On the Estimation of the Polychoric Correlation Coefficient via Markov Chain Monte Carlo Methods

by

Oscar Lorenzo Olvera Astivia

B.A., Univeristy of the Fraser Valley, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate Studies

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2013

Abstract

Bayesian statistics is an alternative approach to traditional frequentist statistics that is rapidly gaining adherents across different scientific fields. Although initially only accessible to statisticians or mathematically-sophisticated data analysts, advances in modern computational power are helping to make this new paradigm approachable to the everyday researcher and this dissemination is helping open doors to problems that have remained unsolvable or whose solution was extremely complicated through the use of classical statistics. In spite of this, many researchers in the behavioural or educational sciences are either unaware of this new approach or just vaguely familiar with some of its basic tenets. The primary purpose of this thesis is to take a well-known problem in psychometrics, the estimation of the polychoric correlation coefficient, and solve it using Bayesian statistics through the method developed by Albert (1992). Through the use of computer simulations this method is compared to traditional maximum likelihood estimation across various sample sizes, skewness levels and numbers of discretisation points for the latent variable, highlighting the cases where the Bayesian approach is superior, inferior or equally effective to the maximum likelihood approach. Another issue that is investigated is a sensitivity analysis of sorts of the prior probability distributions where a skewed (bivariate log-normal) and symmetric (bivariate normal) priors are used to calculate the polychoric correlation coefficient when feeding them data with varying degrees of skewness, helping demonstrate to the reader how does changing the prior distribution for certain kinds of data helps or hinders the estimation process. The most important results of these studies are discussed as well as future implications for the use of Bayesian statistics in psychometrics.

Table of contents

List of tables

List of figures

Acknowledgements

This thesis was the collaborative work of many people who, either directly or indirectly, contributed in various ways to the final product that the reader now has in his or her hands. I would like to thank my advisor Dr. Bruno D. Zumbo who in some obscure and still incomprehensible way managed to change the way I see myself as a scholar and a researcher. With love, patience and understanding he helped me learn to see problems, both inside and outside of academia, for what they really are and not for what they appear to be. I would also like to thank my committee members, Dr. Amery Wu and Dr. Nand Kishor, for taking time from their busy schedules to review this very humble piece of work. I really hope there is at least something remotely interesting in these pages that you can take. Dr. Anita Hubley also has a very special place here for being the very first person to introduce me to the life of a graduate student. I am very happy to have shared these past couple of years with my dear friend Benjamin Shear, my partner-in-crime Tavinder Ark and my dearest friends Yan, Eric and Wen. I don't think you understand just how much you helped me become a better scholar and a better person. On a very personal note I would also like to thank my mother who has always supported me, even when she knew I was making the wrong decisions, and to my husband Alejandro, for teaching me how to work in spite of adversity. Last and not least, I would also like to thank my dear online friends from the online board TalkStats: Dason, Lazar, Dragan, Link, TheEcologist, trinker, Jake, bryangoodrich, GretaGarbo, noetsi, vinux, BGM, victorxstc, hlsmith.. I think those are all "the regulars", right? Bouncing around ideas with you was one of the most awesome experiences I had during my graduate life and I hope you will still be there once I begin my PhD. From spunky to the world I would like to say 'thank you'.

1. Introduction

1.1 Opening remarks

When students in introductory methodology courses are first exposed to the basic ideas of statistical inference (under the Neyman-Pearson paradigm), an all-too-familiar situation occurs the moment that confidence intervals are introduced. In spite of careful explanations, worked-out textbook examples and constant exhortations from instructors and APA taskforces, it is not unusual for a misconception to proliferate regarding their appropriate interpretation (Anderson, Burnham & Thompson, 2000; Thompson, 2002; Wilkinson & TFSI, 1999). Over and over again, claims such as "there is a 95% chance that the (insert the name of your parameter of choice) falls within such and such bounds" get uttered as a way to make sense of what these numbers mean. And over and over again instructors and methodologists have to fight their way into communicating the *correct* interpretation: that if the experiment or study in question were to be repeated under the same circumstances (i.e. same treatment to samples coming from the same population) over and over again to infinity, and 95% confidence intervals were calculated each time, the researcher can expect that 95% of those confidence intervals will contain the true population mean, regression coefficient, factor loading or the parameter estimate needed. Although this explanation is mathematically correct and directly reflects the traditional paradigm of null hypothesis testing, any researcher or data analyst, at any given moment, can raise a very legitimate question: what about *my* 95% confidence interval? What about the uncertainty in *my* dataset and not in these hypothetical infinite replications? It may come as a surprise to some that, unfortunately, very little information can be extracted from these intervals to answer such questions (Agostini, 2003; Berger & Sellke,1987; Gill, 1999). The classical or frequentist viewpoint requires, by its own definition of probability, that such long-run frequency of events must be present (Spanos, 2011). Most of the statistical methods used in everyday research adhere to this interpretation which, up to a certain point, limits the kind of inferences that can be made from the data (Gill, 1999; Lee, 2011). But it

does not have to be this way. An alternative view of probability, with alternative methods of estimation, is beginning to challenge the current position held by classical statistics as the sole approach to data analysis, paving the way for what could become the next new paradigm to guide the process from data collection and knowledge creation.

Bayesian statistics provide a different light under which data can be analysed and parameters can be estimated. Although the underlying technical extensions are very similar to those of classical or frequentist statistics (both have instances of the general/generalised linear model, a notion of statistical inference, etc.) several important differences have contributed to the popularisation and use of Bayesian methods, starting with their conceptualisation of probability. Whereas the relative frequency of events is central in classical statistics, Bayesian statistics relies on a definition of probability that has more to do with states of knowledge or degree of belief (Bunge, 2012; Crovelli, 2011). Most people used to the traditional frequentist definition feel cautious when words such as 'belief' get used in the context of scientific research because of the aura of subjectivity conveyed with it. It is, in these cases, that emphasis should be made on the fact that *subjective* is not the same as *arbitrary,* and that to assume that the scientific (or any human) endeavour is free from people's own intellectual background is indeed quite naïve (Karni, 2011). Mathematically speaking, it has been shown that the conceptualisation of probability as degree of belief is consistent within an axiomatic framework of probability theory (cf. De Finetti, 1974; Cox, 1946; Jaynes, 1957). De Finetti's coherence argument is one of the first articulations of the issue, but later extensions have been made through the use of Jayne's Maximum Entropy principle or Cox's logical consistency reasoning. Even modern results of probability based on measure-theoretic approaches where sample spaces are defined over sigma-algebras have shown to be consistent under the subjective definition of probability (Strzalecki, 2011).

Until recently, most of the advances in Bayesian statistics had been somewhat out of reach for the applied researcher, particularly in the social sciences. As it will be elaborated in subsequent sections,

Bayesian methods of estimation depend upon the integration of probability densities in multiple dimensions whose solution is difficult to find. Advances in computer power have allowed for the development of a family of methods known under the umbrella term of Markov Chain Monte Carlo (MCMC) which have contributed to their dissemination, but there still exist a variety of areas where data analysis is routinely performed but the Bayesian perspective is either ignored or misunderstood.

The social sciences have recently seen a surge both in the development and application of Bayesian methods (Diaconis, 2009; Gelman & Shalizi, 2011). As statistical models increase their complexity (particularly in the realm of latent variables or missing data), the need for a more flexible framework to accommodate them increases as well, opening the doors for the Bayesian paradigm to help inform both the data-analytic and theory-testing process in these areas of knowledge. Item Response Theory (IRT) and Latent Class (LC) models have received particular attention due to the complicated nature of their estimation but, unfortunately, the 'Bayesian Revolution' has not yet gained much traction among psychological and educational researchers when compared to what is being seen now in areas like biology or economics (Kruschke, Aguinis & Joo, 2012).

1.2 Purpose and structure of the thesis

Given the fact that the Bayesian paradigm is still mostly unfamiliar (or completely ignored) by a considerable number of people in the social/behavioural/health sciences and that many of the leading theorists in Bayesian methods only consider problems pertaining to the social sciences "in passing" (so that they provide a general overview of the problem and the solution but no further development of the ideas is done) this thesis explores a very common problem in psychometrics: the estimation of the correlation of ordinal data through the use of the polychoric correlation coefficient. This particular statistic has been widely researched from the frequentist perspective using the method of maximum likelihood, testing its estimation under a wide variety of different conditions to see the impact that each

of them have on the quality of the estimates (conditions such as the number of cutpoints in the latent variable, the skewness of both the observed and latent variables, etc.). Until now, not much work has been done to look at it from a Bayesian perspective though. The primary purpose of this thesis is to extend the work of Albert (1992) who proposed both a flexible mathematical framework to make the estimation of the polychoric correlation tractable and a general overview of the design of a Gibbs sampler that could be used to perform the necessary calculations. Albert himself only shows two brief examples as a demonstration of how his approach could be implemented, but no systematic study is done to test whether his approach holds in the various circumstances that applied researchers regularly find themselves in. Choi, Kim, Chen & Dannels (2011) have recently shown that a different form of Bayesian estimation outperforms traditional maximum likelihood in a variety of settings, but their method is restricted only to the case of the bivariate Gaussian distribution. Albert's method will be investigated further and extended to accommodate left-skewed latent distributions in an effort to help document some of the advantages that switching from frequentist to Bayesian methods has, as well as the instances where the choice of estimation is irrelevant or even when the traditional approach is preferred. Although the Bayesian conceptualisation of probability is still controversial, Bayesian estimation methods are slowly moving at the forefront of statistical analysis which helps make the case as for why practitioners in the social sciences would benefit greatly by starting to become familiar with this new approach.

2. Literature review

2.1 An overview of Bayes' theorem and the Bayesian approach to statistics

Consider the following scenario. One finds her or himself walking down the street and happens to find a coin. The coin is a little bit different from what anyone has come across before, probably left by a visitor from another land. If one were to use that coin for a betting game of sorts, how confident can one be that such coin is fair? Perhaps tossing it a few times to test it would be the best option. If say after ten tosses one finds that one side of the coin came up 7 times and the other 3 times, what can be inferred about the coin? What if after 20 tosses the other side came up 19 times and one side only came up once? Regardless of what the results are, one thing that can be pointed out is that each person's beliefs concerning the fairness of the coin are affected by the very act of testing it. They can be reaffirmed or they can be changed, but what is thought about the uncertainty surrounding the coin is influenced by the act of testing it.

From any introductory methods or statistics course students become acquainted only with the frequentist or classical definition of probability, which implies that probability can be understood as the long-run frequency of sampling events taken from the sample space (Nickerson, 2000). In the case of the previously-mentioned coin, the sample space is simply both sides of the coin (the only two possibilities that can arise in the experiment) and the sampling event is each tossing of the coin. If the coin is fair, after a theoretical infinite number of tosses one should conclude that the proportion of obtaining one side of the coin versus the other one is 0.5. Two things stand out from this conception of probability. The first one is the need for the infinite number of tosses. For any finite number of tosses, regardless how many there are, the average frequency will approximate 0.5 but it will never be 0.5. The second one is that nothing is known about the probability of any specific toss. The coin is only fair on average and the probability that heads or tails will come out is 0.5 in the long-run, but when one is about to make a toss, nothing is known about the probability associated with that specific toss.

A Bayesian twist can be given to the same scenario. Assume that the hypothetical coin-tosser picks up the coin, inspects it and asks him/herself: how big is it? What shape does it have? Does it look perfectly circular or is it slightly elongated? Does it have any kinks that would push it to one side or the other when I flip it? All of these aspects are taken into consideration before the person in question formalises her or his beliefs in a single number, the (subjective) probability that the coin is fair. With each flip of the coin, however, this probability (better interpreted as degree of belief) gets updated and changes. If the coin is assumed to be fair and both sides of it come out roughly the same number of times, there is little need for the coin-tosser to make considerable adjustments to the initially-held beliefs. But what if one side consistently shows up more often than the other one? With each new toss the probability of fairness becomes updated such that after any series of tosses, there is a final assessment of belief in the fairness *conditioned on* the evidence gathered through previous coin tosses. The crux of the Bayesian approach to probability relies on the very fact that the probabilities of outcomes become updated with every new piece of evidence that is collected (Chechile, 2011).

Bayesian statistics owes its name to the work of the English mathematician and Presbyterian minister Thomas Bayes (1702-1761). In 1764, his seminal work "An Essay towards solving a Problem in the Doctrine of Chances" was published posthumously by his friend the moral philosopher Richard Price, where Bayes states the first mathematical formulation of inductive reasoning through his celebrated theorem, Bayes' theorem, in which previous probabilities can be combined to obtain an estimate of the probability of a future event (Stephens, Buskirk, Hayward & Martinez Del Rio, 2005). In its simplest form it can be expressed as follows. For any two non-independent events A and B with probabilities P(A), P(B) and conditional probabilities P(A|B) (read, 'probability of A given B') and P(B|A) the following relationship can be stated through Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

The probability of A given B in this case is the product of the probability of B given A times the probability of A divided by the probability of B. Each term that goes in Bayes' theorem receives a name. P(B|A) is called the 'likelihood', P(A) the 'prior probability' (or simply prior) and P(B) is called the 'evidence'. The French mathematician LaPlace is credited with extending many of the results originally proposed by Bayes (including a re-derivation of the theorem which is now used) although he was mostly unaware of the work done by the English reverend.

Consider the following example to try and bring this mathematical expression to a more concrete level:

$$P(Rain|\ Wet\ Grass) = \frac{P(Wet\ Grass\ |\ Rain)P(Rain)}{P(Wet\ Grass)}.$$

So the probability of inferring rain from finding wet grass requires, first, the absolute probability of finding the grass to be wet. If the grass is not wet then it is an oxymoron to ask for the probability of inferring rain from wet grass. Now, from the spectrum of possibilities surrounding finding wet grass, one needs two probabilities: the probability that it rained and a certain degree of belief regarding the inference of finding wet grass *because* it rained. In this specific example, it could be argued that perhaps the front lawn of some house has a roof on top of it so that the probability that the grass gets wet because it rained is very low but it is very high if the sprinklers are on. Finding the grass to be wet, therefore, does not immediately imply that it rained but it remains in the spectrum of possibilities. All of these events with certain degrees of plausibility get combined together to obtain a final probability assessment so that if one opens the door and sees wet grass it is possible to make the inference that rain could be an explanation for that. Bayes' theorem allows to mathematically proceed from observing 'wet grass given rain' to assessing the probability of 'rain given wet grass'.

One of the best examples of the relevance of this process in the social sciences can be found in Cohen's 1994 critique of null hypothesis testing, "The earth is round, $p < .05$" :

"When one tests Ho one is finding the probability that the data (D) could have arisen if Ho were true, P(D|Ho). If that probability is small then it can be concluded that if Ho is true then D is unlikely. Now what really is at issue, what is always the issue is the probability that Ho is true given the data, P(Ho|D), the inverse probability (…) but that is the posterior probability, available only through Bayes Theorem (p.998)."

The Bayesian approach to statistics has always been controversial (MacCallum, Edwards & Cai, 2012). Even from its inception, Thomas Bayes was so reluctant to let his result be known that it had to be published posthumously because of the implications that it had. These 'inverse probabilities', as they were initially referred to, violated some of the first conceptualisations of probability known in history because they provided evidence for an event that could have never happened before. In general, the probability of A given B is different from the probability of B given A, so $P(A|B) \neq P(B|A)$, but through Bayes' theorem it is possible to estimate this quantity. Most of the major figures in mathematical statistics openly condemned this approach, including Fisher who vehemently and consistently criticised it as a theory "founded upon error and must be wholly rejected" (Fisher, 1970, p.10). Edwards (2004) points out that, as an interesting twist of fate, Fisher would come to realise that in order to solve the shortcomings of the kind of inferences that can be derived from null hypothesis testing he would need to use these "inverse probabilities" (and Bayes' theorem). Unwilling to do so, he developed the concept of "fiducial probabilities" for which he was later criticised due to the logical and mathematical errors that went into developing them. Several authors concur that from Fishers final papers it could be seen that he was not only reconsidering his own position towards Bayesian statistics but also that Bayesian inference was probably the *only* kind of statistically-justifiable inference (Barnard, 1987; Dale, 1999; Fienberg, 1997). Although no one questions the validity of Bayes' theorem today, there are still detractors who question the kind of conclusions that be obtained from switching from classical to Bayesian statistics, particularly because of this conceptualisation of probability as degree of belief (Efron, 1986).

2.2 Bayesian estimation

Until recent days, the development of Bayesian statistics had been confined to theoretical advancements and some limited applications due to the inherent difficulty in the calculations involved. In its more general form, Bayes' theorem can be expressed as:

$$P(\theta|D) = \frac{L(D|\theta)P(\theta)}{\int L(D|\theta)P(\theta)}$$

Where θ is a vector of unknown parameters that will be estimated, $L(\theta|D)$ is the likelihood of the parameters given the data and $P(\theta)$ is the prior probability of the parameters. The main problem with Bayesian estimation is concerned with the calculations that take place in the denominator of the expression, the integral. Integrals can be notoriously difficult to evaluate and, in many cases, numerical approximations are the only possible way to evaluate them, which requires a lot of computer power. Modern computers, however, are now capable of doing an impressive amount of calculations in limited time which has opened the doors for Bayesian statistics to become accessible to the everyday researcher.

Casella and Berger (1987) commented once that Bayesian estimation and inference methods could be conceptualised as following three very general steps. The first one asks for the researcher to specify a statistical model and a prior distribution. Next the estimation is done and finally, as a matter of good practice, the results are compared to those obtained through competing models in order to be able to choose the most appropriate ones.

The first step of model building is where Bayesian statistics finds most of its criticisms. In the numerator of Bayes' theorem is the term referred to as the 'prior probability' or, simply, 'the prior'. This term reflects the degree of belief of the researcher concerning the phenomenon under study before the actual collection of data. The priors can be chosen at will and, depending on the model being analysed, a different choice of priors can generate different parameter estimates and, hence, different conclusions (Dempster, 2005). One of the main claims that detractors of Bayesian statistics make is that by including

the subjective influence of the researcher through the use of priors one is likely to bias the conclusions derived from the analysis of the data (Bem, Utts & Johnson, 2011; Berger, 1990). That only through adopting a completely objective position, the "viewpoint from nowhere" described by Leahey (1991), can one safely proceed into the enterprise of creating science. There exist several arguments against such position, both epistemological and statistical which are beyond the scope of this thesis, although a brief exploration of them will help clarify some of the misconceptions that surround Bayesian inference. From an epistemological point of view it is extremely naïve to believe that the researcher has no impact on his or her object of research. Even pure, hard sciences such as physics acknowledge the presence of an observer effect where the very act of measuring something influences it (Matthews, 2011; Monahan & Jill, 2010; Shen, 2009). An electron, for instance, can behave as a particle or as a wave depending on how it is measured. In the domain of the social sciences the context in which participants are measured can greatly influence the kind of responses they will give, even in seemingly objective measures such as blood pressure of electro encephalogram readings.

Statistically speaking, common frequentist analysis does not go without its own subjective assumptions (Risinger, 2012). The main workhorse of estimation in classical statistics, the method of maximum likelihood, requires the data analyst to *subjectively* choose a likelihood function over which the parameter space will be maximised. In the social sciences it is customary for this likelihood to be Gaussian (with a few exceptions such as in logistic regression) and the software usually makes the choice. However, there are still consequences to choosing the wrong likelihood function for the data, which can range from very minor bias to an absolute lack of consistency of the estimates depending on the complexity of the model (Pawitan, 2001). A second statistical drawback of estimation in classical statistics is that there is no way for the data analyst to aid in the estimation process through the use of data from previous research (Agostini, 2003; Lee, 2011). Say, for example, that one is using a psychological scale from some specific subfield of the discipline which is considered as the 'gold

standard', so inferences derived from its scores have been found to be supported by empirical evidence. Say further that, as it happens commonly in the social sciences, one is using this scale on a small sample. Under the Bayesian framework, the researcher can gather information of the previously-available research concerning this scale to help reduce the uncertainty surrounding the estimates of interest, but under the classical framework all one has is the data and nothing more. Since the frequentist definition of probability requires the idea of infinite sampling and re-sampling, all the researcher has when she or he collects data is one instance of those many samples, regardless of how crude her or his sampling techniques were or how few participants he or she was able to gather. There is no way to bring into the estimation previous information to help in the analysis of data. Bayesian estimation, on the other hand, can help reflect this information through the choice of a prior distribution (Berger, 2000). Things as simple as data plots or having estimates of means and standard deviations from the variables can be brought forward into the prior so that the Bayesian estimates can profit from the work being done before by other researchers.

Once a prior and a model have been chosen (choosing a model works in the same way as one would do in classical statistics: regression, ANOVA, etc.) the second step is to move on to the estimation process. Some analytic solutions exist for a limited number of Bayesian model specifications which usually imply using Bayes' theorem alongside well-behaved integrals. The best example would be inferring the proportion of heads that one obtains after a series of coin tosses. Since the proportion p of heads and tails can be any number in the interval between 0 and 1, a suitable distribution with support over this domain is the beta distribution. The beta distribution is unimodal and governed by two parameters (referred to by convention as 'a' and 'b') which help define its skewness and kurtosis. If one works under the assumption that the coin is fair (so $p = 0.5$) then a equals b and the distribution looks like the normal distribution (that is, symmetric, unimodal and bell-shaped). If one believes that the coin is not fair, the skew of the distribution can be moved to the right (so closer to 0 if one thinks the coin is

biased *against* heads) or to the left (so closer to 1 if the bias is assumed to be *favouring* heads). The beta distribution for inferring a binomial proportion (as this problem is usually known) is very well-behaved analytically because once all the integration and calculations are done, the resulting distribution is also part of the beta family. Distributions whose prior and posterior belong to the same family are known as conjugate distributions and many lend themselves to analytical solutions as opposed to numerical approximations (Greenland, 2001).

Unfortunately, the majority of Bayesian models are not amenable to exact solutions and numerical methods have to be invoked. As it was previously mentioned, the number of simultaneous dimensions over which one would need to integrate increases as a function of the number of parameters in the model. Many of these integrals do not even exist in closed-form (so no symbolic formula can be provided for it) which just increases the complications. Bayesian analysis saw its development stumped by this very fact until the mid-1950s when, for the first time, researchers could implement a series of high-powered approximations that have become collectively known as Markov Chain Monte Carlo (MCMC).

MCMC methods are the collective name of a series of computational algorithms that allow the researcher to sample from high-density regions of probability distributions that either exist in multiple dimensions or have very complex forms (Kruschke, 2010). A useful analogy to how MCMC methods work would be that of a frog jumping across lily pads of several different sizes following a set of simple rules. The jump to the first lily pad is made at random from any part of the pond. If the next lily pad is bigger than the one where the frog is on now, then the frog jumps on it, however, if it is not, the frog only jumps on it probabilistically, taking into account the relative size of the next lily pad to the one it is on now. Say the hypothetical frog is on a lily pad with a size of 7 units and the one where it supposed to jump has a size of 5 units. The imaginary frog would use a fair spinner (marked from 0 to 1 to indicate probabilities) and would only jump to the following lily pad if the spinner falls anywhere between 5/7

(so about 0.71) and 1. If the spinner falls below 5/7 then the frog stays where it is and spins again. Although somewhat simplified, the important result that comes out from it is that after enough jumps and spinnings, the frog will visit each and every lily pad *proportionally to its size* so that it spends more time on bigger lily pads and less time on smaller ones. Recognising the analogy that the pond is the probability distribution and the lily pads are intervals on it, the mathematical miracle of MCMC algorithms with random walks is that, eventually, the algorithm visits every point of the probability distribution. If one were to imagine this in the usual normal case, for instance, the algorithm will tend to move away from the tails (where the probability density is very low) towards the middle of the distribution close to the mean (where the probability density is high). However, the tails and the middle of the distribution will be visited proportionally to how probable events are there, reproducing the probability distribution up to an accuracy proportional to how many iterations (or in more formal terminology "samples") are run by the algorithm. It will only reproduce the distribution perfectly if infinite MCMC iterations are taken, just as maximum likelihood will only yield the exact estimate of a parameter under the assumption of an infinitely large sample size. Sampling from this posterior distribution is at the centre for Bayesian inference and Bayesian estimation, since it is from there that parameters and measures of the uncertainty of those parameters are calculated (Martin, Quinn & Park, 2011).

For instance, the mean of the posterior distribution is usually taken as the best parameter estimate, and its standard deviation as a measure of its variability. If more than one parameter is being calculated, their joint posterior distribution contains all the information needed to estimate them, but it becomes very difficult to find it due to its high-dimensional nature. In such cases, it is possible to convert this high-dimensional problem into multiple low-dimensional problems through sampling from the *conditional* form of the posterior distribution, instead of the joint distribution, although there are times in which even this is not possible and much more complex computations are required (Lynch, 2010).

There exist several algorithms that fall under the umbrella term of 'MCMC'. Two of the most popular ones are the Metropolis-Hastings algorithm and Gibbs sampling, which will be employed in this thesis (Gelman, Carlin, Stern, & Rubin, 2004). Casella and George (1992) provide one of the best overviews of Gibbs sampling, starting by acknowledging the fact that sampling from the joint posterior distribution for many parameters can be very complicated. In this case, the joint distribution of the vector of parameters θ is 'broken down' by conditioning across each parameter, so that instead of sampling simultaneously from a single distribution, several samples are taken from several distributions where each parameter is conditioned on all the other ones. If, for instance, one attempts to find the parameter estimates of a bivariate normal distribution with parameters (μ_1, μ_2) for the mean, (σ_1, σ_2) for the variances and covariance σ_{12} , in order to find μ_1 one would have to sample from the marginal distribution $f(\mu_1 |\ \mu_2, \sigma_1, \sigma_2, \sigma_{12})$. Because it is known that any joint distribution is the product of its marginal distributions, under some general conditions Gibbs sampling can reproduce the joint posterior. In spite of the power of this algorithm, there exist certain caveats to its use. The conditional distribution of the parameters must be known in order to be sampled from, for instance. It is also true that because Gibbs sampling rotates across parameters, convergence problems can exist if these are highly autocorrelated, as it usually happens in time series problems.

The Metropolis algorithm is very similar to what was previously described in the frog-and-lily-pad analogy. One starts with a proposal distribution, say P(X), a target distribution, say T(Y), that needs to be approximated (the posterior in the case of Bayesian analysis) and a specific point from where to start the random walk. A value is sampled from the proposal distribution (call it 'x') and it is evaluated in the target distribution, so T(x) is calculated. Next, a value is generated from a uniform distribution with domain between 0 and 1 and the following rule is applied: if the ratio T(x)/P(x) (usually multiplied times an appropriate normalising constant so that the ratio is always less than 1 for any x) is less than the value generated by the random uniform distribution, then x is accepted and stored. If not, then another

value from P(X) is sampled and the same rule is applied iteratively. Once this is done over and over again, the random walk will have visited enough areas of high density on T(Y) to be able to reproduce such distribution to a degree of accuracy proportional to how many values of 'x' were accepted. This is why the Metropolis algorithm is also known as the *rejection* algorithm (Kruschke, 2010; Lynch, 2010).

Just as with Gibbs sampling, the Metropolis algorithm is not without its own caveats. If the proposal distribution is too narrow and the target distribution too wide-spread, the algorithm can take a long time to explore enough regions of the latter until a representative sample of it has been generated. It is also true that the first samples of the target distribution will only be representative of the area where the random walk began so they need to be discarded. This is known as the 'burn-in' period and depending on the complexity of the target distribution, several 'burn-in' periods might be required in order to be sure that representative samples of the target distribution are been taken (Rupp, Dey & Zumbo, 2004). Convergence can also be an issue for it because there is no direct way to assess whether one has a representative sample of the target distribution or not. This is why further analysis of the samples needs to be performed in order to assure convergence.

MCMC methods helped considerably in the development of Bayesian statistics by making it possible to sample from regions along which the integral of the distribution in the denominator was not well-behaved (Brooks, Gelman, Jones & Meng, 2011). It also allowed for repeated sampling and repeated updating of the likelihood which eventually gives way to the end result of most calculations, the *posterior* distribution. Because the parameters are considered to be random variables under the Bayesian statistical framework (as opposed to those in classical statistics where the parameters are fixed values in the population that need to be estimated), the posterior distribution summarises the uncertainty in the estimation (Morey & Rouder, 2011). The expected value of the probability distribution (or mean) corresponds to the parameter estimate, and a standard deviation of it can be calculated as well, in order to get a measure of the variability involved in the calculations. As an analogue to

frequentist confidence intervals, a *credible* or *high density intervals* can also be obtained to place bounds on the range of plausible values that the parameter estimate can take. These credible intervals do follow the intuitive interpretation where one can be "95% confident" that the true estimate falls within the interval (Gill, 1999).

The last step in the estimation of Bayesian models involves proposing an alternative model (or a series of alternative models) and evaluating the strength of the evidence in favour of each one. One of the main paradigmatic shifts that come with the introduction of a Bayesian perspective to statistics is that the concept of an abstract population whose parameters one is estimating or making generalisations about is not needed (Anderson, Burnham & Thompson, 2000; Kruschke, 2010). The Bayesian perspective conceptualises the uncertainty associated with the estimation through attaching a probability to the parameter one seeks to get at. For classical statistics (and just taking the normal distribution for the sake of example, it could be any distribution) the sample mean, which is itself a random variable, is an estimate of the population mean, which is a fixed-but-unknowable number. It does not change over time or through any process. It *is* that which gives the population's distribution (alongside with the standard deviation) its shape and many of its properties. However, for Bayesian statistics, all the researcher has is her or his beliefs with regards to the statistical model that generated the data and the data itself. The process that generated the data is not static but in constant movement and, therefore, the parameters that contributed to generate this data are also assumed to be changing (Dehghani, Iliev & Kaufmann, 2012). When data is brought in it is factored into the parameter estimation which makes the uncertainty surrounding them change. Perhaps it makes the distribution narrower so that it centres on the most plausible parameter. Or perhaps the data adds to the uncertainty of the model more than it is expected so that the distribution of the parameter becomes more spread out. In either case, the parameters being sought are subject to change according to their posterior distribution and it is not until more data is gathered and brought into the process that an update on the estimation

can be made, either to change it or solidify it. Testing competing models simply implies testing the same data under different assumptions from the researcher regarding how the uncertainty of the parameter behaves (Karni, 2011). By changing the prior, one can test whether the shape of the posterior changes or not and by how much, evidencing the fact that different sets of assumptions can lead to different conclusions. The strength of these conclusions can be calculated through an index called a Bayes factor so that the researcher can decide which model does a better job in making sense of her or his data.

2.3 Applied Bayesian estimation: the case of the polychoric correlation coefficient

Bayesian estimation is beginning to make headway into the realm of the social sciences (Muthén & Asparouhov, 2012). As MCMC methods become more and more widespread, more and more software programs targeted to psychologists, sociologists, education researchers, etc. are beginning to include them either as an alternative or, sometimes, even as a default on the estimation of particularly complicated models. One of such cases where Bayesian methods have made particularly important advances in the social sciences is on its flexibility to handle discrete data, something that has been a challenge in the past before the advent of numerical methods and computational approximations (Agresti, 2007). A specific instance of this problem is reflected in the issue of calculating the Pearson product-moment correlation coefficient when both variables are assumed to come from the same latent, bivariate distribution which becomes discretised, otherwise known as the polychoric correlation coefficient.

Pearson (1900) was the first to introduce the concept of the polychoric correlation when discussing the appropriate analysis of discrete data. The model on which this particular type of correlation is based assumes that the trait or variable one is interested in measuring is truly continuous but is either discretised by the act of measuring or is not directly available to the researcher so only discrete units of it can be obtained. It assumes that two types of parameters control the generation of

observed data: the polychoric correlation coefficient (or collection of those if one is working on a multivariate setting) and the thresholds that divide the variables. The model assumes that if a certain measure crosses a particular threshold value on the continuum it gets placed on a category of higher ordinal value. If there is not enough amount of such measure then what gets recorded is the category of immediate lower value. Figure 1 shows a depiction of the model for a binary case (so the *tetrachoric* correlation) with thresholds set at 0 and the correlation fixed at 0.5. It is important to notice here that the shaded area represents the sections of the marginal univariate normal curves where the proportion of said measure goes past the threshold value.

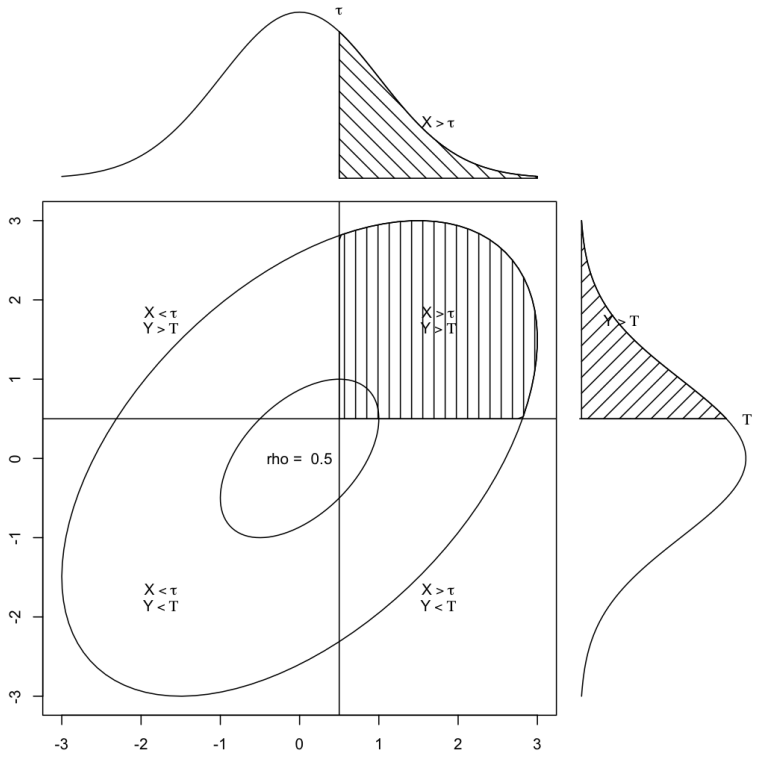

Figure 1: Schematic representation of the tetrachoric correlation coefficient with The correlation set at 0.5 and the thresholds set at 0.5

According to Pearson's colleague Burton Camp (1933), Pearson himself considered that the polychoric correlation coefficient was one of his most relevant contributions to the field of statistics, but its acceptance was not particularly wide-spread because of the difficulties involved in its computation.

Pearson only worked on the 2 X 2 contingency table case with binary data (which would later go on to become the tetrachoric correlation coefficient) and used a particular series expansions (called the tetrachoric series) which was commonly employed to approximate multivariate normal probabilities before other methods of computation were developed. Ritchie-Scott (1918) extended both the theory and computation of the tetrachoric correlation coefficient to any arbitrary r x s frequency table to keep track of the discretisation of the latent continuous variables, coining the term *poly*choric correlation to suggest that the variable would be sub-divided in any arbitrary number of ways and not only four parts, as implied in the name *tetra*choric correlation. Several refinements to the estimation where done in succeeding years, but the real advancement was seen in Olsson (1979) with the introduction of his likelihood equations whose solutions provided estimates for both the thresholds and the correlation coefficient.

Joreskog's (1994) generalisation of Olsson's method of maximum likelihood allowed for the estimation of polychoric correlation matrices, under the assumption of an underlying continuous multivariate normal distribution, and due to the prominence that the analysis of ordinal data was having in structural equation models, more and more simulation research began to appear concerning the estimation of this measure of association under a variety of conditions such as the number of discretisation points (or response options in a Likert-type questionnaire), sample size and unequal distances among the thresholds (e.g. Green, Akey, Fleming, Hershberger & Marquis, 1997; Muthen & Kaplan, 1992; Hutchinson & Olmos, 1998). Although some research indicates that moderate violations of the normality assumption do not influence the estimation of the polychoric correlation coefficient greatly, more recent results do suggest that violations of this assumption beyond skewness and kurtosis do make the case to consider the assumption of bivariate normality with much more regard (Ekstrom, 2008). From most of the literature review, it appears that the size of discretisation intervals (or alternatively whether the latent distribution is skewed or not), the number of discretisation points and

sample size are the three main factors influencing the estimation of the polychoric correlation coefficient through maximum likelihood (Flora & Curran, 2004).

There have primarily been two approaches to the Bayesian estimation of the polychoric correlation on the literature: the one developed by Chen and Choi (2009) and the one used in this thesis by Albert (1992). Chen and Choi (2009) correctly point out that, in real practice, the desirable sample sizes to obtain stable estimates for the polychoric correlation coefficient may not be available to the researcher. In their literature review they signal out the fact that most research on the performance of the maximum likelihood estimation of the polychoric correlation coefficient uses minimum sample sizes of 200 or more participants in simulation studies. They also mention that because most maximum likelihood optimisers are based on hill-climbing methods over the gradient, there is a possibility of settling in a local maximum or non-convergence. They propose two alternatives through Bayesian methods referred to as Expected A Posteriori (EAP) (or the mean of the posterior distribution) and Maximum A Posteriori (MAP) (or the mode of the posterior distribution). In both cases bivariate normality is assumed and the same likelihood equation as proposed by Olsson is employed in the Bayes' theorem step of the calculation. Chen & Choi found that the EAP estimate suffers of shrinkage effect (as it is common with Bayesian estimators which tend to gravitate towards the mean of the prior distribution) and proposed the MAP alternative which is not affected by it. In their study comparing both the traditional maximum likelihood, the EAP and the MAP estimates they focus on a variety of sample sizes (25, 50, 100, 200 and 400), correlation coefficients ranging from 0 to 0.7 and numbers of categories 2, 3, 5 and 7. They found evidence that the MAP estimate outperforms the other two, particularly in cases of small sample sizes and low correlations.

The Albert (1992) method is slightly different in the sense that he was not attempting to empirically demonstrate the advantages or disadvantages that his approach may have when compared to other methods. Albert (1992) developed the mathematical framework to allow for the estimation of

the polychoric correlation coefficient assuming any likelihood or prior distribution (so that one is not necessarily constrained by using the bivariate normal likelihood proposed by Olsson). A general sketch of the Gibbs sampler is provided as well so that the reader can work on coding his or her own and provides the conditional distribution on the cases of a bivariate normal, bivariate t and the bivariate log-normal distribution for skewed cases. In the end he proceeds with two small examples (one with simulated data and one with real, skewed data) of how his method could be used to calculate the polychoric correlation. In both cases sample sizes of 100 were used with categorisations of 3 cutpoints in each variable.

The Albert approach was preferred in this thesis over Choi et. al.'s one because of its flexibility in implementation. MAP and EAP still require the assumption of underlying bivariate normality which could be a reason as for why they were never tested under the condition of skewed distributions or unequally-spaced intervals. They also do not provide the conditional distribution of the bivariate normal distribution given the thresholds and the correlation coefficient so that it is difficult to extend these methods to other underlying distributions. Although the Albert method is more applicable due to its generality, he worked mostly on the mathematical aspects of the derivations and left the testing of its algorithm as avenues for future research. The purpose of this thesis then is to take an approach similar to what Choi & Chen did with his EAP/MAP estimators but following Albert's methodology and test it against the traditional method of maximum likelihood.

2.4 Example of Bayesian estimation of the polychoric correlation coefficient

To help familiarise the reader with the process of Bayesian estimation, a full example of what would constitute one simulation run will be shown with particular attention placed on the steps involved in the estimation stages of the polychoric correlation coefficient. Using the R statistical software package, 50 values were sampled from a standard bivariate normal distribution with correlation set at 0.5. Following the discretisation measurement process, the continuous variables were categorised in two groups for each case: those values greater than the mean of each variable were re-labelled as 1 and those falling below were labelled as 0, in order to obtain a binary variable. This is akin to a 2-question survey to 50 people where the response format is restricted to "YES/NO" or "TRUE/FALSE" answers. Table 1 shows the contingency table of proportions for the simulated data.

		Item 1	
		NO	YES
Item2	NO	*0.28*	*0.22*
	YES	*0.18*	*0.32*

Table 1: Contingency table of the proportions of simulated data with a sample size of 50 and binary response format

A function implementing the Gibbs sampler algorithm developed by Albert (1992) was developed in R. It can either take as an input the raw data or a contingency table of probabilities and outputs the Bayesian estimate of the polychoric correlation, the standard deviation, the standard deviation corrected for dependence between sampled values for the posterior and credible intervals. In order to help the reader gain better insight into the inner workings of the algorithms, time series plots will be presented for the correlation coefficient and the thresholds as well as images of the posterior distribution from which the correlations and thresholds are being sampled.

Table 2 presents some of the most commonly-reported summary statistics to describe the posterior distribution. As it can be seen, the mean (also known as Expected A Posteriori estimator or EAP) yields an estimate that is remarkably close to the set correlation of 0.5, just as the thresholds which

were set at 0. The standard deviation and the time-series corrected standard deviation also gives an idea of the uncertainty surround the estimate, which in this case is moderate.

Figure 2 is a visual representation of the time series plots that depicts the process of sampling from the conditional posterior distribution of the polychoric correlation coefficient. After 10,000 iterations, it is possible to see how the estimate jumps progressively from its initial point of 0 to 0.5 with every new set of iterations. The closer the algorithm gets to the estimate, the more the Markov Chains begin to stabilise around a solution. According to Figure 2, somewhere close to the 8,000th iteration the true value has been achieved and the remaining 2,000 iterations mostly just show how the time series begin to flatten out around the true value of 0.5

Figure 3 is simply an empirical density plot of the sampled values. It more or less mimics the behaviour of the time series plot, as expected, where very little density is located around the left-hand side of the graph and the closer one gets to the value of 0.5, the more density it accumulates until it peaks around the correct estimate. A visual inspection of the empirical posterior density can greatly help the researcher to see whether or not it settled in a proper solution. If the graph had shown cases like a bimodality or positive kurtosis (platykuritc distirbutions), it would have helped the researcher know that the solution on which the Markov Chains settled is not a good-enough solution and further checks would be required.

	Mean	Mode	Standard Deviation	Time Series Standard Deviation	95% Credible Interval
Correlation Estimate	0.51	0.48	0.012	0.01022	(0.392, 0.516)
Threshold 1 Estimate	0.0012	-0.013	0.11	0.0172	(-0.0011, 0.0359)
Thershold 2 Estimate	0.011	0.006	0.17	0.064	(-0.0062, 0.0601)

Table 2: Descriptive statistics of the posterior distribution for the polychoric correlation coefficient.

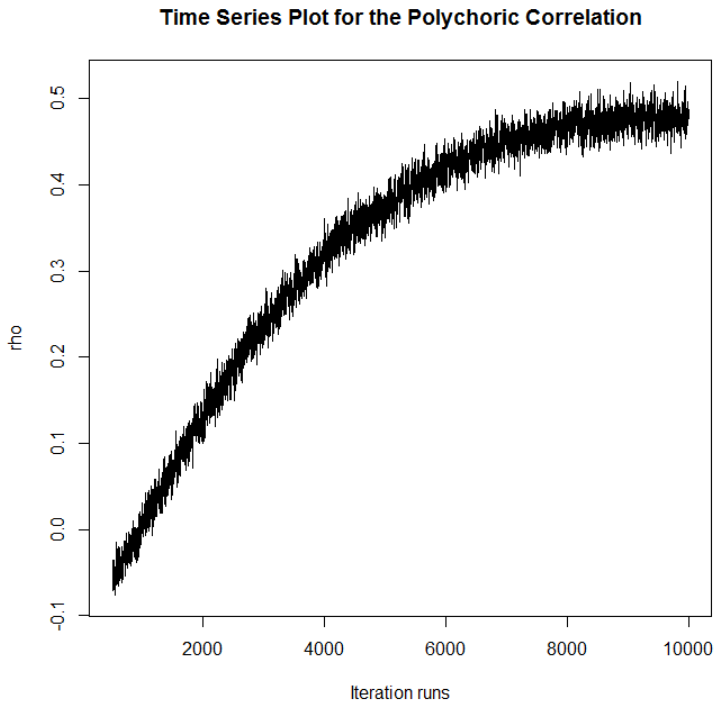

Figure 2: Time series plot of the Gibbs sampler showing the convergence of the Markov Chains towards the polychoric correlation estimate.

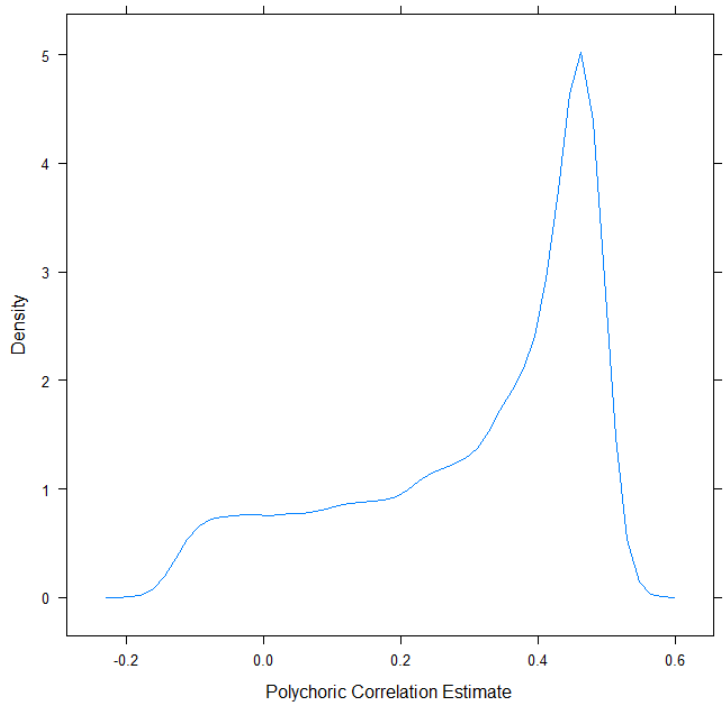

Figure 3: Density plot of the posterior distribution of the polychoric correlation coefficient.

2.4 Summary

From this brief review it can be seen that: (a) Bayesian statistics provides a legitimate alternative to traditional frequentist estimation that is slowly becoming more prominent with the recent increase in affordable computational power, (b) even though Bayesian estimation is starting to become part of the standard methods used in many areas of science, its incursion in the mainstream of the social sciences has either been seen with suspicion or somewhat ignored outside the sphere of the most quantitatively-oriented methodologists, and (c) many of the leading authorities in Bayesian statistics only consider problems pertaining to the social sciences briefly, leaving a whole area open for further development which may or may not be left unexplored depending on whether it gets noticed by people who are both mathematically-sophisticated and have an interest in measurement and data-analytic problems related to psychology, education, sociology, etc. In light of these observations, the purpose of this thesis is to further extend the use of Bayesian methods to estimate the polychoric correlation coefficient.

3. Methods

In order to further investigate the properties of Bayesian estimation methods for the case of the polychoric correlation coefficient, I used Monte Carlo simulations to test the performance of the algorithm under conditions similar to the ones that have been used to test the method of maximum likelihood (cf. Muthen & Hofacker, 1988; Muthen & Kaplan, 1992; Quiroga, 1992; Rigdon & Ferguson, 1991). The data were generated to reflect some of the less-than-ideal circumstances that researchers face in their everyday practice in order to discover how much is it possible to deviate from the theoretical model that substantiates the estimation method and still obtain relatively accurate results. Bayesian and frequentist solutions were compared further then in order to see in which cases one method outperformed the other one or in which cases the choice of method was irrelevant given the type of data being analysed.

I implemented two studies targeted to answer two different questions related to the performance of both estimation procedures for the polychoric correlation coefficient. Study 1 compared Olsson's (1979) maximum likelihood solution to the Bayesian methodology developed by Albert (1992) under a variety of different conditions such as numbers of cutpoints for the latent variable, skewness and sample size. Study 2 used a theoretical extension proposed by Albert to handle skewed data through the use of a log-normal distribution and a slight modification that was made to it which was necessary to handle left-skewed data (the log-normal distribution is always right-skewed). It was used to assess how much skew warrants the use of the log-normal distribution as opposed to the regular normal distribution and the impact that it had to use a model designed for skewed data to analyse symmetric data. Both studies and subsequent data analysis were done in the statistical software package R.

Further details on the derivatives of the conditional likelihood equations and a copy of the R code that implements the Bayesian estimation of the polychoric correlation coefficient can be found on on the Technical Appendix section at the end of this thesis.

3.1 Study 1

This study used simulation designs similar to the ones that have been implemented to test the performance of the method of maximum likelihood. Given that Albert (1992) only used two small simulations as examples of his method and that the Chen & Choi (2011) simulation study did not include skewness as a variable (even though previously research has shown it can influence the estimation), this study attempts complemented both by including skewness as a factor alongside with other theoretically-relevant conditions. A quick overview of the factors and their levels are:

- 5 different sample sizes: N = 15, 25, 50, 100 and 500.
- 6 different, equally-spaced cut points in the latent variable, starting from 2 (binary responses) to up 7.
- 6 levels of skewness for the latent variable (-3, -2, -1, 0, 1, 2, 3).

This results in a 5 X 6 X 6 fully-crossed design with 180 cells. Five hundred samples were generated for each condition and they were analysed using both Olsson's maximum likelihood solution and Albert's Bayesian estimation method defined in the appendix. The following estimates were recorded as dependent variables in this study: the polychoric correlation coefficient, the standard error (for the case of maximum likelihood) and the standard deviation of the posterior distribution (for the case of the Bayesian methods), the empirical skewness of the observed data and the number of convergences for the case of the maximum likelihood algorithm.

3.2 Study 2

This study investigated some of the theoretical developments presented by Albert for the cases of right-skewed data by using a bivariate log-normal prior as well as the necessary modifications that were implemented to handle left-skewed data, described in the appendix. The main aim for this study was to

track the impact that a different prior distribution may have on the estimation of the polychoric correlation coefficient, an issue that is both well-known and widely controversial in Bayesian statistics.

The factors and the levels in this study go as follows:

- 6 levels of skewness of the latent variable (-3, -2, -1, 0, 1, 2, 3)
- 6 different, equally-spaced cut points in the latent variable, starting from 2 (binary responses) to up 7.

For the purpose of making the study more manageable, sample size was fixed at N = 100. Five hundred datasets with those pre-specified conditions were generated and analysed using both the log-normal (right or left skewed, depending on the condition of the study) prior and the normal prior in each combination of conditions. The dependent variables recorded were the polychoric correlation coefficient, the standard deviation for the posterior distribution and the observed skewness from the raw data. Since it is known that changing the prior distribution tends to have an effect on the estimates of the statistics, it was important to therefore document the kind of impact it had (such as biases, consistent over- or underestimation, wider credible regions, etc.) as well as trying to understand under which conditions did the estimation improve when one model was chosen over the other.

3.3 Data generation

The generation of non-normally distributed, correlated data has been a subject of extensive research in Monte Carlo simulations. There exist a wide variety of methods to do so, being the preferred one among social scientists the one published by Vale and Maurelli (1983). This method uses several powers of standard normal variates and joins them together through suitable coefficients in a polynomial equation in order to obtain a joint distribution with certain skewness, kurtosis and correlation structure. Although popular due to the ease of its implementation, it has been criticised before because of how

cumbersome the calculations can become in the cases where many variables need to be generated and also because not all combinations of skewness or kurtosis are possible (Tadikamalla, 1980). It has recently been found as well that the estimates generated through the Vale and Maurelli algorithm are also prone to large variances and in many cases what is specified as the intended population skewness and kurtoses are underestimated (Headrick & Pant, 2012).

In light of these findings and mostly out of algorithmic convenience, a newer, alternative method which is gaining track very quickly was preferred: the use of copula distributions. Copulas are extremely flexible, extremely powerful analytic tools which allow either for the construction or estimation of joint probability distributions. The advantage of using copulas is that it is possible to specify the marginal distributions separately from the dependence structure and, then, join (or *couple*) them to create the multivariate distribution. For a more extensive description of copulas both as data-generators for Monte Carlo simulations and as tools for data analysis please see Joe (1997).

The basic model from which data comes from then is a bivariate Gaussian copula with beta-distributed marginals. The beta distribution was chosen because of the ease with which its skewness can be controlled through its shape parameters. In order to be consistent with the interpretation of the polychoric correlation coefficient, the assumed dependency model is that of a Gaussian copula which preserves the linear relationship between variables. The discretisation process (i.e. the way in which the cut points for the latent variable are generated) follows the same procedure as that of normal categorisation described in Bollen and Barb (1981), where the mean of the distribution is taken as a reference point and then the distribution is divided through equally-spaced intervals that move away from the mean of the distribution towards the extremes. After the data has been sampled from the copula distribution and discretised, it is subjected to analysis both through maximum likelihood estimation and Bayesian estimation and the results are recorded. The data is then discarded and new dataset is re-sampled over and over again until the number of simulation runs is completed for each

condition of the Monte Carlo experiment. Figures 4 to 6 show three examples of different bivariate Gaussian copulas with beta-distributed marginals at varying levels skewness, similar to the data-generating distributions that were used throughout this thesis. These distributions are examples of actual distributions used in this thesis as latent, continuous populations from which data was sampled and discretised. It is important to notice on these figures the skewness on the beta-distributed marginals results on overall skewness in the joint distribution. This becomes particularly evident when contrasting Figure 4 and Figure 6.

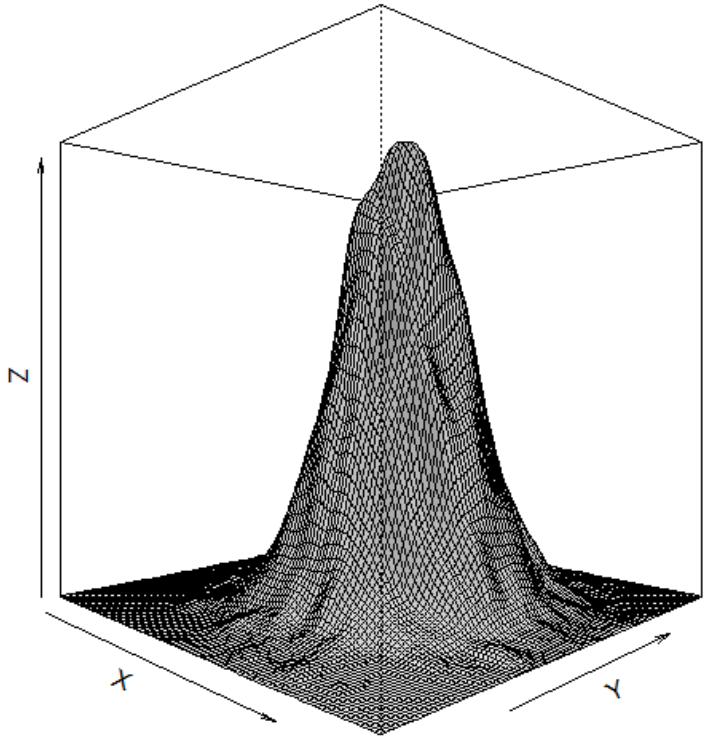

Figure 4: Bivariate Gaussian copula with beta-distributed marginals set at skewness of 0

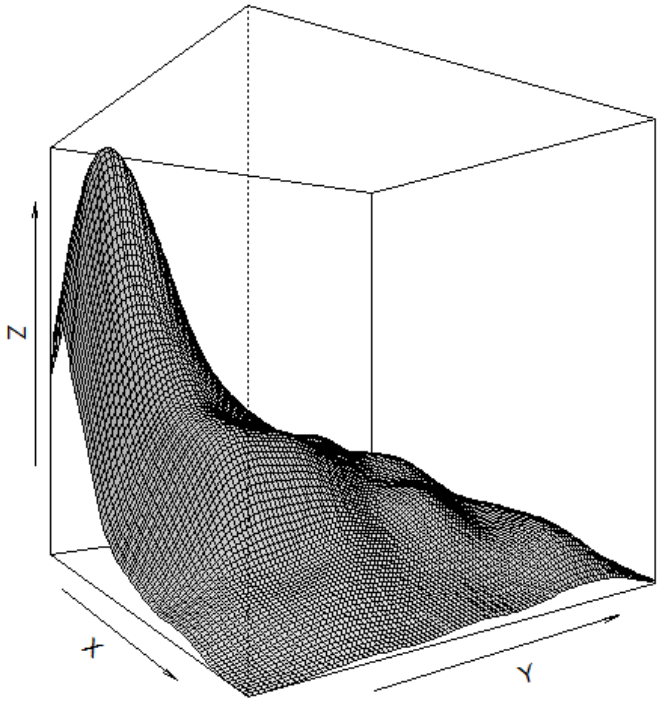

Figure 5: Bivariate Gaussian copula with beta-distributed marginals set at skewness of 1

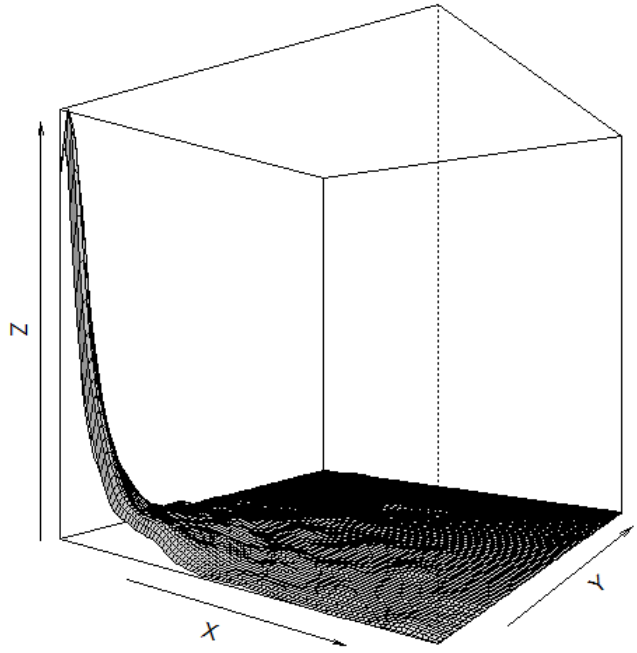

Figure 6: Bivariate Gaussian copula with beta-distributed marginals set at skewness of 2

3.4 Analysis and interpretation of results

The analysis and interpretation of these simulation results were conducted both by following the general guidelines of published simulation research in the social sciences as well as by treating the computer-generated data as results obtained from a controlled experiment (Zumbo & Harwell, 1999). An attempt was made to construct a linear model of the simulation results to further explore the relationships between the simulation conditions and the variables recorded as dependent variables in an effort to try to understand how was it that each condition contributed to the variability and accuracy of the estimates.

The use of a linear model was expected to help inform the interpretation of results by summarising the conditions in which one estimator was better than the other one and by helping to untangle the complex interactions that became present when these types of experiments were performed.

The following measures for study outcomes were employed and became the 'dependent variables' of sorts when the linear model was created to understand the impact that each condition had in the simulation design. All measures were taken from Bandalos (2006).

3.4.1 Measures used in the analysis of results

First, the percentage of convergences from the maximum likelihood estimator was calculated using the convergence rate expressed as:

$$CR = \frac{pcin}{in} \times 100,$$

where CR denotes the convergence rate, *pcin* stands for properly converged iteration number, *in* for the within-condition iteration number (i.e. 500 in all studies) and multiplied by 100 to obtain a percentage. It was expected that this measure would provide further insight as to the simulation conditions where

the ML estimator was so sensitive that not even a proper solution was found given the data it was working with.

Second, mean bias within each experimental condition was calculated as follows:

$$MB = \frac{\sum_{i}^{pcin} \rho_i - \rho}{pcin},$$

Where MB denotes mean bias, ρ_i is the polychoric correlation coefficient in iteration *i*, ρ is the true correlation coefficient (set at 0.5 for throughout conditions in this proposal) and *pcin* is the properly converged iteration number.

The bias of estimators in both theoretical and applied settings can be interpreted in many different ways and its presence (or lack thereof) factors considerably in the preference of one estimator over another. From a statistical viewpoint, the fact that an estimator generates biased results must be mentioned and whenever possible, provide a quantification of such bias if further corrections are necessary (Bandalos, 2006). From a practical perspective, however, the relevance of bias on an estimate is mostly contingent on how much it can change the inferences researchers make from their data. Since this simulation studies can be used to understand the robustness of each estimator to various conditions, I used the criterion suggested by Hoogland and Boomsma (1998) and labelled any bias on the correlation estimate over .05 as unacceptable, regardless of whether it is in the positive or negative direction.

Third, the variability of the estimates was assessed by calculating the Root Mean Squared Error (RMSE) following the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{pcin} (\rho_i - \rho)^2}{pcin}},$$

where *pcin,* ρ_i and ρ are defined as above.

4. Results

4.1 Study 1: Comparison of maximum likelihood and Bayesian estimation for the polychoric correlation coefficient

4.1.1 Convergence rate for the ML estimator

The convergence of the maximum likelihood estimator for the polychoric correlation was tracked at each simulation condition. Consistent with previous results (e.g. Chen & Choi, 2009; Flora & Curran, 2004), there were several instances where no optimal solution was found. Across all conditions, 12.5% of the maximum likelihood solutions did not converge. Sample size was the primary driving cause behind this problem where the highest amount of non-convergences were found at the sample size of 15 (10.7%), followed by N = 25 (8.12%) and N = 50 (5.44%). For sample sizes of 100 and 500 the algorithm converged in all iterations. Skewness and the number of categories also played an important role influencing the performance of the algorithm where, in general, left-skewed data resulted in a higher amount of improper solutions. For the number of categories, results were consistent with the simulation studies by Choi et. al. (2011) who found that as the number of categories increased, the number of convergences decreased. Table 3 summarises these results for a sample size of 15 and with the skewness condition depicted in each column and the number of categories in each row.

				Skewness			
Number of Categories	***-3***	***-2***	***-1***	***0***	***1***	***2***	***3***
2	13.20%	10.11%	9.32%	3.80%	4.27%	5.78%	5.92%
3	11%	9%	8.12%	4.81%	4.32%	5.82%	6.63%
4	12.26%	9.15%	8.50%	4.63%	4.20%	5.88%	6.90%
5	12.13%	10.70%	9.43%	3.21%	5.00%	6.02%	7.33%
6	13.12%	11%	10.12%	4%	5.09%	6.19%	7.49%
7	13.79%	11.90%	10.07%	4.72%	5.58%	6.50%	8.22%

Table 3: Percentage of non-convergences by number of category and Level of Skewness for N = 15

4.1.2 Bias of the maximum likelihood and Bayesian estimates

The bias of the estimation for both the maximum likelihood and Bayesian methods was calculated alongside with the root mean square error to obtain more information about its variability. The detailed summary of the findings regarding bias on the estimate can be found in Table 4. The general findings go as follows:

In general, Albert's Bayesian estimation method greatly outperforms the method of maximum likelihood in small and moderate sample situations. For the cases of N = 15 and N = 25, there was an overall tendency towards a downward bias across conditions for the maximum likelihood method ranging anywhere from 0.03 to 0.23. The mean bias for the Bayesian estimator was towards the positive end, ranging from 0.03 to 0.18 across conditions. In the moderate sample size of N = 50, minor biases were present for the case of symmetric data, but a substantial bias was found in the most extreme skewed cases (so a skewness of the latent variable of +3 or -3), a finding consistent with research on the maximum likelihood estimation of the polychoric correlation (e.g. Muthen & Hofacker, 1988). Bias is also present in Bayesian estimation, but it was somewhat lower that in the case of maximum likelihood and always on the positive direction. For the sample sizes of 100 and 500 the bias is negligible for both algorithms, as per the Hoogland and Boomsma (1998) criterion. Figure 7 shows the overall pattern of the mean bias and how it becomes reduced as sample size increases. These results reflect very closely those found in Choi et.al.'s (2011) simulation studies.

The number of categories improved the performance of both estimation methods as they increased for the cases of either symmetric or slightly skewed data (so a skewness of +1/-1). With moderate and particularly with the severe skewed conditions (so +2/-2 and +3/-3 respectively) the situation reversed and the larger the number of categories, the more substantial was the bias of the estimation, particularly in the cases with small sample sizes. Figures 8 and 9 show the pattern of mean bias as the number of categories increases and also as skewness moves from -3 to +3. In the case of

Table 4: Mean bias results for the maximum likelihood (ML) and Bayesian estimation (Bayes) solution

Number of Categories	*N*	*-3*		*-2*		*-1*		*Skewness 0*		*1*		*2*		*3*	
		ML	**Bayes**	**ML**	**Bayes**	**ML**	**Bayes**	**ML**	**Bayes**	**ML**	**Bayes**	**ML**	**Bayes**	**ML**	**Bayes**
	15	-0.23	0.18	-0.21	0.18	-0.19	0.16	-0.15	0.11	-0.143	0.120	-0.182	0.142	-0.191	0.177
	25	-0.21	0.16	-0.20	0.13	-0.16	0.09	-0.12	0.05	-0.141	0.141	-0.176	0.157	-0.187	0.164
7	***50***	-0.16	0.09	-0.14	0.05	0.06	0.02	0.03	0.002	-0.121	0.132	0.110	0.113	-0.174	0.111
	100	0.03	0.001	0.02	0.005	0.004	0.001	0.001	0.001	-0.02	0.01	-0.052	0.002	0.042	0.002
	500	0.001	0.001	0.001	0.0001	0.0001	0.0001	0.0001	0.0001	0.001	0.001	0.0001	0.0001	0.0001	0.0001
	15	-0.21	0.15	-0.20	0.16	-0.14	0.12	-0.09	0.07	-0.15	0.137	-0.17	0.132	-0.188	0.182
	25	-0.18	0.13	-0.19	0.10	-0.15	0.11	-0.05	0.02	-0.136	0.120	-0.121	0.115	-0.176	0.178
6	***50***	-0.11	0.03	0.05	0.02	0.03	0.01	0.011	0.001	0.11	0.063	0.014	0.025	0.132	0.114
	100	0.002	0.001	0.002	0.001	0.023	0.001	0.001	0.001	0.01	0.001	0.007	0.003	0.006	0.003
	500	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	15	-0.19	0.12	-0.19	0.1	-0.13	0.08	-0.10	0.06	-0.141	0.127	-0.155	0.140	-0.212	0.156
	25	-0.15	0.10	-0.12	0.11	-0.09	0.03	0.04	0.03	-0.122	0.112	-0.131	0.122	-0.160	0.149
5	***50***	0.07	0.07	0.05	0.03	0.04	0.02	0.02	0.015	0.093	0.028	0.048	0.29	-0.100	0.082
	100	0.051	0.0001	0.036	0.001	0.047	0.001	0.001	0.001	0.02	0.001	0.005	0.002	-0.032	0.020
	500	0.0001	0.0001	0.0001	0.0001	0.0001	0.001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	15	-0.20	0.09	-0.15	0.08	-0.11	0.05	-0.06	0.04	-0.132	0.122	-0.162	0.128	-0.200	0.144
	25	-0.10	0.06	-0.05	0.02	-0.05	0.01	-0.03	0.011	-0.066	0.118	-0.100	0.012	-0.134	0.127
4	***50***	0.06	0.05	0.006	0.001	0.007	0.002	0.012	0.001	-0.024	0.012	-0.007	0.001	-0.108	0.090
	100	0.0001	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.0001	0.001	0.001	-0.002	0.0001
	500	0.0001	0.0001	0.0001	0.0001	0.0001	0.001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	15	-0.17	0.09	-0.11	0.10	-0.07	0.05	-0.054	0.032	-0.128	0.115	-0.122	0.120	-0.193	0.120
	25	0.05	0.10	-0.03	0.05	-0.04	0.03	0.066	0.020	0.052	0.031	-0.042	0.012	-0.07	0.02
3	***50***	0.02	0.03	0.001	0.0001	0.01	0.002	0.014	0.011	-0.024	0.011	0.0021	0.0001	-0.03	0.02
	100	0.0001	0.0001	0.0031	0.0001	0.0042	0.0001	0.001	0.001	0.002	0.0001	0.0001	0.0001	0.0001	0.0001
	500	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	15	-0.16	0.17	-0.10	0.08	-0.05	0.03	-0.026	0.021	-0.073	0.041	-0.092	0.074	-0.155	0.128
	25	-0.16	0.04	-0.09	0.03	-0.04	0.02	0.017	0.002	0.032	0.038	-0.101	0.052	-0.152	0.111
2	***50***	0.06	0.005	0.04	0.005	0.001	0.001	0.001	0.001	0.012	0.003	-0.033	0.008	0.093	0.030
	100	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.001	0.001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	500	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

skewness it is possible to see a U-shaped pattern which was a consequence of the scaling of the graph, with the skewness condition of 0 being in the middle of it and, hence, also the section of the graph where bias was at its smallest. For the case of the number of categories, it is possible to see a progressive trend towards minimum bias as the discretisation points for the latent variables increases.

Overall, as it can be seen in Figures 7 to 9, sample size appeared to be the best determinant to ensure an appropriate performance for each algorithm, since at the largest sample sizes of 100 and 500, the estimation was unaffected regardless of the skewness and number of categories that discretised the data. With moderate and smaller sample sizes a complex relationship was present between the number of categories and the skewness of the latent variable, although the mean bias showed that Albert's Bayesian method outperforms the traditional maximum likelihood estimation.

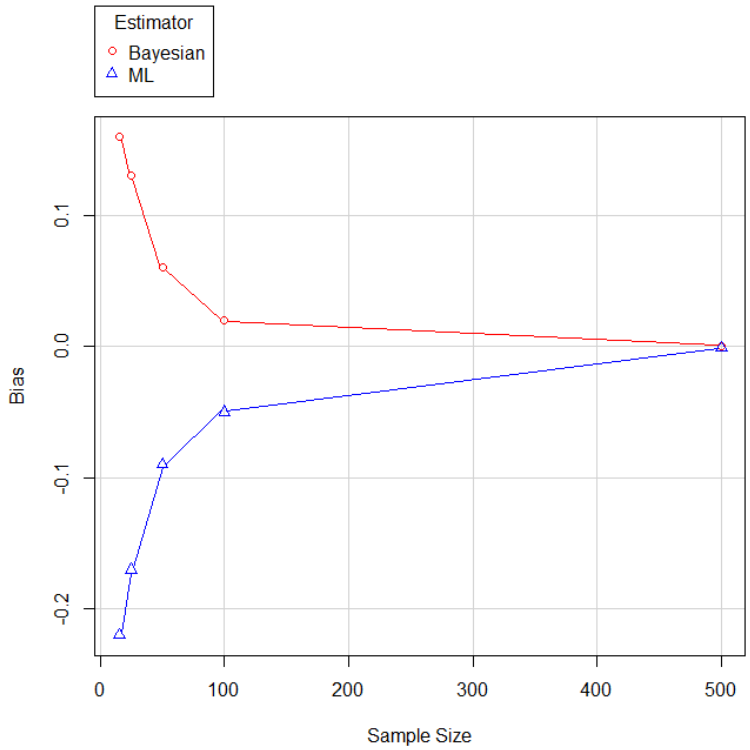

Figure 7 : Mean bias across different sample sizes. Values were averaged across conditions. ML = Maximum Likelihood estimate, Bayes = Bayesian estimate

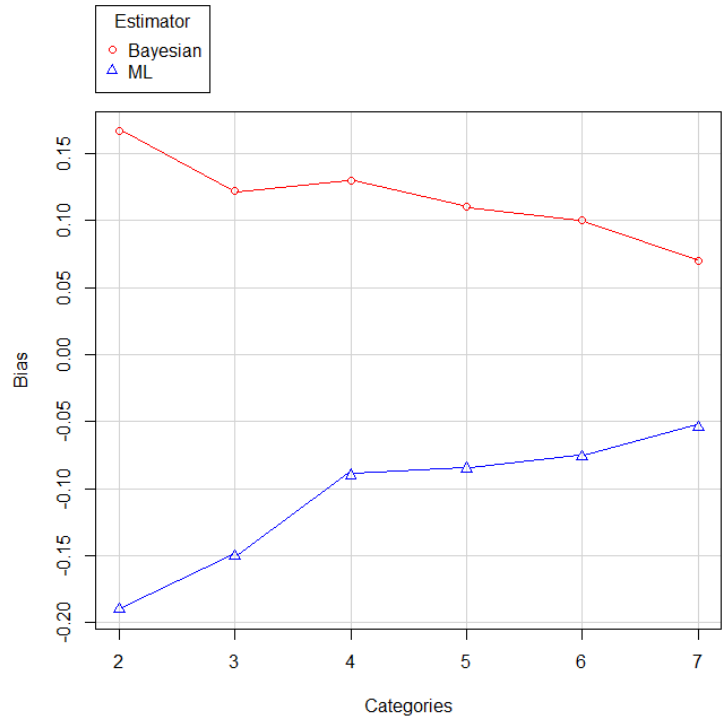

Figure 8 : Mean bias across different numbers of categories. Values were averaged across conditions. ML = Maximum Likelihood estimate, Bayes = Bayesian estimate

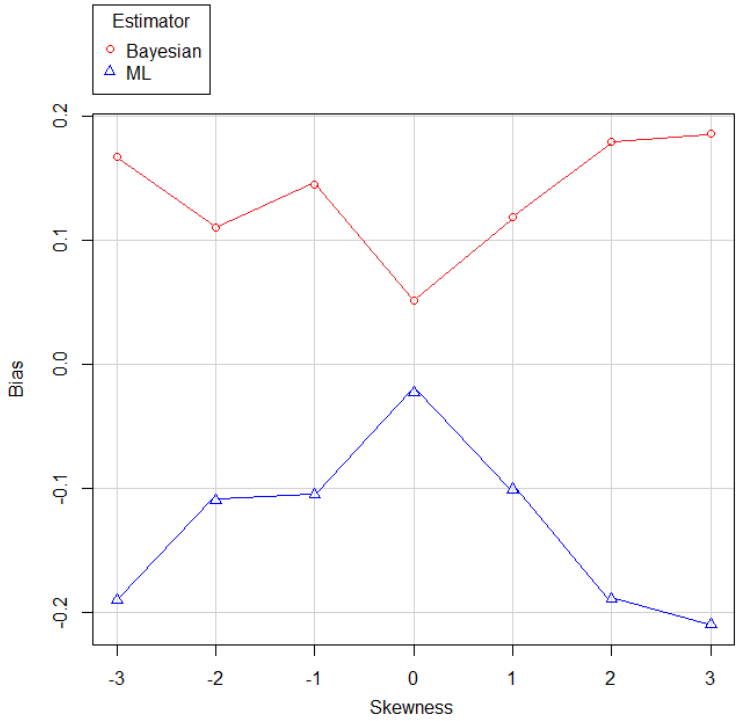

Figure 9 : Mean bias across different skewness levels. Values were averaged across conditions. ML = Maximum Likelihood estimate, Bayes = Bayesian estimate

A linear model was conducted on the results of the bias both for the maximum likelihood and Bayesian estimation cases in order to further explore the impact that each condition had on the magnitude of it. Results for each case including 2-way and 3-way interactions are presented on Table 5 and Table 6. Absent of any other criteria for effect sizes, Cohen's (1988) criterion of an eta squared of 0.1 or greater is considered a large effect size and discussed further. Those were highlighted in the two tables:

	Sum of Squares	*Eta squared*	*Percentage of R-squared (TOTAL R^2 =.685)*
Model			
Intercept	5924.972		
Sample Size (n)	***102.648***	***0.272***	***0.398***
Number of Categories (cats)	***42.325***	***0.112***	***0.164***
Skewness (sk)	***51.28***	***0.136***	***0.199***
n X sk	2.321	0.006	0.009
n X cats	9.214	0.024	0.036
sk X cats	10.425	0.028	0.040
n X sk X cats	***39.653***	***0.105***	***0.154***
Error	118.854		
Total	6301.692		
Corrected Total	376.72		

Table 5: Summary of ANOVA results for the bias of the ML estimator. Eta-squared values greater than .10 are highlighted.

	Sum of Squares	*Eta squared*	*Percentage of R-squared (TOTAL R^2= .830)*
Model			
Intercept	6847.329		
Sample Size (n)	***81.48***	***0.113***	***0.136***
Number of Categories (cats)	***100.92***	***0.140***	***0.168***
Skewness (sk)	***154.79***	***0.214***	***0.258***
n X sk	56.12	0.078	0.094
n X cats	***87.46***	***0.121***	***0.146***
sk X cats	***90.12***	***0.125***	***0.150***
n X sk X cats	28.41	0.039	0.047
Error	122.454		
Total	7569.083		
Corrected Total	721.754		

Table 6: Summary of ANOVA results for the bias of the Bayesian estimator. Eta-squared values greater than .10 are highlighted.

Table 5 shows that there exists a 3-way interaction between all the variables in the model which needs to be further explored in order to make better sense of the complex impact that all three variables have on the bias of the estimation. A non-linear effect from skewness is suspected because of the U-shaped relationship present in Figure 9, so a series of interaction plots were drawn in order to make the relationship appear more evident. Three interaction plots at samples sizes of 15, 50 and 500 in Figures 10 to 12 are presented showing the relationship between skewness and numbers of categories to help make the trend become more apparent:

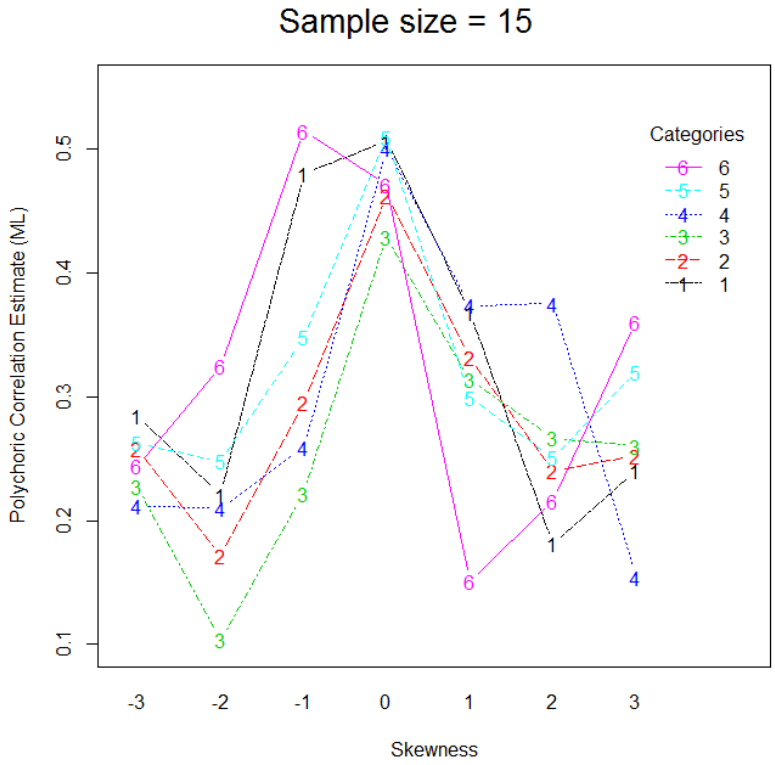

Figure 10 : Interaction plot showing the two-way interaction of the Categories factor and the levels of Skewness factor at the sample size of 15. The true correlation is 0.5

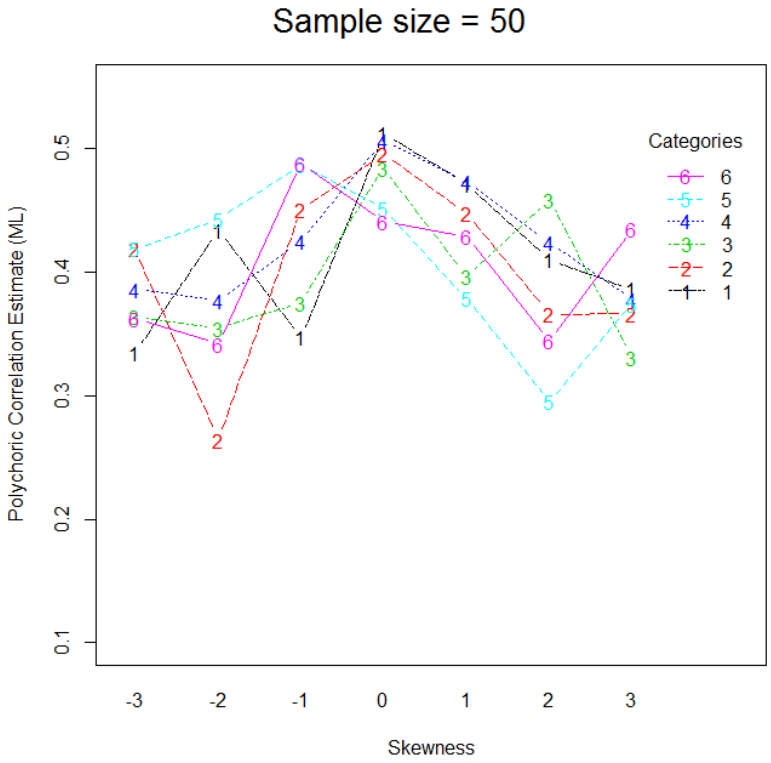

Figure 11 : Interaction plot showing the two-way interaction of the Categories factor and the levels of Skewness factor at the sample size of 50. The true correlation is 0.5

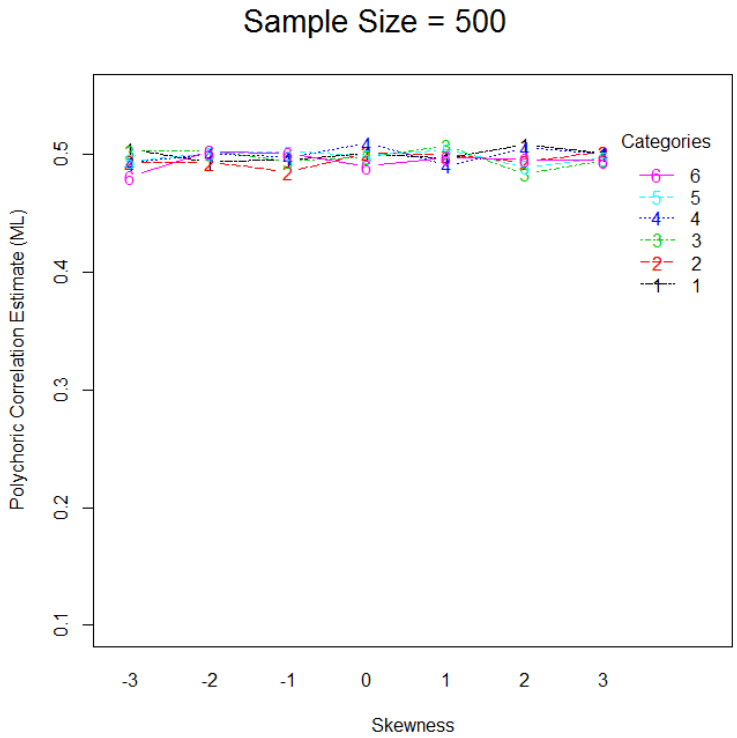

Figure 12 : Interaction plot showing the two-way interaction of the Categories factor and the levels of Skewness factor at the sample size of 500. The true correlation is 0.5

As it can be evidenced through the series of interaction plots on Figures 10 to 12, the sample size factor has an inverted U-shaped relationship with the levels of skewness. At the starting point of the lowest sample size of 15, it is quite evident that the maximum downward bias (also found on Figure 9) tends to gravitate towards the most extreme cases of positive and negative skew. However, as skew moves towards 0, the bias decreases and the marginal means of the polychoric correlation coefficient get closer and closer to the true correlation coefficient of 0.5. The inverted U-relationship is also somewhat maintained at the middle level of sample size of 50, but it has been lessened considerably. The bias surrounding the estimate of the polychoric correlation coefficient is smaller and the curve is much less pronounced, although it is still somewhat evidence that there is a peak at the skewness condition of 0 when compared to the general trend of the other skewness level. At the maximum sample size of 500 the curvilinear relationship has practically disappeared and there is very little variability around the true correlation coefficient of 0.5, as shown on Figure 12.

A potential explanation as for why this curvilinear trend is present could be because of how the contingency table of item proportions is filled-out before it is used in the estimation of the polychoric correlation coefficients. With small and moderate sample sizes there is a higher probability that many cells in such tables will be empty. If the table has very few cells (such as with binary outcomes) chances are at least a few data points will fall in each one helping the algorithm take care of the estimation. Now, on the cases with larger number of categories this becomes progressively worse for small sample sizes, however, it is known that as the number of discretisation points increases, the observed data approximates the hypothesised latent continuous distribution (see Bollen & Barb, 1981). So in cases of large sample sizes, a greater number of categories provides better estimates of the polychoric correlation. The maximum likelihood algorithm works well at the extremes of the discretisation-points (so very few or very large number of categories) but it is not particularly good around the middle.

Because the within-cell variability is so small, statistical inference is not of primary concern on this situation but the variance decomposition properties of ANOVA are utilised in order to further characterise and understand the role that each simulation variable plays in explaining the bias of the correlation coefficient. Eta-squared (η^2) is used in this case to orthogonally decompose the variance on each condition. As shown in Table 5 for the case of maximum likelihood it is possible to see that sample size is the most relevant condition with the highest eta-squared followed by the skewness of the data. Both conditions have been shown in previous research to be relevant to the precision of the estimation of the polychoric correlation coefficient. It is possible to see as well that a 3-way interaction between all three conditions has a greater eta-squared than all two-way interactions, helping to communicate the fact that these variables have complex effect on the bias.

One can see that in Table 6, for the case of Bayesian estimation, skewness shows the highest eta-squared in explaining the magnitude of the bias followed by number of categories and sample size. Here, it is important to consider as a possible explanation the fact that sample size does not influence Bayesian estimation as much as it does maximum likelihood. Nevertheless, the statistical model underlying the data is crucial for Bayesian statistics which is probably why the simulation condition directly related to the misspecification of the model (skewed instead of symmetric data) has the greatest impact in explaining the change in the bias for this estimator. The number of categories is also important to consider here, since as the number of discretisation points increases, the number of threshold parameters that require estimation also increases making the Markov Chains unstable, particularly for cases of skewed data.

4.1.3 Variability and accuracy of the maximum likelihood and Bayesian estimates

The variability of the bias was measured using the Root Mean Squared Error (RMSE). Three very distinct patterns can be identified in Figures 13 to 15. They visually depict the relationship between one of the

three major simulation conditions being studied while averaging across the other two. In Figure 13, the sample size condition reported the maximum RMSE for both the maximum likelihood solution (RMSE = 0.257) and the Bayesian estimate (RMSE = 0.221). This is also the case where the maximum amount of non-convergences was found for the ML estimate, which suggests that optimisation over the likelihood surface proved to be difficult for it. As sample size increased, the RMSE decreased in a quasi-exponential fashion, a result that is also consistent with the findings of Choi et. al. (2011), although the decrease was at a much slower rate than what these authors found. It is also worth noting that as sample size increases, the RMSE for both estimation methods decreases and becomes more and more similar until they reach the maximum of 500, where the differences are almost non-existent. In each case, the line of the Bayesian estimate remained below the one of the maximum likelihood solution, suggesting that Bayesian methods exhibited smaller variability and, hence, a higher level of accuracy around the true correlation of 0.5

For the category condition in Figure 14, it is possible to see that an increase in the number of categories is paired with an increase in the RMSE. This is a pattern consistent with what was found in Table 4, where the bias tended to increase alongside with the number of categories as well as with the amount of non-convergences present in the ML estimate. A potential explanation as for why this might be the case is due to the fact that when the number of categories that discretise the latent bivariate distribution is large, there exists a higher chance for some categories to never be sampled, leaving empty spaces in the contingency tables used to calculate the polychoric correlation coefficient. In these cases, the ML algorithm will have trouble finding a proper solution, generating somewhat disparate results.

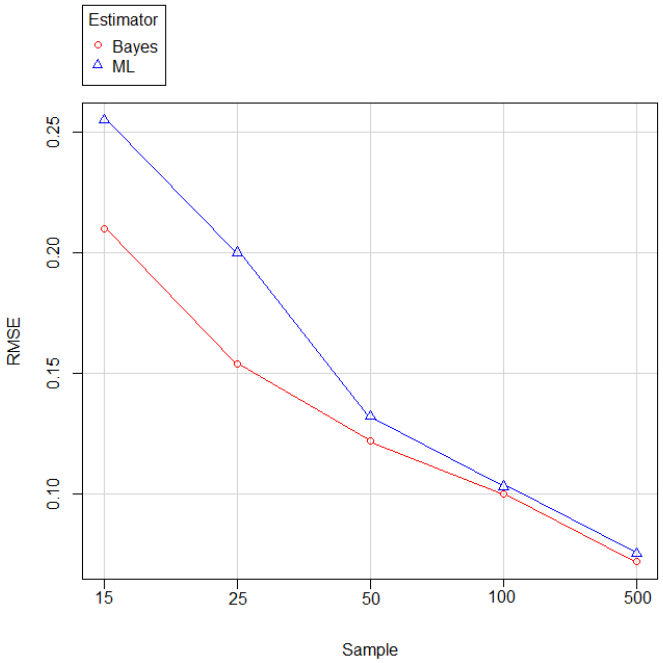

Figure 13 : RMSE across different sample sizes. Values were averaged across conditions. RMSE = Root Mean Squared Error, ML = Maximum Likelihood estimate, Bayes = Bayesian estimate

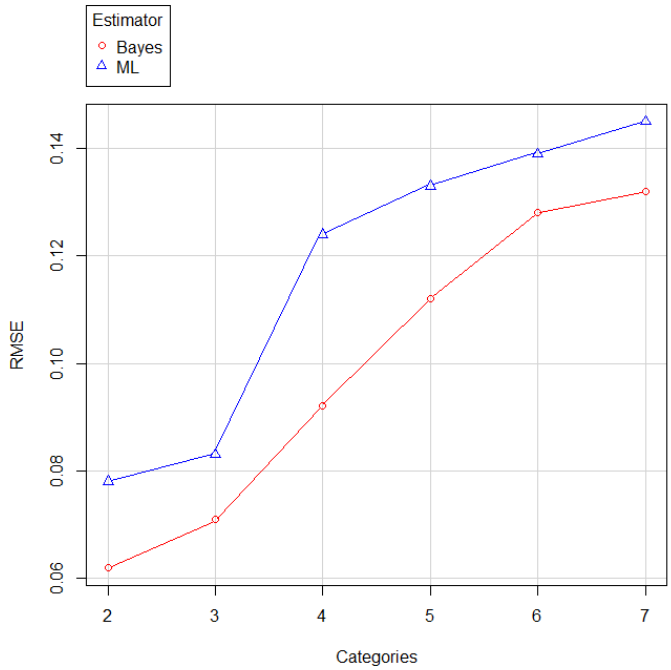

Figure 14 : RMSE across different categories. Values were averaged across conditions. RMSE = Root Mean Squared Error, ML = Maximum Likelihood estimate, Bayes = Bayesian estimate

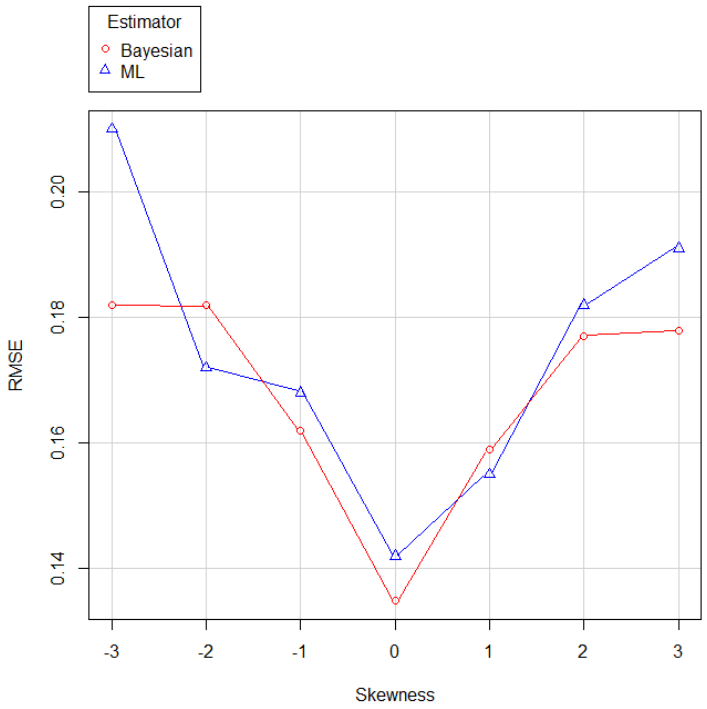

Figure 15 : RMSE across different skewness values. Values were averaged across conditions. RMSE = Root Mean Squared Error, ML = Maximum Likelihood estimate, Bayes = Bayesian estimate

For the case of the Bayesian estimate, a large number of categories imply additional parameters on the conditioning of the posterior distribution form which the correlation coefficients are being sampled, creating difficulties on the convergence of the Markov Chains.

In Figure 15, the skewness variable exhibits a clear U-shape across the different conditions, touching its lowest point when the data is generated from a symmetric distribution, so skewness = 0. Consistent with previous findings on the literature (e.g. Flora & Curran, 2004; Joreskog, 1994), the higher the skewness the more troublesome it is for reliable estimates to be found. On this study condition, no estimator consistently outperformed the other one and the only clearly discernible pattern is that on cases of extreme skewness, the RMSE is at its highest and it progressively decreases as the skewness decreases as well. A potential explanation for this pattern could be the fact that a model that does not match the data-generating mechanism is used to analyse the data, translating into biased and unstable estimates.

4.1.4 Uncertainty of the maximum likelihood and Bayesian estimates

The standard error for the ML estimate and the standard deviation of the posterior distribution for the Bayesian estimate were calculated in order to obtain a better sense of the uncertainty associated with the statistics being calculated. Although the concepts are not interchangeable, the idea behind them is very similar and in many cases both are within the same range.

Figure 16 shows the empirical density plots for both the maximum likelihood and Bayesian estimates of the polychoric correlation across simulation conditions. It can be seen from it that there is more variability for the ML estimate than for the Bayesian one, suggested by the fact that its density plot is skewed to the right, with a tail that extends all the way past -0.5 (so some of the estimates are a full unit below the true correlation of +0.5). It is also possible to see that there is a much higher density on estimates over 0.5 and it is the only graph of the two that extends to the theoretical upper limit of 1. The density of the Bayesian estimates is both more compact (suggesting a lower variability of the estimates) and has a

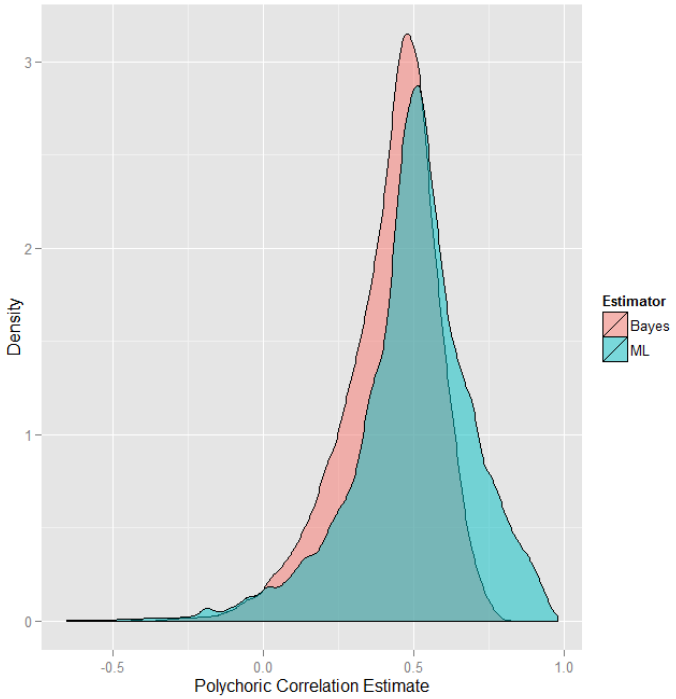

Figure 16: Empirical densities of the ML and Bayesian estimates across simulation conditions

a higher peak around the true correlation of +0.5, suggesting that a larger number of estimates fell closer to it than the ones calculated via maximum likelihood.

A boxplot of the standard errors (for the case of ML) and the standard deviation of the posterior distribution (for the Bayesian estimator) shows that even though the mean standard error for the ML is lower than the mean standard deviation, a lot more uncertainty is present on the maximum likelihood estimates. Standard deviations range from 0.092 to 0.43 whereas standard errors can go from 0.055 to 0.64. Approximately 55.2% of standard errors fall over their mean of 0.183 whereas 44.9% of the standard deviations are located over their mean of 0.188, suggesting that Bayesian estimates are more efficient in controlling the uncertainty involved in their calculation because their standard deviations tend to be smaller.

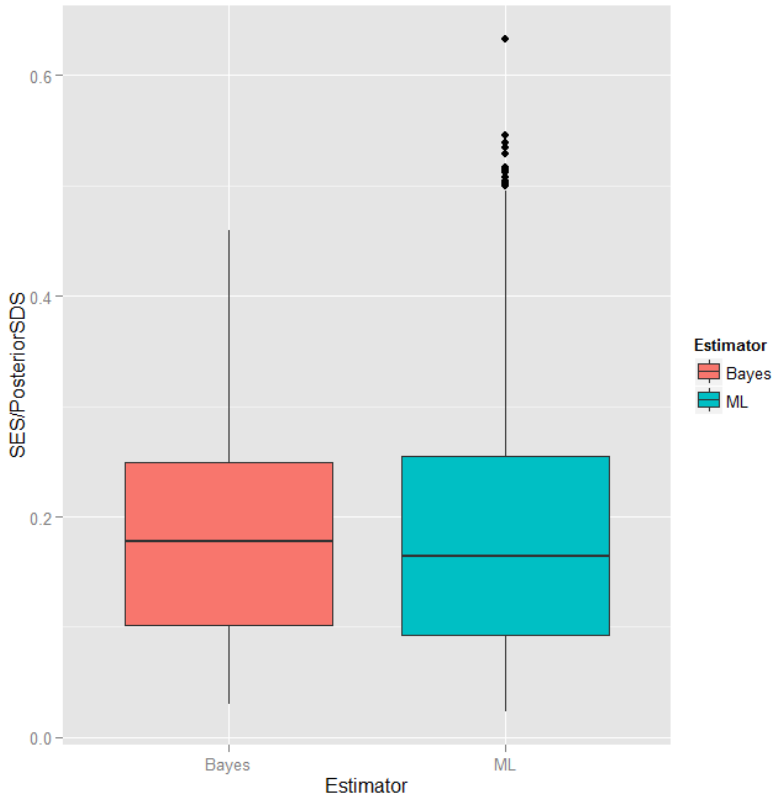

Figure 17: Boxplot of the standard error of the estimate and the standard deviation of the posterior distribution for the ML and Bayesian estimates respectively

4.2 Study 2: Comparison of bivariate normal and bivariate log-normal choice of prior distributions

As it was previously mentioned in the introduction, the choice of prior density influences both the estimate of the statistic being studied and the inferences that can be derived from the data. In the case of the polychoric correlation coefficient, it is known that the nature of the latent, continuous density can have an impact on its estimation but relatively little work has been done in order to extend the assumption of bivariate (or multivariate) normality to accommodate other latent distributions (Ekstrom, 2008). Albert's (1992) framework has a natural extension through the use of a different conditional posterior distribution, the log-normal distribution, which is flexible enough to accommodate latent skewed distributions. From Study 1 it was possible to see that skewness is an important factor on the estimation of the polychoric correlation coefficient. Study 2 attempts to look at how well can Albert's framework handle non-normal cases.

4.2.1 Bias of the estimate

It is known that the skewness of the latent variable reflects on the estimates depending on how the contingency table gets populated (Savalei, 2011). The more skewed the latent variable is, the more certain cells receive the bulk of the proportions while others are left virtually empty or with very few data points, biasing the estimation.

As shown in the previous study, it can be seen that the symmetric prior density (bivariate normal) was able to handle cases with small number of categories and extreme skewness or small levels of skewness if there is a larger number of categories. Bias became more severe at the highest level of skewness or when the number of categories is moderate/large, even with skewness levels of +/-2. In cases such as these ones, it is immediately possible to see that the switch from bivariate normal to bivariate log-normal yields much better estimates in terms of lower bias. Figure 18 shows how the use

of a symmetric prior distribution to analyse data coming from a skewed distribution demonstrates consistent underestimation or over estimation, depending on the direction of the skewness of the marginal distributions. The choice of a skewed prior, on the other hand, shows much smaller bias and it achieves its maximum at the point of zero skewness, where the model used to fit the data is incorrect. Mean bias across replications in this case for the symmetric prior distribution ranged from -.16 for left skew of -3 to to 0.14 for a right skew of +3. For the case of the log-normal (skewed) prior, bias achieves its maximum at the skewness condition of zero, 0.11.

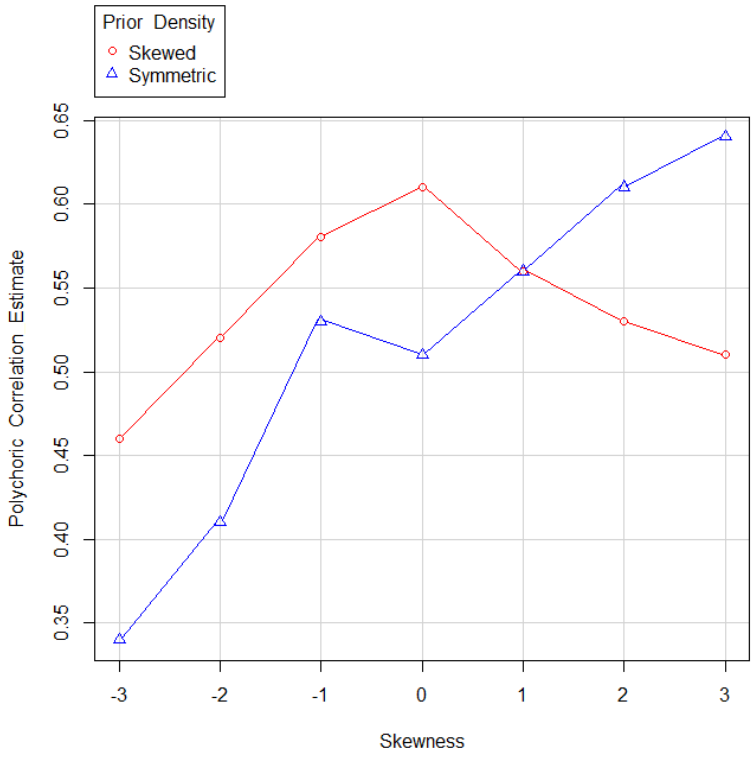

Figure 18: Mean bias across different skewness levels. Values were averaged across conditions. Symmetric = Normal prior density, Skewed = Lognormal prior density

The number of categories also had an effect on the resulting estimates in terms of the choice of prior. As it was detailed before, the level of skewness of the latent variable impacts the way in which the contingency tables are filled-out, swaying the algorithm one way or another and influencing the

estimation. In general, the fewer number of categories the less the impact of the choice of prior on the bias of the estimates. As the number of categories increases the bias on the estimation becomes greater

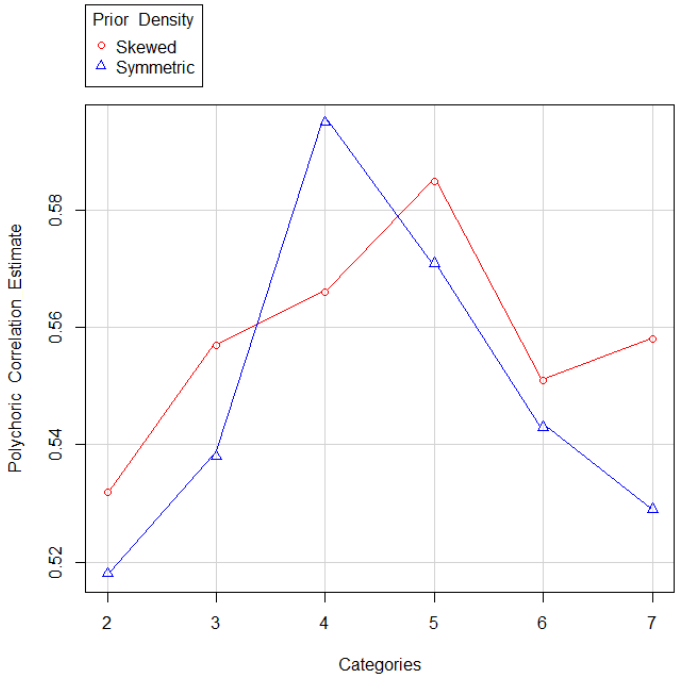

Figure 19: Mean bias across different numbers of categories. Values were averaged across conditions. Symmetric = Normal prior density, Skewed = Lognormal prior density

up until the largest number of categories, where it begins to decrease again, showing an inverted-U shaped relationship in the case of both symmetric and skewed priors. The lines representing both types of priors cross each other two times, so no absolute pattern of over or underestimation can be found regarding the type of distribution used. The consistent overestimation of the Bayesian solution, regardless of the choice of prior, is a result that is consistent with Study 1, although the variability on the estimation is probably due in part because of the averaging across the skewness condition. It is interesting to notice that around the medium range of categories (including the ever-so-popular 5 response format option) the Markov Chains take some of its longest time to converge, requiring several more sample draws from the conditional posterior distribution. For the case of the log-normal prior with 5 categories, Figure 20 shows the time series plot of the sample draws from the posterior averaged across all 500 replications. Although the posterior expected value centres on the true correlation value

of 0.5 with a slight overestimation, it is possible to see that the variability of the samples is *considerably* larger when compared to the time series plot from Figure 2 on page 24 of this thesis, where all the assumptions behind the estimation are in place and the convergence can be easily identified as the number of cycles increases.

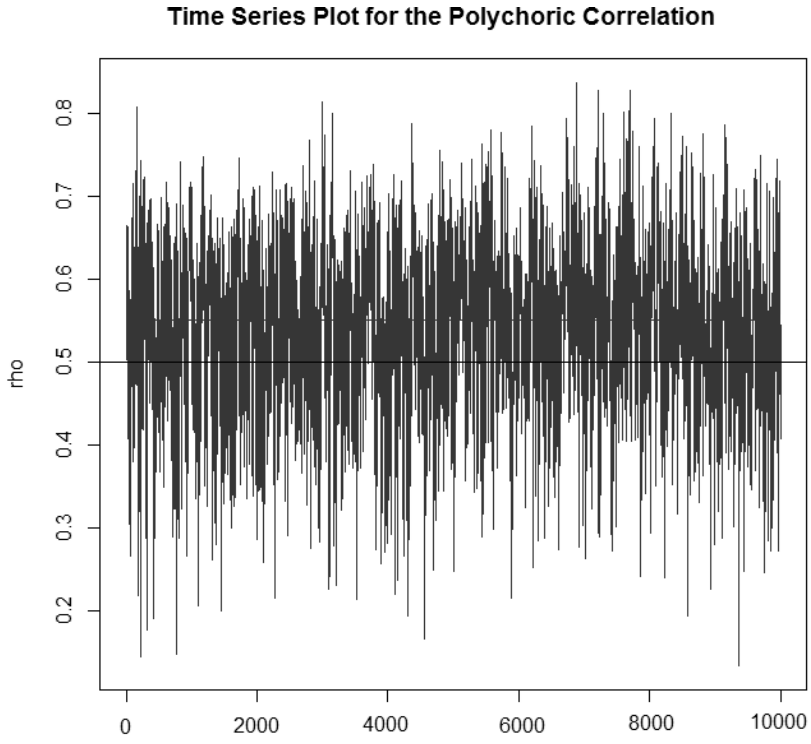

Figure 20: Time series plot of the samples from the posterior distribution of the lognormal prior with 5 categorisation points averaged across 500 replications.

4.2.2 Uncertainty of the estimate

The standard deviations of the posterior distributions for both estimates were recorded in order to obtain more information regarding the variability and uncertainty of the results. Because of the long tail of the lognormal distribution, it is known that that its standard deviation is larger than that of its associated normal analogue (Attfield & Hewett, 1992), however, the standard deviation of the posterior distribution is still an accurate summary of the variance involved in the estimates of the correlation.

Figure 21 shows the empirical density plot of the estimates for the skewed and symmetric prior distributions, helping visualise their variability across replications and conditions.

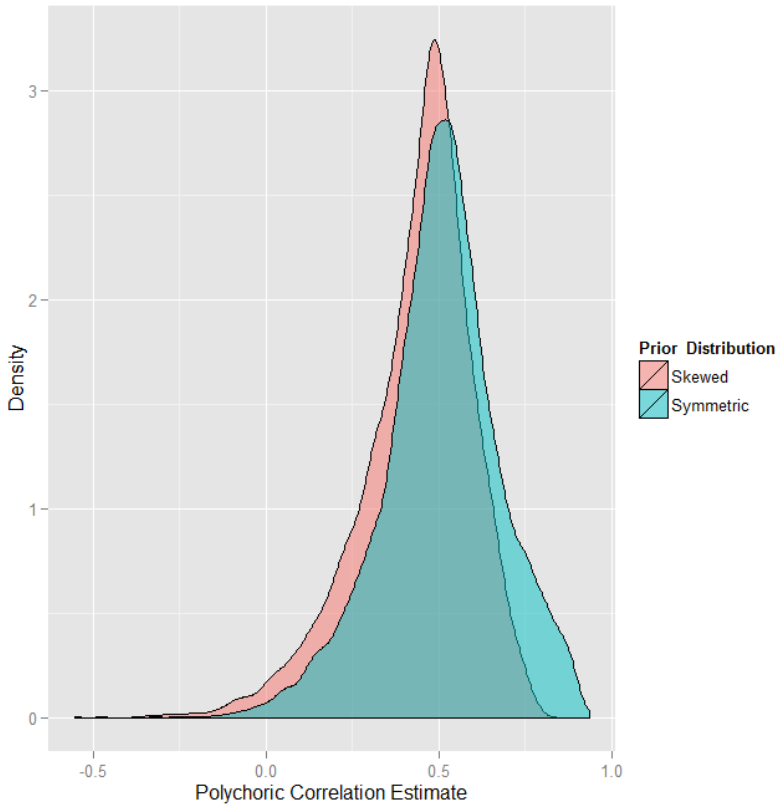

Figure 21: Empirical densities of the Normal and Lognormal prior estimates across simulation conditions.

As it is possible to see, the mean of both empirical distributions are accurately centred on the true correlation of 0.5, although the choice of skewed prior shows a higher peak around it, implying that there is more density on this area of the graph and, therefore, it captures values close to the theoretical estimate more often. One possible explanation as for why this is the case is because across the skewness conditions analysed in Study 2, only one corresponds to a symmetric distribution, whereas the other six conditions imply the continuous distribution is skewed and an appropriate prior is being chosen to analyse it. There is somewhat higher variability on the skewed prior which has both a longer tail on the

left and higher density over the symmetric prior. Regardless of the choice of density no severely extreme values are estimated by either one of them.

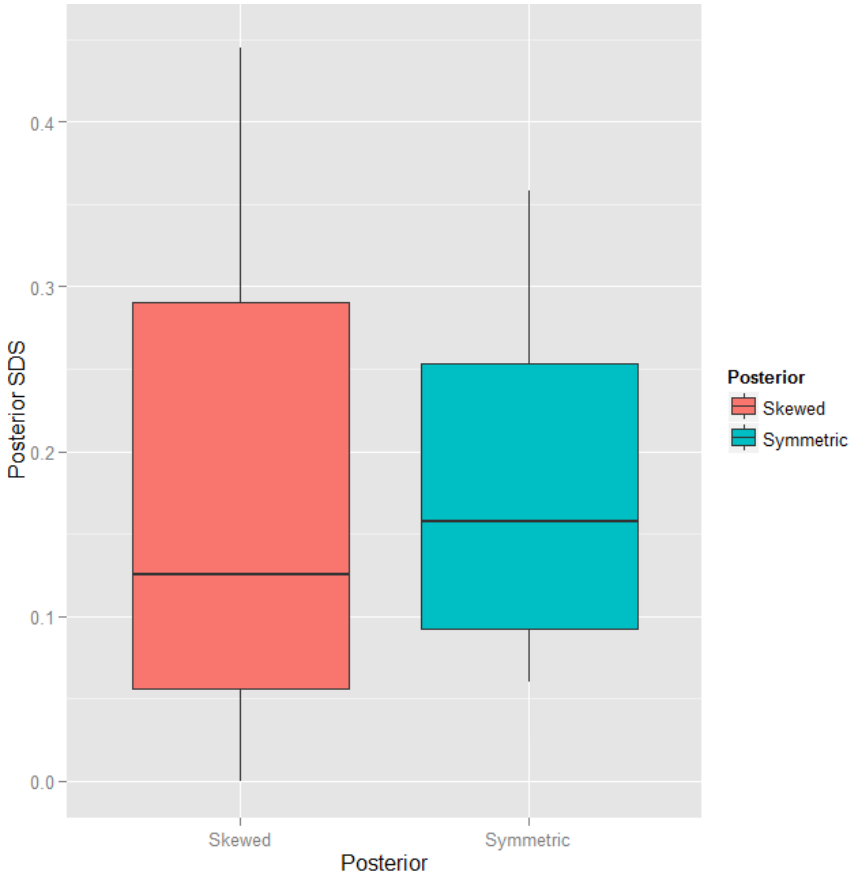

Figure 22: Boxplot of the posterior standard deviations of the normal and lognormal prior

The boxplots of both types of posterior distributions show that even though, on average, the standard deviation for the skewed distribution is smaller than that of the symmetric distribution, the spread of the skewed distribution is considerably greater than that of the symmetric distribution. The difference between the mean standard deviations is 0.03 but the difference between their respective maximums is 0.12, for instance. It is also possible to see that the whiskers of the boxplot representing the lognormal prior extend themselves longer both towards the lower and higher end, suggesting the fact that there are considerably larger numbers for this density, particularly on the higher end of it. From

here it is possible to conclude that there is more variability in the estimates coming from the lognormal prior. Albert (1992) both observed and commented on this situation in the original article where he proposed his Bayesian framework. Upon comparing the results obtained for the lognormal and normal priors, he noticed that the standard deviation of the posterior distribution was somewhat larger than that of the normal distribution, even though he did not offer any potential explanations concerning the reason behind this.

5. Conclusion

The purpose of this thesis was to expand on the previously-published findings regarding the different ways of calculating the polychoric correlation coefficient and how the conditions under which maximum likelihood estimation is tested impact the Bayesian framework developed by Albert (1992). Acknowledging the fact that Bayesian statistics are becoming a competing paradigm to classical statistics, this thesis also aimed at taking a problem pertinent to the social sciences (the analysis of categorical data when a continuous, latent construct is being hypothesised to underpin it) and giving it a Bayesian twist, highlighting the advantages and drawbacks that this new estimation framework has when compared to more traditional methods.

5.1 Summary of Study 1 and Study 2

Study 1 can be framed and interpreted as a robustness study where the performance of the Bayesian algorithm is tested alongside with that of maximum likelihood to understand under which conditions one outperforms the other or both give the same answer. In general, the MCMC approach was found to outperform the ML approach in the some of the most difficult cases (small sample sizes with skewed data and medium/large number of categories) but does not do so consistently. Sample size proved to be a variable of great importance since bias diminished as it increased in both cases. Skewness of the data only becomes an issue in the extreme cases, even with moderately large sample sizes, but can be a problem even in moderate cases if the sample size is small. Maximum likelihood estimation showed consistent problems with underestimation whereas Bayesian methods tended towards overestimation, possibly due to the fact that this type of algorithms tend to be swayed by the mean of the prior distribution. Overall, the Bayesian estimation of the polychoric correlation coefficient tended to be more robust than that of the traditional maximum likelihood, with the exception of cases where extreme skewness was involved due to the fact that these methods are very sensitive to model specification and

the choice of the wrong model can work against finding the appropriate estimate of the statistic being studied.

Study 2 attempts to further understand the Bayesian approach to the estimation of the polychoric correlation coefficient by dwelling in one of the most controversial aspects of Bayesian statistics which is the choice of a prior distribution. By reproducing the same skewness conditions as in Study 1 but performing the analysis with skewed (bivariate lognormal) and symmetric (bivariate normal) priors, it is possible to document the impact that the choice of distribution has in the ultimate estimation of the correlation coefficient. Overall, as expected, an increase in the skewness of the data being analysed goes hand-in-hand with better performance of the lognormal prior distribution over the normal prior. The number of categories seemed to play a complex role as well in the estimation, but no discernible pattern appeared in the simulation study with regards to whether the choice of one prior over the other is preferable, which is probably due to the fact that the number of categories condition was averaged across the skewness condition.

5.2 Conceptual issues arising from the implementation of these studies

During the development and implementation of the simulation studies, two important issues arose that required careful examination before a full implementation of the methodology could be developed. The first one is the discretisation process of the continuous latent variable and its implication on the models used to be analysed and the second one is the role of bias in Bayesian data analysis.

For computer simulation purposes, there exist two general ways in which one can generate skewed discrete data coming from a continuous variable. The first one is to keep the thresholds fixed at equal intervals and manipulate the skewness of the latent variable, which is the Bollen and Barb (1981) method employed throughout this thesis. The second one is to keep the latent variable symmetric (so fix its skewness at 0) and manipulate the distance of the thresholds, so that unequal cutpoints will favour

certain areas of the density of the distribution over others, resulting in samples that are skewed towards such areas (see Zumbo, Gadermann & Zeisser, 2007 for an example). For matters of the data-generation process, the choice of approach is irrelevant because any specified skewness can be achieved by either one of them indistinctively. But for data-analytic purposes, an important distinction between the Bayesian and frequentist approach to data analysis becomes apparent.

For the case of the polychoric correlation coefficient (and the same argument is valid in any other statistical techniques that requires the frequentist paradigm) the method of maximum likelihood *assumes* right from the start that the data comes from a bivariate or multivariate normal distribution which later becomes discretised, echoing the same thought process described in the Bollen and Barb (1981). All the analyses done on the data begin from such assumption and it is that very reason for why the Olsson (1979) likelihood equations are used for the estimation. Under the frequentist paradigm, *the model takes precedence over the data* and in such cases where the data does not fit the model corrections such as transformations, the removal of outliers, etc. are prescribed in order to aid on the estimation.

In Bayesian data analysis, the opposite takes place and *the data takes precedence over the model*, which is why the prior distributions discussed in Section 2.2 of this thesis become so relevant. It is through the use of the prior that the likelihood equation becomes influenced to accommodate the characteristics of the sample one is working with. This is one of the primary reasons for why Bayesian methods have become controversial, which is that they force the researcher to acknowledge that other sources of uncertainty exist which are not contained in the likelihood function, violating the likelihood principle of frequentist statistics that state that all the information needed to draw inferences from the data is contained in either the likelihood function or some transformation of it (Casella & Berger, 1987).

The role of bias in Bayesian statistics highlights an important aspect of the way in which it conceptualises the uncertainty of the data. As elaborated on Section 2.2, the end result of the

estimation process is not a unique number or estimate as in the case of maximum likelihood, but a probability distribution of the most plausible values of such estimate out of which the most plausible value is selected. Usually, this number is the mean or the mode of the posterior distribution but it could well be any other number depending on the purposes of the analysis. If a distribution and not a single value is the end result of Bayesian analysis then, what is the role of bias? Bias from what? And, more importantly, if the Bayesian paradigm does away with the frequentist notion of long-run probability, then what is the final value from which bias can be calculated, particularly when it is known that different prior distributions can generate different final results? In cases like this, it is very important to keep in mind the definition of probability that underlies Bayesian analysis. The posterior distribution is nothing more than a way to measure of the uncertainty of the parameters being estimated and, as such, it still obeys the laws of probability. Among such laws it is true that, as long as one does not end with a degenerate case, an expected value of said distribution exists. The bias that gets calculated is from this expected value (or some other representative measure) and it is the uncertainty around this value that what re-expressed in Bayesian statistics through the posterior probability distribution and not necessarily through a standard error as in the case of ML-based estimates. The bias of a parameter estimate exists and its calculation is meaningful, a fact that can be evidenced by reminding the reader that as the sample size increases, both maximum likelihood and Bayesian solutions will converge towards the same answer (Kruschke, 2010). It is the way in which such bias is understood what changes slightly when jumping from the frequentist to the Bayesian paradigm.

5.3 Contributions and limitations of this thesis

The Bayesian perspective to statistical estimation and analysis is still not very well-known among social scientists, and even many who are specifically trained in quantitative analysis are only familiar with some of its overall concepts. The main aim of this thesis was, therefore, to take on a problem that is

common in psychometrics and attempt to approach it from a Bayesian point of view. Although similar studies have been done before (e.g. Choi et.al., 2011) the first novel contribution of this thesis is the introduction of the skewness condition in the robustness-like study of the MCMC algorithm. Section 3.3 of this thesis elaborates on some of the difficulties to generate data with a pre-specified correlation matrix and different levels of marginal skewness and kurtosis and although modern solutions exist (eg. Headrick & Pant, 2012), most quantitatively-oriented social scientists still rely on the Vale and Maurelli (1983) process to generate their data, regardless of how cumbersome and limited it can be. The use of copulas to generate the desired bivariate distributions can also be cited as a novel contribution since their use has been restricted mostly to the fields of financial mathematics and biostatistics.

Although the Albert (1992) article presents a general overview of how the Gibbs sampler for this polychoric correlation coefficient looks like, many gaps are left for the reader to fill-in, in case he or she is interested in the full implementation of the algorithm. It is particularly complicated to obtain the first, second and third derivatives of the log-likelihoods of the bivariate distributions which are necessary for the variance-stabilising transformations before the Markov chains can be run. The Technical Appendix included at the end of this thesis works through the mathematics that Albert probably worked on before writing his article and provides, in closed form expressions, all the necessary equations that would need to be translated into code if anyone were going to try and re-implement them. These equations were not published but are crucial building blocks in the Bayesian framework herein developed. Last but not least, out of this thesis came a fully-customisable R function capable of handling any kind of discrete data in order to estimate the polychoric correlation coefficient through MCMC. This function can be found in the Technical Appendix. With the appropriate packages installed and running the necessary code beforehand, the function is ready to operate, offering the user a graphic display of the chains, the mean of the posterior distribution as best estimate, the standard deviation of such posterior distribution and the top and bottom 5% empirical quantiles which can be used as credible intervals.

Several limitations are also present in the simulation studies that comprise this thesis. The first one is that although the focus was to test the algorithms under difficult situations such as small sample sizes, not too much attention was paid to the opposite side of the spectrum, larger sample sizes. It is generally true that with larger sample sizes maximum likelihood tends to perform better, but very large sample sizes can also create very large Markov chains which may take longer than expected to converge (Kruschke, 2010). Studying how to optimise the MCM algorithm could have been further explored in this thesis.

Choi et. al. (2011) provided very insightful recommendations by exploring not only the expected a posteriori (EAP, or the mean of the posterior distribution) but also the maximum a posteriori (MAP, or the mode of the posterior distribution) estimators of the polychoric correlation coefficient and found that, in several cases, MAP estimators outperformed both traditional maximum likelihood and EAP. This thesis solely focused on the EAP estimator because Albert provided not guidance in terms of how to calculate MAPs from the conditional distributions he published. Choi et. al. (2011) approach relies exclusively on the log-likelihood of the bivariate normal distribution for which closed-form expressions for the mode exist. Albert (2012) used conditional distributions both for the bivariate normal and the lognormal case for which the estimation of the mode can be complicated and is probably impossible unless numerical methods are used.

Another limitation worth mentioning is the fact that other types of prior distributions could have been used as a manner of conducting a 'sensitivity analysis' of sorts in order to see which type of prior has the most impact on the results, or in which cases the choice of prior has no influence when estimating the polychoric correlation coefficient. For purposes of this thesis, only a lognormal prior distribution was used to control the skewness condition, but distributions with varying degrees of kurtosis or different levels of tail dependence, for instance, could have been used as well to further investigate the assumption of latent bivariate normality and its relationship to the estimation process.

The choice of prior distributions has always been a controversial topic in Bayesian analysis and the more informed a researcher can be about his or her choice of prior the better it is for both the appropriate understanding and use of Bayesian estimation methods.

5.4 Insight into the future directions and challenges of Bayesian statistics in psychometrics

With the advent of modern computer power, Bayesian statistics has come within the reach of the applied researcher who is interested in using this approach to conduct her or his routine analyses. Introductory textbooks that require little technical background have also begun to appear (e.g. Kruschke, 2010), catering to experts in fields outside of statistics in order to help them become familiar with this alternative approach to traditional data analysis. In spite of this, there is still a long way for the Bayesian paradigm to start taking hold, particularly among the social sciences.

Perhaps the greatest obstacle to be found is the education of future generations of methodologists and applied statisticians in the social sciences with regards to Bayesian analysis. Up until now, any student or practitioner wishing to dwell deeply in the new paradigm needs a certain degree of proficiency in the more mathematical aspects of statistics. Most curricula of applied statistics programs for social scientists tend to move away from it, creating a barrier between the Bayesian approach and potential users that could benefit from it. It is also true that many of the programs, as they are, are quite saturated with very relevant content for everyday data analysis, which makes it complicated then to try to extend them to include introductory material related to Bayesian statistics.

A second great obstacle that has even been addressed as one of the ‘open problems’ in Bayesian statistics is the criticism that surrounds choosing prior distributions (Jordan, 2011). The fact that the researcher or data analyst has so much input on the end results of the estimation process troubles a great number of experts in the field, vehemently opposing the use of Bayesian statistics or, at the very least, not recommending their use until many of the properties of their models are better

understood. A lot of misinformation exists with regards to the proper role of the prior in Bayesian estimation but, ultimately, the challenge to the use of this new approach in data analysis does not come from the mathematics underpinning it but from the way in which statistics is so intricately related to the process of creating science. The seeming absolute objectivity that comes from the classical paradigm is a difficult mind set to get away from and a more thorough discourse is needed then not only in the process of analysing data but in how the analysis of data shapes the development and presentation of results that will later go on to become scientific knowledge.

Another important limitation which has been brought forward in the debate to include Bayesian statistics into the mainstream of applied science is that of software issues. At the present moment, there exist very few software programs targeted towards Bayesian analysis that is GUI-friendly. With the exception of a few specific applications (e.g. Netica for Bayesian networks), most software targeted to this area of statistics requires at least some basic familiarity with the process of writing computer code, something not all social scientists feel comfortable with or proficient in. The second issue is the computational difficulties associated with MCMC methods. Although it is true that significant advances have been made on the process of making the implementation of these algorithms more approachable, more research is needed to help with speeding up a lot of the processes inherently involved in the tying up of all the Markov chains. If the applied scientists see just marginal gains and a great loss of time in the use of Bayesian methods for their everyday analyses, it is difficult to expect that they will try to even consider these methods as legitimate alternatives to what classical statistics has to offer.

Bayesian statistics still have a long way to go before it becomes a more mainstream option in standard data analysis and estimation. Nevertheless, the shift has already started and the fact that more and more journal articles, conference papers and introductory textbooks are beginning to appear points towards the fact that, eventually, this new approach could very well become a legitimate alternative to the way in which social scientists analyse and help consolidate scientific knowledge.

References

Agostini, G.D. (2003). *Bayesian Reasoning in Data Analysis: A Critical Introduction*. River Edge: World Scientific.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. (2nd ed.) New York : Wiley.

Albert, J.H.(1992). Bayesian estimation of the Polychoric Correlaiton Coefficient. *Journal of StatisticalComputation and Simulation. 44,* 47-61

Aldrich, A. (2008). R. A. Fisher on Bayes and Bayes' Theorem. *Bayesian analysis, 3(1)*,161–170.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *The Journal of Wildlife Management*, 912–923.

Attfield, M.D. & Hewett, P. (1992). Exact Expressions for the Bias and Variance of Estimators of a Lognormal Distribution. *Journal of the American Industrial Hygiene Association, 53,* 432-435.

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385–427). Charlotte, NC: Information Age Publishing Inc.

Barnard, G. A. (1987). R. A. Fisher—A true Bayesian? *International Statististics Review. 55.* 183–189.

Bem, D., Utts, J. & Johnson, W.O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology, Vol 101(4),* 716-719.

Berger, J.O. (1990). Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference, 25.* 303-328.

Berger, J.O. (2000). Bayesian Analysis: A Look at Today and Thoughts on Tomorrow. *Journal of the American Statistical Association, 95*(452), 1269-1276.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122.

Bollen, K., & Barb, K. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232–239.

Brooks, S., Gelman, A., Jones, G. L., & Meng, X. L. (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL : Chapman & Hall

Bunge, M. (2012). Does Inductive Logic Work? *Evaluating Philosophies*, 115–118

Camp, B. H. (1933). Karl Pearson and Mathematical Statistics. *Journal of the American Statisical. Association. 28* , 395–401.

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician, 46,* 167–174

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, *82*(397), 106–111.

Chechile, R. (2011). Likelihood and posterior identification: Implications for mathematical psychology. *British Journal of Mathematical and Statistical Psychology*, *30*(2), 177–184.

Chen, J., & Choi, J. (2009). A Comparison of Maximum Likelihood and Expected A Posteriori Estimation for Polychoric Correlation Using Monte Carlo Simulation. *Journal of Modern Applied Statistical Methods, 8,* 337-354.

Choi, J., Kim, S., Chen, J., & Dannels, S. (2011). A Comparison of Maximum Likelihood and Bayesian Estimation for Polychoric Correlation Using Monte Carlo Simulation. *Journal of Educational and Behavioral Statistics*, *36*(4), 523–549.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.) New Jersey, NY: Lawrence Erlbaum Associates

Cohen, J. (1994). The earth is round (p< .05). *American Psychologist, 49*, 997–1003.

Crovelli, M. R. (2011). Can Probability be Subjective and Objective at the Same Time-A Reply to Arnold Baise. *Libertarian Papers, 3*, 1.

Cox, RT (1946). Probability, frequency and reasonable expectation. *American Journal of Physics, 14(1)*, 21-33

Dale, A. I. (1999). *A History of Inverse Probability*. (2nd ed) New York : Springer.

De Finetti, B. (1974). *Theory of probability.* New York : Wiley

Dehghani, M. Iliev, R. & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language, 27(1)*, 55-85.

Dempster, A.P. (2005). Bayesian Methods. In P. Armitage and T. Colton (eds.), *Encyclopedia of Biostatistics.* (2nd ed). New York : Wiley

Diaconis, P. (2009). The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, *46*(2), 179–205.

Edwards, A.W.F. (2004). Comment on D. R. Bellhouse "The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth"." *Statistical Science, 19*, 34-37.

Efron, B. (1986). Why isn't everyone a Bayesian? *American Statistician 40*. 1–5.

Ekstrom, J. (2008). *A generalized definition of the polychoric correlation coefficient*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database.

Fienberg, S. E. (1997). Introduction to R. A. Fisher on inverse probability and likelihood. *Statistical Science. 12*. 161.

Fisher, R.A. (1970). *Statistical Methods for Research Workers.* (15th ed.). New York : Macmillan

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods, 9(4),* 466-491

Gelman, A. & Shalizi, C. R. Philosophy and the practice of Bayesian statistics. (2011). Unpublished manuscript

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*(3), 647–674.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4*, 108–120.

Greenland, S. (2001). Putting Background Information about Relative Risks into conjugate Prior Distributions. *Biometrics, 57*, 663–70.

Headrick, T. & Pant, M.D. (2012). Simulating non-normal distributions withspecified L-moments and L-correlations. *Statistica Neerlandica, 66(4),* 422-441.

Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*, 329–367.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling, 5,* 344–364.

Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physical Review, 106* (4): 620–630

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. New York, NY : Chapman & Hall

Jordan, M.I. (2011). What are the open problems in Bayesian statistics? *The ISBA Bulletin, 18(1)*.

Joreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59,* 381–389

Karni, E. (2011). A theory of Bayesian decision making with action-dependent subjective probabilities. *Economic Theory*, *48*(1), 125–146.

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Elsevier.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, *15*(4), 722–752.

Leahey, T.H. (1991). *A history of modern psychology.* Englewood Cliffs, N.J. : Prentice Hall.

Lee, J. J. (2011). Demystify statistical significance—time to move on from the P value to Bayesian analysis. *Journal of the National Cancer Institute*, *103*(1), 2–3.

Lynch, S. M. (2010). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer.

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods, 17*, 340–345.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*(9), 1–21.

Matthews, W. J. (2011). What might judgment and decision making research be like if we took a Bayesian approach to hypothesis testing. *Judgment and Decision Making*, *6*(8), 843–856.

Monahan, T. & Jill A. F. (2010). Benefits of "Observer Effects": Lessons from the Field'. *Qualitative Research 10(3)*, 357–76

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological methods, 16(4)*, 406.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17(3),* 313.

Muthen, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika, 53,* 563–578.

Muthen, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor-analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19–30.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44* , 443–460.

Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford : University Press

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society. 195* , 1–47.

Quiroga, A. M. (1992). *Studies of the polychoric correlation and other correlation measures for ordinal variables*. Unpublished doctoral dissertation, Acta Universitatis Upsaliensis.

Rigdon, E. E., & Ferguson, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28*, 491–497.

Risinger, D.M. (2012). Reservations about likelihood ratios. *Law, Probability and Risk.* First published online August 8, 2012 doi:10.1093/lpr/mgs011

Ritchie-Scott, A. (1918). The correlation coefficient of a polychoric table. *Biometrika,12* , 93–133.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11*, 424–451.

Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations? *Structural Equation Modeling, 18*, 253-273.

Shen, D. (2009). Science and Culture and the Subjectivity Separation from the Perspective of Social Epistemology. *Journal of Huaiyin Teachers College (Social Sciences Edition)*, *6*, 011.

Spanos, A. (2011). A frequentist interpretation of probability for model-based inductive inference. *Synthese*, *2*, 1–31.

Stephens, P. A., Buskirk, S. W., Hayward, G. D., & Martinez Del Rio, C. (2005). Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, *42*(1), 4–12.

Strzalecki, T. (2011). Axiomatic foundations of multiplier preferences. *Econometrica*, *79*(1), 47–73.

Tadikamalla, P.R. (1980) On simulating nonnormal distributions. *Psychometrika, 45*, 273-279.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*, 465–471

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Zumbo, B. D., & Harwell, M. R. (1999). The methodology of methodological research: Analyzing the results of simulation experiments (No. ESQBS-99-2). Prince George, BC: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C.. (2007). Ordinal Versions of Coefficients Alpha and Theta For Likert Rating Scales. *Journal of Modern Applied Statistical Methods*, *6*, 21-29.

Appendix

R SYNTAX

```
library(MASS)
library(mvtnorm)
library(truncnorm)
library(polycor)
library(e1071)
library(copula)

##############################################################
#     Log Likelihood function of Phi and its derivatives
#
#Calculated from subsection "The conditional distribution of
#rho given (c,d,theta)", page 51-52
##############################################################


NegLoglikOfPhi<-function(phi, n, s11, s12, s22){

  -( -n/2 * log(phi) + (n - 2) * log(phi + 1) - ((s11+s22)/8 - s12/4)
* phi - ((s11+s22)/8 + s12/4)/phi )

}


NegGradOfPhi<-function(phi, n, s11, s12, s22){

  -(-n/(2 * phi) + (n - 2)/(phi + 1) - (s11 + s22)/8 + s12/4 + ((s11 +
s22)/8 + s12/4)/phi^2 )

}



DerOfPhi<-function(phi, n, s11, s12, s22){

  L.phi<-( -n/2 * log(phi) + (n - 2) * log(phi + 1) - ((s11+s22)/8 -
s12/4) * phi - ((s11+s22)/8 + s12/4)/phi )
  Grad.phi<-(-n/(2 * phi) + (n - 2)/(phi + 1) - (s11 + s22)/8 + s12/4
+ ((s11 + s22)/8 + s12/4)/phi^2 )
  Hess.phi<-( n/(2 * phi^2) - (n - 2)/(phi + 1)^2 - ((s11 + s22)/4 +
s12/2)/phi^3 )
  ThirdDer.phi<-( -n/phi^3 + 2 * (n - 2)/(phi + 1)^3 + 3 * ((s11 +
s22)/4 + s12/2)/phi^4 )

  list(L.phi = L.phi, Grad.phi = Grad.phi, Hess.phi = Hess.phi,
ThirdDer.phi = ThirdDer.phi)
}
```

```
##Calculating the mode of Phi

NROfPhi<-function(n, s11, s12, s22, start){

  max.try<-100

  while(max.try > 0){
    max.try<-max.try-1
    op<-optim(par = runif(1,-start,20-start), fn=NegLoglikOfPhi,
gr=NegGradOfPhi, lower = 1e-3, upper = 50, method = "Brent", n = n,
s11 = s11, s12 = s12, s22 = s22)
    if(op$convergence == 0){
      return(op$par)
    }
  }

  if(op$convergence != 0){
    print("Error in finding phi")
    stop()
  }

}

############################################################
#     Log Likelihood function of Tau and its derivatives
#
#Calculated from subsection "The conditional distribution of
#rho given (c,d,theta)", page 51-52
############################################################

NegLoglikOfTau<-function(tau, n, s11, s12, s22, nu){

  phi<-(nu * tau + 1)^(1/nu)
  der<-DerOfPhi(phi, n = n, s11 = s11, s12 = s12, s22 = s22)
  L.phi<-der$L.phi
  Grad.phi<-der$Grad.phi
  Hess.phi<-der$Hess.phi

  d.phi<-(nu * tau + 1)^(1/nu - 1)

  L.tau<-L.phi + log(d.phi)

  -L.tau

}

NegGradOfTau<-function(phi, n, s11, s12, s22, nu){
```

```
  phi<-(nu * tau + 1)^(1/nu)
  der<-DerOfPhi(phi, n = n, s11 = s11, s12 = s12, s22 = s22)
  L.phi<-der$L.phi
  Grad.phi<-der$Grad.phi
  Hess.phi<-der$Hess.phi

  d.phi<-(nu * tau + 1)^(1/nu - 1)
  Grad.tau<-Grad.phi * d.phi + (1 - nu)/(nu * tau + 1)

  -Grad.tau

}

DerOfTau<-function(tau, n, s11, s12, s22, nu){

  phi<-(nu * tau + 1)^(1/nu)
  der<-DerOfPhi(phi, n = n, s11 = s11, s12 = s12, s22 = s22)
  L.phi<-der$L.phi
  Grad.phi<-der$Grad.phi
  Hess.phi<-der$Hess.phi

  d.phi<-(nu * tau + 1)^(1/nu - 1)
  d2.phi<-(1 - nu) * (nu * tau + 1)^(1/nu - 2)

  L.tau<-L.phi + log(d.phi)
  Grad.tau<-Grad.phi * d.phi + (1 - nu)/(nu * tau + 1)
  Hess.tau<-Hess.phi * d.phi^2 + Grad.phi * d2.phi - nu * (1 - nu)/(nu
* tau + 1)^2

  list(L.tau = L.tau, Grad.tau = Grad.tau, Hess.tau = Hess.tau)

}


## Calculating the mode of Tau

NROfTau<-function(n, s11, s12, s22, nu, start){

  max.try<-100

  lower<-min((.001^nu-1)/nu, (10^nu-1)/nu)
  upper<-max((.001^nu-1)/nu, (10^nu-1)/nu)

  while(max.try > 0){
    max.try<-max.try-1
    op<-optim(par = runif(1,-start,20-start), fn=NegLoglikOfTau,
gr=NegGradOfTau, method = "Brent", lower = lower, upper = upper, n =
n, s11 = s11, s12 = s12, s22 = s22, nu = nu)
    if(op$convergence == 0){
      return(op$par)
```

```
    }
  }

  if(op$convergence != 0){
    print("Error in finding tau")
    stop()
  }

}

#############################################################
#     Sampling of Rho, Theta, and the Thresholds
#
#Based on section 2.1 and 2.2
#############################################################



SampleRho<-function(c, d, theta1, theta2, rho){

  n<-length(theta1)
  s11<-sum(theta1^2)
  s12<-sum(theta1 * theta2)
  s22<-sum(theta2^2)

  phi0<-NROfPhi(n = n, s11 = s11, s12 = s12, s22 = s22, start =
(1+rho)/(1-rho))

  der<-DerOfPhi(phi0, n = n, s11 = s11, s12 = s12, s22 = s22)
  ThirdDer.phi<-der$ThirdDer.phi
  Hess.phi<-der$Hess.phi

  nu<-1 + phi0 * ThirdDer.phi/(3 * Hess.phi)

  mu<-NROfTau(n = n, s11 = s11, s12 = s12, s22 = s22, nu = nu, start =
(phi0^nu - 1)/nu)

  Hess.tau<-DerOfTau(mu, n = n, s11 = s11, s12 = s12, s22 = s22, nu =
nu)$Hess.tau
  sigma2<-1/(-Hess.tau)

  max.try<-100
  while(max.try > 0){
    max.try<-max.try - 1
    tau<-rnorm(1, mu, sqrt(sigma2))
    if(nu * tau + 1 > 0){
      rho<-((nu * tau + 1)^(1/nu) - 1)/((nu * tau + 1)^(1/nu) + 1)
      max.try<-0
    }
  }
```

```
  if(is.na(rho)){
    stop("Error in SampleRho")
  }


  rho

  }



SampleTheta<-function(rho, c, d, x, y, theta1, theta2){

  n<-length(x)
  new.theta1<-NULL
  new.theta2<-NULL
  tmp<-theta2[1]
  for(i in 1:n){

    tmp1<-rtruncnorm(1, a = c[x[i]], b = c[x[i]+1], mean =
rho*theta2[i], sd = sqrt(1-rho^2))


    if(tmp1 < c[x[i]] || tmp1 > c[x[i]+1]){
      print("error in sampling theta")
      stop()
    }

    new.theta1<-c(new.theta1, tmp1)

    tmp2<-rtruncnorm(1, a = d[y[i]], b = d[y[i]+1], mean = rho*tmp1,
sd = sqrt(1-rho^2))
    if(tmp2 < d[y[i]] || tmp2 > d[y[i]+1]){
      print("error in sampling theta")
      stop()
    }
    new.theta2<-c(new.theta2, tmp2)
  }

  list(theta1 = new.theta1, theta2 = new.theta2)

}

SampleThreshold<-function(theta1, theta2, rho, c, d){

  new.c<-c(-Inf)
  for(i in 2:(length(c)-1)){
    upper.c<-min(c[i+1], min(theta1[theta1>=c[i]]))
    lower.c<-max(c[i-1], max(theta1[theta1<=c[i]]))
```

```
    new.c<-c(new.c, runif(1, lower.c, upper.c))
  }
  new.d<-c(-Inf)
  for(i in 2:(length(d)-1)){
    upper.d<-min(d[i+1], min(theta2[theta2>=d[i]]))
    lower.d<-max(d[i-1], max(theta2[theta2<=d[i]]))
    new.d<-c(new.d, runif(1, lower.d, upper.d))
  }

  new.c<-c(new.c, Inf)
  new.d<-c(new.d, Inf)

  list(c = new.c, d = new.d)

}

## Section 2.2

Gibbs<-function(x, y, iter = 1e4, trace = FALSE){

  library("truncnorm")

  n<-length(x)

  rho<-0.5
  c<-c(-Inf, qnorm(as.vector(cumsum(table(x)))/n, lower.tail = T))
  d<-c(-Inf, qnorm(as.vector(cumsum(table(y)))/n, lower.tail = T))

  tmp.c<-c
  tmp.c[1]<-c[2]-.1
  tmp.c[length(tmp.c)]<-c[length(c)-1]+.1
  tmp.d<-d
  tmp.d[1]<-d[2]-.1
  tmp.d[length(tmp.d)]<-d[length(d)-1]+.1
  theta1<-NULL
  theta2<-NULL
  for(i in 1:n){
    theta1<-c(theta1, runif(1, tmp.c[x[i]], tmp.c[x[i]+1]))
    theta2<-c(theta2, runif(1, tmp.d[y[i]], tmp.d[y[i]+1]))
  }

  gibbs.rho<-NULL
  gibbs.c<-NULL
  gibbs.d<-NULL

  for(i in 1:iter){
    st<-SampleTheta(rho, c, d, x, y, theta1, theta2) # Formula 2.4 in
paper
    theta1<-st$theta1
    theta2<-st$theta2
```

```
    rho<-SampleRho(c, d, theta1, theta2, rho) # Corresponds to
conditional posterior distribution of rho on page 51, includes
formulae 2.2 and 2.3
    scd<-SampleThreshold(theta1, theta2, rho, c, d)
    c<-scd$c
    d<-scd$d
    gibbs.rho<-c(gibbs.rho, rho)
    gibbs.c<-rbind(gibbs.c, c[2:(length(c)-1)])
    gibbs.d<-rbind(gibbs.d, d[2:(length(d)-1)])
    if(trace && i %% 100 == 0){

      cat(i)
      cat(" iterations completed.")
      cat("\n")
    }
  }

  list(rho = gibbs.rho, c = gibbs.c, d = gibbs.d)

}


#############################################################
#     Function to calculate Batch Means
#############################################################



BatchMeans<-function(vals,bs = "sqroot",warn = FALSE){

  N<-length(vals)
  if(N < 1000){
    if(warn){ # if warning
      cat("WARNING: too few samples (less than 1000)\n")
    }
    if(N < 10){
      return(NA)
    }
  }

  if(bs=="sqroot"){
    b<-floor(sqrt(N)) # batch size
    a<-floor(N/b) # number of batches
  }else{
    if(bs=="cuberoot"){
      b<-floor(N^(1/3)) # batch size
      a<-floor(N/b) # number of batches
    }else{ # batch size provided
      stopifnot(is.numeric(bs))
      b<-floor(bs) # batch size
```

```
      if(b > 1){ # batch size valid
        a <- floor(N/b) # number of batches
      }else{
        stop("batch size invalid (bs=",bs,")")
      }
    }
  }

  Ys<-sapply(1:a,function(k) return(mean(vals[((k-1)*b+1):(k*b)])))

  muhat<-mean(Ys)
  sigmahatsq<-b*sum((Ys-muhat)^2)/(a-1)

  bmse<-sqrt(sigmahatsq/N)

  bmse

}


############################################################
#     Utility functions - convert
#Converts a contingency table to two vectors
############################################################



convert<-function(tbl){

  data<-NULL
  for(i in 1:nrow(tbl)){
    for(j in 1:ncol(tbl)){
      data<-rbind(data, matrix(rep(c(i, j), tbl[i, j]), ncol = 2,
byrow = T))
    }
  }
  data<-as.data.frame(data)
  colnames(data)<-c("x", "y")
  data

}


#########################################################################
#######################################
#          Main Function
#
#This function calculates polychoric correlation using the Gibbs
Sampling method
#
```

```
#Arguments:
#u1: It can be a contingency table or a vector of the first ordinal
variable
#u2: If u1 is a vector, then u2 must be a un-null vector given the
second ordinal variable
#iter: number of iterations of Gibbs sampling
#t0: First t0 samplings will be excluded due to non-convergence
#everyN: Only every N-th sampling will be taken
#trace: If TRUE, the program will print out the number of iterations
executed for every 100 iterations
#graph: If TRUE, a graph will be printed showing the sampling history
of the correlation and thresholds
#
#Outputs: rho, c, d
#rho is the vector of Gibbs sampled rho
#c is the vector of thresholds of X
#d is the vector of thresholds of Y
########################################################################
#######################################

polycorGibbs<-function(u1, u2 = NULL, iter = 1e4, t0 = 200, everyN=10,
trace = FALSE, graph = FALSE){

  if(class(u1) == "table"){
    data<-convert(u1)
    x<-data$x
    y<-data$y
  }else{
    if(class(u1) == "integer"){
      if(is.null(u2)){
        print("The argument u2 cannot be null ")
      }else{
        x<-u1
        y<-u2
      }
    }else{
      print("Invalid input")
      return(NULL)
    }
  }

  if (length(unique(x))==1) {
    print("Error: x is a constant vector")
    return(NULL)
  }

  if (length(unique(y))==1) {
    print("Error: y is a constant vector")
    return(NULL)
  }
```

```
  if (length(x) != length(y)){
    print("Please make sure the length of the two input vectors are
the same!")
    return(NULL)
  }

  xd <- sort(unique(x))

  yd <- sort(unique(y))

  x_rec<-x
  y_rec<-y

  for(i in 1:length(x)){
    for (j in 1:length(xd)){
      if (x[i] == xd[j]){
       x_rec[i]<-j
      }
    }
  }

  for(i in 1:length(y)){
    for (j in 1:length(yd)){
      if (y[i] == yd[j]){
       y_rec[i]<-j
      }
    }
  }

  x_rec = as.integer(x_rec)
  y_rec = as.integer(y_rec)

  x<-x_rec
  y<-y_rec

  plc<-Gibbs(x, y, iter, trace)
  rho<-as.matrix(plc$rho, ncol = 1)
  c<-as.matrix(plc$c, ncol = length(unique(x_rec))-1)
  d<-as.matrix(plc$d, ncol = length(unique(y_rec))-1)

  res<-data.frame(rho, c, d)
  res<-res[-(1:t0), ]
  res<-res[seq(1,nrow(res),by= everyN),]
  colnames(res)<-c("rho", paste("c", 1:(length(unique(x))-1), sep =
""), paste("d", 1:(length(unique(y))-1), sep = ""))
  mn<-signif(apply(res, 2, mean), 3)
  md<-signif(apply(res, 2, median), 3)
  sd<-signif(apply(res, 2, BatchMeans), 3)
  qsd<-signif(apply(res, 2, sd), 3)
```

```
  cor<-signif(cor(res), 3)

  cor[lower.tri(cor)]<-NA
  diag(cor)<-NA

  smy<-cbind(mn, md, sd, qsd)
  colnames(smy)<-c("mean", "median", "SD", "Numeric SD")

  cat("Summary\n")
  print(smy)
  cat("\nCorrelation\n")
  print(cor)

  if(graph){
    #par(mfrow = c(1, 1))
    plot(rho, type = "l", xlab = "Round", ylab = "rho", col = "red")
    abline(h = smy["rho", "mean"])
    windows()
    for(i in 1:ncol(c)){
      if(i==1){
        plot(c[, i], type = "l", ylim = c(min(c), max(c)), xlab =
"Round", ylab = "Threshold of X", col = i+1)
        abline(h = smy[paste("c", i, sep = ""), "mean"])
      }else{
        lines(c[, i], type = "l", col = i+1)
        abline(h = smy[paste("c", i, sep = ""), "mean"])
      }
    }
    windows()
    for(i in 1:ncol(d)){
      if(i==1){
        plot(d[, i], type = "l", ylim = c(min(d), max(d)), xlab =
"Round", ylab = "Threshold of Y", col = i)
        abline(h = smy[paste("d", i, sep = ""), "mean"])
      }else{
        lines(d[, i], type = "l", col = i)
        abline(h = smy[paste("d", i, sep = ""), "mean"])
      }
    }
  }
  plc

}
```

Technical Appendix

The following paragraph is just a quick overview of the general framework described in Albert(1992) to estimate the polychoric correlation coefficient. To obtain the posterior distribution of the correlation ρ, it is necessary to first apply the transformation $\phi = \frac{1+\rho}{1-\rho}$ such that $\phi > 0$. Then we proceed to obtain the mode of the distribution of ϕ, denoted as $\hat{\phi}$, using the Newton-Raphson algorithm. To correct for the skewness of the distribution about the mode, it is necessary to apply the power transformation $\tau = (\phi^\nu - 1)/\nu$, described in Albert(1989), where $\nu = 1 + [l'''(\hat{\phi})\hat{\phi}]/[3l''(\hat{\phi})]$. The posterior distribution of τ is approximately $N(\mu, \sigma^2)$, where μ is the mode of the distribution of τ and $\sigma^2 = (-l_2''(\mu))^{-1}$. Here l_2 is the log posterior density of τ. A sampler of ρ can then be developed by, first, generating Z from $N(\mu, \sigma^2)$ and then setting $\rho = ((\nu Z + 1)^{1/\nu} - 1)/((\nu Z + 1)^{1/\nu} + 1)$. This technical appendix details the derivation and implementation of the log likelihoods of ϕ and τ and their first, second, and third derivatives. The exact forms of ν, μ and σ^2 are also shown.

Now that $\phi = \frac{1+\rho}{1-\rho}$, we have $\rho = \frac{\phi-1}{\phi+1}$, and the Jacobian term is $J = |\frac{d}{d\phi}\frac{\phi-1}{\phi+1}| = \frac{2}{(1+\phi)^2}$. We then need to replace ρ with $\frac{\phi-1}{\phi+1}$ in Equation (2.2) in Albert (1992) and multiply it times the Jacobian to obtain the posterior distribution of ϕ or $\pi_1(\phi)$, in Albert's notation. Note that there is a minor typo in Equation (2.2). It should be

$$\pi(\rho|(c, d, \theta)) = C(1 - \rho^2)^{-n/2} exp\{-\frac{1}{2(1 - \rho^2)}(S_{\xi\xi} - 2\rho S_{\xi\eta} + S_{\eta\eta})\}$$

Making the previously-mentioned substitution we have:

$$\pi_1(\phi) = C[\frac{4\phi}{(\phi + 1)^2}]^{-\frac{n}{2}} exp\{-\frac{1}{2\frac{4\phi}{(\phi+1)^2}}(S_{\xi\xi} + S_{\eta\eta} - 2\frac{\phi - 1}{\phi + 1}S_{\xi\eta})\}\frac{2}{(\phi + 1)^2}$$

Taking the logarithm of $\pi_1(\phi)$, we can obtain the log likelihood of ϕ as:

$$l_1(\phi) = C - \frac{n}{2}log\phi + (n-2)log(\phi+1) - \frac{(\phi+1)^2}{8\phi}(S_{\xi\xi} + S_{\eta\eta}) + \frac{(\phi^2-1)}{4\phi}S_{\xi\eta}$$

where C stands for the constant that makes the likelihood proper. Expanding the terms $\frac{(1+\phi)^2}{8\phi}$ as $\frac{1}{8}(\frac{1}{\phi} + 2 + \phi)$, and $\frac{(1-\phi^2)}{4\phi}$ as $\frac{1}{4}(\frac{1}{\phi} - \phi)$, we can then take the derivative and obtain the the first derivative of l_1 as:

$$l_1'(\phi) = -\frac{n}{2\phi} + \frac{n-2}{\phi+1} - \frac{1}{8}(1 - \frac{1}{\phi^2})(S_{\xi\xi} + S_{\eta\eta}) + \frac{1}{4}(1 + \frac{1}{\phi^2})S_{\xi\eta}$$

The second derivative of l_1 is:

$$\begin{aligned} l_1''(\phi) &= \frac{n}{2\phi^2} - \frac{n-2}{(\phi+1)^2} - \frac{1}{8}(-\frac{1}{\phi^3}(-2))(S_{\xi\xi} + S_{\eta\eta}) + \frac{1}{4}(-\frac{2}{\phi^3})S_{\xi\eta} \\ &= \frac{n}{2\phi^2} - \frac{n-2}{(\phi+1)^2} - \frac{1}{4\phi^3}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta}) \end{aligned}$$

and the third derivative of l_1 is:

$$l_1'''(\phi) = \frac{-n}{\phi^3} + \frac{2(n-2)}{(\phi+1)^3} + \frac{3}{4\phi^4}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta})$$

Which is multiplied by ϕ to obtain Albert's suitable re-expression of ν.

$$l_1'''(\phi)\phi = \frac{-n}{\phi^2} + \frac{2(n-2)\phi}{(\phi+1)^3} + \frac{3}{4\phi^3}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta})$$

Through algebra it is possible to obtain and simplify to:

$$\begin{aligned} \nu &= 1 + [l_1'''(\hat{\phi})\hat{\phi}]/[3l_1''(\hat{\phi})] \\ &= 1 + \frac{-\frac{n}{\hat{\phi}^2} + \frac{2(n-2)\hat{\phi}}{(\hat{\phi}+1)^3} + \frac{3}{4\hat{\phi}^3}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta})}{\frac{3n}{2\hat{\phi}^2} - \frac{3(n-2)}{(\hat{\phi}+1)^2} - \frac{3}{4\hat{\phi}^3}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta})} \\ &= \frac{1}{3}\left(\frac{\frac{n}{\hat{\phi}^2} - \frac{2(n-2)(\hat{\phi}+3)}{(\hat{\phi}+1)^3}}{\frac{n}{\hat{\phi}^2} - \frac{2(n-2)}{(\hat{\phi}+1)^2} - \frac{1}{2\hat{\phi}^3}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta})}\right) \end{aligned}$$

Note that $\hat{\phi}$ is found using Newton-Raphson steps of the form $\phi_{i+1} = \phi_i - l_1'(\phi_i)/l_1''(\phi_i)$, as described by Albert(1992).

Now $\tau = \frac{\phi^{\nu}-1}{\nu}$. Replace ϕ with $(\nu\tau+1)^{1/\nu}$ in the $\pi_1(\phi)$ and multiply it times the Jacobian term $(\nu\tau+1)^{1/\nu-1}$, to get the distribution function of τ. The log likelihood of τ is given by:

$$\begin{aligned} l_2(\tau) = c &+ (\frac{2-n}{2\nu} - 1)log(\nu\tau+1) + (n-2)log[(\nu\tau+1)^{1/\nu}+1] \\ &- \frac{1}{8}[(\nu\tau+1)^{\frac{1}{\nu}} + 2 + (\nu\tau+1)^{-\frac{1}{\nu}}](S_{\xi\xi} + S_{\eta\eta}) \\ &+ \frac{1}{4}[(\nu\tau+1)^{\frac{1}{\nu}} - (\nu\tau+1)^{-\frac{1}{\nu}}]S_{\xi\eta} \end{aligned}$$

Taking the first derivative of $l_2(\tau)$ we obtain:

$$\begin{aligned} l_2'(\tau) = &\frac{2-n-2\nu}{2(\nu\tau+1)} + \frac{(n-2)(\nu\tau+1)^{1/\nu-1}}{(\nu\tau+1)^{1/\nu}+1} \\ &- \frac{1}{8}(\nu\tau+1)^{1/\nu-1}(S_{\xi\xi} + S_{\eta\eta} - 2S_{\xi\eta}) \\ &+ \frac{1}{8}(\nu\tau+1)^{-1/\nu-1}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta}) \end{aligned}$$

The second derivative of $l_2(\tau)$ is:

$$\begin{aligned} l_2''(\tau) = &\frac{(2-n-2\nu)(-\nu)}{2(\nu\tau+1)^2} + \frac{(n-2)(\nu\tau+1)^{1/\nu-2}\{1-\nu[1+(\nu\tau+1)^{1/\nu}]\}}{[(\nu\tau+1)^{1/\nu}+1]^2} \\ &- \frac{1}{8}(1-\nu)(\nu\tau+1)^{1/\nu-2}(S_{\xi\xi} + S_{\eta\eta} - 2S_{\xi\eta}) \\ &- \frac{1}{8}(1+\nu)(\nu\tau+1)^{-1/\nu-2}(S_{\xi\xi} + S_{\eta\eta} + 2S_{\xi\eta}) \end{aligned}$$

In this case, the value of μ, the mode of the distribution of τ, is found by using Newton-Raphson with the form of $\tau_{i+1} = \tau_i - l_2'(\tau_i)/l_2''(\tau_i)$. As mentioned previously, $\sigma^2 = (-l_2''(\mu))^{-1}$. Now we have the exact form of $N(\mu, \sigma^2)$ which is the approximate distribution for τ. So for each drawing of ρ, we need to find the values of $\hat{\phi}$, ν, μ, and σ^2, which depend on the samplers of (ξ_h, η_h) for $h = 1, ..., n$.