact that the median is shared across all of these confidence distributions could imply that it would suit as a good estimator for the correlation. This has not been focused on, but similarly to how Bayesian estimators are calculated, estimators based on CDs can be created. This estimator would then be such an estimator under all the CDs described by corollary 3.10.1. There have not been any attempts at testing the estimator.

Expressing the most optimal confidence distribution created using these methods, are challenging as most CDs are expressed implicitly. One can set up an optimality function given the loss function $L(\rho_0, \rho)$, where ρ_0 is the true correlation. Given the density function (34), the risk of the CD is

$$\begin{aligned} R(\rho_0) = \int_0^\infty \int_0^\infty \int_{-1}^1 L(\rho_0, \rho) \int_{-\infty}^\infty \frac{1}{\Gamma^2(n/2)} 2^{-n} \left(g^{-1}\left(\left| g\left(\frac{s_1}{2(1+\rho)} \right) - g\left(\frac{s_2}{2(1-\rho)} \right) \right| + t \right) \right)^{n/2-1} \\ \cdot \left(g^{-1}(t) \right)^{n/2-1} \exp\left(-\frac{1}{2} \left(g^{-1}\left(\left| g\left(\frac{s_1}{2(1+\rho)} \right) - g\left(\frac{s_2}{2(1-\rho)} \right) \right| + t \right) - g^{-1}(t) \right) \right) dt \\ \cdot \left(g'\left(\frac{s_1}{2(1+\rho)} \right) \frac{s_1}{2(1+\rho)^2} + g'\left(\frac{s_2}{2(1-\rho)} \right) \frac{s_2}{2(1-\rho)^2} \right) d\rho \\ \cdot \frac{1}{\Gamma^2(n/2)} \left(4(1-\rho^2) \right)^{-n/2} s_1^{n/2-1} s_2^{n/2-1} \exp\left(-\frac{s_1}{2(1+\rho)} - \frac{s_2}{2(1-\rho)} \right) ds_1 ds_2. \end{aligned}$$

Optimizing the risk wrt. the function g is far from trivial as the risk is expressed using the inverse of g, the derivative of the inverse as well as four integrals. No attempts to solve that problem will be attempted.

An alternative to creating pivots using the minimal sufficient is to use the original data directly. For each data point there is the data generating function

$$X_i = Z_{1,i}, \; Y_i = \rho Z_{1,i} + \sqrt{1-\rho^2} Z_{2,i}, \text{ for } i = 1, ..., n$$

where $Z_{1,i} \sim Z_{2,i} \sim N(0,1)$ (Taraldsen 2020). This holds for all data points and can be reduced to n pivots

$$\frac{Y_i - \rho X_i}{\sqrt{1-\rho^2}} = Z_{2,i}.$$

A similar approach as the APF can be applied to these pivots as well to create the collective pivot

$$\psi\left(\frac{Y_1 - \rho X_1}{\sqrt{1-\rho^2}}, ..., \frac{Y_n - \rho X_n}{\sqrt{1-\rho^2}} \right) = \psi(Z_1, ..., Z_n),$$

where $Z_i \sim N(0,1)$ are independent. As the dimension of the domain of ψ is significantly larger than ϕ, the complexity is greater. Additionally, the shape of the separate pivots is dependent on the values of X_i and Y_i, unlike the case for S_1 and S_2. Designing ψ in order to always give a unique solution, is therefore not a simple task if even possible. Even satisfying the symmetric conditions in definition 3.1 might not be possible as X_i and Y_i are not treated similarly.

A solution is to use the transformation of the data in corollary 3.0.1 instead. On that form, the following pivots are true

$$\frac{X_i + Y_i}{\sqrt{2(1+\rho)}} = Z_{1,i} \sim N(0,1), \; \frac{X_i - Y_i}{\sqrt{2(1-\rho)}} = Z_{2,i} \sim N(0,1).$$

An issue with these pivots is that they can have both positive and negative signs. This implies that for some sets X_i, Y_i, $Z_{1,i}$ and $Z_{2,i}$, a solution does not exist. A possible step is to apply the absolute value function to both pivots. That is

$$\left|\frac{X_i + Y_i}{\sqrt{2(1+\rho)}}\right| = |Z_{1,i}| \sim N(0,1), \quad \left|\frac{X_i - Y_i}{\sqrt{2(1-\rho)}}\right| = |Z_{2,i}| \sim N(0,1).$$

This case is quite similar to the one with the sufficient statistics. In fact, if the pivots are squared, the sum of each type will result in the pivots of the sufficient statistics. This fact will be expressed further. This is not surprising as these steps are the ones used to create the statistics in the first place. It is possible to approach this problem similarly to the one for sufficient statistics. Instead of the function $\phi(x, y)$, this problem needs a more general function $\psi(u_1, v_1, ..., u_n, v_n)$ which takes in all data points (x_i, y_i). The resulting pivot can be seen in the next equation.

$$\psi\left(\left|\frac{X_1 + Y_1}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_1 - Y_1}{\sqrt{2(1-\rho)}}\right|, ..., \left|\frac{X_n + Y_n}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_n - Y_n}{\sqrt{2(1-\rho)}}\right|\right) = \psi(|Z_{1,1}|, |Z_{2,1}|, ..., |Z_{n,1}|, |Z_{n,2}|). \tag{38}$$

As for $\phi(x, y)$, it is necessary to define which functions for ψ that can be used to create confidence distributions.

Definition 3.4. A function $\psi : \mathbb{R}^{+2n} \mapsto I$, where $I = (a, b) \subseteq \mathbb{R}$, is an admissible pivot function or APF if it satisfies the following properties:

1. ψ increases wrt. all u_i
2. ψ decreases wrt. all v_i
3. $\lim_{u_i \to \infty} \psi(u_1, v_1..., u_i, ..., u_n, v_n) = b$ independent of all v_j for at least one i
4. $\lim_{v_i \to \infty} \psi(y_1, v_1..., v_i, ..., u_n, v_n) = a$ independent of all u_j for at least one i.

Theorem 3.13. *Let $x_1, y_1, ..., x_n, y_n$ be n realizations of the distribution (5) and $Z_{1,i}$,$Z_{2,i}$,...,$Z_{2n-1,i}$,$Z_{2n,i}$ are identically and independently standard normally distributed. If $\psi(u_1, v_1, ..., u_n, v_n)$ satisfies definition 3.4 then the inversion of*

$$\psi\left(\left|\frac{X_1 + Y_1}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_1 - Y_1}{\sqrt{2(1-\rho)}}\right|, ..., \left|\frac{X_n + Y_n}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_n - Y_n}{\sqrt{2(1-\rho)}}\right|\right) = \psi(Z_{1,1}, Z_{2,1}, ..., Z_{n,1}, Z_{n,2})$$

wrt. ρ is a model generating function for a confidence distribution of ρ.

Proof. The arguments of this proof is essentially the same as for theorem 3.7. The main difference is that mapping onto I wrt. ρ will now depend on multiple variables. As ρ either converges to 1 or -1, either all the u_i or all the v_i will diverge towards infinity. It is therefore necessary that $\psi(u_1, v_1, ..., u_n, v_n)$ converges to the boundaries of I for at least one u_i or one v_j. □

The definition of an APF in 3.4 is more general than necessary in the definition of the limits. As either all u_i or all v_i will diverge to infinity simultaneously, at different speeds, it is possible to assert the condition that ψ converges to b when all u_i diverges to infinity and likewise for all v_i and the limit a. This is also a reasonable criteria considering the next point that will be brought up.

There is one type of symmetry that is not mentioned in definition 3.1 which often applies to all estimators. As all the data points (x_i, y_i) are assumed to be sampled independently, the ordering of the points should not matter. Similarly to corollary 3.7.1 and theorem 3.8, a theorem for satisfying the symmetry conditions in definition 3.1 as well as the new condition can be assembled.

Corollary 3.13.1. *The symmetry conditions in definition 3.2 are satisfied for the model generating function defined by the inversion of equation (38) if and only if*

1. $\psi(U_1, V_1, ..., U_n, V_n) \sim \psi(V_1, U_1, ..., V_n, U_n)$
2. $\psi(U_1, V_1, ..., U_i, V_i, ..., U_j, V_j, ..., U_n, V_n) \sim \psi(U_1, V_1, ..., U_j, V_j, ..., U_i, V_i, ..., U_n, V_n)$

for all $i, j = 1, ..., n$, where

$$U_i = \frac{|x_i + y_i|}{\sqrt{2(1+P)}}, \; V_i = \frac{|x_i - y_i|}{\sqrt{2(1-P)}}$$

and P is given by the model generating function and inversion of (38) wrt. ρ with data $X_i = x_i$ and $Y_i = y_i$ for $i = 1, ..., n$.

Proof. The proof follows the same line of arguments as for corollary 3.7.1 in terms of the symmetry conditions in definition 3.1. The first criteria of definition 3.1 where the correlation does not change, will be satisfied automatically using the pivots above as the pivots will not change if x_i and y_i are swapped or both change signs simultaneously. If either all x_i or all y_i changes sign, the two separate pivots for each data points (x_i, y_i) will be swapped. This is the same argument used for symmetry conditions of the sufficient statistics S_1 and S_2. As a result, the same condition as in corollary 3.13.1, will have to be true for each pair of pivots. Collectively, this results in condition 1.

This corollary does also have an additional condition. The order of the data points (x_i, y_i) should not alter any estimators if they are independent. In terms of a model generating function that is, if $P = M([x_1, y_1, ..., x_i, y_i, ..., x_j, y_j, ..., x_n, y_n,], Z)$ is a model generating function and the inversion of (38) wrt. ρ at data $X_i = x_i$ and $Y_i = y_i$, $i = 1, ..., n$, then

$$M([x_1, y_1, ..., x_i, y_i, ..., x_j, y_j, ..., x_n, y_n,], Z) \sim M([x_1, y_1, ..., x_j, y_j, ..., x_i, y_i, ..., x_n, y_n,], Z).$$

Under these conditions, condition nr. 2 will necessarily and sufficiently hold. □

Theorem 3.14. *Let ψ satisfy 3.4. The model generating function defined as the inversion of (38) wrt. ρ satisfies the symmetry conditions in definition 3.1 if $\psi(u_1, v_1, ..., u_n, v_n) = h(\psi(v_1, u_1, ..., v_n, u_n))$ and $\psi(u_1, v_1, ..., u_i, v_i, ..., u_j, v_j, ..., u_n, v_n) = \psi(u_1, v_1, ..., u_j, v_j, ..., u_i, v_i, ..., u_n, v_n)$ for any $i, j = 1, ..., n$. h is a decreasing involution, that is h is its own inversion.*

Proof. The proof follows the same lines of arguments as in theorem 3.8. The main difference is that the value of ψ is independent on the ordering of the data points (x_i, y_i). It is self evident that if $\psi(u_1, v_1, ..., u_i, v_i, ..., u_j, v_j, ..., u_n, v_n) = \psi(u_1, v_1, ..., u_j, v_j, ..., u_i, v_i, ..., u_n, v_n)$ then they share model generating function. □

A generalization of the reordering of data points can be expressed using the increasing involutions $h_{i,j}$. $h_{i,j}$ will in that case be such that

$$\psi(..., u_i, v_i, ..., u_j, v_j, ...) = h_{i,j}(\psi(..., u_j, v_j, ..., u_n, v_i, ...)).$$

However, it is not unlikely that it is necessary that all $h_{i,j}$ are the identity function. This assumption was used when formulating theorem 3.14. In either case, the theorem will hold.

From theorem 3.14 there are multiple approaches to formulate pivots. As mentioned earlier, all CDs created using the sufficient statistics can be expressed as

$$\begin{aligned}\psi\left(\left|\frac{X_1+Y_1}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_1-Y_1}{\sqrt{2(1-\rho)}}\right|, ..., \left|\frac{X_n+Y_n}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_n-Y_n}{\sqrt{2(1-\rho)}}\right|\right) \\ = \phi\left(\sum_{i=1}^{n}\left|\frac{X_i+Y_i}{\sqrt{2(1+\rho)}}\right|^2, \sum_{i=1}^{n}\left|\frac{X_i-Y_i}{\sqrt{2(1-\rho)}}\right|^2\right).\end{aligned}$$

A more general approach is

$$\begin{aligned}&\psi\left(\left|\frac{X_1+Y_1}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_1-Y_1}{\sqrt{2(1-\rho)}}\right|, ..., \left|\frac{X_n+Y_n}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_n-Y_n}{\sqrt{2(1-\rho)}}\right|\right) \\ = f\left(\lambda\left(\left|\frac{X_1+Y_1}{\sqrt{2(1+\rho)}}\right|, ..., \left|\frac{X_n+Y_n}{\sqrt{2(1+\rho)}}\right|\right), \lambda\left(\left|\frac{X_1-Y_1}{\sqrt{2(1+\rho)}}\right|, ..., \left|\frac{X_n-Y_n}{\sqrt{2(1+\rho)}}\right|\right)\right),\end{aligned}$$

where f is an APF defined in 3.3. λ is necessarily increasing wrt. all of its variables. It is also necessary that λ keeps the property of the limits given in definition 3.4. Another possibility is to combine each data point (x_i, y_i) into separate pivots and apply a function to each pivot. That is

$$\begin{aligned}&\psi\left(\left|\frac{X_1+Y_1}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_1-Y_1}{\sqrt{2(1-\rho)}}\right|, ..., \left|\frac{X_n+Y_n}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_n-Y_n}{\sqrt{2(1-\rho)}}\right|\right) \\ = \lambda\left(f\left(\left|\frac{X_1+Y_1}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_1-Y_1}{\sqrt{2(1-\rho)}}\right|\right), ..., f\left(\left|\frac{X_n+Y_n}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_n-Y_n}{\sqrt{2(1-\rho)}}\right|\right)\right),\end{aligned}$$

where f and λ is defined as mentioned above. The two latter approaches are ways of combining the pivots. The first separates the data points into two types of pivots which are combined separately across data points. The other approach combines the pivots of each data point first. Each approach will possibly result in two different confidence distributions.

The involution conditions set in theorem 3.14 should be satisfied for both variations of ψ. For the first of the two cases, it is sufficient that f satisfies the involution condition in theorem 3.8 similarly to $\phi(x, y)$. In fact, as stated in theorem 3.10, it is possible to only study cases for f such that $f(x, y) = -f(y, x)$. For the latter of the two, it is not as simple. The condition of λ and f has to be such that $h(\lambda(f(u_1, v_1), ..., f(u_n, v_n))) = \lambda(f(v_1, u_1), ..., f(v_n, u_n))$. Similarly to the other scenario, it is sufficient to study λ and f such that $\lambda(f(u_1, v_1), ..., f(u_n, v_n)) = -\lambda(f(v_1, u_1), ..., f(v_n, u_n))$.

A final approach is based on combining confidence distributions (Singh, Xie, and Strawderman 2005). Each data point can be used to create separate CDs. Because of the similarity between the pivots of the original data (x_i, y_i) and their respective sufficient statistics s_1 and s_2, theorem 3.8 can be applied to

$$\phi\left(\left|\frac{X_i + Y_i}{\sqrt{2(1+\rho)}}\right|, \left|\frac{X_i - Y_i}{\sqrt{2(1-\rho)}}\right|\right) = \phi(|Z_{1,i}|, |Z_{2,i}|).$$

This will result in n confidence distributions. All of these CDs can then be combined using the methods described in (Singh, Xie, and Strawderman 2005) to create a collective CD for all the data. No attempts at the method will be conducted.

It is possible to create multiple distributions using the approaches above. For instant, one can take inspiration from CVCD by creating pivots for each data point using $f(u, v) = u^2/v^2$ and $\psi(u_1, v_1, ..., u_n, v_n) = \lambda(f(u_1, v_1), ..., f(u_n, v_n))$. The separate pivots for each data point will be equivalent with the pivot used to create CVCD using just one data point. Two possible choices for λ is in $\lambda(\xi_1, ..., \xi_n) = \sum_i \xi_i$ or $\lambda(\xi_1, ..., \xi_n) = \prod_i \xi_i$. It is difficult to tell if the composition of λ and f will satisfy the symmetry across any involution for

$$\lambda(\xi_1, ..., \xi_n) = \sum_i \xi_i$$

. For

$$\lambda(\xi_1, ..., \xi_n) = \prod_i \xi_i$$

the involution is $h(x) = 1/x$. If the product is chosen the complete pivot is

$$\prod_{i=1}^{n} \frac{(X_i + Y_i)^2}{(X_i - Y_i)^2} \left(\frac{1-\rho}{1+\rho}\right)^n = \prod_{i=1}^{n} \frac{Z_{1,i}^2}{Z_{2,i}^2}.$$

If the logarithm is applied, then it is apparent that the resulting CD will be symmetric wrt. the reparametrization $z(\rho) = \text{arctanh}(\rho)$, similar to CVCD. Calculating the density function of the resulting CD is challenging. The model generating function is however easily available. Solving the equation above gives that

$$P = \frac{1-\xi}{1+\xi},$$

where

$$\xi = \prod_{i=1}^{n} \left(\frac{Z_{1,i}}{Z_{2,i}} \frac{x_i - y_i}{x_i + y_i} \right)^{\frac{2}{n}}.$$

As the density is not available, it is not as simple to state whether it is unimodal or not. However, it seems to be under simulations, see section 3.7.6.

If λ is chosen to be the sum instead, the pivot is

$$\sum_{i=1}^{n} \frac{(X_i + Y_i)^2}{(X_i - Y_i)^2} \frac{1-\rho}{1+\rho} = \sum_{i=1}^{n} \frac{Z_{1,i}^2}{Z_{2,i}^2}.$$

The model generating function for the resulting CD is then

$$P = \frac{1-\epsilon}{1+\epsilon}, \tag{39}$$

where

$$\epsilon = \frac{\sum_{i=1}^{n} \frac{Z_{1,i}^2}{Z_{2,i}^2}}{\sum_{i=1}^{n} \frac{(x_i+y_i)^2}{(x_i-y_i)^2}}.$$

For the second model, it is necessary to test if the symmetry conditions in definition 3.1 are satisfied. Figure 12 shows the density of the model generating function P under data (x, y) and the model generating function $-P$ under $(-x, y)$. As the figure shows, the two are not similarly distributed and the symmetry is not satisfied.

Both of these model generating functions are on a similar form to the one for CVCD although the distributions should be significantly different. Both of the pivots used to create the CDs will also work in the setting of a common unknown variance just like the CVCD.

3.7.3 Method of regions

There exists an alternative perspective to the method described in the previous subchapter. Instead of inverting a pivot, it is possible to invert a data generating function instead. Take the data generating function for the sufficient statistics S_1 and S_2, defined by (6),

$$S_1 = 2(1+\rho)U_1, \;\; S_2 = 2(1-\rho)U_2,$$

where U_1 and U_2 are independent chi-square distributed variables with n degrees of freedom. Inversion of the data generating function does not exist for all sets S_1, S_2, U_1 and U_2. A solution is to define a set $A \subseteq \Omega_U$ of $U = (U_1, U_2)$ and create the set of all solutions to the data generating function wrt. ρ and $U \in A$. The set of solutions is then defined as

$$B = \{\rho \in (-1, 1) | S_1 = 2(1+\rho)U_1, \;\; S_2 = 2(1-\rho)U_2, (U_1, U_2) \in A\}.$$

If A is chosen carefully, B can be an interval estimator for the correlation, and possibly also a one-sided interval estimator. If the frequentistic coverage of B is sufficiently large given

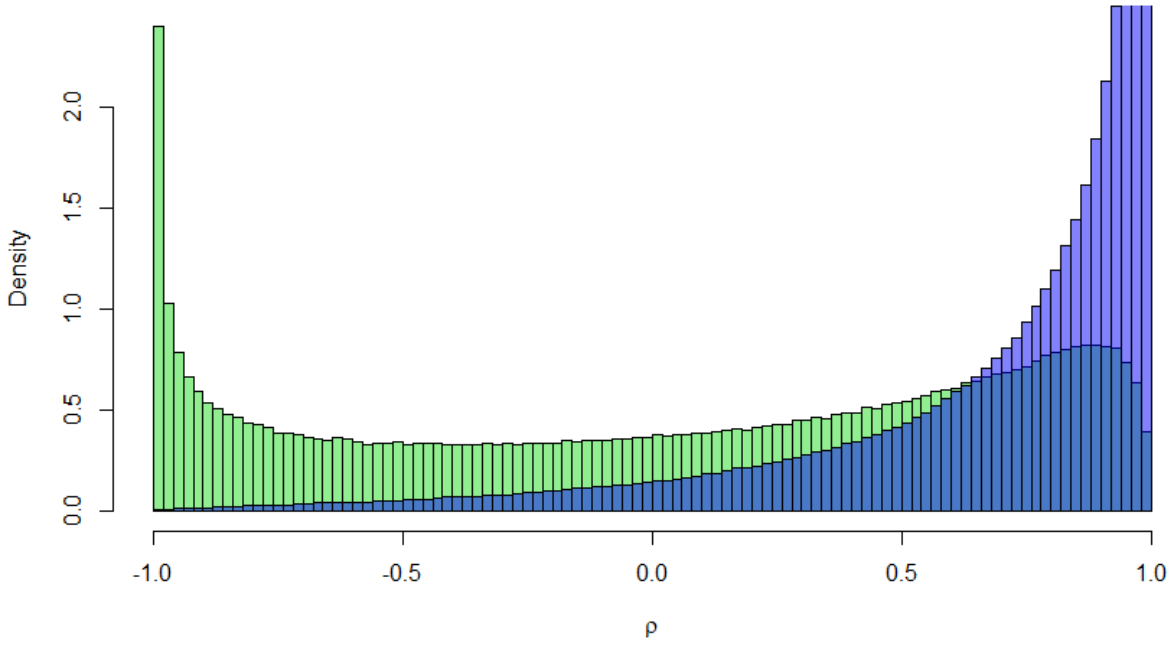


Figure 12: Histogram of the model generating function given in (39). The green histogram is P under data (x, y) and the blue histogram is $-P$ under data $(-x, y)$.

the chosen coverage of A, then B could be a confidence interval estimator. Before studying the coverage, the relation between B and A needs to be studied. Given $S_1 = s_1$ and $S_2 = s_2$, the two solutions for ρ of the two separate equations of the data generating functions are

$$\rho = \frac{s_1}{2U_1} - 1, \ \rho = 1 - \frac{s_2}{2U_2}.$$

The only sets (U_1, U_2) that unites the two solutions are given by the relation

$$U_2 = g(U_1) = \frac{s_2}{4 - \frac{s_1}{U_1}}.$$

The function g is decreasing with respect to U_1 such that $U_1 \geq s_1/4$ and $U_2 \geq s_2/4$. As $U_1 \to \infty$, $\rho \to -1$ and as $U_1 \to S_1/4$, $\rho \to 1$. This will also imply that $U_1 < s_1/4$ or $U_2 < s_2/4$ does not allow for any solution for ρ. The relation between the sets along $U_2 = g(U_1)$ and the corresponding correlation ρ is visualized in figure 13.

If B should be a lower one-sided interval, then A must cover the "point" $(U_1, U_2) = (\infty, S_2/4)$. By defining A as

$$A = \{(U_1, U_2) | U_2 \leq f(U_1; a)\},$$

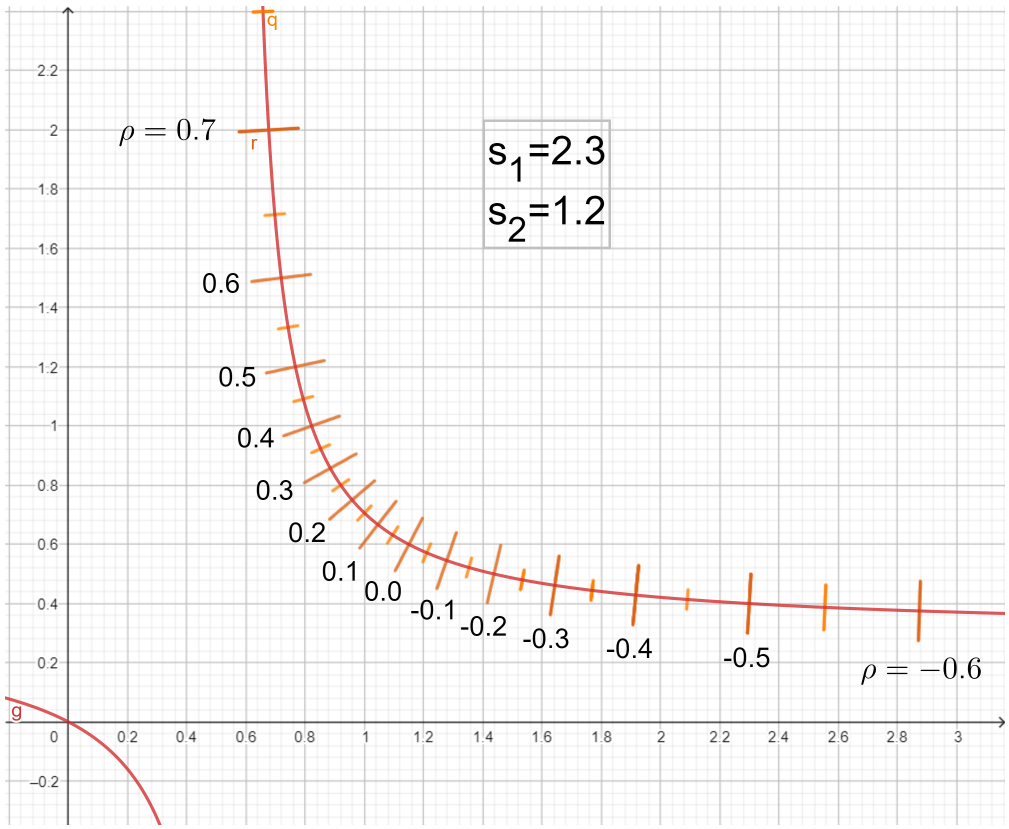


Figure 13: Figure of the line $U_2 = g(U_1)$, where the dashes represents corresponding solutions of the correlation ρ.

where a is a scalar and $f(x; a)$ is an increasing function that diverges to ∞ with x for all permissible a, B will always be a lower one-sided interval. As $g(x)$ is decreasing, f and g will have a unique intersection. This intersection will correspond to some correlation $\hat{\rho}$. Any correlation $\rho \leq \hat{\rho}$ will be covered by B as A will cover all sets along $U_2 = g(U_1)$ with U_1 larger than in the intersection. In other words A will always create lower one-sided intervals. Figure 14 shows the function $f(x; a) = ax$, the set A and the line $U_2 = g(U_1)$.

The next step is to study the coverage of B as a function of the coverage of A. Firstly, it is necessary that A can have any coverage between 0 to 1 by varying a. Secondly, the coverage of A should be increasing with a. If $f(x; a)$ is increasing wrt. a for all x, then if $a_1 < a_2$, $A_1 = \{U|U_2 \leq f(U_1; a_1)\} \subset \{U|U_2 \leq f(U_1; a_2)\} = A_2$. This results in $P(U \in A_1) < P(U \in A_2)$ i.e. the coverage of A increases with a. If the permissible set for a is $(a_{\min}, a_{\max})$, then it is reasonable to define $f(x; a)$ such that it divirges to infinity for all x if $a \downarrow a_{\max}$ and $f(x; a)$ converges to 0 for all x if $a \uparrow a_{\min}$. This will imply that if $a \downarrow a_{\min}$, $A \downarrow \emptyset$ and if $a \uparrow a_{\max}$, $A \uparrow \Omega_U$. The coverage of A can vary therefore vary with a from 0 to 1.

Let all the above criteria be satisfied for $f(x; a)$. To summarize

1. $f(x; a)$ is increasing wrt. x for all a

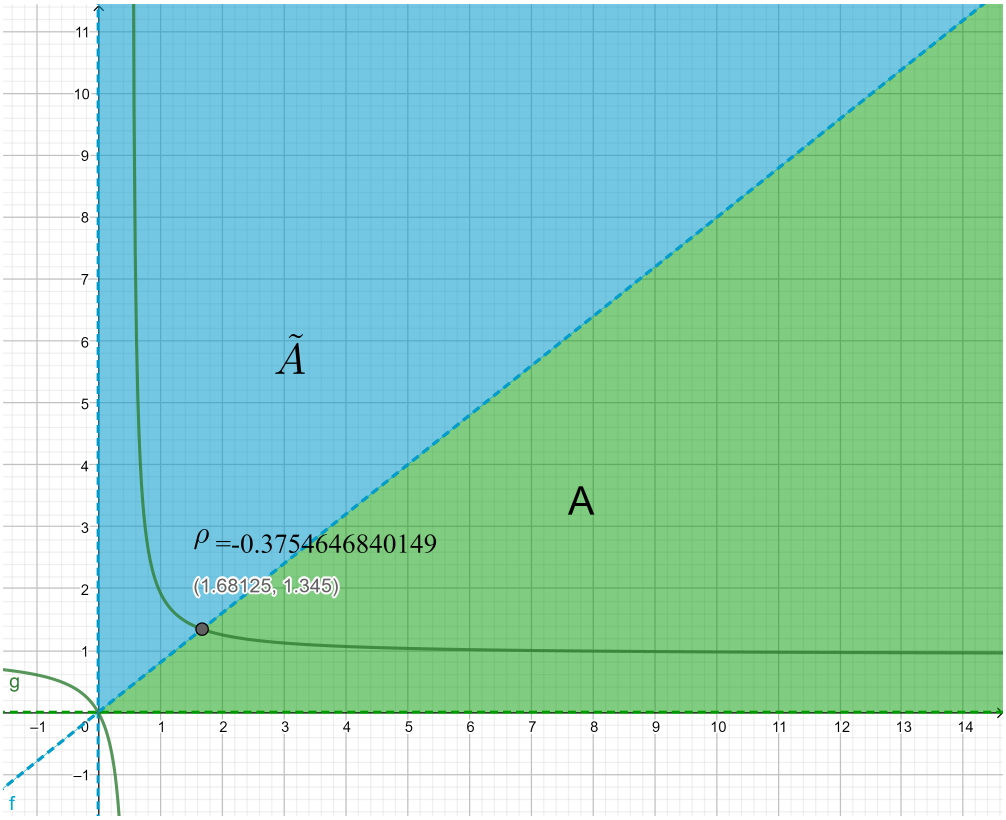


Figure 14: Figure showing different aspects of the method of regions. x-axis is U_1 and y-axis is U_2, however they are interchangable. Green line is $U_2 = g(U_1)$, blue dotted line is $U_2 = aU_1$, green field is the set A and blue field is the complementary set $\tilde{A}$. $S_1 = 2.1$ and $S_2 = 3.7$.

2. $f(x; a)$ is increasing wrt. a for all x
3. $\lim_{x\to\infty} f(x; a) = \infty$ for all a
4. $\lim_{a\uparrow a_{\max}} f(x; a) = \infty$ for all x
5. $\lim_{a\downarrow a_{\min}} f(x; a) = 0$ for all x.

If B covers ρ_0, then the point $(U_1, U_2) = (S_1/(2(1+\rho_0)), S_2/(2(1-\rho_0)))$ satisfies

$$\frac{S_2}{2(1-\rho_0)} \leq f\left(\frac{S_1}{2(1+\rho_0)}; a\right).$$

This means that the probability of B covering ρ_0 is given by

$$P\left(\frac{S_2}{2(1-\rho_0)} \leq f\left(\frac{S_1}{2(1+\rho_0)}; a\right)\right).$$

If ρ_0 is the true parameter then $\frac{S_2}{2(1-\rho_0)} \sim \frac{S_1}{2(1+\rho_0)} \sim U_1 \sim U_2$. As a result

$$P\left(\frac{S_2}{2(1-\rho_0)} \leq f\left(\frac{S_1}{2(1+\rho_0)}; a\right)\right) = P(U_2 \leq f(U_1; a)) = P(U \in A).$$

This means that the frequentistic coverage of B equals the coverage of A. By changing $a \in (a_{\min}, a_{\max})$, one-sided lower confidence interval for the correlation of any level can be created. Potentially, each function f satisfying the criteria can result in different confidence interval estimators.

It is possible to rewrite the confidence intervals into a confidence distribution. In order to do that, an expression for the upper bound of the confidence intervals is needed. This is ρ_0 such that $f(S_1/(2(1+\rho_0)); a) = g(S_1/(2(1+\rho_0)))$ or

$$f\left(\frac{S_1}{2(1+\rho_0)}; a\right) = \frac{S_2}{2(1-\rho_0)}.$$

This can be rewritten into

$$a = \phi\left(\frac{S_1}{2(1+\rho_0)}, \frac{S_2}{2(1-\rho_0)}\right),$$

using the inversion $a = \phi(x, \cdot)$ of $f(x; a)$ with respect to a. ϕ will always exist as f is monotonic with respect to a. Some intrinsic properties of $\phi(x, y)$ is that it is decreasing for x and increasing with y. This is due to y increasing with a and if x increases for constant y, then a has to be decreasing. Additionally, $\phi(x, y)$ will need to cover all values of a in a reasonable manner. $\phi(x, y)$ should converge to $a_{\max}$ if y divirges to infinity for all values for x. Similarily, $\phi(x, y)$ should converge to $a_{\min}$ if y is finite as $x \to \infty$. In terms of the inversion the limit $x \to \infty$ implies that $\phi(x, y) \to a_{\min}$ and $y \to \infty$ implies that $\phi(x, y) \to a_{\max}$. These conditions are equivalent with the definition of an APF in definition 3.3, however with the properties of x and y switched.

Let a be defined such that the coverage of A is α. It is then necessary that $P(U_2 \leq f(U_1; a)) = P(a \leq \phi(U_1, U_2)) = \alpha$. This means that the distribution of a is given by $a = \phi(U_1, U_2)$, where $U_1 \sim U_2 \sim \chi^2_n$. Inserting this into the equation above, we get that the inversion of

$$\phi\left(\frac{S_1}{2(1+\rho_0)}, \frac{S_2}{2(1-\rho_0)}\right) = \phi(U_1, U_2)$$

wrt. ρ defines the model generating function of a confidence distribution. This is the same definition of a confidence distribution as given in theorem 3.7. Using regions will therefore give an alternative argument for the framework presented in definition 3.3 and theorem 3.7. It could be interpreted such that the relation $U_2 = f(U_1; a)$ defines level curves in the space of (U_1, U_2) with level $\phi(U_1, U_2)$. If no solution of the data generating function exists for the pair U_1 and U_2, two alternative values along the level curve of (U_1, U_2), named (V_1, V_2) are chosen as a reasonable alternative. The only V_1 and V_2 along the level curve which allow an inversion of the data generating function wrt. ρ, are the two pivots

$$V_1 = \frac{S_1}{2(1+\rho)}, \; V_2 = \frac{S_2}{2(1-\rho)}.$$

The matching of the two methods, shows a bridge between inverting the data generating function and the pivotal function. In this case, they have a one-to-one relation, which could be the explanation.

In terms of the symmetry conditions in definition 3.2, they are defined slightly different for this method. The idea is that a confidence distribution created by using regions, should not be dependent on how A is defined with respect to the order of U_1 and U_2. Previously A was defined as the set $A = \{(U_1, U_2) | U_2 \leq f(U_1; a)\}$. A just as valid definition would be $\tilde{A} = \{(U_1, U_2) | U_1 \leq f(U_2; b)\}$. An important difference is that the corresponding interval $\tilde{B}$ will then be upper one-sided confidence intervals on the form $(\rho_0, 1)$. If these two definitions of A should create the same distributions, then A and $\tilde{A}$ should be complementary in the space of U. In that case, the lower and upper one-sided intervals for the correlation will also be complementary for all $S_1 = s_1$ and $S_2 = s_2$. The two distributions should therefore align.

A and $\tilde{A}$ are complementary if the boundaries of the two sets match. That is if $U_1 = f(U_2; a) = f(f(U_1; b); a)$. An example of the two is shown in figure 14. For this to be possible $f(\,\cdot\,; a)$ is the inverse of $f(\,\cdot\,; b)$ and vice versa. This will lead to some intrinsic properties of ϕ. First of all, if it is possible to define the relation between U_1 and U_2 as both $U_1 = f(U_2; a)$ and $U_2 = f(U_1; b)$, then the following should be true for the inversion

$$\phi(U_1, U_2) = a \iff \phi(U_2, U_1) = b.$$

Secondly, there needs to be a relation between a and b such that $f(f(\,\cdot\,; b); a)$ is the identity function. Using the inversion wrt. a,

$$\phi(f(U_1; b), U_1) = a, \;\; \phi(f(U_2; a), U_2) = b.$$

In general, $a = \phi(f(\,\cdot\,; b), \,\cdot\,) = h(b)$. Additionally, this relation should go both directions, that is $b = h(a)$, such that $a = h(b) = h(h(a))$. a and b are therefore related across an involution. h is decreasing as $\phi(x, y)$ is decreasing wrt. x and $f(\,\cdot\,; b)$ is increasing wrt. b. If applied further, then

$$\phi(U_1, U_2) = a = h(b) = h(\phi(U_2, U_1)).$$

It is therefore necessary that $\phi(x, y)$ is symmetric along a decreasing involution h, as was given previously in theorem 3.8. Without making the assertion, it seems as if this condition is not only sufficient but also necessary.

A final, and important point is to make sure that the symmetry of this method aligns with the symmetry defined in definition 3.2. One would assume so, as the conditions for the pivotal equation (29) are the same as in section 3.7.2, however a formal bridge should be made. Assume the two scenarios, $S_1 = s_1$, $S_2 = s_2$ and $S_1 = s_2$, $S_2 = s_1$. If $S_1 = s_1$ and $S_2 = s_2$, the set of all (U_1, U_2) such that the data generating function is invertible are along the graph

$$U_2 = g(U_1) = \frac{s_2}{4 - \frac{s_1}{U_1}}.$$

The inverse is

$$U_1 = \tilde{g}(U_2) = \frac{s_1}{4 - \frac{s_2}{U_2}}.$$

If $S_1 = s_2$ and $S_2 = s_1$, then the set of all (U_1, U_2), such that the data generating function is invertible, are along the graph

$$U_1 = g(U_2) = \frac{s_2}{4 - \frac{s_1}{U_2}}.$$

It is apparent that swapping S_1 and S_2 is equivalent to reflecting the graph $U_2 = g(U_1)$ across the line $U_1 = U_2$. One could therefore study the case of $S_1 = s_2$ and $S_2 = s_1$ by swapping U_1 and U_2 and use the data $S_1 = s_1$ and $S_2 = s_2$. There is however a small difference, the correlation decreases with U_1 and increases with U_2. In fact, the correlation can be expressed as

$$\rho = \frac{s_1}{2U_1} - 1 = 1 - \frac{s_2}{2U_2}.$$

If U_1 and U_2 are swapped, then the correlation under the same data is

$$\tilde{\rho} = \frac{s_1}{2U_2} - 1 = 1 - \frac{s_2}{2U_1}.$$

As $U_1 \sim U_2$, $\rho \sim -\tilde{\rho}$. Swapping S_1 and S_2 would therefore result in the same model generating function, however with opposite sign. Figure 3.7.3 illustrates the two scenarios. As seen, under data $S_1 = 2.3$ and $S_2 = 1.2$, the lower confidence distributions for $a = 2.3$ using $f(x; a) = ax$ is $(-1, 0.63)$. The lower confidence distributions for data $S_1 = 1.2$, $S_2 = 2.3$, $b = h(a) = 1/a$ using $f(x; b) = bx$, is $(-1, -0.63)$. The symmetry conditions of the method of regions are therefore equivalent to the symmetry conditions in definition 3.2.

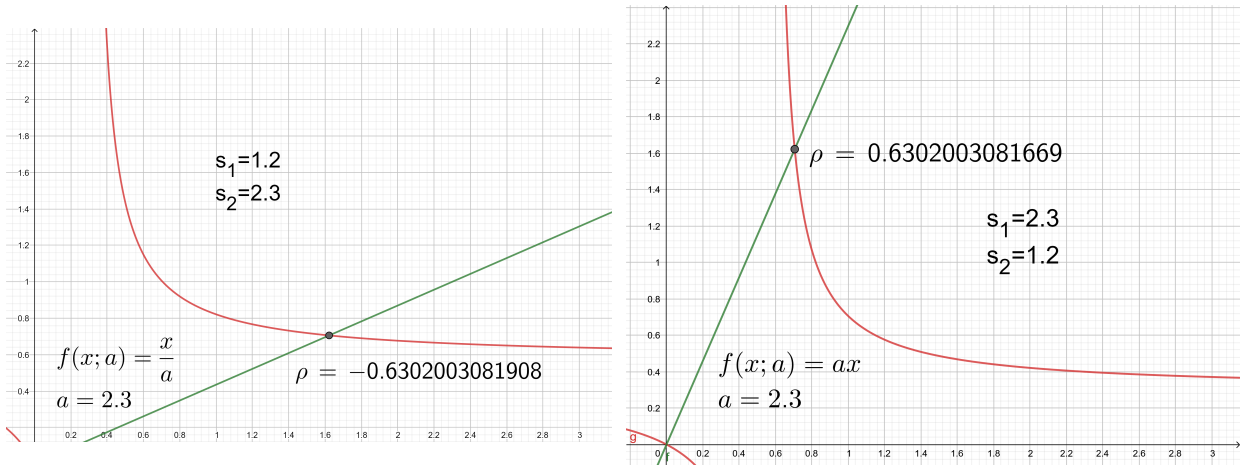


Figure 15: The two figures show the calculation of one-sided intervals under both data $S_1 = 1.2$, $S_2 = 2.3$ and $S_1 = 2.3$, $S_2 = 1.2$.

It would be interesting if a similar approach could be used for the data generating functions posed in 3.7.2 for the original data. This will be a much more complex case, where a relation between $2n$ random variables $Z_{1,i}, Z_{2,i}$, $i = 1, ..., n$ has to be defined. It is not unlikely that this method is applicable for the second data generating function using

3.0.1 and will result in the method using pivots described in definition 3.4, theorem 3.13 and theorem 3.14. That is, use the data generating functions

$$X_i + Y_i = \sqrt{2(1+\rho)}Z_{1,i}\ X_i - Y_i = \sqrt{2(1-\rho)}Z_{2,i}.$$

where (X_i, Y_i) are independently binormally distributed with means 0 and variances 1. Trying to set up the framework will be time consuming and therefore not attempted.

An interesting question, is if such a method is more generally applicable in cases where the inversion of the data generating function does not always exist. Secondly, it is also interesting to see if it will coincide with the method of inverting pivot distributions. The example used in this thesis is also one where the parameter is of one dimension. Is such a method viable if there are multiple unknown parameters? Questions like these would be useful to answer.

3.7.4 Generalized fiducial distribution

An alternative method for inverting the data generating function is the generalized fiducial distribution. The basis for a GFD is the data generating function. For the minimal sufficients, the data generating function is

$$S_1 = 2(1+\rho)U_1,\ \ S_2 = 2(1-\rho)U_2,$$

where U_1 and U_2 are chi-square distributed with n degrees of freedom. Before creating the GFD it is necessary to calculate the Jacobian like function

$$J(S_1, S_2, \rho) = D\left(\nabla_\rho G(\rho, (U_1, U_2))|_{(U_1,U_2)=G^{-1}((S_1,S_2),\rho)}\right),$$

where $\nabla_\rho G(\rho, (U_1, U_2))$ is the gradient matrix calculated with respect to ρ. There is only a scalar parameter, which means that the gradient matrix is the vector

$$\nabla_\rho G(\rho, (U_1, U_2)) = 2(U_1, U_2).$$

The inverse $G^{-1}((S_1, S_2), \rho)$ is the pivots

$$G^{-1}((S_1, S_1), \rho) = \frac{1}{2}\left(\frac{s_1}{1+\rho}, \frac{s_2}{1-\rho}\right).$$

The jacobian J is then given by

$$J(y, \theta) = D\left(A\right),$$

where

$$A = \left(\frac{s_1}{1+\rho}, \frac{s_2}{1-\rho}\right).$$

As mentioned in 2.3.3 the function $D(A)$ is dependent on a choice of norm and can give different solutions. Two possible choices for $D(A)$ following the 2-norm and infinity norm are respectively

$$D_2(A) = \sqrt{\sum_{i=1}^{n} A_i^2},\ \ D_\infty(A) = \sum_{i=1}^{n} |A_i|,$$

when A_i is element i of the vector A. The Jacobian can then be calculated as respectively

$$J_2(y,\theta) = \sqrt{\left(\frac{s_1}{1+\rho}\right)^2 + \left(\frac{s_2}{1-\rho}\right)^2} = \frac{1}{1-\rho^2}\sqrt{\rho^2(s_1^2+s_2^2) + 2\rho(s_2^2-s_1^2) + s_1^2 + s_2^2},$$

$$J_\infty(y,\theta) = \frac{s_1}{1+\rho} + \frac{s_2}{1-\rho} = \frac{1}{1-\rho^2}\left(\rho(s_2-s_1) + s_1 + s_2\right).$$

The absolute value function in the last expression was removed as both $s_1/(1+\rho)$ and $s_2/(1-\rho)$ are positive. For both the 2-norm and the infinity-norm, the symmetry conditions in definition 3.2 are satisfied.

It is also possible to calculate the GFD using the original data and the data generating function

$$X_i = Z_{1,i},\ Y_i = \rho Z_{1,i} + \sqrt{1-\rho^2}Z_{2,i},$$

where $Z_{1,i} \sim Z_{2,i} \sim N(0,1)$. The derivative of the first relation with X_i will necessarily be 0 and will not affect the GFD. We will therefore only look at the gradient using the function for Y_i. The gradient is then

$$\begin{bmatrix} Z_{1,1} - \frac{\rho}{\sqrt{1-\rho^2}}Z_{2,1} \\ \vdots \\ Z_{1,n} - \frac{\rho}{\sqrt{1-\rho^2}}Z_{2,n} \end{bmatrix}.$$

The inverse of the data generating functions with respect to $Z_{1,i}$ and $Z_{2,i}$ are

$$Z_{1,i} = x_i,\ Z_{2,i} = \frac{y_i - \rho x_i}{\sqrt{1-\rho^2}}.$$

For each i

$$Z_{1,i} - \frac{\rho}{\sqrt{1-\rho^2}}Z_{2,i} = \frac{1}{1-\rho^2}\left(x_i(1-\rho^2) - \rho(y_i - \rho x_i)\right) = \frac{x_i - \rho y_i}{1-\rho^2}.$$

The Jacobians under the 2-norm and infinity-norm are then respectively

$$J_2 = \sqrt{\sum_{i=1}^n \left(\frac{x_i - \rho y_i}{1-\rho^2}\right)^2} = \frac{1}{1-\rho^2}\sqrt{\sum_{i=1}^n (x_i - \rho y_i)^2},$$

$$J_\infty = \sum_{i=1}^n \left|\frac{x_i - \rho y_i}{1-\rho^2}\right| = \frac{1}{1-\rho^2}\sum_{i=1}^n |x_i - \rho y_i|.$$

One issue is that the Jacobian will treat x_i and y_i differently. For instance, in the infinity norm, if $|x_i| \geq |y_i|$, then the value of ρ will not affect the sign of $x_i - \rho y_i$. If $|y_i| \geq |x_i|$, then the value of ρ will affect the sign of $x_i - \rho y_i$. The distribution of ρ will therefore not satisfy the symmetry conditions in definition 3.1.

A different approach to a data generating function for the original data is based on corollary 3.0.1. This relation can be rewritten into $X_i + Y_i = \sqrt{2(1+\rho)}Z_{1,i}$ and $X_i - Y_i = \sqrt{2(1-\rho)}Z_{2,i}$, where $Z_{1,i} \sim Z_{2,i} \sim N(0,1)$. By inverting them wrt. X_i and Y_i, we get the data generating function

$$X_i = \frac{1}{2}\left(\sqrt{2(1+\rho)}Z_{1,i} + \sqrt{2(1-\rho)}Z_{2,i}\right),\ Y_i = \frac{1}{2}\left(\sqrt{2(1+\rho)}Z_{1,i} - \sqrt{2(1-\rho)}Z_{2,i}\right).$$

As earlier, the Jacobian can be obtained by inverting with respect to $Z_{1,i}$ and $Z_{2,i}$,

$$Z_{1,i} = \frac{x_i + y_i}{\sqrt{2(1+\rho)}},\ Z_{2,i} = \frac{x_i - y_i}{\sqrt{2(1-\rho)}},$$

and differentiate the data generating function wrt. ρ

$$\frac{1}{2}\left(\frac{Z_{1,i}}{\sqrt{2(1+\rho)}} - \frac{Z_{2,i}}{\sqrt{2(1-\rho)}}\right),\ \frac{1}{2}\left(\frac{Z_{1,i}}{\sqrt{2(1+\rho)}} + \frac{Z_{2,i}}{\sqrt{2(1-\rho)}}\right).$$

Combining the two, the vector A consists of n pairs of the two following expressions

$$\frac{1}{2}\left(\frac{x_i + y_i}{2(1+\rho)} - \frac{x_i - y_i}{2(1-\rho)}\right),\ \frac{1}{2}\left(\frac{x_i + y_i}{2(1+\rho)} + \frac{x_i - y_i}{2(1-\rho)}\right).$$

Both of these can be rewritten into

$$\frac{1}{2(1-\rho^2)}(y_i - \rho x_i),\ \frac{1}{2(1-\rho^2)}(x_i - \rho y_i)$$

respectively.

The two Jacobian are

$$J_2 \propto \frac{1}{1-\rho^2}\sqrt{\sum_{i=1}^{n}(x_i - \rho y_i)^2 + \sum_{i=1}^{n}(y_i - \rho x_i)^2}$$

for the 2-norm, and

$$J_\infty \propto \frac{1}{1-\rho^2}\left(\sum_{i=1}^{n}|x_i - \rho y_i| + \sum_{i=1}^{n}|y_i - \rho x_i|\right)$$

for the infinity-norm. These Jacobian are similar to the previous ones found for the original data. The main difference is that these will satisfy the symmetry conditions in definition 3.1. By rewriting the 2-norm Jacobian, it is apparent that it is similar to the 2-norm Jacobian of the sufficient statistics. If the expression inside the square root of the 2-norm is rewritten, we get that

$$\rho^2\sum_{i=1}^{n}(x_i^2 + y_i^2) - 4\rho\sum_{i=1}^{n}(x_i y_i) + \sum_{i=1}^{n}(x_i^2 + y_i^2),$$

where $\sum_{i=1}^{n}(x_i^2 + y_i^2) = t_1$ and $\sum_{i=1}^{n}(x_i y_i) = t_2$ are minimal sufficient statistics on the form (3.3). As noted in section 3.3, they can be expressed as s_1 and s_2. Inserting this into the expression above, the 2-norm for the original data can be expressed using the minimal sufficient statistcs

$$\rho^2 t_1 - 4\rho t_2 + t_1 = \rho^2 \frac{s_1 + s_2}{2} - \rho(s_1 - S_2) + \frac{s_1 + s_2}{2}.$$

The Jacobian for the 2-norm is in that case

$$J_2 \propto \frac{1}{1-\rho^2}\sqrt{\rho^2(s_1 + s_2) - 2\rho(s_1 - s_2) + s_1 + s_2},$$

which is similar to the Jacobian for the sufficient statistics, however with s_1 and s_2 instead of s_1^2 and s_2^2. Rewriting the infinity-norm Jacobian is more of a challenge as the shape of J_∞ varies greatly with the data points (x_i, y_i). The numerator for the Jacobian is piece wise linear.

3.7.5 Bayesian posteriors as confidence distributions

In order for a posterior distribution to be a confidence distribution, the credible intervals created using the posterior have to satisfy the definition of confidence intervals. Given the cumulative distribution of a posterior distribution $\Pi(\rho)$, the lower one-sided credible interval of level p is given by $(-1, \rho_p]$, where $\Pi(\rho_p) = p$. Unfortunately, none of the cumulative functions for any of the posterior distributions in section 3.6 have been found. Testing whether they are confidence distributions or not is therefore limited to numerical methods.

An opposite position is to find confidence distributions that can be defined as posterior distributions. If the form of the distributions is a result of (32), then it is possible to study which g that can result in a Bayesian distribution. One initial note to make is that, in order to have a posterior distribution, the posterior needs to include a prior. Unless it is the uniform prior, the prior will have to be a function of the parameter and not the data. This is a necessary condition as long as the prior is not uniform.

Say that the density function of $\phi = g(U_1) - g(U_2)$ is given by $h(\phi)$. The density of ρ is then calculated using the transform $\phi(\rho) = g(s_1/(2(1+\rho))) - g(s_2/(2(1-\rho)))$. That is, the density of the CD is

$$f(\rho) = h(\phi(\rho))\left|-g'\left(\frac{s_1}{2(1+\rho)}\right)\frac{s_1}{2(1+\rho)^2} - g'\left(\frac{s_2}{2(1-\rho)}\right)\frac{s_2}{2(1-\rho)^2}\right|$$

$$= h\left(g\left(\frac{s_1}{2(1+\rho)}\right) - g\left(\frac{s_2}{2(1-\rho)}\right)\right)\left|g'\left(\frac{s_1}{2(1+\rho)}\right)\frac{s_1}{2(1+\rho)^2} + g'\left(\frac{s_2}{2(1-\rho)}\right)\frac{s_2}{2(1-\rho)^2}\right|.$$

The density consists of two factors. The first is $h(\phi(\rho))$. The only case for which $h(\phi(\rho))$ can produce a factor that is only dependent on ρ and not the data, is when $h(\phi)$ is constant. Which means that ϕ is uniformly distributed. The remaining factor would in that case have

to include the density function of the base model. The other option is that the second factor will give a prior distribution of ρ. That is

$$\left| g'\left(\frac{s_1}{2(1+\rho)}\right)\frac{s_1}{2(1+\rho)^2} + g'\left(\frac{s_2}{2(1-\rho)}\right)\frac{s_2}{2(1-\rho)^2}\right| = \gamma(s_1,s_2)\pi(\rho).$$

First, let $\gamma(s_1,s_2) \propto 1$. The only possible solution for g is that $g(x) = x^{-1}$. This solution can be found directly, but also using differential equations (derivative of $\gamma(s_1,s_2)\pi(\rho)$ wrt. s_1 and s_2 is 0). Given $g'(x) = x^{-1}$, $\delta(\rho) = 1/(1+\rho)+1/(1-\rho) = 2/(1-\rho^2)$. That is $g(x)$ has to be on the form $g(x) = \ln x$. This is an admissible function and will result in the CD named CVCD. Unfortunately, the density is not a posterior distribution of the binormal with known means 0 and known variances 1. It is however the marginal posterior distribution using the prior $\frac{1}{\sigma^2(1-\rho^2)}$ under the binormal with known means 0 and unknown common variance σ^2, see Appendix **??**. The alternative is that $\gamma(s_1,s_2)$ is not constant. This implies that the two following factorizations should hold,

$$g'\left(\frac{s_1}{2(1+\rho)}\right)\frac{s_1}{2(1+\rho)^2} = \gamma_1(s_1)\delta(\rho)$$

and

$$g'\left(\frac{s_2}{2(1-\rho)}\right)\frac{s_1}{2(1-\rho)^2} = \gamma_2(s_2)\delta(\rho).$$

However, this will further imply that the ratio of the two is independent of ρ, that is

$$\frac{g'\left(\frac{s_1}{2(1+\rho)}\right)\frac{s_1}{2(1+\rho)^2}}{g'\left(\frac{s_2}{2(1-\rho)}\right)\frac{s_1}{2(1-\rho)^2}} = \frac{s_1}{s_2}\left(\frac{1-\rho}{1+\rho}\right)^2\frac{g'\left(\frac{s_1}{2(1+\rho)}\right)}{g'\left(\frac{s_2}{2(1-\rho)}\right)} = f(s_1,s_2).$$

The solution with respect to g has to be such that

$$\frac{g'\left(\frac{s_1}{2(1+\rho)}\right)}{g'\left(\frac{s_2}{2(1-\rho)}\right)} \propto \left(\frac{1+\rho}{1-\rho}\right)^2.$$

The solution is then $g'(x) = x^{-2}$ and $g(x) = x^{-1}$. x^{-1} is not an admissible solution. Either way, this solution will also correspond with the uniform prior.

The only possible prior that will result in a confidence distributions on the form (32) is given by the uniform distribution. Because of a lack of any proof of uniqueness, there might exist suitable confidence distributions that are not on the form derived from (32). Priors that lead to confidence distributions can therefore exist.

3.7.6 Comparing the CDs

Before conducting any analysis on the confidence distribution, it is useful to visualize and compare them. The density alone can show issues and benefits with the various distributions.

The exact confidence distributions proposed in section 3.7 are the ones that will be studied in this subchapter. Characteristics one can look for is the spread of the density and whether the density is unimodal, which means it has a unique maximum.

All the exact CDs presented in section 3.7 are visualized in figure 16 and 17. From the figures, it is clear that the only multimodal distribution is diffCD, which has two modes. The others seem to have only one mode. For the UVCD and CVCD, an exact expression for the density exists, which can be used to verify that claim. For CD1, that is not the case and can only be assumed based on a series of simulations.

It is also possible to compare the CDs with posterior distributions. A figure of all the posteriors alongside the diff CD can be seen in figure 18. The figure shows that the posteriors and the diff CD hold similar shapes.

The fiducial distributions for the sufficient statistics are also shown in figure 19. It is apparent that the fiducial distribution are not as focused as the difference confidence distribution.

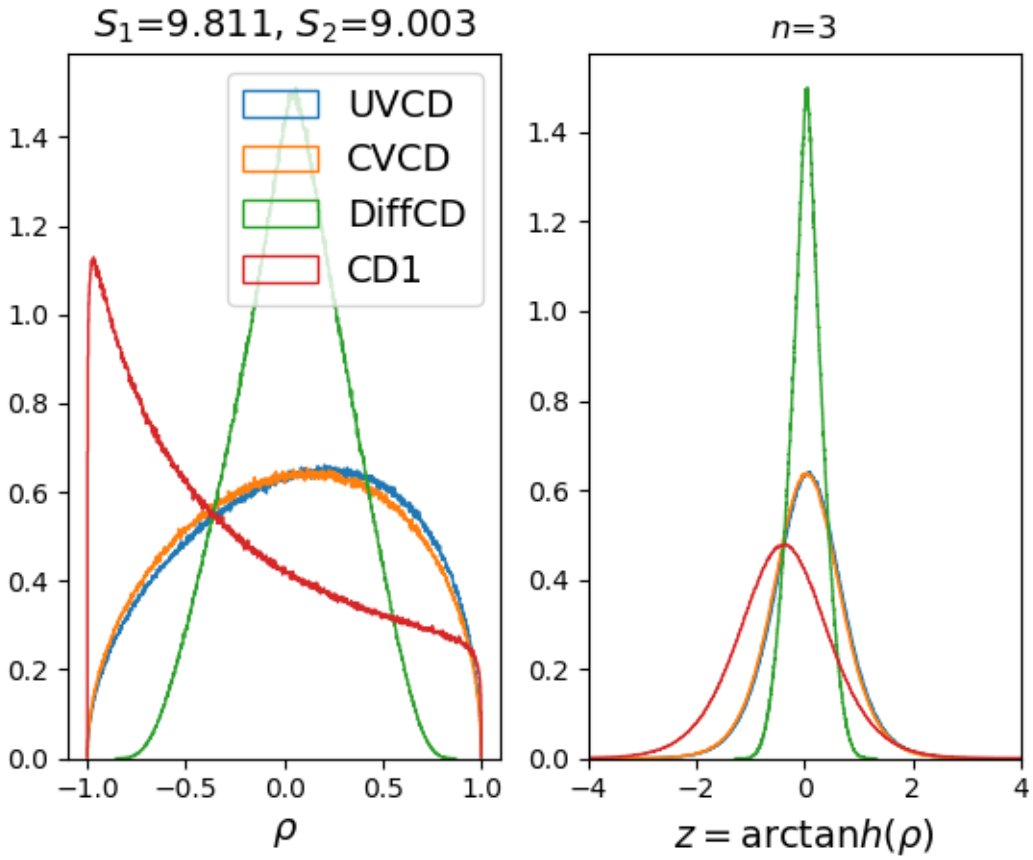


Figure 16: Sampled density for all four CDs proposed in section 3.7 under the parameter ρ and $z(\rho) = \text{arctanh}(\rho)$. The densities are given data set 1 in the Appendix A.1.

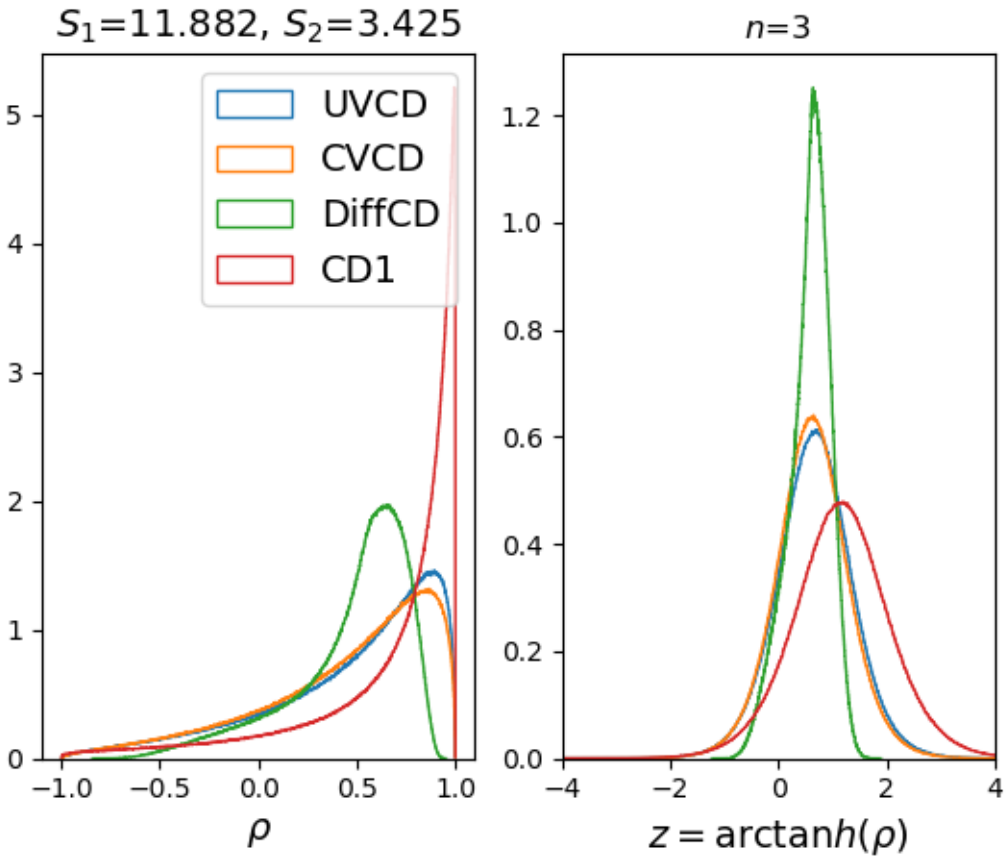


Figure 17: Samples density for all four CDs proposed in section 3.7 under the parameter ρ and $z(\rho) = \text{arctanh}(\rho)$. The densities are given data set 3 in the Appendix A.1.

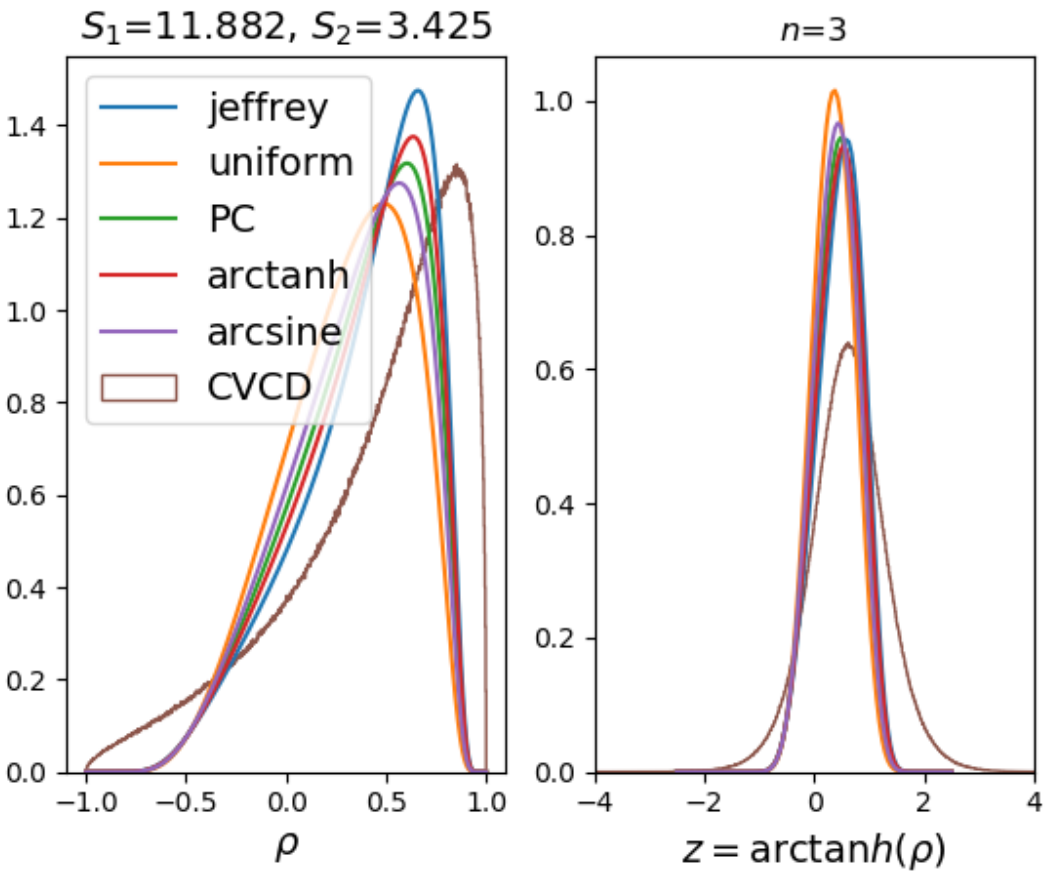


Figure 18: Density of all posterior distributions and the CVCD as functions of ρ (left) and $z(\rho) = \text{arctanh}(\rho)$ (right). The densities are given data set 3 in Appendix A.1

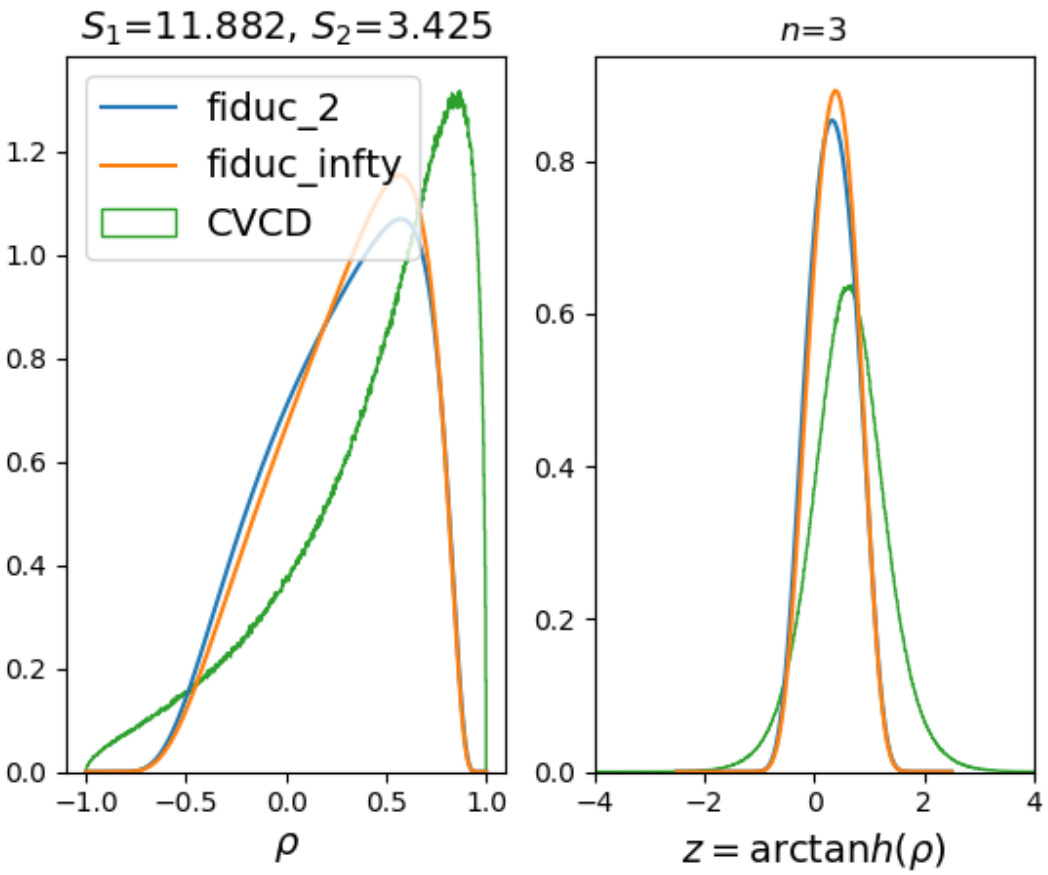


Figure 19: Density of the two fiducial distributions in theorem 3.5 (fiduc_2) and theorem 3.6 (fiduc_2) alongside the CVCD. The densities are given data set 3 in Appendix A.1

3.8 Frequentist point estimators

The following estimators will be used in this report. For simplicity, $SS_X = \sum_{i=1}^n x_i^2$, $SS_Y = \sum_{i=1} y_i^2$ and $SS_{XY} = \sum_{i=1}^n x_i y_i$. Each estimator is obtained from (Fosdick and Raftery 2012).

The **empirical correlation**, also known as sample correlation.

$$\hat{\rho}_{EMP} = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}}. \tag{40}$$

This is the maximum likelihood estimator when the means are known and variances are unknown.

The **empirical correlation with variance 1**,

$$\hat{\rho}_{EMP2} = \begin{cases} -1, & \text{for } SS_{XY} < -n \\ \frac{SS_{XY}}{n}, & \text{for } -n < SS_{XY} < n \\ 1, & \text{for } n < SS_{XY}. \end{cases} \tag{41}$$

The final choice is the **maximum-likelihood estimator** for known mean and variance.

$$\hat{\rho}_{MLE} := \text{solution of } \rho^3 - \frac{SS_{XY}}{n}\rho^2 - \frac{n - SS_X - SS_Y}{n}\rho - \frac{SS_{XY}}{n} = 0.$$

3.8.1 Empirical correlation with variance 1

It is possible to use the empirical correlation where it is also assumed that the variance is 1. With that in mind, the empirical correlation would be

$$\frac{SS_{XY}}{n}.$$

However, the estimator are no longer bounded between -1 and 1. The solution is to simply truncate the estimator at these bounds. Which gives the estimator in (41).

3.8.2 Maximum-likelihood estimator

The maximum-likelihood estimator when all means and variances are known, is given as the solution of a third degree polynomial equation. The equation is

$$\rho^3 - \frac{SS_{XY}}{n}\rho^2 - \frac{n - SS_X - SS_Y}{n}\rho - \frac{SS_{XY}}{n} = 0.$$

Figure 20 shows the polynomial that determines the MLE. All the possible solutions are known. They are

$$\hat{\rho}_{MLE,1} = \frac{SS_{XY}}{3n} + \frac{2^{1/3}\psi}{3n\left(\gamma + \sqrt{4\psi^3 + \gamma^2}\right)^{1/3}} - \frac{\left(\gamma + \sqrt{4\psi^3 + \gamma^2}\right)^{1/3}}{3 \cdot 2^{1/3} n},$$

$$\hat{\rho}_{MLE,2} = \frac{SS_{XY}}{3n} - \frac{(1+i\sqrt{3})\psi}{3\cdot 2^{2/3} n\left(\gamma + \sqrt{4\psi^3+\gamma^2}\right)^{1/3}} + \frac{(1-i\sqrt{3})\left(\gamma+\sqrt{4\psi^3+\gamma^2}\right)^{1/3}}{6\cdot 2^{1/3} n},$$

$$\hat{\rho}_{MLE,3} = \frac{SS_{XY}}{3n} - \frac{(1-i\sqrt{3})\psi}{3\cdot 2^{2/3} n\left(\gamma + \sqrt{4\psi^3+\gamma^2}\right)^{1/3}} + \frac{(1+i\sqrt{3})\left(\gamma+\sqrt{4\psi^3+\gamma^2}\right)^{1/3}}{6\cdot 2^{1/3} n},$$

where

$$\psi = -3n(n - SS_X - SS_Y) - SS_{XY}^2,$$
$$\gamma = -36n^2 SS_{XY} + 9nSS_X SS_{XY} + 9nSS_Y SS_{XY} - 2SS_{XY}^3$$

(Fosdick and Raftery 2012). Out of the three solutions, it is possible that some are complex numbers. Which solutions that are complex or not, is fully dependent on ψ and γ. In the case displayed in figure 20, there are only one real solution. If there are multiple real solutions, the one that maximizes the likelihood is naturally the MLE estimator. A small note is that there are methods of determining which solution is the MLE, see (Fosdick and Raftery 2012).

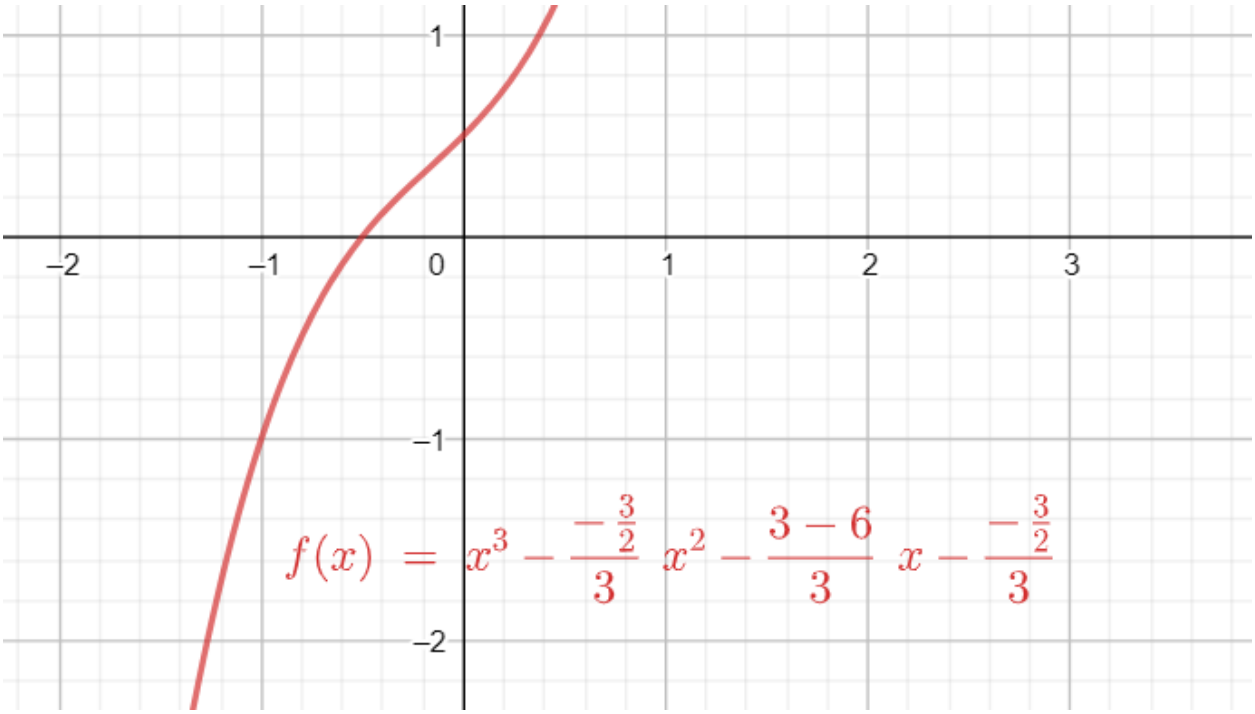


Figure 20: Plot of polynomial that determines the MLE.

3.8.3 Symmetry conditions for the estimators

All the estimators will satisfy the symmetry conditions in definition 3.1 and definition 3.2.

For the empirical correlation, changing the sign or swapping the X_i and Y_i will affect SS_X or SS_Y. The sign of SS_{XY} will however change if and only if either all the X_i or all

the Y_i changes sign. This implies that the symmetry conditions should hold for both the empirical correlation and the empirical correlation for known variances 1.

For the MLE, one can study the equation that defines it. That is

$$\rho^3 - \frac{SS_{XY}}{n}\rho^2 - \frac{n - SS_X - SS_Y}{n}\rho - \frac{SS_{XY}}{n} = 0.$$

Let $\hat{\rho}$ be a solution for the equation. If SS_{XY} changes sign, then $-\hat{\rho}$ would be a solution, as seen below

$$\begin{aligned}&(-\hat{\rho})^3 + \frac{SS_{XY}}{n}(-\hat{\rho})^2 + \frac{n - SS_X - SS_Y}{n}\hat{\rho} + \frac{SS_{XY}}{n}\\ &= -\left(\hat{\rho}^3 - \frac{SS_{XY}}{n}\hat{\rho}^2 - \frac{n - SS_X - SS_Y}{n}\hat{\rho} - \frac{SS_{XY}}{n}\right) = 0.\end{aligned}$$

This implies that the symmetry conditions holds for the MLE.

3.9 Bayesian point estimators

The second batch of estimators is the Bayesian estimators.

The first estimator is the **posterior mean**.

$$\hat{\rho}_E = E(\rho).$$

This estimator is the minimizer of both expected mean squared error, but also the expected Kullback-Leibler loss $\kappa(f(\,\cdot\,|\rho)||f(\,\cdot\,|\rho_0)$. The proofs can be seen in section 3.10.2 and section 3.10.3.

The second choice of estimator is the **posterior median**.

$$\hat{\rho}_M = \text{Median}(\rho).$$

The median is the minimizer of both the expected absolute error and the expected Fisher information metric. The proof for both are given in section 3.10.1.

The third estimator is denoted as the **FI2** estimator.

$$\hat{\rho}_{FI2} = f^{-1}\Big(\ E\big(\ f(\rho)\ \big)\ \Big),$$

where

$$f(x) = \sqrt{2}\arctan(\frac{\sqrt{2}x}{\sqrt{1+x^2}}) - \arcsin x.$$

The estimator is the minimizer of the expected squared Fisher information metric. Proof in section 3.10.2.

The fourth estimator is denoted as the **KL2** estimator.

$$\begin{aligned}\hat{\rho}_{KL2} := \text{solution of } &-\rho_0\frac{1 - E(\rho)\rho_0}{1-\rho_0^2} + \frac{1}{2}\rho_0 E(\ln(1-\rho^2)) - \rho_0\ln(1-\rho_0^2) + \rho_0\\ &+\frac{E(\rho) - \rho_0 E(\rho^2)}{1-\rho_0^2} + \frac{1}{2}E(\rho)\ln(1-\rho_0^2) - \frac{1}{2}E(\rho\ln(1-\rho^2)) - E(\rho) = 0\end{aligned}$$

Proof in section 3.10.4.

The final estimator is the **maximum a posteriori** estimator, or **MAP**.

$$\hat{\rho}_{MAP} = \arg\max_{\rho_0} g(\rho),$$

where $g(\rho)$ is the posterior of ρ.

3.10 Proofs of Baysian estimators

This subsection will go through the results that determined the choice of Bayesian estimators.

3.10.1 Fisher information metric and MAE as loss

The following proof shows that the median is the minimizer of the mean absolute error and the Fisher information metric.

Corollary 3.14.1. *Let $f(x)$ be a monotone continuous function and X be distributed by the density $\pi(x)$ on the space Ω_X. Then the median of X is the unique minimizer of*

$$\int_{\Omega_X} |f(x) - f(x_0)|\pi(x)dx$$

with respect to x_0.

Proof. Firstly, the expected loss can be written as

$$\int_{-\infty}^{\infty} |f(x) - f(x_0)|\pi(x)dx = \int_{-\infty}^{x_0} (f(x_0) - f(x)))\pi(x)dx - \int_{x_0}^{\infty} (f(x_0) - f(x))\pi(x)dx$$

$$= f(x_0)(2\Pi(x_0) - 1) - \int_{-\infty}^{x_0} f(x)\pi(x)dx - \int_{\infty}^{x_0} f(x)\pi(x)dx,$$

where $\Pi(\rho)$ is the cumulative function of $pi(x)$. The extreme points of the expected loss function can be found by taking the derivative

$$f'(x_0)(2\Pi(x_0) - 1) + 2f(x_0)\pi(x_0) - f(x_0)\pi(x_0) - f(x_0)\pi(x_0) = f'(x_0)(2\Pi(x_0) - 1) = 0.$$

As long as $f'(x) \neq 0$ for all x, the only solution to the problem is $\Pi(x_0) = 1/2$ or x_0 equals the median of $\pi(x)$.

The next step is to show that x_0 is a minimizer. As there is only one expremal point, it is only necessary to study the twice derivative of the expected loss at the extremal point. If it is positive, then the point is a minimum. The second derivative wrt. x_0 is

$$f''(x_0)(2\Pi(x_0) - 1) + f'(x_0)\pi(x_0).$$

At the median, $2\Pi(x_0) - 1 = 0$. The remaining part is necessarily positive. In other words, the solution $x_0 = Median(X)$ has to be the minimizer. □

Corollary 3.14.2. *The median of Θ is the minimizer of expected Fisher information metric if the Fisher information is non-zero on the sample space of Θ. That is, the median is a minimizer if there does not exist a θ such that*

$$\frac{\partial}{\partial\theta}l(\theta) = 0$$

for all x.

Proof. The proof follows directly from 2.2.1 and 3.14.1. □

3.10.2 MSE and squared Fisher information metric as loss

The following corollary given the estimator for minimizing MSE and squared Fisher information metric.

Corollary 3.14.3. *Let $f(x)$ be a monotone continuous function and X be distributed by the density $\pi(x)$ on the space Ω_X. Then $f^{-1}(Ef(X))$ with respect to $\pi(x)$ is the minimizer of*

$$\int_{\Omega_X} \big(f(x) - f(x_0)\big)^2\pi(x)dx$$

with respect to x_0.

Proof. First, the expected loss is

$$\begin{aligned}\int_{-\infty}^{\infty}(f(x_0) - f(x))^2\pi(x)dx &= \int_{-\infty}^{\infty}(f(x_0)^2 - 2f(x_0)f(x) + f(x)^2)\pi(x)dx \\ &= f(x_0)^2 - 2f(x_0)E(f(X)) + E(f(X)^2).\end{aligned}$$

The extreme points can be found by taking the derivative of the expected loss with respect to x_0. This gives the equation

$$2f(x_0)f'(x_0) - 2f'(x_0)E(f(X)) = 2f'(x_0)(f(x_0) - E(f(X))) = 0.$$

As long as $f'(x_0) \neq 0$ for all x, the only solution is $f(x_0) = E(f(X))$ or $x_0 = f^{-1}(E(f(X)))$.

The next step is to show that x_0 is a minimizer. As there is only one expremal point, it is only necessary to study the twice derivative of the expected loss at the extremal point. If it is positive, then the point is a minimum. The second derivative wrt. x_0 is

$$f''(x_0)(f(x_0) - E(f(X))) + f'(x_0)^2.$$

At the minimizer, $f(x_0) - E(f(X)) = 0$. The remaining part is necessarily positive. In other words, the solution $x_0 = f^{-1}\left(E(f(X))\right)$ is the minimizer. □

3.10.3 Kullback-Leibler divergence as loss

Corollary 3.14.4. *Let $\pi(\rho)$ be the distribution of the correlation ρ in a binormal distribution with known mean and variance. Then the posterior mean is the minimizer of the expected Kullback-Leibler divergence. That is*

$$E(\rho) = \arg\min_{\rho_0} E\left(\kappa\Big(f(\,\cdot\,|\rho)||f(\,\cdot\,|\rho_0)\Big)\right)$$

Proof. The expected loss using Kullback-Leibler divergence as loss function is

$$\int_{-1}^{1}(-\frac{1}{2}\ln\frac{1-\rho^2}{1-\rho_0^2}+\frac{1-\rho_0\rho}{1-\rho_0^2}-1)f(\rho)d\rho = -\frac{1}{2}E(\ln(1-\rho^2))+\frac{1}{2}\ln(1-\rho_0^2)+\frac{1-\rho_0 E(\rho)}{1-\rho_0^2}-1.$$

The extreme points can be calculated by finding the derivative with respect to ρ_0

$$-\frac{\rho_0}{1-\rho_0^2}+\frac{-2\rho_0(1-\rho_0 E(\rho))-E(\rho)(1-\rho_0^2)}{(1-\rho_0^2)^2} = -\frac{(E(\rho)-\rho_0)(1+\rho_0^2)}{(1-\rho_0^2)^2} = 0.$$

The only real solution to this equation is for $\rho_0 = E(\rho)$. That is, the estimator is equal to the expected value of ρ given the distribution $f(\rho)$.

The next step is to show that ρ_0 is a minimizer. If we let $\rho_0 \to 1$, then the Kullback-Leibler divergence goes to infinity. The solution $\rho_0 = E(\rho)$ is therefore a minimizer. □

A small note, there are reasons to believe that the above is true of the Kullback-Leibler divergence of any distribution from the exponential family. This will in that case be because of the connection between Kullback-Leibler divergence and Bregman divergence. The minimizer of the expected Bregman divergence is always the mean. For further reading, see (*Exponential family*) and (*Bregman divergence*). It is also possible that this can hold for the curved exponential family, as the binormal with known means and variances are part of it.

3.10.4 Squared Kullback-Leibler divergence

Corollary 3.14.5. *Let $\pi(\rho)$ be the distribution of the correlation ρ in a binormal distribution with known mean and variance. Then the minimizer of the expected squared Kullback-Leibler divergence solves the equation*

$$\begin{aligned} -\rho_0\frac{1-E(\rho)\rho_0}{1-\rho_0^2}+\frac{1}{2}\rho_0 E(\ln(1-\rho^2))-\rho_0\ln(1-\rho_0^2)+\rho_0 \\ +\frac{E(\rho)-\rho_0 E(\rho^2)}{1-\rho_0^2}+\frac{1}{2}E(\rho)\ln(1-\rho_0^2)-\frac{1}{2}E(\rho\ln(1-\rho^2))-E(\rho)=0 \end{aligned} \tag{42}$$

with respect to ρ_0.

The minimizer of the squared Kullback-Leibler divergence can be found by solving the equation

$$\frac{\partial}{\partial \rho_0}\int_{-1}^{1}\kappa^2(\rho||\rho_0)f(\rho)d\rho = 2\int_{-1}^{1}\kappa(\rho||\rho_0)\frac{\partial}{\partial \rho_0}\kappa(\rho||\rho_0)f(\rho)d\rho = 0. \tag{43}$$

From section 3.10.3, we know that

$$\frac{\partial \kappa(\rho|\rho_0)}{\partial \rho_0} = -\frac{(\rho-\rho_0)(1+\rho_0^2)}{(1-\rho_0^2)^2}.$$

By inserting the expressions into (43), the equation for the minimizer ρ_0 is

$$\int_{-1}^{1} 2\kappa(\rho|\rho_0)\left(-\frac{(\rho-\rho_0)(1+\rho_0^2)}{(1-\rho_0^2)^2}\right)f(\rho)d\rho = 0.$$

The factors $(1+\rho_0^2)$ and $1/(1-\rho_0^2)^2$ are both independent of ρ and non-zero on (-1,1). They will therefore have no affect on the solution and can be removed. The remaining equation is then

$$\int_{-1}^{1}\left(-\rho_0\frac{1-\rho\rho_0}{1-\rho_0^2}+\rho_0\frac{1}{2}\ln\frac{1-\rho^2}{1-\rho_0^2}+\rho_0+\rho\frac{1-\rho_0\rho}{1-\rho_0^2}-\rho\right)f(\rho)d\rho = 0$$

By splitting up the terms and calculating the expectancy, the equation is

$$\begin{aligned}-\rho_0\frac{1-E(\rho)\rho_0}{1-\rho_0^2}+\frac{1}{2}\rho_0 E(\ln(1-\rho^2))-\frac{1}{2}\rho_0\ln(1-\rho_0^2)+\rho_0\\ +\frac{E(\rho)-\rho_0 E(\rho^2)}{1-\rho_0^2}+\frac{1}{2}E(\rho)\ln(1-\rho_0^2)-\frac{1}{2}E(\rho\ln(1-\rho^2))-E(\rho)=0.\end{aligned}$$

No explicit expressions for the solution have been found.

3.10.5 Estimators with regards to priors

As all the priors are part of the conjugate family given in equation (22), it can be seen as the posterior distribution without any information. It is therefore of interest to apply the estimators to the conjugate priors to see the estimates with no information given. This can imply the bias of the prior. Some initial comments to make about the priors is that they are all symmetric around 0 and some are improper. For the proper priors symmetry implies that the expectation of any odd function is 0. Similarly, the median will be 0. This implies that $\hat{\rho}_E = \hat{\rho}_M = \hat{\rho}_{FI2} = 0$. For the improper priors, neither the median nor the expectation in general exists.

3.10.6 Additional comments about the Bayesian estimators

There are some additional comments to be made about the Bayesian estimators. These comments are made based on studying the estimators for different observed data. General assumptions will be made.

The Bayesian estimators generally follow a specific order for any prior. The order from smallest to largest is $\hat{\rho}_E$, $\hat{\rho}_M$, $\hat{\rho}_{FI2}$ and $\hat{\rho}_{MAP}$. There are no proof of this, however I have yet to see a counter-example. Figure 21 shows an example. The order is believed to be a result of the posterior's shapes. For all the posteriors, the density is skewed towards larger correlations, as can be seen in figure 6.

Secondly, the median and the minimizer of the squared Fisher information metric, $\hat{\rho}_{FI2}$ will hold similar values. Whether they are equal or not remains to be proven.

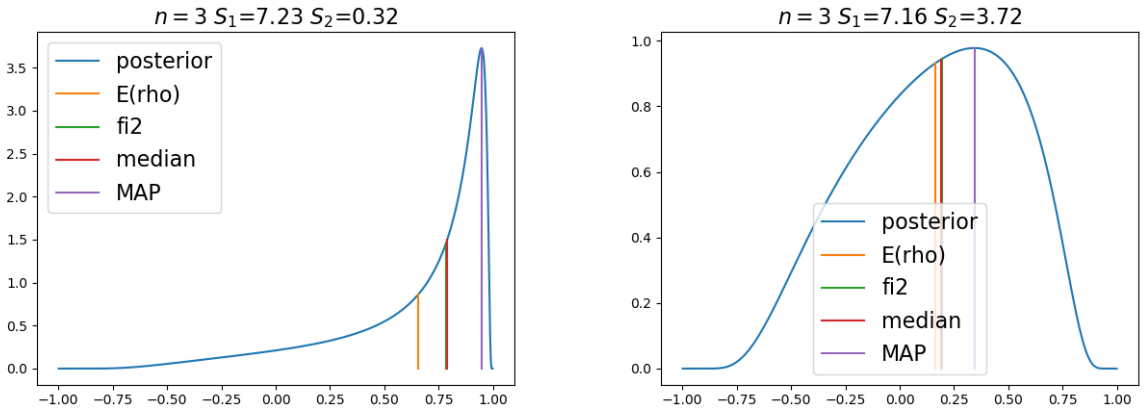


Figure 21: Bayesian estimates with the uniform prior and $n = 3$ data points. The simulated data can be found in A.1.

All the Bayesian estimators will satisfy the symmetry conditions in definition 3.2. Firstly, all the posterior distributions satisfy the symmetry. Let $L(\rho, \hat{\rho})$ be any loss function. The Bayesian risk of the estimator h is

$$R(S_1, S_2, \hat{\rho}) = \int_{-1}^{1} L(x, \hat{\rho})\pi(x|S_1, S_2)dx.$$

Let $\hat{\rho} = h(s_1, s_2)$ be the minimizer of the risk for data $S_1 = s_1$ and $S_2 = s_2$. Under the data $S_1 = s_2$ and $S_2 = s_1$, the risk is given as

$$R(s_2, s_1, \hat{\rho}) = \int_{-1}^{1} L(x, \hat{\rho})\pi(x|s_2, s_1)dx.$$

If the loss function is such that $L(x, \hat{\rho}) = L(-x, -\hat{\rho})$ for all $x, \hat{\rho} \in (-1, 1)$, then the risk equals

$$R(s_2, s_1, \hat{\rho}) = \int_{-1}^{1} L(-x, -\hat{\rho})\pi(-x|s_1, s_2)dx = \int_{-1}^{1} L(x, -\hat{\rho})\pi(x|s_1, s_2)dx.$$

The mimimizer would necessarily be $h(s_2, s_1) = -h(s_1, s_2)$ and the symmetry conditions are satisfied. A final step is to show that for all the loss functions, $L(x, y) = L(-x, -y)$. For the

MAE and MSE, the criteria necessarily holds. For the Fisher information metric, the criteria will hold if $|f(x)| = |f(-x)|$. As f is odd, this is satisfied. Finally, the Kullback-Leibler divergence will also satisfy the condition as $\kappa(x||y)$ is a function of x^2, y^2 and xy.

4 Data analysis

The data analysis of this thesis will consist of two main parts. The first is point estimation of the correlation and the other is confidence distribution for the correlation. Both will have a focus on calculating average risk in order to evaluate the different point estimators and distribution estimators. These tests will be taken from a frequentist perspective using frequentist versions of average risk as seen in (1) and (4). For each estimator, multiple tests will be conducted to calculate the performance for different situations, that is for different true values for the correlation. In addition to calculating risks for the distribution estimators, the frequentist coverage will be tested for multiple of them.

There are two approaches that are used for the analysis. One approach is to use strictly numerical methods to estimate different quantities. The other approach is to use samples from a distribution in order to estimate quantities. The latter will create samples using either model generating functions developed in the theory section or Markov Chain Monte Carlo (MCMC).

Calculating the average risk for any estimator, either point or distribution, might not have an explicit expression. The risk under some true parameter ρ_0 will therefore be calculated by iterating the following three steps:

1. Simulate the data
2. Calculate the estimators, either point or distribution, given the data
3. Calculate the loss or risk of the estimator.

For each value of ρ_0, these steps will be applied iteratively a given number of times, often 10^5 times. The average risk is the mean of the losses or risks calculated. The algorithm for testing frequentistic coverage will be discussed further in section 4.3.1. All numerical analysis is produced using python.

4.1 Simulation of data

The data can either be expressed as the original binormal data or the minimal sufficient statistics. In most cases, the minimal sufficient will suffice, however some estimator uses the original data. In either case, the package *numpy.random* is used for simulations.

4.2 Estimation of Bayesian point estimators

Estimating the Bayesian point estimators in this thesis requires numerical methods. It is possible to use MCMC to calculate most of the estimators, but this can be very time consuming. The alternative is to use strictly numerical methods.

The first barrier for each estimator is that even the normalizing constant for any of the posteriors are known. This can be solved by integrating over the posterior using any kind of quadrature. For the point estimators in question, different approaches are necessary. Some

can be calculating by integration, but most might require some kind of optimization. Each method will be discussed separately in the following subchapters.

The numerics will be conducted primarily using the python package *scipy*. Quadrature is from the package *scipy.integrate* under the function *quad*. The function *fsolve* from *scipy.optimize* is used for solving equations and the function *mimize* from the same package is used for solving mimization problems.

4.2.1 Numerical estimation of $\hat{\rho}_E$

The mean is one of the simpler estimators to calculate as the only numerical process needed is a numerical integration. As the normalization constant is not known, it is crucial to calculate that as well using quadrature.

4.2.2 Numerical estimation of $\hat{\rho}_M$

The median poses more difficulty as it uses the cumulative function, which is not available analytically. It is however possible to find the median by solving the equation $\Pi(\rho) - 1/2 = 0$ numerically, where $\Pi(\rho)$ is the cumulative distribution function estimated numerically using a quadrature. This will mean that the algorithm is dependent on the convergence of two numerical methods which can become both time consuming and unstable. A good initial guess is therefore important to ensure convergence.

4.2.3 Numerical estimation of $\hat{\rho}_{FI2}$

In order to calculate the minimizer of squared Fisher information metric, both $E(f)$ and f^{-1} of the function 10 needs to be calculated. As neither of these have clear analytical expressions, it is necessary to use numerical estimations. Firstly $E(f)$ can be calculated using quadrature. Finally, the estimator can be approximated by solving the equation

$$f(\hat{\rho}_{FI2}) - E(f) = 0$$

numerically. This method will naturally be less time consuming than the median, as it is only necessary to do numeric integration ones. A good initial guess is also useful, however not as necessary as the convergence is mostly defined by the properties of $f(x)$ which is fairly well behaved.

4.2.4 Numerical estimation of $\hat{\rho}_{KL2}$

For $\hat{\rho}_{KL2}$ it is most useful to use the equation given in (42). Each of the expectations can be calculated using quadrature which leaves an equation that can be solved numerically.

4.2.5 Numerical estimation of $\hat{\rho}_{MAP}$

The last estimator is then the MAP estimator. This estimator was calculated by numerical optimization, that is by minimizing the negative posterior. Choosing a good initial guess

for the optimization is essential as the posterior can be very thin in some situations. If the initial guess misses the spike of the curve in these situations, the algorithm can have a too slow convergence.

4.2.6 Choices of initial guesses for point estimators

For both the MAP estimator and the median estimator, an optimization problem has to be solved. This will require an initial guess which can heavily affect the convergence of the numerical methods. The two estimators have two different choices of initial guesses based on fast but crude and inaccurate methods. Both will divide the parameter space into equidistant points. The choice was set to $m = 50$ points.

For the MAP, the initial guess is the point with the largest density value. For the median, a discretization of the density is created over each point and is normalized. The initial guess for the median is then the median of the discretized distribution.

4.3 Simulating confidence distributions

As mentioned in the introduction to this chapter, calculating the risk analytically for the distribution estimators will generally pose a challenge. Approximations to the risk can be calculated using samples drawn from the distribution estimator. For the posterior distributions, Markov Chain Monte Carlo (MCMC) with Metropolis-Hastings algorithm is necessary (JCGM 2008b, p. 5.9.6). The technicalities of MCMC and Metropolis-Hastings algorithm will not be explained, just like the numerical estimation for point estimators. Important information about MCMC is that it can be used to simulate from a distribution using the density function. This method is a common and flexible tool but it is computationally heavy. If a more direct method of simulating from a distribution is available, it should be chosen. For all of the exact confidence distributions it is possible to establish a model generating function. In some cases, it is expressed indirectly as the solution to an equation. These equations are generally on the form

$$\phi\left(\frac{S_1}{2(1+\rho)}, \frac{S_2}{2(1-\rho)}\right) = \phi(U_1, U_2),$$

where $U_1 \sim U_2 \sim \chi^2_n$. The correlation can be calculated by sampling U_1 and U_2 and solving the equation above using some numerical equation solver. An advantage is that the solution is known to lie in the interval (-1,1) and where the left hand side is decreasing as a function of ρ. There are methods that utilizes such knowledge to improve performance. The method found to perform the best is Brent's method brentq. In python, the package *scipy.optimize* offers the method under the function named *brentq*. Although solving an equation for each simulated point can be cumbersome, it is significantly more efficient than using MCMC in many cases. This is especially true due to the speed of Brent's method.

4.3.1 Testing confidence of distribution estimators

Given a distribution estimator with distribution function $C(x)$ it is possible to create interval estimators of different levels given by $I = (a, b]$, where $a \leq b$. The level of these intervals are given by $C(b) - C(a) = \alpha$. For any distribution estimator it is possible to study whether it is able to create confidence intervals. That is if the frequentistic coverage of the interval is at least as large as the level of the interval. This is a frequentist perspective and the coverage will be a function of the true parameter value ρ_0. If the interval estimators are confidence intervals independent on ρ_0 the distribution estimator can create confidence interval estimators. To be clear, a beneficial and general test is to see whether a distribution estimator is a confidence distribution as defined in 2.7. In that case, it is sufficient to test whether all one-sided interval estimators are confidence distributions.

The following test can be used to test both one-sided and two-sided interval estimators, with a small alteration. For simplicity, the test will be written specifically for one-sided intervals.

The coverage of the interval estimators given true correlation ρ_0 consists of m simulations of the data, where the coverage of the interval estimator is tested for the data. Each simulation consists of the following steps

1. Simulate data given ρ_0
2. Calculate and save $C(\rho_0)$
3. If $C(\rho_0) \leq \alpha$, the α level interval estimator covers ρ_0.

The amount of simulations where ρ_0 is covered gives the frequentist coverage. For one-sided α level interval estimator, a point ρ is covered if $C(\rho) \leq \alpha$. In that case, calculating $C(\rho_0)$ for each distribution is sufficient to testing the coverage for any level. If one wants to do the same for two-sided intervals, a point ρ is in an α level two-sided interval estimator if $\alpha/2 \leq C(\rho) \leq 1 - \alpha/2$. An upside to using $C(\rho_0)$ to test for coverage, is that it is not dependent on α. If all $C(\rho_0)$ for each simulation is stored, the coverage can be calculated for each level.

4.4 Problems of the data analysis

As mentioned in section 4.2.6 about the initial choice, some of the numerical methods are less stable. Problems occurred more frequently with the convergence of the estimator $\hat{\rho}_{KL2}$. Because of lack of time, finding methods to ensure the convergence of the estimations were not finished. This can imply that the data is not trustworthy. Neither the results nor the further discussion will include the $\hat{\rho}_{KL2}$.

4.5 Results

4.5.1 Comparing point estimators

Generally throughout the results, the Bayesian estimators will be named something similar to "uniformE". The names imply the use of prior and the use of estimator. The names are generally written "<name of prior><name of type of estimator>".

In this problem there are 19 estimators, 16 of which are Bayesian (not including $\hat{\rho}_{KL2}$). The average loss of each of the estimators has been estimated for $\rho = 0.0,\ 0.1,\ ..., 0.9$, for both sample size 3 and 10. Presenting all of these results here will only hide the important information. In order to present useful characteristics of the estimators, only some results are shown in figures. The full data sets will not be presented in this report, see appendix A.2.

The following four figures shows the loss of various estimators. Each figure will include four plots, each showing the loss under a specific loss function. Here, FIM stands for Fisher Information Metric and KL stands for Kullback-Leibler divergence. Figure 22 shows the loss of the frequentist estimators. The goal of this figure is to show the characteristics of the frequentist estimators. Figure 23 shows the loss of the Bayesian estimators using uniform prior. The goal of this figure is to demonstrate the behaviour of the types of estimators for any given prior. Figure 24 shows the loss of all the posterior means of each prior. The goal of this figure is to demonstrate the relation between different priors using the same estimators. Figure 25 shows the loss of the posterior mean with uniform prior, posterior median with Jeffreys prior and empirical correlation with variance 1. The goal of this figure is to show the relationship between Bayesian estimators and the special empirical correlation.

Finally, the next three figures will demonstrate the distribution of various estimators. These comparisons will be used to show how some estimators are better than others. The distribution can also be used to compare estimators. Figure 26 shows the distribution of the posterior mean with uniform prior alongside the frequentist estimators. Figure 27 displays how the Bayesian estimators will change with correlation and sample size. Lastly, figure 28 shows the distribution of both the posterior median and $\hat{\rho}_{FI2}$ with uniform prior.

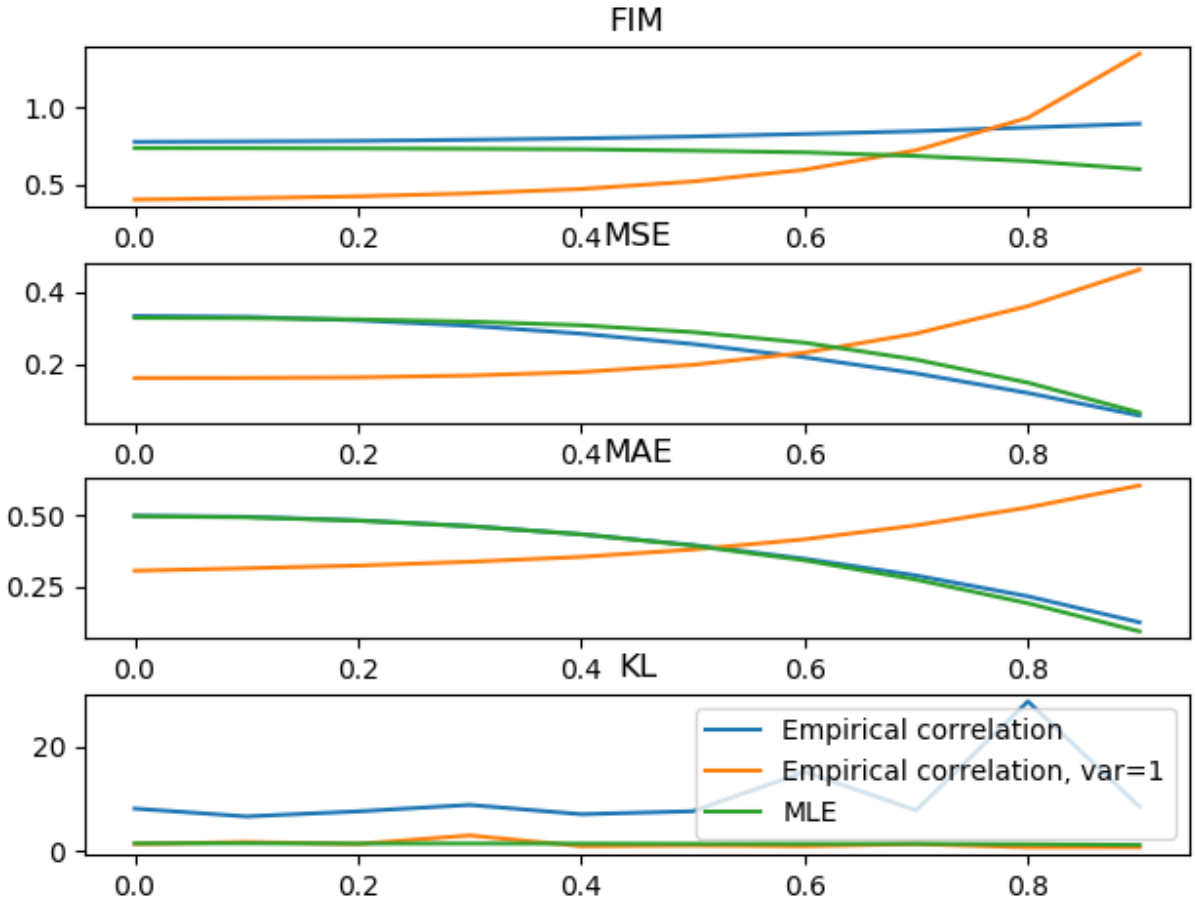


Figure 22: Loss of frequentist estimators for $n = 3$ data points

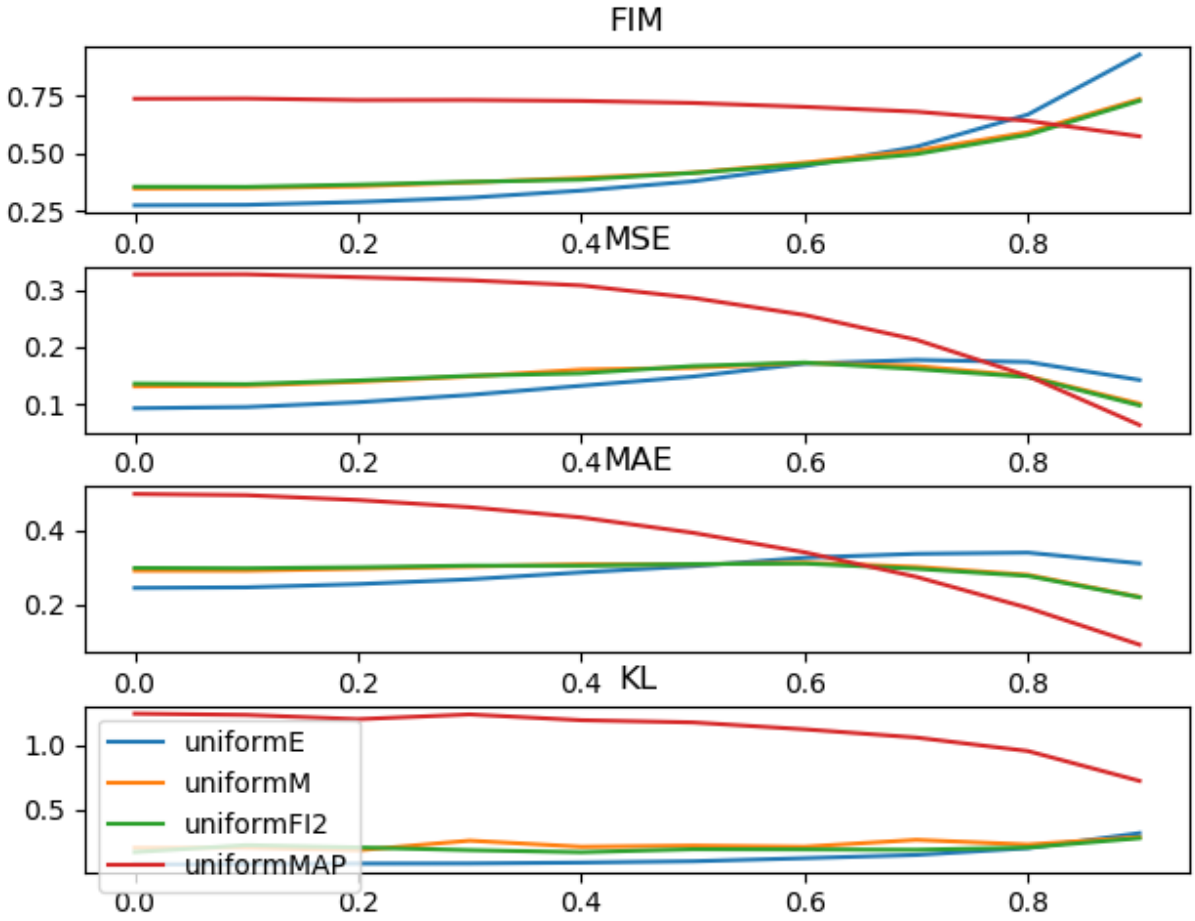


Figure 23: Loss of the Bayesian estimator with uniform prior for $n = 3$ data points

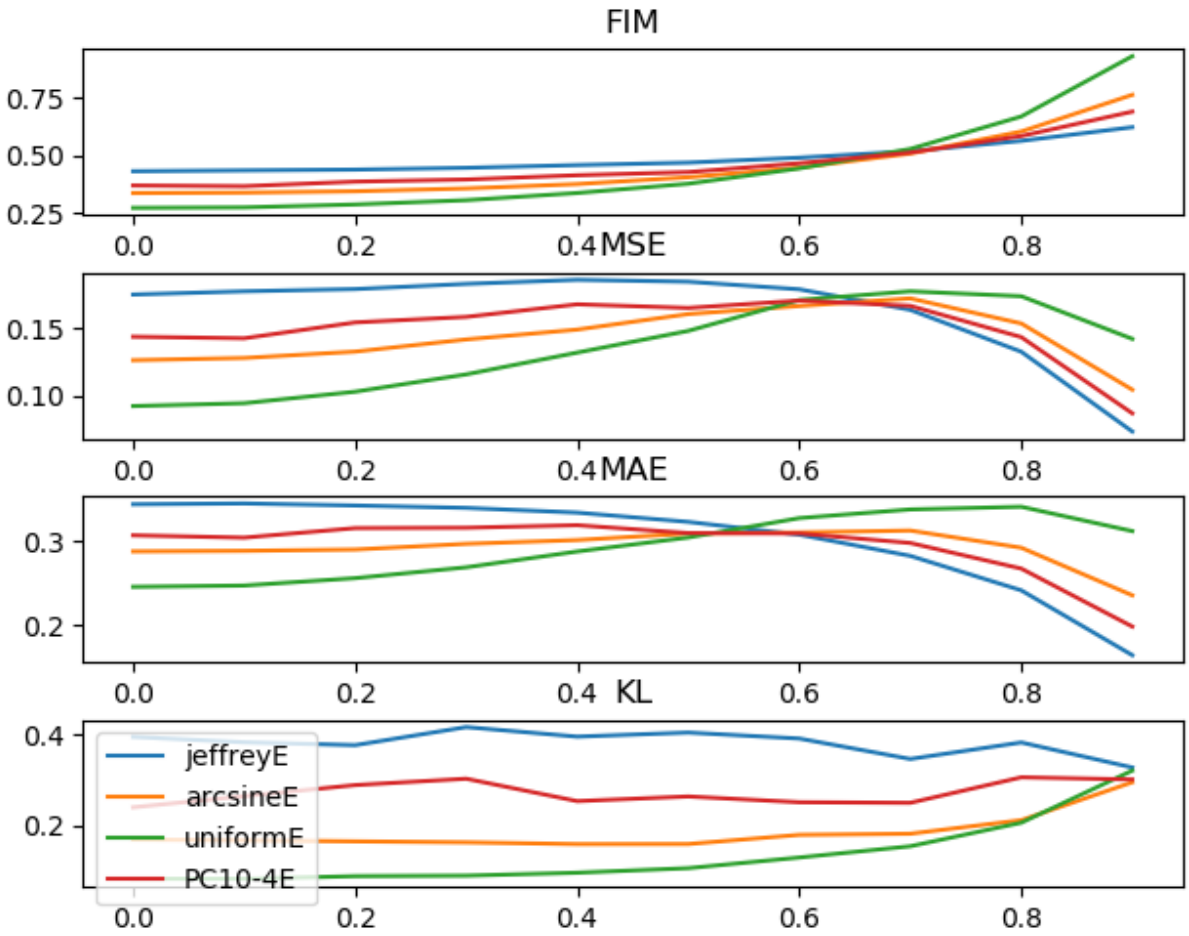


Figure 24: Loss of the posterior means for $n = 3$ data points

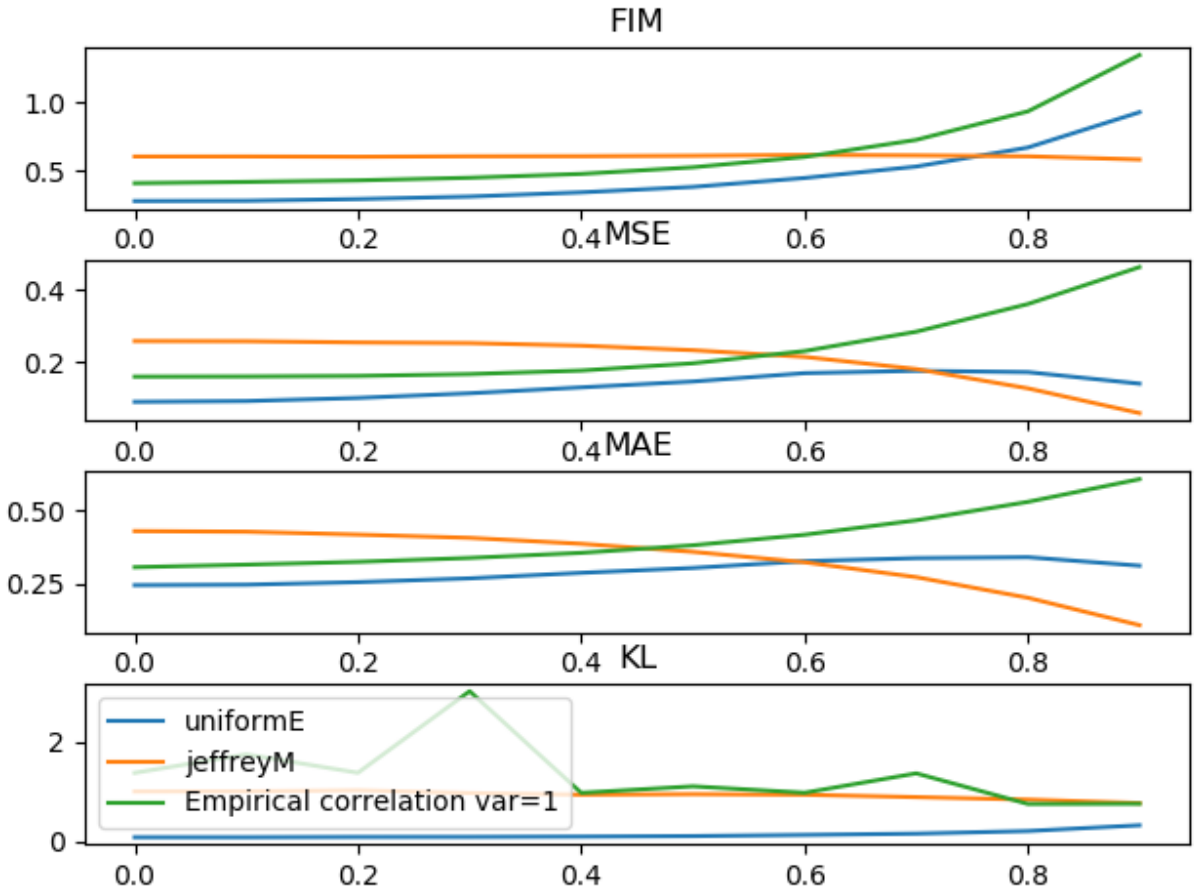


Figure 25: Loss of the posterior mean with uniform prior, posterior median with Jeffreys prior and empirical correlation with variance 1.

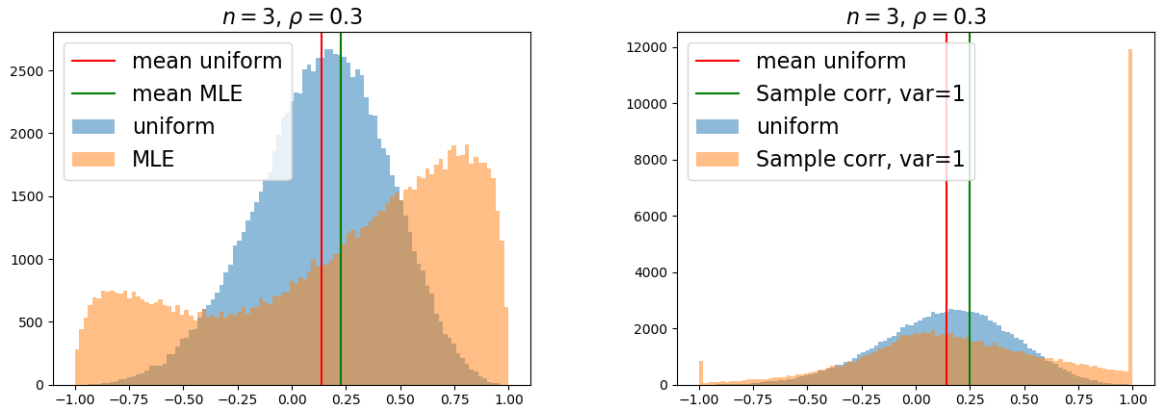


Figure 26: Distribution of the posterior mean for uniform prior, and both the MLE and empirical correlation with variance 1.

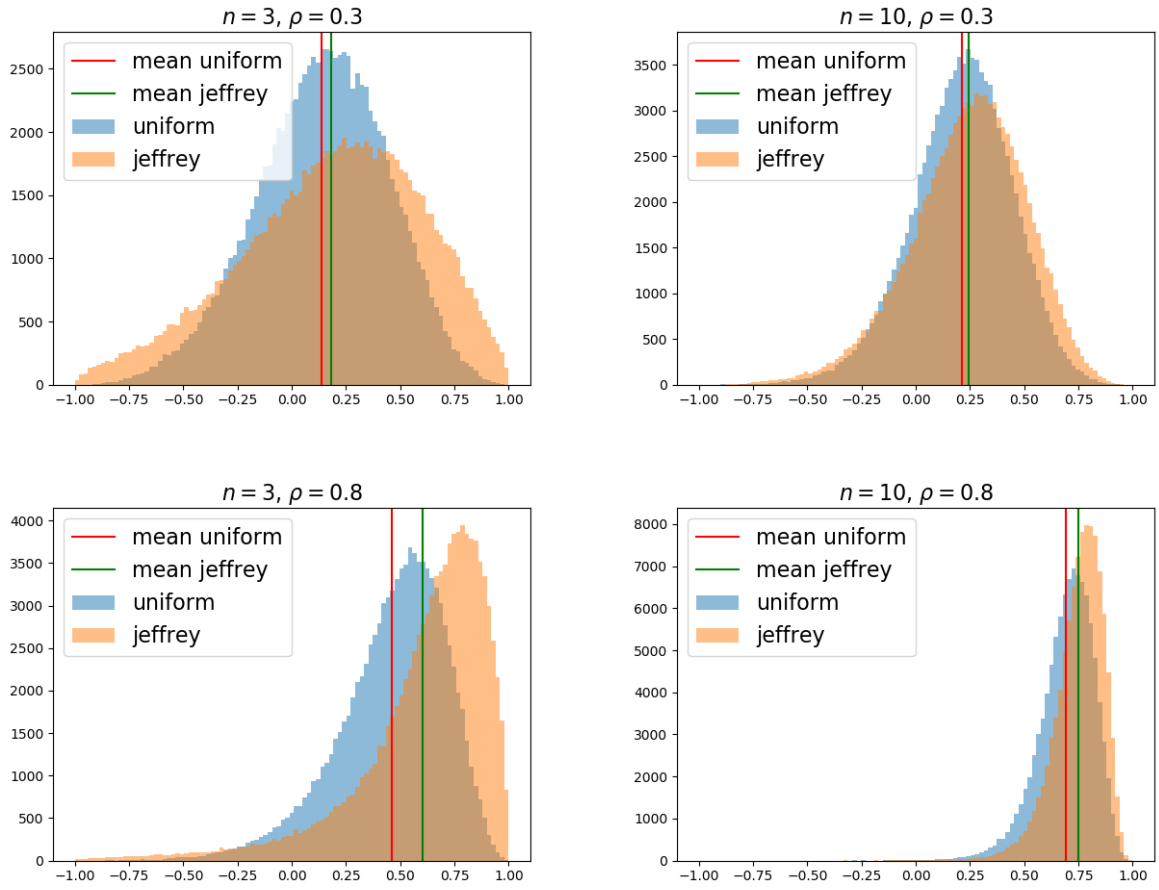


Figure 27: Distribution of the posterior mean for both Jeffreys prior and uniform prior.

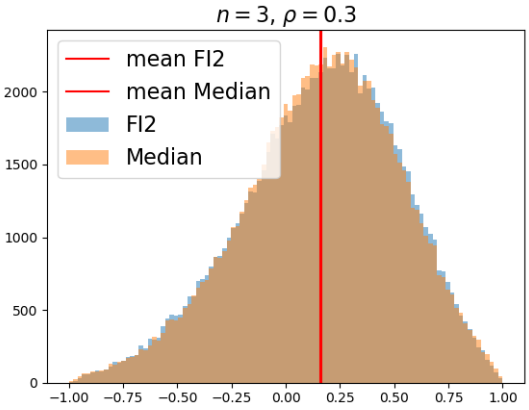


Figure 28: Distribution of the posterior median and $\hat{\rho}_{FI2}$ for uniform prior.

4.5.2 Testing coverage of distribution estimators

The frequentist coverage can be plotted as a function of the levels α. If the distribution is an exact confidence distribution, the empirical coverage equals the assumed level α. The distribution is also a CD if the empirical coverage is at least the assumed level α. The coverage is displayed in figure 29, 30 and 31 as the difference between the frequentistic coverage and the assumed level. These results are calculated using 1000 simulations. More simulations would be optimal, however time consuming. In each figure, all posteriors are displayed. The coverage error of the fiducial distributions from equation (24) and (25) can be seen in figure 4.5.2, 4.5.2 and 4.5.2. Fiduc2 and fiducinf are the names of the 2-norm fiducial and the infinity-norm fiducial, respectively.

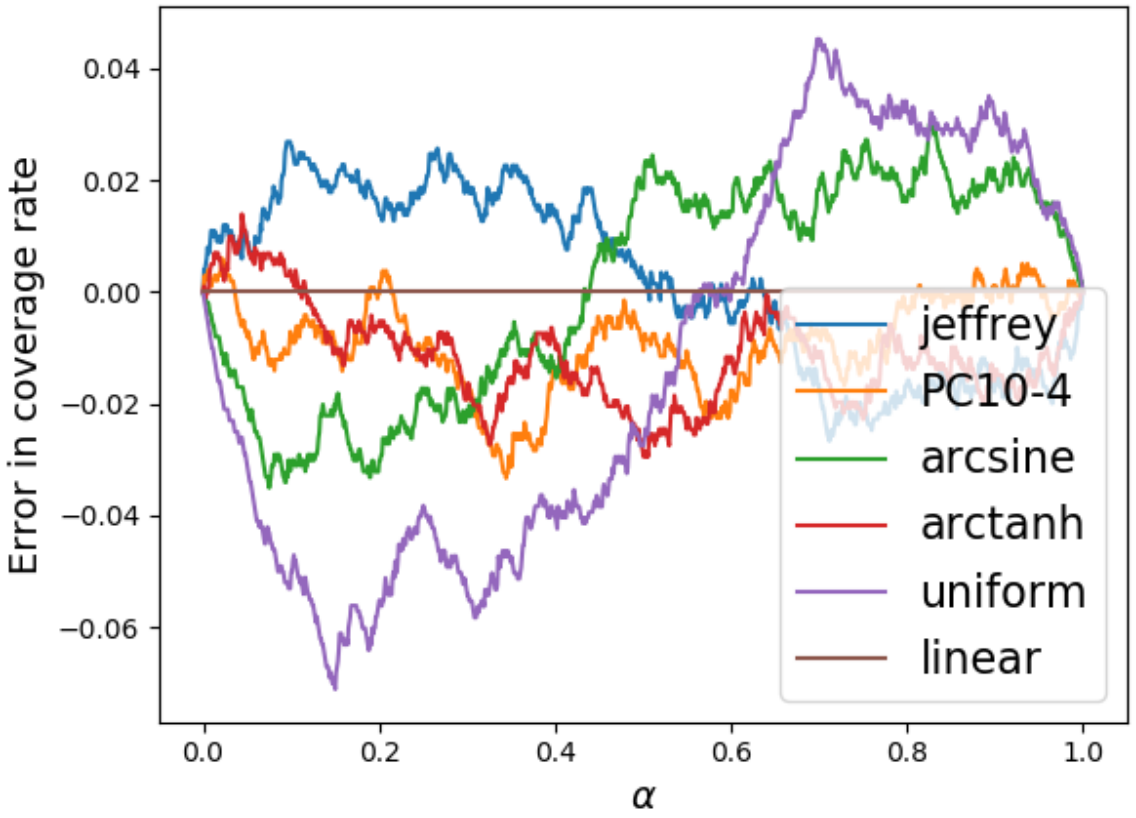


Figure 29: Error in frequentist convergence as a function of the levels α for various posterior distributions under $n = 3$ data points sampled for correlation $\rho = 0.0$. The error is calcualted as the difference between frequentist coverage and level of one-sided interval estimators.

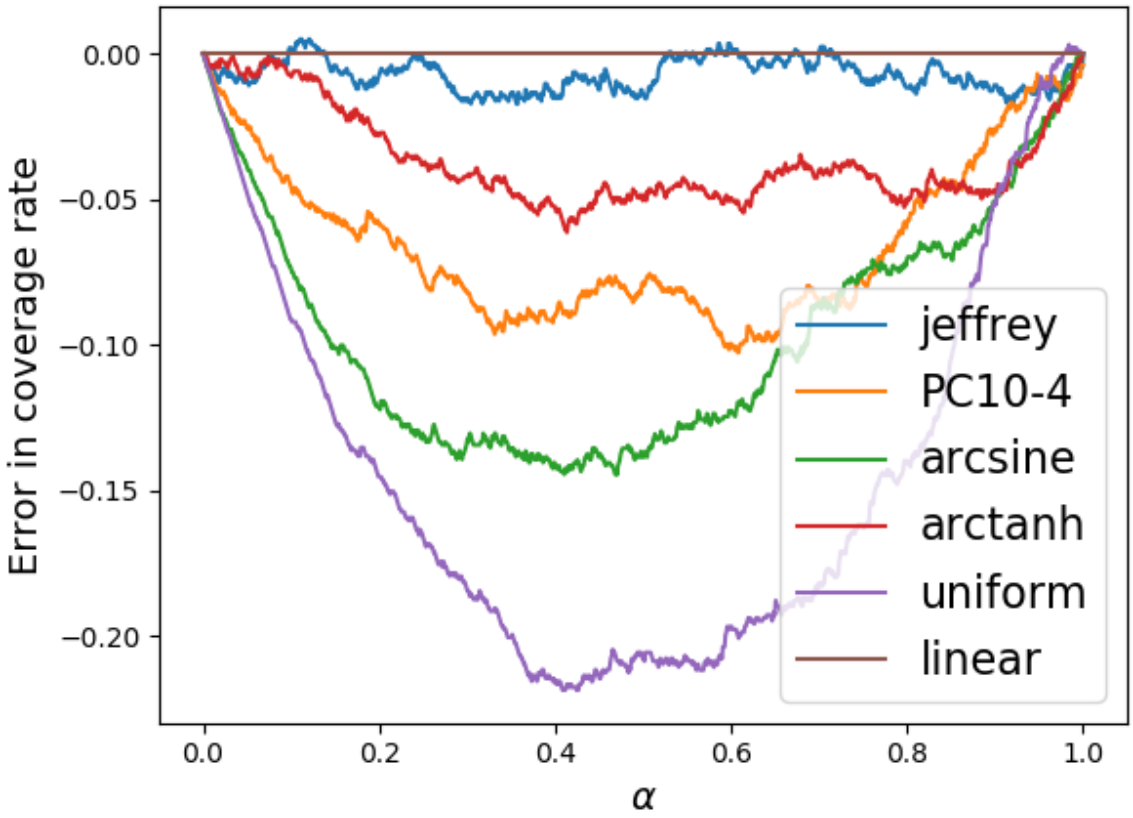


Figure 30: Error in frequentist convergence as a function of the levels α for various posterior distributions under $n = 3$ data points sampled for correlation $\rho = 0.5$. The error is calcualted as the difference between frequentist coverage and level of one-sided interval estimators.

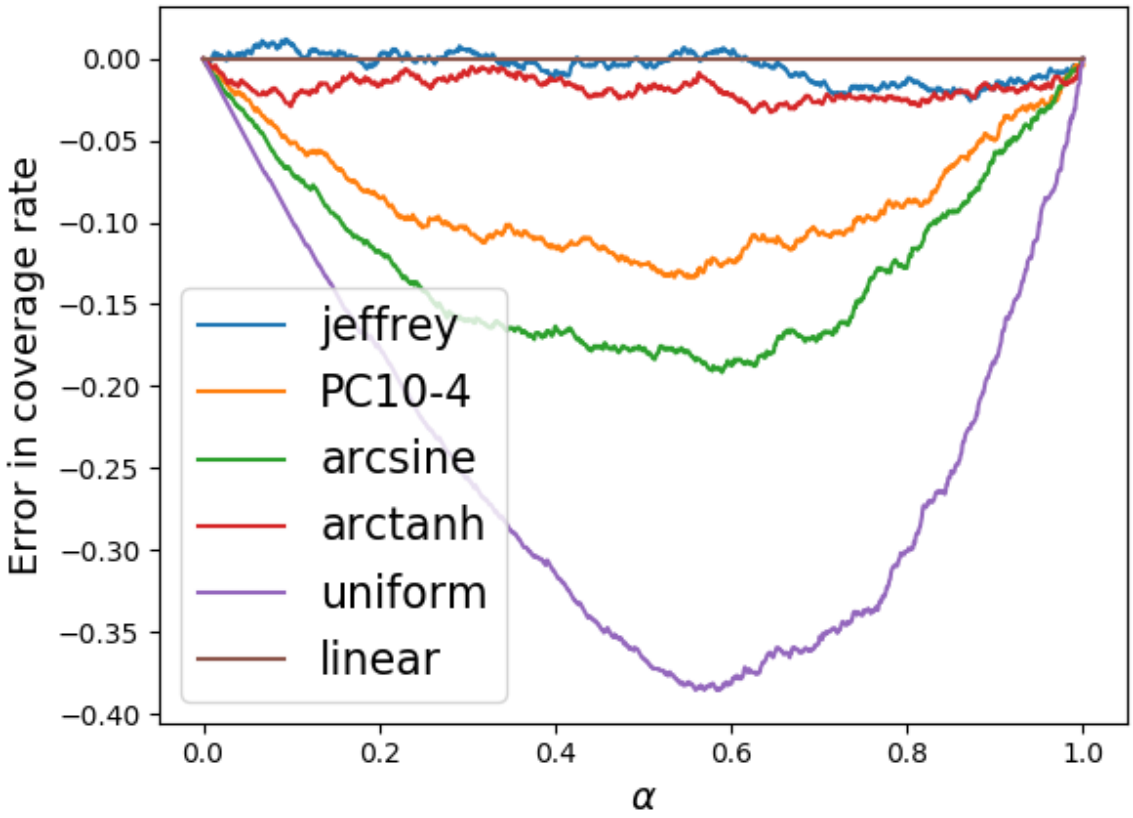


Figure 31: Error in frequentist convergence as a function of the levels α for various posterior distributions under $n = 3$ data points sampled for correlation $\rho = 0.8$. The error is calcualted as the difference between frequentist coverage and level of one-sided interval estimators.

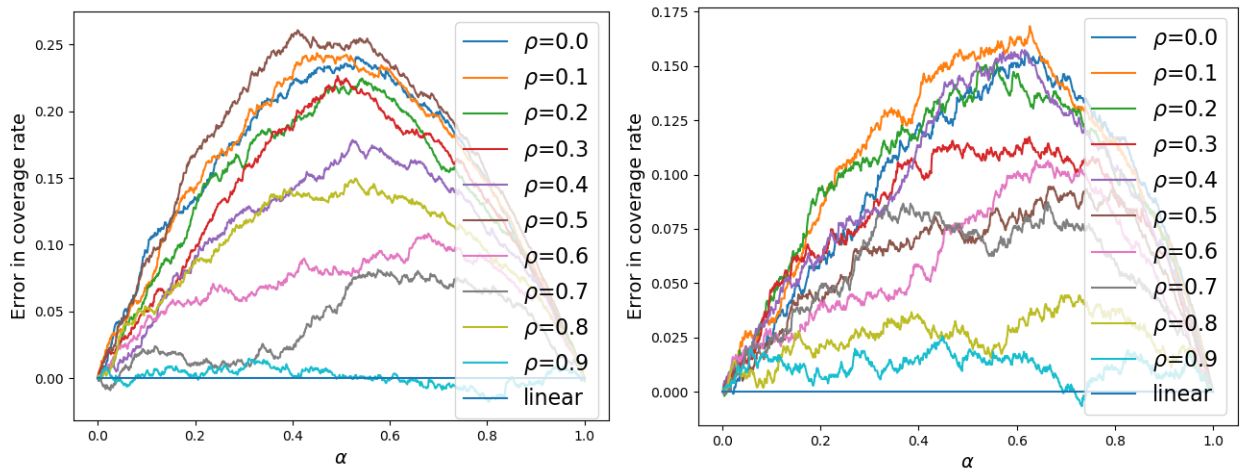


Figure 32: Figures of the error in frequentist convergence as a function of the levels α for two GFDs using sufficient statistics. The figures are for $n = 3$ data points using the 2-norm (left) and infininty-norm (right). The error is calcualted as the difference between frequentist coverage and level of the interval estimator after 1000 simulations.

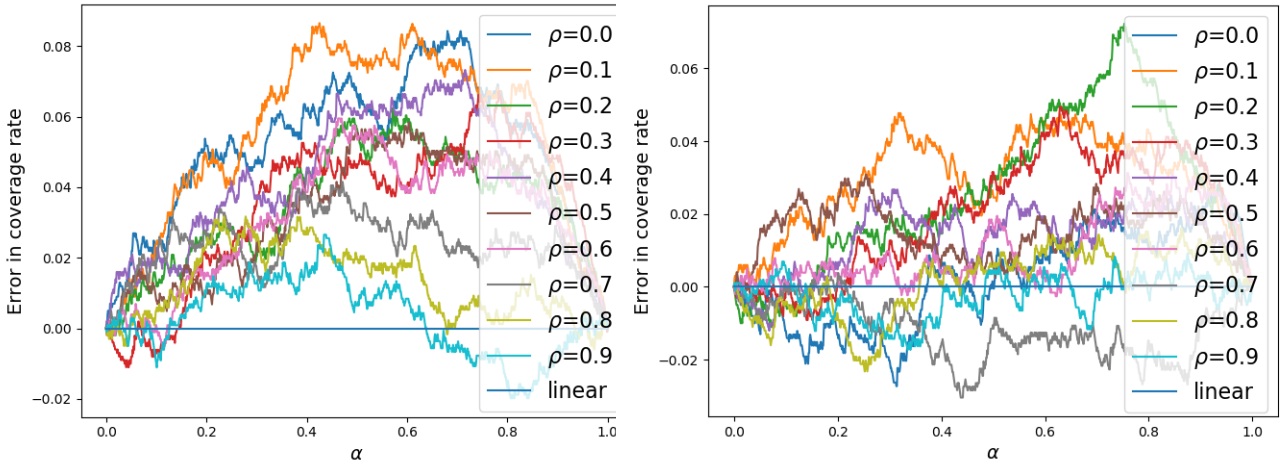


Figure 33: Figures of the error in frequentist convergence as a function of the levels α for two GFDs using sufficient statistics. The figures are for $n = 10$ data points using the 2-norm (left) and infininty-norm (right). The error is calcualted as the difference between frequentist coverage and level of the interval estimator after 1000 simulations.

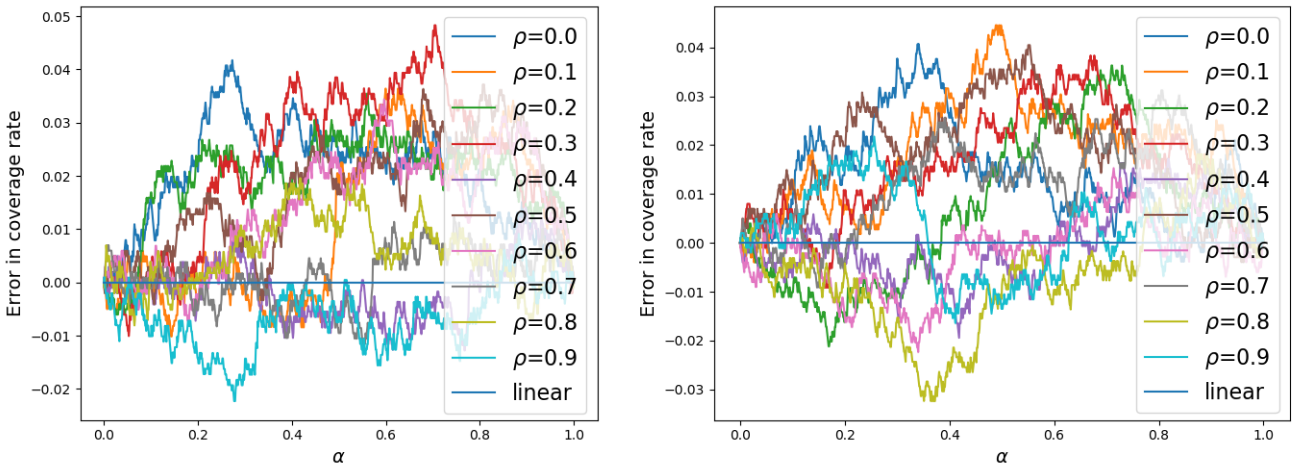


Figure 34: Figures of the error in frequentist convergence as a function of the levels α for two GFDs using sufficient statistics. The figures are for $n = 20$ data points using the 2-norm (left) and infininty-norm (right). The error is calcualted as the difference between frequentist coverage and level of the interval estimator after 1000 simulations.

4.5.3 Comparing confidence distributions

As for the point estimators, the confidence distributions will be compared with respect to risk. Risk is calculated using simulations as presented earlier. Each risk will be presented as a function of the true correlation. As none of the Bayesian distributions are confidence distributions, they will not be included. The fiducial distribution will however be relevant as they can create confidence intervals. Each of the risks are presented for each of the four loss functions

1. MSE
2. z-MSE
3. Fisher information metric
4. Kullback-Leibler divergence.

Each figure have a different data size varying from 3, 10 and 20 points.

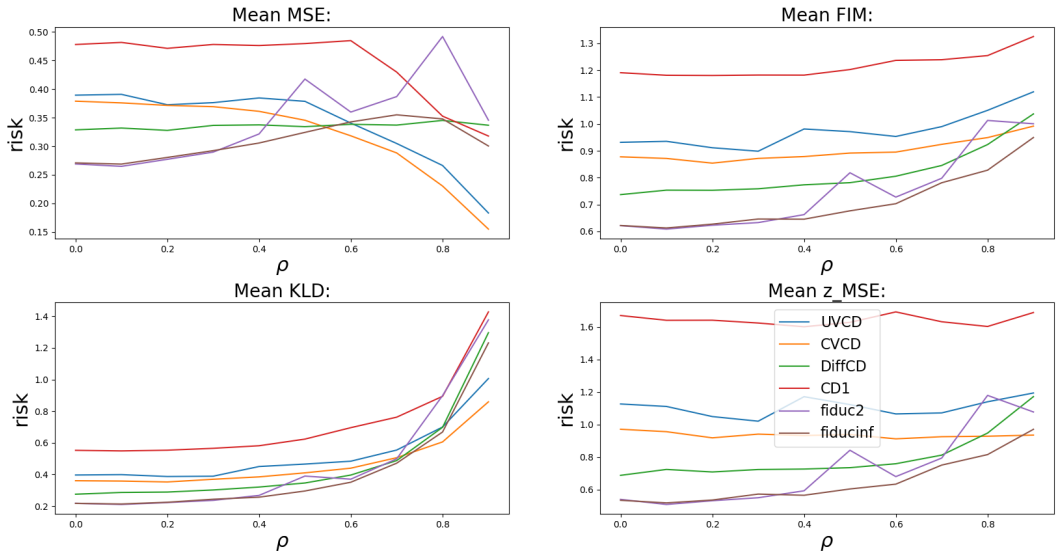


Figure 35: Four plots of the total risks of the five exact confidence distributions and the two fiducial distributions with $n = 3$ data points. Each plot has a different risk based on the loss function

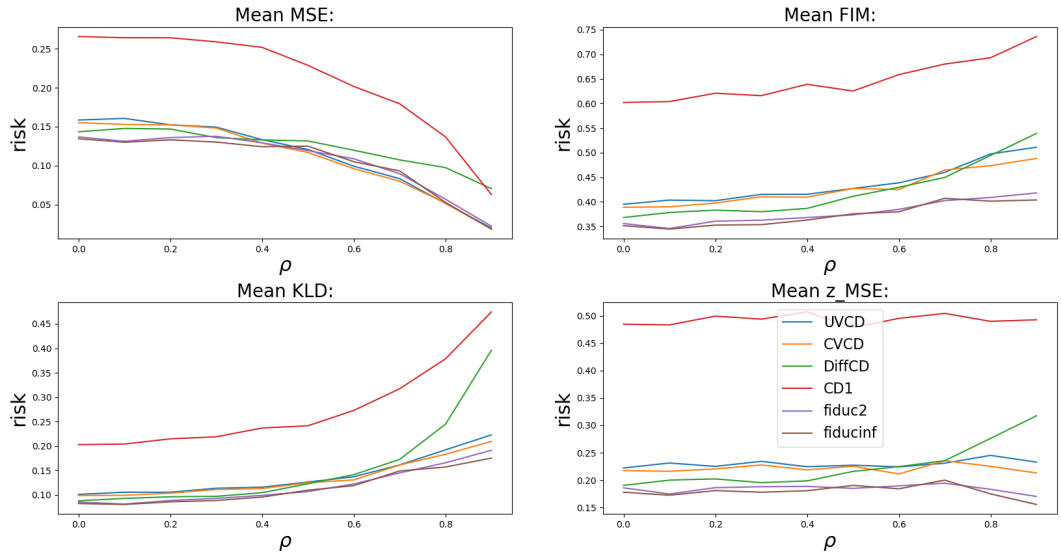


Figure 36: Four plots of the total risks of the five exact confidence distributions and the two fiducial distributions with $n = 10$ data points.

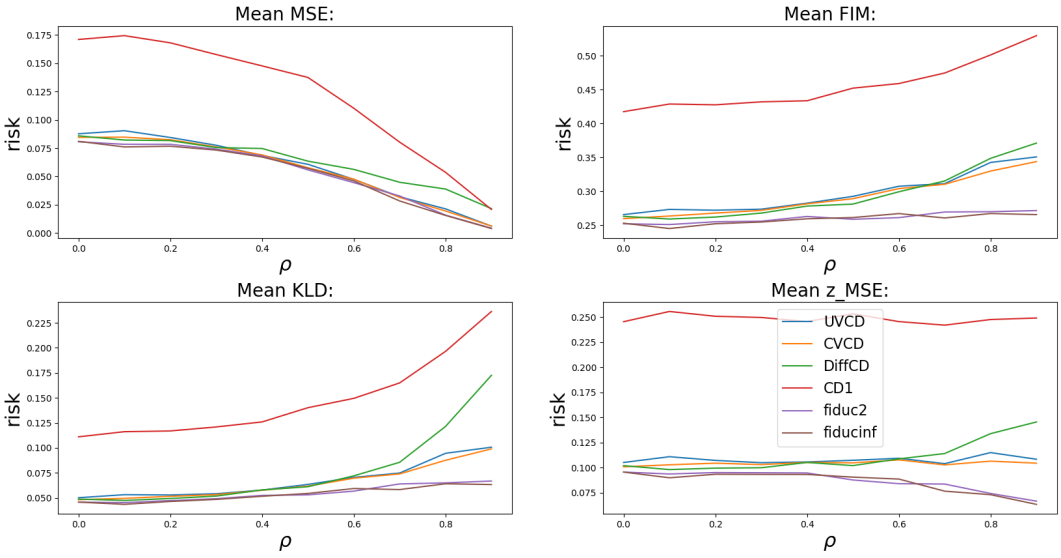


Figure 37: Four plots of the total risks of the five exact confidence distributions and the two fiducial distributions with $n = 20$ data points.

5 Discussion

The following chapter will discuss the results of the data analysis. Any conclusion is based on multiple types of visualizations, where the most important are shown in the Results part of the Data Analysis chapter.

5.1 Performance of point estimators

Each Bayesian point estimator is defined by the choice of loss function and prior. If either the loss function or prior is chosen, then it is possible to order the estimators by risk. It so happens that this order is independent of either the loss or the prior chosen. For instance, the posterior mean will always perform better for small correlations than the posterior median if they have the same prior. Another example is that an estimator of the uniform prior will always perform better for small correlations than one for Jeffreys prior if they minimize the same Bayesian risk. The effect of the choice of loss function and prior can therefore be discussed separately.

5.1.1 Performance of the Bayesian point estimators

The first thing to note about the Bayesian point estimators, are that the MAP performs the poorest. Despite it performing well for large correlations, the loss is too significant for most other correlations. Therefore, the MAP will not be discussed further as a reasonable candidate.

Of the remaining Bayesian point estimators, the properties are very similar. The biggest difference is which estimators that favors larger or smaller correlations. The posterior median and $\hat{\rho}_{FI2}$ are almost indistinguishable. Figure 28 illustrates the similarities. Whether they are actually equal has not been proven. Either way, as their characteristics are almost identical, the discussion will only focus on the posterior median after this point. Any comments about the posterior median should apply for the FI2 estimator as well.

The remaining two estimators to compare are the posterior mean and the posterior median. In section 3.10.6, it is argued that the posterior mean is never larger in size than the posterior median. As a result of this property, the spread around 0 of the posterior mean is also smaller than the posterior median for any correlation. This implies that for smaller correlations, the average loss of the mean is smaller than that of the medians. This can be seen in the risks of figure 23. An important question is then for how large portion of the correlations will the mean perform the best? From the results of the data analysis, the mean will perform the best until about $\rho = 0.6$ for $n = 3$ data points and $\rho = 0.4$ for $n = 10$ data points. In the latter case, the difference in risk is sufficiently smaller. The risks of the two types of estimators are also very similar for $\rho \in (0.6, 0.7)$ when $n = 3$.

Interestingly, the mean has almost always smaller average Kullback-Leibler divergence. An interpretation is that it is almost always better to approximate the true model with a model that underestimates the correlation. Using this logic, the posterior mean is a better choice than the posterior median.

5.1.2 Discussing the choice of prior

Similar arguments as in section 5.1.1 can be used for the choice of prior distributions. The main difference between each prior is how much they weigh larger correlations. This ordering can be seen in figure 6. For smaller correlations, the uniform prior performs best, followed by arcsine, PC and finally Jeffreys prior. Performance is opposite for larger correlations. Figure 27 exemplifies the difference of the prior choice. The example is for the posterior mean, however, the same holds for all other types of estimator. It also highlights how the differences in the estimators decrease as the data size increases.

Similar to the types of point estimators, the Kullback-Leibler divergence prefers smaller correlations. The uniform prior will always perform best in that regard. Using the Kullback-Leibler divergence, the best choice of prior is therefore the uniform prior.

5.1.3 Comparing the Bayesian and frequentist estimators

Finally, the frequentist estimators will be compared to the Bayesian. The Bayesian estimators will generally perform better. Keep in mind that the MLE is equal to the MAP for the uniform prior. As noted in section 5.1.1, the MAP performs poorly. For the same reason, the MLE also performs poorly. As seen in figure 22, the performance of the empirical correlation is similar to the MLE. This will also imply that the empirical correlation will perform poorly. The remaining frequentist estimator is then the empirical correlation with variance 1.

The empirical correlation with variance 1 does compare to the Bayesian estimators for small correlations. This can be seen in figure 25. In the figure, it is compared to posterior mean with uniform prior and posterior median with Jeffreys prior. These two estimators are supposed to be the opposite limits of the Bayesian estimators, without MAP. From the figure, it is possible to argue that the special empirical correlation can be used as an alternative to other Bayesian estimators. However, the performance does deteriorate significantly for larger correlation. Additionally, it is categorically worse than the posterior mean with uniform prior.

If the sample size is 10, the above arguments still hold. In that case, the empirical correlation with variance 1 is actually outperformed by the posterior median with Jeffreys prior. On the other hand, the MLE will perform significantly better. Despite this, it will be categorically outperformed by the posterior median with Jeffreys prior. The Bayesian estimators will therefore outperform the frequentist for 10 data points, however the difference is not large in many cases.

5.1.4 Final comments on point estimators

All Bayesian estimators, except the MAP, outperform or perform as well as the frequentist estimators. While I do believe the posterior mean with uniform prior is the best choice, other choices of Bayesian estimators can perform better. If one assumes that the correlation is large, another estimator might be better without significant loss. Keep in mind, if it is known that the correlation is large, a subjective prior might be a better choice. The question

will come down to preferences. Generally, I will advocate for choosing less complex models over more complex ones. This is not an uncommon position to take. Criterion such as the AIC, which is based on the Kullback-Leibler divergence, are examples of this. I will therefore go back to the argument used earlier, based on the Kullback-Leibler divergence, the posterior mean with uniform prior is the best choice of estimator.

5.2 Performance of distribution estimators

5.2.1 Coverage properties of posterior and fiducial distributions

Figure 29, 30 and 31 shows graphs of the frequentist coverage of one sided interval estimators against their level created for each posterior and fiducial distribution. None of the Bayesian posterior have sufficient frequentist coverage to create one-sided confidence intervals. In general, there seems to be no consistent method for creating confidence intervals across all true correlations. Out of the five choices of priors in this thesis, Jeffreys prior is the one performing the best in terms of coverage. Even at only 3 data points, the posterior will resemble a CD as the greatest error in coverage based on the simulations is 5%. Despite this, it is clear that Jeffreys prior does not give a confidence distribution.

The fiducial distribution are similar in the sense that they do not satisfy definition 2.7 of a confidence distribution. However, it seems to be possible to create two-sided confidence intervals of all levels consistently using the fiducial distribution. These intervals will have a frequentist coverage far greater than the level of the intervals. This can be a sign that the intervals are not as efficient as they could be if they were exact. As the results are based on only 1000 simulations, there might be slight inaccuracies. As it is not certain the coverage is always sufficient, as seen in figure 4.5.2, it does however seem like it. More simulations should be used to be make a stronger conclusion. The distributions will however be seen as having sufficient coverage. As the data size increases, it seems as though the two-sided frequentist coverage is converging towards the assumed level.

When studying the fiducial distribution, there is a possible need to expand the definition of a confidence distribution posed in definition 2.7. While precise, it is also a very narrow definition. Taraldsen 2021 expresses in Definition 1 an expansion of the confidence distribution that will include all distributions that can generate confidence intervals consistently, even if the coverage does not span $(0, 1)$. Such a definition would cover the fiducial distributions here. Definition 2.7 is however important in it's own regard. It will describe a class of distributions that have a greater flexibility in creating confidence intervals.

5.2.2 Comparing confidence distributions

When studying the confidence distributions, the fiducial distributions will be included as well. This is due to their ability to produce confidence interval consistently, even though they are not confidence distribution wrt. definition 2.7.

The risk of the confidence distribution will be used to order them from best to worst performing out of the bunch. As the risk varies with the true correlation, the ordering will

consider the overall performance. It is possible to combine all the risks for different true parameters into a collective risk. The two approaches of minimizing maximum risk, (2), and average weighted risk, (3), can be applied. An issue is that the risk for all true correlations is not available, only at some points where the risk is simulated. Additionally, there are no clear choice for weights to use. The focus will therefore be on a more overall evaluation of the risk.

The risks are seen in figures 35, 36 and 37. With very small data sizes, the diff CD will perform better for most true correlations. However, as the correlation increases, it's performance will significantly deteriorate. In fact, as the data size increases, the ill-performance of the distribution increases. The benefits of using the diff CD will also be reduced as the data size increases. It can therefore be useful in very small data samples, but if the data size increases, other CDs are more reliable.

For the remaining CDs, the UVCD and CVCD is uniformally better than CD1 with a significant margin. Amongst the former two, UVCD is uniformally outperformed by CVCD. This is not as surprising as CVCD utilizes additional information about the data. The difference is significantly reduced when the data size increases. For $n = 10$ and $n = 20$ data points, they are almost interchangeable. In terms of the densities of the distributions, it is clear that they generally are very similar. However, in the situations that they differ, the UVCD performs better than CVCD. CD1 is the distribution that performs the worst of them all. It is possible that this is a sign that the general method in theorem 3.14 does not perform as well as the more specific theorem 3.8.

The risk of both fiducial distributions behaves very similarly to the diffCD. In fact, they are slightly better. This would imply that they are a good source of confidence intervals. One should remember the limitations of the interval estimators, as they can only be two-sided with equal probability on each side. These interval will also have a too large coverage compared to the level of the intervals. As the data size increases to $n = 10$ and $n = 20$, they will perform similarly to both CVCD and the UVCD. If fact, the fiducial distributions seems to perform the best. The difference is not large and it might be better to choose the CVCD as a safer choice due to it's beneficial properties. The fiducial distributions does however look like good candidates. However, one should also run larger simulations for the risk of the fiducial distributions to ensure that the coverage is sufficient for all true correlations.

One would assume that there exists an exact CD that can outperform CVCD as it does not use all the available information about the model. However, the distribution is based on the minimal sufficient statistics, which could imply that the improvement is minimal. Despite this, it would still be of interest to find other choices of g in corollary 3.10.1 such that a better performing confidence distribution can be created. This is mainly important in cases with a small sample size. If the sample size is larger the advantage of knowing the variance is reduced as the information from the data is sufficient. In that case, CVCD is a good candidate for a confidence distribution because of its properties. It is a unimodal distribution and symmetric under the transformation $z(\rho) = \operatorname{arctanh}(\rho)$. In terms of the expanded model with common unknown variance, it is also a marginal posterior distribution for the correlation using the prior $\frac{1}{\sigma^2(1-\rho^2)}$, see page 47. This confidence distribution is

therefore a good choice for uncertainty quantification even without the assumption that the variances are 1.

6 Conclusion

In terms of point estimation, multiple objective priors and Bayesian estimators have been introduced. The analysis showed that for small data sizes these will generally out-perform conventional frequentist estimators under multiple choices of loss. Out of the Bayesian estimators, the posterior mean with uniform prior is recommended for small sample sizes. Other choices of Bayesian estimators are possible. These can improve the loss for larger correlations, but perform worse for smaller correlations. The maximum a posteriore estimator is the only choice that is not recommended.

In terms of uncertainty quantification, the most consistent confidence distribution is the distribution named CVCD. When there is a very small data size, n=3, then the diff CD can be used instead. There is however a great potential for creating confidence distributions using theorem 3.7 and theorem 3.13, which can result in a more powerful distribution than CVCD. As the data size increases it is assumed that much improvement on CVCD is unlikely as the data will contain sufficient information.

A Appendix

A.1 Data sets for visualization

Data set 1: $n = 3$, $S_1 = 9.83$, $S_2 = 9.01$

	X	Y
1	-1.73	-0.41
2	-0.85	0.57
3	2.28	-0.01

Data set 2: $n = 3$, $S_1 = 7.16$, $S_2 = 3.71$

	X	Y
1	0.20	1.08
2	-0.39	0.35
3	-1.95	-0.40

Data set 3: $n = 3$, $S_1 = 11.87$, $S_2 = 3.44$

	X	Y
1	1.01	1.19
2	-0.74	-0.23
3	2.12	0.35

Data set 4: $n = 3$, $S_1 = 6.97$, $S_2 = 0.61$

	X	Y
1	0.98	0.96
2	-0.11	0.67
3	0.87	0.82

A.2 Data set for results

The full results will not be presented here as there is too much information. The results for point estimation can be found in https://github.com/OlavHel/Prosjektoppgave and the results for the distribution estimators can be found in https://github.com/OlavHel/Masteroppgave.

For the point esitmators, the results can be visualized using the file plotting_results.py. For the distribution esitmators, the two files plot_risks.py and view_CD_samples.py can be used to visualize the risks and the coverage respectively.

A.3 Code

The code is separated into two "projects" First, the code used for the point estimation can be found on github under the url: https://github.com/OlavHel/Prosjektoppgave. This was created for the project thesis in the fall semester. It has not been updated since then.

The code for distribution estimation can be found in https://github.com/OlavHel/Masteroppgave. This has been created for the master thesis. There are some overlap as certain files for the point estimation were reused for the distribution estimation. All files found in the code for master thesis will be the newest versions of any file.

Because of a lack of time, no code will be put here. The most important files to study for the distribution estimators are the files MCMC_test2.py and simulate_CD.py for the simulations. The file test_region.py will also include a function for simulating from any CD defined by theorem 3.7 with APF from corollary 3.10.1.

References

Berger, James O., José M. Bernardo, and Dongchu Sun (2009). "The formal definition of reference priors". In: *The Annals of Statistics* 37.2, pp. 905 –938. DOI: 10.1214/07-AOS587. URL: https://doi.org/10.1214/07-AOS587.

Bernardo, Jose (Dec. 2005). "Reference Analysis". In: *Handbook of Statistics* 25. DOI: 10.1016/S0169-7161(05)25002-2.

Beta function Calculator. URL: https://keisan.casio.com/exec/system/1180573394 (visited on 06/23/2021).

Bioche, Christele and Pierre Druilhet (Aug. 2016). "Approximation of improper priors". In: *Bernoulli* 22.3, pp. 1709–1728. DOI: 10.3150/15-BEJ708. URL: https://doi.org/10.3150/15-BEJ708.

BIPM. *BIPM homepage*. URL: https://www.bipm.org/en/home.

Bregman divergence. https://en.wikipedia.org/wiki/Bregman_divergence#Properties. URL: https://en.wikipedia.org/wiki/Bregman_divergence#Properties.

Brockwell, Peter J. and Richard A. Davis (2016). *Introduction to Time Series and Forecasting*. 3rd ed. Springer Texts in Statistics. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-29854-2.

Casella, George and Roger L. Berger (2002). *Statistical Inference*. Duxbury, Thomson Learning.

Consonni, Guido et al. (June 2018). "Prior Distributions for Objective Bayesian Analysis". In: *Bayesian Anal.* 13.2, pp. 627–679. DOI: 10.1214/18-BA1103. URL: https://doi.org/10.1214/18-BA1103.

Exponential family. https://en.wikipedia.org/wiki/Exponential_family#Relative_entropy. URL: https://en.wikipedia.org/wiki/Exponential_family#Relative_entropy.

Fosdick, Bailey K. and Michael D. Perlman (2016). "Variance-stabilizing and Confidence-stabilizing Transformations for the Normal Correlation Coefficient with Known Variances". In: *Communications in Statistics - Simulation and Computation* 45.6, pp. 1918–1935. DOI: 10.1080/03610918.2014.882948. eprint: https://doi.org/10.1080/03610918.2014.882948. URL: https://doi.org/10.1080/03610918.2014.882948.

Fosdick, Bailey K. and Adrian E. Raftery (2012). "Estimating the Correlation in Bivariate Normal Data With Known Variances and Small Sample Sizes". In: *The American Statistician* 66.1. PMID: 23378667, pp. 34–41. DOI: 10.1080/00031305.2012.676329. eprint: https://doi.org/10.1080/00031305.2012.676329. URL: https://doi.org/10.1080/00031305.2012.676329.

Gamma function Calculator. URL: https://keisan.casio.com/exec/system/1180573444 (visited on 06/23/2021).

Gradshteyn, I.S and I.M Ryzhik (2007). *Table of integrals, series and producs*. 7th ed. Elsevier Academic Press, p. 348.

Hannig, Jan (2009). "On Generalized Fiducial Inference". In: *Statistica Sinica* 19.19, pp. 491–544. URL: https://hannig.cloudapps.unc.edu/publications/Hannig2009.pdf.

Hannig, Jan et al. (2016). "Generalized Fiducial Inference: A Review and New Results". In: *Journal of the American Statistical Association* 111.515, pp. 1346–1361. DOI: 10.

1080/01621459.2016.1165102. eprint: https://doi.org/10.1080/01621459.2016.1165102. URL: https://doi.org/10.1080/01621459.2016.1165102.

JCGM (2008a). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*.

— (2008b). *Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method*.

Karr, Alan F. (1993). *Probabiliy*. Springer, New York, NY. DOI: https://doi.org/10.1007/978-1-4612-0891-4.

Lehmann, Erich L. and George Casella (1998). *Theory of Point Estimation*. 2nd ed. Springer Texts in Statistics. Springer, New York, NY. DOI: https://doi.org/10.1007/b98854.

Marsden, Jerrold and Alan Weinstein (1985a). *Calculus I*. 2nd ed. Undergraduate Texts in Mathematics. Springer, New York, NY. DOI: https://doi.org/10.1007/978-1-4612-5024-1.

— (1985b). *Calculus II*. 2nd ed. Undergraduate Texts in Mathematics. Springer, New York, NY. DOI: https://doi.org/10.1007/978-1-4612-5026-5.

Robert, Christian P. (1996). "Intrinsic losses". In: *Theory and Decision* 40.2, pp. 191–214. ISSN: 1573-7187. DOI: 10.1007/BF00133173. URL: https://doi.org/10.1007/BF00133173.

Schervish, Mark J. (1995). *Theory of Statistics*. 1st ed. Springer Texts in Statistics. Springer, New York, NY. DOI: https://doi.org/10.1007/978-1-4612-4250-5.

Schweder, Tore and Nils Lid Hjort (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/CBO9781139046671.

Shao, Jun (2003). *Mathematical Statistics*. 2nd ed. Springer Texts in Statistics. Springer, New York, NY. DOI: https://doi.org/10.1007/b97553.

Simpson, Daniel et al. (Feb. 2017). "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors". In: *Statist. Sci.* 32.1, pp. 1–28. DOI: 10.1214/16-STS576. URL: https://doi.org/10.1214/16-STS576.

Singh, Kesar, Minge Xie, and William E. Strawderman (2005). "Combining information from independent sources through confidence distributions". In: *The Annals of Statistics* 33.1, pp. 159 –183. DOI: 10.1214/009053604000001084. URL: https://doi.org/10.1214/009053604000001084.

Stover, Christopher. *Little-O Notation*. https://mathworld.wolfram.com/Little-ONotation.html.

Taraldsen, Gunnar (2020). *Confidence in Correlation*. Tech. rep. http://dx.doi.org/10.13140/RG.2.2.23673.4 DOI: 10.13140/RG.2.2.23673.49769.

— (2021). *Joint Confidence Distribution*. Tech. rep. DOI: 10.13140/RG.2.2.33079.85920.

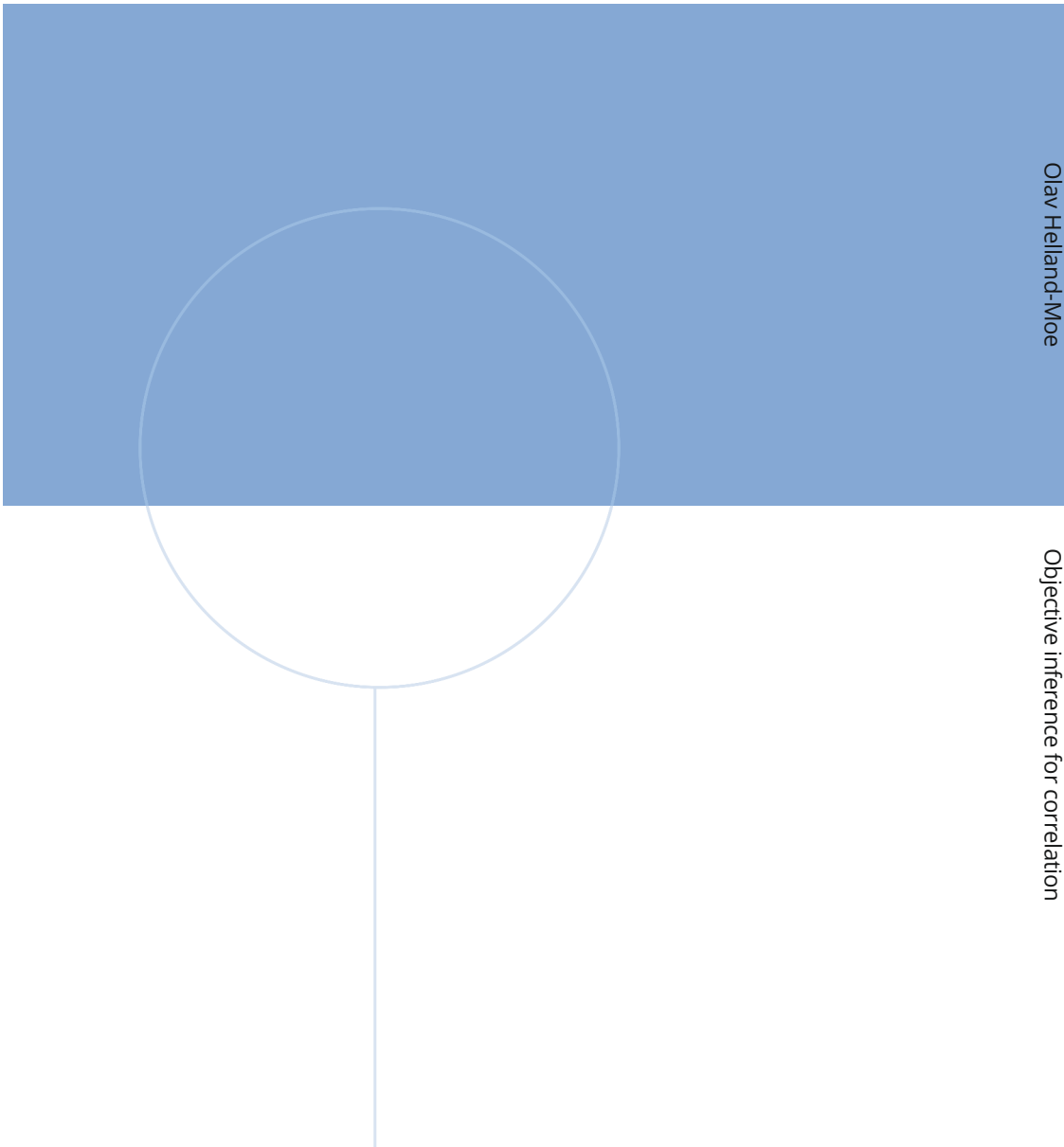
Taraldsen, Gunnar, Jarle Tufto, and Bo H. Lindqvist (2018). *Statistics with improper posteriors*. arXiv: 1812.01314 [math.ST].

Taylor, Stephen (2019). "Financial Return Distributions Using the Fisher Information Metric". In: *Entropy* 21, p. 110. DOI: https://doi.org/10.3390/e21020110. URL: https://www.mdpi.com/1099-4300/21/2/110#framed_div_cited_count.

Weisstein, Eric. W. *Beta Distribution*. URL: https://mathworld.wolfram.com/BetaDistribution.html.

Weisstein, Eric W. *Beta prime distribution*. URL: https://mathworld.wolfram.com/BetaPrimeDistribution.html.

Weisstein, Eric W. *Gamma distribution*. URL: https://mathworld.wolfram.com/GammaDistribution.html.



Olav Helland-Moe

Objective inference for correlation

NTNU
Norwegian University of
Science and Technology