Estimating the Correlation in Bivariate Normal Data with Known Variances and Small Sample Sizes [1]

Bailey K. Fosdick and Adrian E. Raftery
Department of Statistics
University of Washington

Technical Report No. 591

January 29, 2012

[1]Bailey K. Fosdick is Graduate Research Assistant and Adrian E. Raftery is Professor of Statistics and Sociology, both at the Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322 (Email: bfosdick@u.washington.edu/raftery@u.washington.edu). This work was supported by NICHD grant R01 HD54511. Raftery's research was also partially supported by NIH grant R01 GM084163, and NSF grants ATM0724721 and IIS0534094. The authors thank Sam Clark, Jon Wellner and the Probabilistic Population Projections Group at the University of Washington for helpful comments and discussion.

Abstract

We consider the problem of estimating the correlation in bivariate normal data when the means and variances are assumed known, with emphasis on the small sample case. We consider eight different estimators, several of them considered here for the first time in the literature. In a simulation study, we found that Bayesian estimators using the uniform and arc-sine priors outperformed several empirical and exact or approximate maximum likelihood estimators in small samples. The arc-sine prior did better for large values of the correlation. For testing whether the correlation is zero, we found that Bayesian hypothesis tests outperformed significance tests based on the empirical and exact or approximate maximum likelihood estimators considered in small samples, but that all tests performed similarly for sample size 50. These results lead us to suggest using the posterior mean with the arc-sine prior to estimate the correlation in small samples when the variances are assumed known.

KEYWORDS: Arc-sine prior; Bayes factor; Bayesian test; Maximum likelihood estimator; Uniform prior; Jeffreys prior.

1 INTRODUCTION

Sir Francis Galton defined the theoretical concept of bivariate correlation in 1885, and a decade later Karl Pearson published the formula for the sample correlation coefficient, also known as Pearson's r (Rodgers and Nicewander, 1988). The sample correlation coefficient is still the most commonly used measure of correlation today as it assumes no knowledge of the means or variances of the individual groups and is the maximum likelihood estimator for the correlation coefficient in the bivariate normal distribution when the means and variances are unknown.

In the event that the variances are known, information is lost by using the sample correlation coefficient. We cannot simply substitute the known variance quantities into the denominator of the sample correlation coefficient since that results in an estimator that is not the maximum likelihood estimator and has the potential to fall outside the interval $[-1, 1]$. When the variances are known, we seek an estimator that takes advantage of this information.

Kendall and Stuart (1979) noted that conditional on the variances, the maximum likelihood estimator of the correlation is the solution of a cubic equation. Sampson (1978) proposed a consistent, asymptotically efficient estimator based on the cubic equation that avoided the need to solve the equation directly. In a simulation study, we found that when the true correlation is zero and the sample size is small, the variances of these estimators are undesirably large. This led us to search for more stable estimates of the correlation, which condition on the known variances and perform well when sample sizes are small.

Our interest in this problem arose in the context of probabilistic population projections. Alkema et al. (2011) developed a Bayesian hierarchical model for projecting the total fertility rate (TFR) in all countries. This model works well for projecting the TFR in individual countries. However, for creating aggregated regional projections, there was concern that excess correlation existed between the country fertility rates that was not accounted for in the model. To investigate this we considered correlations between the normalized forecast errors in different countries, conditional on the model parameters. Often there were as few as five to ten data points to estimate the correlation. For each pair of countries, these errors were treated as samples from a bivariate normal distribution with means equal to zero and variances equal to one. Determining whether the correlations between the countries are non-zero, and if so estimating them, is necessary to assess the predictive distribution of aggregated projections.

In Section 2 we describe the estimators we consider, in Section 3 we give the results of our

simulation study, and in Section 4 we discuss alternative approaches.

2 ESTIMATORS OF CORRELATION

Let (X_i, Y_i), $i = 1, \ldots, n$ be independent and identically distributed observations from a bivariate normal distribution with means equal to zero, variances equal to one, and correlation unknown. We let $SSx = \sum_{i=1}^n X_i^2$, $SSy = \sum_{i=1}^n Y_i^2$, and $SSxy = \sum_{i=1}^n X_i Y_i$ and consider eight estimators of the correlation.

The first estimator is the maximum likelihood estimator for bivariate normal data when the variances are unknown. We refer to this as the *sample correlation coefficient* even though we have conditioned on the means being zero. This estimator is defined as follows:

$$\hat{\rho}^{(1)} = \frac{\frac{\sum_{i=1}^n X_i Y_i}{n}}{\sqrt{(\frac{\sum_{i=1}^n X_i^2}{n})(\frac{\sum_{i=1}^n Y_i^2}{n})}} = \frac{SSxy}{\sqrt{SSx\ SSy}}.$$

The second estimator is a modification of the first estimator, where we assume the variances are known to be equal to one. We name this estimator the *empirical estimator with known variances* and define it as:

$$\hat{\rho}^{(2)} = \frac{\sum_{i=1}^n X_i Y_i}{n} = \frac{SSxy}{n}.$$

This estimator is unbiased yet is not guaranteed to fall in $[-1, 1]$, especially for small samples. This unappealing property motivated us to define the third estimator called the *truncated empirical estimator with known variances*, $\hat{\rho}^{(3)}$, where the second estimator is truncated at -1 if it falls below -1 and at 1 if it falls above 1.

The *maximum likelihood estimator (MLE)* when the means are known to be zero and variances are known to be one is the fourth estimator. This estimator is found by solving the cubic equation

$$0 = \rho^3 - \rho^2 \frac{SSxy}{n} - \rho \frac{(n - SSx - SSy)}{n} - \frac{SSxy}{n}, \tag{1}$$

which results from setting the derivative of the log-likelihood equal to zero. If we define

$$\psi \equiv \psi(SSx, SSy, SSxy) = -3n(n - SSx - SSy) - SSxy^2, \quad \text{and}$$

$$\gamma \equiv \gamma(SSx, SSy, SSxy) = -36n^2 SSxy + 9nSSx \times SSxy + 9nSSy \times SSxy - 2SSxy^3,$$

then the three roots of this equation can be written fairly compactly, as follows:

$$\rho_1^{(4)} = \frac{SSxy}{3n} + \frac{2^{1/3}(\psi)}{3n\left(\gamma + \sqrt{4(\psi)^3 + (\gamma)^2}\right)^{1/3}} - \frac{\left(\gamma + \sqrt{4(\psi)^3 + (\gamma)^2}\right)^{1/3}}{3 \times 2^{1/3} n},$$

$$\rho_2^{(4)} = \frac{SSxy}{3n} - \frac{\left(1 + i\sqrt{3}\right)(\psi)}{3 \times 2^{2/3} n\left(\gamma + \sqrt{4(\psi)^3 + (\gamma)^2}\right)^{1/3}} + \frac{\left(1 - i\sqrt{3}\right)\left(\gamma + \sqrt{4(\psi)^3 + (\gamma)^2}\right)^{1/3}}{6 \times 2^{1/3} n},$$

$$\rho_3^{(4)} = \frac{SSxy}{3n} - \frac{\left(1 - i\sqrt{3}\right)(\psi)}{3 \times 2^{2/3} n\left(\gamma + \sqrt{4(\psi)^3 + (\gamma)^2}\right)^{1/3}} + \frac{\left(1 + i\sqrt{3}\right)\left(\gamma + \sqrt{4(\psi)^3 + (\gamma)^2}\right)^{1/3}}{6 \times 2^{1/3} n}.$$

Kendall and Stuart (1979) noted that at least one of the roots above is real and lies in the interval $[-1, 1]$. However, it is possible that all three roots are real and in the admissible interval, in which case the likelihood can be evaluated at each root to determine the true maximum likelihood estimate. Based on whether $(SSxy/n)^2$ is bigger than $3(SSx/n + SSy/n - 1)$, and whether $\gamma/(2\psi)$ is bigger than 1, Madansky (1958) specified conditions under which each of the three roots is the maximum likelihood estimate.

Sampson (1978) acknowledged the effort involved in computing the maximum likelihood estimate when the variances are known and proposed an asymptotically efficient estimator of the correlation based solely on the coefficients in the cubic equation (1). Sampson's estimator does not necessarily fall in the interval $[-1, 1]$ so he suggested truncating the estimate to lie in the interval, as was done with the empirical estimator with known variances. This less computationally intensive estimator is referred to as *Sampson's truncated MLE approximation*, $\hat{\rho}^{(5)}$, and is the fifth estimator we consider.

The remaining three estimators are Bayesian. Our sixth estimator is the *posterior mean*

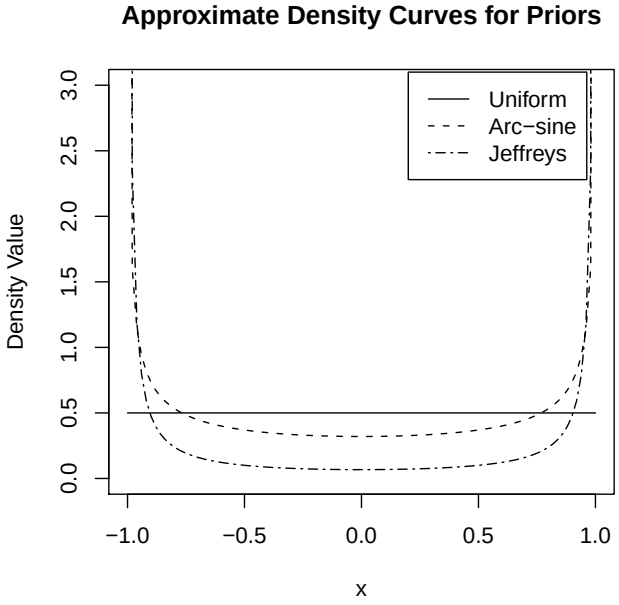

Figure 1: The density of each of the priors for the Bayesian estimators are shown. The Jeffreys curve is an approximation since it is not integrable on $[-1, 1]$. Observe that the arc-sine and Jeffreys priors are very similar, but the Jeffreys puts more weight on extreme values.

assuming a uniform prior, which has the form:

$$\hat{\rho}^{(6)} = E[\rho|X, Y] = \frac{\int_{-1}^{1} \frac{\rho}{2} \left(\frac{1}{2\pi\sqrt{1-\rho^2}} \right)^n \exp\left(-\frac{1}{2(1-\rho^2)} [SSx - 2\rho SSxy + SSy] \right) d\rho}{\int_{-1}^{1} \frac{1}{2} \left(\frac{1}{2\pi\sqrt{1-\rho^2}} \right)^n \exp\left(-\frac{1}{2(1-\rho^2)} [SSx - 2\rho SSxy + SSy] \right) d\rho}$$

where $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$. The denominator is the integral of the likelihood of the bivariate normal data multiplied by $1/2$, representing the Uniform$(-1, 1)$ prior, while the numerator is the same but with the integrand multiplied by ρ for the expectation.

Jeffreys (1961) described the improper prior, conditional on the variances, as:

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

This prior was the basis for the seventh estimator: the *posterior mean assuming a Jeffreys prior*, $\hat{\rho}^{(7)}$.

Finally, Jeffreys (1961) noted that the arc-sine prior,

$$\lambda_{\text{arc-sine}}(\rho) = \frac{1}{\pi} \frac{1}{\sqrt{1-\rho^2}},$$

is similar to the Jeffreys prior, but integrable on $[-1, 1]$. The *posterior mean assuming an arc-sine prior*, $\hat{\rho}^{(8)}$, represents the eighth, and final, estimator investigated.

Each of these priors is shown in Figure 1. The curve for the Jeffreys prior is an approximation since it is not integrable on $[-1, 1]$. Note that the arc-sine distribution on ρ is equivalent to placing a generalized beta (2, 1, 0.5, 0.5) of the first kind on $|\rho|$ (McDonald (1984)). Similarly, the uniform prior corresponds to a generalized beta (1, 1, 1, 1) of the first kind on $|\rho|$. Of these estimators, the empirical estimator with known variances and truncated empirical estimator with known variances are, to our knowledge, proposed here for the first time.

3 SIMULATION STUDY

3.1 Estimating the Correlation

Samples of sizes 5, 10, and 50 were generated from a bivariate normal distribution with means equal to zero, variances equal to one, and a specified correlation value. The estimators were first evaluated for positive and negative values of the correlation and were all found to be symmetric. Thus values of the correlation were sampled uniformly from symmetric intervals on $[-1, 1]$ to analyze how the estimators performed for different magnitudes of correlation. The estimators were compared based on root mean squared error using one million samples. The results are shown in Table 1.

Numerical issues arose when computing the integrals involved in the posterior mean estimators in cases where the true correlation value was extremely close to one in magnitude. To handle this, a tolerance of $10^{-6} \times n$ was put on the value of $|SSx + SSy \pm 2SSxy|$ since $SSx + SSy \pm 2SSxy = 0$ signifies a correlation of ∓ 1, respectively. When this tolerance was satisfied, the correlation estimate was given the appropriate value of 1 or -1. This approximation was used about ten times out of one million in the [0, 1] interval and thirty times out of one million in the [0.75, 1] interval for each sample size.

For the first column, since the correlations were drawn uniformly from the interval $[-1, 1]$, the Bayesian estimator assuming a uniform prior will have the lowest mean squared error according to theory. In samples of size 5, the uniform and arc-sine priors had superior performance over the entire $[-1, 1]$ interval compared to the other estimators, with a root mean squared error of about 0.3. The empirical estimator with known variances performed least well, whereas the maximum likelihood estimator and sample correlation coefficient performed similarly, with the sample correlation coefficient doing slightly better. This suggests that in small sample sizes, knowing the variances yields no improvement when using the maximum

Table 1: Root mean squared errors multiplied by 1000 are shown for each estimator based on one million simulated data sets (n=sample size). The estimators with the smallest root mean squared error are shown in bold for each sample size and each true correlation interval.

				$\lvert\rho\rvert$		
n	Estimator	[0,1]	[0,.25]	[.25,.50]	[.50,.75]	[.75,1]
5	Sample Correlation Coeff	352	442	406	326	172
	Emp w/ Known Var	516	452	479	529	595
	Trunc Emp w/ Known Var	387	419	399	369	358
	MLE	373	464	437	352	**161**
	Sampson's MLE Approx	382	462	435	357	232
	Mean w/ Uniform Prior	**297**	**289**	**315**	332	244
	Mean w/ Jeffreys Prior	311	358	354	**319**	182
	Mean w/ Arc-sine Prior	299	316	330	325	213
10	Sample Correlation Coeff	240	311	280	213	101
	Emp w/ Known Var	365	319	338	373	421
	Trunc Emp w/ Known Var	299	314	312	295	274
	MLE	248	334	295	**203**	**72**
	Sampson's MLE Approx	249	333	295	206	90
	Mean w/ Uniform Prior	**216**	**241**	**246**	227	124
	Mean w/ Jeffreys Prior	222	277	261	208	92
	Mean w/ Arc-sine Prior	217	254	251	219	109
50	Sample Correlation Coeff	104	139	122	88	39
	Emp w/ Known Var	163	143	151	167	188
	Trunc Emp w/ Known Var	150	143	151	161	145
	MLE	100	142	117	**75**	**29**
	Sampson's MLE Approx	100	142	117	**75**	**29**
	Mean w/ Uniform Prior	**97**	**129**	116	82	33
	Mean w/ Jeffreys Prior	98	135	**115**	78	30
	Mean w/ Arc-sine Prior	98	131	116	80	32

likelihood estimator.

However, when the correlations are decomposed by magnitude, a different story is told. For extreme correlation values, the sample correlation coefficient, maximum likelihood estimator and posterior mean assuming a Jeffreys prior had the smallest root mean squared errors. The Jeffreys prior is highly concentrated at extreme correlation values so we would expect it to outperform the other Bayesian estimators in the last interval. The posterior mean assuming an arc-sine prior and that assuming a uniform prior had root mean squared errors 1.3 and 1.5 times as large as that for the MLE, or best estimator. Conversely, at low values of correlation, the uniform and arc-sine posterior mean estimates had significantly lower root mean squared error than all other estimators. The posterior median estimators for each of the priors was also considered. Overall they performed very similar to the posterior mean estimates and hence are not included here.

In general, one does not know the magnitude of the correlation to be estimated, so an estimator that performs well for all levels of correlation is desired. Both the posterior mean assuming an arc-sine prior and that assuming a uniform prior had routinely low root mean squared error values when compared to the other estimators and were fairly consistent across the different correlation magnitudes. Therefore, we concluded that these should be the methods of choice for small sample sizes. One might argue that if estimating large correlations accurately is of greater interest then the posterior mean assuming the arc-sine prior should be used since it outperforms that with a uniform prior at extreme correlations.

As the sample size increased from 5 to 10 and from 10 to 50, the root mean squared errors decreased for all estimators, as expected. For samples of size 50, the root mean squared errors for correlations on the entire interval $[-1, 1]$ were low and effectively the same for all estimators except the empirical estimators when the variances are known. However, the estimators' performances by magnitude of the correlation still varied as in the case of samples of size 5.

Sampson's truncated approximation of the maximum likelihood estimator performed similarly to the maximum likelihood estimator for smaller sample sizes and almost identically for the larger sample sizes. This is because, as the sample size increases, the probability of the cubic equation having more than one real root goes to zero. Thus, large samples make it easier to use properties of cubic equations to pinpoint the correct MLE root.

Figure 2 shows the first 5,000 samples of each estimator's correlation estimates and the true correlation values for samples of size 5. Notice that the empirical estimate with known variances often lay outside the range $[-1, 1]$. In addition, for small values of the correla-

Table 2: 95% Significance Test bounds for testing if $\rho > 0$ for the non-Bayesian estimators when $\rho = 0$ based on one million simulated data sets.

Sample Size	5	10	50
Sample Correlation Coeff	0.729	0.522	0.233
Emp w/ Known Var	0.731	0.518	0.232
Truncated Emp w/ Known Var	0.731	0.518	0.232
MLE	0.754	0.565	0.241
Sampson's MLE Approx	0.756	0.566	0.241

tion, the empirical estimates, maximum likelihood estimates and Sampson's estimates were extremely variable, spanning most of the interval $[-1, 1]$. The Bayesian estimates showed a closer association overall between the true correlation value and the estimates, especially when the true correlation was small. However, there was some curvature in the tails of the plots for the Bayesian estimators, suggesting that the estimators typically underestimate the magnitude of the correlation when the true correlation is high. This is to be expected, as the Bayesian approach shrinks estimators away from the extremes.

3.2 Hypothesis Tests

Estimating the value of the correlation is important, but often with small sample sizes our interest is not in its actual value but simply in whether or not it is non-zero. We often have knowledge about the sign of the correlation between two variables. Here we consider the case when we are interested in testing if the correlation is positive.

One way of testing this is to look at the confidence bounds of the estimators. A level 0.05 test of whether the true correlation is positive can be derived by generating numerous samples of independent bivariate normal random variables with means equal to zero and variances equal to one, calculating a correlation estimate for each sample, and determining the sample 95% quantile of the correlations. A level 0.05 test then rejects the hypothesis that the correlation is zero in favor of the alternative that it is positive if the estimate obtained is greater than the 95% quantile, i.e. the significance test bound. Table 2 shows the 95% significance test bounds for all non-Bayesian estimators based on one million simulations with $\rho = 0$. For example, for the sample correlation coefficient, the significance test bound for samples of size 5 is 0.73, indicating that about 5% of the samples resulted in an estimated correlation value greater than 0.73.

For Bayesian tests, Jeffreys (1935, 1961) developed ideas based on Bayes factors for test-

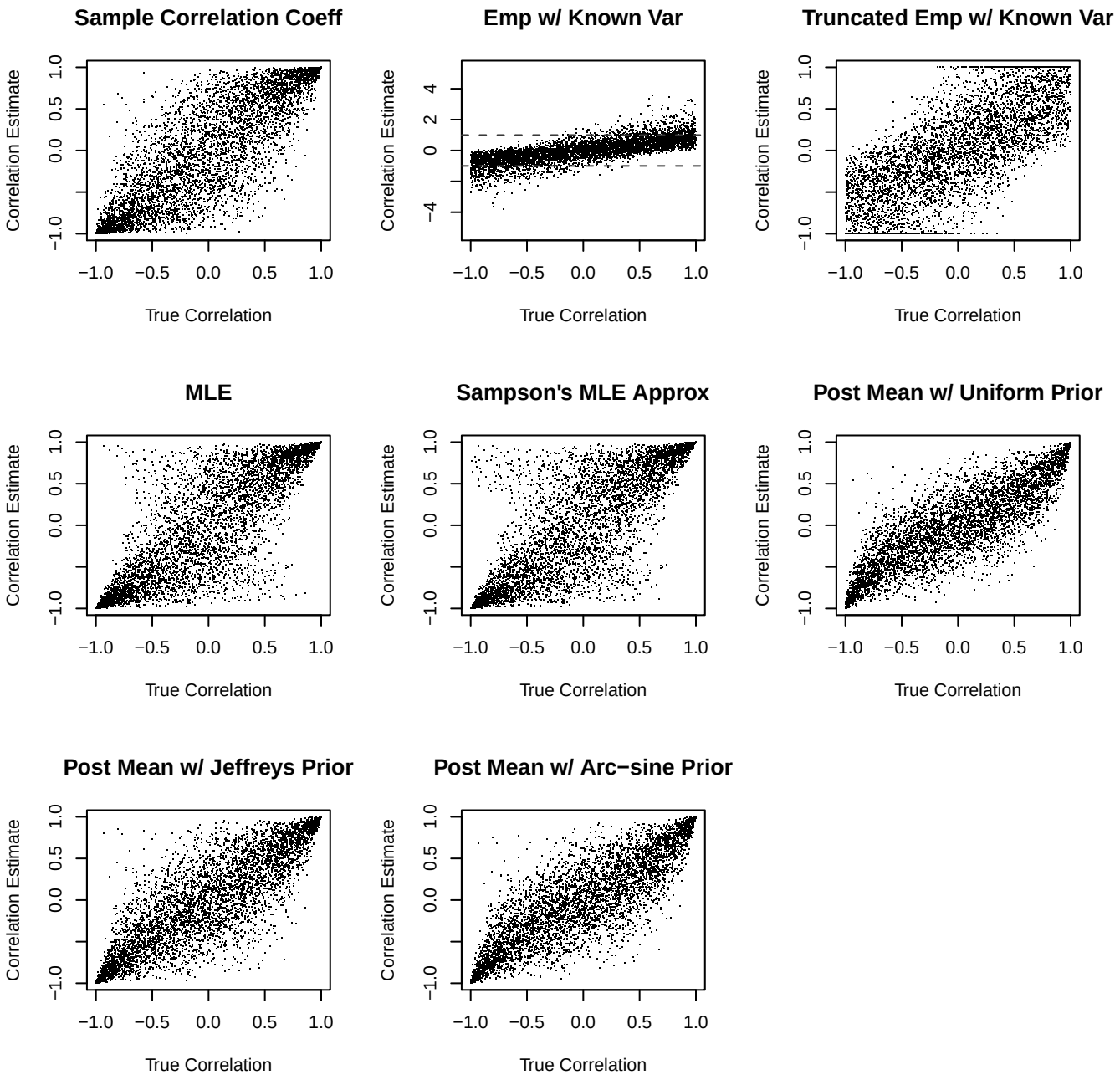

Figure 2: For samples of size 5, the true and estimated correlation values for each estimator is shown above for the first 5,000 samples. The dotted lines in the empirical with known variances plot mark the admissible interval $[-1, 1]$.

Table 3: Value of c such that the Bayes factor has 5% probability of exceeding c if the true value of ρ is 0 (i.e. $P(B_{10} > c|\rho = 0) = 0.05$) based on one million simulated data sets.

Sample Size	5	10	50
Uniform Prior	2.701	2.304	1.238
Arc-sine Prior	2.275	1.715	0.817

ing/deciding between two models; see also Kass and Raftery (1995). A Bayes factor, B_{10}, is the ratio of the probability of the data under the alternative model to the probability of the data under the null model. Equivalently, it is the ratio of the posterior odds for the alternative against the null model, to its prior odds. A test that rejects the null hypothesis when $B_{10} > 1$ minimizes the sum of the probabilities of Type I and Type II errors if the prior odds between the models are equal to one.

However, if we wish to fix the probability of a Type I error at 0.05 for example, we can generate data under the null model and determine the value c such that the probability under the null model that the Bayes factor is greater than c is 0.05. A level 0.05 test is then carried out for the null model against the alternative model by rejecting the null model if the Bayes factor is greater than c. This method was used with $\rho = 0$ as the null hypothesis and $\rho > 0$ as the alternative hypothesis to compare the performance of the Bayesian and non-Bayesian methods when the Type I error is fixed at 0.05. Note that the Bayes factor is

$$B_{10} = \frac{P(X, Y|\rho > 0)}{P(X, Y|\rho = 0)} = \frac{\int_0^1 p(X, Y|\rho)p(\rho|\rho > 0)d\rho}{p(X)p(Y)} = \frac{2\int_0^1 p(X, Y|\rho)p(\rho)d\rho}{p(X)p(Y)} \tag{2}$$

where $p(\rho)$ is one of the three prior distributions for ρ and the denominator is the product of the marginal probabilities assuming $\rho = 0$, or independence. The factor of two in equation (2) is due to the fact all prior distributions are centered at zero. The Bayes factor is undefined for the Jeffreys prior so we do not consider it here forward.

Table 3 shows the values of c obtained for the various prior distributions and sample sizes. We see that as sample size increased, the values of c decreased since the amount of evidence for the null increased. Also, the values of c for the arc-sine prior were much greater than those for the uniform prior, reflecting the fact that the arc-sine prior places more weight on extreme correlation values.

Table 4 shows the power when the true correlation was uniformly generated from various intervals for each of the non-Bayesian significance tests and the tests based on Bayes factors. In samples of size 5 the Bayesian tests had the greatest power over the entire $[0, 1]$ interval and

Table 4: Average power multiplied by 1000 over intervals for ρ when testing $\rho = 0$ vs $\rho > 0$ at the 0.05 significance level based on one million simulated data sets. For the non-Bayesian estimators, the significance test bounds found in Table 2 were used. The Bayesian tests were based on the Bayes factors using the value of c listed in Table 3. The tests with the largest power are shown in bold for each sample size and each correlation interval.

				ρ		
n	Test Based on	[0,1]	[0,.25]	[.25,.50]	[.50,.75]	[.75,1]
5	Sample Correlation Coeff	383	81	187	423	839
	Emp w/ Known Var	288	**89**	**198**	348	517
	Trunc Emp w/ Known Var	288	**89**	**198**	348	517
	MLE	356	69	141	352	862
	Sampson's MLE Approx	355	69	140	350	861
	Uniform Prior	**397**	82	196	**442**	869
	Arc-sine Prior	**397**	81	192	440	**875**
10	Sample Correlation Coeff	529	106	**326**	708	977
	Emp w/ Known Var	441	**110**	302	562	793
	Trunc Emp w/ Known Var	441	**110**	302	562	793
	MLE	505	88	262	682	**990**
	Sampson's MLE Approx	505	88	260	680	**990**
	Uniform Prior	**534**	105	325	722	985
	Arc-sine Prior	**534**	104	323	**723**	987
50	Sample Correlation Coeff	770	**250**	833	998	**1000**
	Emp w/ Known Var	759	245	800	993	**1000**
	Trunc Emp w/ Known Var	759	245	800	993	**1000**
	MLE	768	238	834	**999**	**1000**
	Sampson's MLE Approx	768	238	834	**999**	**1000**
	Uniform Prior	**772**	**250**	**838**	998	**1000**
	Arc-sine Prior	**772**	**250**	**838**	**999**	**1000**

for the most extreme correlation values. For the smaller correlation values, all tests, except possibly those based on the MLE and Sampson's MLE, performed about the same. The tests based on the arc-sine prior and uniform prior performed similarly for all correlation values and sample sizes. As sample size increased, the difference between the powers of the tests based on the MLE and Sampson's MLE and all others decreased.

As mentioned, tests based on the Bayes factor are optimal in that they minimize the sum of the probabilities of Type I and Type II errors when simulating from the prior. For this reason the uniform prior performs best over the entire interval [0,1] for all sample sizes. Table 5 shows the average value of the Type I and Type II error probabilities when the standard rule of rejecting the null hypothesis when the Bayes factor is greater than one is used. This optimal Bayesian method is compared with the significance test bound procedure for the non-Bayesian estimators via this average error measure. The Bayesian tests had the smallest average error

Table 5: Average error probability, [Type I + Type II]/2, when testing if $\rho = 0$ versus $\rho > 0$, multiplied by 1000, based on one million simulated data sets. The error probabilities for the non-Bayesian tests are based on 0.05 level significance tests and the Bayesian test error probabilities are based on rejecting the null hypothesis that $\rho = 0$ if the Bayes factor is greater than 1. The tests with the smallest average error are shown in bold for each sample size and each correlation interval.

				ρ		
n	Test Based on	[0,1]	[0,.25]	[.25,.50]	[.50,.75]	[.75,1]
5	Sample Correlation Coeff	333	485	431	313	106
	Emp w/ Known Var	381	480	426	351	267
	Trunc Emp w/ Known Var	381	480	426	351	267
	MLE	347	490	455	349	94
	Sampson's MLE Approx	348	491	455	350	95
	Uniform Prior	**284**	**460**	**351**	**210**	113
	Arc-sine Prior	289	469	374	225	**88**
10	Sample Correlation Coeff	261	472	362	171	37
	Emp w/ Known Var	304	470	374	244	129
	Trunc Emp w/ Known Var	304	470	374	244	129
	MLE	272	481	394	184	**30**
	Sampson's MLE Approx	273	481	395	185	**30**
	Uniform Prior	**235**	**446**	**291**	**128**	76
	Arc-sine Prior	240	458	319	132	51
50	Sample Correlation Coeff	140	400	109	26	25
	Emp w/ Known Var	145	402	125	29	25
	Trunc Emp w/ Known Var	145	402	125	29	25
	MLE	141	406	108	25	25
	Sampson's MLE Approx	141	406	108	25	25
	Uniform Prior	**139**	**389**	**100**	33	32
	Arc-sine Prior	142	411	114	**21**	**20**

for samples of size 5. The MLE and Sampson's MLE approximation performed very similarly to the Bayesian tests at the extreme correlation values.

At larger sample sizes, the tests performed effectively equally well. For the extreme correlation values with samples of size 50, all tests have essentially 100% power so their average error achieves its lower bound at one-half the Type I error rate. Notice again that the tests based on the arc-sine prior had slightly smaller average error than that assuming a uniform prior at extreme correlation values and that its performance on the entire interval $[0, 1]$ was close to the uniform, which was best.

4 DISCUSSION

We have considered the estimation of the correlation in bivariate normal data when the means and variances are assumed known, with emphasis on the small sample situation. Using simulation, we found that the posterior mean using a uniform prior or an arc-sine prior consistently outperformed several previously proposed empirical and exact and approximate maximum likelihood estimators for small samples. The arc-sine prior performed similarly to the uniform prior for small values of ρ, and better for large values of ρ in small samples. This suggests using the posterior mean with the arc-sine prior for estimation when it is important to identify extreme correlations.

For testing whether the correlation is zero, we carried out a simulation for positive values of ρ within specified intervals, and found that Bayesian tests had smaller average error than the non-Bayesian tests when $n = 5$. With $n = 50$, however, all the tests performed similarly.

Spruill and Gastwirth (1982) derived estimators of the correlation when the data are normal but the variables are contained in separate locations and cannot be combined. Their work combines the data into groups based on the value of one variable to obtain an estimate of the correlation. This differs from the more usual situation considered here where both variables are available in their sampled pairs.

Estimation of the sample correlation coefficient with truncation was investigated by Gajjar and Subrahmaniam (1978). However, it is the underlying distribution that is assumed to be truncated instead of the estimator as here.

Data sets and distributions for which use of the sample correlation coefficient is inappropriate were investigated by Carroll (1961). Norris and Hjelm (1961) considered estimation of correlation when the underlying distribution is not normal, and Farlie (1960) considered it for general bivariate distribution functions. Since we limit ourselves to the bivariate normal distribution, we did not consider these estimators.

Olkin and Pratt (1958) derived unbiased estimates of the correlation in the case when the means are known and the case when all parameters are unknown. This addresses different situations to the one we have considered, where the variances are also assumed known.

Others have considered estimating the correlation in a Bayesian framework for the bivariate normal setting. Berger and Sun (2008) addressed this problem using objective priors whose posterior quantiles match up with the corresponding frequentist quantiles. Ghosh et al. (2010) extended these results by considering a probability matching criterion based on highest posterior density regions and the inversion of test statistics. However, in both cases the focus

was on matching frequentist probabilities rather than estimation accuracy.

Much of the other Bayesian correlation work relates to estimation of covariance matrices. Barnard et al. (2000) discussed prior distributions on covariance matrices by decomposing the covariance matrix into $\Sigma = SRS$ where $S = diag(\sigma)$ is a diagonal matrix of standard deviations and R is the correlation matrix. With this, one can use the prior factorization $p(\sigma, R) = p(\sigma)p(R|\sigma)$ to specify a prior on the covariance matrix. Barnard et al. (2000) suggest some default choices for the prior distribution on R that are independent of σ. Specifically they mention the possibility of placing a uniform distribution on R, $p(R) \propto 1$, where R must be positive definite. The marginal distributions of the individual correlations are then not uniform. Alternatively, for a $(d \times d)$ matrix R one can specify

$$p(R|\nu) \propto |R|^{\frac{1}{2}(\nu-1)(d-1)-1} \left(\prod_{i=1}^{d} |R_{ii}|^{-\nu/2} \right), \qquad \nu \geq d,$$

where R_{ii} is the ith principal submatrix of R. This is the marginal distribution of R when Σ has a standard inverse-Wishart distribution with ν degrees of freedom and results in the following marginal distribution on the pairwise correlations

$$f(r_{ij}|\nu) \propto (1 - r_{ij}^2)^{\frac{\nu-d-1}{2}}, \qquad \text{where } |r_{ij}| \leq 1$$

Uniform marginal distributions for all pairwise correlations comes from the choice $\nu = d + 1$. Note that for $\nu = 2$ and $d = 2$, this prior reduces to the arc-sine prior. This is the boundary case that is the most diffuse prior in the class. Barnard et al. (2000) discussed using these priors for shrinkage estimation of regression coefficients and a general location-scale model for both categorical and continuous variables. Zhang et al. (2006) focused on methods for sampling such correlation matrices.

Liechty et al. (2004) considered a model where all correlations have a common truncated normal prior distribution under the constraint that the resulting correlation matrix be positive definite. They also considered the model where the correlations or observed variables are clustered into groups that share a common mean and variance. Chib and Greenberg (1998) assumed a multivariate truncated normal prior in the context of a multivariate probit model, and Liu and Sun (2000) and Liu (2001) assumed a Jeffreys' prior on R in the context of a multivariate probit and multivariate multiple regression model.

A number of advances have been made with respect to estimation of the covariance matrix treating the variances as unknown, unlike here. Geisser and Cornfield (1963) developed posterior distributions for multivariate normal parameters with an objective prior, and Yang

and Berger (1994) focused on estimation with reference priors. Geisser (1965), Tiwari et al. (1989), and Press and Zellner (1978) derived posterior distributions of the multiple correlation coefficient using the prior from Geisser and Cornfield, an informative beta distribution, and diffuse and natural conjugate priors assuming fixed regressors, respectively. It is possible that some of these ideas regarding prior specification of covariance matrices could be applied to the present setting or be used to extend this work to the multivariate setting.

References

Alkema, L., A. E. Raftery, P. Gerland, S. J. Clark, F. Pelletier, T. Buettner, and G. Heilig (2011). Probabilistic Projections of the Total Fertility Rate for All Countries. *Demography 48*(3), 815–839.

Barnard, J., R. McCulloch, and X. Meng (2000). Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Application to Shrinkage. *Statistics Sinica 10*, 1281–1311.

Berger, J. O. and D. Sun (2008). Objective Priors for the Bivariate Normal Model. *Annals of Statistics 36*, 963–982.

Carroll, J. B. (1961). The Nature of the Data, or How to Choose a Correlation Coefficient. *Psychometrika 26*, 347–372.

Chib, S. and E. Greenberg (1998). Analysis of Multivariate Probit Models. *Biometrika 85*, 347–361.

Farlie, D. J. G. (1960). The Performance of Some Correlation Coefficients for a General Bivariate Distribution. *Biometrika 47*, 307–323.

Gajjar, A. V. and K. Subrahmaniam (1978). On the Sample Correlation Coefficient in the Truncated Bivariate Normal Population. *Communications in Statistics - Simulation and Computation 7*, 455–477.

Geisser, S. (1965). Bayesian Estimation in Multivariate Analysis. *The Annals of Mathematical Statistics 36*, 150–159.

Geisser, S. and J. Cornfield (1963). Posterior Distributions for Multivariate Normal Parameters. *Journal of the Royal Statistical Society. Series B (Methodological) 25*, 368–376.

Ghosh, M., B. Mrkherjee, U. Santra, and D. Kim (2010). Bayesian and Likelihood-based Inference for the Bivariate Normal Correlation Coefficient. *Journal of Statistical Planning and Inference 140*, 1410–1416.

Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. *Proceedings of the Cambridge Philosophy Society 31*, 203–222.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

Kass, R. E. and A. E. Raftery (1995). Bayes Factors. *Journal of the American Statistical Association 90*, 773–795.

Kendall, S. M. and A. Stuart (1979). *The Advanced Theory of Statistics* (4 ed.), Volume 2. MacMillan Publishing Co., Inc.

Liechty, J. C., M. Liechty, and P. Muller (2004). Bayesian Correlation Estimation. *Biometrika 91*, 1–14.

Liu, C. (2001). Discussion: Bayesian Analysis of Multivariate Probit Model. *Journal of Computational and Graphical Statistics 10*, 75–81.

Liu, C. and D. X. Sun (2000). Analysis of Interval Censored Data from Fractionated Experiments using Covariance Adjustments. *Technometrics 42*, 353–365.

Madansky, A. (1958). On the Maximum Likelihood Estimate of the Correlation Coefficient. *Rand Corporation, Santa Monica, Calif. Report No. P-1355*.

McDonald, J. B. (1984). Some Generalized Functions for the Size Distribution of Income. *Econometrica 52*(3), 647–665.

Norris, R. C. and H. F. Hjelm (1961). Nonnormality and Product Moment Correlation. *The Journal of Experimental Education 29*, 261–270.

Olkin, I. and J. W. Pratt (1958). Unbiased Estimation of Certain Correlation Coefficients. *The Annals of Mathematical Statistics 29*, 201–211.

Press, S. J. and A. Zellner (1978). Posterior Distribution for the Multiple Correlation Coefficient with Fixed Regressors. *Journal of Econometrics 8*, 307–321.

Rodgers, J. L. and W. A. Nicewander (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician 42*, 59–66.

Sampson, A. R. (1978). Simple BAN Estimators of Correlations for Certain Multivariate Normal Models with Known Variances. *Journal of the American Statistical Association 73*, 859–862.

Spruill, N. L. and J. L. Gastwirth (1982). On the Estimation of the Correlation Coefficient from Grouped Data. *Journal of the American Statistical Association 77*, 614–620.

Tiwari, R. C., S. Chib, and S. R. Jammalamadaka (1989). Bayes Estimation of the Multiple Correlation Coefficient. *Communications in Statistics - Theory and Methods 18*, 1401–1413.

Yang, R. and J. O. Berger (1994). Posterior Distributions for Multivariate Normal Parameters. *Annals of Statistics 22*, 1195–1211.

Zhang, X., W. J. Boscardin, and T. R. Belin (2006). Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables. *Journal of Computational and Graphical Statistics 15*, 880–896.